ГОСУДАРСТВЕННЫЙ
ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А.С. ПУШКИНА

СЛАВЯНСКАЯ КУЛЬТУРА: ИСТОКИ, ТРАДИЦИИ, ВЗАИМОДЕЙСТВИЕ

XXIV Кирилло-Мефодиевские чтения
(24 мая 2023 г., Москва)

Материалы
Международной
научно-практической конференции

Москва
2023

ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА
им. А. С. ПУШКИНА

ФИЛОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

III Костомаровский форум
(24–25 мая 2023 г., Москва)

СЛАВЯНСКАЯ КУЛЬТУРА: ИСТОКИ, ТРАДИЦИИ, ВЗАИМОДЕЙСТВИЕ

Материалы Международной
научно-практической конференции
«XXIV Кирилло-Мефодиевские чтения»
(24 мая 2023 г., Москва)

Сборник статей

Москва
2023

УДК 8.80
ББК 80.4
С 47

Рекомендовано к изданию Редакционно-издательским советом Государственного института русского языка им. А. С. Пушкина. Протокол № 8 от 3 марта 2023 г.

Рецензенты:
О.В. Шаталова, доктор филологических наук, профессор, зав. кафедрой славянских языков ГОУ ВО МО «Московский государственный областной университет»;
Т.В. Кудоярова, кандидат педагогических наук, доцент кафедры русской словесности и межкультурной коммуникации ФГБОУ ВО "Гос. ИРЯ им. А.С. Пушкина".

С 47 **Славянская культура: истоки, традиции, взаимодействие.** Материалы Международной научно-практической конференции «XXIV Кирилло-Мефодиевские чтения» (24 мая 2023 г.): сборник статей / гл. ред. В.И. Карасик. – Москва : Гос. ИРЯ им. А.С. Пушкина, 2023. – 734 с.

ISBN 978-5-98269-313-6

В сборнике представлены статьи, посвященные актуальным вопросам языкознания, литературоведения, отечественной и мировой истории и культуры. Обсуждаются вопросы методики преподавания филологических дисциплин и русского языка как иностранного, функционирования русского языка в современных медиа. Затрагиваются проблемы лингвокультурологии и межкультурной коммуникации, истории и поэтики русской и зарубежной литературы. Обсуждается современный литературный процесс и современное состояние славянских языков и культур.

УДК 8.80
ББК 80.4

ISBN 978-5-98269-313-6

Ольховская Александра Игоревна
Государственный институт русского языка им. А. С. Пушкина
aleksandra_olhovskaya@mail.ru

Зельникова Анна Артемовна
Государственный институт русского языка им. А. С. Пушкина
anna.zelnikova@mail.ru

Математические инструменты измерения коллокаций в аспекте изучения фразеологии

В статье описаны основные математические инструменты, позволяющие измерить степень близости компонентов коллокации – MI, MI3, t-score, logDice. Авторы определяют параметры, потенциально влияющие на показатели мер, формируют группы фразеологизмов на их основе и осуществляют подсчеты мер для разных групп идиом с использованием данных корпуса RuTenTen. Полученные результаты позволили сформулировать особенности применения разных мер при изучении фразеологизмов.

Ключевые слова: фразеологизм, коллокация, мера устойчивости, MI, MI3, T-score, logDice.

С появлением корпусов лингвистика вышла на новый этап своего развития, важной чертой которого является активное применение квантитативных методов. Автоматически извлекаемая из корпуса статистика касается не только отдельных языковых единиц, но и их сочетаний. В связи с этим важным направлением корпусных исследований стало изучение комбинаторного потенциала слов и в частности коллокаций.

Термин *коллокация* трактуется в современной лингвистике неоднозначно [4, 6]. В широком смысле слова коллокация понимается как «комбинация двух или более слов, имеющих тенденцию к совместной встречаемости» [6: 343]. М.В. Копотев считает, что коллокация – это принятое в корпусной лингвистике наименование устойчивых словосочетаний, при этом подчеркивает, что понятие «коллокация» существенно шире, чем понятие «фразеология». К примеру, предложно-падежную форму в Москву можно считать коллокацией, но никак не фразеологизмом [3: 99].

Несмотря на несоизмеримость понятий «коллокация» и «фразеологизм», потенциально именно коллокации – естественный источник пополнения фразеологического состава языка. Соответственно, корпусная методика изучения коллокаций может частично применяться при изучении фразеологии.

Данная работа преследует две взаимосвязанных задачи: 1) описание существующих в корпусной лингвистике математических методов измерения силы коллокаций и 2) выборочную проверку данных методов на русских фразеологизмах с целью определения их эффективности и границ применения.

Для выявления коллокаций в корпусной лингвистике используются специальные математические инструменты, получившие название *меры устойчивости*. Принцип их работы строится на предположении о том, что реальная встречаемость двух слов, устойчиво связанных в речи, должна быть выше математически ожидаемой. При этом математически ожидаемая встречаемость вычисляется перемножением частотностей двух заданных слов.

В настоящий момент в корпусной лингвистике существует несколько мер устойчивости. Наиболее простой считается MI (англ. Mutual information 'взаимная информация'), которая показывает вероятность, с которой два объекта (два слова) окажутся рядом в некотором объеме данных (корпусе текстов).

Формула для вычисления этой меры выглядит следующим образом: $MI = log_2 \frac{f(n,c) \times N}{f(n) \times f(c)}$, где MI – объем информации; n – коллокат 1 (ключевое слово); c – коллокат 2; f(n,c) – частота встречаемости коллоката 1 в паре с коллокатом 2; f(n), f(c) – абсолютные частоты коллокатов в корпусе; N – общее число словоформ в корпусе [3, 6].

Таким образом, в теории можно предположить следующую зависимость: чем чаще два слова оказываются рядом друг с другом, тем выше их взаимная устойчивость и тем частотней употребляется выражение, которое с высокой долей вероятности окажется фразеологическим.

Особенность MI, по словам М.В. Хохловой, состоит в том, что она показывает завышенные результаты для редких сочетаний слов [5: 167]. Следовательно, для фразеологизмов, которые употребляются не очень часто или появились в языке недавно, мера будет работать не вполне корректно.

Чтобы преодолеть этот недостаток, формула MI была усовершенствована эмпирическим путем в работах M. Oakes за счет возведения в куб частоты употребления коллокации:

$MI^3 = log_2 \frac{f^3(n,c) \times N}{f(n) \times f(c)}$ (цит. по [там же]).

Еще один инструмент измерения силы коллокаций – t-score. Он вычисляется по формуле: $t - score = \frac{f(n,c) - \frac{f(n) \times f(c)}{N}}{\sqrt{f(n,c)}}$ [3, 6].

К особенностям этой меры можно отнести то, что с опорой на нее выделяются в первую очередь коллокации с наиболее частотными единицами, к которым в большинстве случаев принадлежат служебные слова. Поэтому при анализе биграмм по t-score «необходимо задавать список стоп-слов, чтобы «отбросить" самые частотные слова, сочетания с которыми неизменно окажутся в самом верху таблицы: предлоги, местоимения или союзы» [2: 142].

Наконец, последняя мера, которую мы осветим в данной статье, – logDice. Она представляет собой модификацию меры Дайса – одного из наиболее известных алгоритмов для поиска n-словных сущностей в тексте и, в отличие от него, не завышает значений для редко встречающихся сочетаний. Вот как выглядит формула меры logDice: $logDice = 14 + log_2 \frac{2f(n,c)}{f(n)+f(c)}$ [5: 167].

Следует отметить, что все перечисленные меры устойчивости обладают рядом существенных недостатков. Во-первых, они плохо приспособлены к измерению силы словосочетаний, включающих в свой состав более двух компонентов. Во-вторых, при их использовании не учитывается омонимия и многозначность, что создает определенные трудности при анализе фразеологизмов, имеющих аналоги с прямым значением, ср. *кот в мешке* – ‘что-либо неизвестное’ и ‘животное в мешке’. В-третьих, все они являются не абсолютными, а относительными величинами – значение, полученное в ходе вычислений, приобретает смысл только на фоне мер других словосочетаний с этим же словом. Четкого порога, который бы отделял статистически значимые словосочетания от статистически незначимых, не существует. Исследование В.П. Захарова и С.Ю. Богдановой показали, что «в диапазоне значений меры MI от 0 до 1 не были найдены словосочетания, которые можно было бы причислить к устойчивым» [1: 106], однако на практике статистически незначимыми могут оказаться словосочетания с гораздо бóльшей мерой устойчивости, ср.: MI *зеленое море* = 3,5; *фиолетовое море* = 3,5; *розовое море* = 1,6.

Перейдем непосредственно к практической части, заключающейся в сопоставлении показателей всех рассмотренных мер применительно к фразеологизмам разных типов. Материалом исследования послужили устойчивые обороты из «Фразеологического словаря русского литературного языка» А.И. Федорова, «Словаря русских фразеологических неологизмов» В.М. Мокиенко, Е.В. Генераловой (рукопись), а также из газетного корпуса НКРЯ, вычлененные авторами посредством корпусного поиска по маркерам новизны. Чтобы определить специфику каждого статистического инструмента, нами были отобраны критерии, которые гипотетически могут оказывать влияние на корпусную статистику.

1. Стиль речи (в укрупненном масштабе): 1.1) книжный, 1.2) разговорный

2. Наличие в составе фразеологизма лексики разной употребительности: 2.1) хронологически ограниченной лексики: 2.1.1) неологизмов, 2.1.2) архаизмов и историзмов, 2.2) общеупотребительной лексики

3. Функция фразеологизма: 3.1) номинативная, 3.2) экспрессивная.

В соответствии с обозначенными параметрами нами были сформированы 7 групп фразеологизмов для их оценки с помощью четырех мер устойчивости. Данные о частоте ключевого слова и его коллокации были взяты из наиболее

объемного на данный момент корпуса русского языка – RuTenTen. Значения мер рассчитывались вручную. В приведенных ниже таблицах вы можете увидеть среднее арифметическое мер для фразеологизмов из каждой категории.

Таблица. *Средние значения MI, MI3, t-score, logDice для фразеологизмов разных групп (на основании данных корпуса RuTenTen)*

Параметры	Наполнение группы	MI	MI3	T-score	logDice
1. Стиль речи					
1.1. Книжный стиль	достигнуть апогея, социальные лифты, терновый венец, баловень фортуны, агенты кремля	12,9	31	46,2	6,4
1.2. Разговорный стиль	гнать пургу, развести руками, пинать балду, белая зависть, собаку съесть	9,4	31,2	61	6,1
2. Наличие в составе фразеологизма лексики разной употребительности					
2.1.1. Неологизмы	драйверы роста, постсоветское пространство, запастись попкорном, разрулить ситуацию, кошмарить бизнес	10	30,3	40,6	5
2.1.2. Архаизмы и историзмы	мелкая сошка, ахиллесова пята, жрецы Фемиды, осваивать азы, попасть впросак	10,4	27	44	5
2.2. Общеупотребительная лексика	гроза морей, эффективный менеджер, акула пера, новый русский, женская логика	5,7	30,2	80	5,5
3. Функция фразеологизма					
3.1. Номинативная	анютины глазки, Большая медведица, группа риска, информационное поле, ледовая арена	11,25	39,1	136,5	9,1
3.2. Экспрессивная	закидать тапками, житья нет, выбивать деньги, сорвать башню, обломать рога	8,2	24,3	18,5	3,4

Наиболее рельефные данные были получены в отношении параметра «Функция фразеологизмов». Самые высокие показатели по всем мерам устойчивости, кроме MI, выявлены у фразеологизмов, обладающих номинативной функцией (*анютины глазки, Большая медведица, группа риска* и др.). Вероятно, это объясняется тем, что по своей устойчивости эти выражения сопоставимы со словами. В то же время фразеологизмы с экспрессивной функцией, напротив, характеризуются самыми низкими показателями мер t-score и logDice.

Данные по параметру «Наличие в составе фразеологизма лексики разной степени употребительности» в общем и целом подтвердили специфику мер MI, MI3 и t-score, тем не менее наши наблюдения позволили несколько уточнить эти особенности. Так, если ученые указывали на завышение значений MI для редких словосочетаний, то наши расчеты показали, что завышение происходит также в тех случаях, когда редким является только один из компонентов словосочетания, а не все словосочетание в целом. В качестве редких единиц в сформированной выборке выступали хронологически ограниченные слова – неологизмы, архаизмы и историзмы. Показатели MI фразеологизмов с такими единицами оказались ощутимо выше, чем у выражений с общеупотребительной лексикой: 10 – неологизмы; 10,4 – архаизмы/историзмы; 5,7 – общеупотребительная лексика.

Мера MI3, призванная исправить недостаток своего предшественника, вполне успешно справляется со своей задачей. Доказательством этому служит то, что разница в показателях различных по своему составу выражений не является существенной, ср.: 30,3 – неологизмы; 27 – архаизмы/историзмы; 30,2 – общеупотребительная лексика. Как и в предыдущем случае, сглаживающее действие MI3 распространяется не только на редкие словосочетания, но и на словосочетания с редкими компонентами.

Мера t-score показывает более высокие значения для фразеологизмов, содержащих общеупотребительную лексику: 40,6 – неологизмы; 44 – архаизмы / историзмы; 80 – общеупотребительная лексика. Это согласуется с наблюдениями ученых о том, что данная мера ориентирована прежде всего на коллокации с высокочастотными словами.

Разница между показателями мер по параметру «Стиль» оказалась несущественной. Возможно, это обусловлено тем, что RuTenTen, будучи интернет-корпусом, более или менее сбалансированно представляет книжную и разговорную речь. Между тем есть вероятность того, что преимущественно «книжные» по содержанию корпуса, в частности, НКРЯ, будут демонстрировать бóльшие показатели у фразеологизмов книжного стиля.

Следует отметить, что разные меры ведут себя неодинаково в отношении стилевой маркированности идиом. Показатели меры MI несколько выше у книжных выражений (12,9 – книжные; 9,4 – разговорные), в то время как показатели t-score – напротив, выше у разговорных оборотов (46,2 – книжные; 61 – разговорные), что также может свидетельствовать о влиянии состава фразеологизма на высоту мер.

Средние величины, полученные для разных мер, позволяют обозначить приблизительный порог, отделяющий фразеологические единицы от коллокаций. Можно предположить, что для причисления устойчивого оборота к фразеологизму его MI должна быть выше 5, MI3 – выше 25, t-score – выше 40, LogDice – выше 3. Низкие значения t-score в отношении экспрессивных фразеологизмов нуждаются в отдельном осмыслении.

Таким образом, статистические меры исследования коллокаций могут применяться при изучении фразеологии. Полученные результаты будут максимально эффективны в лексикографической практике при корпусном отборе фразеологизмов и экспериментальной проверке степени их устойчивости. Особенно востребованными данные методы могут стать при работе с новой фразеологией, не зафиксированной в лингвистических источниках и нуждающейся в скорейшем описании.

Литература

1. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». – СПб.: СпбГУ. РИО. Филологический факультет, 2013. – 148 с.

2. Захаров В.П., Хохлова М.В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии. – 2010. – № 9 (16). – С. 137−143.

3. Копотев М.В. Введение в корпусную лингвистику: Учебное пособие для студентов филологических и лингвистических специальностей университетов. – Прага: Animedia, 2014. – 128 с.

4. Палийчук Д.А. Проблема определения понятия «коллокация» в современной лингвистике // Евразийский гуманитарный журнал. – 2022. – № 1. – С. 20–25.

5. Хохлова М.В. Сопоставительный анализ статистических мер на примере частеречных преференций сочетаемости существительных // Компьютерная лингвистика и вычислительные онтологии. – 2017. – Вып. 1. – С. 166–171.

6. Хохлова М.В. Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia. Серия «Инструментарий русистики». Т. 34. – Хельсинки: Yliopistopaino, 2008. – С. 343–357.

Olkhovskaya A.I.
Pushkin State Russian Language Institute

Zelnikova A.A.
Pushkin State Russian Language Institute

Mathematical tools for measuring collocations in the aspect of phraseology research

The article describes the main mathematical tools allowing measuring the proximity of collocation components – MI, MI3, t-score, LogDice. The authors determine the parameters affected potentially the indicators of measures, form groups of idioms based on them, and calculate measures for different groups using the data of the RuTenTen corpus. The results made it possible to formulate the features of the different measures application in the research of phraseology.

Keywords: idiom, collocation, lexical association measures, MI, MI3, t-score, LogDice.