Д. В. Дроздова
г. Минск, Республика Беларусь

СТАТИСТИЧЕСКИЕ МЕРЫ, ИСПОЛЬЗУЕМЫЕ ПРИ ВЫЯВЛЕНИИ СИНТАГМ (на примере корпусного менеджера Sketch Engine)

Актуальность темы определяется ростом интереса лингвистов к проведению исследований на материале корпусов и перспективам применения современных статистических методов для анализа языковых явлений в новом формате. Исследование языковой репрезентации этнических меньшинств в газетном дискурсе, который материально представлен в виде корпусных данных, открывает возможность проследить социокультурные трансформации норм и ценностей.

На основе анализа сочетаний слов с точки зрения регулярности и частотности употребления в определенном типе текстов можно установить коммуникативно релевантные компоненты значения и проследить семантические трансформации, которые слова претерпевают в процессе употребления.

Под лингвистическим корпусом текстов в данной статье понимается «большой, представленный в машиночитаемом формате, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [1, с. 5]. Работа с корпусом проводится в корпусном менеджере – специализированной поисковой системе, включающей программные средства для поиска данных в корпусе, получения статистической информации и предоставления результатов пользователю в удобной форме [1, с. 50].

Внедрение в лингвистику корпусов текстов и использование корпусных менеджеров принесло новые методы изучения и освоения проблемы сочетаемости. Одним из таких инструментов являются статистические методы. Как указывает Н. Д. Андреев, статистические характеристики языковых единиц в речи оказываются весьма важным фактором при описании языкового материала [2]. Большие массивы текстов позволяют проанализировать закономерности в сочетании слов, и статистические меры становятся основным методом подобного анализа.

В данной статье мы обращаемся к статистическим мерам корпуса текстов для выявления регулярных сочетаний номинации этнической группы *black* ‘черные’. Выявление таких сочетаний позволит установить прагматический компонент значения данной номинации.

В фокусе внимания находятся сочетания, которые обладают такими свойствами, как повторяемость, специфичность значения, структурная устойчивость и экспрессивность. В концепциях Ш. Балли и В. В. Виноградова подобные сочетания определяются как синтагмы. Данный термин подчеркивает синтаксическую природу синтагмы как речевой единицы и ее системный характер (синтагматические отношения – одни из основных в лексической системе языка). По мнению А. А. Реформатского, «синтагмой является соче-

тание двух членов, связанных тем или иным отношением с неравноправной направленностью членов, где один член является определяемым, а другой – определяющим» [3, с. 137].

В рамках корпусной лингвистики, принят термин коллокация, который может быть определен как статистически устойчивое словосочетание [4, с. 138]. Исходя из определений, приведённых выше, синтагма и коллокация обладают одинаковыми свойствами: устойчивость, наличие главного и зависимого слов (ключевого слова и коллоканта), повторяемость, в связи с чем в настоящей работе данные термины являются взаимозаменяемыми.

В качестве материала исследования избран подкорпус американских новостных статей в корпусе английского языка English Web 2021 (enTenTen21). Тексты новостных статей нацелены не только на передачу информации, но и на передачу определенных мнений и оценок, и, как следствие, формирование общественного мнения. В языковом аспекте выражается мировозренческий, и на примере подобных языковых явлений можно отследить социальные тенденции, например, дискриминацию, стереотипизацию, этноцентризм, политкорректность. Инструментом анализа выступил корпус-менеджер Sketch Engine. Выбор данного инструментария обусловлен наличием функции вывода списков коллокаций по отдельным синтаксическим моделям с указанием силы связи между лексемами.

Программное обеспечение корпусного менеджера позволяет исследовать семантические свойства слов путем выявления контекста использования слова, что в дальнейшем может быть отражено в численном виде [5].

Word Sketch – это автоматически полученная из корпуса сводка грамматического и коллокационного поведения слова. Word Sketch впервые появился в 1999 году для составления словаря английского языка Macmillan для продвинутых учащихся. С тех пор они были интегрированы в корпусный инструментарий Sketch Engine, подготовлены для пятнадцати языков и широко используются для лексикографии [6].

Целью инструмента Word Sketch является полное и исчерпывающее описание грамматического и коллокационного поведения слова. Иными словами, задачей является показать все слова, которые обычно используются в сочетании с каждым ключевым словом: существительные, глаголы, прилагательные, наречия и предлоги [6].

Статистические меры, используемые в корпусном менеджере Sketch Engine для поиска коллокантов, основаны на вероятности появления двух слов вместе при анализе корпуса языка.

На разных этапах развития корпусного менеджера Sketch Engine применялись несколько статистических мер: MI-score, Association Score, Dice, logDice.

Элементы формул статистических мер:

N – размер корпуса;

$\mathbf{F_A}$ – количество случаев употребления ключевого слова во всем корпусе (размер конкорданса);

F_B – количество случаев употребления коллокации во всем корпусе;

F_{AB} – количество случаев употребления коллокации в конкордансе;

R – грамматическое отношение между словами;

||w1, R, w2|| – количество случаев употребления ключевого слова с коллокантом и их грамматическое отношение;

||w1, R, ∗|| – количество случаев употребления ключевого слова с грамматичеким отношением к любому коллоканту;

||∗, ∗, w2 || – количество случаев употребления коллоканта с любым ключевым словом в любых грамматических отношениях;

||∗, ∗, ∗|| – количество случаев употребления любого ключевого слова в любых грамматических отношениях с любым коллокантом.

Мера MI-score – коэффициент взаимной информации, направленный на выявление в корпусе редкие коллокации. Следовательно, чем реже встречается отдельная коллокация в корпусе, тем больше вес ее. Но следует отметить, если частота сочетания мала, использование данной формулы может привести к некорректным результатам [7, с. 349].

$$\log_2 \frac{f_{AB}N}{f_A f_B}$$

После, статистическая мера MI-score сменилась методом Association Score (AScore). По сравнению с предыдущей метрикой, AScore принимает во внимание грамматические отношения между ключевым словом и коллокантом.

$$AScore(w_1, R, w_2) = \log \frac{||w_1, R, w_2|| \cdot ||*, *, *||}{||w_1, R, *|| \cdot ||*, *, w_2||} \cdot \log(||w_1, R, w_2|| + 1)$$

С 2006 года, отмечая зависимость AScore от размера корпуса и новые исследования, статиститическая мера для создания скетчей была изменена на вариацию коэффициента Дайса – logDice [8].

LogDice – статистическая мера для выявления совместного появления в коллокации. Данная метрика основана на частотности ключевого слова и его коллоката, в отличии от предыдущих метрик, logDice не зависит от размера корпуса, следовательно можно использовать для сравнения результатов с разных корпусов. При переводе иноязычной терминологии используется данная метрика для подбора наиболее оптимального перевода.

$$14 + \log_2 \text{Dice}\left(\frac{||w_1, R, w_2||}{||w_1, R, *||}, \frac{||w_1, R, w_2||}{||*, *, w_2||}\right) = 14 + \log_2 \frac{2 \cdot ||w_1, R, w_2||}{||w_1, R, *|| + ||*, *, w_2||}$$

По умолчанию Word Sketch в корпусном менеджере Sketch Engine сортируется так, чтобы наиболее типичные словосочетания находились вверху. Оценка logDice используется для определения того, насколько типичным (или сильным) является коллокация [9].

Значения статистической меры logDice имеют следующие особенности:

1) Теоретический максимум равен 14 в случае, когда все появления ключевого слова совпадают с коллокантом, а все появления коллоканта совпадают с ключевым словом;

2) Значение 0 означает, что на 16 000 ключевых слов или 16 000 коллокантов приходится менее 1 совпадения коллокации;

3) Оценка не зависит от общего размера корпуса. Оценка объединяет относительные частоты коллокации в целом по отношению к ключевому слову и коллоканту [10].

Невозможно установить универсальный порог между слабыми и сильными коллокациями, поскольку каждое слово ведет себя по-разному. Основная цель статистической меры – отсортировать коллокации по их типичности или силе, а не решить, является ли словосочетание слабым или сильным [11].

Наш анализ синтагм, включающих в себя наименования представителей афроамериканской этно-расовой группы, был проведен в корпусном менеджере Sketch Engine. При выявлении коллокантов в инструменте Word Sketch в корпусном менеджере Sketch Engine указывается не только количество случаев появления коллокации, но и результат статистической меры logDice.

В настоящем исследовании было выбрано наименование представителей афроамериканской этно-расовой группы – *Black* ‘черные’. Выбор наименований обусловлен анализом новостных газет *The Boston Globe, Star Tribune, USA Today, The New York Times* в период с 2020 по 2022 гг. [12].

В сравнительном анализе было выявлено численное превалирование случаев употребления представителей афроамериканской этно-расовой группы в позиции объекта, в сравнении с предложениями, в которых они представлены в качестве субъекта действия. Глаголы *to lynch* ‘расправляться самосудом’ (15 случаев коллокации), *to enslave* ‘поработить’ (53 случая), *to harass* ‘преследовать’ (13 случаев) с определением *black* ‘черные’ встретились в подкорпусе американских новостных статей, что часто апеллируют к сложившимся стереотипам, слабого, подверженного опасности.

В контекстах, в которых наименование *black* ‘черные’ упоминалось в качестве субъекта действия, глагол *to migrate* ‘мигрировать’, который так же указывает на совершение действия под влиянием кого-либо или каких-либо обстоятельств (см. рис. 1).

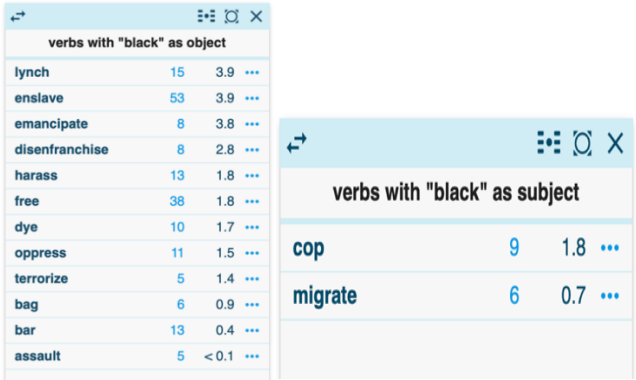

Рис 1.

Следует отметить наличие погрешности в статистическом инструментарии, так как в коллокации *black* (в качестве подлежащего) + глагол слово *cop* было распознано в качестве глагола ‘схватить’, но при подробном анализе контекстов, была выявлена погрешность в разметке, в которой *cop* должно быть размечено как существительное ‘полицейский’ в структуре прилагательное + существительное (см. рис. 2).

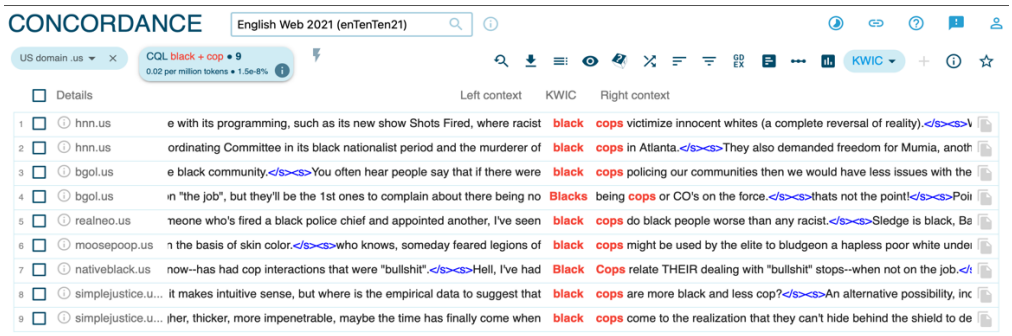

Рис 2.

Прилагательные, используемые с *black* ‘черные’ также подчеркивают социальное положение представителей афроамериканской этно-расовой группы, *inferior* ‘низший’ (5 случаев употребления), *incapable* ‘неспособный’ (5 случаев употребления) указывают на неравенство в обществе, неспособность к совершению действия, с низким уровнем интеллекта (см. рис. 3).

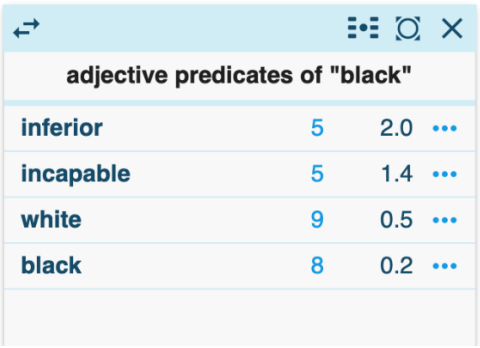

Рис 3.

Таким образом, существует множество разнообразных мер выявления коллокаций, каждая из которых направлена на решение определенных практических задач. Процесс выявление коллокантов в корпусном менеджере Sketch Engine претерпевал изменения. На данном этапе развития используется статистическая мера logDice. К сожалению, на данном этапе развития статистических метрик, существуют ряд проблем, а именно выявляются случаи объединения в коллокацию слов, разделенных запятой, что создает погрешность в результатах анализа. При анализе синтагм с ключевым словом *black* ‘черные’, в качестве наименования представителя афроамериканской этно-расовой группы, были выявлены прагматические значения слабого, подверженного опасности, неспособного к чему-то.

ЛИТЕРАТУРА

1. *Захаров, В. П.* Корпусная лингвистика : учебник для студентов, обучающихся по направлению подготовки бакалавров и магистров / В. П. Захаров, С. Ю. Богданова. –2-е изд. – СПб : Изд-во СПбГУ, 2013. – 144 с.
2. *Андреев, Н. Д.* Статистико-комбинаторные методы в теоретическом и прикладном языковедении / Н. Д. Андреев. – Л. : Наука, 1967. – 403 с.
3. *Реформатский, А. А.* Введение в языковедение / А. А. Реформатский / под ред. В. А. Виноградова. – М. : Аспект Пресс, 1996. – 536 с.
4. *Захаров, В. П.* Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке / В. П. Захаров, М. В. Хохлова // Компьютерная лингвистика и интеллектуальные технологии. – М. : РГГУ, 2010. – Вып. 9 (16). – С. 137–143.
5. *Kilgarriff, A.* The Sketch Engine: Ten Years On, Lexicography [Electronic resource] / A. Kilgarriff. – Mode of access: https://www.sketchengine.eu/wpcontent/uploads/The_Sketch_Engine_2014.pdf. – Date of access: 18.02.2024.
6. *Kilgarriff, A.* A quantitative evaluation of Word Sketches [Electronic resource] / A. Kilgarriff. – Mode of access: https://www.euralex.org/elx_proceedings/Euralex2010/019_Euralex_2010_1_KILGARRIFF%20KOVAR%20KREK%20SRDANOVIC%20TIBERIUS_A%20Quantitative%20Evaluation%20of%20Word%20Sketches.pdf. – Date of access: 18.02.2024.
7. *Хохлова, М. В.* Особенности статистических мер при выделении биграмм / М. В. Хохлова // Корпусная лингвистика – 2017 : тр. Международной конференции. – СПб. : Изд-во СПбГУ, 2017. – С. 349–354.
8. *Curran, J. R.* From Distributional to Semantic Similarity [Electronic resource] / J. R. Curran. – Mode of access: https://era.ed.ac.uk/bitstream/handle/1842/563/IP030023.pdf;jsessionid=93DA7C858C403C9CF846B55AC8F47923?sequence=2. – Date of access 20.02.2024.
9. Log-likelihood and effect size calculator [Electronic resource]. – Mode of access: https://ucrel.lancs.ac.uk/llwizard.html. – Date of access: 18.02.2024.
10. Sketch Engine [Electronic resource]. – Mode of access: https://www.sketchengine.eu/. – Date of access: 18.02.2024.
11. *Rychlý, P.* A Lexicographer-Friendly Association Score / P. Rychlý // Recent advances in Slavonic natural language processing / ed. P. Sojka, A. Horák. – Brno : Masaryk University, 2008. – P. 6–9.
12. *Дроздова, Д. В.* Лексическое выражение расовой принадлежности в американском газетном дискурсе / Д. В. Дроздова // Материалы ежегодной научной конференции студентов и магистрантов университета, 15–16 апреля 2021 г. : в 4 ч. / редкол.: Н. Е. Лаптева (отв. ред.) [и др.]. – Минск : МГЛУ, 2021. – Ч. 3. – С. 176–177.