Родионова Е.С. Отбор информативных параметров при атрибуции стихотворных пьес Мольера // Материалы XXXVI Международной филологической конференции (12 – 17 марта 2007 г.). – СПб : Филол. фак. С.-Петерб. гос. ун-та, 2007. – Вып. 10 : Прикладная и математическая лингвистика / под ред. Т. Г. Скребцовой. С. 67–74

Е. С. Родионова
Санкт-Петербургский государственный университет

ОТБОР ИНФОРМАТИВНЫХ ПАРАМЕТРОВ ПРИ АТРИБУЦИИ СТИХОТВОРНЫХ ПЬЕС МОЛЬЕРА

Вопрос об авторстве комедий Мольера был поднят в начале XX века французским поэтом Пьером Луи[1], 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.[2,3] По другой версии, в поддержку которой приводятся различные литературные, биографические и языковые доказательства, Корнель является автором всех произведений Мольера.[4,5] Помимо Пьера Корнеля в числе возможных авторов пьес Мольера упоминаются также драматург Филипп Кино и поэт Шапель.[6] Наконец, существует общепринятая, официальная точка зрения, согласно которой именно Мольер и есть автор своих произведений.[7]

Большинство исследователей творчества Мольера при решении вопроса об атрибуции его произведений рассматривают 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 случае 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[8], проверка атрибуционной гипотезы выполняется средствами теории распознавания образов и предусматривает реализацию двух независимых этапов: 1- отбор информативных параметров, 2- процедура 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»[9]. В нашем исследовании 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.

При формировании алфавита классов необходимо установить объем каждого априорного класса. Атрибутируемые объекты и априорные классы представляют собой стихотворные пьесы 17 века, характеризующиеся специфическими пунктуационными знаками и синтаксисом, для их анализа нами были использованы специальные лингвистические процедуры по разграничению речевого материала. Применение одного лишь строгого формально-пунктуационного метода структуризации текста показало в данном случае недостаточную лингвостилистическую надежность. При формально-пунктуационном подходе предложение определяется как «последовательность символьных цепочек и пунктуационных знаков между ними от одного конечного знака до другого, где «символьная цепочка» - это текстовое графическое изображение словоупотребления, а «конечный знак» представляет собой составной пунктуационно-пространственный отрезок текста, позволяющий формально распознавать в строке ситуацию «конец

предложения»»[10]. В нашем случае в роли «конечных знаков» могут выступать такие пунктуационные знаки, как «.», «;», «!» и «?». Знаки «.» и «;» служат символами конца предложения вне зависимости от их правого окружения: они позволяют формально разграничивать предложения вне зависимости от их месторасположения в строке (середина или конец) и от того, с какой буквы (прописной или строчной) начинается следующий фрагмент текста. Знаки «!» и «?» безусловно фиксируют конец предложения тогда, когда они оказываются в конце строки. Все те случаи, когда они расположены в середине строки, определяются как неоднозначные, в анализируемых текстах за знаками «!» и «?» в середине строки следуют слова, начинающиеся со строчных букв. Для установления границ предложения в этом случае требуется не формальный, а смысловой подход. Мы не считаем знаки «!» и «?» символами окончания предложений в том случае, если они выделяют одно - два междометия или эмоциональный оборот речи.

Любому стилистическому исследованию художественной прозы обычно предшествует процедура разбиения речевого материала на авторскую речь и чужую речь, при которой именно авторская речь составляет дальнейший предмет изучения. Однако в нашем случае именно авторская речь, представленная в пьесах перечислением персонажей пьесы, описанием декораций, указаниям к действиям актеров и именами героев соответствующих реплик, исключается из дальнейшего изучения. Чужая речь в пьесах оформляется двумя способами:

1- в виде отдельных реплик участников разговора после указания героя-автора соответствующей реплики: CELIE Ah ! n'espérez jamais que mon coeur y consente. (Molière. Sganarelle. I, 1);

2- она может входить в состав реплики героя и тогда заключается в кавычки: Puisqu'elle a sur mon coeur un pouvoir absolu, Il lui suffit de dire : "Ainsi je l'ai voulu". (Corneille. La Galerie du Palais. VIII, 2).

В текст, подлежащий дальнейшему изучению, мы включаем текст, составляющий реплики героев, то есть чужую речь, оформленную первым способом, и не включаем предложения с чужой речью, оформленных способом вторым.

Таким образом, 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

При формировании параметрического пространства в задачах распознавания образов очень важно сделать правильный выбор исходных параметров. К настоящему времени количество параметров, предлагаемых разными исследователями для параметризации стилей, доходит до нескольких сотен. В истории диагностики стилевых различий можно

выделить тенденцию перехода от параметров лексического уровня, которые можно характеризовать как недостаточные и ограниченные[11], к параметрам, отображающим синтаксическую структуру языка. Если лексический состав текста теснейшим образом связан с темой, и при имитации текста лексического сходства добиться несложно, то синтаксическая структура текста носит более скрытый характер, и имитация латентных синтаксических характеристик практически невозможна.

Исходя из этих соображений, был составлен априорный словарь из 51 параметра[12], релевантных для описания французского языка, который «представляет собой в значительной степени стандартизованный набор, полученный путем унификации и стандартизации известных средств квантитативного описания стилей, предложенных разными авторами»[13].

На языке параметров из априорного словаря параметров были описаны априорные классы, для чего были сделаны прикидочные случайные выборки объемом по 100 предложений. В результате определения значений 51 параметра для априорных классов были сформированы две объектно-признаковые матрицы данных, и были вычислены статистические характеристики: среднее арифметическое ($\bar{X}_i$) и стандартное отклонение (σ_i), для каждого класса.

При формировании набора информативных параметров мы применили схему Бонграда[14], 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:

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

По формуле 1 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 связная объектно-признаковая матрица, вектор-строки которой соответствуют параметрам, а вектор-столбцы – предложениям. Полученная матрица данных имеет размерность Nxn, где N=200, а n=51.

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

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

Выборочные коэффициенты корреляции представляют собой косинус угла α_{jk} в N-мерном пространстве между векторами x^j и x^k:

$\rho_{jk}=\cos\alpha_{jk}$ (3)

Коэффициенты корреляции принимают значения, которые лежат в интервале [-1;1].

Размерность матрицы 51x51, и она обладает следующими свойствами:

1- $\rho_{jk}=\rho_{kj}, j,k=\overline{1,n}$

2- $\rho_j=1, j=k$

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

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

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

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

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

Затем были вычислены критерии эффективности каждого параметра:

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

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
X02	Число элементарных предложений
X04	Число сочиненных предложений
X21	Число спрягаемых форм глагола
X31	Число подлежащих
X32	Число местоимений-подлежащих

На языке информативных параметров были описаны априорные классы и атрибутируемые объекты, и, таким образом, был завершен первый этап работы по проверке атрибуционной гипотезы, заключающийся в определении информативных параметров и подготовке к процедуре распознавания. Проверка атрибуционной гипотезы на этапе процедуры распознавания будет осуществлена в ходе дальнейшего исследования.

[1] Louys P. L'auteur d'Amphitryon // le Temps, 16 octobre 1919.
[2] Louys P. Molière est un chef-d'œuvre de Corneille // Comedia, 7 novembre 1919 .
[3] Labbé D. Corneille dans l'ombre de Molière. Histoire d'une recherche. Paris – Bruxelles, Les Impression nouvelles, 2003.
[4] Poulaille, Henry. Corneille sous le masque de Molière. Paris, B. Grasset, 1957.
[5] Vergnaud F. Appendice II // Wouters H., Christine de Ville de Goyet. Molière ou l'auteur imaginaire? Edition Complèxe, 1990.
[6] Wouters H., Christine de Ville de Goyet. Molière ou l'auteur imaginaire? Edition Complèxe, 1990.
[7] Forestier G. Le dossier «Corneille-Molière». http://www.crht.org/?Dossiers/Le+dossier+Corneille-Moli%E8re
[8] Марусенко М.А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л.: изд-во ЛГУ, 1990.
[9] Большая советская энциклопедия, 1969-1978. Т.16, с.154.
[10] Гринбаум О.Н. Компьютерные аспекты стилеметрии / /Прикладное языкознание. Отв. Ред. А.С. Герд. Спб, 1996. С. 456.
[11] 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.
[12] 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 слов 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
[13] Марусенко М.А., Бессонов Б.А., Богданова Л.М. и др. В поисках потерянного автора: Этюды атрибуции. СПб.: Филологический ф-т СПбГУ, 2001. С. 9.
[14] Бонгард М.М. Проблемы узнавания. М., 1967.