Родионова Е.С. Параметризация стилей: отбор информативных параметров при атрибуции пьес Мольера // Вестник Санкт-Петербургского университета. Сер. 9. Филология. Востоковедение. Журналистика. – Вып. 2 – Ч. 2. – 2007. – С. 61–67

Е. С. Родионова

ПАРАМЕТРИЗАЦИЯ СТИЛЕЙ: ОТБОР ИНФОРМАТИВНЫХ ПАРАМЕТРОВ ПРИ АТРИБУЦИИ ПЬЕС МОЛЬЕРА

Одной из задач стилеметрии – прикладной филологической дисциплины, занимающейся измерением стилевых характеристик, – является атрибуция анонимных и псевдонимных текстов. При решении вопроса об атрибуции какому-либо автору спорного произведения необходимо, чтобы аргументы характеризовали его с трех сторон: биографической, идеологической и стилистической[1]. Фиксация стиля атрибутируемого произведения и сопоставление его со стилем предполагаемых авторов является неотъемлемой частью задачи атрибуции, которая невозможна без выявления специфических языковых признаков текста.

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, фразеологические обороты и выражения. При стилистическом анализе исследователи уделяли внимание лишь лексическому составу текста, который очень тесно связан с темой и содержанием, и, следовательно, его оценка не может являться достаточным критерием при атрибуции текста.

Первым отечественным ученым, использовавшим математический аппарат для решения задачи атрибуции считается Н.А. Морозов, опубликовавший в 1915 году статью «Лингвистические спектры» [2], 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. Таким образом, в этой работе впервые была обозначена необходимость анализа и других уровней языка, помимо лексического. Для определения стиля Н.А. Морозовым были использованы такие количественные характеристики, как комбинации употреблений предлогов *в* и *на* и частицы *не*.

Первый отклик с критикой на статью Н.А. Морозова был получен от академика А.А. Маркова, который в работе «Об одном применении статистического метода», [3] 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».[4] Самым трудным моментом при определении авторства с использованием математических методов, по его мнению, является «выделение непререкаемых, бесспорных качеств индивидуального стиля в их структурном единстве».[5] Таким образом, В.В. Виноградов обозначил важнейшую задачу, возникающую при атрибуции анонимного произведения –

фиксацию стиля автора и формирование рабочего словаря параметров распознающей системы.

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[6], П. Вашака[7], Т.А. Якубайтиса[8] и других, разрабатывались методы статистики применительно к лексике, а также к грамматике.

С 80-х годов XX века появляются все новые методы атрибуции, основанные на анализе синтаксических структур, которые интенсивно развиваются и в настоящее время. В 1981 году И.П. Севбо опубликовала работу[9], посвященную графическому представлению синтаксических структур в виде деревьев зависимостей. В 1983 году Г.Я. Мартыненко[10] провел изучение синтаксических структур в рамках типичных фраз и предложений, отношения зависимости и однородности были представлены в этой работе в виде стрелочно-скобочной записи. В работе М.А. Марусенко[11] проблема атрибуции была впервые решена методами распознавания образов на основе индивидуальных характеристик авторского стиля.

С этого времени можно отметить все возрастающий интерес исследователей к применению компьютерной обработки данных при анализе текстов как в синтаксическом, так и в грамматическом, морфемном, лексическом аспектах. Автоматическая обработка данных применялась в исследованиях Г. Хетсо[12, 13], Л.И. Бородкина и Л.В. Милова[14] и других. За последние несколько лет в сети Интернет появились программы, предназначенные для распознавания автора присылаемого пользователем текста или выдачи списка наиболее близких к нему по стилистике авторов из числа входящих в некоторый заранее заданный перечень «эталонных» авторов. Такими программами

являются «ЛингвоАнализатор» Д.В. Хмелева[15] и разработанный на его основе «Атрибутор»[16]. В качестве стилевых признаков в них используются бинарные буквосочетания.

Итак, в истории развития научной мысли в области параметризации авторского стиля можно выделить следующие основные тенденции: постепенный переход от анализа лексического уровня к анализу синтаксическому, переход от одномерных классификаций к описанию объектов в многомерном признаковом пространстве и все более широкое использование компьютерной обработки данных.

Одним из последних исследований, основанном на автоматической обработке текста на лексическом уровне, является работа французского специалиста по анализу речи Д. Лаббе. В 2001 году он представил новый математический метод атрибуции, основанный на анализе лексического состава текстов и вычислении их меры близости или удаленности друг от друга[17]. Д. Лаббе применил новый метод при атрибуции театра Мольера.[18]

Вопрос об авторстве комедий Мольера был поднят в начале XX века французским поэтом Пьером Луи[19], и с течением времени все больше исследователей уделяют пристальное внимание этой проблеме. На сегодняшний день существует несколько гипотез об истинном авторе пьес, приписываемых Мольеру.

Согласно первой выдвинутой гипотезе лучшие пьесы Мольера в стихах была написаны известным французским драматургом Пьером Корнелем.[20] По другой версии, в поддержку которой приводятся различные литературные, биографические и языковые доказательства, Корнель является автором всех произведений Мольера.[21, 22] Помимо Пьера Корнеля в числе возможных авторов пьес Мольера упоминаются также драматург Филипп Кино и поэт Шапель.[23] Наконец, существует общепринятая, официальная точка зрения, согласно которой именно Мольер и есть автор своих произведений. [24]

В ходе проверки гипотезы, согласно которой пьесы Мольера написал Пьер Корнель, Д. Лаббе сделал вывод о принадлежности 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, 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. Проблема атрибуции пьес, подписанных Мольером, на сегодняшний день является очень актуальной, и в нашем исследовании проверка всех существующих гипотез будет осуществлена по методу[26], предусматривающему применение многомерного статистического анализа при фиксации стиля предполагаемых авторов и атрибутируемых текстов.

Д. Лаббе, как и большинство исследователей творчества Мольера, при решении вопроса об атрибуции его произведений рассматривал все работы драматурга, которые представляют собой крайне неоднородную совокупность пьес, написанных как в стихах, так и в прозе. Ввиду того, что предметом самых ожесточенных споров стало авторство именно стихотворных шедевров Мольера, и, исходя из требования соблюдения жанрово-стилевой однородности текстов, нам представляется разумным анализировать только комедии, написанные в стихах. Таким образом, класс атрибутируемых объектов в нашей работе составляют 13 комедий в стихах, приписываемых Мольеру.

При формулировании атрибуционной гипотезы необходимо учитывать все мнения, высказанные различными исследователями и в число возможных авторов включить Мольера, Пьера Корнеля, Филиппа Кино и Шапеля. Однако, исходя из соображений жанровой однородности исследуемого материала, мы исключили из этого списка поэта

Шапеля, поскольку, будучи автором небольших стихотворений, он не написал ни одной пьесы.

Существующие противоречивые гипотезы могут быть представлены в виде следующей литературно-критической атрибуционной гипотезы:

Нулевая гипотезва (H_0): тексты пьес Мольера полностью принадлежат Мольеру и не принадлежат никому из возможных авторов (Корнелю, Кино).

В случае опровержения нулевой гипотезы необходимо будет осуществить проверку сложной альтернативной гипотезы:

(H_a^1): тексты пьес Мольера полностью принадлежат Корнелю.

(H_a^2): тексты пьес Мольера являются совместным произведением Мольера, Корнеля, Кино с определенной долей участия каждого из них.

(H_a^3): в создании пьес Мольера помимо вышеуказанных принимали участие один или несколько неизвестных авторов. В этом случвае необходимо будет попытаться определить число авторов и возможную долю участия каждого из них.

В соответствии с положенной в основу данной работы методикой атрибуции анонимных и псевдонимных произведений, проверка атрибуционной гипотезы выполняется средствами теории распознавания образов и предусматривает реализацию двух независимых этапов: отбор информативных параметров и процедуру распознавания. В настоящей статье описывается выполнение первого этапа: отбор информативных параметров из априорного словаря параметров.

Значение термина «параметр» существенно меняется в зависимости от контекста, в котором он употребляется, в общем же случае параметром называют «величину, значения которой служат для различения элементов некоторого множества между собой»[27]. В нашем исследовании информативные параметры должны разделять два априорных класса:

Корнеля - Ω (Pierre Corneille) и Кино - Ω (Philippe Quinault). Поскольку все произведения Мольера считаются спорными, априорного класса работ Мольера не существует. Можно будет предположить, что пьесы Мольера действительно написал Мольер в том случае, если они не будут атрибутированы ни Корнелю, ни Кино, и статистический анализ покажет, что класс атрибутируемых объектов является достаточно однородным по своему составу. Итак, был сформирован алфавит классов, мощность (число текстов) и объем (число предложений) которого представлены в табл.1.

Таблица 1. Объем априорных классов

Класс	Мощность	Объем
Ω (Pierre Corneille)	11	11103
Ω (Philippe Quinault)	3	3125

В нашем исследовании априорный словарь параметров, который «представляет собой в значительной степени стандартизованный набор, полученный путем унификации и стандартизации известных средств квантитативного описания стилей, предложенных разными авторами», состоит из 51 параметра[28], релевантных для описания французского языка.

На языке параметров из априорного словаря параметров были описаны априорные классы, для чего были сделаны прикидочные случайные выборки объемом по 200 предложений. В результате определения значений 51 параметра для априорных классов были сформированы две объектно-признаковые матрицы данных, и были вычислены статистические характеристики: среднее арифметическое ($\overline{X}_i$) и стандартное отклонение (σ_i), для каждого класса.

При формировании набора информативных параметров мы применили схему Бонгарда[29], предусматривающую двухступенчатое свертывание параметрического пространства.

На первом этапе было произведено разбиение априорного набора информативных параметров на подмножества параметров, релевантных и не релевантных для различения априорных классов. Релевантность параметров для различения двух априорных классов определялась по t-критерию Стьюдента:

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}} \qquad (1)$$

По формуле 2 было выделено 5 параметров (X2, X4, X21, X31, X32), для которых значение t-критерия оказалось больше 1,96 при уровне значимости α=0.05. Это небольшое число параметров вполне может сформировать информативный набор, однако было решено попытаться еще больше свернуть параметрическое пространство и воспользоваться вторым этапом схемы Бонгарда, предусматривающим обработку корреляционной матрицы связей параметров.

Из двух имеющихся у нас объектно-признаковых матриц соответствующих априорных классов была составлена единая объектно-признаковая матрица, вектор-строки которой соответствуют параметрам, а вектор-столбцы – предложениям. Полученная матрица данных имеет размерность Nxn, где N=400, а n=51.

На основе объектно-признаковой матрицы была сформирована корреляционная матрица связей параметров, элементами которой являются выборочные коэффициенты корреляции

$R=\{r_{jk}\}_{j,k=1,n}$, где n =51 (2)

По этой матрице были определены средняя внутригрупповая корреляция (формула 4) и средняя внегрупповая корреляция (формула 5) каждого параметра:

$$\bar{r}^{m} = \frac{\left(\sum_{i=1}^{m} |r_{ij}| - 1\right)}{m-1}, \qquad (4)$$

где m=5, r_{ij}-коэффициент корреляции в матрице;

$$\bar{r}^{n-m} = \frac{\left(\sum_{i=1}^{n} |r_{ij}| - 1\right) - \left(\sum_{i=1}^{m} |r_{ij}| - 1\right)}{n-m-1}, \qquad (5)$$

где n=51, m=5, r_{ij}-коэффициент корреляции в матрице.

Затем были вычислены критерии эффективности каждого параметра:

$$E_j = \frac{\bar{r}_j^{n-m}}{\bar{r}_j^{m}}. \qquad (6)$$

Результаты приведены в табл. 2.

Таблица 2. Критерий эффективности.

Параметр	$\bar{r}^{n-m}$	$\bar{r}^{m}$	E_i
X2	0,372	0,824	0,452
X4	0,220	0,657	0,335
X21	0,378	0,804	0,470
X31	0,351	0,806	0,436
X32	0,256	0,672	0,382

Как видно из таблицы, нет ни одного параметра, для которого значение критерия эффективности было бы больше единицы, более того, $0.33<E_j<0.47$, это свидетельствует о тесной связи между всеми пятью параметрами. Поскольку нет поводов для того, чтобы убрать из полученного набора какой-либо параметр, в информативный набор параметров были включены все полученные на первом этапе параметры. Итак, рабочий словарь системы включает пять диагностирующих параметров, представленных в табл. 3.

Таблица 3. Информативные параметры

Параметр	Наименование параметра
X01	Число слов в простом самостоятельном предложении
X11	Число элементарных предложений без номинативного подлежащего
X22	Число именных форм глагола

X33	Число групп однородных членов
X34	Число членов однородных групп

В настоящей работе стиль текстов, представляющих априорные классы и атрибутируемые объекты, описан в многомерном признаковом пространстве пятью информативными параметрами, характеризующими тексты в синтаксическом аспекте. Таким образом, завершен первый этап работы по проверке атрибуционной гипотезы. Следующий этап, заключающийся в процедуре распознавания, будет осуществлен в ходе дальнейшего исследования.

[1] Берков П.Н. Об установлении авторства анонимных и псевдонимных произведений XVIII века. // Русская литература. 1958. №2, с. 43-61.
[2] 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.
[3] Марков А.А. Об одном применении статистического метода. // Известия Имп.Акад.наук, серия VI, Т.X, N4, 1916, с.239.
[4] Виноградов В.В. Проблема авторства и теория стилей. М., 1961, с. 85.
[5] Там же, с. 73
[6] Гришунин А.Л. Опыт обследования употребительности языковых дублетов в целях атрибуции // Вопросы текстологии. Вып. 2 / Под ред. В.С. Нечаевой. М., 1960.
[7] Вашак П. Длина слова и длина предложения в текстах одного автора // Вопросы статистической стилистики / Под. Ред. Б.Н. Головина. Киев, 1974.
[8] Якубайтис Т.А., Скляревич А.Н. Корреляционная характеристика частей речи в связных текстах. Рига, 1980.
[9] Севбо И.П. Графическое представление синтаксических структур и стилистическая диагностика. Киев, 1981.
[10] Мартыненко Г.Я. Многомерный синтаксический анализ художественной прозы // Структурная и прикладная лингвистика. Л.: Изд-во ЛГУ, 1983. Вып. 2, с. 58-72.
[11] Марусенко М.А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л.: Изд-во ЛГУ, 1990.
[12] Хетсо Г. Проблема авторства в романе «Тихий Дон» // Scando-slavica. Т. 24. 1978.
[13] Хетсо Г. Принадлежность Достоевскому: К вопросу об атрибуции Ф.М. Достоевскому анонимных статей в журналах «Время» и «Эпоха». 1986.
[14] Милов Л.В. и др. От Нестора до Фонвизина. Новые методы определения авторства. М., 1994.
[15] http://www.rusf.ru/cgi-bin/fr.cgi
[16] http://www.textology.ru
[17] Labbé C., Labbé D. Inter-textual distance and authorship attribution Corneille and Molière. // Journal of Quantitative Lingustics. 2001. Vol. 8. №3. P. 213-231.
[18] Labbé D. Corneille dans l'ombre de Molière. Histoire d'une recherche. Paris – Bruxelles: Les Impression nouvelles, 2003. C. 71-102.
[19] Louys P. L'auteur d'Amphitryon // le Temps, 16 octobre 1919.
[20] Louys P. Molière est un chef-d'œuvre de Corneille // Comedia, 7 novembre 1919 .
[21] Poulaille H. Corneille sous le masque de Molière. Paris, B. Grasset, 1957.
[22] Vergnaud F. Appendice II // Wouters H., Christine de Ville de Goyet. Molière ou l'auteur imaginaire? Edition Complèxe, 1990.
[23] Wouters H., Christine de Ville de Goyet. Molière ou l'auteur imaginaire? Edition Complèxe, 1990.
[24] Forestier G. Le dossier «Corneille-Molière». http://www.crht.org/?Dossiers/Le+dossier+Corneille-Moli%E8re
[25] Labbé D. Corneille et Molière. // Séminaire du Groupe Langues Information Représentations (13 janvier 2004). P. 2.
[26] Марусенко М.А., 1990.
[27] Большая советская энциклопедия, 1969-1978. Т.16, с.154.

[28] X1 - число слов в простом самостоятельном предложении; X2 - число элементарных предложений; X3 - число главных предложений; X4 - число сочиненных предложений; X5 - число сочиненных предложений без спрягаемой формы глагола; X6- число подчиненных предложений; X7- число подчиненных предложений 1-й степени; X8 - число подчиненных предложений 2-й степени ; X9 - число подчиненных предложений 3-й степени; X10 - число подчиненных предложений 4-й и высших степеней; X11 - число элементарных предложений без номинативного подлежащего; X12 - число подчиненных предложений без спрягаемой формы глагола; X13 - число вставных предложений; X14 - число охватывающих предложений; X15 - число слов 1-й группы (знаменательных); X16 - число слов 2-й группы (служебных); X17 - число имен существительных; X18 - число имен прилагательных; X19 - число местоимений; X20 - число имен числительных; X21 - число спрягаемых форм глагола; X22 - число именных форм глагола; X23 - число наречий; X24- число предлогов; X25- число союзов; X26- число подчинительных союзов; X27- число сочинительных союзов; X28 - число предикативов; X29 - число прямых дополнений; X30 - число косвенных дополнений; X31 - число подлежащих; X32 - число местоимений- подлежащих; X33 - число групп однородных членов; X34 - число членов однородных групп; X35 - число однородных сказуемых; X36 - число однородных групп дополнений; X37 - число причастных оборотов; X38 - число членов причастных оборотов; X39 - число распространенных причастных определений; X40 - число членов распространенных причастных определений; X41 - число согласованных определений; X42 - число причастий - согласованных определений; X43 - число несогласованных определений; X44 - число существительных - несогласованных определений; X45 - число обособленных членов; X46 - число членов в группах обособленных членов; X47 - число существительных без группы; X48 - число групп имен существительных; X49 - число членов групп имен существительных; X50 - число знаменательных слов в группах имен существительных; X51 - число служебных слов в группах имен существительных

[29] Бонгард М.М. Проблемы узнавания. М., 1967.