АПРОБАЦИЯ МЕТОДОВ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ ПРИ АТРИБУЦИИ ТЕКСТА В РАМКАХ СУДЕБНОГО АВТОРОВЕДЕНИЯ

Хоменко Анна Юрьевна

выпускница НГЛУ им. Н.А. Добролюбова, специалист;

студентка 1 курса магистратуры НИУ ВШЭ НФ по специальности «Компьютерная лингвистика»,

эксперт-лингвист, эксперт-фоноскопист,

специалист отдела лингвистических, фоноскопических и видеофонографических экспертиз НПО «Эксперт Союз» (сертификаты по специальностям «Исследования продуктов речевой деятельности», «Исследования голоса и звучащей речи»),

г. Нижний Новгород

e-mail: lili-th89@narod.ru

Аннотация. В статье идёт речь о поиске наиболее объективной методики атрибуции текста для судебного авторoведения. В рамках исследования была сделана попытка интеграции интерпретационных методов анализа текста с методами математической статистики и теории вероятности.

Annotation. The article deals with the search for the most objective method of text attribution for forensic expertise. In the study was made an attempt to integrate interpretation methods of text analysis with the methods of mathematical statistics and probability theory.

Ключевые слова: судебное авторoведение, атрибуция текста, анализ языковой личности автора текста, анализ квазисинонимичных лексем, методы математической статистики, методы теории вероятности.

***Key words**:* forensic expertise, attribution of the text, analysis of the linguistic identity of the text author, kvasiconvertible terms analysis, methods of mathematical statistics, methods of probability theory.

Проблемы атрибуции текстов в судебном автороведении на современном этапе его развития стоят очень остро. Компетентным государственным органам часто требуется выяснить принадлежность того или иного текста, причём тексты на экспертное исследование предоставляют очень разные: от учебных пособий объёмом более пятисот страниц до расписок, занимающих не более половины страницы. Это затрудняет выработку единой универсальной методики авторизации текстового материала. Эксперты при атрибуции текстов используют очень разные методики. Так, проблемы методического характера при производстве речеведческих экспертиз, к коим и относятся автороведческие исследования, зачастую становятся камнем преткновения, ведь основное, что использует любая экспертиза, связанная с продуктами речевой деятельности, - это интерпретация языкового знака; а эта интерпретация может сильно варьироваться у различных экспертов.

Спорами, касающимися различий в интерпретации одного и того же языкового знака различными экспертами и обусловлено стремление к оптимизации и объективизации методик анализа, применяющихся в автороведческих экспертизах.

На наш взгляд, именно методы математической статистики и основные постулаты теории вероятности могут помочь как в оптимизации, так и в объективации результатов исследования в области судебного автороведения.

Так, **цель** данной работы – определить, могут ли методы математической статистики и стилеметрического анализа успешно применяться в автороведении; можно ли на их основе создать универсальную безошибочную методику атрибуции текста *любого объёма*.

Материалы и методика исследования. Материалом для исследования послужили художественные тексты заведомо известных авторов, поскольку

целью работы является определение того, *работоспособна ли методика* стилеметрического анализа для текстов различных стилей и объёмов. Для определения этого необходимо было использовать уже авторизованный, «проверенный» материал, то есть тот материал, который объективно сможет показать плюсы и минусы методики. Так, материалом послужила следующая речевая продукция:

1. Тестовая выборка (ТВ) – выборка, на основе которой строилась исходная модель, это в терминологии судебной экспертизы – сравнительный материал. ТВ представляла собой тексты С.Д. Довлатова, представленные в Национальном корпусе русского языка (электронный ресурс Интернет: http://www.ruscorpora.ru/), за исключением текста «Наши» (1983 г.) (ЭТ1). Этот текст рассматривался как экспериментальный, то есть текст, у которого якобы не определён автор.

Таким образом, объём ТВ – 330709 слов.

2. ЭТ1 (экспериментальный текст №1) – текст С.Д. Довлатова «Наши» (1983 г.). Объём - 21230 слов.

В качестве второго экспериментального текста использовался не текст С.Д.Довлатова, а текст другого автора.

3. ЭТ2 (экспериментальный текст №2) – текст В.П. Астафьева «Затеси» (1999 г.). Объём - 15168 слов.

Тексты ЭТ1 и ЭТ2 выбраны для апробации методики, поскольку они, с одной стороны, имеют достаточно важные сходные художественные и экстралингвистические характеристики (близкий к публицистическому стиль написания, высокий уровень автобиографичности текстов, тематика – описание советской действительности, время действия – советский период, обширная аудитория читателей), с другой – принадлежат разным авторам, имеющим различные идиостили.

В качестве основы для методики анализа было положено исследование Е.С. Родионовой «Лингвистические методы атрибуции и датировки литературных произведение (К проблеме «Мольер - Корнель»)» [1]. Методика Е.С Родионовой

была совмещена с методикой анализа языковой личности по Ю.Н. Караулову [2], методикой квантитативного анализа незнаменательных и стилистически немаркированных лексем и квазисинонимов А.Н. Баранова [3] и некоторыми постулатами теории вероятности.

Апробация методики.

I. Построение атрибуционных гипотез об авторстве спорных текстов ЭТ1 и ЭТ2:

$H_{0/1}$ – автор ТВ и ЭТ1 – одно лицо, то есть автор ТВ и ЭТ1 – С.Д. Довлатов (по закону транзитивности: если автор ТВ – С.Д. Довлатов, а автор ЭТ1 и ТВ – одно лицо, то автор ЭТ1 – тоже С.Д. Довлатов).

$H_{1/1}$ – авторы ТВ и ЭТ1 – разные лица, то есть автор ЭТ1 не С.Д. Довлатов (если автор ТВ – С.Д. Довлатов, а авторы ЭТ1 и ТВ – разные лица, то автор ЭТ1 – не С.Д. Довлатов).

$H_{0/2}$ – автор ТВ и ЭТ2 – одно лицо, то есть автор ТВ и ЭТ2 С.Д. Довлатов (по закону транзитивности: если автор ТВ – С.Д. Довлатов, а автор ЭТ1 и ТВ – одно лицо, то автор ЭТ2 – тоже С.Д. Довлатов).

$H_{1/2}$ – авторы ТВ и ЭТ2 – разные лица, то есть автор ЭТ2 не С.Д. Довлатов (если автор ТВ – С.Д. Довлатов, а авторы ЭТ2 и ТВ – разные лица, то автор ЭТ2 – не С.Д. Довлатов).

II. Анализ языковой личности (ЯЛ).

1. Анализ ЯЛ автора ТВ, то есть ЯЛ С.Д. Довлатова.

Анализ ЯЛ необходим в данной работе для того, чтобы определить параметры для построения математических моделей сравнительного материала и спорного текста.

В ходе анализа были выделены следующие релевантные для исследования характеристики (под релевантными понимаются такие фрагменты ЯЛ, которые можно вербализовать в виде одной лексемы или одной синтаксической особенности; более того, эта лексема или синтаксическая особенность не должна быть чересчур индивидуально маркированной, то есть в анализ нельзя включать окказионализмы, авторские неологизмы и пр.; все характеристики должны быть, с

одной стороны, общеупотребимыми, с другой – встречаться в произведениях автора, идиостиль которого анализируется, и отражать его взгляд на мир):

1) *вербально-семантический* уровень: местоимения «*я*», «*мы*», «*ты*», «*они*». Выделены, исходя из наличия соотношения в прозе Довлатова проблемы субъекта говорения и субъекта действия, то есть автора и героя;

2) *лингвокогнитивный* уровень*:*

а) «плохо», «хорошо» - лексемы, маркирующие отношение к действительности;

б) «*тёмный», «белый», «светлый*» - лексемы, имплицитно маркирующие отношение к действительности и создающие её образ;

в) «город», «чемодан» - лексемы, вербализующие значимые для Довлатова концептуальные сущности. Образ города присутствует во многих произведениях Довлатова, являясь символом определённого типа сознания. Чемодан же является у Довлатова символом дороги, перемещения;

3) *мотивационный* уровень: *«пусть», «бы», «так», «пожалуй», «ладно», «ну»* - модификаторы субъективной модальности допущения. Данные экспликаторы модальности взяты, поскольку именно они, исходя из исследований Топтыгиной Е.Н. [4], являются «центральными модификаторами» семантического поля допущения. Перечисленные частицы, действительно, чаще других формальных материальных экспликаторов субъективной модальности допущения встречаются в текстах Довлатова, представленных в НКРЯ.

К перечисленным параметрам был добавлен ещё ряд. Так, анализировались также репрезентанты субъективной модальности удивления «*ах*», «*разве*» и «*неужели*». Эти лексемы выделены как одни из центральных, эксплицирующих удивление. Добавлены также лексемы, репрезентирующие модальность ограничения: «*только*», «*лишь*», «*почти*», - и модальность возражения «*всё-таки*».

Параметрами для построения математических моделей являются также и лексемы, вербализующие некоторые фрагменты ЯЛ В.П. Астафьева. Наличие этих параметров необходимо, поскольку принимается постулат о непохожести

идиостилей различных авторов. Соответственно, нужно выявить, действительно ли элементы, отражающие особенности видения мира одним автором, являются настолько показательными, что посредством их стилеметрического анализа можно отличить тексты этого автора от текстов другого. Гипотетически, именно наличие элементов, репрезентирующих ЯЛ разных авторов, должно привести к тому, что характеристики ЯЛ одного будут значимы в квантитативном отношении настолько, что помогут выявить принадлежность того или иного текста именно этому автору. Как следствие, сходство ТВ и ЭТ1 и различие ТВ и ЭТ2 должны доказать успешность использования предлагаемой методики для атрибуции письменных текстов.

Из перечисленного выше понятно, что необходимо обзорно проанализировать не только ЯЛ автора ТВ и ЭТ1, но и ЯЛ автора ЭТ2, то есть В.П. Астафьева.

2. В ходе анализа ЯЛ В.П. Астафьева были выделены следующие релевантные для исследования параметры:

1) *вербально-семантический* уровень: использование сочинительных союзов «*а», «и», «но*» в начале предложения;

2) *лингвокогнитивный* уровень*:*

а) «грусть», «грустный», «грустно» - лексемы, имплицитно («*грусть*», «*грустный*») и эксплицитно («*грустно*») маркирующие отношение к действительности. Эти лексемы отражают в том числе аксиологические оценки в прозе Астафьева;

б) «детство», «родина» - лексемы, вербализующие значимые для Астафьева концептуальные смыслы;

в) «молчание», «молчаливый» - лексемы, имплицитно маркирующие отношение к действительности и создающие её образ;

3) *мотивационный* уровень: «*видно*» - лексема, эксплицирующая модальность допущения, неуверенности, предположения.

Видим, что при анализе структуры ЯЛ Довлатова С.Д. и Астафьева В.П. в работе выделяются сходные фрагменты ЯЛ (на каждом уровне реализована

попытка взять те лексемы, которые репрезентируют сходные зоны ЯЛ: лингвокогнитивный уровень – оценка действительности, образ действительности; мотивационный уровень – модальность допущения, неуверенности). Это должно повысить качество моделей.

В общей сложности было взято 35 параметров для исходных моделей.

III. Квантитативные и стилеметрические преобразования данных, полученных в результате анализа ЯЛ.

1) Определение выборочных частот, то есть был произведён механический подсчёт того, сколько раз параметр реализуется в ТВ, ЭТ1, ЭТ2.

2) Определение средневыборочной частоты каждого параметра по формуле (1).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_i + \ldots + x_n)$$

формула 1, -

где x_i - *i*-й элемент выборки, *n* – объём выборки.

3) Определение отклонения выборочных частоты от средневыборочной частоты (среднеквадратического отклонения рассчитывается по формуле (2)).

$$s = \sqrt{\frac{n}{n-1}\sigma^2} = \sqrt{\frac{1}{n-1}\sum_{i-1}^{n}(x_i - \bar{x})^2}$$

формула 2, -

где σ^2 — дисперсия; x_i — *i*-й элемент выборки; *n*— объём выборки; $\bar{x}$ — среднее арифметическое выборки (средневыборочная частота).

4) Поиск вероятной ошибки в определении средней частоты по формуле (3) (для α – 0,2 и вероятности 0,8 при (n – 1) степеней свободы (35-1=34): **t = 1,3070**).

$$L = \frac{t\sigma}{\sqrt{k}}$$

формула 3, -

где *t* - табличный коэффициент (t-критерий Стьюдента); σ-среднеквадратичное отклонение; *k* - объём выборки.

Для ТВ ошибка составляет 0,002272751.

Для ЭТ1 - 0,008969957

Для ЭТ2 - 0,010611997.

Естественно, для каждого параметра значимость этой ошибки различна. Тем не менее, можно говорить о том, что для большинства лексем (для синтаксических особенностей, как то: сочинительные союзы в начале предложения) ТВ ошибка не очень велика, а вот для ЭТ1 и ЭТ2 ошибка ощутима. Поэтому в работе учитывается эта ошибка.

5) Определение релевантных параметров для конечных моделей. Определяются по t-критерию Стьюдента (4). Уровень значимости α – 0,2. Критическое значение – в таблице пересечение уровня степеней свободы (количества параметров - 1) и вероятности 0,8.

$$t = \frac{|\overline{x_1} - \overline{x_2}|}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}}$$

формула 4, -

где $\overline{x_1}$, x_2 - средние арифметические; σ_1^2, σ_2^2 - стандартное отклонение; n_1, n_2 - объёмы выборок.

По результатам исследования выделены следующие релевантные для моделей параметры:

- модель ТВ и ЭТ1: *грусть, сочинительный союз "но" в начале предложения, пусть, грустный, разве, ах, белый, неужели;*

- модель ТВ и ЭТ2: *молчаливый, грусть, молчание, сочинительный союз "и" в начале предложения, сочинительный союз "а" в начале предложения, сочинительный союз "но" в начале предложения, всё-таки, видно.*

Релевантными для построения моделей в настоящей работе считаются параметры, числовые показатели которых наиболее близки к табличному значению t-критерия (1, 3070).

Важно, что ни одно значение параметра не превысило значение t-статистики. Для исследования это означает, что полученные результаты будет иметь точность менее изначально заявленной, то есть менее 80%. Интересно, что значения параметров для ТВ и ЭТ2 ближе к t-критерию, чем значения для ТВ и ЭТ1. Это говорит о том, что характеристики для сравнения ТВ и ЭТ2 выбраны более удачно, чем для сравнения ТВ и ЭТ1.

IV. Переход от реальных объектов к их математическим моделям (как для текстов-образцов (ТВ), так и для спорных текстов(ЭТ1), (ЭТ2)), то есть описание выделенных в ходе предшествующего анализа параметров с помощью условной сигнатуры. Формирование матриц данных.

Математические модели и матрицы данных для ТВ и ЭТ1 и ТВ и ЭТ2 представлены в Таблицах 1 и 2.

Таблица 1.

Математическая модель ТВ и ЭТ1

Параметр	**Класс**	
	$\Omega_{ТВ}$	**$\Omega_{ЭТ1}$**
	$/x_i \pm s_i/$	$/x_i \pm s_i/$
$X_{1/1}$	0,019	0
$X_{2/1}$	0,003	0
$X_{3/1}$	0,124	0
$X_{4/1}$	0,07	0
$X_{5/1}$	0,141	0,014
$X_{6/1}$	0,068	0,007
$X_{7/1}$	0,19	0,021
$X_{8/1}$	0,061	0,007

Где:

$X_{1/1}$- грусть

$X_{2/1}$- сочинительный союз "но" в начале предложения

$X_{3/1}$- пусть

$X_{4/1}$- грустный

$X_{5/1}$- разве

$X_{6/1}$- ах

$X_{7/1}$- белый

$X_{8/1}$- неужели

$\Omega_{ТВ}$ – модель ТВ;

$\Omega_{ЭТ1}$ – модель ЭТ1;

x_i – средневыборочная частота с учётом ошибки про вычислении средневыборочной частоты;

s_i – среднеквадратическое отклонение.

Таблица 2.

Математическая модель ТВ и ЭТ2

Параметр	**Класс**	
	$\Omega_{ТВ}$	**$\Omega_{ЭТ2}$**
	$/x_i \pm s_i/$	**$/x_i \pm s_i/$**
$X_{1/2}$	0,014	0,000
$X_{2/2}$	0,019	0,000
$X_{3/2}$	0,044	0,000
$X_{4/2}$	0,023	0,000
$X_{5/2}$	0,023	0,000
$X_{6/2}$	0,035	0,000
$X_{7/2}$	0,004	0,000
$X_{8/2}$	0,424	0,008
$X_{9/2}$	0,117	0,008

Где:

$X_{1/2}$- молчаливый

$X_{2/2}$- грусть

$X_{3/2}$- молчание

$X_{4/2}$- сочинительный союз "и" в начале предложения

$X_{5/2}$- грустно

$X_{6/2}$- сочинительный союз "а" в начале предложения

$X_{7/2}$- сочинительный союз "но" в начале предложения

$X_{8/2}$- всё-таки

$X_{9/2}$- видно

$\Omega_{ТВ}$ – модель ТВ;

$\Omega_{ЭТ2}$ – модель ЭТ2;

x_i – средневыборочная частота с учётом ошибки про вычислении средневыборочной частоты;

s_i – среднеквадратическое отклонение.

V. Сравнение моделей ТВ и ЭТ1 и ТВ и ЭТ2, соответственно. Для сравнения моделей используется коэффициент корреляции между однородными параметрами модели, определяемый по формуле (5).

$$r_{XY} = \frac{\mathrm{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}},$$

формула 5, -

где $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ - средние значения выборок.

Этот коэффициент показывает, насколько близки две модели. Чем ближе значение этого коэффициента к 1, тем более сходны модели в качественном отношении, что говорит и о близости характеристик текстов.

Коэффициент корреляции между числовыми значениями матриц ТВ и ЭТ1 равен 0,783448911306154.

Коэффициент достаточно близок к единице, что говорит о сходстве качественных характеристик моделей ТВ и ЭТ1, то есть идиостиль ТВ схож с идиостилем ЭТ1.

Это позволяет сделать вывод о том, что идиостиль текстов ТВ (текстов заведомо известного автора – С.Д. Довлатова) схож с идиостилем спорного текста «Наши» (ЭТ1) (текста, автор которого по условиям эксперимента неизвестен) настолько, что можно говорить об атрибуции текста «Наши» как текста, принадлежащего перу С.Д. Довлатова. Однако вероятность принадлежности произведения «Наши», исходя из исследования, ниже 80% (вывод сделан, исходя

из того, что числовые значения параметров конечной модели, не превышают критического значения t-критерия Стьюдента, просчитанного для восьмидесятипроцентной вероятности принадлежности текста определённому автору).

Коэффициент корреляции между числовыми значениями матриц ТВ и ЭТ2 равен 0,81432738421146.

Коэффициент достаточно близок к единице, что говорит о близости качественных характеристик моделей ТВ и ЭТ2, то есть по результатам исследования идиостиль ТВ схож с идиостилем ЭТ2.

Это говорит о том, что идиостиль текстов ТВ схож с идиостилем спорного текста «Затеси» (ЭТ2) настолько, что можно говорить об атрибуции текста «Затеси» как текста, принадлежащего перу С.Д. Довлатова. Однако вероятность принадлежности произведения «Затеси», исходя из исследования, ниже 80%

Так, анализ показал, что математическая модель текста «Затеси» несколько ближе к модели текстов С.Д. Довлатова, взятым в качестве тестовой выборки, то есть материала для построения исходной, образцовой, сравнительной модели, чем математическая модель произведения «Наши».

VI. Выводы о том, какие из выстроенных в начале исследования гипотез нашли своё подтверждение.

Подтвердились следующие гипотезы:

$H_{0/1}$ – автор ТВ и ЭТ1 – одно лицо, то есть автор ТВ и ЭТ1 - С.Д. Довлатов (по закону транзитивности: если автор ТВ – С.Д. Довлатов, а автор ЭТ1 и ТВ – одно лицо, то автор ЭТ1 – тоже С.Д. Довлатов).

$H_{0/2}$ – автор ТВ и ЭТ2 – одно лицо, то есть автор ТВ и ЭТ2 - С.Д. Довлатов (по закону транзитивности: если автор ТВ – С.Д. Довлатов, а автор ЭТ1 и ТВ – одно лицо, то автор ЭТ2 – тоже С.Д. Довлатов).

Выводы о соответствии полученных результатов действительности. Из раздела VI апробации методики видно, что своё подтверждение по результатам исследования нашла как гипотеза, которая соответствует реальной действительности ($H_{0/1}$ - автор произведения «Наши» - С.Д. Довлатов; и это

правда), так и гипотеза, которая не имеет ничего общего с этой действительностью ($H_{0/2}$ – автор произведения «Затеси» - С.Д. Довлатов, что на самом деле неправда, поскольку автором «Затесей» является В.П. Астафьев).

Анализ достоверности методики. Получается, что методика, *используемая в современном автороведении, в том числе и судебном*, работает, так сказать, лишь наполовину, то есть из двух случаев в одном она работает, а во втором – нет. Грубо, методика работает в 50% случаев, то есть вероятность её срабатывания ½ или 50% (если оценить вероятность (p) «срабатывания» методики более точно, то получится, что она укладывается в интервал [0; 0,552786]). То есть по результатам исследования напрашивается вывод о том, что это нерабочая методика. На первый взгляд применить эту методику, значит почти то же, что выбрать наугад из двух вариантов ответа (да или нет) на вопрос, является ли автором определённого текста конкретное лицо или нет. Но это не совсем так, ведь в работе Е.С. Родионовой [1] была доказана работоспособность методики.

Что же необходимо для успешной работы этой методики?

Анализ качества выборок и качества моделей. Рекомендации по улучшению рабочих характеристик методики. На наш взгляд, причин того, что методика в реальных условиях не работает, несколько.

В методике, предложенной в настоящей работе для атрибуции текста, сделана попытка совмещения нескольких методологий, а именно: анализа языковой личности автора текста и стилеметрического исследования текста. Причём, анализ языковой личности использовался как инструмент для подбора параметров для квантитативного исследования. В данном случае нет никакого противоречия с методикой судебного автороведения, основанной на методах стилеметрии, поскольку для подбора параметров как таковых рекомендаций нет. Для этого подбора можно использовать любые подходящие алгоритмы, которые отвечают постулату о том, что любому автору присущ свой собственный стиль, то есть такие, которые выявляют индивидуальные особенности произведения.

Тем не менее, рассматриваемая в настоящей работе методика изначально была основана на теории распознавания образов и апробирована с

использованием синтаксических и морфологических параметров в широком смысле. В работе Родионовой Е.С. «Лингвистические методы атрибуции и датировки литературных произведение (К проблеме «Мольер - Корнель»)», например, исследуются такие параметры, как число элементарных предложений, число сочинительных предложений, число спрягаемых форм глагола, число подлежащий и пр. [1]. Всё это характеристики достаточно обобщённого вида, речь не идёт о конкретных словах, эксплицирующих какой-то класс, исследуются целые синтаксические и морфологические классы. В данной же работе в качестве параметров выбраны определённые репрезентанты, причём описывающие даже не целый класс, а фрагменты этого класса, как то фрагмент языковой личности на вербально-семантическом уровне (личные местоимения «*я», «ты», «они*»), на лингвокогнитивном уровне («*чемодан*», «*город*»), на мотивационном уровне («*пожалуй*», «*ладно*», «*ну*»).

Таким образом, проблема методологии отбора параметров отразилась на качестве первичной выборке, то есть на том, какие параметры были взяты за основу идентификации. Более того, после сопоставления качественных характеристик параметров, становится понятно, что и количество параметров для исследования было недостаточным. При апробации методики на материале пьес Мольера и Корнеля, а также на текстах Пушкина число параметров было: для исследований по Мольеру и Корнелю – 51, по Пушкину – 49. В данной работе исследовалось 35 параметров. Кажется, что цифры сопоставимы, но, как показала практика, 35 узкоспециализированных характеристик оказалось недостаточно для построения объективных текстовых моделей.

Соответственно, для использования методики в условиях реальной действительности для текстов большого объёма можно дать следующие рекомендации:

1) число параметров для идентификации автора по письменному речевому произведению должно быть не менее 45 – 50 единиц;

2) параметры должны представлять собой обширные синтаксические и морфологические классы, например, большинство

материальных репрезентантов субъективной модальности (вводные слова, модальные частицы, междометия, конструкции с именительным представления).

3) отбор параметров должен происходить на основе более глубокого анализа языковой личности автора текста-образца, причём в большем объёма именно на мотивационном уровне (возможно, также на вербально-семантическом).

Для анализа текстов малого объёма в рамках рассматриваемой методики по результатам исследования можно дать иные рекомендации. Как показало исследование, для конечных моделей параметры выбираются с помощью расчётов по формуле (4). В этой формуле большое значение имеет числовое значение s (среднеквадратического отклонения). Чем меньше это отклонение и его пропорциональное соотношение со средневыборочной частотой, тем больше подходит параметр.

Такая тенденция приводит к тому, что в параметры для модели попадают характеристики, которые вообще не встречаются в тексте. Само по себе это не критично, поскольку теоретически реализация какого-то параметра может присутствовать в одном тексте и отсутствовать в другом, что и будет дифференцировать эти тексты. Тем не менее, в нашем случае это не сработало. То, что в модель были включены параметры со столь малыми случаями реализации, привело к непреодолению их числовыми значениями критического порога, заданного t-критерием Стьюдента.

Сказанное выше позволяет говорить о том, что даже выбранные конечные критерии не смогли обеспечить первоначально заявленную восьмидесятипроцентную вероятность отнесения того или иного текста к перу определённого автора; она оказалась гораздо ниже при тех условиях, которые были реализованы в эксперименте.

Таким образом, можно сделать вывод о том, что для текстов такого объёма, который присущ ТВ, ЭТ1, ЭТ2, при выборе параметров на основе предложенного

алгоритма (на основе анализа фрагментов языковой личности на всех уровнях) методика работает некорректно.

Тем не менее, эта методика (с подбором критериев на основе анализа всех уровней языковой личности автора) отлично подойдёт для идентификации автора небольшого по объёму (не более 10 страниц) письменного произведения. То есть при условии, что спорный текст будет около 10 страниц, текст-эталон также будет иметь приблизительно такой же объём, эти тексты будут схожи в стилистическом отношении (хотя бы с точки зрения функциональных стилей), у средневыборочных частот выбранных параметров не будет такого большого разброса (не будет ситуации, что разброс составляет от 10281 реализации до 0 реализаций, как в настоящем исследовании), сами частоты между собой будут сопоставимыми, соответственно, соотносимыми будут и их среднеквадратические отклонения, предложенная методика будет вполне подходящей для атрибуции данных текстов.

Значит, для применения указанной методики без внесения в саму методику изменений можно дать следующие рекомендации:

1) объёмы текста-образца и спорного текста должны быть близки и не должны превышать 10 страниц печатного текста каждый;

2) тексты должны быть схожи с точки зрения функциональных стилей;

3) методику можно дополнить вычленением из двух текстов (эталонного текста, то есть сравнительного образца, и спорного текста), так называемых, квазисинонимичных лексем.

Важно, что методика, предложенная в работе, была лишь создана на основе уже имеющейся методики квантитативной лингвистики по атрибуции текстов [1] с применением постулатов Н.А. Морозова, А.Н. Баранова [3], иных исследователей, занимающихся проблемами автороведения, а также с использованием методов анализа языковой личности автора текста [2].

Основным её концептуальным отличием от методики Е.С. Родионовой, например, является замена итерационного алгоритма на конечном этапе

атрибуции на вычисление коэффициента корреляции между числовыми характеристиками моделей. Это было сделано не случайно. В работе изначально была сделана попытка минимизировать оперативные затраты, то есть оптимизировать процесс на этапе соотнесение какого-либо текста с автором. Для решения задачи оптимизации итерационный алгоритм был заменен на операцию вычисления коэффициента корреляции между параметрами математической модели.

После того как оказалось, что методика работает неудовлетворительно, была совершена попытка применить именно итерационный алгоритм сравнения моделей, дабы определить, не кроется ли некорректность методики именно в замене этого итерационного алгоритма на иной.

Тем не менее, исследование показало, что итерационный алгоритм с пересчётом весов в данном случае также не может считаться релевантным, поскольку уже на нулевой итерации видно, что вновь ни одно числовое значение параметра не превышает критического значения t-критерия Стьюдента. Для того чтобы применить итерационный алгоритм, необходимо было бы снова принять ближайшее к этому значению в качестве эталонного, то есть принимать эмпирически высчитанный критерий, который бы не был ничем обоснован. Это ещё более понизило качество того решения, которое было бы принято.

Исходя из этого, можно дать следующие рекомендации:

1) попытка использования итерационного алгоритма на этапе сравнения конечных математических моделей возможна лишь при условии преодоления числовыми значениями параметров критического значения на этапе определения релевантных параметров для дальнейшей идентификации; при этом логичным было бы использовать итерационный алгоритм в рамках его применения как инструмента для построения дерева решений;

2) это в свою очередь возможно лишь при изменении изначального набора атрибутов моделей либо объёма спорного и образцового текстов, то есть при условии соблюдения рекомендаций, данных выше;

3) для разработки V этапа методики имеет смысл попробовать также метод наименьшей энтропии (далее также должно быть построено дерево решений) или линейной регрессии.

Таким образом, видим, что методика может считаться рабочей для определенного объёма текстового материала при условии соблюдения рекомендаций именно для этого объёма текста. Тем не менее, универсальной эту методику назвать нельзя.

СПИСОК ЛИТЕРАТУРЫ

1. **Родионова Е.С.** Лингвистические методы атрибуции и датировки литературных произведение (К проблеме «Мольер - Корнель»). Автореферат дис. канд. филологич. Наук. URL: http://epir.ru/pragmat!/projects/corneille/files/autoreferat.pdf.
2. **Караулов Ю.Н.** Русский язык и языковая личность. /Отв. ред. член-кор. Д.Н. Шмелев. – М.: Наука, 1987. – 263 с.
3. **Баранов А.Н**. Введение в прикладную лингвистику: Учебное пособие, 2001. URL: http://rudocs.exdat.com/docs/index-418941.html.
4. **Топтыгина Е.Н.** Средства выражения субъективно-модальных значений предположения и допущения в современном русском языке, 2003. URL: http://www.dissercat.com/content/sredstva-vyrazheniya-subektivno-modalnykh-znachenii-predpolozheniya-i-dopushcheniya-v-sovrem#ixzz2N3EDjXgv