Родионова Е.С. Методы атрибуции художественных текстов // Структурная и прикладная лингвистика. Вып. 7: Межвуз. сб. / Под ред. А.С. Герда. – СПб.: Изд-во С.-Петерб. Ун-та, 2008. С.118–127.

Е. С. Родионова

МЕТОДЫ АТРИБУЦИИ ХУДОЖЕСТВЕННЫХ ТЕКСТОВ

Проблема установления авторства анонимных и псевдонимных текстов связана как с историко-филологическими, так и с естественно-техническими науками, среди которых особое значение при решении вопроса об атрибуции принимают статистика, теория вероятностей, семиотика и другие. При этом постановка задачи и применение результатов атрибуции относятся к литературоведческой сфере, а аппарат и методы получения результатов к сфере математической, требующей использования современных научных теорий и вычислительных средств.

При решении вопроса об атрибуции какому-либо автору спорного произведения необходимо, чтобы аргументы характеризовали его с трех сторон: биографической, идеологической и стилистической [Берков, 1958, сс.184-185]. Фиксация стиля атрибутируемого произведения и сопоставление его со стилем предполагаемых авторов является неотъемлемой частью задачи атрибуции.

В классификации принципов атрибуции, составленной академиком В.В. Виноградовым, были выделены 6 объективных и 5 субъективных принципа атрибуции [Виноградов, 1961]. Лингвостилистический анализ с применением математического аппарата является наиболее плодотворным объективным методом атрибуции, поскольку разнообразие математических методов анализа объектов различной природы позволяет успешно использовать их в практике установления авторства. Любое стилистическое исследование будет носить субъективный характер без учета количественных показателей, именно количественная оценка однородности или неоднородности состава и структуры сравниваемых языковых систем составляет основу лингвостилистического анализа, без которого невозможна объективная атрибуция.

До 70-х годов XX века в практике атрибуции доминировали историко-документальные и филологические методы исследования. Для выявления авторских особенностей применялась субъективная методика атрибуции, в соответствии с которой субъективно отбирались внешние детали авторского стиля, такие как любимые слова, термины, выражения.

Применение математико-статистических методов было начато в конце XIX века в целях атрибуции произведений античных авторов. Работы Кэмпбелля [Campbell, 1867] и Люгославского [Lutoslawski, 1897] основывались на установлении «оригинальных» слов и на позиционном расположении определяющих и определяемых слов.

Первым отечественным ученым, использовавшим математический аппарат для решения задачи атрибуции считается Н.А. Морозов, опубликовавший в 1915 г. статью «Лингвистические спектры» [Морозов, 1915], в которой впервые обосновал идею о том, что при описании стиля необходимо использование не одного аспекта, а нескольких. Кроме того, в отличие от предшествующих исследователей, филологов-классиков, опиравшихся при атрибуции на частоту употребления знаменательных слов, Н.А. Морозов полагал, что для индивидуального стиля писателя показательными являются именно служебные слова, которые не связаны с темой и содержанием книги. Метод, предложенный Н.А. Морозовым, лег в основу многих исследований по лексическому составу языка писателей и встречается в современных исследованиях. Вместе с тем он не может служить основой для достоверного стилистического исследования, поскольку не выходит за рамки лексического анализа и состава предложения.

Одними из первых критические отзывы на эту работу дали А.А. Марков [Марков, 1916] и В.Э. Сеземан [Сеземан, 1918, с.73], со стороны которых сам метод лингвистических спектров возражений не получил. Вместе с тем А.А. Марков отметил недостаточную проверку Н.А. Морозовым устойчивости предлагаемых количественных характеристик и предостерег последующих исследователей от использования таких характеристик, которые при увеличении объема текста сходятся к средним характеристикам русского языка.

Наиболее четко необходимость отказа от субъективных методов атрибуции стала ощущаться в 50-60-е годы XX века. В этот период наметилась тенденция угасания традиционной методики. Общее положение дел в теории атрибуции было охарактеризовано академиком В.В. Виноградовым как кризисное. Он отмечал, что субъективные методы атрибуции уже отжили свой век, и «решение проблемы авторства по отношению к литературе нового времени требует от исследователей глубоких объективных знаний системы индивидуального стиля конкретного автора» [Виноградов, 1961, с.85].

С 60-70-х годов XX века при описании индивидуального стиля лингвоматематические методы стали применяться все шире, благодаря чему накапливались данные о свойствах единиц языка, и формировался специальный научный аппарат атрибуции текстов. В работах А.Л. Гришунина [Гришунин, 1960], П. Вашака

[Вашак, 1974], В.И. Батова [Батов, 1977] и других, разрабатывались методы статистики применительно к лексике, а также к грамматике.

Т.А. Якубайтис и А.Н. Скляревичем была проведена типология научно-технических, поэтических, драматических и других текстов по числу повторений какой-либо части речи [Якубайтис, Скляревич, 1982]. В ходе этой работы были сделаны выводы о структурности системы частей речи, и было обосновано утверждение о том, что достоверность атрибуции типов текстов возрастает с ростом количества анализируемых признаков. Это последнее утверждение, сделанное авторами исследования, является ошибочным: с увеличением числа анализируемых признаков достоверность атрибуции возрастает далеко не всегда. При использовании большого количества коррелируемых признаков возможно появление «шума», который может затруднять анализ и сильно искажать результаты. Высокая эффективность применения набора диагностирующих признаков может быть достигнута лишь при их независимости друг от друга, то есть при отсутствии между ними сильной корреляции. Для определения набора информативных параметров требуется специальная процедура, которая во время написания рассматриваемой работы в подобных задачах еще не применялась.

Еще одно исследование лексики текста было проведено А.П. Василевичем, опубликовавшим в 1981 году работу, посвященную изучению употребления цветонаименований [Василевич, 1981]. Основой анализа цветообозначения в поэзии и прозе 19-20 веков стали такие величины, как индекс лексической оригинальности (отношение числа редких к числу частых слов), индекс морфологической оригинальности (отношение сложных слов к простым) и индекс насыщенности (частота употребления цветонаименований).

Использование специально разработанных индексов для оценки лексической структуры текста было обусловлено стремлением разработать новый универсальный аппарат для объективного анализа лексики. Многие ученые продолжают заниматься разработкой все новых индексов, мер и оценок лексического состава, отличающихся теми или иными достоинствами и недостатками, и не рассматривают уже готовые существующие решения в смежных областях знания, например в математике, где для описания объектов различной природы давно и успешно применяется теория распознавания образов.

Одно из последних исследований по атрибуции текстов с использованием специально разработанной формулы вычисления «межтекстового расстояния» было проведено французским специалистом по анализу речи Д. Лаббе в 2001 году [Labbé, Labbé, 2001]. Вычисление «межтекстового расстояния» подразумевает анализ

лексического состава двух текстов и определение меры их близости или удаленности друг от друга, и определяется как «сумма разностей частот всех вокабул из наименьшего текста и из всех возможных выборок равных наименьшему тексту, которые можно извлечь из большего текста» [Labbé, Labbé, 2001, p.217]:

$$D_{Va,b(u)} = \sum_{V \in (A,B)} \left| F_{ia} - E_{ia(u)} \right|, \quad \text{где}$$

F_{ia} - частота вокабулы i из множества A, $E_{ia(u)}$ - математическое ожидание вокабулы i.

Д. Лаббе применил новый метод определения авторства в исследовании, посвященном сравнению лексического состава театров Мольера и П. Корнеля. Лексический анализ текстов происходил с помощью автоматической процедуры морфологического анализа, которая предусматривала представление каждого слова в виде записи, состоящей из трех компонентов: словоформы, вокабулы и соответствующей части речи. По полученным данным было вычислено «межтекстовое расстояние», и результаты были представлены в виде древовидной классификации. После анализа полученного дерева зависимостей Д. Лаббе сделал вывод о принадлежности П. Корнелю лучших пьес Мольера в стихах, составляющих примерно половину театра Мольера.

Метод, предложенный Д. Лаббе, имеет ряд существенных недостатков, связанных как с полной автоматизацией лексического разбора текста, при котором не учитываются синтаксические роли слов в предложении и морфологические категории, так и с отсутствием вероятностного подхода. Анализ одного лишь лексического уровня также не может служить достаточно достоверным критерием атрибуции текстов, поскольку при подделке или имитации текста лексического сходства добиться легче всего. Несомненно, что атрибуция, основанная на использовании такого несовершенного математического и теоретического аппарата не может считаться доказанной.

К 80-м годам XX века было опубликовано множество работ, описывающих лексический состав языка в различных аспектах. Опыт квантитативно-лингвистических исследований был обобщен в монографии Ю. Тулдавы [Тулдава, 1987]. Эксперименты по выявлению возможности классификации текстов с помощью кластер-анализа, описанные в этой работе, показывают, насколько остро к этому времени встал вопрос о поиске адекватного решения проблемы отбора информативных параметров. В монографии была сформулирована идея о связи признаков, которая являлась предпосылкой для разработки математического аппарата оценки связей между параметрами и использования для этой цели теории распознавания образов.

В 70-е – 80-е годы XX века в отечественной лингвистике был проведен ряд исследований, посвященных квантитативно-структурному изучению текстов на синтаксическом уровне. Интерес исследователей к проведению стилистического анализа именно в синтаксическом аспекте был, прежде всего, обусловлен пониманием в современной науке стиля как категории структурно-синтаксической. Кроме того, использование синтаксического анализа подразумевает комплексный подход к анализу текста благодаря выделению специфических признаков как на синтаксическом, так и на лексическом, фразеологическом и морфологическом уровнях. Исходя из этих соображений, в 1981 году И.П. Севбо опубликовала работу, посвященную графическому представлению синтаксических структур в виде деревьев зависимостей [Севбо, 1981]. В 1983 году Г.Я. Мартыненко провел изучение синтаксических структур в рамках типичных фраз и предложений, отношения зависимости и однородности были представлены в этой работе в виде стрелочно-скобочной записи [Мартыненко, 1983]. Математические методы атрибуции, основанные на анализе синтаксических структур, являются наиболее эффективными и интенсивно развиваются в настоящее время. Однако методы стилистической диагностики, основанные на анализе деревьев зависимостей и деревьев составляющих, связаны с характеристиками предложения, а не текста. Такими характеристиками предложения являются как диагностические параметры, предложенные И.П. Севбо, так и меры сложности, анализируемые в работах Г.М. Мартыненко. Разработка усредненного графа для стиля анализируемого автора абсолютно невозможна, а значит, невозможно и применение анализа графов для характеристики текста в целом, но именно анализ текста должен лежать в основе эффективной методики фиксации авторского стиля. Тот же недостаток, связанный с ограничением метода характеристикой предложения, отличает и метод атрибуции, разработанный в 1994 году группой ученых под руководством Л.В. Милова [Милов и др., 1994]. Обработка текстового материала древних текстов, описанная в совместной монографии, заключалась в построении графов «сильных связей» по матрице частот парной встречаемости грамматических классов слов.

Период с конца 70-х годов XX века до настоящего времени отмечен бурным развитием вычислительной техники и программного обеспечения, в связи с чем все больше исследователей проявляют интерес к применению компьютерной обработки данных при анализе текстов, как в синтаксическом, так и в грамматическом, морфемном, лексическом аспектах. Одна из первых методик установления авторства, основанная на анализе текста с автоматизированным получением частотных словарей и статистических данных, была предложена в работе норвежского филолога Г. Хетсо в 1978 году [Хетсо, 1978]. Отметим здесь одну из самых серьезных на наш взгляд ошибок, допущенных при

разработке этого метода, которая заключается в использовании параметра средняя длина предложения. В 90-е годы XX века на кафедре математической лингвистики СПбГУ группой молодых ученых под руководством профессора М.А. Марусенко был проведен подробный анализ значений, принимаемых этим параметром на выборках различного объема. Оказалось, что графики распределения длин предложения обладают такой характерной особенностью, как «многовершинность», которая становится все более явной с увеличением объема выборки. В данном случае явление «многовершинности» свидетельствует о том, что этот параметр состоит из нескольких параметров, характеризующих различные типы предложения. Таким образом, параметр средняя длина предложения «является статистически бессмысленным, так как представляет собой смесь распределений» [Марусенко, 2003, с.116]. В статистике давно были разработаны специальные методы работы со смесью распределений. С помощью процедуры «расщепления» происходит выделение отдельных распределений для дальнейшей обработки данных. Без этой процедуры анализ, проведенный с использованием такого параметра, не может считаться результативным.

На сегодняшний день автоматическая обработка текстов используется в большей или меньшей степени во всех современных исследованиях, посвященных разработке новых методик атрибуции. Вполне объяснимое стремление ученых к применению автоматической стилистической диагностики и автоматизированного поиска индивидуальных характеристик авторского стиля приводит к тому, что предпочтение в стилистических исследованиях отдается анализу любых других языковых уровней, кроме синтаксического. Подавляющее большинство исследований по лингвистической стилистике по-прежнему посвящено лексическому анализу, а именно изучению лексики синтаксиса и поэтической семантики, при этом синтаксис писателя остается в тени. Причинами недостаточного внимания к синтаксическому анализу в современной стилистике являются объективные трудности, связанные как с поиском типических характеристик авторского стиля, так и с обработкой содержащихся в них информации [Севбо, 1981, с.97]. Если грамотный синтаксический анализ, выполненный путем компьютерной обработки данных, представляет собой отдельную сложную задачу, не решенную до настоящего времени, то анализ лексического уровня стал менее трудоемким благодаря автоматизированному составлению частотных словарей. Разрабатываются и методы автоматизации обработки текстовых данных при грамматическом и морфологическом анализе. Однако зависимость стилистического анализа от компьютерной обработки данных и от методов, для нее предназначенных, приводит к

упрощению методологической основы исследований, что, в конечном итоге, делает методы атрибуции текста менее эффективными.

Итак, обзор истории развития научной мысли в области параметризации авторского стиля позволяет выделить следующие основные тенденции: переход от одномерных классификаций к описанию объектов в многомерном признаковом пространстве, все более широкое использование компьютерной обработки данных, а также возникший в последние десятилетия интерес исследователей к применению синтаксического анализа при описании авторского стиля.

История развития методов атрибуции привела к пониманию того, что эффективный метод стилистического анализа в целях определения авторства должен обладать следующими характеристиками:

1. С помощью метода стилистического анализа должны определяться характеристики текста, а не отдельного предложения;
2. Описание текста должно охватывать разные уровни языковой системы, и помимо лексического состава текста должна рассматриваться и его структура;
3. Необходимо применение многомерных классификаций.

Кроме того, изучение связей между параметрами показало, что простое увеличение числа параметров не приводит к увеличению эффективности анализа и, следовательно, необходим специальный математический аппарат для оценки связей между параметрами. Отбор информативных параметров должен заключаться в исключении «шумовых» параметров, обладающих сильной корреляцией друг с другом.

Всем этим требованиям отвечает такой готовый математический аппарат, как теория распознавания образов. Методы распознавания образов были впервые применены при атрибуции анонимных и псевдонимных произведений на основе индивидуальных характеристик авторского стиля в работе М.А. Марусенко в 1990 году [Марусенко, 1990].

В данной работе текст рассматривается как сложный лингвистический объект, характеризующийся обширным инвентарем элементов и многоуровневостью анализа. Исходя из требования адекватного его описания, в основу нового метода атрибуции анонимных и псевдонимных произведений был положен многомерный статистический анализ, представленный в его наиболее развитой форме – теории распознавания образов. В терминах распознавания образов стиль определяется как «набор свойств (параметров), характеризующих состав, способы объединения и статистико-вероятностные закономерности употребления речевых средств, образующих данную разновидность языка» [Марусенко, 1990, с.17]. Набором свойств, характеризующих структуру текста в синтаксическом аспекте, становится в данном случае совокупность информативных

параметров, чей состав определяется путем выполнения специальной процедуры отбора информативных параметров для каждого конкретного случая.

Разработанный М.А. Марусенко математический аппарат был положен в основу нашего исследования, посвященного решению проблемы принадлежности драматургических работ, приписываемых Мольеру [Родионова, 2007]. На этапе описания атрибутируемых объектов на языке параметров из априорного словаря параметров была применена ручная обработка данных. Статистическая обработка полученных данных и построение на их основе матрицы корреляции производилось с помощью компьютерных программ. Данный метод атрибуции анонимных и псевдонимных произведений представляется на сегодняшний день наиболее перспективным в силу применения многомерной классификации, основанной на теории распознавания образов, и описания индивидуального авторского стиля в синтаксическом аспекте.

Список литературы

1. Campbell L. The Sophisties and Polilicus of Plato. Oxford. 1867.
2. Labbé C., Labbé D. Inter-textual distance and authorship attribution Corneille and Molière. // Journal of Quantitative Lingustics. 2001. Vol. 8. №3. pp.213-231.
3. Lutoslawski W. The origin and growth of Plato's logic. London. 1897.
4. Батов В.И., Сорокин Ю.А. Опыт построения методики для установления авторства текстов // Изв. АН СССР. Сер. Лит. И языка. 1977. Т. 36. №4.
5. Берков П.Н. Об установлении авторства анонимных и псевдонимных произведений XVIII века. // Русская литература. 1958. №2.
6. Василевич А.П. Цветонаименования как характеристика языка писателея // Лингвистика текста и стилистика: Уч. Зап. Тартуского гос. Университета. Вып. 585 / Под ред. Ю.А. Тулдава. Тарту, 1981.
7. Вашак П. Длина слова и длина предложения в текстах одного автора // Вопросы статистической стилистики / Под. Ред. Б.Н. Головина. Киев, 1974.
8. Виноградов В.В. Проблема авторства и теория стилей. М., 1961.
9. Гришунин А.Л. Опыт обследования употребительности языковых дублетов в целях атрибуции // Вопросы текстологии. Вып. 2 / Под ред. В.С. Нечаевой. М., 1960.
10. Марков А.А. Об одном применении статистического метода. // Известия Имп.Акад.наук, серия VI, Т.X, N4, 1916.
11. Мартыненко Г.Я. Многомерный синтаксический анализ художественной прозы // Структурная и прикладная лингвистика. Л.: Изд-во ЛГУ, 1983. Вып. 2. С. 58-72.
12. Марусенко М.А. Атрибуция анонимных и псевдонимных текстов как типичная задача распознавания образов // Историография и источниковедение отечественной истории. СПб, 2003. Вып. 3.
13. Милов Л.В., Бородкин Л.И., Иванова Т.В. и др. От Нестора до Фонвизина: новые методы определения авторства. М., 1994.
14. Морозов Н.А. Лингвистические спектры: Средство для отличения плагиатов от истинных произведений того или другого известного автора: Стилеметрический этюд// Изв. Отд. русского языка и словесности Имп. Акад. Наук. Т. XX. Кн. 4. 1915.
15. Родионова Е.С. Параметризация стилей. Отбор информативных параметров при атрибуции стихотворных пьес Мольера // Вестник Санкт-Петербургского университета. Сер. 9, вып. 2, ч. 2. – Спб, 2007. С. 61-67.

16. Севбо И.П. Графическое представление синтаксических структур и стилистическая диагностика. Киев, 1981.
17. Сеземан В.Э. «Лингвистические спектры» г. Морозова и Платоновский вопрос // Изв. Отд. русского языка и словесности Имп. Акад. Наук. Т. XXII. Кн.2. 1918.
18. Тулдава Ю.А. Проблемы и методы квантитативно-системного исследования лексики. Таллинн, 1987.
19. Хетсо Г. Проблема авторства в романе «Тихий Дон» // Scando-slavica. Т. 24. 1978.
20. Якубайтис Т.А., Скляревич А.Н. Вероятностная атрибуция типа по нескольким морфологическим признакам. Рига, 1982.