Известия Гомельского государственного университета
имени Ф. Скорины, №5 (44), 2007

УДК 004.02:004.9

Метод реализации стилеметрических исследований на основе применения аппарата математической теории распознавания образов

В. Г. Родченко

Введение

Стиль в обобщенном представлении относят к универсальным понятиям. В различных сферах человеческой деятельности он трактуется по-разному. Например, для композитора стиль может определяться совокупностью черт, близостью выразительных художественных способов и средств музыкального отображения окружающей реальной или идеальной действительности, которые представляют единство конкретного творческого направления. В управленческой деятельности понятие стиля руководителя связывается с совокупностью приемов деятельности и поведения.

Проблема исследования авторского стиля относится к числу важнейших филологических задач, решаемых в рамках научной текстологии. Она же является и серьезной междисциплинарной проблемой, связанной не только с циклом гуманитарных дисциплин, но и с достаточно широким кругом естественно-научных дисциплин, к которым относятся теория вероятностей, прикладная статистика, информатика [1].

Представление об индивидуальной творческой манере писателя, об его авторском стиле фактически начало складываться с момента появления самой литературы. Априори считается, что индивидуальный стиль автора обладает характерными признаками, благодаря которым он легко отличается от стилей других авторов, а потому относительно легко распознается [2]. Опыт практического исследований по выделению соответствующих признаков авторского стиля в своей эволюции прошел целый ряд этапов. Первые попытки научного поиска индивидуальных “отпечатков” авторов строились на весьма оптимистичных предположениях о том, что уже на предварительном этапе анализа авторского текста подготовленному эксперту удастся выделить характерные признаки [3, 4]. Затем пришло понимание сложности и неоднозначности проблемы. Академик В.В.Виноградов в 1961 в монографии “Проблема авторства и теория стилей” отмечал, что “*субъективные методы определения автора отжили или, во всяком случае, отживают свой век. Осуществленные с применением субъективно-эстетических или субъективно-психологических, а также субъективно-идеологических приемов удачные, т.е. оказавшиеся правильными, атрибуции – результат интуиции или личной одаренности изыскателей, а не итог правильно найденного пути поисков и открытий в области изучения проблем авторства*” [5].

Повышение уровня объективности стилеметрических исследований напрямую связано с развитием методов, основанных на использовании современного математического аппарата прикладной статистики и компьютерного анализа данных. Появилась возможность путем проведения объемных вычислительных экспериментов анализировать и устанавливать закономерности функционирования разнообразных признаков авторского стиля.

Идентификация конкретного автора фактически предполагает распознавание его индивидуального стиля, а потому весьма логичным выглядит стремление воспользоваться методами и алгоритмами математической теории распознавания образов для объективного исследования и выделения формальных дескрипторов индивидуального авторского стиля.

В данной статье предлагается метод реализации стилеметрических исследований, который базируется на использовании аппарата математической теории распознавания образов. Предусматривается обязательное выполнение отдельной процедуры, связанной с анализом информативности исходных признаков с точки зрения описания индивидуальных стилисти-

ческих особенностей конкретного автора. Такой подход позволяет осуществлять исследования функционирования авторского стиля как в рамках компаративного анализа текстов различных авторов, так и на основе произведений одного писателя. Это дает возможность изучения динамики изменения стилистических приемов в творчестве отдельного автора.

Описание метода реализации стилеметрических исследований

Процесс проведения стилеметрических исследований начинается с определения круга авторов и их текстов, которые будут использованы в качестве исходного материала для реализации дальнейшей процедуры компаративного анализа. Следует учитывать, что в случае изучения динамики изменения стиля одного автора на протяжении определенных интервалов наблюдений круг авторов может ограничиваться и одним представителем. Фактически этот первый этап связан с конкретизацией решаемой задачи и в первую очередь для его реализации привлекаются специалисты в области текстологии. Отметим, что, например, при проведении стилистического анализа Б.Я.Слепак предлагает выделять вариабельность трех порядков [6]. Под вариабельностью первого порядка подразумевается наличие у отдельных авторов признаков, существенно отличающих одного писателя от другого. Вариабельность второго порядка отражает существование различий стиля одного автора в его различных текстах. Вариабельность третьего порядка касается анализа стиля автора в рамках одного произведения.

Второй этап исследований предполагает уже взаимодействие текстологов и специалистов в области математической теории распознавания образов и компьютерного анализа данных. В первую очередь их совместными усилиями определяется требуемый уровень детализации конкретного стилеметрического исследования. После этого формируется конечный список авторских текстов (произведений). На основе отобранных текстов в дальнейшем будет проводиться процедура формального описания классов в многомерном признаковом пространстве. Далее формируется пространство наблюдений и определяется процедура преобразования этого пространства в формализованное пространство признаков, собственно которое и ориентировано на дальнейшее использование математического аппарата. В результате выполнения этого этапа формируется алфавит классов $A=\{A_1,A_2,...,A_k\}$ и априорный словарь признаков $P=\{P_1,P_2,...,P_n\}$.

С точки зрения прикладной статистики любой априорный словарь признаков (АСП) является выборкой из генерального словаря. Это означает, что можно сформировать *приемлемый* для реализации стилеметрических исследований АСП, но нельзя сформировать *оптимальный* словарь [7]. Многочисленные опыты практического построения АСП свидетельствует о том, что в априорный словарь попадут такие признаки, которые не обеспечивают разделение формальных образов авторов в соответствующем признаковом пространстве. Эти признаки наоборот способствуют “размыванию” образов. В таком случае речь идет о *неприемлемом варианте словаря признаков* для идентификации индивидуального авторского стиля.

Третий этап исследований начинается с того, что на основе априорного словаря $P=\{P_1,P_2,...,P_n\}$ реализуется процедура представления каждого отдельного авторского фрагмента текста для каждого класса $A_1,A_2,...,A_k$. При этом формальное описание j-го текстового фрагмента i-го автора задается в виде вектора-столбца $(x_{1j}, \dots , x_{nj}, i)^T$. Объединение всех m_i векторов i-го класса образует матрицу размерности n x m_i:

$$X^i_{n\,x\,m_i} = \begin{pmatrix} x^i_{11} & x^i_{12} & \dots & x^i_{1m_i} \\ x^i_{21} & x^i_{22} & \dots & x^i_{2m_i} \\ \dots & \dots & \dots & \dots \\ x^i_{n1} & x^i_{n2} & \dots & x^i_{nm_i} \end{pmatrix}, \text{ где } i=\overline{1,k}\,;\ j=\overline{1,m_i}\,.$$

Эта матрица представляет собой формальное описание соответствующего i-го класса в многомерном априорном признаковом пространстве.

Результатом выполнения этого этапа исследований будет являться классифицированная обучающая выборка $X_{n\ x\ m} = \bigcup_{i=1}^{k} X^{i}_{n\ x\ m_i}$ (где n – количество признаков априорного словаря, а $m=m_1+m_2+\ldots+m_k$), которая получается при объединении всех матриц $X^{i}_{n\ x\ m_i}$.

Четвертый этап исследований предусматривает сепарирование на три вида признаков из априорного словаря $P=\{P_1,P_2,...,P_n\}$ по степени их информативности с точки зрения отражения ими индивидуальных и устойчивых характеристик авторского стиля.

Если для очередного признака P_i в результате компаративного анализа всех пар выборок из разных классов выполнились соответствующие критерии однородности, то этот признак будет относиться к первому виду. Характерной особенностью таких признаков является то, что их значения фактически подчиняются одному и тому же закону распределения, а значит они, с одной стороны, не несут в себе разделяющей классы функции, а с другой стороны, демонстрируют в данном случае “одинаковость” стилей авторов.

Когда в результате компаративного анализа всех пар выборок из разных классов оказалось, что, с одной стороны, внутри каждого отдельного класса значения признака P_i слабо варьируют около среднего значения, а с другой стороны, для всех без исключения пар выборок из разных классов ни разу не выполнились соответствующие критерии однородности, то признак будет относиться ко второму виду. Отличительной чертой таких признаков является то, что по своей природе они несут разделяющую классы функцию, поскольку внутри отдельного класса варьируют слабо (тем самым обеспечивают компактность образа), а при межклассовом сравнении демонстрируют неоднородность (тем самым обеспечивают разделение образов классов в признаковом пространстве). Отметим, что именно признаки второго вида и обеспечивают формальное описание индивидуальных особенностей стилей исследуемых авторов.

В случае, когда признак P_i не был отнесен ни к первому, ни ко второму виду, его сепарируют к третьему виду. Признаки этого последнего вида характеризуются хаотичностью поведения в первую очередь при межклассовом сравнении и “размывают” формальные образы авторского стиля в соответствующем многомерном признаковом пространстве.

Пятый этап исследования связан с формированием итогового заключения и содержательной интерпретацией полученных результатов.

Описание алгоритма

Алгоритмом метода реализации стилеметрических исследований предусматривается выполнение следующей последовательности шагов:

Шаг 1. Текстологи формируют соответствующий целям исследований алфавит классов $A=\{A_1,A_2,...,A_k\}$. Совместно со специалистами в области компьютерного анализа данных они определяют исходное пространство наблюдений и на его основе строят априорный словарь признаков $P=\{P_1,P_2,...,P_n\}$.

Шаг 2. Каждый класс A_i (где $i=\overline{1,n}$) изначально определяется совокупностью текстовых фрагментов соответствующего автора. В свою очередь, каждый отдельный текстовый фрагмент с формальной точки зрения является объектом и на основе признаков из априорного словаря описывается в многомерном признаковом пространстве в виде вектора-столбца $x=(x_1,x_2,...,x_n)^T$, где x_j – значение j-го признака.

Шаг 3. Путем объединения всех соответствующих векторов из всех классов формируется классифицированная обучающая выборка, которая представляет собой прямоугольную матрицу, состоящую из n строк и m столбцов (где $m=m_1+m_2+\ldots+m_k$, а m_i – количество объектов i-го класса). При этом для каждого класса $A_i \subset A$ (где $i=\overline{1,k}$) формируется матрица $X^{i}_{n\ x\ m_i}$ размерности $n\ x\ m_i$, где m_i – число объектов i-го класса.

Шаг 4. Все признаки из априорного словаря $P=\{P_1,P_2,...,P_n\}$ последовательно анализируются и сепарируются на три вида: $P^{(1)}=\{P_1^{(1)},P_2^{(1)},...,P_{n1}^{(1)}\}$, $P^{(2)}=\{P_1^{(2)},P_2^{(2)},...,P_{n2}^{(2)}\}$, $P^{(3)}=\{P_1^{(3)}, P_2^{(3)},..., P_{n3}^{(3)}\}$, где $P= P^{(1)} \cup P^{(2)} \cup P^{(3)}$ и $n= n_1+n_2+n_3$. При этом непосредственно процедура сепа-

рирования выполняется следующим образом: *очередной признак P_i (где $i=\overline{1,n}$) классифицируется к одному из трех по правилу: 1) если для всех пар классов соответствующий критерий однородности не показал существенного различия между выборками значений этого признака для двух сравниваемых классов, то P_i относится к первому виду; 2) если для всех пар классов соответствующий критерий однородности показал существенное различие между выборками значений этого признака для двух сравниваемых классов, то P_i является признаком второго вида; 3) если для признака P_i не выполнилось ни одно из двух предыдущих условий, то его позиционирует к третьему виду.*

Шаг 5. Проводится содержательная интерпретация полученных на предыдущем шаге результатов. При этом признаки первого вида формально описывают устойчивые закономерности, присущие стилям всех исследуемых текстов, и демонстрируют стилистическую схожесть авторов. Признаки второго вида фактически формально описывают особенности авторского стиля, и именно они "отражают" индивидуальное лицо автора. Оставшиеся признаки третьего вида относятся к фоновым.

Заключение

Применение аппарата математической теории распознавания образов позволяет реализовать новый подход к сложной проблеме анализа стиля автора. Для организации соответствующих исследований авторский текст предлагается разбивать на фрагменты и в дальнейшем каждый текстовый отрывок рассматривать как многомерный объект в специально сформированном признаковом пространстве.

В статье представлен метод реализации стилеметрических исследований, который предусматривает процедуру формирования исходной классифицированной выборки и процедуру сепарирования признаков по степени их информативности на три вида. В итоге выделяются признаки, которые характеризуют индивидуальные особенности стиля того или иного автора.

Использование предложенного метода позволяет значительно сократить влияние субъективного фактора при проведении стилеметрических исследований и предоставляет возможность одновременно параллельно анализировать большое число разнообразных параметров текста, получая в конечном итоге более объективную и развернутую картину закономерностей функционирования авторского стиля.

Abstract. In compliance with the mathematics-based point of view any author's text can be described by means of a large number of various formal characteristics. However, only subset of these characteristics tends to conform to mechanism of functioning of the individual author's style. Experience of textual studies proves the fact that the issue of search and isolation of formal descriptive signs of individual author's style remains to be an edge-cutting issue. To conduct stylometric studies the application of the apparatus of mathematical theory of pattern recognition is offered in the paper.

Литература

1. Мартыненко, Г. Я. Основы стилеметрии / Г. Я. Мартыненко // Ленинград: Издательство Ленинградского университета, 1988.

2. Марусенко, М. А. Атрибуция анонимных и псевдоанонимных литературных произведений методами распознавания образов / М. А. Марусенко // Ленинград: Издательство Ленинградского университета, 1990.

3. Морозов, Н. А. Лингвистические спектры: Средство для отличения плагиатов от истинных произведений того или другого известного автора: Стилеметрический этюд / Н. А. Морозов // Известия отделения русского языка и словесности Императорской Академии Наук, 1915 – Т. XX, – Кн.4. – С. 93–134.

4. Гришунин, А. Л. Опыт обследования употребительности языковых дублетов в целях атрибуции / А. Л. Гришунин // Москва: Издательство АН СССР. – Вопросы текстологии, 1960, Выпуск 2. – С. 146–195.

5. Виноградов, В. В. Проблема авторства и теория стилей / В. В. Виноградов // Москва, 1961.

6. Слепак, Б. Я. Некоторые теоретико-методологические предпосылки качественно-количественной концепции стиля / Б. Я. Слепак // Вопросы сопоставительной и прикладной лингвистики. Ученые записки Тартуского государственного университета, Тарту, 1982. – Выпуск 619.

7. Родченко, В. Г. Об одном методе реализации процедуры обучения при построении системы распознавания образов / В. Г. Родченко // Известия Гомельского государственного университета имени Франциска Скорины, 2006. – № 4(37). – С. 73–76.

Гродненский государственный
университет имени Янки Купалы

Поступило 12.05.07