Известия Воронежского государственного педагогического университета. 2026. № 3 (312). С. 239–247.
Izvestia Voronezh State Pedagogical University. 2026. (3), 239–247.

Научная статья
УДК 81'33
DOI: 10.47438/2309-7078_2026_3_239.

УСТОЙЧИВОСТЬ ИДИОЛЕКТНОГО СИГНАЛА ПРИ СМЕНЕ КОММУНИКАТИВНОЙ РОЛИ (НА МАТЕРИАЛЕ ТЕКСТОВ В ЖАНРЕ «ПИСЬМО-УГРОЗА»)

Ирина Валентиновна Швецова[1, 2]

Экспертно-криминалистический центр МВД России[1]
Москва, Россия
Воронежский государственный педагогический университет[2]
Воронеж, Россия

[1] *Главный эксперт отдела лингвистических экспертиз управления экспертиз цифровой и речевой информации*
[2] *Аспирант кафедры русского языка, современной русской и зарубежной литературы,*
ORCID ID: 0009-0007-2496-6976, тел.: +7-499-745-79-45, e-mail: ishvetcova6@mvd.ru

Аннотация. В статье оценивается устойчивость идиолектного сигнала[3] при контролируемом изменении коммуникативной роли автора. Материалом исследования послужил сформированный на базе Экспертно-криминалистического центра МВД России[4] экспериментальный корпус, состоящий из 270 текстов в жанре «письмо-угроза». Тексты для данного корпуса создавались 90 авторами в соответствии с тремя задаваемыми коммуникативными условиями. В зависимости от реализованной коммуникативной роли в исследовании выявляются наиболее надежные стилеметрические признаковые каналы, указывающие на устойчивость идиолектного сигнала. Отмечается неоднородность проявления сигнала, в частности, при переходе к третьему условию – к роли, предполагающей негативное отношение к какой-либо группе лиц, объединенных по национальному признаку. Подчеркивается целесообразность применения для анализа коротких текстов в жанре «письмо-угроза» компактной многокомпонентной модели, сочетающей символьные, морфологические и структурно-пунктуационные идиолектные признаки. Обозначаются при этом границы ее применимости: в условиях отсутствия сведений об авторах выделенные моделью схожие тексты требуют последующего качественного экспертного анализа.

Ключевые слова: идиолект, коммуникативная роль, стилеметрия, вариативность письменной речи, автороведческая экспертиза, символьные *n*-граммы, преобразование Хеллингера, пермутационный дисперсионный анализ, авторонезависимая проверка.

Для цитирования: *Швецова И. В.* Устойчивость идиолектного сигнала при смене коммуникативной роли (на материале текстов в жанре «письмо-угроза») // Известия Воронежского государственного педагогического университета. 2026. № 3. С. 239–247. DOI: 10.47438/2309-7078_2026_3_239.

Введение

Устойчивость и вариативность относятся к числу основополагающих свойств языка. Язык должен быть достаточно устойчивым для того, чтобы обеспечивать взаимопонимание его носителей, и одновременно вариативным для того, чтобы позволить говорящему в зависимости от условий общения выбирать более подходящие языковые средства. Идиолект, с учетом данной дихотомии, свойственной языку в целом, нельзя рассматривать как неизменный набор признаков, проявляющихся с одинаковой интенсивностью во всех текстах одного автора. В настоящем исследовании мы опираемся на концепцию Т. А. Литвиновой, согласно которой устойчивые и вариабельные черты субъекта идиолектной деятельности рассматриваются в их неразрывной связи [6].

Изучение языковой вариативности, по замечанию В. Г. Гака, «не может ограничиваться анализом

[3] Под идиолектным сигналом в настоящей работе понимается совокупность речевых признаков (в том числе количественных), относительно устойчивых при смене коммуникативных условий и позволяющих статистически различать тексты одного и разных авторов.

[4] Далее – ЭКЦ МВД России.

© Швецова И. В., 2026

отдельных фактов или групп фактов на различных уровнях языковой системы» [2, с. 366]. Исследователь выделяет три группы факторов, обусловливающих языковую вариативность. Наиболее релевантными для настоящей работы представляются две из них – «внутриязыковые (системно-структурные)» и «внешние по отношению к системе языка, но внутренние по отношению к языку в целом и к его функционированию: логические и психологические факторы, связанные с общими закономерностями человеческого общения и мышления» [2, с. 369–370].

К подобным факторам можно отнести различные условия речевой деятельности, например, тему, жанр текста, адресата, коммуникативную цель, регистр, степень подготовленности текста, речевую стратегию, избранную автором для построения текста. Необходимость их учета неоднократно подчеркивалась как в работах, опирающихся на качественно-описательный подход при изучении индивидуальных особенностей речи [1; 5 и др.], так и на количественно-статистический [9; 10; 14 и др.]. Существенно усиливает вариативность речи сознательная стилистическая маскировка, что, как отмечают М. Бреннан, Р. Гринштадт в своем экспериментальном исследовании, значительно снижает идиолектный сигнал [11]. Вместе с тем остается открытым вопрос, полностью ли разрушается идиолектный сигнал при изменении коммуникативной роли автора, или его признаки перераспределяются между разными уровнями организации текста.

Тексты жанра «письмо-угроза» относятся к эпистолярному поджанру, в котором форма письма используется для реализации деструктивного речевого действия. Такие тексты криминалистически значимы и могут быть объектом судебной авторoведческой экспертизы. По формальным признакам письма-угрозы, написанные в рамках одной темы, одного жанра и адресованные однотипной группе получателей (представителям государственных и иных структур), являются сопоставимыми объектами для проведения указанной экспертизы. Подобные тексты при этом подчинены жесткому жанровому канону, проявляющемуся в использовании большого количества речевых клише и шаблонных конструкций, а также имеют малый объем, что снижает количество устойчивых идиолектных признаков. Кроме того, как показал предварительный анализ писем-угроз [3], субъект, создавший такой текст, может целенаправленно изменять лексику, синтаксические конструкции, структуру, регулировать степень экспрессивности, искажать графическое оформление или – что особенно важно для настоящего исследования – выбирать способ самопрезентации, отличный от своего.

Обозначенные особенности писем-угроз (типизированный жанр, однотипный адресат, малый объем) в соответствии с критериями действующих методик ЭКЦ МВД России [см. подробнее: 4; 8] относят письма-угрозы к объектам, имеющим ограниченную пригодность для проведения авторoведческого исследования. В качестве дополнительного инструмента для проверки сохранения устойчивых идиолектных признаков в данных текстах нами были применены количественно-статистические методы. С целью проведения исследования на базе научно-исследовательской лаборатории ЭКЦ МВД России по изучению возможностей использования методов компьютерной лингвистики в экспертно-криминалистической деятельности[5] был сформирован экспериментальный корпус текстов-угроз, в котором принадлежность текстов заранее известна; каждый автор представлен несколькими текстами, созданными в разных коммуникативных условиях.

Таким образом, целью настоящего исследования является установить, какие формализуемые (стилеметрические) группы признаков сохраняют идиолектный сигнал при смене коммуникативной роли, а также определить признаки, пригодные для последующего анализа экспертом реальных анонимных писем-угроз.

Материал исследования

Экспериментальный корпус состоит из 270 текстов в жанре «письмо-угроза», созданных 90 авторами в соответствии с тремя задаваемыми коммуникативными условиями – написать текст от лица 1) своего пола, 2) противоположного пола, 3) любого пола, который выражает негативное отношение к какой-либо группе лиц, объединенных по национальному признаку. Первоначально корпус включал 273 текста, созданных 91 автором[6]. В ходе проверки корпуса установлено, что у автора с условным номером 50 тексты, представленные для первого и второго задания, полностью совпадают. С целью сохранения сбалансированной структуры материала все три текста данного участника были исключены из дальнейшего анализа.

Методика исследования

Единицей анализа выступала пара текстов. Для каждого автора формировались три пары текстов, созданных в разных условиях: «1–2» (сочетание текстов первого и второго задания), «1–3» (сочетание текстов первого и третьего задания), «2–3» (сочетание текстов второго и третьего задания). Среди данных пар выделялись положительные, то есть пары текстов, принадлежащие одному автору, и отрицательные – пары текстов, принадлежащие разным авторам. Отрицательные пары формировались из текстов разных авторов одной гендерной группы с разницей в возрасте не более пяти лет. Такое выравнивание было необходимо для снижения вероятности того, что модель будет различать тексты преимущественно по социально-демографическим характеристикам. Число положительных и отрицательных пар оказалось несбалансированным, что учитывалось при обучении моделей посредством весовых коэффициентов.

Признаки, ввиду неоднородности источников сходства, рассчитывались на параллельных представлениях текста. Так, исходная версия (*text_raw*) сохранялась без изменений; техническая версия (*text_technical*) унифицировала пробелы, кавычки и тире, а также приводила различные формы обезличивания к единому маркеру <PLACEHOLDER>;

[5] Лаборатория создана в соответствии с приказом ЭКЦ МВД России № 178 от 27 декабря 2024 г. Руководитель лаборатории – заместитель начальника ЭКЦ МВД России А. В. Громова.

[6] Описание процедуры формирования корпуса, социально-демографический состав выборки представлены в работе: [7].

деобфусцированная[7] версия (*text_deobfuscated*) дополнительно устраняла графическую маскировку: например, замену кириллической буквы на визуально похожую букву латиницы; идиолектная версия (*text_idiolect*) строилась поверх деобфусцированной и заменяла URL, электронную почту, телефоны, даты и числа на категориальные маркеры <URL>, <EMAIL>, <PHONE>, <DATE_TIME>, <NUM> с сохранением позиции элемента внутри структуры текста. Признаки графического уровня при этом предварительно извлекались в отдельный канал по исходному тексту, что позволяло сохранить их для самостоятельного анализа и не утратить при нормализации. Разделение текстовых представлений позволило избежать смешения задач анализа.

Морфологическая аннотация выполнялась автоматически с использованием модели *Universal Dependencies (UDPipe)*. Нестандартная лексика, орфографические ошибки и элементы графической маскировки могли снижать точность автоматической морфологической разметки, в связи с чем подобные последовательности были исключены из морфологических и структурно-пунктуационных представлений текста, но сохранены в символьных признаковых пространствах.

В анализ вошли девять признаковых пространств, различающихся по уровню языковой организации и размерности и ранее апробированных в диссертационном исследовании Т. А. Литвиновой [6]: 1) UPOS – характеризует общий частеречный состав текста; 2) FULLPOS – учитывает часть речи совместно с набором морфологических характеристик; 3) MORPH_VALUE – отражает отдельные значения грамматических категорий; 4) POS_PUNCT_3 и 5) POS_PUNCT_4 – учитывают сочетание частей речи и функциональных классов пунктуационных знаков; 6) WORD_PUNCT_3 и 7) WORD_PUNCT_4 – представляют собой вариант, в котором все словесные элементы обозначены единым маркером, знаки препинания – маркерами завершения предложения и прочей пунктуацией; 8) CHAR_3 и 9) CHAR_4 – отражают символьные признаки, триграммы и тетраграммы. Например, предложение «Отправь мне деньги.» в соответствии с указанными пространствами включает следующий набор признаков: 1) глагол + местоимение + существительное; 2) глагол повелительного наклонения, 2-го лица, единственного числа + местоимение 1-го лица, единственного числа, дательного падежа + существительное женского рода, неодушевленное, в винительном падеже, множественного числа; 3) повелительное наклонение + 2-е лицо + 1-е лицо + единственное число + дательный падеж + женский род + неодушевленность + винительный падеж + множественное число; 4) глагол + местоимение + существительное, местоимение + существительное + знак завершения предложения; 5) глагол + местоимение + существительное + знак завершения предложения; 6) слово + слово + слово, слово + слово + знак завершения предложения; 7) слово + слово + слово + знак завершения предложения; 8) отп + тпр + пра + рав + авь + вь_[8] + ь_м + _мн + мне + не_ + е_д + _де + ден + ень + ньг + ьги + ги.; 9) отпр + тпра + прав + равь + авь_ + вь_м + ь_мн + _мне + мне_ + не_д + е_де + _ден + день + еньг + ньги + ьги.

Для снижения зависимости признаков от объема текста использовались относительные частоты с последующим преобразованием Хеллингера, ослабляющим доминирование наиболее частотных признаков и сохраняющего информативность редких единиц. Парное сходство текстов оценивалось косинусной мерой, поскольку она учитывает направление векторов, а не их абсолютную величину, и хорошо применима к разреженным признаковым представлениям. При неотрицательных компонентах векторов значения этой меры находились в диапазоне от 0 до 1. Показатель сходства рассчитывался отдельно для каждого признакового пространства. В связи со значительным различием размерностей пространств абсолютные значения сходства между ними непосредственно не сопоставлялись; сравнивалась способность каждого пространства ранжировать пары текстов одного и разных авторов.

Анализ проводился в два последовательных этапа, на каждом из них использовалась одна и та же схема разделения данных: пятикратная перекрестная проверка, повторенная десять раз; всего пятьдесят разбиений. Разбиение выполнялось по авторам, в результате чего три текста одного участника в каждом разбиении целиком относились либо к обучающей, либо к тестовой части. Таким образом, тексты авторов из тестовой выборки не встречались модели на этапе обучения. В данном случае проверялось не узнавание уже знакомых участников, а сама работа модели на новых, ранее не представленных авторах. Различие между этапами состояло в способе формирования признакового пространства.

На первом этапе девять признаковых пространств сравнивались между собой. Здесь оценивалась различительная способность отдельных блоков признаков, влияние коммуникативного условия на структуру каждого пространства и устойчивость идиолектного сигнала. Словари признаков на этом этапе строились один раз по всему экспериментальному корпусу на предшествующем шаге подготовки данных и дальше не менялись. Внутри каждого разбиения по обучающим авторам определялись только два параметра – поправка на средний эффект коммуникативного условия и порог классификации. Поправка рассчитывалась следующим образом. Для каждого из трех заданий находилось отклонение его среднего вектора от общего среднего вектора обучающих текстов, после чего полученные отклонения применялись и к обучающим, и к тестовым текстам. Такая коррекция позволяла отделить собственно межавторские различия от общего сходства текстов, написанных в одном и том же условии. Поскольку словарь здесь строился по корпусу в целом, результаты первого этапа характеризуют сравнительную различительную способность признаковых пространств, а не их итоговую переносимость на новых авторов.

На втором этапе проводилась строгая авторонезависимая проверка. В нее вошли каналы, отобранные по итогам первого этапа (CHAR_3, CHAR_4,

[7] Деобфусцированная (деобфускация; от англ. obfuscation – затемнение, запутывание) – восстановленная после искусственных искажений форма текста.

[8] Знаком «_» обозначен пробел.

MORPH_VALUE, WORD_PUNCT_3, FULLPOS), и их компактные комбинации, в том числе трехкомпонентная модель (CHAR_4, MORPH_VALUE, WORD_PUNCT_3). Принципиальное отличие этого этапа в том, что признаковое пространство здесь строилось заново внутри каждого разбиения и только по обучающим авторам. Для каждого признака подсчитывалось, в скольких документах обучающей части он встречается. В словарь включались признаки, встретившиеся не менее чем в трех обучающих текстах (для символьных *n*-грамм) или не менее чем в двух (для морфологических и структурно-пунктуационных блоков). Тестовые документы проецировались в этот обучающий словарь, преобразование Хеллингера выполнялось уже после отбора признаков. Каналы объединялись взвешенной логистической моделью по значениям парного сходства. Стандартизация значений, обучение модели и выбор порога также выполнялись только на обучающей части; дисбаланс классов компенсировался весом положительного класса. Таким образом, тестовые авторы не участвовали ни в формировании признакового пространства, ни в настройке модели. Все признаки извлекались с использованием скриптов на языке R. Анализ также производился в данной среде.

До описанных двух этапов выполнялась предварительная проверка наличия идиолектного сигнала, которая опиралась на три укрупненных канала сходства: распределение служебных слов, профили символьных *n*-грамм и числовые стилометрические характеристики текста. Каждая пара текстов описывалась тремя значениями сходства, которые объединяла логистическая модель. Так же, как и на основных этапах, авторы делились на обучающих и тестовых. Результаты здесь носили диагностический характер и не рассматривались в качестве окончательной оценки, поскольку словари и параметры каналов строились по корпусу в целом, до разбиения.

В качестве показателей качества модели на обоих этапах использовались ROC AUC и Average Precision. Показатель ROC AUC характеризует общую способность модели присваивать парам текстов одного автора более высокие значения сходства, чем парам текстов разных авторов. Average Precision (AP) отражает качество ранжирования положительных пар (текстов одного автора), показывая, насколько высоко они располагаются в списке пар, упорядоченном по убыванию сходства. Поскольку Average Precision зависит от доли положительных примеров, дополнительно рассчитывалось ее отношение к этой доле. Порог классификации устанавливался отдельно для каждой обучающей выборки как 95-й процентиль распределения значений сходства отрицательных пар; тестовые данные при его выборе не использовались.

Оценка влияния коммуникативного условия выполнялась в рамках первого этапа анализа с помощью пермутационного дисперсионного анализа на основе матрицы попарных расстояний между текстами. Для исключения постоянных межсубъектных различий из признаковых векторов каждого автора предварительно вычитался средний вектор, рассчитанный по трем коммуникативным условиям. Перестановки проводились только между текстами одного и того же автора. При проверке результатов по девяти признаковым пространствам применялась поправка Холма на множественные сравнения. Для оценки индивидуальной реакции каждого автора на смену коммуникативной роли использовался индекс, отражающий изменение узнаваемости в парах текстов, включавших третье коммуникативное условие, по сравнению с базовым переходом между первым и вторым условиями. Узнаваемость определялась как процентиль значения сходства текстов одного автора относительно распределения значений сходства сопоставимых пар текстов разных авторов. Направленные различия проверялись с помощью парного критерия Уилкоксона по всем авторам с поправкой Холма на множественные сравнения. Согласованность индивидуальной различимости между разными пространствами признаков оценивалась с помощью ранговой корреляции Спирмена.

Результаты исследования

Полученные результаты предварительного анализа трех укрупненных каналов подтвердили наличие идиолектного сигнала. Наиболее информативным отдельным каналом оказались символьные *n*-граммы, вторым по различительной способности – совокупность числовых стилометрических характеристик, в то время как распределение служебных слов обладало сравнительно слабой самостоятельной различительной способностью. Символьные признаки рассматривались как основной потенциальный источник устойчивого идиолектного сигнала, морфологические и структурно-пунктуационные характеристики как признаки, способные его дополнить. Вместе с тем укрупненные каналы были неоднородны по составу и не позволяли установить, какие именно уровни языковой организации обеспечивают наблюдаемое авторское сходство, в связи с чем был осуществлен переход к детализированным признаковым пространствам. Далее приведены результаты первого этапа анализа – сравнительного исследования девяти признаковых пространств.

Влияние коммуникативного условия оказалось статистически значимым во всех девяти признаковых пространствах. Эффект сохранялся после поправки на множественные сравнения, хотя его величина была небольшой: условие объясняло от 1,4 до 6,0 % вариативности признаков. Наиболее выраженным он оказался в пространстве морфологических значений, наименее – в последовательностях частей речи с пунктуацией. То, что эффект коммуникативного условия оказался статистически значимым, но сравнительно небольшим по величине, само по себе имеет важное содержательное значение. Условие действительно перестраивает речевой материал, но не является доминирующим фактором, определяющим сходство текстов.

Сравнение девяти признаковых пространств до и после устранения среднего эффекта коммуникативного условия приведено на рис. 1.

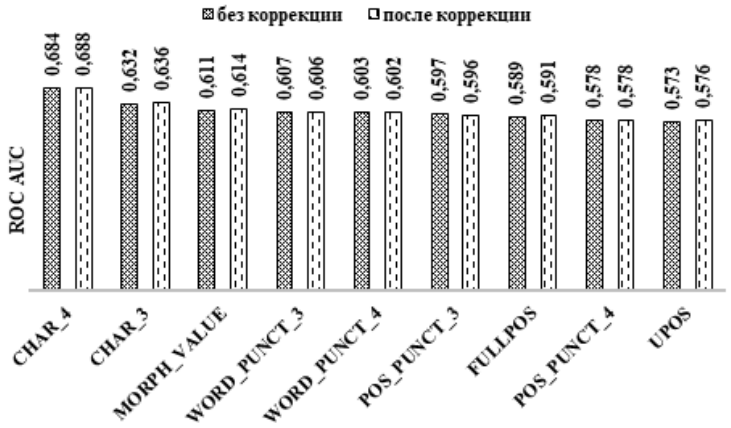


Рис. 1. ROC AUC девяти признаковых пространств до и после устранения среднего эффекта коммуникативного условия (этап 1 – сравнительный анализ)

В сравнительном анализе первого этапа наибольшую различительную способность продемонстрировали символьные тетраграммы (CHAR_4): показатель ROC AUC составил 0,688, Average Precision – 0,383. Далее следовали символьные триграммы (CHAR_3), затем – отдельные морфологические признаки (MORPH_VALUE). Наименьшую различительную способность показали частеречный состав текста и последовательность частей речи и знаков препинания (UPOS, POS_PUNCT_4). Приведенные оценки получены при фиксированных признаковых словарях, сформированных по корпусу в целом, и характеризуют относительный порядок признаковых блоков; они не являются результатами финальной интеграционной модели. Окончательные оценки отдельного канала и интегрированных моделей, полученные на втором этапе в строгой авторонезависимой проверке, представлены ниже.

Устранение среднего эффекта коммуникативного условия практически не повлияло на различительную способность исследованных пространств. Так, изменение показателя ROC AUC находилось в пределах от 0,002 до +0,005. Максимальное улучшение (+0,0047) отмечено для символьных тетраграмм. Такой незначительный сдвиг свидетельствует о том, что выявляемый идиолектный сигнал не сводится к общему сходству текстов, созданных в одних и тех же коммуникативных условиях, а отражает устойчивые индивидуальные языковые предпочтения автора. Переход от символьных триграмм к тетраграммам сопровождался заметным ростом различительной способности. Более длинные символьные сочетания позволяют учитывать более протяженные локальные закономерности текста, в которых могут одновременно проявляться морфемные, лексические и пунктуационные особенности авторской речи.

Различительная способность оценивалась отдельно для трех типов сопоставлений коммуникативных условий – «1–2», «1–3» и «2–3» (рис. 2).

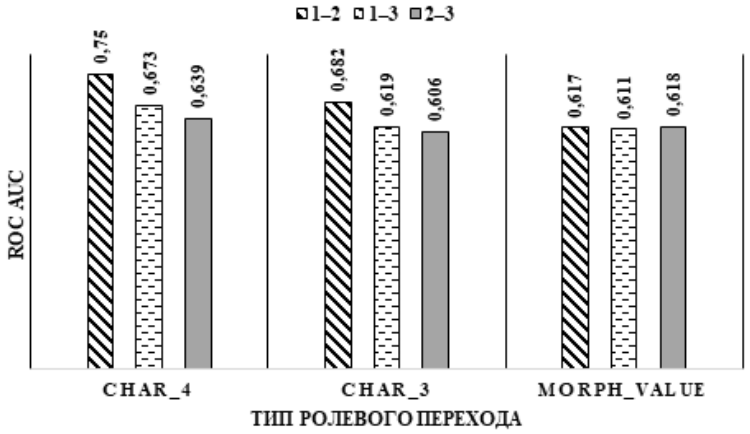


Рис. 2. Различительная способность (ROC AUC) символьных *n*-грамм и морфологических признаков при различных типах переходов коммуникативной роли (этап 1 – сравнительный анализ)

В пространстве символьных тетраграмм (CHAR_4) наблюдалась четкая градация качества модели в зависимости от типа ролевого перехода: показатель ROC AUC снижался с 0,75 (для пар «1–2») до 0,673 («1–3») и 0,639 («2–3»). Average Precision составила 0,461, 0,397 и 0,339 соответственно. При средней доле положительных пар около 0,165 эти значения превышали базовый уровень случайного ранжирования в 2,81, 2,41 и 2,06 раза. В пространстве морфологических значений различительная способность практически не зависела

от типа перехода. В пространстве обобщенных последовательностей «слово – знак препинания» (WORD_PUNCT_3 и WORD_PUNCT_4), напротив, наблюдалась обратная тенденция: пары «1–2» различались слабее, чем сопоставления с участием третьего коммуникативного условия.

Для каждого автора и каждого типа ролевого перехода рассчитывался процентиль сходства текстов одного автора относительно распределения сходства сопоставимых пар текстов разных авторов. Средний уровень идиолектной узнаваемости варьировался в зависимости от признакового пространства. Максимальный средний уровень индивидуальной узнаваемости (0,704) был зафиксирован для символьных тетраграмм, минимальный (0,581) – для частеречного состава текста. Тексты одних авторов демонстрировали стабильно высокую узнаваемость в большинстве признаковых пространств, тогда как тексты других участников неизменно оставались в нижней части распределения, то есть оставались трудноразличимыми, что соответствует результатам, полученным в работах Т. А. Литвиновой [12; 13]. Ранжирование авторов по степени узнаваемости оказалось согласованным внутри структурно близких признаковых пространств. Наиболее высокие коэффициенты ранговой корреляции Спирмена наблюдались между триграммами и тетраграммами последовательностей «слово – знаки препинания» (ρ=0,945), между символьными триграммами и тетраграммами (ρ=0,88), а также между разными вариантами последовательностей частей речи со знаками препинания (ρ=0,825). Между каналами, различающимися по структуре, согласованность снижалась до 0,25–0,27. Полученная согласованность показывает, что индивидуальная различимость не сводится исключительно к выбору конкретной длины *n*-грамм, хотя степень согласованности существенно зависит от типа признакового пространства.

Переход к третьему коммуникативному условию продемонстрировал выраженную специфику на разных уровнях языковой организации. В пространстве символьных тетраграмм (CHAR_4) средняя узнаваемость авторов при включении третьего условия снижалась: для перехода «1–2» этот показатель составлял 0,776, для переходов с участием третьего коммуникативного условия (среднее по парам «1–3» и «2–3») – 0,668. Напротив, в пространствах последовательностей «слово – знак препинания» наблюдался противоположный эффект – узнаваемость автора возрастала. Статистическая значимость этих разнонаправленных изменений была подтверждена критерием Уилкоксона с поправкой Холма для трех признаковых пространств: CHAR_4, WORD_PUNCT_3 и WORD_PUNCT_4. Для символьных триграмм (CHAR_3) аналогичное снижение узнаваемости проявилось лишь на уровне описательной тенденции.

Анализ векторов смещения показал, что доля авторов, чье речевое поведение отклонялось от группового тренда, существенно различалась в зависимости от блока признаков. В пространстве символьных тетраграмм таких участников было немного, тогда как на уровне структурно-пунктуационных последовательностей их доля оказалась значительно выше. Наличие общей тенденции изменения текста при широком индивидуальном разбросе подтверждает, что смена коммуникативной роли задает общий вектор языковой перестройки, поверх которого реализуются индивидуальные стратегии речевой адаптации автора.

Принципиально важно, что исходный уровень узнаваемости автора не предопределяет направление его реакции на смену роли. Среди субъектов с изначально высокой узнаваемостью встречались как те, кто утратил индивидуальную особенность речи в третьем условии, так и те, у кого она, напротив, усилилась. Таким образом, выраженность базового идиолектного сигнала и характер ролевой адаптации выступают как две самостоятельные характеристики речевого поведения, требующие раздельного учета при экспертном анализе.

Второй этап состоял в строгой авторонезависимой проверке отобранных каналов и их компактных комбинаций: словарь признаков и все параметры модели строились внутри каждого разбиения только по обучающим авторам. Такая проверка полностью воспроизвела иерархию каналов первого этапа. Наилучший результат среди отдельных пространств показали символьные тетраграммы: ROC AUC = 0,682, Average Precision = 0,367, что при средней доле положительных пар около 0,165 соответствует превышению базового уровня случайного ранжирования в 2,23 раза. За ними по качеству следовали символьные триграммы (ROC AUC = 0,631; AP = 0,345) и морфологические значения (ROC AUC = 0,611; AP = 0,284). Высокая позиция символьных тетраграмм не объясняется формированием признакового словаря по корпусу в целом.

Идиолектный сигнал не ограничивался символьным уровнем. Так, при полном исключении символьных *n*-грамм объединение признаковых пространств (морфологических значений, последовательностей «слово – знак препинания» и признаков, учитывающих часть речи и морфологические характеристики слова) обеспечило ROC AUC, равный 0,617, при Average Precision 0,327, то есть примерно в два раза превышающей долю положительных пар в тестовой выборке. Таким образом, индивидуальные языковые предпочтения автора проявляются также в грамматическом и структурно-пунктуационном оформлении текста, что подтверждает многокомпонентный характер идиолектного сигнала.

Результаты оценки эффективности отдельного канала CHAR_4 и интегрированных моделей, полученные в строгой авторонезависимой проверке, представлены на рис. 3.

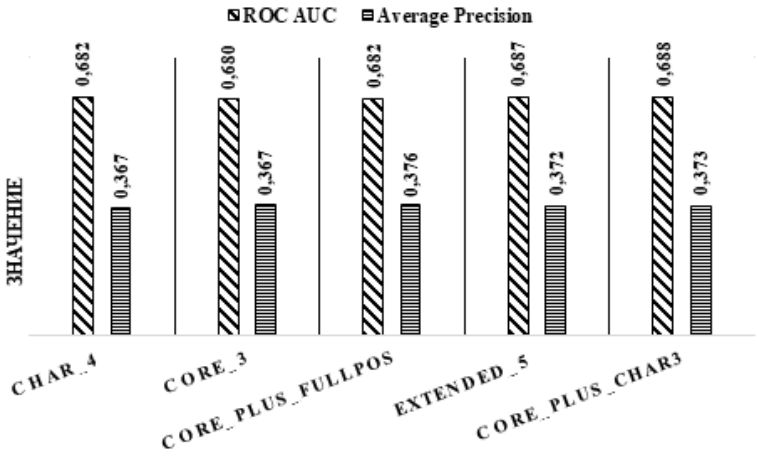


Рис. 3. Результаты строгой авторонезависимой проверки отдельного канала CHAR_4 и интегрированных моделей (этап 2)

Отдельный канал CHAR_4 показал значение ROC AUC 0,682 при Average Precision 0,367. Лидирующие интегрированные модели показали практически одинаковую эффективность. Максимальное среднее значение ROC AUC 0,688 при Average Precision = 0,373 показывает четырехблочная модель core_plus_char3 (CHAR_4, CHAR_3, MORPH_VALUE и WORD_PUNCT_3). Расширение ее состава до пяти блоков за счет включения полных морфологических характеристик слова (extended_5) практически не меняет результат (ROC AUC = 0,687; Average Precision = 0,372). Вариант дополнения базовой трехкомпонентной модели полными морфологическими характеристиками вместо символьных триграмм (core_plus_fullpos) дает близкий результат (ROC AUC = 0,682; Average Precision = 0,376). Трехкомпонентная модель core_3 (CHAR_4, MORPH_VALUE и WORD_PUNCT_3) показывает сопоставимое значение ROC AUC (0,680) при наибольшем показателе Average Precision (0,376). Различия интегрированных моделей и отдельного канала CHAR_4 по ROC AUC минимальны: они не превышают 0,007, проявляются лишь в 72 % разбиений, в части случаев вовсе отсутствуют.

Иная картина наблюдается при анализе Average Precision. Включение морфологических и структурно-пунктуационных признаков (core_3) дает прирост порядка +0,0096, сохраняющийся в 88 % тестовых разбиений. Для прикладной задачи предварительного отбора текстов писем-угроз этот эффект особенно важен, поскольку качество модели в таком сценарии определяется тем, насколько плотно положительные пары текстов сосредоточены в верхней части ранжированного списка, передаваемого на экспертную оценку.

Одновременное включение однотипных признаков, в частности, совмещение триграмм и тетраграмм или добавление полных морфологических характеристик, приводит к информационной избыточности и не повышает качества классификации. Преимущество дополнения символьного канала морфологическими и структурно-пунктуационными признаками связано прежде всего с приростом Average Precision. Трехкомпонентная модель core_3 (CHAR_4, MORPH_VALUE и WORD_PUNCT_3) дает при этом компактное многоуровневое представление сигнала – символьного, грамматического и структурно-пунктуационного, сохраняющее возможность его содержательной интерпретации.

Выводы

Проведенное исследование показало, что смена коммуникативной роли не разрушает идиолектный сигнал автора, сигнал перераспределяется между различными уровнями языковой организации. Наиболее устойчивым и надежным источником индивидуальной различимости выступают символьные тетраграммы. Включение третьего коммуникативного условия проявлялось как канал-специфический эффект: в пространстве символьных тетраграмм узнаваемость автора снижалась, в структурно-пунктуационных последовательностях, напротив, возрастала. Полученная разнонаправленность эффектов показывает, что изменение коммуникативной роли неодинаково затрагивает различные уровни организации текста. Возможные причины такой неоднородности требуют отдельного исследования.

Экспериментальные данные подтвердили нецелесообразность механического объединения максимального количества признаков вследствие их высокой информационной избыточности. Для задачи предварительного отбора может быть использована компактная трехкомпонентная модель core_3, объединяющая символьный, морфологический и структурно-пунктуационный каналы. Такая модель обеспечивает небольшой, но более устойчивый прирост Average Precision и позволяет сочетать статистическую надежность с содержательной экспертной оценкой текстов.

В экспертной практике данный подход, включающий автоматизированные методы анализа, может быть использован для предварительного исследования массива анонимных писем-угроз. Модель ранжирует тексты, выделяет ограниченный круг наиболее схожих пар, которые далее детально анализирует эксперт.

Конфликт интересов

Автор декларирует отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Библиографический список

1. Бахтин М.М. Проблема речевых жанров // Бахтин М.М. Эстетика словесного творчества. М. : Искусство, 1979. С. 237-280.
2. Гак В.Г. Языковые преобразования. М. : Школа «Языки русской культуры», 1998. 768 с.
3. Громова А.В., Литвинова Т.А., Ошкуков С.С., Швецова И.В. Криминалистическое исследование текстов жанра «письмо-угроза» компьютерными и лингвистическими методами // Известия Воронежского государственного педагогического университета, 2025. № 2. С. 139–147.
4. Громова А.В., Манянин П.А., Ростовская А.В. и др. Идентификация автора по текстам, создаваемым в ходе переписок в мессенджере: Методические рекомендации. М. : ЭКЦ МВД России, 2022. 108 с.
5. Караулов Ю.Н. Русский язык и языковая личность: монография / Ю. Н. Караулов ; отв. ред. Д. Н. Шмелев ; АН СССР, Отд-ние лит. и яз. М. : Наука, 1987. 261 с.
6. Литвинова Т.А. Проблемы стилеметрической идентификации субъекта идиолектной деятельности : дис. ... д-ра филол. наук.: 10.02.21 / Литвинова Татьяна Александровна. СПб., 2021. 430 с.
7. Литвинова Т.А., Швецова И.В. Корпус экспериментальных текстов, содержащих угрозу: состав, структура, общие частотные характеристики и области применения // Известия Воронежского государственного педагогического университета, 2024. № 3. С. 160–168.
8. Назарова Т.В., Манянин П.А., Громова А.В. и др. Диагностика половозрастных и индивидуально-личностных характеристик автора нерукописного текста: методические рекомендации. М.: ЭКЦ МВД России, 2014. 112 с.
9. Bell A. Language style as audience design // Language in Society, 1984. Vol. 13. P. 145–204. URL: https://www.researchgate.net/publication/316496165_Language_Style_as_Audience_ Design (дата обращения: 11.08.2026).
10. Biber D. Variation across Speech and Writing. Cambridge. Cambridge University Press. 1988. 299 p. URL: https://archive.org/details/variationacrosss0000bibe (дата обращения: 15.08.2026).
11. Brennan M., Greenstadt R. Practical attacks against authorship recognition techniques // Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence Conference. AAAI Press, 2009. P. 1521–1526. URL: https://cdn.aaai.org/ocs/i/iaai0020/257-3903-1-PB.pdf (дата обращения: 15.08.2026).
12. Litvinova T. Stylometrics features under domain shift: do they really "context-independent"? // Speech and Computer. SPECOM 2020. Lecture Notes in Computer Science. Vol. 12335. Cham: Springer, 2020. P. 279–290. URL: https://link.springer.com/chapter/10.1007/978-3-030-60276-5_28 (дата обращения: 15.08.2026).
13. Litvinova T.A., Litvinova O.A. Cross-modal authorship attribution in Russian texts // Quantitative Approaches to Universality and Individuality in Language / ed. by M. Yamazaki, H. Sanada, R. Kцhler et al. Berlin; Boston: De Gruyter Mouton, 2023. P. 73–88. URL: https://doi.org/10.1515/9783110763560-007 (дата обращения: 17.09.2026).
14. Stamatatos E. On the robustness of authorship attribution based on character n-gram features // Journal of Law and Policy. 2013. Vol. 21. P. 421–439. URL: https://www.researchgate. net/publication/303443999_On_the_robustness_of_authorship_attribution_based_on_character_n-gram_features (дата обращения: 15.08.2026).

References

1. Bahtin, M.M. (1979 *Problema rechevyh zhanrov* [The problem of speech genres]. Moscow, Art. P. 237-280. (In Russian).
2. Gak, V.G. (1998) *Yazykovye preobrazovaniya* [Linguistic transformations]. Moscow, The School of «Languages of Russian Culture». 768 p. (In Russian).
3. Gromova, A.V., Litvinova, T.A., Oshkukov, S.S., Shvecova, I.V. (2025) Kriminalisticheskoe issledovanie tekstov zhanra «pis'mo-ugroza» komp'yuternymi i lingvisticheskimi metodami [Forensic analysis of texts in the «threat letter» genre using computer and linguistic methods]. *Izvestija Voronezhskogo gosudarstvennogo pedagogicheskogo universiteta*. (2), 139–147. (In Russian).
4. Gromova, A.V., Manjanin, P.A., Rostovskaja, A.V. i dr. (2022). *Identifikacija avtora po tekstam, sozdavaemym v hode perepisok v messendzhere: Metodicheskie rekomendacii* [Identification of the author from texts created during correspondence in a messenger: Methodological recommendations]. Moscow, Forensic science centre of the Ministry of the Interior of Russian Federation, 108 p. (In Russian).
5. Karaulov, Yu.N. (1987) *Russkij yazyk i yazykovaya lichnost'* [Russian Language and Linguistic Personality]. Moscow, The science. 261 p. (In Russian).
6. Litvinova, T.A. (2021) *Problemy stilemetricheskoj identifikacii sub"ekta idiolektnoj deyatel'nosti*. Diss. d-ra filol. nauk. [Problems of stylistic identification of the subject of idiolectal activity]. Saint Petersburg. 430 p. (In Russian).
7. Litvinova, T.A., Shvecova, I.V. (2024) Korpus eksperimental'nyh tekstov, soderzhashchih ugrozu: sostav, struktura, obshchie chastotnye harakteristiki i oblasti primeneniya [A corpus of experimental texts containing threats: composition, structure, general frequency characteristics, and areas of application]. *Izvestija Voronezhskogo gosudarstvennogo pedagogicheskogo universiteta*. (3), 160–168. (In Russian).
8. Nazarova, T.V., Manjanin, P.A., Gromova, A.V. et al. (2014). *Diagnostika polovozrastnyh i individual'no-lichnostnyh harakteristik avtora nerukopisnogo teksta: metodicheskie rekomendacii* [Diagnostics of age,

gender and individual-personal characteristics of the author of a non-handwritten text: methodological recommendations]. Moscow, Forensic science centre of the Ministry of the Interior of Russian Federation. 112 p. (In Russian).

9. Bell, A. (1984). Language style as audience design. *Language in Society*. 13 (2), 145–204.

10. Biber, D. (1988). *Variation across Speech and Writing*. Cambridge. Cambridge University Press. 299 p.

11. Brennan, M., Greenstadt, R. (2009). Practical attacks against authorship recognition techniques. In: *Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence Conference*. AAAI Press, pp. 1521–1526.

12. Litvinova, T. (2020). Stylometrics Features Under Domain Shift: Do They Really "Context-Independent"?. *Speech and Computer. SPECOM 2020. Lecture Notes in Computer Science*. Vol. 12335. Cham, Springer, pp. 279–290.

13. Litvinova, T.A., Litvinova, O.A. (2023). Cross-modal authorship attribution in Russian texts. In: Yamazaki, M., Sanada, H., Kцhler, R. et al. (eds.) *Quantitative Approaches to Universality and Individuality in Language*. Boston, De Gruyter Mouton, pp. 73–88.

14. Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*. 21 (2), 421–439.

Поступила в редакцию 31.07.2026
Подписана в печать 30.09.2026

Original article
UDC 81'33
DOI: 10.47438/2309-7078_2026_3_239.

THE STABILITY OF THE IDIOLECT SIGNAL WHEN CHANGING THE COMMUNICATIVE ROLE (BASED ON THE MATERIAL OF TEXTS IN THE GENRE THREAT LETTER)

Irina V. Shvetsova[1, 2]

Forensic Science Centre of the Ministry of the Interior of Russian Federation[1]
Moscow, Russia
Voronezh State Pedagogical University[2]
Voronezh, Russia

[1]*Senior Expert of the Linguistic Expertise Section of the Digital and Speech Information Expertise Department*
[2]*Postgraduate Student of the Department of Russian language, Modern Russian and Foreign Literature, ORCID ID: 0009-0007-2496-6976, tel.: +7-499-745-79-45, e-mail: ishvetcova6@mvd.ru*

Abstract. The article assesses the stability of the idiolectal signal under controlled changes in the author's communicative role. The study was based on an experimental corpus consisting of 270 texts in the ''letter-threat'' genre, compiled at the Forensic science centre of the Ministry of the Interior of Russian Federation. The texts for this corpus were created by 90 authors in accordance with three specified communicative conditions. Depending on the realized communicative role, the study identifies the most reliable stylistic marker channels that indicate the stability of the idiolectal signal. The signal's manifestation is noted to be heterogeneous, particularly when transitioning to the third condition – to a role that implies a negative attitude towards a group of people united by national origin. The expediency of using a compact multi-component model combining symbolic, morphological, and structural-punctuational idiolectal features for the analysis of short texts in the ''threat letter'' genre is emphasized. At the same time, the boundaries of its applicability are indicated: in the absence of information about the authors, the similar texts identified by the model require subsequent qualitative expert analysis.

Key words: idiolect, communicative role, stylometry, variability of written speech, author identification, character *n*-grams, Hellinger transformation, permutation dispersion analysis, author-independent verification.

Cite as: Shvetsova, I.V. (2026). The stability of the idiolect signal when changing the communicative role (based on the material of texts in the genre threat letter). *Izvestia Voronezh State Pedagogical University*. (3), P. 239–247 (In Russ., abstract in Eng.). DOI: 10.47438/2309-7078_2026_3_239.

Received 31.07.2026
Accepted 30.09.2026