раздел **ФИЛОЛОГИЯ и ИСКУССТВОВЕДЕНИЕ**

УДК 81’233

# СТАТИСТИЧЕСКИЕ МЕТОДЫ В ИССЛЕДОВАНИИ ТЕКСТОВ

**© В. А. Гречачин**

*Башкирский государственный университет*
*Россия, Республика Башкортостан, 450076 г. Уфа, ул. Заки Валиди, 32.*

*Тел.: +7 (937) 789 28 97.*
*Email: Vitaley.grechachin@gmail.com*

*Целью данной статьи является рассмотрение статистических методов в контексте исследования текстов и определение возможностей их применения.*

*Основное внимание уделено квантитативным характеристикам лингвистических единиц. Рассмотрены основные разновидности частоты употребления лингвистических единиц. Выявлены способы подсчета частоты употребления лингвистических единиц; определены возможности приложения рассмотренных статистических методов в сопоставительных исследованиях текстов; рассмотрены возможности использования описательной статистики для проведения лингвистических исследований. На материале небольшого корпуса текстов проведен анализ особенностей употребления тех или иных частей речи в произведениях нескольких авторов, что может быть использовано для исследований авторского стиля и художественной картины мира произведений. Кроме того, рассмотрен исследовательский потенциал различных визуализаций статистических данных, полученных благодаря частотному анализу единиц в текстах.*

*В рамках данной работы получены данные, которые могут быть использованы для дальнейшего исследования потенциала статистических методов в лингвистических исследованиях.*

***Ключевые слова**: статистика, квантитативная лингвистика, корпусная лингвистика, частотность лингвистических единиц.*

Количественные методы в науке возникли в начале XX в. Тогда к ним прибегали прежде всего в естественно-научных и социологических исследованиях. Позже, в середине XX в., статистика стала преобладать над другими методами количественного подхода к исследовательским данным. С течением времени квантитативные методы смогли занять в инструментарии исследователей свое место и обозначились принципиальные отличия между ними и качественными методами [5].

Главной особенностью любых количественных исследований является тот факт, что сбор данных представляет собой отдельную задачу, стоящую перед исследователем. Прежде всего, тот должен выделить параметры, или переменные, которые подлежат измерению. Но он не может предсказать степень важности той или иной выбранной переменной. Определение важности переменных относится к задаче отбора признаков (feature selection) в статистике и машинном обучении. Кроме того, важность переменных можно определить при помощи агрегирования данных. Количественные характеристики переменных, полученные в ходе сбора данных, оформляются в виде матрицы, которая, в свою очередь, является предметом исследования статистическими методами.

Следующий этап заключается в интерпретации полученных результатов.

Положение о том, что качественные характеристики исследуемого объекта имеют цифровое выражение как в единичном ряду, так и во взаимосвязи, находит свое отражение в работе Дернеи [10], где автор называет квантитативные методы «meaning in numbers».

Особую роль квантитативные методы играют в современном мире. Благодаря сети Интернет количество данных, подлежащих подсчету, растет невероятными темпами. Размер Интернета можно выразить количеством веб-страниц (например, англоязычный сегмент оценивается в 4.58 млрд отдельных веб-страниц). Основная информация представлена в текстовом виде. Поэтому анализ текстовой информации – одно из важнейших направлений в современной науке.

Если в распоряжении исследователя имеется определенный набор текстов, которые он хочет проанализировать при помощи квантитативных методов, то ему нужно решить, что извлечь из текста в качестве переменных. Кроме того, стоит позаботиться о репрезентативности получаемых данных о количественных характеристиках переменных. Например, делать выводы о количественном распределении переменных в поэтических текстах на основании анализа выборки, состоящей из 5 стихотворений, невозможно. Конечно, такую попытку предпринять можно, но результаты анализа нерепрезентативной выборки не будут соответствовать характеристикам генеральной совокупности, которой являются все поэтические тексты.

В статистике выделяют несколько типов переменных. Прежде всего, это количественные переменные. В текстах ими могут стать различные лингвистические единицы, которые можно посчитать: фонемы, морфемы, слова, словосочетания. Посчи-

тав количество представителей этих языковых единиц в текстах, мы получим для них частотную характеристику, затем сможем описать структуру текстовых данных, имеющихся в нашем распоряжении.

В лингвистическом контексте в статистике применимо также понятие «категориальные переменные» Категориальная переменная – это переменная, принимающая одно из заданных значений. Например, категориальной переменной может выступать часть речи, так как мы заранее задаем значение для всех частей речи.

Основное допущение, на котором основывается наше исследование, заключается в следующем: «Текст есть последовательность независимых реализаций случайной величины *X*» [1]. Это предположение относится к любым лингвистическим единицам, но для определенности положим, что значениями случайной величины *X* являются слова (словоформы) *x1, x2, x2, …,* $x_n$. Слова, в свою очередь, составляют словарь – множество *V*. Чтобы говорить о значимых различиях в частоте употребления той или иной переменной в различных выборках, приведем еще одно предположение, на котором основывается наше исследование: «Каждой языковой единице *x* может быть сопоставлена вероятность *p* ее употребления в корпусе текстов *X*» [1]. Таким образом, *p1, p2, p3, …,* $p_n$ – это вероятности, с которыми *x* принимает значения *x1, x2, x3, …,* $x_n$. Попробуем интуитивно разобраться в описанном выше. Важной для нашего исследования величиной, на основании которой мы можем сделать вывод о существенности или несущественности различия в частоте употреблении того или иного слова в нескольких корпусах текстов, является вероятность *p*, которая может быть вычислена для любого слова *x*, входящего в корпус *X* и которая отражает вероятность употребления этого слова *x* в корпусе *X*.

Таким образом, для каждого *x* в корпусе *X* мы можем получить различные характеристики, основываясь на описательной статистике. Абсолютная частота $n$ словоформы $x_n$ в корпусе *X* – это целочисленное значение, отражающее общее количество употреблений $x_n$. Приведем таблицу абсолютных частот наиболее употребительных слов в русском и башкирском корпусах поэтических текстов XX в.

Средняя частота характеризует обобщенное значение переменной [Васнев, 2001] и прежде всего используется для сравнения нескольких совокупностей признаков. Пусть корпус *X* состоит из *N* текстов *t1, t2, t3, …, tn*. Тогда средняя арифметическая частота словоформы *xn* рассчитывается по формуле:

$$\bar{x}_w = \frac{\sum x_n}{N}$$

где $\Sigma x_n$ – сумма частот словоформы *w* в *n* текстах, *N* – количество текстов, где наблюдается употребление словоформы.

Кроме средней арифметической величины, которая используется чаще всего, существуют также средняя квадратическая, средняя гармоническая, средняя геометрическая, средняя кубическая [11].

Важной величиной для статистики лингвистических единиц является и относительная частота. В статистике относительные показатели используют для проведения сравнительного анализа, а также обобщения и синтеза. Относительная частота показывает отношение количества словоупотреблений определенной словоформы в наблюдаемом тексте или наблюдаемых текстах к общему количеству словоформ в этом тексте или этих текстах. Относительная частота рассчитывается по формуле:

$$\omega_w = \frac{n_w}{N}$$

где $n_w$ – это количество словоупотреблений *w*, а *N* – общее количество слов в тексте или текстах.

Сравним график абсолютных частот частей речи в текстах нескольких русских поэтов XX в. (*рис.* 1) и график относительных частот частей речи в тестах этих же поэтов (рис. 2). Корпус был предобработан для проведения статистических исследований в соответствии с основными методами компьютерной лингвистики [3].


*Рис. 1.* Абсолютная частота употребления частей речи некоторыми поэтами.

Данные, обследуемые нами, состоят из произведений Б. Л. Пастернака, А. А. Ахматовой, О. Э. Мандельштама, М. И. Цветаевой, количество которых в нашем наборе равно 531, 939, 679, 1468 соответственно.

Для начала опишем фигуры этих графиков. Перед нами 4 гистограммы, показывающие количественное соотношение частей речи в текстах поэтов. Цвет прямоугольников соотносится с частями речи, а их высота показывает количество словоупотреблений определенной части речи в текстах (*рис.* 1) и величину относительной частоты частей речи (*рис.* 2).

Данный тип визуализации (*рис.* 1) позволяет нам сделать выводы о продуктивности того или иного поэта. Таким образом, можно сказать, что М. Цветаева написала больше слов по всем частям речи, нежели Б. Пастернак, А. Ахматова и О. Мандельштам. Сделаем вывод о том, что абсолютная частота лингвистических единиц, в данном случае частота слов по частям речи, показывает продуктивность автора. Также мы можем подсчитать абсолютную частоту употребления лингвистических единиц не только у отдельно взятых авторов, но и, например, в определенные временные промежутки.

Теперь обратимся к следующему графику (*рис.* 2) и попытаемся эксплицировать относительный показатель в контексте сопоставительного анализа различных текстовых данных. Во-первых, фигуры графиков значительно отличаются между собой. График на *рис.* 1 не отражает продуктивность какого-либо автора. Как можно видеть, значения усреднились. Размах значений на первом графике равен ~ 57 000, в то время как размах на втором графике значительно меньше ~ 0.35. Вторая визуализация отражает соотношение слов по частям речи у каждого автора. На основании результатов подсчета относительной величины во втором графике мы не можем сказать о том, что в стихах М. Цветаевой больше существительных, чем у остальных.

Максимальное значение для существительных (NOUN) теперь принадлежит Л. Пастернаку. Но и это не говорит нам о том, что у него существительных больше, чем у других. Значение NOUN у этого поэта показывает одновременно, что он употребил больше существительных относительно других частей речи в своих текстах и что он по сравнению с М. Цветаевой, А. Ахматовой и О. Мандельштамом чаще использовал существительные, чем другие части речи. Это может говорить о стилистических особенностях текстов Б. Пастернака. Например, мы можем сделать вывод о большей предметности языковой картины мира в его произведениях, в то время как в текстах А. Ахматовой наблюдается высокий показатель после NOUN относительной величины для полных прилагательных (ADJF) [Напреенко, 2014]. Распределение относительных частот переменной ADJF по данным авторам показывает, что значение ADJF в текстах А. Ахматовой выше, чем у других писателей. Но при этом значение NOUN существенно ниже. На основании этих данных можно сделать вывод об образности авторского языка А. Ахматовой, большей эмоциональной окрашенности лексики, но при этом меньшей предметности относительно некоторых ее современников.

Таким образом, мы можем сделать вывод о том, что относительные показатели играют важную роль в сопоставительном изучении письменного дискурса. В рамках нашего исследования будем исходить из того, что применение статистического подсчета относительных показателей различных лингвистических единиц является отправной точкой в создании естественно-научного фундамента в сопоставительных исследованиях поэтического текстового пространства XX в. в русском и башкирском языках.


Рис. 2. Относительная частота употребления частей речи некоторыми поэтами.

Следует отметить, что при анализе относительных показателей нужно остерегаться различного рода обобщений. Относительные величины не могут описать объект исследования, они лишь являются инструментом сравнительного анализа нескольких наборов данных.

В рамках данного исследования мы рассмотрели основные статистические методы, которые могут быть приложены в сопоставительных исследованиях текстов. Мы определили перспективы подобных методов, некоторые из них уже нашли применение в работах по стилометрии [6–7]. На примере данного исследования мы продемонстрировали, как инструменты статистики помогают в анализе целых массивов текстов.

*Работа выполнена в рамках поддержанного РФФИ проекта №17-04-00193 «Исторический корпус башкирского языка».*

## ЛИТЕРАТУРА

1. Арапов М. В. Квантитативная лингвистика. М.: Наука, 1988. 184 с.
2. Баранов А. В. Введение в прикладную лингвистику. М.: Эдиториал УРСС, 2001. [Электронная книга]. Вайсгербер Й. Л. Родной язык и формирование духа. М., 2004. 232 с.
3. Гречачин, В. А. К вопросу о токенизации текста // Международный научно-исслед. журнал. 2016. №6(48). Ч. 4. С. 25–27. doi: 10.18454/IRJ.2016.48.070.
4. Гржибовский А. М. Корреляционный анализ // Экология человека. 2008. №9 URL: https://cyberleninka.ru/article/n/korrelyatsionnyy-analiz
5. Кащеева А. В. Квантитативные и качественные методы исследования в прикладной лингвистике // Социально-экономические явления и процессы. 2013. №3(049). URL: https://cyberleninka.ru/article/n/kvantitativnye-i-kachestvennye-metody-issledovaniya-v-prikladnoy-lingvistike (дата обращения: 19.10.2018).
6. Кочеткова Н. А. Статистические языковые методы. Коллокации и коллигации // Новые информационные технологии в автоматизированных системах. 2013. №16. URL: https://cyberleninka.ru/article/n/statisticheskie-yazykovye-metody-kollokatsii-i-kolligatsii.
7. Напреенко Г. В. Идентификация текста по его авторской принадлежности на лексическом уровне (формально-колич. модель) // Вестн. Том. гос. ун-та. 2014. №379. URL: https://cyberleninka.ru/article/n/identifikatsiya-teksta-po-ego-avtorskoy-prinadlezhnosti-na-leksicheskom-urovne-formalno-kolichestvennaya-model (дата обращения: 19.10.2018).
8. Ибрагимова В. Л., Фаткуллина Ф. Г. Основные принципы исследования словарного состава современного русского языка // Вестник БашГУ. 2010. №2. С. 320–324.
9. Collins M. Three generative, lexicalized models for statistical parsing // In Proceedings of ACL 35. 1997.
10. Dornyei Z. Research Methods in Applied Linguistics. OUP, 2007. [Electronic book].
11. Diez D. M., Barr C. D., Cetinkaya-Rundel, M. OpenIntro statistics. CreateSpace, 2012. Т. 12.
12. Harris Z. Methods in Structural Linguistics. Chicago: University of Chicago Press, 1951. [Electronic book].
13. Jurafsky D., Martin J. H. Speech and Language Processing. NJ: Prentice Hall, 2000. [Electronic book].

*Поступила в редакцию 05.09.2018 г.*

## STATISTICAL METHODS IN THE STUDY OF TEXTS

**© V. A. Grechachin**

*Bashkir State University*
*32 Zaki Validi Street, 450076 Ufa, Republic of Bashkortostan, Russia.*

*Phone: +7 (937) 789 28 97.*
*Email: vitaley.grechachin@gmail.com*

The aim of the article is to consider statistical methods in the context of text studies and to determine the possibilities of their application. The main attention was paid to the quantitative characteristics of linguistic units. The difference in the frequency of use of linguistic units was considered. The ways of calculating the frequency of use of linguistic units were described. The possibilities for application of the considered statistical methods in comparative studies of texts were determined. The possibilities of using descriptive statistics for linguistic studies were considered. The work is based on the material of a small corpus of texts; the analysis of the peculiarities of the use of certain parts of speech in the works of several authors was carried out; it can be used to study the author's style. In addition, authors of the article considered the research potential of various visualizations of statistical data obtained through the analysis of frequency of using units in texts. The data gathered for the analysis can be used in further studies considering the potential of application of statistical methods in linguistics.

**Keywords:** statistics, quantitative linguistics, corpus linguistics, frequency of using linguistic units.

Published in Russian. Do not hesitate to contact us at bulletin_bsu@mail.ru if you need translation of the article.

### REFERENCES

1. Arapov M. V. Kvantitativnaya lingvistika [Quantitative linguistics]. Moscow: Nauka, 1988.
2. Baranov A. V. Vvedenie v prikladnuyu lingvistiku [Introduction to applied linguistics]. Moscow: Editorial URSS, 2001. [Elektronnaya kniga]. Vaisgerber I. L. Rodnoi yazyk i formirovanie dukha. Moscow, 2004.
3. Grechachin, V. A. Mezhdunarodnyi nauchno-issled. zhurnal. 2016. No. 6(48). Pt. 4. Pp. 25–27. doi: 10.18454/IRJ.2016.48.070.
4. Grzhibovskii A. M. Ekologiya cheloveka. 2008. No. 9 URL: https://cyberleninka.ru/article/n/korrelyatsionnyy-analiz
5. Kashcheeva A. V. Sotsial'no-ekonomicheskie yavleniya i protsessy. 2013. No. 3(049). URL: https://cyberleninka.ru/article/n/kvantitativnye-i-kachestvennye-metody-issledovaniya-v-prikladnoy-lingvistike (data obrashcheniya: 19.10.2018).
6. Kochetkova N. A. Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh. 2013. No. 16. URL: https://cyberleninka.ru/article/n/statisticheskie-yazykovye-metody-kollokatsii-i-kolligatsii.
7. Napreenko G. V. Vestn. Tom. gos. un-ta. 2014. No. 379. URL: https://cyberleninka.ru/article/n/identifikatsiya-teksta-po-ego-avtorskoy-prinadlezhnosti-na-leksicheskom-urovne-formalno-kolichestvennaya-model (data obrashcheniya: 19.10.2018).
8. Ibragimova V. L., Fatkullina F. G. Vestnik BashGU. 2010. No. 2. Pp. 320–324.
9. Collins M. In Proceedings of ACL 35. 1997.
10. Dornyei Z. Research Methods in Applied Linguistics. OUP, 2007. [Electronic book].
11. Diez D. M., Barr C. D., Cetinkaya-Rundel, M. OpenIntro statistics. CreateSpace, 2012. Vol. 12.
12. Harris Z. Methods in Structural Linguistics. Chicago: University of Chicago Press, 1951. [Electronic book].
13. Jurafsky D., Martin J. H. Speech and Language Processing. NJ: Prentice Hall, 2000. [Electronic book].

*Received 05.09.2018.*