В. П. Чепига

К ПРОБЛЕМЕ АТРИБУЦИИ РОМАНОВ, НАПИСАННЫХ ПОД ПСЕВДОНИМОМ ЭМИЛЬ АЖАР: ЛЕКСИЧЕСКОЕ БОГАТСТВО СЛОВАРЯ, МЕТОД «МЕЖТЕКСТОВОГО РАССТОЯНИЯ», МЕТОД РАСПОЗНАВАНИЯ ОБРАЗОВ

Работа представлена кафедрой романской филологии Санкт-Петербургского государственного университета. Научный руководитель – доктор филологических наук, профессор М. А. Марусенко

В статье рассматриваются лексические и синтаксические методы атрибуции анонимных и псевдонимных произведений применительно к романам, написанным под псевдонимом Эмиль Ажар. Приводятся основные выводы к указанной проблеме, полученные в результате комплексного исследования, базой которого является атрибуция анонимных и псевдонимных произведений методом теории распознавания образов.

The article touches upon some lexical and syntactic methods of attribution of anonymous and pseudonymous texts relating to the novels that were written under the name of Emile Ajar. The author presents the conclusions gained from the complex research on the basis of attribution of anonymous and pseudonymous works by means of the theory of pattern recognition.

Изучение индивидуальных авторских стилей является комплексной филологической задачей, результаты разрешения которой могут иметь как теоретическое, так и прикладное значения и применяться в различных областях знаний; в текстологических, стилистических исследованиях, в том числе при установлении авторства псевдонимных и анонимных произведений.

Ромен Гари (1914–1980), один из крупных французских писателей XX в., привлекает внимание широкого круга читателей и исследователей прежде всего своей экспериментальной деятельностью в области художественной литературы.

Наибольший интерес вызывает проблема авторства произведений, написанных им под псевдонимом Эмиль Ажар в 1974–1980 гг. Стили двух авторов настолько отличны друг от друга, что при жизни писателя у его современников и критиков не возникало сомнений в том, что Р. Гари и Э. Ажар являются разными людьми. После выхода в свет романа П. Павловича «Человек, кото-

рому верили»[1], а также посмертной публикации Р. Гари «Жизнь и смерть Эмиля Ажара»[2] литературный мир потряс тот факт, что Р. Гари и Э. Ажар могут быть одним и тем же человеком. К тому же Ромен Гари оказался единственным писателем, дважды получившим Гонкуровскую премию.

В настоящей статье приведены некоторые результаты первого отечественного исследования сравнительного анализа стилей и установления авторства псевдонимных произведений Э. Ажара, а именно: «Голубчик», «Жизнь впереди» и «Страхи царя Соломона», методом распознавания образов и некоторыми другими методами атрибуции.

Рассмотрим автоматизированный метод атрибуции произведений, основанный на лексическом анализе текста, а именно метод, легший в основу программы «Lexico3», которая была разработана в 2003 г. группой сотрудников лаборатории SYLED – CLA2T (Сорбонна, университет Париж-3, Франция). Функциональный диапазон программы включает в себя анализ длины предложения, частотности, лексического богатства словаря, т. е. основных лексикометрических текстовых составляющих. Безусловно, применительно к нашей проблематике индекс разнообразия лексики становится показателем особой значимости, поскольку произведения Э. Ажара отличаются стилистическим своеобразием, но при этом характеризуются относительной «бедностью» словаря. Наоборот, в романах Р. Гари прослеживается тенденция к обновлению слов.

После отбора текстов и проведения лексического анализа были получены следующие результаты (см. рис. 1):

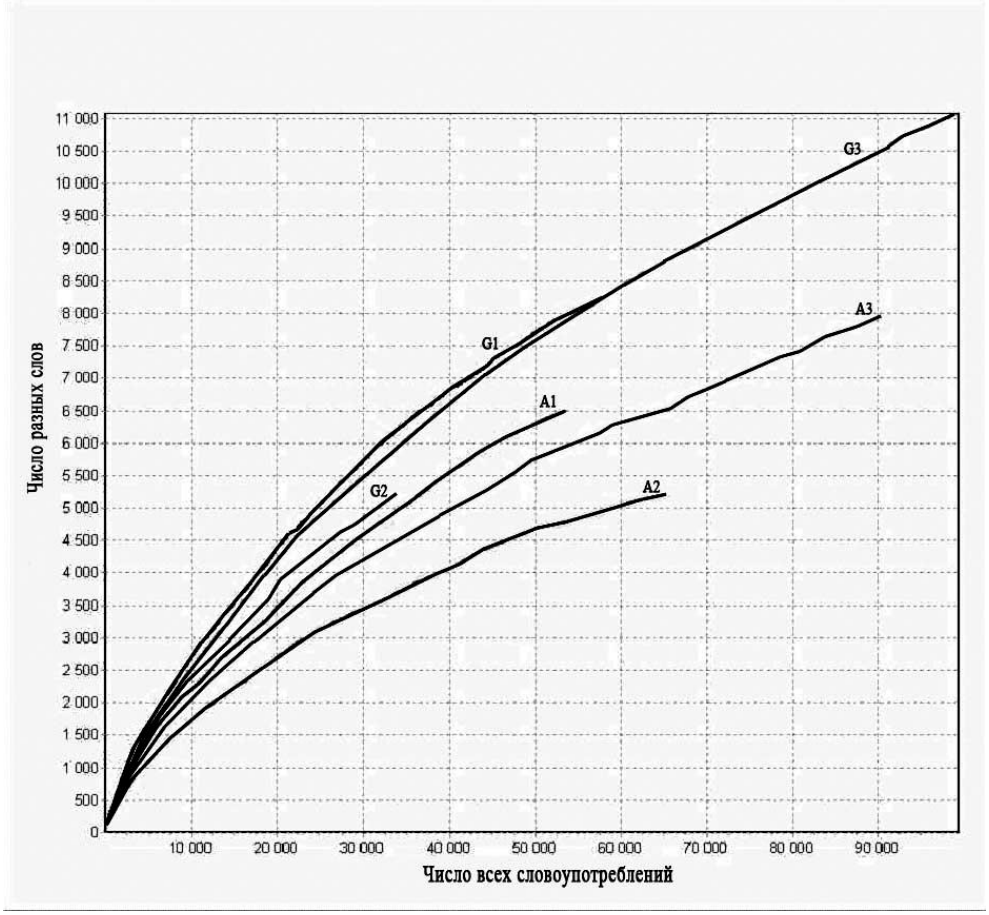

Рис. 1. Лексическое богатство словаря Р. Гари и Э. Ажара

Характерным отличительным признаком авторского языка является индекс разнообразия лексики, вычисленный для Р. Гари (G1 – «Дальше ваш билет недействителен», 0,204, G3 – «Воздушные змеи», 0,183 и G2 – «Свет женщины», 0,156) и Э. Ажара (A1 – «Голубчик», 0,137, A3 – «Страхи царя Соломона», 0,133 и A2 – «Жизнь впереди», 0,117). Индекс разнообразия лексики для исследованных текстов лежит в интервале [0,177–0,204]. Чем выше значение индекса, тем больше разнообразие слов, употребляемых авторами. Следовательно, «Дальше ваш билет недействителен» (Р. Гари, верхняя кривая) в этом смысле наиболее «богатый» текст из всех шести, за ним следуют «Воздушные змеи» (Р. Гари) и «Свет женщины» (Р. Гари), а «Голубчик», «Страхи царя Соломона» и «Жизнь впереди» (Э. Ажар), имея близкие значения, занимают соответственно четвертое, пятое и шестое места, существенно отставая от предыдущих.

Действительно, имеет место разделение класса из шести романов Р. Гари на два подкласса, имеющих близние значения, – Ω (Гари) и Ω (Ажар). Но в данном случае вопрос об атрибуции псевдонимных романов не может быть решен, поскольку богатство словарного запаса писателя не является непременным условием высокого художественного уровня автора. В данном случае мы имеем дело с особенностью стилистической манеры писателя. Как отмечает У. Эко, «…возможно, что для некоторых писателей увеличение словаря приведет, как это ни странно, к обеднению стиля»[3].

Еще одним из последних исследований, основанном на автоматической обработке текста на лексическом уровне, является работа французского специалиста по анализу речи Д. Лаббе, который предложил новый математический метод атрибуции, основанный на анализе лексического состава текстов и вычислении их меры близости или удаленности друг от друга[4]. Метод «межтекстового расстояния» заключается в следующем.

Даны два текста *A* и *B*. Абсолютное расстояние между двумя текстами будет равно разности между объединением этих двух текстов (A∪B) и их пересечением (A∩B).

Д. Лаббе выводит формулу абсолютного межтекстового расстояния:

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

Исследователь определяет расстояние между двумя текстами как сумму разностей частот всех вокабул из наименьшего текста и из всех возможных выборок, равных наименьшему тексту, которые можно извлечь из большего текста, т. е., некоторым образом, происходит проецирование на маленький текст уменьшенной копии большего текста, что позволяет сравнить их напрямую.

Относительное межтекстовое расстояние вычисляется по формуле:

$$Drel_{(a,b)} = \frac{\sum_{V \in (A,B)} \left| F_{ia} - E_{ia(u)} \right|}{\sum_{V \in A} F_{ia} + \sum_{V \in B} E_{ia(u)}} .$$

Мера расстояния между текстами *Drel* принимает значения от нуля до единицы. Д. Лаббе трактует возможные значения следующим образом[5]:

Drel = 0: рассматриваемые тексты содержат одинаковые вокабулы, встречающиеся с одинаковой частотой;

Drel = 1: рассматриваемые тексты не имеют ничего общего;

Drel = 0,5: половина вокабул у двух текстов общая;

Drel ≤ 0,2: тексты принадлежат одному автору;

0,2 < *Drel* < 0,25: либо тексты написаны одним автором, либо имело место тесное литературное сотрудничество, либо тексты написаны двумя разными людьми;

Drel > 0,25: либо один автор написал тексты на разные темы, либо два автора, живущие в одно время, посвятили свои ра-

боты одной теме. В этом случае Д. Лаббе считает необходимым прибегать к другим атрибуционным методам;

Drel > 0,4: либо у текстов разные авторы, либо один автор написал тексты, относящиеся к разным жанрам.

Д. Лаббе применил данный метод при исследовании лексического состава театральных пьес Мольера и П. Корнеля. (П. Корнелю при этом атрибутируется ряд пьес Мольера.) В качестве иллюстрации работы метода, Д. Лаббе был также исследован лексический состав четырех произведений Э. Ажара и четырех романов Р. Гари[6]. Им были получены следующие результаты: «межтекстовое расстояние» между романами «Жизнь впереди» и «Страхи царя Соломона» (Э. Ажар) и романом «Бремя души» (Р. Гари), равное 0,327 и 0,360 соответственно, является значимым, в связи с чем указанные романы попадают в интервал [0,25–0,4] и близки верхнему пределу (*Drel* > 0,4): т. е. либо у текстов разные авторы, либо один автор написал тексты, относящиеся к разным жанрам. Среднее «межтекстовое расстояние» для двух авторов равно 0,285 (*Drel* > 0,25), а значит, необходимо прибегнуть к другим атрибуционным методикам. При этом среднее «межтекстовое расстояние» внутри произведений самого Э. Ажара равно 0,237, на основании чего Д. Лаббе делает вывод о том, что все четыре произведения Э. Ажара являются стилистически выдержанными, близкими друг другу, объединенными одним автором (0,2 < *Drel* < 0,25).

Предложенный метод имеет тот же недостаток, что и методы, связанные только с анализом лексики, не принимающие во внимание иные характеристики текста.

На наш взгляд, эффективный метод анализа стиля в целях определения авторства должен обладать рядом характеристик, а именно: исследовать текст в его совокупности; подвергать анализу разные языковые уровни; центральной процедурой метода должна стать многомерная классификация объектов. Всем требованиям, применительно к проблеме атрибуции анонимных и псевдонимных произведений, отвечает такой готовый математический аппарат, как теория распознавания образов.

Опыт реализации конкретных задач по атрибуции литературных произведений показал высокую эффективность указанного метода[7].

По результатам биографических данных и сравнительного исследования рукописей Р. Гари и Э. Ажара сформированы гипотезы по атрибуции псевдонимных произведений. При этом нулевая гипотеза сформулирована следующим образом: (H_0) – все тексты романов Э. Ажара полностью принадлежат Р. Гари. Альтернативные гипотезы сформулированы таким образом: (H_a^1) – тексты романов Э. Ажара являются произведением П. Павловича и (H_a^2) – тексты романов Э. Ажара являются совместным произведением Р. Гари и П. Павловича с определенной долей участия каждого из них. В результате априорный алфавит классов распознающей системы включил два класса: Ω (Гари), Ω (Павлович)[8]. Для определения координат эталонов априорных классов сформирован корпус текстов, отвечающий требованиям синхронии и жанрово-стилевой однородности.

Из работ Р. Гари, П. Павловича и Э. Ажара были отобраны произведения, представляющие собой романы, написанные от первого лица. Проведена процедура разбиения речевого материала на авторскую и прямую речь, при которой именно авторская речь явилась дальнейшим предметом изучения. Параметрическое пространство составили 54 параметра из априорного словаря параметров, и следующий этап исследования заключался в отборе из полученного априорного словаря небольшого числа информативных параметров. Для определения информативного набора параметров был проведен эксперимент по описанию априорных классов на языке параметров из априорного словаря параметров, для чего были сделаны прикидочные

случайные выборки объемом в 200 предложений для каждого априорного класса. Результаты эксперимента были представлены в виде объектно-признаковых матриц данных размерностью n x N=200 x 54, где n – число параметров, а N – число объектов. Общее число элементов матриц данных составило по 10 800.

Ввиду большого объема выборки было использовано обратное построение матриц данных, при котором набор значений параметров представляется соответствующей строкой матрицы, а значения всех параметров на объектах (предложениях) – соответствующим столбцом. Затем были вычислены статистические характеристики анализируемых объектов: среднее арифметическое ($\bar{x}_i$) и стандартное отклонение (σ_i) для каждого класса. На первом этапе произошло разбиение априорного набора информативных параметров на два подмножества параметров, релевантных и нерелевантных для различения априорных классов. Для этого была произведена автоматическая классификация параметров в одномерном пространстве. Релевантность параметров для различения априорных классов была определена по t-критерию Стьюдента, пороговое значение которого при уровне значимости α = 0,05 равно 1,96. Процесс классификации заключался в сравнении значений критерия с пороговым значением. Если наблюдаемое значение критерия больше порогового, то параметр относился к числу информативных, в противном случае он исключался из дальнейшего рассмотрения.

Статистически значимой оказалась разность средних для шестнадцати параметров (наблюдаемые значения t-критерия больше критического). Все остальные параметры оказались нерелевантными для различения априорных классов.

В дальнейшем подмножество диагностических параметров формировалось на основе группы из этих параметров, релевантных для разделения известных классов.

На основе объектно-признаковой матрицы была сформирована корреляционная матрица связей параметров, элементами которой являются выборочные коэффициенты корреляции. Полученная матрица данных имеет размерность 54 x 54. Содержательный критерий информативности набора параметров заключается в слабой корреляции информативных параметров между собой и сильной их корреляции с остальными параметрами, не вошедшими в эту группу. На основе корреляционной матрицы были определены средняя внутригрупповая корреляция и средняя внегрупповая корреляция каждого параметра. Затем были вычислены критерии эффективности каждого параметра.

Подмножество из шестнадцати параметров разбилось на подмножество из трех параметров (число союзов, число предлогов и число прямых дополнений), значения, для эффективности которых лежат в интервале [0,509–0,951], и на подмножество из тринадцати параметров, значения, для эффективности которых лежат в интервале [0,318–0,467].

Таким образом, из подмножества шестнадцати параметров удалось выделить подмножество из трех параметров, удовлетворяющих требованиям сильной корреляции с другими параметрами совокупности и способности к разделению объектов на классы. В дальнейшем в эксперименте по классификации текстов каждый объект (роман) характеризовался набором из трех диагностических (информативных) параметров, а классификация объектов производилась в трехмерном пространстве, осями которого явились данные параметры. При определении координат распознаваемых объектов и эталонов априорных классов были применены методы выборочного обследования текста.

Необходимые объемы выборок для определения координат атрибутируемых объектов и эталонов классов определялись по наибольшим значениям парамет-

ров. В настоящем эксперименте они соответствуют значениям параметра X24 – число союзов. Были сформированы матрицы данных атрибутируемых объектов и априорных классов.

Применяемый алгоритм распознавания включил в себя детерминированную и вероятностную атрибуции. Детерминированный алгоритм выявил существенные статистические различия между атрибутируемыми объектами и обоими априорными классами, в результате чего три романа оказались не отнесенными ни к одному априорному классу.

Реализация вероятностного алгоритма распознавания предусматривает преобразование исходной матрицы данных в матрицу евклидовых расстояний между объектами, подлежащими распознаванию, и эталонами априорных классов.

При наличии двух априорных классов решающее правило сформулировано следующим образом:

$$X_i \in \Omega_2 \text{ , } P(X_i \in \Omega_2) > \sum_{n=1}^{2} P(X_i \in \Omega_2).$$

Таким образом,

$$X_i \in \Omega_j \text{ , } \quad P(X_i \in \Omega_j) > 0{,}5 \text{ ,}$$

т. е. при данном решающем правиле пороговое значение вероятности принадлежности объекта к одному из априорных классов должно превышать 0,5. По результатам вероятностной атрибуции классу Ω (Гари) смогли быть атрибутированы все три объекта: «Голубчик», «Жизнь впереди» и «Страхи царя Соломона».

Заключительным этапом процедуры распознавания становится оценка качества полученной классификации, могущая иногда повлечь за собой корректировку полученных классов. Оценка качества классификации основана на выявлении закономерностей для каждого класса, которые могут быть определены как подмножества из заданного множества объектов, значения параметров которых одинаковы для большинства объектов анализируемого класса и отличаются от значений параметров других классов. Правильность и правомочность полученных результатов после проведения вероятностной атрибуции были проверены, исходя из оценки однородности априорного класса до и после добавления к нему атрибутируемых объектов. Критерием однородности послужило вычисление среднего квадрата расстояния между объектами $\bar{d}^2(\Omega_N)$. Была построена матрица данных, объединившая исходные данные трех объектов, составляющих класс Ω (Гари), и всех трех атрибутируемых этому классу объектов. В результате эксперимента, поскольку среднеквадратичный разброс $\bar{d}^2$ $(\Omega_{Гари}) > \bar{d}^2$ $(\Omega_{Гари,\ Ажар})$, сделан вывод о том, что полученная классификация объектов улучшилась и приблизилась к естественной, что подтвердило результаты работы вероятностного алгоритма и принадлежность априорному классу трех атрибутируемых объектов.

В результате атрибуции удалось установить, что между объектами априорного класса Ω (Гари) и атрибутируемыми романами «Голубчик», «Жизнь впереди» и «Страхи царя Соломона» присутствуют статистически значимые различия; путем вероятностной атрибуции установлена принадлежность указанных романов одному писателю (Ромену Гари) [9].

Представляется возможным отойти от рассмотрения Р. Гари и Э. Ажара как одного человека, творившего под двумя разными псевдонимами, и разграничить такие термины, как «писатель» и «автор». Данное разграничение предложено французским литературоведом Ф. Вернье. Писателем она предлагает признать индивидуума, писательство для которого является профессиональной деятельностью. У такого человека есть личная, творческая, социальная истории, к которым мы можем обратиться в литературоведческом исследовании. Автора же можно рассматривать как личность, которая раскрывается путем исследования конкретного литературного наследия, сти-

листики, интенциональности его произведений[10]. Если принять данную точку зрения, то мы можем рассматривать Р. Гари и Э. Ажара как двух разных авторов, объединенных одним писателем – Роменом Гари.

Тот факт, что все результаты получены путем вероятностной атрибуции, может быть интерпретирован следующим образом: Р. Гари, скрывавший свое авторство, намеренно старался изменить стиль псевдонимных произведений. Именно этим может объясняться отсутствие результатов детерминированной атрибуции. Но поскольку писатель не может значительно изменять характеристики латентных синтаксических структур при создании больших объемов текста, то вероятностная атрибуция смогла однозначно указать на истинного автора.

Результаты исследования служат еще одним показателем того, что характеристики индивидуального авторского стиля носят объективный характер и не зависят от воли пишущего.

ПРИМЕЧАНИЯ

[1] *Pavlowitch P.* L'Homme que l'on croyait. P.: Fayard, 1981.

[2] *Gary R.* Vie et mort d'Йmile Ajar. P.: Gallimard, 1981.

[3] *Eco U.* En quoi l'usage de l'ordinateur complexifie la genиse d'un texte? // L'Йcriture et le souci de la langue. Louvain-la-Neuve: Academia-Bruylant, 2007. P. 174.

[4] *Labbй C., Labbй D.* Inter-textual distance and authorship attribution Corneille and Moliиre // Journal of Quantitative Linguistics. 2001. Vol. 8. N 3. P. 213–231.

[5] *Labbй D.* Corneille dans l'ombre de Moliиre. Histoire d'une recherche. Paris; Bruxelles: Les Impressions nouvelles, 2003. P. 14.

[6] *Labbй C., Labbй D.* R. Gary et Й. Ajar // Corneille a йcrit 16 piиces reprйsentйes sous le nom de Moliиre: Version prйliminaire. Grenoble, 2007. P. 146–153.

[7] *Марусенко М. А.* Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л.: Изд-во ЛГУ, 1990.

[8] Поль Павлович является двоюродным племянником Р. Гари. При жизни последнего, П. Павлович выступал в роли Э. Ажара и был «живым воплощением» мистификации Р. Гари. Для нашего исследования необходимо привлечь тексты романов П. Павловича, так как он единственный является вероятным автором произведений, написанных под псевдонимом Э. Ажар.

После издания книги в 1981 г. «Человек, которому верили» П. Павлович уходит в тень и начинает снова публиковаться только в конце 1990-х гг. Романы П. Павловича признаны произведениями своеобразного и талантливого писателя. Таким образом, неясными остаются мотивы этого автора при его преднамеренном отказе в 1974 г. от собственной литературной карьеры.

[9] Атрибуция произведений Э. Ажара методом распознавания образов полностью представлена в статье: *Чепига В. П.* Проблема «Ромен Гари – Эмиль Ажар»: атрибуция романов, опубликованных под псевдонимом Эмиль Ажар // Известия РГПУ им. А. И. Герцена: Аспирантские тетради. 2007. Вып. 19. С. 284–290.

[10] *Vernier F.* L'ange de la thйorie // Paragraphes. Montrйal: Presses universitaires de Montrйal, 2004. P. 120, 122.