В. С. Андреев

ФОРМАЛЬНЫЕ МАРКЕРЫ ИЗМЕНЕНИЙ СТИЛЯ Г. ЛОНГФЕЛЛО

В работе рассматриваются лингвистические характеристики стихотворного текста, которые позволяют зафиксировать изменения идиостиля известного американского поэта-романтика Г. Лонгфелло и дифференцировать различные этапы его творчества. При помощи многомерного дискриминантного анализа выявляются интегральные для всего творчества поэта признаки и признаки, обладающие дифференциальной силой.

V. Andreyev

FORMAL MARKERS OF CHANGES IN H. LONGFELLOW'S STYLE

The article studies linguistic characteristics of verse text, which reveal changes in the style of a famous American romantic poet H. Longfellow and distinguish different stages of his creative work. The use of multivariate discriminant analysis makes it possible to single out characteristics that are integral for the whole period of the poet's creative activity, as well as characteristics possessing differential force.

В настоящее время все большее внимание уделяется поиску лингвистических параметров индивидуального стиля с целью автоматической группировки текстов по авторству, гендеру, жанру, времени создания и другим основаниям[1]. В этих исследованиях ставится задача отыскать маркеры стиля, которые бы характеризовали «свой» класс текстов, отличая его от другого или других классов. Поиск таких дифференциальных признаков, как правило, основывается на выявлении различий, в частотности лингвистических параметров в сопоставляемых текстах.

Признаковое пространство исследований по дискриминации идиостиля разных авторов обычно меняется от исследователя к исследователю и охватывает чрезвычайно широкий набор параметров. Наиболее часто в число таких параметров, используемых при дискриминации стилей, входят морфологические (частеречная отнесенность, морфологические формы глаголов и/или других частей речи, частотность служебных слов и др.), лексические (частотность конкретных лексем или классов слов, насыщенность словаря, типы и частотность аффиксов и др.), синтаксиче-

ские (различные характеристики, отражающие длину и структуру предложений, тип и характер словосочетаний и др.) и производные (агрегированные) признаки. Последние представляют собой различные комбинации (комплексы) исходных морфологических, синтаксических и лексических характеристик[2]. Кроме того, достаточно часто используются различные параметры, отражающие длину слов в буквах, специфику пунктуации и др.[3]

Следует отметить, что признаки, оказывающиеся эффективными в одних случаях, часто теряют дифференциальную значимость – в других. Так, Дж. Тамбуратзис и его коллеги в ряде исследований по дифференциации стиля, используя такие параметры, как различные виды отрицания (8 видов), целый ряд лексических параметров (17), признаки, отражающие морфологические формы глаголов и частеречную отнесенность слов в текстах (33), а также рад признаков по пунктуации, длине слов в буквах, длине предложений в словах и т. д. (27), и, применив различные методы анализа, такие как дискриминантный, кластерный, факторный и др., в конечном итоге приходят к выводу о релевантности только небольшой части признаков, из которых в число наиболее эффективных у них попали частеречные параметры и ряд морфологических характеристик (различные формы глаголов). Вместе с тем они не обнаружили сколько-нибудь высокой эффективности служебных слов[4]. С другой стороны, именно этот параметр, отражающий использование служебных слов, выделяется как основной стилистический маркер в целом ряде других работ[5].

Достаточно часто большое внимание при дискриминации стиля уделяется различным лексическим параметрам[6]. В то же время ряд авторов указывает на целесообразность исключения лексики из анализа такого вида[7]. В исследованиях М. А. Марусенко и его коллег приоритет отдается целому комплексу синтаксических агрегированных признаков[8]. С другой стороны, в целом ряде исследований синтаксические параметры были исключены из исследования, а удачные результаты по дискриминации стилей дал учет достаточно упрощенных параметров, таких как средняя длина слов, измеренная в буквах[9].

Список исследований, в которых различные характеристики подходят для дискриминации одних текстов, но оказываются слабыми или вовсе нерелевантными факторами для решении схожих задач на ином материале (и тем более – на другом языке), можно было бы продолжить, однако в любом случае вывод ясен: пока не найдено универсальных лингвистических маркеров дифференциации идиостилей. По-видимому, для каждой конкретной задачи необходимо осуществить целый ряд экспериментов, прежде чем будет получен подходящий набор дискриминирующих параметров.

В нашей работе ставится задача определить, имело ли место изменение стиля у американского поэта-романтика Лонгфелло в течение его жизни, которое можно было бы установить путем анализа лингвистических признаков его текстов. Эта задача, в целом, соответствует проблематике указанных выше работ по атрибуции текстов, однако вместе с тем имеет и одно существенное отличие. Если в большинстве работ по стилеметрии, посвященных определению авторства либо дискриминации текстов, сопоставляются тексты (или группы текстов), принадлежащие различным авторам, то в нашем случае сопоставление должно быть проведено для текстов одного и того же автора, но написанных им в разное время. Это отличие определяет два следствия. Во-первых, нас будут интересовать не только те признаки, которые дифференцируют разные периоды творчества, но и те, которые выступают как интегральные для всей творческой деятельности поэта. Во-вторых, признаки, привлекаемые к исследованию, должны быть релевантными именно для стихотворных текстов и

быть интерпретируемыми в рамках имеющейся парадигмы понятий лингвистического и стиховедческого анализа.

В отличие от работ, посвященных дискриминации текстов прозаических произведений, работ по поиску маркеров для дифференциации стихотворных произведений неизмеримо меньше. Одним из первых, либо вообще первым исследователем, кто систематически использовал многомерные методы анализа для поиска параметров, автоматически различающих классы стихотворных текстов, является В. С. Баевский. Рассмотрим используемые им параметры.

Для разграничения стиля различных периодов творчества Пушкина, Пастернака и Гумилева В. С. Баевский использует такие признаки, как разнообразие метрики, тип размера, количество стихов с пропуском ударения на 1-м и 2-м иктах, наличие разрывов стихотворных строк синтаксическими паузами, количество стихов с неточными и приблизительными рифмами, количество строф по отношению к количеству стихов; количество разных строфических форм по отношению к количеству стихов[10]. В ряде других своих исследований В. С. Баевский привлек несколько иной набор признаков, в которые входили количество размеров, количество текстов с усложненной метрикой, количество пропусков ударения на сильных местах, количество стихов, разорванных синтаксической паузой, количество точных и приблизительных рифм, строфических форм и строф, завершающихся женской или дактилической клаузулой[11].

Таким образом, в разработанной В. С. Баевским схеме отражены метро-ритмические, рифменные, строфические признаки и характеристики стихотворного синтаксиса.

В нашей работе мы используем ценный опыт исследований В. С. Баевского, однако сделали ряд существенных изменений в признаковой схеме. Эти изменения сводятся к тому, что было уменьшено количество чисто стиховедческих признаков и увеличена представленность лингвистических параметров. Разумеется, анализ стихотворного текста не может обойтись без некоторого набора наиболее существенных признаков стиховедческого ряда. Их представленность в нашем исследовании составляет немногим больше четверти от всего количества признаков (26%). В число таких стихотворных параметров нашей признаковой схемы мы включили ритмические и синтаксические («стихотворный синтаксис») признаками.

В состав ритмических характеристик (их 6) входят: число пропусков ударений на первой, второй и последней сильных позициях. Сильной является позиция, на которую согласно метрической схеме должно падать ударение. Для ямба сильными являются четные слоги. Кроме того, нами привлечены такие важные характеристики стихотворной строки, как тип анакрусы и клаузулы. Анакруса – это группа слогов перед первой сильной позицией, клаузула – группа слогов после последней сильной позиции. И анакруса, и клаузула характеризуют протяженность стиха в его инициальной и финальной частях. Наиболее распространенной в текстах Лонгфелло является безударная анакруса длиной в один слог и мужская клаузула, при которой ударный рифмующийся слог является последним в строке. Отклонения у этих характеристик от ожидаемых величин (приращение/усечение анакрусы, внесхемное ударение на анакрусе, увеличение клаузулы, например, в случае женской рифмы) меняют интонационную структуру стиха и являются важными маркерами изменений в стихотворном тексте.

Поэтический синтаксис (3 признака) представлен количеством стихотворных переносов (enjambements), разрывов строки синтаксической паузой и эмфатически маркированных стихотворных строк (строк, заканчивающихся восклицательным, вопросительным знаками либо многоточием).

В число других привлекаемых нами характеристик входят морфологические (частеречная отнесенность слов в тексте) и синтаксические признаки (синтаксическая функция). Оба параметра оказались чрезвычайно важными при анализе стихотворных текстов, что было показано в целом ряде работ известных стиховедов[12] и весьма эффективными при дискриминации стилей ряда прозаических авторов, о чем указывалось выше.

Базой морфологических характеристик являются пять морфологических классов, определяемых в терминах следующих частей речи: существительное, глагол, прилагательное, наречие, местоимение.

Мы не включили никаких лексических параметров, поскольку в нашем случае они оказались бы контекстно связанными, т.е. различали бы не столько стиль автора, сколько темы, к которым он обращался в различные периоды жизни[13].

Проводя эксперименты по оптимизации списка признаков, мы сочли целесообразным внести уточнения относительно места стихотворной строки, где может быть манифестирован тот или иной признак. Такими позициями в стихе являются его сильные позиции, на которые, как было сказано выше, должны падать ударения в соответствии с используемой метрической схемой. В нашем исследовании мы выделяем первую, последнюю сильные позиции, а также середину стихотворной строки. Финальная позиция, как правило, является у Лонгфелло рифмующейся, что увеличивает значимость заполняющих ее элементов. Инициальная часть стиха определяет его ритмическую напряженность[14]. Средняя часть строки выделяется нами как «остаточная». Она непостоянна по своей длине и в зависимости от размера в лирических произведениях Лонгфелло может включать от одной до трех сильных позиций. Следует признать, что вследствие различной длины строк середина стиха является менее эксплицитно выделимой. Так, если в 10-сложной строке о середине можно говорить с уверенностью, то в 6-сложной строке ее можно выделить лишь условно, а к строке из четырех слогов понятие середины неприменимо вовсе. В соответствии с этим мы включили в базовый список параметров локализацию признаков только относительно первой и последней сильных позиций, а срединную локализацию привлекали только как дополнительный параметр.

Синтаксические признаки (16 характеристик) включают количество полных и частичных инверсий, простых предложений в составе сложносочиненных и сложноподчиненных предложений. Кроме того, мы также сочли целесообразным отразить синтаксическую функцию слов, замещающих указанные выше сильные позиции в стихе.

Ниже приводятся основные группы параметров, которые составили базу признаковой схемы. Признаки принадлежат четырем уровням исследования, их число равно 35. Они могут быть представлены в виде таблицы (табл.1), отражающей их локализацию в строке. Ряд признаков не ограничен какой-либо частью строки, эти признаки являются макроструктурными, характеризующими всю стихотворную строку целиком.

Следует отметить, что данный список признаков был получен в результате целого ряда экспериментов по классификации и дискриминации стихотворных текстов, а также сопоставления оригиналов стихотворных произведений и их переводов[15].

В качестве основного метода исследования нами применяется дискриминантный анализ[16]. Термин «дискриминантный анализ» относится к нескольким связанным статистическим процедурам, которые могут использоваться совместно либо выборочно в зависимости от поставленных целей. Среди его главных задач – сопоставление различных классов объектов (в нашем случае групп текстов, написанных на разных этапах творчества поэта) одновременно по большому числу характеристик с целью выявления дискриминирующих их признаков.

Таблица 1

Признаки, используемые в исследовании

Разделы	Локализованные признаки		Макроструктурные признаки
	Первая сильная позиция	Последняя сильная позиция	
Морфология	Часть речи (5)	Часть речи (5)	
Синтаксис	Члены предложения (6)	Члены предложения (6)	Состав сложносочиненных и сложноподчиненных предложений. Тип инверсии (4)
Ритмика	Пропуск икта Анакруса (4)	Пропуск икта Клаузула (2)	
Поэтический синтаксис			Перенос. Разрыв строки синтаксической паузой. Эмоциональная маркированность стиха (3)

Этот метод в настоящее время успешно применяется в зарубежной лингвистике для атрибуции и автоматической классификации текстов[17]. Вместе с тем в отечественной лингвистике эта методика не нашла пока сколько-нибудь значительного применения.

Следует отметить, что дискриминантный анализ, кроме решения нашей основной задачи по определению того, имело ли место лингвистически релевантное изменение стиля Лонгфелло, может также быть использован для проверки получаемых результатов путем выяснения точности дискриминации[18].

В соответствии с биографическими данными Лонгфелло было выделено четыре периода его творческой деятельности: первый (начало двадцатых годов XIX в.), второй (с 1826 г. до начала 1840-х гг.), третий (1843–1865 гг.), четвертый (период после окончания Гражданской войны в США).

Все лирические произведения этих периодов, вошедшие в прижизненные сборники произведений поэта[19] общим объемом более 2100 строк, были описаны при помощи указанных базовых признаков, а также ряда дополнительных параметров (см. далее). К анализу привлекались только соизмеримые по размеру произведения не менее 8 и не более 70 строк.

Дискриминантный анализ показал, что выделенные периоды творчества достаточно эксплицитно различаются между собой по стилю. Для базового набора признаков точность дифференциации составила чуть более 91%. Это означает, что для 91% произведений Лонгфелло базовые признаки успешно определяют период их создания. И это весьма хороший показатель.

В то же время имеется целый ряд интегральных признаков, определяющих общую устойчивость стиля поэта. В их число входят ритмические признаки начала строки (количество пропусков первого и второго икта, длина анакрусы), морфологические (количество глаголов и местоимений в обеих исследуемых позициях, количество прилагательных в первой и существительных в последней сильных позициях), ряд синтаксических признаков (количество сказуемых в первой и последней сильных позициях, количество прямых дополнений в последней сильной позиции, число придаточных в сложносочиненных предложениях), а также один признак поэтического синтаксиса (количество эмфатически маркированных строк).

Рассмотрим более подробно те признаки, которые являются наиболее значимыми (входят в дискриминантную модель) для дифференциации стиля Лонгфелло на различных этапах творчества.

В таблице 2 отражается значимость различных групп признаков для дискриминации.

Таблица 2

Значимость различных групп признаков

Группа признаков	Доля от общего числа признаков данной группы, %	Доля от общего числа попавших в модель признаков, %
Морфологические	40	24
Синтаксические	57	47
Ритмика	50	17
Поэтический синтаксис	67	12

Как следует из табл. 2, наибольший вклад в дифференциацию текстов различного времени создания вносят синтаксические и морфологические признаки.

Теперь рассмотрим участие признаков в разграничении периодов по их локализации в стихотворной строке (признаки начала, конца стиха, не локализованные макроструктурные параметры). Данные о вкладе этих групп признаков в дифференциацию периодов представлены в табл. 3.

Таблица 3

Значимость признаков различной локализации

Группа признаков	Доля от общего числа попавших в модель признаков, %
Начала строки	35
Конца строки	35
Макроструктурные	30

Как мы видим, вклад этих трех групп характеристик примерно одинаков, с некоторым превосходством локализованных параметров.

В целях экспериментальной проверки релевантности морфологических характеристик у слов в различных позициях привлекается ряд дополнительных параметров. В их число вошли морфологические признаки: а) середины строки; б) среднее количество слов различной частеречной принадлежности в каждом тексте в целом (вне локализации по месту в строке). Добавление морфологических признаков середины строки к базовому набору признаков не только не улучшило, но даже несколько ухудшило точность дискриминации (88,3%). Замена же морфологических признаков, определяемых у слов по сильным позициям, на обобщенные показатели по всему произведению привела к еще большему падению точности (83%). Это подтверждает релевантность первой и последней сильных позиций при поиске дискриминирующих факторов стихотворного текста.

В целом полученные результаты свидетельствуют о наличии значимых изменений идиостиля Лонгфелло с точки зрения лингвистических параметров его текстов и об эффективности выделенной признаковой схемы в плане определения лингвистических маркеров, дифференцирующих разные периоды творчества поэта, а также для выявления интегральных черт его стиля.

ПРИМЕЧАНИЯ

[1] *Мартыненко Г. Я.* Основы стилеметрии. – Л.: Изд.-во Ленингр. ун-та, 1988; *Марусенко М. А., Бессонов Б. А., Богданова Л. М., Аникин М. А., Мясоедова Н.Е.* В поисках потерянного автора: Этюды атрибуции. Сер. Филологические исследования. – СПб.: Филологический факультет СПбГУ, 2001; *Koppel M, Argamon S., Shimoni A. R.* Automatically Categorizing Written Texts by Author Gender //

Literary & Linguistic Computing. – 2002. – Vol. 17. – P. 401–412; *Хьетсо Г., Густавссон С., Бекман Б., Гил С.* Кто написал «Тихий Дон»? (Проблема авторства «Тихого Дона»). – М., 1989.

[2] *Biber D.* Dimensions of Register Variation: A Cross-Linguistic Comparison. – Cambridge: Cambridge University Press, 1995; *Karlgren, J.* Stylistic Experiments in Information Retrieval // T. Strzalkowski (ed.) Natural Language Information Retrieval. – Dordrecht: Kluwer, 1999. – P. 147–166.

[3] *Mikros G., Carayannis G.* Modern Greek corpus taxonomy // Proceedings of the 2nd International Conference on Language Resources and Evaluations, Athens, Greece, 2000, 31 May – 2 June. – Vol. 3. – P. 129–134.

[4] *Tambouratzis G., Markantonatou S., Hairetakis N., Vassiliou M., Carayannis G., Tambouratzis D.* Discriminating the Registers and Styles in the Modern Greek Language – Part 2: Extending the Feature Vector to Optimize Author Discrimination // Literary and Linguistic Computing. – 2004. – Vol. 19. – N 2. – P. 221–242.

[5] *Burrows, J.F.* Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style // Literary and Linguistic Computing. – 1987. – N 2. – P. 61–70; *Mosteller F., Wallace D. L.* Applied Bayesian and Classical Inference. The Case of the Federalist Papers. – 2nd edition. – New York: Springer-Verlag, 1984; *Tse E. E., Tweedie F. J, Frischer B. D.* Unravelling the Purple Thread: Function Word Variability and the Scriptores Historiae Augustae // Literary and Linguistic Computing. – 1998. – N 13, 3. – P. 141–149.

[6] *Holmes D. I.* Authorship attribution // Computers and the Humanities. – 1994. – N 28 (1). – P. 87–106; *Gurney P. J., Gurney L. W.* Authorship attribution of the Scriptores Historiae Augustae // Literary and Linguistic Computing. – 1998. – N 13 (3). – P. 119–31.

[7] *Stamatatos E., Fakotakis N., and Kokkinakis G.* Computer-based authorship attribution without lexical measures // Computers and the Humanities. – 2001. – N 35 (2). – P. 193–214.

[8] *Марусенко М. А.* Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. – Л.: Изд-во Ленингр. ун-та, 1990.

[9] *Kelih E., Antić G., Grzybek P., Stadlober E.* Classification of Author and/or Genre? The Impact of Word Length // C. Weihs and W. Gaul (Eds.) Classification – The Ubiquitous Challenge. – Heidelberg: Springer, 2005. – P. 498–505.

[10] *Баевский В. С.* Лингвистические, математические, семиотические и компьютерные модели в истории и теории литературы. – М.: Языки славянской культуры, 2001. – С. 186.

[11] *Баевский В. С.* Пастернак – лирик. – Смоленск: Траст-Имаком, 1993. – С. 59.

[12] *Гаспаров М. Л.* Лингвистика стиха // Славянский стих: Стиховедение, лингвистика и поэтика. – М.: Наука, 1996. – С. 5–17; *Скулачева Т. В.* Ритм и грамматика в стихе // Славянский стих. Лингвистическая и прикладная поэтика: Лингвистическая и прикладная поэтика: Материалы международной конференции 23–27 июня 1998 г. – М.: Языки славянской культуры, 2001. – С. 121–129.

[13] *Mikros G.K., Argiri E.K.* Investigating topic influence in authorship attribution // Proceedings of the SIGIR 2007. Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection. – Amsterdam, 2007.

[14] *Красноперова М. А.* Основы реконструктивного моделирования стихосложения (на материале ритмики русского стиха). – СПб.: Издательство С.-Петербургского университета, 2000.

[15] *Андреев В. С.* Классификация стихотворных текстов Уитьера // Известия Российского государственного педагогического университета им.А. И. Герцена. – 2007. – № 7 (28). – С. 24–35; *Андреев В. С.* Динамика идиостиля Дж. Г. Уитьера // Вестник Тамбовского университета. Сер. Гуманитарные науки. – Тамбов, 2007. – Вып. 12 (56). – С. 266–270.

[16] *Клекка У. Р.* Дискриминантный анализ // Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. – С. 80–138.

[17] *Holmes D. I., Forsyth D.* The Federalist Revisited: New Directions in Authorship Attribution // Literary and Linguistic Computing. – 1995. – Vol. 10. – P. 111; *Karlgen J., Cutting D.* Recognizing text genres with simple metrics using discriminant analysis // Procedeings of COLING 94. – Kyoto, 1994. – P. 1071–1075; *Murata M.* Identify a Text's Genre by Multivariate Analysis – Using Selected Conjunctive Words and Particle-phrases // Proceedings of the Institute of Statistical Mathematics. – 2000. – Vol. 48. – P. 311–326.

[18] *Warner R. M.* Applied Statistics. – Los Angeles; London: Sage Publications, 2008. – P. 650–701.

[19] The Works of Henry Wardsworth Longfellow / With an Introduction by Alvert Glover and Bibliography. – Ware: Wadsworth Editions Ltd., 1994.