Департамент образования и науки города Москвы
Государственное автономное образовательное учреждение
высшего образования города Москвы
«Московский городской педагогический университет»
Институт иностранных языков
Кафедры языкознания и переводоведения

Басанин Кирилл Олегович

Исследовательский потенциал систем Big Data (на материале анализа предикатов *seem, appear*)

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Направление подготовки – 45.03.02 «Лингвистика»
Профиль подготовки – Перевод и переводоведение – английский язык
(очная форма обучения)

Руководитель ВКР:
кандидат педагогических наук, доцент
Гулиянц Светлана Борисовна

Рецензент:
кандидат филологических наук, доцент
Трухановская Наталья Сергеевна

Зав. выпускающей кафедрой:
доктор филологических наук, профессор
Сулейманова Ольга Аркадьевна

Москва
2020

Аннотация. В представленной работе анализируется возможность применения системы Big Data для исследования предикатов, на примере *to seem* и *to appear*. Изучая применение больших данных в лингвистике, стоит отметить, что потенциал этой системы в данном поле на сегодняшний день изучен слабо, что определяет актуальность выбранной темы. Материалом диссертационного исследования послужили системы Big Data, а также словосочетания, предложения, тексты разнообразных тематик и дискурсов с глаголами *to seem* и *to appear*, отобранные с их помощью.

Исследование показало, что развитие систем Big Data, включающих корпусные данные и данные разнообразных поисковых систем, открывает новые возможности и позволяет ставить и решать различные исследовательские задачи оптимальным образом.

Ключевые слова: Big Data; предикаты to seem и to appear; лингвистика, потенциал систем.

Abstract. The aim of the research is to analyze the possibility of Big Data systems usage for predicate research, for example, such as: *to seem, to appear*.

Considering Big Data usage in linguistics, it is worth mentioning that the potential of the systems in this field is currently underexplored, which determines the relevance of the research topic. Phrases, sentences, texts on various topics and discourses with *to seem* and *to appear*, selected with the help of Big Data systems, are used as study material.

The research suggests that the development of Big Data systems, including corpus data and data from various searching systems, provides new opportunities and allows us to set and solve different research goals in optimal way.
Keywords: Big Data; predicates *to seem* and *to appear*; linguistics; systems potential.

СОДЕРЖАНИЕ

Введение

В настоящей работе представлено исследование потенциала систем Big Data на материале анализа предикатов (*seem, appear*).

Актуальность исследования. Многие эксперты

разделяют мнение о том, что ускорение роста данных стало объективной реальностью. Цифровые технологии проникли в жизнь современного человека. Источники, генерирующие огромные объемы информации, сегодня встречаются повсеместно: от смартфонов и социальных сетей до данных с многочисленных устройств измерения. Объем данных по различным аспектам жизни растет, и в то же время растут возможности хранения информации.

Уже в начале века отмечалось, что на устройствах хранения информации данные накапливаются слишком быстро, и скорость их накопления превышает скорость обработки [Gray 2004]. Сегодня доступной емкости на носителях уже не хватает для хранения объемов данных из многочисленных источников информации. Одновременно с этим разница между количеством данных, которое может быть обработано, и возможностями хранения также быстро растет [Tauheed 2013].

Системы Big Data или *большие данные* это не только хранение больших объемов информации. Они также предоставляют инструменты для решения такой сложной задачи, как анализ огромного объема разрозненных и слабоструктурированных данных. Большой интерес к теме вызван широким распространением информационных технологий в общем, и систем Big Data в частности. Исследование и применение больших данных в различных областях научного знания показывает возможность использования этих систем и в науке о языке.

В.П. Захаровым анализируются поисковые системы Интернета как инструменты лингвистических исследований,

описываются экспериментальные исследования устойчивости словосочетаний и способы их количественной оценки в синхронии и диахронии. Он считает, что количественная оценка лингвистических данных и математические методы их обработки представляют большой интерес для лингвистов [Захаров 2003,2015]. В работе О.В. Нагель анализируется языковой материал корпуса с точки зрения потенциала обучения, также рассматривается применение методов прикладной лингвистики в методике преподавании иностранного языка. Корпусные методы изучения иностранного языка сочетают в себе целый ряд преимуществ: аутентичность, междисциплинарность, адаптацию к конкретным целевым группам и задачам. По мнению О.В. Нагель, это делает их эффективным дополнением к традиционным образовательным технологиям [Нагель 2008]. Однако потенциал систем Big Data в области лингвистики на данный момент раскрыт не полностью, что и определяет актуальность выбранной темы.

Новизна исследования заключается в: использовании традиционных для лингвистики методов компонентного анализа в сочетании с интерпретацией данных, полученных из национальных корпусов текстов; получении новых и уточнении существующих описаний семантики данных языковых единиц; анализе возможностей применения различных систем Big Data для исследований в области предикатов.

Теоретической базой послужили отдельные положения, представленные в трудах специалистов в области: предикатов (Е.В. Ильчук, Ю.Д. Апресян, О.Н.

Селиверстова, Л.В. Щерба, З. Вендлера и др.); систем Big Data (В.П. Захаров, О.В. Нагель, А.С. Большаков, О.В. Журенков, Д. Бойд, К. Кроуфорд, О.А. Сулейманова, В.В. Демченко и др.).

Объект исследования: исследовательский потенциал систем Big Data.

Предмет исследования: потенциал систем Big Data для исследования предикатов *to seem* и *to appear*.

Цель исследования: выявить потенциал систем Big Data для исследования предикатов.

Задачи исследования:

1) описать, что такое *большие данные* (Big Data);

2) провести анализ лингвистических исследований, в которых использовались системы Big Data, выявить их специфику и особенности;

3) описать семантические особенности предикатов *to seem* и *to appear*;

4) применить системы Big Data для исследования предикатов *to seem, to appear;*

5) проанализировать полученные в ходе исследования данные и сделать выводы.

Материалом исследования стали 420 примеров употребления английских предикатов *to seem* и *to appear* из Корпуса современного американского английского языка, Британского национального корпуса текстов, а также результаты десяти анкетирований носителей английского языка. В работе использовались сервисы Google Books Ngram Viewer, Google Trends, Google docs – для сбора анкет информантов, SentiStrength, электронный тезаурус WordNet

и электронные словари: Большой Оксфордский Словарь, Словарь Вебстера, Кембриджский словарь, Этимологический онлайн словарь и другие.

Методы исследования: на первом этапе была изучена история возникновения систем Big Data, описано их использование в различных областях. На втором этапе исследования проанализированы работы, в которых представлено, что такое предикативность, описана разница в семантике предикатов *to seem* и *to appear.* На третьем на примере предикатов *to seem* и *to appear* изучались возможности систем больших данных для проведения лингвистических исследований (уточнения семантического значения предикатов, определения частотности их употребления, поиска коллокаций, подтверждения или опровержения высказанных исследователями предположений, определение эмоциональной окраски). Для уточнения семантики предикатов применяется триангуляционный подход, включающий в себя составление запросов в поисковых системах, корпусный и семантический эксперименты. На четвертом этапе полученные данные были обобщены и описаны.

Теоретическая значимость исследования заключается в описании исследовательского потенциала систем Big Data на материале анализа предикатов. Полученные в ходе исследования результаты, могут стать основой для дальнейших научных работ в сфере предикатов, а также использования информационных технологий в лингвистических исследованиях.

Практическая значимость исследования состоит в том, что полученные результаты могут быть использованы в преподавании практических и теоретических дисциплин, например, «Практический курс первого иностранного языка», «Теория перевода», «Языкознание», «Теоретическая грамматика»). Они будут полезны при составлении учебных и справочных пособий по теории и практике перевода, при написании дипломных и курсовых работ, статей по сходной тематике.

Структура исследования. Выпускная квалификационная работа состоит из введения, двух глав, заключения, списка использованной литературы и приложений.

Во введении обосновывается актуальность исследования, определяются его объект, цели и задачи, раскрываются практическая и теоретическая значимость, перечисляются используемые методы исследования.

В первой главе «Исследовательский потенциал систем Big Data» анализируется, что такое Big Data, а также описывается потенциал и опыт применения этих систем в научных, лингвистических и лингвокогнитивных исследованиях.

Во второй главе «Использование систем Big Data для анализа предикатов *to seem* и *to appear*» рассматривается понимание предикативности в лингвистике, описываются и анализируются результаты применения систем Big Data для исследования предикатов *to seem*, *to appear*.

В заключении сформулированы основные выводы по результатам исследования.

Библиография представляет собой полный перечень используемой литературы и состоит из 86 наименований, из них 22 на иностранном языке.

Приложения включают дополнительный материал к тексту выпускной квалификационной работы. В них представлены методики анализа массива данных, список и описание наиболее популярных лингвистических корпусов, результаты поиска в корпусных менеджерах и др.

Апробация исследования. Основные теоретические и практические положения данной работы нашли свое отражение в статьях:

- Басанин К.О. Исследовательский потенциал систем Big Data // Электронный сборник статей по материалам LXXXI студенческой международной научно-практической конференции. – Новосибирск: Изд. ООО «СибАК». – 2019. – № 9 (81). – С. 12-19.
- Басанин К.О. Исследовательский потенциал систем Big Data в лингвистике и методике преподавания иностранного языка // Сборник статей по материалам конференции «Иностранный язык. Методические вопросы подготовки конкурентоспособного выпускника» (в печати).

Исследование также было представлено на 54 Международном лингвистическом коллоквиуме в форме постерного доклада.

Глава I. Исследовательский потенциал систем Big Data

В первой главе рассматриваются системы Big Data, даются их характеристики, описываются особенности и сфера применения. На сегодняшний день Big Data – это работающий набор технологий, которые используются в тех сферах деятельности человека, где требуется собирать и анализировать огромные массивы информации. Описывается использование и потенциал Big Data в физике, климатологии, генетике и т.д. Более глубоко рассматриваются предпосылки возникновения корпусной лингвистики, существующие подходы к изучению этой области научного знания, а также приводятся типология лингвистических корпусов и возможности конкорданса (анализ частотности слов в языке, определение значения слова в национальном лингвистическом корпусе по его контексту, изучение словоупотребления и грамматических сторон языка).

1.1 Системы Big Data – понимание и применение

В современном мире цифровых технологий постоянно увеличивающиеся объемы информации требуют новых решений для организации ее анализа и хранения. Источником данных может служить непрерывный поток

информации от сообщений из соцсетей до информации со всевозможных устройств измерения, например, датчиков находящихся в океане. Так, даже с ограничением количества символов в сообщении, социальной сетью Твиттер ежедневно генерируется восемь терабайт данных. Сбор всех подобных данных для последующей обработки означает, что возникнет потребность хранения тысяч петабайт информации. При изменении состава такого рода данных, к примеру, при запуске новых сервисов, установке улучшенных датчиков или создании новых маркетинговых кампаний, возникают дополнительные трудности. Повсеместное распространение вышеперечисленных технологий и абсолютно новых способов применения разнообразных веб-сервисов и устройств стало началом проникновения больших данных почти во все сферы человеческой деятельности.

Согласно Кэмбриджскому словарю: данные – это информация, особенно факты и числа, собранные для последующего использования при принятии решений (часто в электронной форме), пригодная для хранения и использования компьютером [Cambridge dictionary].

Впервые термин Big Data или *большие данные* был употреблен в 2008 году Клиффордом Линчем, редактором журнала «Nature». В статье он рассказывал о многообразии данных и феномене бурного роста их количества. Большими данными обычно называют громадные массивы информации, неопределенные и неоднородные по своей структуре. Большие данные это не просто неструктурированная информация, они имеют определенную структуру. Поскольку данные поступают из разнообразных источников и

представляют собой отличные друг от друга или вовсе неизвестные сведения, их структура довольно сложная [Кравченко, Крюкова 2016].

Согласно сервису Google Trends активный рост употребления словосочетания Big Data начался с 2011 года (см. Рис. 1).

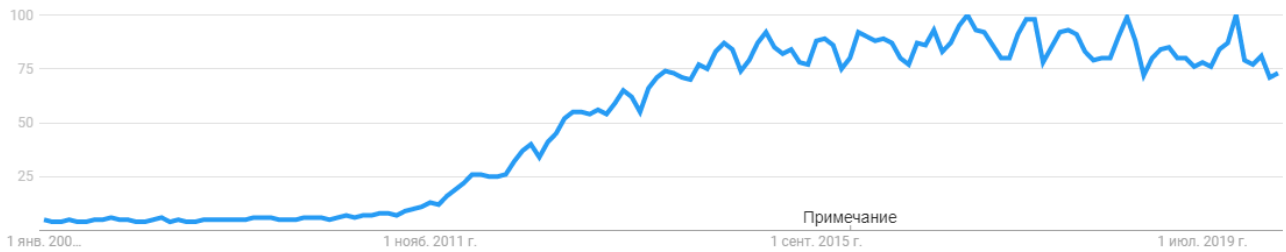


Рис.1.Частотность употребления словосочетания Big Data

Обычно большие данные описываются при помощи следующих характеристик:

- Volume – объем. Эта характеристика является самым важным и наиболее заметным параметром аналитических процессов на основе больших данных. Слово *большой* показывает, что 90% всех мировых данных было получено в течение последних десяти лет, благодаря взрывному росту компьютерных технологий.
- Velocity – быстрота реакции. Скорость принятия решения – это время между моментом получения определенных данных и моментом принятия решения, исходя из полученной информации. Это основной фактор, определяющий эффективность структуры больших данных. Новые технологии способны обрабатывать огромные объемы данных в реальном или

почти реальном времени. Благодаря этому компании могут адекватно и быстро реагировать на любые изменения.

- Variety – разнообразие форм. Структуры Big Data, содержащие разнородные и неупорядоченные данные, могут быть представлены в виде текста, информации, полученной от датчиков, по запросам видео- и аудиоданных, маршрутов навигации. Они также могут содержать данные, которые требуют времени и соответствующих технологий для преобразования в форму, доступную для обработки и анализа [Новиков 2013].

Помимо вышеперечисленных характеристик, в некоторых источниках выделяются еще две:

- Veracity – достоверность (аутентичность). Анализ данных является инструментом для оценки их надежности, наиболее важной характеристикой, которая может служить основанием для принятия важных решений. Однако большие объемы данных могут быть ненадежными из-за отсутствия связи между информационными элементами, их неполноценности или скрытого состояния. Современные информационные системы должны иметь возможность различать, оценивать и классифицировать различные массивы данных, чтобы поддерживать определенный уровень надежности.
- Variability – изменчивость. Противоречивость наборов данных может препятствовать их обработке и управлению ими [Харин 2017].

Однако, по мнению некоторых исследователей, система имеет свои слабые стороны. Д. Бойд и К. Кроуф выделили следующие несовершенства систем Big Data:

1) ошибочное ощущение объективности, поскольку в центре обработки находится субъективная по своей природе интерпретация найденных закономерностей;

2) большое количество данных может быть минусом, поскольку не все они касаются изучаемой проблемы. Также огромное количество информации может быть лишено корреляций;

3) данные могут потерять смысл, если будут рассматриваться вне контекста;

4) нерелевантная информация искажает смысл данных;

5) ошибки репрезентативности и измерения – в таком случае программа задает неправильный алгоритм, вследствие чего данные будут невалидными [Boyd, Crawford 2012].

Помимо технических проблем существует ряд этических. Их решение является более проблематичным. Этические проблемы можно объединить в следующие группы:

1) закрытость данных для некоторых слоев населения;

2) отслеживание информации о пользователе в сети, как в реальном времени, так и его истории;

3) нарушение приватности [Одинцов 2017].

Принимая во внимание эти этические проблемы, существует точка зрения, которая подразумевает общедоступность больших данных, что в свою очередь благоприятно влияет на увеличение осведомленности в сфере

информационных технологий, среди обычных пользователей и работников научной сферы.

Взаимодействие систем Big Data с качественными методами исследования общественного мнения является наиболее эффективным на данный момент. Поскольку анализируя только статистические данные нельзя точно выявить, какая переменная является зависимой, а какая нет. Также возможно наличие третьего фактора, который может оказывать влияние на два остальных параметра. Поэтому к тем данным, которые можно получить через анализ Big Data необходимо качественное дополнение [Радченко, Николаев 2018].

Технологии на основе больших данных на сегодняшний день используются повсеместно. Перечислим некоторые сферы деятельности человека, в которых они нашли применение. Big Data в торговой сфере – это сбор данных о предпочтениях потребителей: анализ опросов, совершенных покупок, обзоров товаров в Интернет-магазинах, телефонных разговоров с центрами обработки вызовов клиентов. Собранная информация помогает компаниям понять, почему одни продукты пользуются спросом, а другие нет.

В сфере коммунальных услуг большие данные обеспечивают анализ информации, поступающей от приборов учета, например, от различных счетчиков. Этот метод сбора информации может уменьшить человеческий фактор и, как следствие, количество ошибок. Это также облегчает анализ большого количества поступающей информации.

В телекоммуникациях – это вся внутренняя информация с подключенных к сети устройств. Данные геолокации,

история поиска и посещения различных сайтов сети Интернет. В случае необходимости, весь трафик может быть проанализирован, как при запросе от самого пользователя, так и при запросе от правоохранительных органов. У пользователей смартфонов IPhone и Android есть приложения, которые используют технологию распознавания лиц для различных задач. При использовании такой системы на большом предприятии, можно оптимизировать контроль сотрудников.

Большие данные нашли применение в автомобильной индустрии. Автомобильные бренды активно используют системы Big Data. С помощью данных, полученных с тестовых прототипов, они помогают выявить проблемные области конструкции еще на стадии проекта. В течение дальнейшей эксплуатации, благодаря информации о неисправностях от владельцев и СТО по всему миру, можно быстро исправить проявившиеся дефекты конструкции.

Применение технологий больших данных в финансовой сфере дает банкам возможность проводить собственный анализ кредитных рейтингов для существующих клиентов, используя широкий спектр данных, в том числе данных о чеках, сбережениях, кредитных картах, ипотеках и инвестициях.

Реализация технологий Big Data в медицинской сфере дает медикам возможность собирать данные для более тщательного изучения болезни, чтобы увидеть, какое лечение более эффективно, выявить закономерности и получить другую важную информацию, которая может помочь пациентам.

Системы Big Data также применяются полицией. Департамент полиции Лос-Анджелеса использует систему собственной разработки. Она анализирует отчеты о преступлениях за конкретный период времени и с помощью определенных алгоритмов вычисляет места с наибольшей вероятностью совершения правонарушений.

Большие данные используются Интернет-корпорациями, например, Яндекс. Компания разработала алгоритмы для определения целевой аудитории для трансляции рекламы, мониторинга ситуации с трафиком, оптимизации выдачи результатов поиска и музыкальных рекомендаций.

Несмотря на сравнительно небольшой период функционирования систем Big Data, уже существуют оценки их эффективности на основе реальных примеров. Эксперты в области энергетики утверждают, что технологии Big Data могут повысить эффективность генераторов на 95-98%, благодаря более правильному распределению мощностей. Структура здравоохранения США может экономить до 300 миллиардов долларов.

В России доступны программы от ведущих производителей: Cisco, HP, IBM, Microsoft, Oracle, Apache. Однако на данный момент проектов по реализации немного. Российский рынок только начинает использовать преимущества данной технологии, но большинство аналитиков прогнозируют взрывной рост технологий Big Data.

Корпорация EMC в 2013 году провела опрос среди российских компаний. В ходе исследования было установлено, что применение систем Big Data существенно

улучшает процессы принятия решений, упрощает управление рисками и повышает конкурентоспособность компании:

- по мнению 70% респондентов, анализ данных их компании поможет принимать более обоснованные решения;
- 45% опрошенных уверены, что руководство использует результаты анализа Big Data для принятия важных бизнес решений;
- 55% респондентов согласны, что технологии анализа Big Data помогут в выявлении и предотвращении кибератак [Иванов, Вампилова 2014].

Существует много различных методов анализа массивов данных, основанных на инструментах, заимствованных из статистики и информатики. Постоянно ведется работа над созданием новых и совершенствованием существующих методов. Некоторые из них могут быть успешно использованы не только для больших данных, но и для небольших массивов информации. Однако точность полученных данных напрямую зависит от анализируемого массива – чем он разнообразнее и объемнее, тем точнее будут полученные на выходе данные. В списке ниже перечислены наиболее популярные методы из разных отраслей.

Association Rule Learning – методики для выявления взаимосвязей между переменными величинами в больших массивах данных.

Data Fusion and Data Integration – набор методик, с помощью которого комментарии пользователей социальных сетей сопоставляются с результатами продаж в режиме

реального времени.

Machine Learning или искусственный интеллект – создание алгоритмов самообучения на основе анализа эмпирических данных.

Natural Language Processing (**NLP**) – набор заимствованных из информатики и лингвистики методик распознавания естественного языка человека.

Полный список и описание всех методик можно найти в приложении (см. Приложение 1).

Человек сталкивается с большими данными каждый день. На современном этапе развития системы Big Data представляет собой набор технологий, которые могут быть полезны во многих областях человеческой жизни. Большие данные помогают собирать и анализировать огромные объемы информации, которая активно используется представителями автомобильной промышленности, Интернет-корпорациями, полицией, работниками системы здравоохранения и др.

1.2 Потенциал систем Big Data в научных исследованиях

Технологии анализа информационных массивов ускоряют исследования в различных областях: от астрофизики и генетики до социологии и лингвистики.

Физика как наука имела дело с огромными объемами данных еще до того как само понятие Big Data было сформулировано. В 1960-х годах представители субъядерной физики, стремясь исследовать частицы, из которых состоит

вселенная, впервые стали использовать компьютеры для сбора, моделирования и анализа данных. Усилия по обмену и обработке количества данных, сгенерированных в CERN – крупнейшей лаборатории физики элементарных частиц в мире – в конечном итоге привели к созданию Всемирной паутины (в 1989 году).

Сегодня Большой Адронный Коллайдер (LHC) – главный ускоритель частиц CERN – производит 1 миллиард столкновений частиц в секунду. Такие испытания дают представление о фундаментальных составляющих вселенной, генерируя более 30 петабайт данных в год. Эти данные доступны сообществу тысяч физиков по всему миру практически в режиме реального времени через крупнейшую в мире распределенную вычислительную инфраструктуру, известную как Worldwide LHC Computing Grid. Хотя эта система в настоящее время хорошо работает с CERN, будущие обновления и потенциальные преемники LHC будут производить на несколько порядков больше данных. Перед учеными стоит задача справиться с таким потоком данных.

Эта проблема касается и других устройств, генерирующих и анализирующих большие данные. Например, новый телескоп Square Kilometre Array, цель которого – ответить на фундаментальные вопросы о происхождении и эволюции Вселенной. Ожидается, что он будет производить около 15 терабайт информации за одну ночь. Одной из задач наблюдения неба является поиск гравитационных линз. До недавнего времени большая часть линз была открыта случайно. Астроном Карло Энрико Петрильо и его коллеги обучили искусственный интеллекту

искать те самые гравитационные линзы, и результаты превзошли ожидания. Очень внимательный и эффективный исследователь может просматривать около тысячи снимков в час, нужный объект обнаруживается с частотой приблизительно один раз в 30 000 галактик. То есть человек, работающий неделю без сна и отдыха, может найти около 5 или 6 линз за всю свою жизнь. Нейронная сеть команды Petrillo всего за 20 минут анализирует 21 789 снимков, используя только один старый компьютер. Пока точность компьютерного интеллекта не абсолютна. Но из 761 потенциальных гравитационных линз, выбранных компьютером, людьми были отмечены 56 наиболее вероятных. Считается, что до трети из них могут оказаться гравитационными линзами, то есть при работе в безостановочном режиме нейронная сеть позволит находить одну линзу в минуту, несмотря на то, что в прошлом за 40 лет ученые обнаружили всего чуть более сотни таких объектов [Keating 2015].

Ожидается, что благодаря тесному сотрудничеству с промышленностью и открытому обмену знаниями эти и многие другие крупные научные проекты приведут к развитию компьютерных технологий. Результатами сотрудничества станут более совершенные суперкомпьютеры и методы анализа больших данных, более энергоэффективные вычислительные методы.

Большие данные не являются чем-то новым для аэрокосмической индустрии. Датчики для сбора телеметрии самолетов использовались уже в эпоху двоичных данных и собирали такую информацию как скорость, высота, тангаж и

т.д. С помощью современных датчиков, на основе существующих повреждений, можно предсказать их развитие и соответствующим образом менять интервалы технического обслуживания. Авиационные власти проделали большую работу по использованию такого рода данных и информации о катастрофах для повсеместного повышения стандартов безопасности.

Высокопроизводительные вычисления используются для достижения более высокой степени точности при создании аэродинамических и гидродинамических моделей. Marussia, успешная гоночная команда Формулы 1, использует суперкомпьютер для дополнительной точности в проектировании своих автомобилей. Так команда может работать с бюджетом в тридцать миллионов фунтов, а не с бюджетом в 150 миллионов фунтов, как большинство команд F1 [Scott J 2012].

Прогностическая аналитика может быть полезна в качестве инструмента для расчета физики мягких тел. Исследовательская группа по информатике в Университете Брауна использовала компьютерное моделирование физики мягких тел и прогностической аналитики для определения поведения диартродиальных суставов (суставов, которые имеют широкий диапазон движения) в теле человека [Marai 2007].

В области генетики работа ученых с системами Big Data связана с расшифровкой человеческих генов, предсказанием болезней или склонности к спорту, созданием виртуальных (и после этого реальных) моделей растений. Технологические достижения позволили ученым быстро создавать, хранить и

анализировать данные, которые до недавнего времени собирались годами. Например, Национальные институты здравоохранения запустили проект «Big Data to Knowledge» и «Precision Medicine Initiative» с целью разработки генетически ориентированного лечения в рамках индивидуализированной медицины для улучшенной профилактики, раннего выявления и лечения распространенных сложных заболеваний. Планируется, что это будет реализовано путем сбора и объединения электронных медицинских карт и данных около миллиона американцев.

Редактирование генов – это набор технологий, которые позволяют ученым изменять ДНК организма путем добавления, удаления или изменения генетического материала в определенных местах генома. Существует несколько подходов к редактированию генов, один из них – CRISPR. Используя Big Data, большие вычислительные системы и CRISPR, Ричард Кандасами объединил современные технологии с более традиционным и воспроизвел принцип реакции иммунной системы на присутствие вируса в клетке [Midling 2017].

В настоящее время климатические исследования являются одной из приоритетных областей, поскольку изменение климата влияет на общество в целом. Существует потребность в изучении изменчивости погоды и климата с очень высокой точностью. Сегодня используются сложные модели для прогнозирования изменчивости погоды и климата, генерируя огромное количество многомерных цифровых данных. Поэтому потребность в

высокопроизводительных и облачных вычислениях для проведения исследований климата возрастает.

Объединение климатологии и прогностической аналитики привело к появлению совершенно новой области, которую некоторые называют погодной аналитикой. Эта сфера характеризуется сбором больших данных, связанных с погодой и климатом, а затем использованием прогнозной аналитики для прогнозирования будущих погодных условий или урожая. Хотя область находится только в начале своего развития, предприятия уже показывают свою заинтересованность.

Большие данные также играют важную роль в гуманитарных науках. Например, команда психологов из Центра позитивной психологии при Университете Пенсильвании во главе с Мартином Селигманом провела несколько экспериментов по анализу контента в сервисах Facebook и Twitter. В исследовании ученые проанализировали 148 миллионов твитов, чтобы предсказать смертность от сердечных заболеваний в округе США. Слова, связанные с гневом и негативным отношением, оказались факторами риска. Более того, этот прогноз оказался более точным, чем тот, который был сделан на основе 10 обычных факторов риска, таких как курение или диабет. Позднее, с использованием этой технологии, была составлена карта рисков, где по округам были отмечены уровни благосостояния, депрессии, доверия и других состояний. Анализ текстовых сообщений в Интернете требует тесной работы с лингвистами, предоставляя им много данных для

анализа: изменение языка и стиля общения, использование сленга и отмирание понятий.

А.С. Большаков и О.В. Журенков разработали методику сбора и анализа данных из открытых источников World Wide Web. Инструментом поиска стал веб-сервис Яндекс: поиск по блогам. В работе исследовалось использование неструктурированных текстовых данных Big Data для разработки новых подходов к оценке внутриполитической и экономической ситуации в государствах. Поиск по ключевым словам осуществлялся сразу на двух языках: русском и английском. Следующим этапом стала визуализация полученных результатов. По результатам проведенного исследования, авторы делают вывод, что возможность извлечения данных из информационного пространства WWW позволяет проводить исследования на новом уровне, а полученные в работе результаты могут быть применены для поиска альтернативных подходов в гуманитарных исследованиях [Большаков, Журенков 2017].

Большие данные оказывают влияние на все области науки: от физики и генетики до социальных и гуманитарных. Системы Big Data предлагают широкие возможности для исследований и в смежных с ними науках. Чтобы использовать большие данные в полном объеме, ученым необходимо придерживаться гибкого подхода к Big Data и быть в курсе последних достижений в инструментах их анализа.

1.3 Применение систем Big Data в лингвистике

Использование систем Big Data в лингвистике – это прежде всего сфера корпусной лингвистики. Она возникла в 60-х годах двадцатого века и основывалась главным образом на материалах английского языка. Через некоторое время корпуса стали появляться и на основе других языков. Первый электронный корпус был разработан в 1963 году учеными Г. Кучерой и В.Н. Фрэнсисом из Университета Брауна [Иванов 2014]. Он насчитывал около миллиона слов и состоял из текстов популярных жанров англоязычной литературы. Корпус был дополнен приложением, которое содержало некоторые статистические распределения, частотный и алфавитно-частотный указатели [Козлова 2013]. В широком смысле корпус это комбинация текстов одного и / или нескольких языков, которые связаны определенными параметрами. К началу 90-х годов двадцатого века одновременно с формированием понятийного аппарата корпусная лингвистика стала отдельной областью лингвистического знания.

Под корпусной лингвистикой понимается раздел языкознания, который занимается поиском закономерностей функционирования языка при помощи лингвистического корпуса и анализа [Сысоев 2011]. Ее основной характеристикой (в сравнении с традиционной) является изучение языка, а не речи. Целью – описание языка в том виде, в котором он проявляет себя в речи, представленной в виде специально подобранного корпуса текстов. В ней предпочтительны квантативные (т.е. количественные методы), в то время как традиционная лингвистика

предпочитает квалитивные (т.е. качественные методы). Квалитативный анализ выявляет общие закономерностей, но не предоставляет их точное количественное описание. Особенностью квантитативного анализа является интерпретация статистических закономерностей и большой объем выборки. Квантификация данных создает условия для использования средств математического анализа и для работы над ними. Она включает в себя анализ частотного распределения, корреляций между переменными, ассоциаций, сопряженности и кластерный анализ. А.Н. Баранов считает, что при определении частей контент-анализа и их распознавании в тексте не всегда можно добиться полной объективности. В то время как при экспликации и обработке данных обеспечить следование строгим стандартам возможно почти всегда [Баранов 2001].

В корпусной лингвистике работа с лингвистическими данными проводится в том виде, в каком они встречались в контексте; традиционная лингвистика предпочитает искусственные примеры из изолированных от текста словоупотреблений. В первой предпочтительнее применение индуктивных методов обработки словесного материала, вторая опирается на дедуктивные методы обработки.

Текст в корпусной лингвистике рассматривается как некая физическая сущность, в традиционной – как абстракция. Основное внимание уделяется форме, а не содержанию. Тексты рассматриваются в глобальной основе, а не локальной перспективе. При этом традиционная лингвистика предпочитает логические рассуждения,

корпусная использует вероятностные методы и статистику для первичной обработки материала.

Д. Синклар называет корпус совокупностью неотредактированных естественных текстов, отобранных по определенному критерию для наиболее полного представления языка или его вариаций [Sinclair 1991]. Отдельно им выделяется основной принцип отбора текстов для корпуса – естественность. Важно, чтобы язык в тексте был близок к тому виду, в котором он существует в повседневной речи. В дальнейшем понятие корпуса конкретизируется: «Корпус – это не просто речь носителя языка, а нечто, созданное исследователем. Это характеристики речи, как правило, разных пользователей, предназначенные для изучения и формулирования дальнейших выводов о типичном использовании языка» [Stubbs 2001]. Сегодня корпус – большой, представленный в электронном виде, структурированный и размеченный массив языковых данных, предназначенных для решения определенных лингвистических задач [Захаров 2011].

Н.В. Владимов определяет лингвистический корпус следующим образом: массив, отобранных по определенным характеристикам и собранных в единую систему текстов. Они могут быть письменными или являться транскриптами телепередач и радиопередач. Корпус может состоять из текстов на определенном языке, одного или нескольких авторов, разных жанров, относящихся к особому промежутку времени. Таким образом, цель создания корпуса влияет на его состав [Владимов 2005].

Нельсон Фрэнсис выделяет 4 основных признака лингвистического корпуса, среди которых: machine readable form – обязательное расположение на машинном носителе; sampling – определенный стандарт для размещения словесного материала на электронном носителе, позволяющий применять программы для его обработки, отбора и поиска; representativeness – набор требований, по которым создавался корпус; design criteria – окончательный размер [Francis 1991].

Во всех представленных определениях авторами делается акцент на таких особенностях как: тексты представлены в электронном виде, языковые данные имеют особую разметку, и есть возможность распределения языкового материала по принадлежности к жанру, тематике, году создания и т.д.

Существующие корпуса могут быть разделены на три категории: со свободным доступом, частично свободным доступом и закрытые, коммерческие. Первый тип включает в себя достаточно малое количество существующих корпусов. Примером является Национальный корпус русского языка. Ко второму типу относится большая часть существующих на данный момент корпусов. Корпус современного американского английского языка и Британский национальный корпус позволяют незарегистрированным пользователям совершать только 50 поисковых запросов. К третьему типу относится Банк английского языка, предоставляющий возможность бесплатного пользования в течение первого месяца. В приложении представлен полный

список и описание наиболее популярных лингвистических корпусов (см. Приложение 2).

Следующей отличительной особенностью лингвистического корпуса является наличие разметки. Под ней понимается приписывание текстам и их компонентам специальных меток. Они могут быть: лингвистическими, то есть описывающими лексические, грамматические и прочие характеристики; экстралингвистическими, внешними и структурными [Захаров 2015]. Разметка также бывает морфологической (осуществляется с помощью специальных программ автоматического морфологического анализа) и синтаксической (подразумевает указание синтаксической структуры для каждого предложения).

Создание целиком размеченного корпуса – это сложный и трудоемкий процесс, который требует усилий большого количества лингвистов. По этой причине создание большого текстового корпуса обычно осуществляется исследовательскими группами в специализированных институтах для дальнейшего использования при решении различных прикладных и научных задач [McEnery, Gabrielatos 2006].

Еще одним признаком лингвистического корпуса является его репрезентативность. Данная характеристика оценивается по изменению относительной частоты рассматриваемого явления при увеличении выборки. Если относительная частота явления от прибавления каждого последующего фрагмента текста будет изменяться все меньше и меньше, то корпус репрезентативен [Кибрик 2006]. С помощью репрезентативности неструктурированный набор

разных текстов превращается в корпус текстов, пригодный для проведения лингвистического исследования.

Стоит отметить важность такой особенности корпуса как простота использования. Корпус должен упрощать процесс исследования и сокращать временные затраты на его проведение, а не путать пользователей сложными алгоритмами поиска.

Процесс создания текстового корпуса состоит из двух последовательных этапов. Это сбор текстов и последующая их разметка (см. Рис. 2).

Рис. 2. Пример основных этапов создания текстового корпуса [Рыков 2002]

Современная корпусная лингвистика располагает большим количеством всевозможных вариантов корпусов. Существующее многообразие корпусов объясняется многообразием исследовательских и прикладных задач, для которых эти корпуса разрабатываются, и спецификой языкового материала, на котором они основаны. Важным аспектом будущего исследования является выбор соответствующего целям и задачам корпуса.

В научном сообществе на данный момент не существует общепринятой типологии корпусов. Представим некоторые из типологий. В.В. Рыков выделяет следующие корпуса:

1) по степени организации и структурированности: электронный архив (система структурированного хранения электронных документов); электронная библиотека (упорядоченная коллекция разнородных электронных документов, имеющая средства поиска и навигации); корпус (собрание текстов, по определённому принципу); подкорпус (подмножество текстов, ограниченное определенными метатекстовыми признаками: автором, произведением, временем создания или жанром и типом текста);

2) по хронологическому признаку: синхронический (пример использования языка в момент определенного отрезка времени); мониторинговый (позволяет пользователю в любое время обратиться к коллекции текстов или к их части); диахронический (динамика развития языка в течение некоторого периода);

3) по типу разметки: простой (неаннотированный); аннотированный (содержит данные, не являющиеся частью текста, но несущие какую-то информацию о нём);

4) по языковому признаку: одноязычный (в состав входят тексты одного языка, например его варианты и инварианты); двуязычный (содержит тексты двух языков); многоязычный (предоставляет доступ к текстам двух языков и более);

5) по способу применения и использования: исследовательский (предназначены для изучения всевозможных аспектов функционирования системы языка); иллюстративный (создаются после проведения научного исследования для подтверждения и

обоснования уже полученных результатов); параллельный (например, английский текст и его переводы на другие языки могут формировать такой корпус или быть частью большего корпуса параллельных текстов) [Рыков 2002].

В.П. Захаров предлагает свою типологию корпусов. Он также подразделяет корпуса на несколько видов: по типу языковых данных, хронологическому признаку, типу разметки и доступности (см. Рис.3).

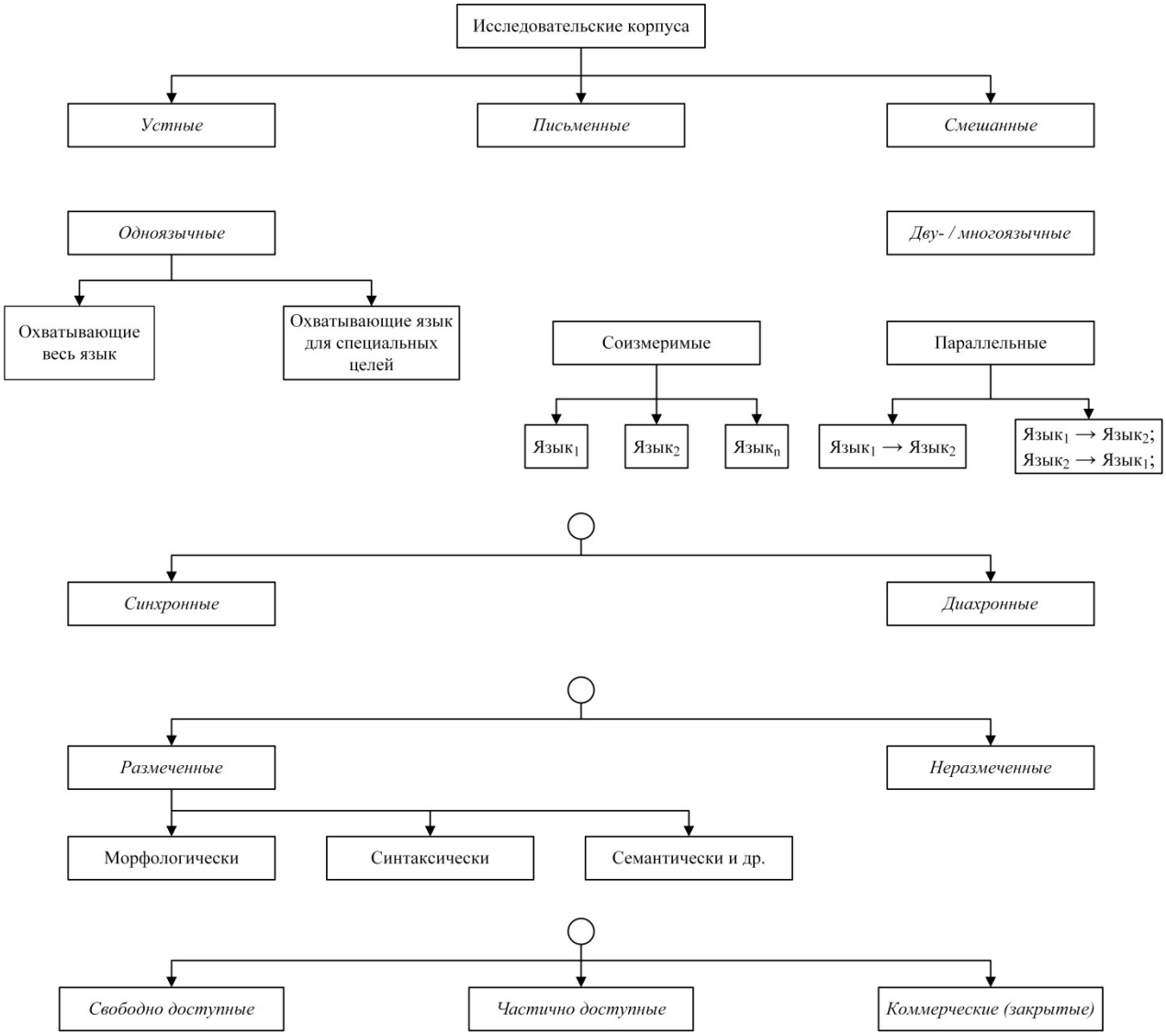


Рис.3. Типология корпусов по В.П. Захарову

По типу языковых данных корпуса бывают: устными, письменными и смешанными. Первый устный корпус разрабатывался на материале американского английского языка и появился в начале 80-х гг. (the Brown University Standard Corpus of Present Day American English). Корпусы устной речи – это специальные коллекции тщательно отобранных текстовых отрывков, «произнесенных многочисленными говорящими при различных акустических условиях» [Гвишиани 2008]. Процесс сбора устных данных более трудоемкий и напряженный. К ним относятся the London Lund Corpus (LLC), the Cambridge and Nottingham Corpus of Discourse in English (CANCO DE), the Santa Barbara Corpus of Spoken American English (SBCSAE) и др. Большую часть составляют письменные корпуса, например, Мангеймский корпус немецкого языка, и смешанные, содержащие и письменные, и устные тексты. Чаще всего это национальные корпуса.

Одноязычные корпуса могут охватывать весь язык или только его часть. Примером является корпус медицинских текстов на английском языке, насчитывающий 1.5 млн. слов – CEEM (Corpus of Early English Medical Writing). Двуязычные и многоязычные корпуса предоставляют тексты параллельно. Примером многоязычного корпуса может служить European Corpus Initiative (ECI) общим объемом более 100 млн. слов, состоит из текстов европейских языков, но содержит и японский, турецкий, русский и др.

С помощью диахронного корпуса можно проследить развитие языкового феномена или языка в целом на протяжении какого-либо временного отрезка. Примером

может служить Thesaurus Indogermanischer Text und Sprachmaterialien, в котором представлены индоевропейские тексты разных эпох. Синхронные корпуса предоставляют текстовый материал для анализа состояния языка как системы в определенный момент времени [Козлова 2013].

Неразмеченным корпусом называют массив текстов с определенным количеством упоминаний нужной единицы. Данные, полученные из анализа материала такого вида корпусов, полезны только для изучения языка со статистической точки зрения. Корпуса с разметкой предлагают гораздо больше возможностей для лингвистического анализа [Там же].

В национальном корпусе объединяются тексты разных типов и жанров. Некоторые из них могут содержать аудио и видео материалы. Объем влияет на разнообразие и точность разметки. От этих факторов напрямую зависит эффективность программного обеспечения и ценность корпуса как лингвистического ресурса. Национальный корпус дает справки, относящиеся к таким областям как: словарный запас, грамматика, история языка, акцентология. Новейшие компьютеры ускоряют и упрощают лингвистическую обработку массивов текстов, позволяя выявить закономерности в структуре и развитии языка, о существовании которых ученые раньше не догадывались или не могли обосновать. Самыми популярными национальными корпусами являются the British National Corpus (более 100 млн. словоупотреблений), the American National Corpus (23 млн.) и Национальный корпус русского языка (600 млн.).

Основными потребителями национальных корпусов являются исследователи-лингвисты. Но корпус используют не только профессиональные исследователи языка. Точная статистика о языке конкретной эпохи или определенного автора может представлять интерес для историков, филологов и других представителей области гуманитарного знания. Национальный корпус играет важную роль в преподавании языка. В наше время все больше учебников и учебных планов ориентируются на корпус. С его помощью школьник, учитель, редактор, писатель или журналист могут быстро и эффективно выяснить особенности использования незнакомого слова или грамматической формы.

Противоположностью национальным корпусам выступают специальные, созданные для решения конкретных лингвистических задач. Несмотря на широкую применимость универсальных текстовых корпусов, часто возникает необходимость в использовании узкоспециализированных коллекций для решения частных задач [Svartvik 2007, Tognini-Bonelli 2001]. Специализированный корпус – это специфический корпус, созданный для отображения определенного подъязыка. Например, the Corpus of Professional Spoken American English (CPSA) состоит из транскриптов коммуникативных ситуаций из политической и академической профессиональных областей. Среди такого типа корпусов можно отметить и корпусы одного автора, например, The Shakespeare Corpus, Корпус словаря языка Достоевского.Еще одним видом специальных корпусов являются ученические корпуса (Learner Corpora). В них входят тексты или аудио-записи лиц, изучающих язык как

иностранный [Захаров 2005]. Они создавались на стыке 80-х и 90-х годов прошлого века. Среди ученических корпусов выделяют следующие типы: коммерческие, академические, ученические.

Подводя итог обзору существующих корпусов, отметим, что приведенные выше примеры и данные не описывают в полной мере их многообразие. Раздел корпусной лингвистики находится в постоянном развитии и пополняется более совершенными разработками с учетом поставленных лингвистических задач. Для нужд корпусной лингвистики, которая в своих исследованиях сталкивается с объемными выборками текстов естественного языка, были разработаны особые виды программного обеспечения – коммерческие компьютерные программы конкордансеры (WordCruncher, LEXA, MicroConcord, CorpusWorkbench (CQP), TACT), а также компьютерные программы, разработанные для специфических процедур анализа.

Конкорданс – программа для поиска заданных языковых единиц в больших массивах текста и анализа закономерностей в языке. Результатом поиска являются несколько отрывков из различных текстов, в которых присутствует искомая единица. С их помощью можно определить значение слова или выражения в контексте или проанализировать употребление в языке. Например, употребление глаголов *to look* и *to watch* в Британском национальном корпусе (BNC) можно проследить на рисунке 4 [Aarts, Meijs 1984].

1	come to Purley she offered to come and	**look**	after me. She was a real treasure, of yeoman
2	and I'm sure Ellie can be persuaded to	**look**	after the children for a bit. If I contribute
3	till tomorrow, Miss Jeannie, and	**look**	after yourself." And with those weak words
4	This is the problem of how to	**look,**	and not the problem of what to look for. It
5	stores, and he picked it up, took a last	**look**	around the underground gallery, sniffing the
6	to detect as you might imagine. You can	**look**	as fit as a fiddle and yet be bloodless. My test
7	above the hob is that it is easier to	**watch,**	as you don't have to bend, and allows the
8	heavy lift and fullness. He kept a sharp	**watch**	but made no attempt to conceal his coming
9	slid gently back and forth along the gold	**watch**	chain slung across his wide chest. His
10	have been nothing pleasanter than to	**watch**	Christmas browsing, while one bore gently on
11	of us know the remarks made about the	**Watch**	Committee – and they are not always too kind

Рис.4. Употребление слов look и watch в БНК

Поиск можно использовать для уточнения словоупотребления и формулировки правил использования данного слова в языке, а также для изучения грамматики. Некоторые исследователи интерпретируют конкорданс как вертикальный список случаев использования слова в алфавитном порядке в электронном корпусе текстов. Слово находится вместе с его правым и левым окружением [Шаров 2003].

Работа с корпусом осуществляется с помощью других программных средств – корпус-менеджеров – систем поиска, включающих в себя программные средства для получения статистики и поиска данных в корпусе. Результаты поиска выдаются в удобном формате Key Word in Context (KWIC) – ключевое слово в контексте. Эта процедура позволяет представлять результаты в виде горизонтальных строк с поисковым словом посередине (см. Рис. 5). Существуют форматы Keyword alongside Context (KWAC) – ключевое слово вместе с контекстом и Keyword out of Context (KWOC) – ключевое слово вне контекста (см. Приложение 3).

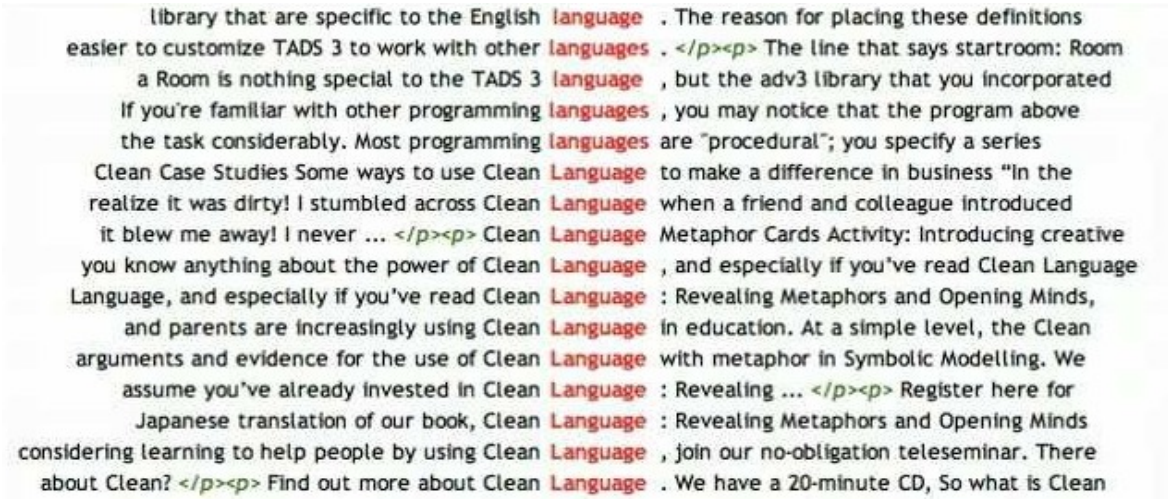


Рис.5. Формат Key Word in Context

Корпусные менеджеры предоставляют следующие возможности поиска: поиск словоформ по леммам, поиск словоформ по набору морфологических признаков, поиск конкретных словоформ, поиск группы словоформ в виде разрывной или неразрывной синтагмы, вывод результатов поиска с указанием заданного контекста и последующее сохранение нужных материалов. По результатам поиска искомая единица представляется в ее контекстном окружении и сопровождается статистической информацией.

Корпусные менеджеры должны соответствовать ряду общих требований: производить поиск отдельных слов и словосочетаний, строить полные конкордансные списки, сохранять и распечатывать результаты, осуществлять поиск по шаблонам, сортировать списки по нескольким критериям, отображать найденные словоформы в широком контексте, работать с отдельными файлами и неограниченными по размеру корпусами, быстро обрабатывать запросы и выдавать результаты, поддерживать различные форматы текстовых данных (txt, doc, rtf, html, xml) и быть интуитивно понятными в использовании [Захаров 2011].

С развитием систем корпусный метод начал предоставлять новые возможности в области исследования лексических и грамматических моделей. Благодаря корпусному анализу у лингвистов появилась возможность более глубоко исследовать метафоры. Предполагается, что образное значение слова, согласно данным словарей, может выступать в качестве маркера метафоры, необходимого для корпусного анализа [Deignan 2005]. Результатом исследования А. Дейнан стала разработка тематически организованного словаря на материале Английского Банка [Захаров 2005]. Таким образом, одним из преимуществ корпусных данных является то, что аутентичные примеры употребления слова (буквальные и небуквальные) могут быть изучены в контексте. Они дают исчерпывающую информацию о контекстных связях. Конкретный контекст дает возможность понять, в каком значении (образном или буквальном) употреблено слово или фраза. Благодаря данным корпуса, можно увидеть отличие метафорических выражений от буквальных по их форме.

Сегодня корпуса находят свое применение и в качестве экспериментальной базы для проверки гипотез и теорий лингвистами-теоретиками. В прикладной лингвистике компьютерные корпуса применяются для обучения иностранному языку и решения задач прикладного характера. Компьютерные лингвисты анализируют закономерности, полученные с помощью материалов корпуса, для создания компьютерных моделей языка. В социолингвистических исследованиях они используются для изучения разнообразия языков, к примеру, регистров и

социолектов. Корпуса также используются в таких сферах, как литературоведение, переводоведение, судебная лингвистика и др.

Еще одна область, находящаяся в прямом контакте с корпусной лингвистикой – создание и анализ ученических корпусов (Learner Corpora). Они предоставляют важную для учителя информацию о распространенных ошибках в употреблении лексических, грамматических и синтаксических единиц. Проанализировав данные, учитель может оптимизировать процесс обучения, сделав упор на пробелы учеников в конкретных областях знания. Важным аспектом для методики преподавания языка является использование параллельных корпусов. Они позволяют получить доступ к переводным эквивалентам слова, фразы или синтаксической конструкции. Такой инструмент способствует изучению иностранного языка в соответствии со стратегией коммуникативного обучения.

С увеличением компьютерных мощностей создатели программного обеспечения разработали новый тип словарей – электронный словарь. Электронный словарь, с технической точки зрения, это определенная база данных на каком-либо информационном носителе, предоставляющая закодированный список словарных статей, позволяющая осуществлять быстрый поиск нужных слов. Большой объем, удобство и скорость пользования стали возможными благодаря машинному механизму поиска, основанному на технологии Big Data. Это дает возможность постоянно обновлять базу данных словаря и создавать новые тематические варианты словарей.

На сегодняшний день самым большим корпусом может считаться сам Интернет (Web as Corpus), так как в нем представлено огромное количество текстов в электронной форме, находящихся в открытом доступе. Средствами доступа к этому корпусу считаются поисковые системы, например, Google или Yandex [Чернякова 2011]. Однако тексты в Интернете расположены не структурированно, поэтому исследователю сложно сформулировать лингвистически правильный запрос. Национальные лингвистические корпуса в плане использования удобнее и по этой причине нашли свое применение.

Вышеуказанные возможности использования корпусов не являются исчерпывающими. Параллельно с совершенствованием технологий Big Data развиваются и лингвистические корпуса, растет их исследовательский потенциал. Благодаря этому сегодня корпус стал бесценным лингвистическим ресурсом.

Выводы по главе I

В первой главе было описано, что такое Big Data, перечислены сферы применения, представлена история появления и становления лингвистических корпусов, корпусной лингвистики и программ конкорданс.

Появление и развитие систем больших данных оказало влияние на многие сферы научной деятельности. Как в технических, так и в гуманитарных науках системы Big Data предоставляют возможность анализа огромных массивов информации и открывают новые перспективы для проведения исследований.

Методы корпусной лингвистики становятся наиболее популярными в современных лингвистических исследованиях. Полученные после анализа корпуса данные, открывают исследователям сведения о закономерностях языка и отдельных типах текста. Корпусная лингвистика предоставляет широкий выбор корпусов: от крупных национальных корпусов с разметкой, таких как Национальный Корпус Русского Языка, до корпусов одного автора, например The Shakespeare Corpus.

Для анализа данных лингвистического корпуса требуется соответствующее программное обеспечение. Работа с корпусом может осуществляться с помощью специальных программных средств – конкордансеров и корпус-менеджеров. Конкордансеры используются в анализе больших массивов текста (например, корпуса), для поиска слов и последующего выявления закономерностей. Под корпусным менеджером понимается специальная систему поиска, включающая программные средства для получения статистики и поиска данных в корпусе.

Глава II. Возможности систем Big Data для анализа предикатов *to seem* и *to appear*

Обращение к корпусам текстов при анализе предикатов способствует более быстрой и объективной обработке языковых данных, что в свою очередь позволяет получить достоверные статистические данные о частотности их использования, информацию о нюансах семантических значений и особенностях употребления. В главе представлено понимание предиката и предикативности, описана семантика глаголов *to seem* и *to appear*, проанализирован потенциал систем Big Data для проведения лингвистических исследований.

2.1 Понимание предикативности в лингвистике

К началу XX-го века многие лингвисты, философы и психологи задавались вопросом о конкретизации понятия *предикативность*. Считалось, что весь окружающий мир состоит из фактов и событий, а не из предметов или вещей. В языке для выражения факта используется предложение, а его центром является предикат [Рассел 1959].

В первобытном обществе людей весь язык состоял из «предложений с выраженным в слове одним только сказуемым» [Потебня 2003]. Современные лингвисты поддерживают эту точку зрения и предполагают, что в эпоху первобытного человека единицей общения выступало слово-предложение. Под ним понимается неделимое синкретичное образование, совмещающее слабо выраженную психологическую двучленность и сочетающее предикативную и номинативную функции [Гречко 2003].

В работах по изучению внутренней речи подчеркивается преобладание в ней абсолютной и постоянной предикативности. Считается, что внутренняя речь опирается на семантику, в то время как синтаксис и фонетика сведены к минимуму, максимально упрощены [Выготский 2003].

В.А. Дорошевский отмечал, что одной из основных функций человеческого мозга является способность к предикации. Она выступает своеобразным фильтром информации, поступающей от внешних раздражителей. После такой фильтрации окружающий мир предстает перед человеком как движение взаимосвязанных элементов,

подчиняющихся порядку в категориях предметов и отношений, а не как беспорядочный хаос. [Дорошевский 1973].

Лингвистический энциклопедический словарь дает следующее определение предикативности – это ключевой признак предложения, относящий информацию к действительности и формирующий единицу, предназначенную для сообщения; такая синтаксическая категория, которая определяет функциональную специфику предложения [Ярцева 1990].

Значение общей категории предикативности заключается в соотнесении содержания предложения с действительностью. К этой категории в первую очередь относятся модальные отношения [Виноградов 2001].

В вышеперечисленных определениях понятие *предикативность* в основном относится к синтаксической теории предложения. Однако оно может рассматриваться и с лексической точки зрения. В таком случае предикат рассматривается как семантическое понятие. При этом подходе предикатные значения реализуются при помощи морфем, лексем и словосочетаний, а основное внимание уделяется глаголу и другим формам, которые могут выступать в функции сказуемого, что не всегда отражает полную картину их функционирования.

Отмечается, что предикаты – это особые семантические сущности языка, они типизируются языком не в форме словарных единиц, глаголов, а в форме пропозициональных функций и соответствующих им структурных схем предложений [Степанов 2010].

В современном языкознании представлено несколько семантических классификаций предикатов. Рассмотрим некоторые из них. В рамках русского языка Л.В. Щербой выделяются три типа предикатов:

1) со значением действия – передается полнозначными глаголами: *бегать, играть, работать*;

2) со значением состояния – связка + *жаль, в состоянии, печален, надо*: *был в состоянии, стал печален, было жаль*;

3) со значением качества – связка + прилагательное: *является непреодолимым, оставаться нетронутым, казаться большим* [Потапова 2017].

Широко используется классификация, предложенная З. Вендлером для английских предикатов. Им были выделены четыре основных класса. Общепринятых русских названий не сформировалось, поэтому предложены английские названия:

1) activities – слова со значением деятельности: *The soup was boiling*;

2) achievements – со значением достижения: *I was finding it hard to finish*;

3) accomplishments – со значением исполнения: *I finished my book in two weeks;*

4) states – состояния: *It lasted for three hours* [Там же].

В работе О.Н. Селиверстовой была представлена более широкая классификация предикативных типов, предложены некоторые изменения в интерпретации предикатов, выделен класс предикатов «состояния». Автор отмечает связь предикативных типов с признаком соотнесенности с

непосредственным протеканием во времени. В классификации перечислены предикаты со значением:

1) действия: *читать, бежать*;

2) состояния: *входить в список*;

3) процесса: *тонуть, расти*;

4) потенциальности: *может справиться, должен приехать*;

5) нахождения в пространстве: *лежать, стоять*;

6) качества и набора качеств: *раньше она была красива*;

7) класса и связи: *курить* (как ежедневное действие);

8) результата и факта: *встретить, найти* [Селиверстова 1982].

Будучи понятием логики и языкознания, предикат обозначает характеристики субъекта. Это не всякая информация о субъекте, а указание на признак предмета, его состояние и отношение к другим предметам. В языкознании для этого понятия используется термин *сказуемое*. Это позволяет избежать терминологического смешения логических и грамматических категорий [Пащенко 2006].

Сказуемое является одним из главных членов предложения и дает информацию о том, что происходит с предметом или о том, что делает какое-либо одушевленное существо. Со сказуемым ассоциируется формальный аспект этого члена предложения, а с предикатом – содержательный [Арутюнова 1980]. В предложении предикат может представляться только признаковым значением, в то время как сказуемое допускает любой вид информации.

В языкознании понятие *сказуемое* используется при обозначении ядерного компонента состава предложения,

соответствующего сообщаемому (англ. predicate, исп. predicado франц. predicat, итал. predicato). Предикат находится в предикативном отношении к субъекту. Отношения, связывающие субъект и предикат (подлежащее и сказуемое), называются предикативными или предикативной связью. Состав этих понятий включает в себя предикативность. Это синтаксическая категория, формирующая предложение и соотносящая содержание предложения с действительностью, что делает его единицей сообщения. Любому предложению свойственна предикативность, и она делает предложение предложением [Пащенко 2006].

Можно сделать вывод что, понятие *предикативность* одновременно затрагивает области философии, психологии и лингвистики. В плане значения оно объемнее, чем понятие *предикат*, поскольку предикативность является экстралингвистическим термином.

В английском языке предикат выражается личной формой глагола, которая согласуется с подлежащим в числе и лице. Исходя из значения его компонентов, предикат может обозначать действие, состояние, качество или отношение к какому-либо действию или состоянию, приписанному субъекту. Эти различные значения находят свое выражение в структуре предиката и лексическом значении его составляющих [Там же].

Со структурной точки зрения предикаты делятся на два основных типа: простой предикат и составной. Н.А. Кобрина предлагает свою классификацию (см. Рис. 6).

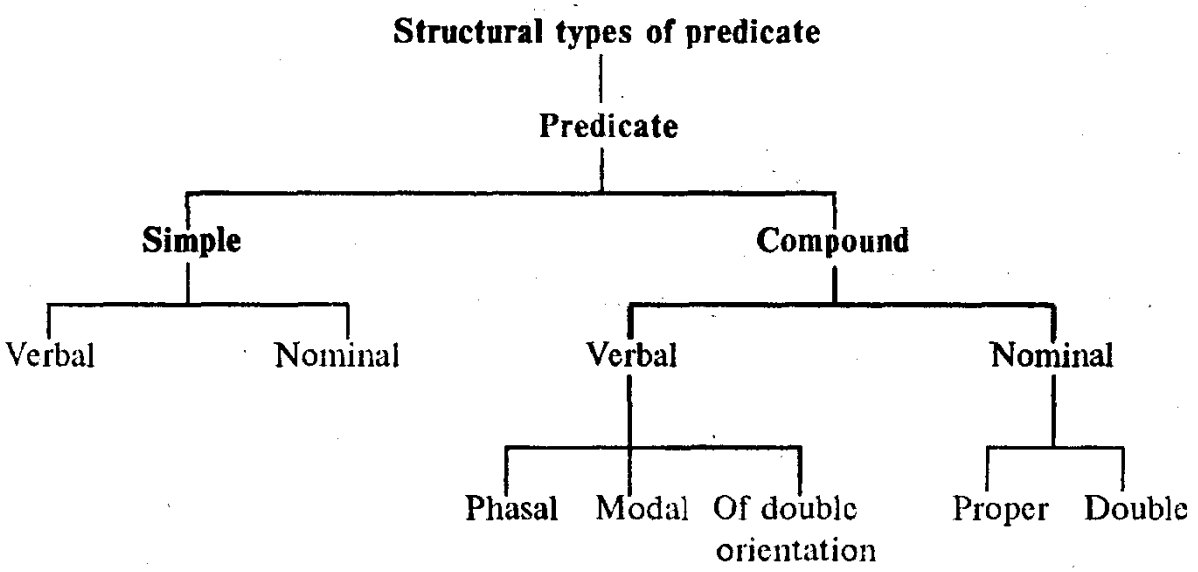


Рис.6. Типы предикатов

Простой и составной предикаты подразделяются на подгруппы – глагольные и именные. Составное глагольное сказуемое бывает: фазовым, модальным и двойным. Составное именное сказуемое делится на двойное и собственное.

С точки зрения смысла наиболее важной частью составного предиката является смысловая часть, поскольку она содержит информацию, выраженную подлежащим. С точки зрения структуры, наиболее важной частью предиката является первая, поскольку она выражается личной формой глагола и несет грамматическую информацию о субъекте, числе, времени, модальности и т.д. всего предиката.

Выделяют предикативные конструкции с инфинитивом и причастием. Первые подразделяются на три подгруппы: трехчленная глагольная конструкция с инфинитивом, инфинитивный оборот с предлогом for, субъектный инфинитивный оборот. Предикативные конструкции с причастием подразделяются на: трехчленные глагольные конструкции с причастием I и причастием II, субъектные,

независимые и предложные причастные обороты [Берман 1993].

В параграфе было представлено понимание предикативности, приведены существующие классификации предикативных типов, указаны их синтаксические функции и перечислены предикативные конструкции.

2.2 Семантика предикатов *to seem* и *to appear*

Глаголы *to seem* и *to appear* обладают сложной семантической структурой. Обобщенное значение этих предикатов можно выразить следующим образом: они сообщают об отображении действительности, которая сформировалась в сознании субъекта при ее восприятии [Ильчук 2004]. Опишем каждый из глаголов.

С помощью электронной версия этимологического словаря английского языка мы можем проследить этимологию глагола. В XII веке в английском языке он имел следующие значения: *быть подходящим, уместным, годным*. В древнескандинавском языке слово *soeta* означало: *уважать, примеряться, приспосабливаться*. Этот глагол образовался от прилагательного *soemr – подходящий* в Прото-Германском языке или глаголов *somiz , söme – быть правильным, благопристойным* в старом Датском языке [Этимологический онлайн словарь].

Онлайн версия программы морофлогического анализа слов, реализованная на основе словаря Мюллера, дает следующие определения:

1) *Seem* – казаться, представляться: ***I seem*** *to hear somebody crying* – ***Мне послышалось / показалось****, что кто-то плачет.*

2) *Seem* – употребляется как глагол-связка: *She* ***seems*** *happy – Она* ***выглядит*** *счастливой* [Англо-русский словарь Мюллера].

Представим информацию из исследований, в которых анализируется семантика этого предиката.

To seem указывает на известную предположительность оценки, на отсутствие у субъекта восприятия уверенности в том, что создавшееся у него впечатление правильно:

He ***seemed*** *surprised at the news* [Апресян 2000].

To seem приписывает возможное отклонение оценки от истины каким-то особенностям субъекта восприятия [Там же].

To seem (в ситуациях, когда речь идет о впечатлениях, относящихся к какой-либо воспринимаемой характеристике объекта) употребляется для обозначения иллюзий, связанных с ощущениями субъекта, а также создаваемых туманом, воздухом, звуком:

> *A coin* ***seems*** *larger when it is placed on the tongue than when it is held in the palm of the hand* [Ильчук 2004].

To seem несет информацию о таких цветовых впечатлениях, которые не образуют четкого зрительного образа:

> *The more distant it is the hazier and bluerer it* ***seems*** [Там же].

To seem сообщает о таких впечатлениях относительно размера, формы, удаленности объекта, которые не формируют четкий зрительный образ, а устанавливаются в результате сравнения:

The moon ***would seem*** *to be several times farther away than when it is midnight* [Там же].

Глагол *to seem* (в значении вторичного восприятия) передает информацию об: отображении, относящимся к какой-либо внутренней характеристике объекта, формирующейся при восприятии субъектом таких данных, которые можно считать признаками устанавливаемой характеристики, а не внешними проявлениями; признаках во внешнем виде, не связанных очевидным образом с устанавливаемой характеристикой; субъекте, если тот не уверен, что подобное впечатление будет получено любым другим наблюдателем и допускает сомнение в его правильности; впечатление зависит от субъекта.

Глагол *to seem* (в ситуациях, когда отображение определяется особенностями субъекта) сообщает о том, что эмоциональное состояние субъекта меняет восприятие ситуации / объекта на более глубоком уровне, а не на уровне внешних проявлений:

If you go to three places, the holiday ***seems*** *to last at least three times as long* [Там же].

Опишем семантику второго предиката. Электронная версия этимологического словаря английского языка дает следующую информацию: современный английский глагол *to appear* связан со старым французским глаголом *aparoir, aperer* (XIII век) – *появиться, открыться взгляду* и с

латинским глаголом *apparere – появиться в поле зрения, показаться* [Этимологический онлайн словарь].

Обратимся к словарю Мюллера, в котором указаны следующие значения глагола *appear*:

1) показываться, появляться: *Sure enough, the ghost **appeared** on the balcony;*

2) выступать на сцене: *To **appear** in the character of Othello*;

3) выступать (официально, публично): *To **appear** for the defendant*;

4) предстать (перед судом): *More than 1000 witnesses from all over the world have been called to **appear** before the Tribunal;*

5) выходить, издаваться, появляться (в печати): *Some press organs had ceased to **appear** for purely commercial reasons, generally bankruptcy;*

6) производить впечатление, казаться: *Strange as it may **appear***; *you **appear** to forget*;

7) явствовать: *It **appears** from this* [Англо-русский словарь Мюллера].

Представим информацию из исследований, в которых анализируется семантика предиката *to appear.*

To appear показывает сложность оценки, вероятность того, что она не отражает существующую на самом деле ситуацию. Глагол приписывает возможное отклонение оценки от истины особенностям воспринимаемого объекта, его внешности, которая может быть обманчива, или, если этот объект является человеком – он может быть заинтересован в том, чтобы создать ложное впечатление:

*He **appeared** as helpless as achild* [Апресян 2000].

To appear (сообщающий о впечатлениях, относящихся к какой либо воспринимаемой характеристике объекта) употребляется по отношению к такой группе денотатов, которые имеют зрительную форму воплощения:

*The sky **appeared** to be darker than* [Ильчук 2004].

Сообщает о ярких, контрастных цветовых иллюзиях, которые имеют выраженную принадлежность к данному цветовому спектру, что согласуется с признаком качества отображения:

*Illuminated from the front and seen against a dark background the water **appears** bluish* [Там же].

Глагол выбирается в тех случаях, когда впечатление об объекте представляет собой хорошо структурированный зрительный образ:

*Stars do not **appear** to us as perfect spots, but as small irregularly shaped figures* [Там же].

Глагол *to appear* (в значении вторичного восприятия) сообщает о том, что: отображение, относящееся к внутренней характеристике, формируется при восприятии субъектом таких данных, которые можно считать внешними проявлениями указанной характеристики; внешние проявления связаны с внутренней характеристикой; субъект уверен в правильности своего впечатления; впечатление не зависит от субъекта.

Глагол *to appear* (в случаях, когда отображение особенностями субъекта) показывает, что эмоциональное состояние субъекта определенным образом изменяет восприятие внешнего облика проявлений:

*He was in love and therefore she **appeared** to him so perfect in every respect* [Там же].

В большинстве случаев предикаты *to seem* и *to appear* являются взаимозаменяемыми и переводятся на русский язык глаголами несовершенного вида – казаться, оказываться.

*It **seemed** to go down very well – **Кажется** все закончилось благополучно* [COCA].

*Jane **appears** to have heard some terrible news – **Кажется** Джейн узнала ужасные новости* [COCA].

Исключением являются случаи, когда глагол *to appear* обозначает появление кого-либо или чего-либо:

*He **appeared** out of nowhere* [COCA].

Согласно словарю MW Dictionary of Synonyms, эти предикаты взаимозаменяемы и не имеют видимой разницы в значении [Словарь Вебстера]. Вместе с тем отмечается, что даже в таких фразах *to seem* обозначает мнение, основанное на субъективных впечатлениях и личном отношении, а не на объективных признаках. Глагол *to appear* может означать, что мнение основано на общем визуальном впечатлении (например, как в случаях с глаголом *to look*), но иногда *to appear* предполагает искаженное восприятие, которое может быть создано оптическим обманом или ограниченным углом зрения [Там же].

О наличии различий в значении предикатов свидетельствуют исследования Ю.Д. Апресян, Е.В. Ильчук и Т.И. Семеновой. У глаголов *to seem* и *to appear* выделяются несколько вариантов значения, а именно: случаи, когда *to*

seem и *to appear* сообщают о формировании отображения с помощью синтезирующего восприятия ($seem_1$ / $appear_1$), и случаи, в которых отображение формируется частично с помощью восприятия и с помощью логической обработки данных ($seem_2$ / $appear_2$). По мнению Е.В. Ильчук, понятие синтезирующее восприятие характеризует способ построения отображения или некоторой мысленной картины действительности, и в отличие от других глаголов восприятия, относящихся к глаголам проецируемого отображения, глаголы *to seem* и *to appear* указывают на конструирование образа, а не на построении в сознании проекции объекта.

В варианте синтезирующего восприятия различаются несколько подвариантов:

- *to seem / to appear* $_{1/1}$ – сообщают о различных впечатлениях, которые относятся к какой либо непосредственно воспринимаемой характеристике объекта (цвет, форма, размер);
- *to seem / to appear* $_{1/2}$ – несут информацию о вторичном восприятии;
- *to seem / to appear* $_{1/3}$ – отображение предопределяется не только свойствами самого объекта, но и особенностями субъекта [Ильчук 2004].

Перейдем к рассмотрению характеристик, разграничивающих глаголы *to seem* и *to appear* и описанию их значения. Выделяются следующие признаки:

1) признак учитывающий качество отображения или характер данных, на основании которых формируется отображение;

2) признак глубины обработки данных;

3) признак субъективности / объективности;

4) признак произвольности / непроизвольности [Ильчук 2006].

Рассмотрим первый признак, поскольку он является ведущим и обуславливает остальные. В зависимости от варианта значения данных глаголов, он может несколько менять интерпретацию. В соответствии с ним, глагол *to seem*$_1$ показывает, что отображение не имеет четкого образа, либо неочевидно. В свою очередь глагол *to appear*$_1$ свидетельствует о наличии четкого, как правило, зрительного образа:

The blurred hands ***seemed*** *to indicate nearly half past six* [COCA].

Данный пример подтверждает сделанное выше утверждение, поскольку в нем, наличие помех затрудняет восприятие объекта.

В отличие от глагола *to seem*, глагол *to appear* выбирается в тех случаях, когда в контексте подчеркивается явность, определенность фиксируемой характеристики:

When the sun is low, the water on the north and south sides ***should appear*** *noticeably darker than on the east and west sides* [BNC].

Согласно исследованию, проведенному Е.В. Ильчук, в то время как *to seem*$_2$ несет информацию об отображении, сформированном на основании нечетких исходных данных или данных, не связанных очевидным образом с устанавливаемой характеристикой, глагол *to appear*$_2$, показывает то, что отображение формируется на основании четких, но поверхностных исходных данных, связанных

очевидным образом с приписываемой объекту характеристикой [Ильчук 2006]. Так глагол *to seem*$_2$ может использоваться в тех случаях, когда отображение формируется на основании смутно осознаваемых данных, всплывающих в сознании, например, воспоминаний и в ситуациях, предполагающих осмысление собственных чувств. Глагол *to appear*$_2$, наоборот, в таких случаях не употребляется:

*I **seem** to have changed my mind, he thought gloomily* [BNC].

Когда речь идет не столько о нечетком, сколько о неочевидном характере данных, употребляется глагол *to seem*. *To appear*, в свою очередь, подчеркивает, что между исходными данными и отображением существует связь. Такие различия употребления глаголов можно проиллюстрировать на примере их использования с пропозициями, содержащими чужое сознание [Ильчук 2004].

*Larry had been watching John's mood and he **seemed** to make a sudden decision* [COCA].

Поскольку предикаты *to seem* и *to appear* относятся к средствам вербализации понятия кажимости, в их семантике присутствуют все признаки, характерные для данного феномена [Семенова 2007]. Когнитивную основу кажимости определяют три признака: 1) двуплановость – совмещение реального и кажущегося миров; 2) наличие наблюдателя / самонаблюдения; 3) восприятие (чувственное) ситуации наблюдающим [Арутюнова 1999]. Кроме вышеназванных признаков, некоторые исследователи выделяют еще два: условие восприятия (чувственно-эмоциональное и

физическое состояние воспринимающего) и объект восприятия [Семенова 2007].

Согласно Большому Оксфордскому Словарю, глагол *to seem* (казаться), имеет следующие основные лексико-семантические варианты:

1) *to have a semblance or appearance; to appear to be, to be apparently (what is expressed by the complement*) – иметь сходство или видимость, казаться, быть явным (что выражается посредством дополнения): *It* ***seems*** *clear that there has been a mistake.*

2) *with infinitive: to appear to be or to do something* – употребляется с инфинитивом: казаться чем-то или делать что-то: *She* ***seems to be*** *a smart woman* [Большой Оксфордский Словарь].

Согласно определениям, в первом случае глагол *to seem* выражает предположение с низкой степенью достоверности; во втором лексико-семантическом варианте, напротив, значение глагола *to seem казаться* является выражением предположения с высокой степенью достоверности.

Словарь Вебстера (Webster's new international dictionary (2nd edition) предлагает следующие лексико-семантические варианты глагола *to seem*: представлять внешние признаки таким образом, чтобы они заставили говорящего или других воспринимающих предположить один из них (быть, действовать, идти, и так далее): представить все знаки, указания, относящиеся к делу; быть очевидной правдой или чьими-то впечатлениями, либо мнением; казаться, притворяться; представить похожие или какие-то явные признаки; сообщаться официально или услышать слухи, стать

известным [Словарь Вебстера]. Вышеупомянутые значения глагола *to seem* позволяют выделить варьирование его значений от оттенков неуверенности, колебания до предположения, основанного на явных очевидных признаках [Абдусаламова 2011].

Для глагола *to appear* предлагаются следующие лексико-семантические варианты значений: появиться в поле зрения, стать заметным; начать свое существование, проявить свои признаки впервые; создавать впечатление о ком-либо или о каком-либо действии (синоним *seem*) [Большой Оксфордский Словарь]. Как видно из дефиниций, глагол *to appear* употребляется, когда речь идет о более четком, физически ощутимом восприятии, которое часто происходит впервые.

Ю.Д. Апресян также уверен, что в тех случаях, когда речь идет об объекте или ситуации (не о человеке), и когда впечатление об этом объекте или ситуации основано на их внутренних характеристиках (то есть не на их физических свойствах, которые могут быть восприняты нашими чувствами), вышеописанное семантическое различие исчезает:

*The news **seemed / appeared** to be very important*

[Апресян 2000].

Таким образом, в параграфе были проанализированы семантические особенности глаголов восприятия *to seem* и *to appear*. Исследование источников показало, что их значения разграничены четырьмя признаками. Считается, что данные глаголы относятся к глаголам синтезирующего восприятия. Была рассмотрена точка зрения, согласно которой предикаты

to seem и *to appear* относятся к средствам вербализации понятия кажимости.

2.3 Исследование возможностей систем Big Data с помощью предикатов *to seem* и *to appear*

Системы больших данных дают исследователю большой набор возможностей, которые могут существенно помочь в работе. Опишем некоторые из них.

Практически во всех словарях и исследованиях утверждается, что предикат *to appear* встречается реже, чем *to seem*, и является немного более формальным. С помощью сервиса Google Books Ngram Viewer, который дает возможность строить графики частотности языковых единиц на основе большого количества печатных источников в период времени с XVI века и до сегодняшнего дня, проанализируем употребление глаголов *to seem* и *to appear* в диахронии, за период с 1800 до 2008 года. Результат представлен на рисунке 7.

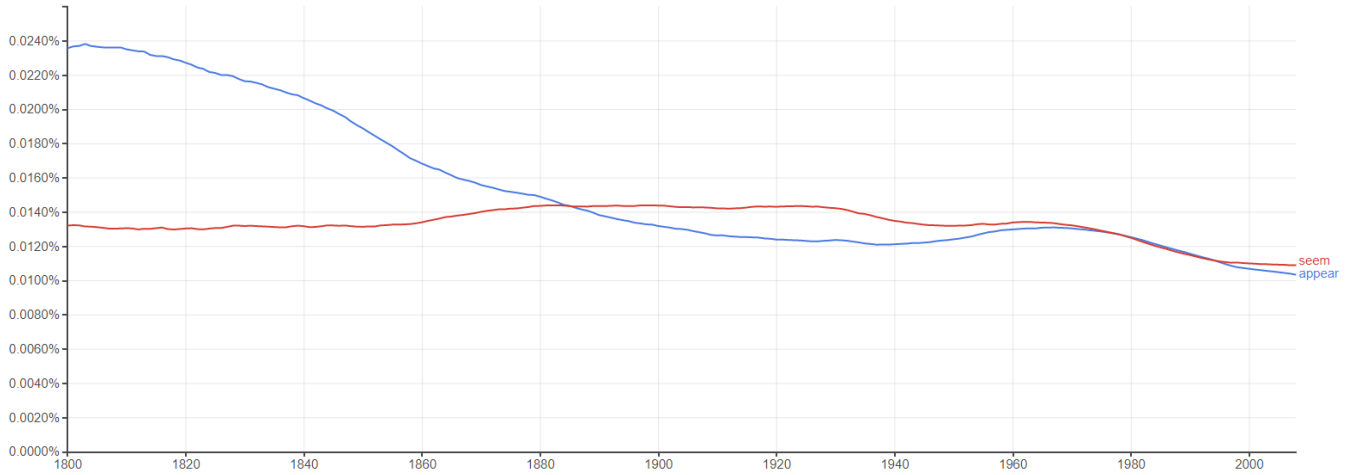


Рисунок 7. Частотность употребления глагола seem и appear с 1800г. до 2008г.

Судя по полученным данным, можно утверждать, что два века назад более частотным глаголом был *to appear*, но в начале XX века он уступил первое место глаголу *to seem*. Следовательно, подтверждается мнение о том, что *to seem* встречается чаще, чем *to appear*, хотя разница в частотности употребления данных языковых единиц на сегодняшний день незначительна.

С помощью системы GoogleTrends проанализируем, насколько часто предикаты *to seem* и *to appear* используются в разных странах. Для этого введем поисковые запросы формата *to seem* и *to appear* в поисковую строку. Рассматриваемый период времени – с 2004 г. по 2020 г. Результатом будет график динамики популярности для конкретной страны и таблица популярности для всех стран. Значения для каждой страны представлены в виде баллов. 100 баллов означают местоположение с наибольшей долей популярности запроса, 50 баллов – местоположение, уровень популярности запроса в котором вдвое ниже, чем в первом. 0 баллов означает местоположение, по которому недостаточно данных о рассматриваемом запросе. Данные представлены в таблице (см. Таблица 1).

Таблица 1

Популярность глаголов to seem и to appear по странам

Глагол	**Google Trends популярность по регионам**
To seem	Соединенные Штаты - 100 Великобритания - 76 Австралия - 71 Россия - 10
To appear	Соединенные Штаты - 80 Австралия - 62 Великобритания - 54 Россия - 4

Основываясь на полученных данных, можно утверждать, что наибольшее количество запросов относится к Соединенным Штатам, а наименьшее к России. Во всех рассматриваемых регионах запросы с глаголом *to seem* являются более популярными, чем с глаголом *to appear*.

Ранее при анализе семантики предикатов было найдено утверждениие, что разница между *to appear* и *to seem* заключается в том, откуда возникает неопределенность. При употреблении глагола *to seem* эта неопределенность исходит от наблюдателя / субъекта как носителя опыта. При использовании глагола *to appear* неопределенность вызвана характеристиками / чертами наблюдаемого человека или предмета и может подразумевать попытку обмана [Апресян 2000].

Маркером речевой ситуации, подразумевающей попытку обмана, может быть конструкция *tried to appear / to seem* с прилагательным. Рассмотрим частотность с помощью корпусов BNC и COCA. С помощью поисковых запросов формата *tried to appear _j** и *tried to seem _j**, найдем нужные нам конструкции. Результатом поиска будет список предложений с искомыми единицами и их общее число (см. Таблица 2 и Приложение 4).

Из результатов видно, что конструкция *tried to appear + adj* встречается чаще, чем *tried to seem + adj* – 39 и 16 употреблений соответственно. Таким

Таблица 2

Частотность использования конструкций *tried to appear + adj / tried to seem + adj*

	BNC	COCA
Tried to appear + adj	7	32
Tried to seem + adj	0	16

образом, можно предположить, что *to appear* подразумевает попытку обмана чаще, чем *to seem*.

Программа SentiStrength помогает проанализировать предложения с глаголами *to seem* и *to appear* с точки зрения положительной / отрицательный окраски речи. Всего было проанализировано 25 предложений с глаголом *to seem* и 25 предложений с глаголом *to appear*. Результат представлен в таблице 3, полный список примеров в приложении (см. Приложение 15 и Приложение 16).

Таблица 3

Эмоциональная окраска предложений с глаголами to seem и to appear

	Положительные	Отрицательные	Нейтральные
To seem	6	7	12
To appear	2	8	15

Согласно полученным данным, оба глагола чаще всего нейтральны. Однако глагол *to appear* в четыре раза чаще встречается в предложениях с отрицательной эмоциональной окраской в сравнении с предложениями с положительной эмоциональной окраской.

Электронный тезаурус WordNet предоставляет все возможные толкования слов и показывает взаимосвязи между ними. С его помощью можно получить ссылки на производные или более общие понятия, найти синонимы, автоматически объединенные в смысловые группы.

Проанализировав глаголы с помощью данного ресурса, убеждаемся, что *to seem, to appear*, имеют тесную семантическую связь. *Seem, appear* – произвести определенное впечатление или иметь определенный внешний вид:

*She **seems** to be sleeping;*

*This **appears** to be a very difficult problem;*

This project looks fishy [словарь WordNet]

Стоит отметить, что ввиду своего ограниченного функционала, данный ресурс не подходит для серьезных лингвистических исследований. Остальные материалы, полученные с помощью ресурса, представлены в приложении (см. Приложение 7).

Частотность употребления предикатов *to seem* и *to appear* можно также уточнить средствами систем национальных корпусов. Для проведения исследования BNC и COCA были выбраны как самые репрезентативные, объемные и удобные в использовании корпуса английского языка. Отличия в полученных результатах обусловлены как разницей в семантике глаголов, так и неодинаковой наполненностью корпусов (примерно 560 млн. слов в COCA и 100 млн. слов в BNC).

Для определения частотности употребления в каждый из корпусов текстов вводится искомый глагол. В целях получения более точного результата для каждой формы глагола вводился отдельный поисковый запрос формата *seem*, *seems*, *seemed* и так далее. В результате были получены данные о количестве употребления одной конкретной формы (см. Рис. 8).

*Рисунок 8. Результат поискового запроса **see**m в COCA.*

По результатам последовательного исследования всех форм предиката *to seem* в обоих корпусах были получены

следующие данные о частотности их употребления (см. Таблица 4).

Таблица 4

Частотность употребления глагола to seem и его форм в корпусах COCA/BNC

	Seem	Seems	Seemed	Seeming	Will seem	To seem	Would seem	Общее кол-во
COCA	161011	214654	144632	4906	881	2641	8522	537247
BNC	116290	200541	219177	378	106	255	1616	59636

Из полученных результатов видно, что как в COCA, так и в BNC, наиболее частотными являются формы *seems* и *seemed*. Наименее частотной в обоих корпусах является форма будущего времени *will seem*.

Используя системы больших данных можно уточнить нюансы употребления любых языковых единиц. Для определения коллокаций предиката *to seem* в корпусах текстов BNC и COCA вводился поисковый запрос формата *seem**, что позволило включить в поиск формы *seems, seemed, seeming*. Далее с помощью функции *Collocates* был получен список коллокатов, упорядоченный по частотности их употребления (см. Рис. 9 и 10).

		CONTEXT	FREQ		ALL	%	MI	
1		TO	30245		2565070	1.18	4.23	
2		LIKE	976		145999	0.67	3.41	
3		LIKELY	767		22798	3.36	5.74	
4		UNLIKELY	365		5480	6.66	6.73	
5		QUITE	361		39516	0.91	3.86	
6		ALMOST	272		30043	0.91	3.85	
7		CLEAR	247		24762	1.00	3.99	
8		RATHER	211		41348	0.51	3.02	
9		REASONABLE	203		6091	3.33	5.73	
10		POSSIBLE	178		33339	0.53	3.09	
11		STRANGE	145		6196	2.34	5.22	
12		APPROPRIATE	139		11346	1.23	4.28	
13		IMPOSSIBLE	136		6761	2.01	5.00	
14		CERTAIN	132		21589	0.61	3.28	
15		ENDLESS	102		1512	6.75	6.75	
16		OBVIOUS	100		8234	1.21	4.27	
17		ODD	99		4255	2.33	5.21	
18		SENSIBLE	92		2656	3.46	5.78	
19		UNABLE	86		6090	1.41	4.49	
20		DETERMINED	82		7504	1.09	4.12	

Рис.9. Коллокация предиката seem в Британском национальном корпусе текстов

Проанализируем коллокацию предиката *to seem* с помощью национальных корпусов. Представим несколько примеров с глаголом *to seem* и самими часто употребляемыми коллокациями в корпусе BNC. Остальные примеры можно найти в приложении (см. Приложение 5).

Countries now ***seem to*** *think that monetary-policy measures are their only option* [BNC].

And something appeared to them which ***seemed like*** *tongues of fire* [BNC].

It seems likely *that, in many cases, they are correct, but that these symptoms are not due to what doctors would normally regard as allergy* [BNC].

Далее представлены несколько примеров с глаголом *to seem* и самими часто употребляемыми коллокациями в корпусе COCA (см. Рис. 10). Остальные примеры представлены в приложении (см. Приложение 5). .

Your truck ***seemed to*** *be running just fine yesterday, but you got in today and you're going nowhere* [COCA].

It seems like *CNN, NBC, CBS, and ABC are ignoring this* [COCA].

		CONTEXT	FREQ		ALL	%	MI	
1		TO	245351		25557793	0.96	4.13	
2		LIKE	47076		2368863	1.99	5.18	
3		PRETTY	2428		222515	1.09	4.31	
4		LIKELY	2364		189242	1.25	4.51	
5		ALMOST	2201		279516	0.79	3.84	
6		UNLIKELY	1777		27384	6.49	6.89	
7		LESS	1632		339506	0.48	3.13	
8		CLEAR	1616		195945	0.82	3.91	
9		QUITE	1595		181484	0.88	4.00	
10		IMPOSSIBLE	1325		58014	2.28	5.38	
11		ENDLESS	1144		17196	6.65	6.92	
12		OBVIOUS	1118		56768	1.97	5.17	
13		REASONABLE	1084		37114	2.92	5.73	
14		POSSIBLE	1045		213993	0.49	3.15	
15		STRANGE	998		55570	1.80	5.03	
16		ODD	979		29449	3.32	5.92	
17		HAPPY	908		158440	0.57	3.38	
18		PARTICULARLY	867		104067	0.83	3.92	
19		FAIR	829		86564	0.96	4.13	
20		APPROPRIATE	760		59641	1.27	4.54	

Рис. 10. Коллокация предиката seem в Корпусе современного американского английского языка

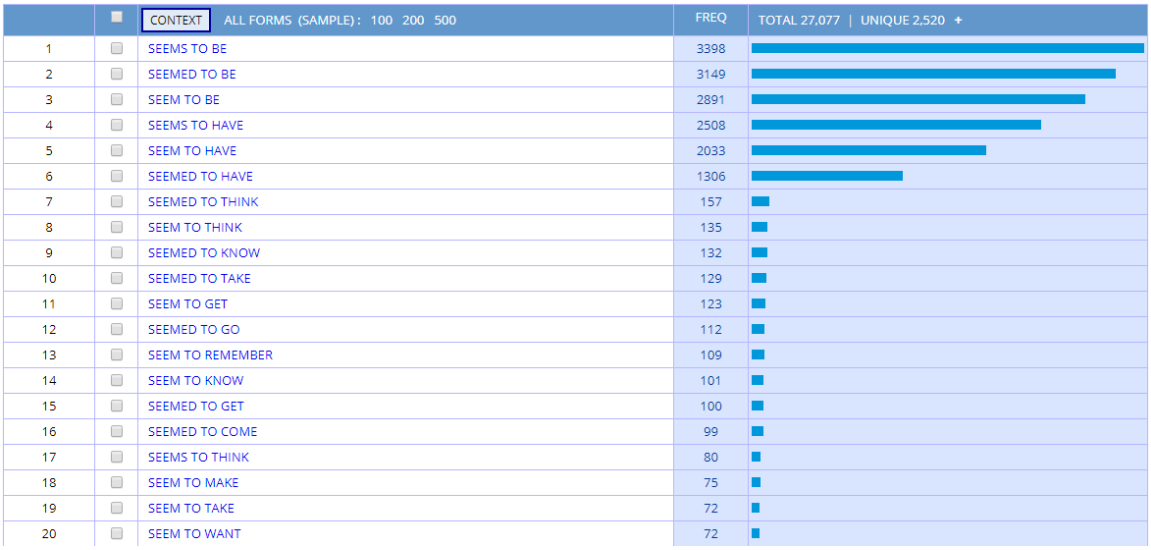
Из полученных данных видно, что наиболее частотным словосочетанием является *seem(s/ed)+to*:

If they ***seemed to*** *be alright then…* [COCA].

Actually the enterprise ***seems to*** *have fizzled out* [COCA].

It doesn't ***seem to*** *me that it would be reasonable* [BNC].

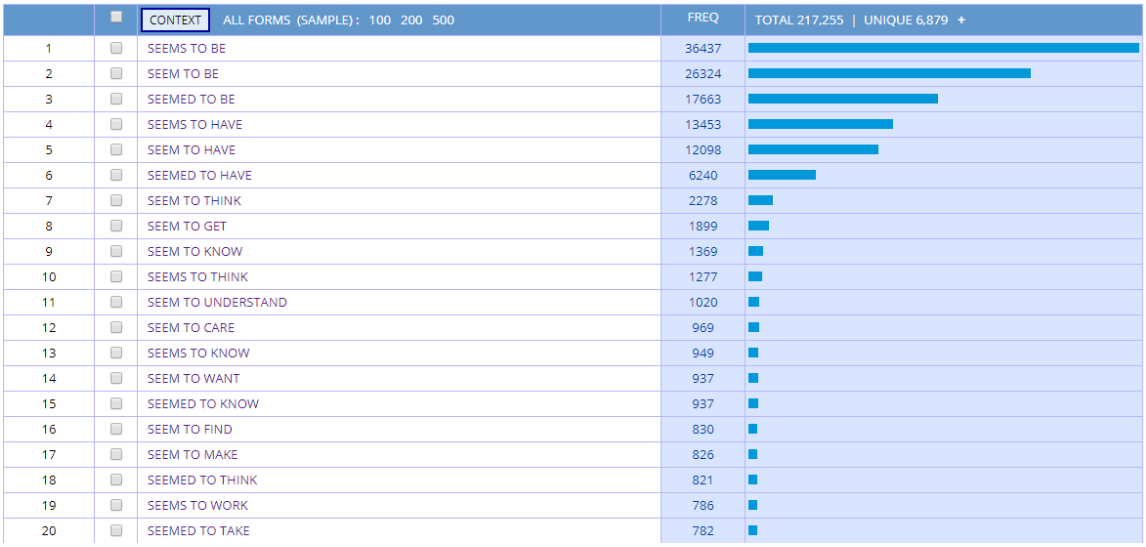
Чаще всего после конструкции *seem(s/ed) + to* следует глагол в той или иной форме. С помощью запроса формата *seem* _to _v** узнаем количество употреблений таких конструкций, и какие именно глаголы употребляются чаще всего (см. Рис. 11 и 12).



		CONTEXT ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL 27,077 | UNIQUE 2,520 +
1		SEEMS TO BE	3398	
2		SEEMED TO BE	3149	
3		SEEM TO BE	2891	
4		SEEMS TO HAVE	2508	
5		SEEM TO HAVE	2033	
6		SEEMED TO HAVE	1306	
7		SEEMED TO THINK	157	
8		SEEM TO THINK	135	
9		SEEMED TO KNOW	132	
10		SEEMED TO TAKE	129	
11		SEEM TO GET	123	
12		SEEMED TO GO	112	
13		SEEM TO REMEMBER	109	
14		SEEM TO KNOW	101	
15		SEEMED TO GET	100	
16		SEEMED TO COME	99	
17		SEEMS TO THINK	80	
18		SEEM TO MAKE	75	
19		SEEM TO TAKE	72	
20		SEEM TO WANT	72	

Рис. 11. Частотность seem(s/ed)+V в Британском национальном корпусе текстов

Видно, что в корпусе BNC самыми распространенными глаголами оказались *to be, to have.*



		CONTEXT ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL 217,255 | UNIQUE 6,879 +
1		SEEMS TO BE	36437	
2		SEEM TO BE	26324	
3		SEEMED TO BE	17663	
4		SEEMS TO HAVE	13453	
5		SEEM TO HAVE	12098	
6		SEEMED TO HAVE	6240	
7		SEEM TO THINK	2278	
8		SEEM TO GET	1899	
9		SEEM TO KNOW	1369	
10		SEEMS TO THINK	1277	
11		SEEM TO UNDERSTAND	1020	
12		SEEM TO CARE	969	
13		SEEMS TO KNOW	949	
14		SEEM TO WANT	937	
15		SEEMED TO KNOW	937	
16		SEEM TO FIND	830	
17		SEEM TO MAKE	826	
18		SEEMED TO THINK	821	
19		SEEMS TO WORK	786	
20		SEEMED TO TAKE	782	

Рис. 12. Частотность seem(s/ed)+V в Корпусе современного американского английского языка

По результатам видно, что самыми распространенными глаголами в корпусе COCA стали *to be, to have.*

Суммируя данные обоих корпусов, можно утверждать, что наиболее распространенными глаголами являются *to be* и *to have.* Приведем несколько примеров.

Scotland ***seem to have*** *got away with it at the moment* [BNC].

But it is a task with which John ***seems to be*** *coping remarkably well* [BNC].

The details of the decision ***seem to be*** *misrepresented here* [COCA].

The problem is they don't ***seem to have*** *to report to anyone* [COCA].

В этой конструкции *to seem,* как правило, указывает на отсутствие уверенности говорящего в том, что создавшееся впечатление правильное (см. Приложение 13).

Проанализируем коллокацию предиката *to appear* с помощью национальных корпусов. По результатам последовательного исследования всех форм предикатов в обоих корпусах, были получены определенные данные, которые после обработки и обобщения представлены в таблице (см. Таблица 5).

Таблица 5

Частотность употребления глагола appear и его форм в корпусах COCA/BNC

	Appear	Appears	Appeared	Appearing	Will appear	To appear	Would appear	Общее кол-во
COCA BNC	73937 10597	71161 7480	65539 10032	10746 1394	4137 520	10658 1665	4283 1064	240461 32752

Согласно полученным данным, можно утверждать, что наиболее частотными являются формы *appear, appears* и *appeared*. Однако в сравнении с аналогичными формами глагола *to seem*, формы глагола *to appear* уступают в количественном соотношении в два и более раза. Наименее частотной в обоих корпусах является форма будущего времени *will appear*.

Сравним данные по обоим глаголам. Суммарное количество употреблений глагола *to seem* в корпусах COCA и BNC – 596 883, а глагола *to appear* – 273 213. Таким образом, еще раз подтверждаются данные полученные в ходе использования сервиса Google Books Ngram Viewer и предположение о более высокой частотности употребления глагола *to seem*.

Рассмотрим коллокацию слова *to appear* в национальных корпусах. Последовательность работы была такая же, как с предыдущим предикатом. С помощью функции *Collocates* был получен список коллокатов, упорядоченный по частотности их употребления (см. Рис. 13 и 14).

		CONTEXT	FREQ		ALL	%	MI	
1	☐	TO	11015		2565070	0.43	3.53	
2	☐	BEFORE	575		84032	0.68	4.20	
3	☐	REGULARLY	40		3787	1.06	4.82	
4	☐	SUDDENLY	33		10919	0.30	3.02	
5	☐	BESIDE	22		5349	0.41	3.46	
6	☐	ALONGSIDE	21		3172	0.66	4.15	
7	☐	UNLIKELY	20		5480	0.36	3.29	
8	☐	SOMEWHAT	19		4498	0.42	3.50	
9	☐	TWICE	19		5988	0.32	3.09	
10	☐	REASONABLE	19		6091	0.31	3.06	
11	☐	WILLING	16		3897	0.41	3.46	
12	☐	CALM	14		2753	0.51	3.77	
13	☐	BRIEFLY	14		3152	0.44	3.57	
14	☐	ANYWHERE	13		3917	0.33	3.15	
15	☐	REMOTE	10		2772	0.36	3.27	
16	☐	UNWILLING	9		952	0.95	4.66	
17	☐	SIMULTANEOUSLY	9		1715	0.52	3.81	
18	☐	RELUCTANT	9		1931	0.47	3.64	
19	☐	UNEXPECTEDLY	8		864	0.93	4.63	
20	☐	FOOLISH	8		1088	0.74	4.30	

Рис. 13. Коллокация предиката to appear в Британском национальном корпусе текстов

Далее представлены несколько примеров с глаголом *to appear* и самими часто употребляемыми коллокациями в корпусе BNC. Остальные примеры можно найти в приложении (см. Приложение 6). .

*Now they **appear to** be trying to take the people of Norris Green down the same road and we refuse to go* [BNC].

*She must be told to **appear before** the Committee or be forced to take the* [BNC].

*Since the mid-1970s these clocks **appeared regularly** at auction and been valued consistently at around $4,000-6,000* [BNC].

		CONTEXT	FREQ		ALL	%	MI	
1		TO	74875		25557793	0.29	3.44	
2		ON	17772		6901553	0.26	3.25	
3		BEFORE	2583		822638	0.31	3.54	
4		WITHIN	673		255537	0.26	3.28	
5		REGULARLY	254		25660	0.99	5.19	
6		POISED	184		6594	2.79	6.69	
7		UNLIKELY	178		27384	0.65	4.59	
8		FREQUENTLY	161		40068	0.40	3.89	
9		HEADED	151		38983	0.39	3.84	
10		ALONGSIDE	144		17336	0.83	4.94	
11		BESIDE	144		34247	0.42	3.96	
12		SOMEWHAT	135		41949	0.32	3.57	
13		ANYWHERE	133		54040	0.25	3.18	
14		SLIGHTLY	129		55454	0.23	3.10	
15		INCREASINGLY	124		44981	0.28	3.35	
16		BRIEFLY	122		19386	0.63	4.54	
17		WEAK	107		36461	0.29	3.44	
18		ONSTAGE	87		4668	1.86	6.11	
19		CALM	83		37887	0.22	3.02	
20		FEES	72		25422	0.28	3.39	

Рис. 14. Коллокация предиката appear в Корпусе современного американского английского языка

Далее представлены несколько примеров с глаголом *to appear* и самими часто употребляемыми коллокациями в корпусе COCA. Остальные примеры представлены в приложении (см. Приложение 5).

*Especially since there didn't **appear to** be any damage to the photo or frame* [COCA].

*The same information about Miller **appears on** a variety of other websites, including at the American Legion and Project Vote Smart, where it specifies that he served in the U.S. Army* [COCA].

*The fisherman repeated the words, and the fish **appeared before** him* [COCA].

Можно увидеть, что самым частотным словосочетанием является *appear(s/ed)+to*:

*The pace of job growth **appears** to be slowing down in San Francisco* [COCA].

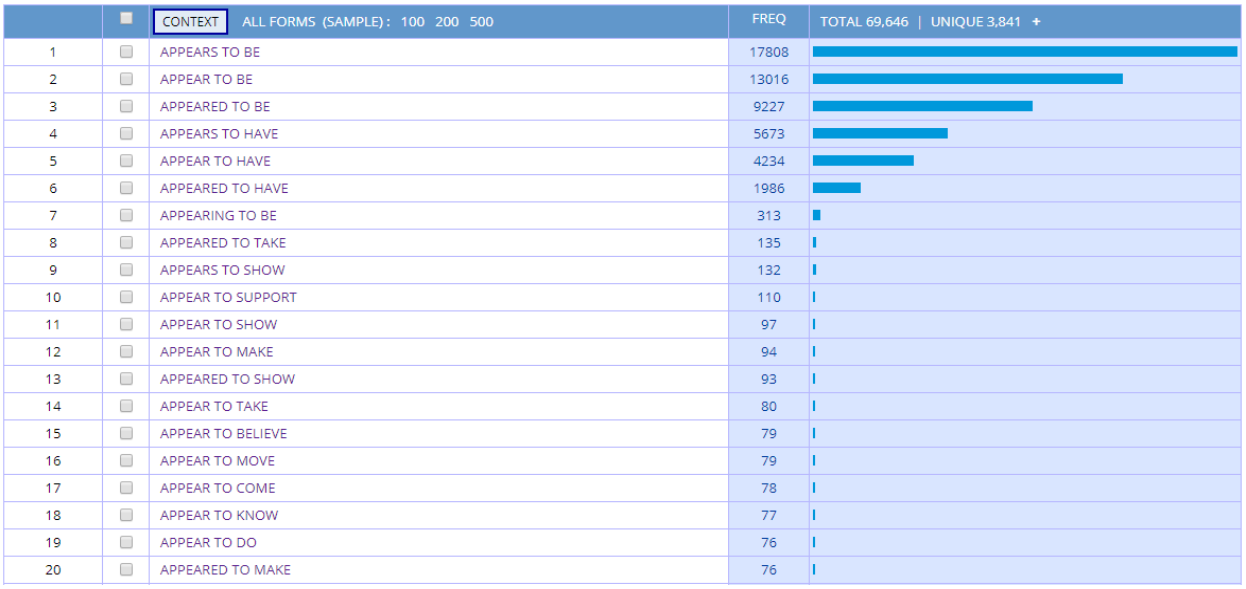
*Yet the stock returns for Abbott don't **appear** to be as impressive* [BNC].

После конструкции *appear(s/ed) + to* чаще всего следует глагол в той или иной форме. С помощью запроса формата *appear* _to _v** узнаем количество употреблений конструкций с глаголами, а также выясним, какие именно глаголы употребляются чаще всего (см. Рис. 15).

		CONTEXT ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL 10,288 | UNIQUE 1,313 +
1		APPEAR TO BE	1821	
2		APPEARS TO BE	1792	
3		APPEARED TO BE	1375	
4		APPEARS TO HAVE	1006	
5		APPEAR TO HAVE	872	
6		APPEARED TO HAVE	463	
7		APPEARING TO BE	47	
8		APPEARED TO TAKE	28	
9		APPEAR TO MAKE	23	
10		APPEARED TO OFFER	23	
11		APPEARED TO MAKE	22	
12		APPEAR TO OFFER	19	
13		APPEARED TO CONFIRM	19	
14		APPEAR TO SHOW	18	
15		APPEARED TO GIVE	18	
16		APPEAR TO TAKE	17	
17		APPEARED TO ACCEPT	16	
18		APPEARING TO HAVE	15	
19		APPEAR TO BELIEVE	14	
20		APPEAR TO DO	14	

Рис. 15. Частотность appear(s/ed)+V в Британском национальном корпусе текстов

Видно, что в корпусе BNC самыми распространенными глаголами оказались *to be, to have* (см. Рис. 16).



		CONTEXT ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL 69,646 | UNIQUE 3,841 +
1		APPEARS TO BE	17808	
2		APPEAR TO BE	13016	
3		APPEARED TO BE	9227	
4		APPEARS TO HAVE	5673	
5		APPEAR TO HAVE	4234	
6		APPEARED TO HAVE	1986	
7		APPEARING TO BE	313	
8		APPEARED TO TAKE	135	
9		APPEARS TO SHOW	132	
10		APPEAR TO SUPPORT	110	
11		APPEAR TO SHOW	97	
12		APPEAR TO MAKE	94	
13		APPEARED TO SHOW	93	
14		APPEAR TO TAKE	80	
15		APPEAR TO BELIEVE	79	
16		APPEAR TO MOVE	79	
17		APPEAR TO COME	78	
18		APPEAR TO KNOW	77	
19		APPEAR TO DO	76	
20		APPEARED TO MAKE	76	

Рис. 16. Частотность appear(s/ed)+V в Корпусе современного американского английского языка

По результатам видно, что самыми распространенными глаголами в корпусе COCA стали *to be, to have.*

Согласно полученным данным можно утверждать, что самыми распространенными глаголами являются *to be, to have.*

Meanwhile waxwings certainly ***appear to be*** *making their way back north after the big flocks recorded here in early December generally dispersed southwards* [BNC].

The public's political knowledge ***appears to have*** *increased* [BNC].

The most common cause ***appears to be*** *heat stress arising from climate change* [COCA].

It's a strategy that ***appears to have*** *worked in the past* [COCA].

Анализ примеров позволяет утверждать, что в данной конструкции глагол *to appear* показывает, что мнение

основано на общем впечатлении об объекте / ситуации (см. Приложение 14).

Отметим сходство в сочетаемости предикатов *to seem* и *to appear*, поскольку у обоих глаголов наиболее частотной конструкцией является *seem / appear(s/ed) + to*. Более того, схожей является сочетаемость полученной конструкции со стоящими далее глаголами, которыми стали *to be* и *to have*.

В следующих по частотности употреблений конструкциях, проявляются различия. В то время как глагол *seem(s/ed)* часто используется с наречиями, например: *likely, unlikely, almost*, после глагола *apeear(s/ed)* чаще следуют предлоги: *on, before, within.*

Определим стилевую стратификацию изучаемых предикатов *to seem* и *to appear*. Проведем анализ частотности употребления глаголов в различных стилях и уточним, могут ли они являться стилистическими маркерами для какого-либо типа дискурса. Для этого в корпусах BNC и COCA в разделе *Chart* задается поисковый запрос формата *seem** и *appear**. Результатом поиска является гистограмма, отображающая количество употреблений искомого глагола по разделам и подразделам (см. Рис. 17-20).

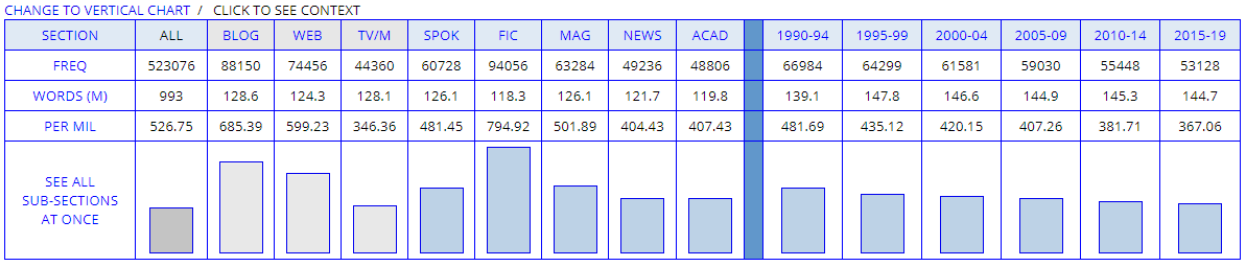


Рис. 17. Определение стратификации глагола to seem в COCA

CHANGE TO VERTICAL CHART / CLICK TO SEE CONTEXT

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	59122	4053	17250	4205	4277	8361	10958	10018
WORDS (M)	100	10.0	15.9	7.3	10.5	16.5	15.3	20.8
PER MIL	591.22	406.78	1,084.27	579.04	408.64	506.88	714.73	480.82
SEE ALL SUB-SECTIONS AT ONCE								

Рис. 18. Определение стратификации глагола to seem в BNC

CHANGE TO VERTICAL CHART / CLICK TO SEE CONTEXT

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	221371	26094	33174	8382	17038	28657	34779	29857	41390	28842	27305	26854	24822	24019	28261
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8	139.1	147.8	146.6	144.9	145.3	144.7
PER MIL	222.93	202.89	283.08	65.45	135.08	242.19	275.82	245.25	345.52	207.41	184.77	183.22	171.25	165.35	195.25
SEE ALL SUB-SECTIONS AT ONCE															

Рис. 19. Определение стратификации глагола to appear в COCA

CHANGE TO VERTICAL CHART / CLICK TO SEE CONTEXT

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	NON-ACAD	ACADEMIC	MISC
FREQ	29507	775	3962	2089	3358	5943	7089	6291
WORDS (M)	100	10.0	15.9	7.3	10.5	16.5	15.3	20.8
PER MIL	295.07	77.78	249.04	287.66	320.84	360.29	462.38	301.94
SEE ALL SUB-SECTIONS AT ONCE								

Рис. 20. Определение стратификации глагола to appear в BNC

Анализ глагола *seem(s/ed)* в Корпусе современного американского английского показал, что частотность слова составляет 532076 единиц. Согласно графику частотности, можно сделать вывод, что глагол *to seem* чаще всего употреблялся в 1990-1994 годах. В последующие годы разница в количестве употреблений составляет около 1000 в год. Глагол *to seem* чаще используется в Интернете, в блогах и на веб-сайтах. Наибольшие значения показывают категории *Дискуссия* 93662 и *Разное* 34388. Второй категорией по частотности является *Художественная литература*. Наименьшее число употреблений зафиксировано в сфере телевидения – 44752 употреблений. Количество употреблений в разговорном дискурсе составляет 60728.

Больше примеров представлено в приложении (см. Приложение 8).

*It **seems** like it may be a really good starter book for reporters or photographers who are just getting started* (Blog).

*It **seems** only those films are in that universe* (Web).

*So it **seems** that even Buster believes the story that happened in 1982* (TV).

*Well it **seemed** to be the motif* (Spoken).

*The boy, blond and pale, **seemed** no older than nine, though in Europe Richard had found it difficult to judge* (Fiction).

*Either way, **it seems** trade talks are back on* (Magazine).

*Do not be afraid to call this what it **seems** to be* (Newspaper).

*We also found that spatial memory **seemed** to be more affected at older ages* (Academic).

Анализ данных Британского национального корпуса текстов показал, что частотность глагола *seem(s/ed)* составляет 59122 единицы, информация о частотности употребления по годам отсутствует. Раздел *Интернет* также не представлен в данном корпусе. Глагол *to seem* чаще всего используется в художественной литературе – с частотностью 17250 единиц. Самое большое количество употреблений в данном подкорпусе приходится на прозу – 17110. Вторым по количеству употреблений является академический дискурс с количеством использований 10958. Чаще всего глагол употребляется в области гуманитарных наук 3491 и

социологии 3446. Количество употреблений в разговорном дискурсе составляет 4076. Больше примеров представлено в приложении (см. Приложение 9).

*Though we do **seem** to be doing quite well in numbers* (Spoken).

*He looks round, **seems** doubtful, then the explanation* (Fiction).

*He **seems** to have a natural ability which encourages plants to grow well* (Magazine).

*Something **seemed** to fall off the cockpit* (Newspaper).

He seemed a very pleasant person but he spoke no Bengali (Nonacademic).

*These are radical suggestions which may **seem** impractical; but the status quo may be indefensible* (Academic).

*These two opposite conditions of temperature **seemed** impossible to reconcile* (Misc).

Глагол *to seem* часто употребляется в британской прозе и научном дискурсе, реже всего используется сфере телевидения. Также *to seem* часто употребляется в американской художественной литературе и Интернете. Глагол *to seem* часто встречается в разговорном стиле в обоих корпусах.

Далее в Корпусе современного американского английского проанализируем глагол *appear(s/ed)*. Исследование показало, что частотность использования слова составляет 221371 единицу. Согласно графику частотности, можно сделать вывод, что глагол *to appear* чаще всего употреблялся в 1990-1994 годах и в последние годы –

2015-2019. В промежутке между ними разница в количестве употреблений составляет около 2000 в год. Глагол *to appear* чаще используется в Интернете (74604). Он часто встречается в академическом дискурсе – 41390 употреблений. Здесь больше всего употреблений приходится на гуманитарные науки (7859), социологию (7499) и технические науки (5238). Третьей по частотности категорией являются *журналы* – 34779. Так, глагол *to appear* чаще всего встречается в колонках о науке и технологиях (9158), новостях (8119) и здоровье (4676). Самое маленькое количество употреблений зафиксировано в сфере телевидения – 8382 примера. Количество употреблений в разговорном дискурсе составляет 17038. Примеры представлены в приложении (см. Приложение 10).

This is not as straightforward as it might at first ***appear*** (Blog).

The full name of the victim has still not been released, but it ***appears*** *her last name was Romero* (Web).

It ***appears*** *your sister lied to me* (TV).

As we mentioned earlier, we asked the mayor to ***appear****, but like her predecessor, Rahm Emmanuel, she declined* (Spoken).

Before she has to answer, Mallory ***appears*** *in the kitchen doorway* (Fiction).

Note that eczema in children is more likely to ***appear*** *in the bends of elbows and knees* (Magazine).

Apparently I am mistaken as it ***appears*** *it is the president who writes the laws, then Congress approves them* (Newspaper).

The idea is clear, but it ***appears*** *that it is not well presented by the figure* (Academic).

Анализ данных Британского национального корпуса текстов показал, что частотность глагола *appear(s/ed)* составляет 29507 единиц, информация о частотности употребления по годам отсутствует. Раздел *Интернет* не представлен. Глагол *to appear* чаще всего используется в академическом дискурсе – 7089 единиц. Самое большое количество употреблений в данном подкорпусе приходится на гуманитарные науки (1888), юриспруденцию (2025) и социологию (1917). Вторым является неакадемический дискурс с количеством использований 5943. В этом разделе лидером является группа художественных дисциплин (1772). Количество употреблений в разговорном дискурсе составляет 775. Больше примеров представлено в приложении (см. Приложение 11).

Also, the roof is leaking in several places, there are cracks ***appearing*** *on the parapet* (Spoken).

But Sven Hjerson ***appeared*** *not to have heard* (Fiction).

At about 1.40pm cars started to ***appear*** *and people were seen walking up the road* (Magazine).

The problem is that when a new trick ***appears****, everyone is trying to see what advantages they can get from it* (Newspaper).

*At first the social worker **appeared** willing and helpful* (Non-academic).

*Subsequently a little girl **appeared** in the doorway with the little boy* (Academic).

*The safer the loan **appears** to be, the lower, in general, the interest rate* (Misc).

Глагол *to appear* часто употребляется в американском и британском академическом дискурсе, а именно в гуманитарных науках. Также этот глагол часто употребляется в американских журналах. *To appear* чаще встречается в американском разговорном дискурсе, чем в британском. Представим все полученные данные в таблицах (см. Таблицы 6 и 7).

Таблица 6

Стилевая стратификация глаголов to seem/to appear в корпусе COCA

COCA	Блог	Веб	ТВ/ фильм	Устн.	Худ.лит.	Журн.	Газета	Академ.
To seem	88150	74456	44360	60728	94056	63284	49236	48806
To appear	26094	33174	8382	17038	29657	34779	29857	41390

Таблица 7

Стилевая стратификация глаголов to seem/to appear в корпусе BNC

BNC	Устн.	Худ.лит.	Журнал	Газета	Не академ.	Академ.	Разное
To seem	4053	17250	4205	4277	8361	10958	10018
To appear	775	3962	2089	3358	5943	7089	6291

Сравнивая полученные данные, можно отметить, что оба глагола часто встречаются в сети Интернет. Глагол *to seem*

часто используется в художественной литературе, в отличие от глагола *to appear,* который является более частотным в академическом дискурсе. В устной речи более распространенным является *to seem*.

Для того чтобы выявить возможность взаимозаменяемости глаголов, было проведено анкетирование среди носителей языка. Участникам предлагалась анкета, содержащая 20 предложений с глаголами *to seem / to appear*. Примеры были отобраны с помощью поискового запроса *seem** и *appear** в корпусах COCA и BNC. Каждый пункт анкеты содержит оригинальное предложение и то же самое предложение, но с другим глаголом. Респонденту необходимо было отметить каждое из них как верное / неверное / возможное. Пример из анкеты представлен в таблице 8, полные результаты анкетирования размещены в приложении (см. Приложение 17).

Таблица 8

Пример анкетирования с целью выяснить взаимозаменяемость глаголов

	Correct	Wrong	Possible
I **seem** plagued by convoluted sentences today -- sorry!			
I **appear** plagued by convoluted sentences today -- sorry!			

Всего в анкетировании приняли участие 5 носителей языка из США, Великобритании, Австралии и Канады. Два респондента являются преподавателями английского языка. Результаты представлены в таблице 9.

Таблица 9

Результаты анкетирования с целью выяснить взаимозаменяемость глаголов

	Согласился с оригиналом	Согласился с неоригиналом	Выбрал оба варианта	Отметил оригинал как возможное	Отметил не оригинал как возможное
Респондент 1	9	5	6	1	0
Респондент 2	7	10	3	0	0
Респондент 3	5	1	14	1	2
Респондент 4	6	6	8	4	2
Респондент 5	4	5	11	1	1

Согласно полученным данным, можно сделать вывод, что многие респонденты затруднились с определением оригинального предложения, часто выбирая оба варианта как правильные. Кроме того, некоторые опрашиваемые испытывали затруднения при выборе и отмечали, что данные глаголы чаще всего взаимозаменяемы. Случаи, когда респондент выбирал оба варианта, часто обусловлены тем, что вариант предложения с глаголом *to appear*, является более формальным. Таким образом, проведенный опрос носителей английского языка подтверждает взаимозаменяемость глаголов и более высокую степень формальности глагола *to appear*.

Для того, чтобы доказать или опровергнуть утверждение о том, что *to seem* указывает на неопределенность исходящую от наблюдателя / субъекта, в то время как *to appear* указывает на неопределенность, вызванную особенностями

объекта, был проведен опрос носителей языка. Для этого с помощью поискового запроса *seem** и *appear**, были отобраны примеры употребления глаголов *to seem, to appear* в корпусах COCA и BNC. Анкета состоит из 20 пунктов, по 1 предложению в каждом. Респонденту был предложено предложение и три варианта ответа неопределенность от субъекта / неопределенность от объекта / затрудняюсь ответить. Пример представлен в таблице 10, полная анкета представлена в приложении (см. Приложение 18).

Таблица 10

Пример анкетирования с целью выяснить откуда исходит неопределенность

Uncertainty comes from:	Subject/person	Object/ situation	Not sure
I kept in touch with Joe -- there were no other friends. But they **appeared** to me as if seen through the wrong end of a telescope, muted and unreal			

В анкетировании поучаствовали те же самые 5 информантов. Результаты представлены в таблице 11.

Таблица 11

Результаты анализа значений возникновения неопределенности

Респондент 1	To seem	To appear
Неопределенность исходит от субъекта	6	2
Неопределенность исходит от объекта	1	6
Затрудняюсь ответить	5	

Респондент 2	To seem	To appear
Неопределенность исходит от субъекта	4	5
Неопределенность исходит от объекта	5	5
Затрудняюсь ответить	1	

Респондент 3	To seem	To appear
Неопределенность исходит от субъекта	6	3
Неопределенность исходит от объекта	1	2
Затрудняюсь ответить	8	

Респондент 4	To seem	To appear
Неопределенность исходит от субъекта	7	3
Неопределенность исходит от объекта	3	7
Затрудняюсь ответить	0	

Респондент 5	To seem	To appear
Неопределенность исходит от субъекта	9	3
Неопределенность исходит от объекта	1	7
Затрудняюсь ответить	0	

Таким образом, только анкеты респондентов 3 и 5 обладают показателями, соответствующими выдвинутому выше утверждению об источнике неопределенности (неопределенность исходит от субъекта *to seem* > неопределенность исходит от объекта *to appear*). В анкетах 1 и 4 эти показатели равны. В анкете респондента 2 показатели противоположны исходному утверждению. Учитывая результаты проведенного анкетирования носителей языка, нельзя с уверенностью утверждать, что фактор неопределенности влияет на выбор глагола.

Проведенное с помощью Big Data исследование доказывает наличие большого потенциала этих систем, который еще не в полной мере используется в лингвистике. При этом стоит отметить, что не все возможности больших данных для изучения предикатов были описаны и проанализированы в работе. Например, не представлена информация о возможности использования данных из

двуязычных корпусов для анализа особенностей перевода предикатов и выявления закономерностей. Также еще не анализировалась возможность использования диахронических корпусов для изучения развития значения предикатов *to seem* и *to appear*. Это может стать перспективой для дальнейшего изучения возможностей систем Big Data для лингвистических исследований.

Выводы по главе II

Во второй главе описано, как менялось понимание предикативности, приведены существующие классификации предикатов, перечислены предикативные конструкции и указаны их синтаксические функции; подробно представлена

семантика предикатов *to seem* и *to appear*, рассмотрена частотность их употребления в диахронии, этимология и др. Также проведен анализ потенциала некоторых систем Big Data для проведения лингвистических исследований, обозначены перспективы работы.

На основании результатов проведенного корпусного анализа глаголов *to seem* и *to appear* выделены их семантические свойства. В большинстве случаев они взаимозаменяемы и переводятся на русский язык глаголами несовершенного вида – казаться, оказываться. Однако существуют тонкости в их употреблении. *To appear* является более формальным и имеет более низкую частотность употребления. *To seem* обозначает мнение, основанное на субъективных впечатлениях и личном отношении, и приписывает погрешность оценки неким особенностям субъекта восприятия. Глагол *to appear* может означать, что мнение основано на общем визуальном впечатлении, приписывает погрешность оценки особенностям воспринимаемого объекта. В то время как глагол *to appear* выражает только идею появления как перцептивного события, глагол *to seem* передает все виды восприятия.

Оба глагола могут свидетельствовать о том, что говорящий высказывает предположение или не уверен в ситуации. Многие различия в значениях глаголов лежат в области синтезирующего восприятия, то есть в случаях, когда речь идет о цвете, форме, размер, четкости / нечеткости, эмоциональном состоянии. При этом *to seem* часто указывает на неуверенность говорящего в том, что создавшееся у него впечатление правильно; *to appear*

указывает на то, что мнение основано на общем впечатлении об объекте или ситуации.

Предположение о том, что при использовании глагола *to seem* неопределенность исходит от наблюдателя, а при использовании глагола *to appear* неопределенность вызвана характеристиками человека или объекта, в ходе исследования подтверждено не было.

Анализ сочетаемости показал сходство в сочетаемости слов *to seem* и *to appear*. У обоих глаголов, наиболее частотной конструкцией является *seem / appear(s/ed) + to*. Схожей является сочетаемость полученной конструкции со стоящими далее глаголами, которыми чаще всего были *to be* и *to have*. В следующих по частотности употреблений конструкциях выявлены различия: глагол *seem(s/ed)* часто используется с наречиями, например: *likely, unlikely, almost*, а после глагола *apeear(s/ed)* чаще следуют предлоги: *on, before, within.*

В результате определения стилевой стратификация предикатов *to seem* и *to appear*, а также анализа частотности их употребления в различных стилях и дискурсах, можно сделать вывод, что в устной речи предпочтение отдается глаголу *to seem*, а глагол *to appear* является более формальным. Оба глагола часто встречаются в сети Интернет. Глагол *to seem* часто употребляется в художественной литературе, а также в исследованиях (гуманитарные науки и социология). Глагол *to appear* чаще употребляется в американском и британском академическом дискурсе (социология, гуманитарные и технические науки). Глагол часто встречается в журналах, в колонках о науке и

технологиях, новостях и здоровье. Реже всего оба глагола используются в сфере телевидения.

Анализ предложений с глаголами *to seem* и *to appear* в программе SentiStrength показал, что чаще всего данные глаголы встречаются в предложениях, имеющих нейтральную окраску.

Заключение

В представленной работе уточнен потенциал систем Big Data для лингвистических исследований. Для этого рассмотрено понимание предикативности, описана семантика предикатов *to seem* и *to appear,* проведено исследование их семантических свойств и значений, указана частотность и сочетаемость, обозначены перспективы исследования.

Проведенный анализ научных исследований показал, что предоставляемые системами Big Data возможности находят свое применение во всех областях научного знания. В современных лингвистических исследованиях наиболее популярными становятся методы корпусной лингвистики. В работе представлены несколько классификаций корпусов с их подробным описанием. Сделан вывод, что с помощью информации, полученной после анализа корпусных данных, исследователям открываются прежде недоступные сведения о закономерностях языка и отдельных типах текста. Работа с

корпусом осуществляется с помощью специальных программных средств – конкордансеров и корпус-менеджеров.

Применение систем Big Data для исследования предикатов позволило получить данные о частотности их употребления, этимологии, коллокации, семантике и особенностях использования, как в разных станах, так и в различных жанрах, дискурсах и стилях. При этом с помощью корпусного анализа глаголов были получены новые данные о частотности употребления схожих по семантике глаголов в различных функциональных стилях.

Проведенный анализ теоретических источников и корпусных данных позволил прийти к выводу, что предикаты *to seem* и *to appear* часто являются взаимозаменяемыми и переводятся на русский язык глаголами *казаться, оказываться*. При этом *to seem* является более частотным и менее формальным.

Некоторые исследователи указывают на различия в семантических значениях данных предикатов, однако при проведении анализа корпусных данных, не все они нашли подтверждение. Так предположение о том, что при использовании глагола *to seem* неопределенность исходит от наблюдателя, а при использовании глагола *to appear* неопределенность вызвана характеристиками человека или объекта, в ходе исследования не подтвердилось.

Исследование также показало сходную сочетаемость глаголов *to seem* и *to appear*. Наиболее частой конструкцией стало *seem / appear(s/ed) + to*. Схожей является сочетаемость полученной конструкции со стоящими далее глаголами,

которыми чаще всего были *to be* и *to have*. В следующих по частотности употреблений конструкциях выявлены различия: глагол *seem(s/ed)* часто используется с наречиями, например: *likely, unlikely, almost*, а после глагола *appear(s/ed)* чаще следуют предлоги: *on, before, within.*

Описывая перспективы исследования, была отмечена возможность использования данных из двуязычных корпусов для анализа особенностей перевода предикатов и выявления закономерностей. Видится возможным использование диахронических корпусов для изучения развития семантики предикатов *to seem* и *to appear*. Также перспективой дальнейшего исследования являются: изучение возможностей других программных средств, предоставляемых системами Big Data; уточнение специфики лингвистических исследований с применением Big Data.

Подводя итог, можно сделать вывод, что системы Big Data открывают широкие перспективы для проведения различных лингвистических исследований. Полученные в ходе этой работы данные, могут стать основой для дальнейшего исследования предикатов, а также применения в лингвистических исследованиях разнообразных технологий систем Big Data (например, различных типов текстовых корпусов, Google Trends, Google Books Ngram Viewer и др.).

Библиография

I. ЛИТЕРАТУРА

1.Апресян Ю.Д. Синонимический ряд выглядеть, казаться. II Ю.Д. Апресян, О.Ю.Богуславская, Т.В.Крылова, И.Б.Левонтина, Е.В. Урысон и др. Новый объяснительный словарь синонимов русского языка. Под общим рук.акад. Ю.Д.Апресяна. Вып. 2. М., 2000. – С.61-88.

2.Англо-русский синонимический словарь / Апресян Ю.Д ., Ботякова В.В., Латышева Т.Э. и др.: под рук. Розенмана А.И. и Апресяна Ю.Д. М.: рус. яз., 2000. – 544 с.

3.Арутюнова Н.Д. Язык и мир человека. М.: Языки русской культуры, 1999. – 896 с.

4.Арутюнова Н.Д. Сокровенная связка (к проблеме предикативного отношения) // Известия АН СССР / Серия литературы и языка. – Т. 39. – 1980. – №4. – С. 392-393.

5.Афанасьева О.В. Адъективный класс лексики в современном английском языке и формы его языковой репрезентации: диссертация доктора филологических наук: 10.02.04. – Москва, 1994. – 395 с.

6.Баранов А.Н. Корпусная лингвистика / Баранов А.Н. // Введение в прикладную лингвистику. – М.: Едиториал УРСС, 2001. – С. 51-52.

7.Берман И.М. Грамматика английского языка. Курс для самообразования. – М., «Высшая школа», 1993. – С. 162-165.

8.Виноградов В.В. Основные вопросы синтаксиса предложения (на материале русского языка) // Введение в языкознание. Хрестоматия. – М.: Аспект Пресс, 2001. – С. 204-222.

9.Владимов Н.В. Корпусный подход к решению переводческих проблем: На материале письменных

переводов с русского языка на английский: дис. ... кан. филол. наук. – М., 2005. – 198 с.

10. Выготский Л.С. Мышление и речь (Извлечения) / Психолингвистика в очерках и извлечениях: Хрестоматия: – М.: Академия, 2003. – С. 280-285.

11. Гвишиани Н.Б. Практикум по корпусной лингвистике: Учеб.пособие по английскому языку/ Н.Б. Гвишиани. – М.: Высшая школа, 2008. – 191 с.

12. Горина О.Г. Использование технологий корпусной лингвистики для развития лексических навыков студентов-регионоведов в профессионально-ориентированном общении на английском языке: Дис. ... кан. пед. наук. – Спб., 2014. – 308 с.

13. Гречко В.А. Теория языкознания. – М.: Высшая школа, 2003. – 375 с.

14. Дорошевский В. Элементы лексикологии и семиотики. – М.: Прогресс, 1973. – С. 182-183.

15. Захаров В.П. Лингвистика больших корпусов / В.П. Захаров // Компьютерная лингвистика и вычислительные онтологии: сборник научных статей. – Спб.: НИУ ИТМО, 2015. – С. 82-93.

16. Захаров В.П. Поисковые системы Интернета как инструмент лингвистических исследований / В.П. Захаров // Русский язык в Интернете: Сб. статей. – Казань: Отечество, 2003. – С. 48-59.

17. Захаров В.П. Корпусная лингвистика: учебник для студентов гуманитарных вузов / В.П. Захаров, С.Ю. Богданова – Иркутск: ИГЛУ, 2011. – 161 с.

18. Ильчук Е.В. Мышление и восприятие сквозь призму языка (на материале английского языка). М.: Прометей, 2004. – 263 с.

19. Ильчук Е.В. Некоторые типы эпистемической модальности в английском языке: автореферат дис. кандидата филологических наук: 10.02.04 / Ин-т языкознания. – Москва, 1990. – 24 с.

20. Ильчук Е.В. Основные направления когнитивизма // Лингвистика на рубеже эпох: доминанты и маргиналии. Сборник статей. Вып. 2. / Сост. – О.А.Сулейманова и Н.Л. Огуречникова. – М.: МГПУ, 2004. – С.18-29.

21. Ильчук Е.В. Сравнительный анализ показателей эпистемической модальности в английском языке, выраженных глаголами восприятия (seem и appear) и модальными словами (probably) // Семантико-прагматические и социолингвистические аспекты изучения языка. Конференция молодых научных сотрудников и аспирантов. Тезисы докладов. – М.: Институт языкознания РАН, 1990. – С. 30-34.

22. Ильчук Е.В Представление когнитивных процессов в семантике английских глаголов: автореф. дис. на соиск. учен.степ. докт. филол. наук (10.02.04) / Ильчук Елена Вячеславовна; МПГУ. – Москва, 2006. – 46 с.

23. Кибрик А.Е., Брыкина М.М., Леоньев А.П., Хитров А.Н. Русские посессивные конструкции в свете корпусно-статистического исследования // Вопросы языкознания. 2006. – Вып. 1 – С. 16-45.

24. Козлова Н.В. Лингвистические корпуса: определение основных понятий и типология / Н.В. Козлова //

Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. – 2013. – Т. 11, Вып. 1. – С. 79-88.

25. Кравченко В.О., Крюкова А.А. «Большие данные» – практические аспекты и особенности // Academy. – 2016. – №6(9). – С. 65-67.

26. Лакофф Дж. Метафоры, которыми мы живем / Дж. Лакофф, М. Джонсон. – М.: Едиториал УРСС, 2004. – 256 с.

27. Лингвистический энциклопедический словарь / под ред. В.Н. Ярцева. – М.: Советская энциклопедия, 1990. – 685 с.

28. Новиков Д.А. Большие данные: отБраге к Ньютону // Проблемы управления. – 2013. – № 6. – С. 15-23.

29. Одинцов А.В. Социология общественного мнения и вызов BigData // Мониторинг общественного мнения: Экономические и социальные перемены. – 2017. – № 3. – С. 30-43.

30. Пащенко Ю.А. Предикативность и предикат в лингвистике и логике / Ю.А. Пащенко //вестник ТГПИ. – Таганрог: ТГПИ им. Чехова, 2006. – С. 70-72.

31. Потапова Т.В. Понятие предикативности в языкознании // Вестник Таганрогского института имени А.П. Чехова. – 2017. – № 6. – С. 30-35.

32. Потебня А.А. Мысль и язык (извлечения) / Психолингвистика в очерках и извлечениях: Хрестоматия. – М.: Академия, 2003. – С.100-101.

33. Плунгян В.А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики / В.А. Плунгян // Русский язык в научном

освещении №2 (16). – М.: РАН ИРЯ им. Виноградова, 2008. – С. 7-20.

34. Радченко И.А, Николаев И.Н. Технологии и инфраструктура BigData. – СПб: Университет ИТМО, 2018. – 52 с.

35. Рассел Б. Здравый смысл и ядерная война / Пер. с англ. В.М. Карзинкина. М.: Изд-во иностранной литературы, 1959. – 77 с.

36. Рыков В.В. Корпус текстов как реализация объектно-ориентированной парадигмы // Труды Международного семинара «Диалог 2002». – М.: Наука, 2002. – С. 59-61.

37. Семантические типы предикатов / под ред. О.Н. Селиверстовой. – М.: Наука, 1982. – С. 88-91.

38. Семенова Т.И. Лингвистический феномен кажимости. Иркутск, 2007. – 237 с.

39. Смирницкий А.И. Синтаксис английского языка. – М., Издательство литературы на иностранных языках, 1957. – 285 с.

40. Степанов Ю.С. В трехмерном пространстве языка: Семиотические проблемы лингвистики, философии, искусства. – М.: Книжный дом «Либроком», 2010. – 133 с.

41. Сулейманова О.А. Использование BIG DATA в экспериментальных лингвокогнитивных исследованиях: анализ семантической структуры глагола shudder / О.А. Сулейманова, В.В. Демченко // Когнитивные исследования языка. – Тамбов: Общероссийская общественная организация "Российская ассоциация лингвистов-когнитологов", 2018. – С. 466-472.

42. Сулейманова О.А. Экспланаторный потенциал теории классов для лингвистического исследования, порядок следования определений / О.А. Сулейманова, И.М. Петрова // Филология: научные исследования. – Москва: ООО "НБ-Медиа", 2018. – С. 52-64.

43. Сысоев П.В. Компетенция учителя иностранного языка в области использования информационно-коммуникационных технологий: определение понятий и компонентный состав/П.В. Сысоев // Иностранные языки в школе. – 2011. – № 6. – С. 16-20.

44. Харин А.В. Экосистема анализа больших данных hadoop: магистерская диссертация / А.В. Харин; Уральский федеральный университет имени первого Президента России Б. Н. Ельцина, Институт „Высшая школа экономики и менеджмента“, Кафедра анализа систем и принятия решений. – Екатеринбург, 2017. – 110 с.

45. Чернякова Т.А. Использование лингвистического корпуса в обучении иностранному языку // Язык и культура. 2011, №4. – С. 119-125.

46. Шаров С.А. Представительный корпус русского языка в контексте мирового опыта / С.А. Шаров // НТИ. Сер.2. – 2003. – №6. – С. 9-17.

47. Шевчук В.Н. Электронные ресурсы переводчика: Справочные материалы для начинающего переводчика. – М., 2010. – 44 с.

48. Aarts J., Meijs. W. Corpus Linguistics: Recent developments in the Use of Computer Corpora in English Language Research / J. Aarts, W. Meijs // Amsterdam: Rodopi. – 1984. – 425 p.

49. Boyd D., Crawford K. Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. //Information, Communication & Society. – 2012. – Vol. 15. – № 6. – pp. 662-679.

50. Deignan A.A. corpus-linguistic perspective on the relationship between metonymy and metaphor // Style. – 2005. – Vol. 39(1) – pp. 72-91.

51. Francis N.W. Language Corpora B.C. Directions in Corpus Linguistics// Proceedings of Nobel Symposium 82. Stockholm, 1991. – pp. 17-35.

52. Gray J. The Next Database Revolution // SIGMOD Conference. – 2004. – pp. 1-4.

53. McEnery T., Gabrielatos C. English Corpus Linguistics / T. McEnery, C. Gabrielatos // The Handbook of English Linguistics: McMahon-Blackwell Publishing, 2006 – pp. 33-72.

54. Meyer Ch.P. English Corpus Linguistics. An introduction / Ch.P. Meyer// Cambridge University Press, 2004. – 168 p.

55. Sinclair J.M. Linear Unit Grammar: Integrating Speech and Writing: Studies in Corpus Linguistics / J.M. Sinclair // Amsterdam/Philadelphia: John Benjamins, 2006. – 185 p.

56. Sinclair J. Corpus, Concordance, Collocation// Oxford: Oxford University Press, 1991. – 179 p.

57. Stubbs M. Words and phrases: corpus studies of lexical semantics// Oxford: Oxford University Press, 2001. – pp. 239-240.

58. Svartvik J. Corpus linguistics 25+ years on / J.Svartvik// Amsterdam, NY 2007. – pp. 11-27.

59. Tauheed F., Nobari S., Biveinis L., Heinis T., Ailamaki A. Computational Neuroscience Breakthroughs through Innovative Data Management. // ADBIS. – 2013. – pp. 14-27.

60. Tognini-Bonelli E. Corpus Linguistics at Work / E. TogniniBonelli// Amsterdam: John Benjamins Publishing Company, 2001. – 224 p.

II.ЭЛЕКТРОННЫЕ РЕСУРЫ

61. АбдусаламоваМ.М. Модальность предположения (на примере глаголов английского языка) // Известия ДГПУ. Общественные и гуманитарные науки. 2011. – №4. [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/n/modalnost-predpolozheniya-na-primere-glagolov-angliyskogo-yazyka (дата обращения: 07.10.2019).

62. Большаков А.С., Журенков О.В. Применение технологий BigData в гуманитарных исследованиях // Актуальные проблемы прикладной информатики в образовании, экономике, государственном и муниципальном управлении: сборник трудов международной научной конференции. – Барнаул, 2017. [Электронный ресурс]. – Режим доступа: https://elibrary.ru/item.asp?id=32406751 (дата обращения: 22.02.2020).

63. Захаров В.П. Лингвистические средства информационного поиска в Интернете – Библиосфера, 2015. [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/v/lingvisticheskie-sredstva-informatsionnogo-poiska-v-internete (дата обращения: 27.12.2018).

64. Иванов П.Д. Технологии BigData и их применение на современном промышленном предприятии / П.Д. Иванов, В.Ж. Вампилова // Инженерный журнал: наука и инновации. Вып.8. – М.: МГТУ им. Баумана, 2014. [Электронный ресурс]. – Режим доступа:http://engjournal.ru/catalog/it/asu/1228.html (дата обращения: 26.04.2020).

65. Нагель О.В., Корпусная лингвистика и ее использование в компьютеризированном языковом обучении // Язык и культура. 2008. – №4. – С.53-59. [Электронный ресурс]. – Режим доступа: http://cyberleninka.ru/article/n/korpusnaya-lingvistika-i-ee-ispolzovanie-v-kompyuterizirovannom-yazykovom-obuchenii (дата обращения: 25.11.2019).

66. Феномен bigdata // Век качества. 2014. – №4. – С.54-59. [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/n/fenomen-big-data (дата обращения: 10.12.2019).

67. Тиндал С. Большие данные: все, что вам необходимо знать. PCWeek/RE, 2012. – №25. – С. 18-22. (810). [Электронный ресурс]. – Режим доступа: https://www.itweek.ru/upload/iblock/803/PCWeekRE_2012_N25_web.pdf (дата обращения 03.06.2019).

68. Banerjee A., Monteleoni C. (2014). Climate change: challenges for machine learning. [Электронный ресурс]. – URL: http://www-users.cs.umn.edu/~banerjee/talks/BanerjeeMonteleoniNIPSTutorial2014.pdf (дата обращения: 13.02.2020)

69. Farias T., et al. (2008, October 29). A high performance massively parallel approach for real time deformable body physics simulation. [Электронный ресурс]. –

URL: https://www.researchgate.net/publication/221306607_A_High_Performance_Massively_Parallel_Approach_for_Real_Time_Deformable_Body_Physics_Simulation (дата обращения: 07.03.2020)

70. Gantz J., Reinsel D., Rydning J. The Digitization of the World. FromEdgetoCore. 2018. [Электронный ресурс]. – URL: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (дата обращения 16.02.2020).

71. Keating A. (2015, November 11). Machine learning could solve riddles of galaxy formation. [Электронный ресурс]. – URL: http://phys.org/news/2015-11-machine-riddles-galaxy-formation.html (дата обращения 16.02.2020).

72. Marai G. (2007 May). Thesis. [Электронный ресурс]. – URL: http://cs.brown.edu/research/pubs/theses/phd/2007/marai.pdf (дата обращения 16.02.2020).

73. Midling, A. (Jan, 2017). Using big data to understand immune system responses. [Электронный ресурс]. – URL:https://phys.org/news/2017-01-big-immune-responses.html (дата обращения 22.02.2020).

74. Scott J. (2012, July). High-performance computing drives high-performance F1 cars to success. [Электронный ресурс]. – URL: http://www.computerweekly.com/feature/High-performance-computing-drives-high-performance-F1-cars-to-success (дата обращения 22.02.2020).

75. Sterling, J. (Oct, 2017). Fully enabling genome-editing system for crop improvement launched. [Электронный ресурс]. – URL: https://www.genengnews.com/insights/fully-enabling-

genome-editing-system-for-crop-improvement-launched/ (дата обращения 22.02.2020).

76. NOAA (2015, January 5). NOAA announces significant investment in next generation of supercomputers. [Электронный ресурс]. – URL: http://www.noaanews.noaa.gov/stories2015/20150105_supercomputer.html. (дата обращения 15.03.2020).

III.ИНФОРМАЦИОННО–СПРАВОЧНЫЕ РЕСУРСЫ

77. Американский корпус английского языка [Электронный ресурс]. – URL:https://corpus.byu.edu/coca/

78. Британский корпус английского языка [Электронный ресурс]. –URL:https://corpus.byu.edu/bnc/

79. Брауновский корпус [Электронный ресурс]. – URL: http://corpus.leeds.ac.uk/

80. Большой Оксфордский Словарь [Электронный ресурс]. – URL: https://www.oxfordlearnersdictionaries.com/

81. Словарь Вебстера [Электронный ресурс]. – URL: https://www.merriam-webster.com/

82. Словарь современного английского языка Лонгмана[Электронный ресурс]. – URL: https://www.ldoceonline.com/

83. Кембриджский словарь [Электронный ресурс]. – URL: https://dictionary.cambridge.org/dictionary/english/data

84. Электронный тезаурус WordNet [Электронный ресурс]. – URL: http://wordnetweb.princeton.edu/perl/webwn

85. Англо-русский словарь Мюллера [Электронный ресурс]. – URL: https://starling.rinet.ru/morph.htm

86. Этимологический онлайн словарь [Электронный ресурс]. – URL: https://www.etymonline.com/

Приложение 1

Список и описание всех методик анализа массивов данных

Network Analysis. Набор методик анализа связей между узлами в сетях. Применительно к социальным сетям позволяет анализировать взаимосвязи между отдельными пользователями, компаниями, сообществами и т.п.

Optimization. Набор численных методов для редизайна сложных систем и процессов для улучшения одного или нескольких показателей. Помогает в принятии стратегических решений, например, состава выводимой на рынок продуктовой линейки, проведении инвестиционного анализа и проч.

Pattern Recognition. Набор методик с элементами самообучения для предсказания поведенческой модели потребителей.

Predictive Modeling. Набор методик, которые позволяют создать математическую модель наперед заданного вероятного сценария развития событий. Например, анализ базы данных CRM-системы на предмет возможных условий, которые подтолкнут абоненты сменить провайдера.

Regression. Набор статистических методов для выявления закономерности между изменением зависимой переменной и одной или несколькими независимыми. Часто применяется для прогнозирования и предсказаний. Используется в data mining.

Sentiment Analysis. В основе методик оценки настроений потребителей лежат технологии распознавания естественного языка человека. Они позволяют вычленить из общего информационного потока сообщения, связанные с интересующим предметом (например, потребительским продуктом). Далее оценить полярность суждения (позитивное или негативное), степень эмоциональности и проч.

Signal Processing. Заимствованный из радиотехники набор методик, который преследует цель распознавания сигнала на фоне шума и его дальнейшего анализа.

Spatial Analysis. Набор отчасти заимствованных из статистики методик анализа пространственных данных – топологии местности, географических координат, геометрии объектов. Источником больших

данных в этом случае часто выступают геоинформационные системы (ГИС).

Statistics. Наука о сборе, организации и интерпретации данных, включая разработку опросников и проведение экспериментов. Статистические методы часто применяются для оценочных суждений о взаимосвязях между теми или иными событиями.

Supervised Learning. Набор основанных на технологиях машинного обучения методик, которые позволяют выявить функциональные взаимосвязи в анализируемых массивах данных.

Simulation. Моделирование поведения сложных систем часто используется для прогнозирования, предсказания и проработки различных сценариев при планировании.

Time Series Analysis. Набор заимствованных из статистики и цифровой обработки сигналов методов анализа повторяющихся с течением времени последовательностей данных. Одни из очевидных применений – отслеживание рынка ценных бумаг или заболеваемости пациентов.

Unsupervised Learning. Набор основанных на технологиях машинного обучения методик, которые позволяют выявить скрытые функциональные взаимосвязи в анализируемых массивах данных. Имеет общие черты с **Cluster Analysis**.

Visualization. Методы графического представления результатов анализа больших данных в виде диаграмм или анимированных изображений для упрощения интерпретации облегчения понимания полученных результатов.

Приложение 2

Список и описание наиболее популярных лингвистических корпусов

Название	Состав	Доступ	Разметка
Русский язык			
Национальный корпус русского языка, http://www.ruscorpora.ru	Более 500 млн слов. Кроме основного корпуса содержит газетный, параллельный, диалектный, поэтический, обучающий, устной речи, акцентологический и мультимедийный (пополняется)	Свободно доступный, оффлайновая версия недоступна, однако для свободного пользования предоставляется случайная выборка предложений из корпуса со снятой омонимией объемом 180 тыс. словоупотреблений	Морфологическая (для 6 млн слов со снятой морфологической омонимией), морфо-синтаксическая со снятой омонимией
Хельсинкский аннотированный корпус русских текстов ХАНКО, http://www.ling.helsinki.fi/projects/hanco/	Содержит тексты журнала «Итоги» (пополняется)	Свободно доступный	Морфологическая и синтаксическая
Машинный фонд русского языка, http://cfrl.ru/	Содержит тексты русской прозы, поэзии и драматургии XIX–XX вв., подкорпус текстов российских газет 90 гг. XX в., произведения русских историков XIX–XX вв., а также подкорпус по фольклору (русские народные сказки А. Н. Афанасьева)	Свободно доступный	Морфологическая (частично)
Regensburg Russian Diachronic Corpus (RRuDi), http://rhssl1.uni-regensburg.de/SlavKo/korpus/rrudi-new/	Содержит тексты на церковнославянском и древнерусском языках (пополняется)	Свободный доступ для выполнения исследовательских задач предоставляется после подписания лицензионного соглашения	Морфосинтаксическая (большинство текстов проверено вручную)

Название	Состав	Доступ	Разметка
Английский язык			
BYU-BNC: British National Corpus, созданный Марком Дэвисом, http://corpus.byu.edu/bnc/	100 млн слов британского варианта английского языка (1980–1993 гг.)	Свободный доступ для выполнения исследовательских задач после несложной процедуры регистрации на сайте	Морфологическая (можно искать конкретную словоформу, все формы одной лексемы по возможным начальным формам, словосочетания, выбранные грамматические формы лексемы)
Corpus of Contemporary American English (COCA), созданный Марком Дэвисом, http://corpus.byu.edu/coca/	Более 450 млн слов американского варианта английского языка (1990–2012 гг.). Содержит в одинаковых пропорциях тексты разговорной речи (скрипты более чем 150 ТВ- и радиопередач), художественной литературы, публицистики (популярные журналы и газеты), а также тексты академических журналов (пополняется)	Свободный доступ для выполнения исследовательских задач после несложной процедуры регистрации на сайте	Морфологическая (можно искать конкретную словоформу, все формы одной лексемы по возможным начальным формам, словосочетания, выбранные грамматические формы лексемы)
Corpus of Historical American English (COHA), созданный Марком Дэвисом, http://corpus.byu.edu/coha/	Более 400 млн слов американского варианта английского языка (1810–2009 гг.). Содержит тесты художественной литературы и публицистики	Свободный доступ для выполнения исследовательских задач после несложной процедуры регистрации на сайте	Морфологическая (можно искать конкретную словоформу, все формы одной лексемы по возможным начальным формам, словосочетания, выбранные грамматические формы лексемы)
Bank of English, http://www.collinslanguage.com/content-solutions/wordbanks	Более 553 млн слов различных вариантов английского языка, сбалансировано по разным жанрам (пополняется)	Коммерческий, пробная версия предоставляется бесплатно на один месяц после процедуры регистрации	Частеречная с элементами морфологической

Название	Состав	Доступ	Разметка
Brown Corpus, http://corpus.leeds.ac.uk/protected/	Первый представительный корпус. Состоит из 500 прозаических фрагментов в 2 000 слов, взятых из текстов, опубликованных в США в 1961 г.	Свободно доступный с сайта университета Лидс (100 примеров использования)	Морфологическая и синтаксическая
Немецкий язык			
Мангеймский корпус немецкого языка, DeReCo, http://www.ids-mannheim.de/kl/projekte/korpora/	Самый представительный корпус немецкого языка, поддерживаемый Институтом немецкого языка (Мангейм). Более 5,4 млрд слов. Содержит тексты художественной, научной и научно-популярной литературы, периодики, а также подкорпус устной речи	Свободно доступный после регистрации на сайте и подписания лицензионного соглашения. Требуется установка специальной программы – оболочки COSMAS II	Частичная морфологическая. Можно искать конкретную словоформу, все формы одной лексемы по возможным начальным формам, словосочетания
LIMAS, http://korpora.zim.uni-duisburg-essen.de/Limas/	Более 1 млн словоупотреблений. Состоит из 500 текстов 33 различных рубриках	Свободно доступный	Поиск по слову, контексту, фразе
Корпус Берлинско-Бранденбургской Академии наук DWDS, http://www.dwds.de	Около 1,8 млрд слов. Содержит тексты художественной литературы XX–XXI вв., периодики (Berliner Zeitung, Bild, Süddeutsche Zeitung, Tagesspiegel, WELT, ZEIT), устной речи и др. В разработке корпус текстов 1650–1900 гг.	Свободно доступный после регистрации на сайте	Можно искать конкретную словоформу, все формы одной лексемы по возможным начальным формам, словосочетания
Многоязычные корпуса			
TITUS, http://titus.uni-frankfurt.de/indexe.htm	Тезаурус материалов по индоевропейским языкам (древнее, среднее и для ограниченного количества языков современное состояние)	Свободно доступный. Тексты доступны для поиска, просмотра и скачивания	Возможен поиск по грамматическим формам слова
European Parliament Proceedings Parallel Corpus, http://www.statmt.org/europarl	Корпус слушаний парламента (1996–2011 гг.). Тексты на всех языках европейского парламента	Свободно доступный для скачивания	–