© Н.Н. ЖУРАВЛЕВА

УДК 81

ПРИМЕНЕНИЕ КОЛИЧЕСТВЕННЫХ МЕТОДОВ ПРИ АНАЛИЗЕ СТИЛЯ АВТОРА И РЕШЕНИИ ПРОБЛЕМ АТРИБУЦИИ

АННОТАЦИЯ. В этой статье представлен краткий обзор количественных методов лингвистического анализа идиостиля в диахроническом аспекте. Рассматриваются работы известных авторов, занимающихся применением статистических и компьютерных методов при решении проблемы атрибуции текста.

SUMMARY. The article presents a brief overview of the quantitative methods of linguistic analysis of style in the diachronic aspect. It considers the works of notable authors that deal with the application of statistical and computer methods to the solution of the problems of text attribution.

КЛЮЧЕВЫЕ СЛОВА. Стиль, атрибуция, количественные методы, статистические методы, корпус тестов, конкорданс.

KEY WORDS. Style, attribution, quantitative methods, statistical methods, corpus of texts, concordance

Цель настоящей статьи — сделать краткий обзор развития количественных методов в лингвистике и выяснить, каким образом исследователи применяли их для выявления особенностей стиля того или иного автора или решения проблем атрибуции.

Прежде чем проанализировать развитие количественных методов анализа авторского стиля, необходимо рассмотреть определение стиля.

По определению Г. Хердана, стиль — общая характеристика индивидуального способа выражения личности в языке. Стиль понимается им как подсознательный фактор, которому автор не может не подчиняться. Следовательно, языковое выражение является в меньшей степени намеренным выбором, как это может показаться на первый взгляд [1; 12]. Человек сам не осознает своего стиля, и его можно распознать так же четко, как и отпечатки пальцев, если только он не намерен скрыть его [2; 54].

Приведем еще одно, упрощенное, определение стиля, предложенное В. Винтером. Стиль может быть охарактеризован как система периодически повторяющихся выборок из перечня произвольных черт языка. Типы выборки могут быть различными: абсолютное исключение произвольных элементов, обязательное включение произвольных черт куда-либо еще, различные степени включения особого варианта без полного исключения конкурирующих черт [3; 3].

Таким образом, из этих двух определений следует, что стиль — это система периодически повторяющихся выборок, характеризующая индивидуальный способ

выражения в языке конкретного человека. Именно такое определение оправдывает применение количественных методов при анализе авторского стиля.

Существует множество различных количественных методов анализа стиля автора. Рассмотрим их в диахроническом аспекте.

История современной статистической стилистики начинается в середине XIX в., когда английский математик Аугустус де Морган в 1851 г. высказал предположение, что различные авторы могут быть определены посредством скрытых статистических черт. Рассматривая проблемы греческой прозы, Морган утверждал, что средняя длина слов в произведении автора может быть характерной чертой авторского стиля. Однако, насколько нам известно, сам де Морган никаких вычислений не делал [цит. по 4; 368].

В середине XIX в. также существовала группа ученых, разрабатывающая так называемый метод «стилометрики». Они подсчитывали количество повторений определенного слова и изменение размера в стихах. Исследователи представляли свои результаты как среднюю величину или процентное соотношение. Школа достигла своей высоты с основанием (около 1874 г.) нового Шекспировского общества. Среди членов Общества были, например, Ф.Г. Фриари (О метрических тестах, применяемых к драматической поэзии Шекспира» (Freagy, 1874)), Дж. К. Инграм (1874), Ф.У. Фурнивал (1887). Главным результатом их работы было открытие медленного, но постоянного изменения стиля в течение 22-х лет, за которые Шекспир написал 36 пьес, начиная с 1589 г., когда ему было 26, и заканчивая 1612 г., когда ему было 48 [цит. по 4; 369].

Термин «стилометрия» («стилометрика») был также использован германским исследователем В. Диттенбергером (1880), который сделал попытку решить проблему атрибуции и хронологии диалогов Платона. Он исследовал частоту употребления слов, особенно служебных, в текстах Платона. Позже его исследования на различных материалах продолжили Е. Зеллер (1887), Ф. Када (1901), Ц. Риттер (1903), последний сравнивал Платона и Гете статистически. [цит. по 4; 369]

Идеи де Моргана были развиты в работах Т.К. Менденхалла, американского геофизика, который, между 1887 и 1901 гг., изучал длину слов в английском. Он понял, что дистрибуция слов различной длины имеет больше возможностей сопоставления стилей, чем простой арифметический способ, предложенный де Морганом. Первая работа Менденхалла «Характеристика состава» (1887 г.) была выдающимся продвижением по направлению к современному стилостатистическому подходу. Он исследовал разницу между литературными стилями Диккенса и Теккерея с точки зрения дистрибуции длины слова и давал примеры других произведений в современных и классических языках. Все его результаты были показаны в форме диаграмм («спектров слов», как он их называл), к сожалению, без списка исконных чисел. В более поздней статье Менденхалл использовал дистрибуцию частотности длины слова в исследовании авторства пьес Шекспира. Он показал, что в каждом отрывке из пьес Шекспира было больше слов из четырех букв, чем из трех. В то же время у Бэкона было больше слов из трех букв, чем из четырех. Следовательно, дистрибуция длины слов у Бэкона значительно отличалась от Шекспировской.

Опубликованные работы Менденхалла не привлекали большого внимания в то время. В тот ранний период статистической стилистики можно назвать лишь небольшое количество работ. Например, Л.А. Шерман (1888) и Х.А. Паркер (1896) изучали длину предложений. К. Хилрет (1897) внес новый вклад в проблему Шекспир — Бэкон. В. Лутоставский (1897) использовал статистические методы в установлении хронологии диалогов Платона, Л. Франк (1909) писал о частотности цветообозначений в работах Гете. П. Парцингер (1911) изучал эволюцию стиля Цицерона. [цит. по 4; 370]

В России Н.А. Морозов поднял проблему отличия плагиата от оригинальных работ известных авторов. В 1915 г. он опубликовал статью «Лингвистические спектры». Предшествующие ему исследователи опирались, главным образом, на частоту употребления знаменательных слов. Н.А. Морозов, применяя простые вычислительные способы, рассматривал частоту употребления служебных слов и их вариаций в индивидуальных текстах [5].

В 20-е гг. XX в. можно назвать только несколько серьезных исследователей по стилостатистике, таких как Р.Е. Паркер (1925), З.Е. Чендлер (1928), М. Пэрри (1928), и, в особенности, А. Бусман (1925), автора так называемого соотношения глагол-прилагательное. [цит. по 4; 371]

В 30-е гг. XX в. был сделан новый шаг в применении статистических методов в стилистике такими лингвистами, как Дж. В. Флетчер (1934), рассматривавшим развитие стиля Спенсера, Г.М. Боллинг (1937), с критическим эссе по статистическому исследованию языка Гомера, Дж. Б. Кэрролл (1938), поднимавший проблему разнообразия словаря, и У.Г. Юл (1938), первым исследовавший дистрибуцию длины предложений как статистическую характеристику стиля.

Именно с него начинается применение современных статистических методов в стилистике. С этого периода применение статистических методов в исследовании стиля распространяется по всему миру. Резко возрос интерес к статистической лингвистике, особенно в 1960-70 гг. (Дж. Б. Кэрролл, Г. Хердан, Х.Х. Сомерс, Ч. Мюллер, Б. Келман, Л.Т. Милик, Дж. Мистрик, Л. Долежел, К.Б. Уильямс, Б.Н. Головин, Й. Краус, М.Н. Кожина и др.) [цит. по 4; 371]

Именно в этот период возникают и развиваются разнообразные идеи анализа авторского стиля.

С точки зрения Л. Долежела, основы статистической теории стиля могут быть суммированы в простом утверждении: стиль — это теоретико-вероятностное понятие. Долежел уверен в том, что тот или иной признак не может безоговорочно определяться привычками автора. Вероятностное распределение признака является скорее тенденцией. Таким образом, вероятностное понятие позволяет описать стиль не как неизменную привычку, а скорее предпочтение той или иной формы выражения. Общий характер стиля вызван скорее *степенью* наличия или отсутствия определенного способа выражения, чем его исключительным использованием или полным подавлением. Следовательно, вероятностный подход раскрывает гибкий характер стилистических черт: эти черты противостоят любому описанию с точки зрения необходимости или с точки зрения строгих правил или запретов [6; 11]. В то же время это говорит о том, что исследователь при выявлении того или иного доминирующего признака должен проверять, действительно ли признак характеризует именно автора, а не эпоху, в которой было написано произ-

ведение, жанр или сюжет. Для этого нужно не только рассматривать в сопоставительном аспекте несколько произведений данного автора, но и, как предлагает Х.Х. Сомерс, анализировать различия между тестируемым текстом и работами известных авторов. [7; 128].

С появлением компьютеров в 60-е гг. XX в. стало возможным их применение в лингвистических исследованиях, в частности, в решении проблемы авторства и идиостиля. Привлечение внимания лингвистов к компьютеру связано с его способностью хранить большие объемы информации, в нашем случае корпуса текстов, и находить употребления слов, групп слов, повторяющихся слогов и т.д. Таким образом, компьютер может обрабатывать огромные объемы информации в доли секунды. С.И. Сиделоу и У.А. Сиделоу ввели новый термин «вычислительная стилистика» («computational stylistics»), под которым они понимают количественно строгое и глубокое изучение стиля в естественном языке. Вычислительная стилистика имеет обширное практическое применение для различных сфер — от машинного перевода и автоматического реферирования до общественных и гуманитарных наук. Одним из частных ее применений, безусловно, является исследование идиостиля и решение проблем атрибуции [8; 1].

М.Х.Т. Элфорд в статье «Применение компьютера в изучении языка и в определении авторства» пытается решить проблему атрибуции следующим образом: он делает вывод о том, что слова, являющиеся низкочастотными в общем измерении, становятся высокочастотными в частном измерении. Таким образом, компьютерные данные показывают, что если низкочастотное слово однажды встретилось в тексте, то дальнейшая частотность его употребления будет примерно в десять раз больше, чем в общем употреблении [9; 84-85]. Такой вывод позволяет утверждать, что можно выявить стилистические особенности автора на основании частотности употребления определенных лексем.

К подобным выводам приходит и Г. Хердан в работе «Quantitative Linguistics» (Квантитативная лингвистика): стиль, по его мнению, может быть охарактеризован постоянным соотношением между однородностью и разнообразием частотности слов [10; 71]. Г. Хердан говорит о следующих соотношениях: специальная лексика/общая лексика, специальные случаи употребления/общие случаи употребления, специальная лексика/ общие случаи употребления [1; 20-22].

С началом применения компьютеров в лингвистическом исследовании возникла проблема построения конкордансов. Конкорданс — список контекстов, в которых встречается определенное слово или последовательность слов. Программа конкорданс является базовым инструментом корпусной лингвистики, превращающим электронные тексты в базу данных, которая может быть исследована [11]. С помощью данных программ осуществляется поиск отдельного слова, поиск словосочетаний, отдельных частей слов, данная программа способна создавать список коллокаций, а также собирать данные о частотности употребления [12; 57]

Одним из ученых, занимающихся построением конкордансов, был Д. Росс, объяснявший проблему их построения тем фактом, что компьютеру достаточно трудно распознать части речи. Он критикует фреймовый метод Эллегарда, который предлагает распознавать их при помощи порядка слов. Так, например, существительное — это часть речи, стоящая между детерминативом и другой функциональной частью речи, а прилагательное — это часть речи, которая

может находиться между предлогом и существительным. Трудности возникают, когда слово имеет несколько категорий. Вместо этого он предлагает свою программу конкорданса «EYEBALL», в которой маркируются только слова, находящиеся в фокусе внимания, и каждая новая фраза или придаточное предложение рассматривается в изоляции от остальных. Еще одно отличие данной программы от других — включение функциональных маркеров, которые делают синтаксическое описание значительно более законченным, чем только категориальные маркеры [13; 88].

В 1970-90-е гг. все больше исследователей проявляют интерес к применению компьютерной обработки данных при анализе текстов, как в синтаксическом, так и в грамматическом, лексическом аспектах.

Обязательное применение автоматической обработки данных лежит в основе работ Ю.В. Сидорова, И.О. Тарнопольской, Д.В. Хмелева. В исследованиях текстов, проводимых под руководством Л.В. Милова, атрибуция текстов проводится при помощи построения графов «сильных связей» по матрице частот парной встречаемости грамматических классов слов и происходит при помощи специальной компьютерной программы [14].

Одним из наиболее современных отечественных лингвистов, занимающихся статистикой, является Г.Я. Мартыненко. В 1988 г. он написал монографию «Основы стилеметрии» и на протяжении более чем двадцати лет занимается статистическим методами в лингвистике. Наиболее поздним его исследованием является исследование теории так называемого «золотого сечения», придуманного еще пифагорейцами, в лингвистике. Рассматривая синтаксические структуры с точки зрения меры синтаксической сложности, ритмические структуры, соотношение однократных и многократных лексем, он приходит к выводу о том, что все рассмотренные им соотношение регулируются законом золотого сечения [15]. Примечательным здесь может показаться то, что с помощью данного закона он попытался рассмотреть не один уровень, а несколько (фонемный, морфологический, синтаксический), что позволяет провести анализ стиля с разных сторон.

Другим современным отечественным лингвистом, занимающимся статистическими методами атрибуции текста, является М.А. Марусенко. Ему принадлежит идея теории распознавания образов. Он разделяет процедуру атрибуции на три относительно самостоятельных этапа:

- формирование литературно-критической гипотезы,
- проверка литературно-критической атрибутивной гипотезы методами теории распознавания образов,
- интерпретация результатов проверки атрибутивной гипотезы.

В данной работе статистико-вероятностные методы анализа языка и стиля произведения используются автором для проверки атрибутивной гипотезы [16; 25]

Таким образом, стиль нами понимается как теоретико-вероятностное понятие, а именно — система повторяющихся выборок, характеризующих способ выражения в языке, чаще всего подсознательный. Для анализа стиля мы не можем искать исключительное использование или полное подавление того или иного признака, нужно лишь определять *степень* его наличия или отсутствия. Для этого наиболее удобными будут являться количественные методы, а и наиболее применимой методика конкорданс (список контекстов)

СПИСОК ЛИТЕРАТУРЫ

1. Herdan, G. The advanced theory of language as choice and chance. Berlin: Springer-Verlag, 1966. 365 p.

2. Booth, A.D.; Brandwood, L.; Cleave, J.P. Mechanical resolution of linguistic problems. London: Butterworks scientific publications, 1958. 306 p.

3. Winter, W. Styles as dialects // Statistics and style / Edited by Lubomir Doležel, Richard W. Bailey. New York: American Elsevier Publishing Company, 1969. P. 3-9.

4. Tuldava, J. Stylistics, author identification // Quantitative linguistics: an international handbook / Edited by Reinhard Köhler, Gabriel Altmann, Raĭmond Genrikhovich Piotrovskiĭ. Berlin, New York: de Gruyter, 2005. P.368-387.

5. Морозов Н.А. Новое орудие объективного исследования документов. (Лингвистические спектры как средство для отличения плагиатов от истинных произведений того или иного известного автора и для определения их эпохи). URL: http://www.textology.ru/libr/Morozov.htm

6. Doležel, L. A framework for the statistical analysis of style // Statistics and style / Edited by Lubomir Doležel, Richard W. Bailey. New York: American Elsevier Publishing Company, INC, 1969. P. 10-25

7. Somers, H.H. Statistical methods in literary analysis // The computer in literary style. Introductory essays and studies / Edited by Jacob Leed. Kent, Ohio, USA: Kent State University press, 1966. P. 128- 140.

8. Sedelow, S.Y.; Sedelow, W.A., Jr. A preface to computational linguistics // The computer in literary style. Introductory essays and studies / Edited by Jacob Leed. Kent, Ohio, USA: Kent State University press, 1966. P. 1-13.

9. Alford, M.H.T. Computer assistance in language learning and in authorship identification // Statistics and style / Edited by Lubomir Doležel, Richard W. Bailey. New York: American Elsevier Publishing Company inc, 1969. P. 77-86.

10. Herdan, G. Quantitative linguistics. London: Butterworths, 1964. 284 p.

11. McEnery, T., Wilson, A. Corpus linguistics. URL: http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus1/1fra1.htm.

12. Филипенко Т.В. Использование методов корпусной лингвистики при анализе семантики идиом// Вестник МГУ. Сер. 19. Лингвистика и межкультурная коммуникация. 2004. №1. С.84-88.

13. Ross, D. Beyond the concordance: algorithms for description of English clauses and phrases // The computer and literary style / Edited by A. J. Aitken, R.W. Bailey, N. Hamilton-Smith. Edinburgh: University Press, 1973. P. 85- 99.

14. От Нестора до Фонвизина. Новые методы определения авторства./ под ред. Л.В. Милова. М.: Прогресс, 1994. 378 с.

15. Мартыненко Г.Я. Золотое сечение в нумерологии текста. URL: http://numbernautics.ru

16. Марусенко М.А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л.: Издательство Ленинградского университета, 1990. 168 с.