# Индивидуально-авторские характеристики стиля (прикладной аспект)

***Поротикова Яна Вадимовна***

студентка 4 курса (бакалавриат)
Национальный исследовательский университет Высшая школа экономики

Работа, посвященная исследованию проблемы атрибуции текстового материала и разработке методики идентификации автора письменного текста на материале Интернет – ресурса, на современном этапе развития науки и техники является актуальной, так как именно с решением автороведческих задач связан на сегодняшний день целый ряд лингвоэкспертных исследований.

Исследование, описанное в настоящей работе, проведено на следующем материале:

- тексты, используемые для заполнения контента Интернет-сайта top10.nn.ru (тематика текстов: ведущий на свадьбу);
- короткие публицистические тексты, размещённые в сети Интернет на ресурсе Живой Журнал (Электронный ресурс: bakushinskaya.livejournal.com) (тематика текстов: рассказы о жизни);
- короткие тексты, находящиеся в открытом доступе, опубликованные на странице девушки в социальной сети «ВКонтакте» (Электронный ресурс: https://vk.com/nadezhda1234 ) (тематика текстов: танцы, преподавание)

Объектом исследования являются выбранные массивы текстов. Предмет исследования – основные идентификационные особенности письменной языковой личности авторов трёх текстовых блоков, указанных выше.

Цель предпринятого исследования – атрибуция спорных текстов на основе авторизованного текстового материала с использованием интерпретационных и автоматических методов анализа. Для достижения поставленной цели необходимо решить ряд задач:

1. Изучение теоретической базы;
2. Анализ массивов текстов;
3. Отбор параметров в авторизованных текстах, подходящих в качестве материала для анализа речевых компетенций различных авторов спорных текстов;

4. Анализ текстов с помощью чисто лингвистических методик с целью подготовки материала для интерпретационного и автоматического исследования;
5. Автоматическая обработка текстового материала с целью получения объективных стилостатистических данных;
6. Интеграция результатов автоматического и интерпретационного исследований с целью получения объективной модели письменной языковой личности авторов различных текстов;
7. Оценка полученных результатов с целью определения атрибуции неавторизованных текстов.

В ходе исследования были применены следующие методы: метод сплошной выборки, метод интерпретации, собственно лингвистический метод анализа, квантитативный метод, метод стилеметрического анализа, методы математической статистики и теории вероятности.

Практическая значимость обусловлена возможностью использования результатов исследования в лингвистических экспертизах, которые имеют цель определить авторство каких-либо текстов.

Методика атрибуции текстового материала и разработки методики идентификации автора письменного текста, предложенная в данной работе, основывается на совмещении двух методик: лингвистической и машинной.

В более детализированном виде последовательность анализа будет следующая:

- Построение атрибуционных гипотез об авторстве спорных текстов (тех текстов, автора которых необходимо определить) (H0, H1)
- Анализ письменной языковой личности автора текстов-образцов (тех текстов, автор которых заведомо известен) и письменной языковой личности автора спорных текстов (тех текстов, автора которых необходимо определить).
- применение методики количественного анализа квазисинонимичных лексем, описанная А.Н.Барановым в его работе «Введение в прикладную лингвистику» во второй главе «Оптимизация когнитивной функции языка» в разделе «Авторизация текста». Сущность методики заключается в том, чтобы выявлять авторские предпочтения при выборе из групп квазисинонимов — близких по значению слов или устойчивых словосочетаний (фразеологизмов). Однако данная методика будет иметь следующие дополнения:
- Подготовка для машинного анализа: описание массивов текстов, определение объёма и количества выборок из массивов текстов. Описание количественных показателей этих выборок.

- Частеречная разметка слов в массивах. Такая разметка осуществлялась с помощью программы Mystem - морфологический анализатор русского языка с поддержкой снятия морфологической неоднозначности, разработанный Ильёй Сегаловичем в компании «Яндекс»1.
- Определение выборочных частот. Механический подсчёт того, сколько раз параметр реализуется в каждой выборке. Определение средневыборочной частоты (1).

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

формула 1, -

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

- Определение отклонения выборочных частоты от средневыборочной частоты (определение среднеквадратического отклонения (2), (3)).

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

формула 2, -

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

формула 3, -

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

- Определение релевантных параметров для дальнейшей идентификации. Определяется с помощью квадратичного отклонения разности двух средних частот (4).

$$\varepsilon^{1,2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

формула 4, -

$\sigma_1^2$, $\sigma_2^2$- дисперсии двух средних выборок; $n_1$, $n_2$ - объёмы выборок.

Полученная величина с помощью данной формулы сравнивается с разностью двух средних частот, и если окажется, что данная разность более, чем в 3 раза превосходит её квадратичное отклонение σ, значит данный

1 Электронный ресурс: https://tech.yandex.ru/mystem/

параметр имеет существенное расхождение частот и для дальнейшего анализа и построения модели он отвергается.

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

- Сравнение двух моделей: модели текстов-образов, описывающей некоторые закономерности языковой личности заведомо известного автора, и модели спорных текстов, описывающей некоторые закономерности языковой личности неизвестного автора. Для сравнения моделей используется коэффициент корреляции между однородными параметрами (5). Этот коэффициент показывает, насколько близки две модели. Чем ближе значение этого коэффициента к 1, тем более близки модели в качественном отношении.

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

формула 5, -

где $\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$ , $\bar{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t$ - средние значения выборок.

В ходе экспериментального исследования было доказано авторство спорных текстов по выбранной методике. Так, методика, включающая лингвистический и автоматический анализ оправдала себя, так как 1) интерпретационная часть алгоритма (анализ письменной языковой личности автора текста с помощью лингвистических методик), с одной стороны, может быть использована в качестве способа отбора параметров для математической модели письменной языковой личности автора, а с другой – с её помощью можно проверить правильность выводов машинного анализа при условии отбора параметров с помощью других методик или использовать её как первичный метод атрибуции с последующей проверкой результатов с помощью стилостатистических методик; 2) машинная составляющая алгоритма всегда объективирует результат исследования.

В процессе исследования был решён ряд концептуально значимых для работы задач:

- была создана текстовая база, собрана коллекция текстов, подходящая в качестве материала для анализа речевых компетенций различных авторов;

- на основе анализа специальной литературы был выработан свой, аутентичный алгоритм анализа письменного текстового материала,

основанный на уже имеющихся научных изысканиях и включающий интерпретационное и автоматическое исследования;

- была проведена апробация выработанной методики на разнородном материале;
- была произведена интеграция результатов автоматического и интерпретационного исследований с целью получения объективной модели письменной языковой личности авторов различных текстов.
- на основе анализа с помощью лингвистического и машинного анализа были установлены авторы спорных текстов.

Полученные при анализе с помощью описанного алгоритма результаты показали релевантность каждого из этапов методики. Таким образом, при установлении авторства спорных или неавторизованных текстов невозможно обойтись как без лингвистических методик, так и без объективации этих результатов с помощью машинного метода.

**Литература:**


1. Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. – М.: Эдиториал УРСС, 2001.
2. Теория вероятностей и математическая статистика : учеб- ник для бакалавров / А. М. Попов, В. Н. Сотников ; под ред. проф. А. М. Попова. — М. : Издательство Юрайт, 2011. — 440 с. — Серия : Бакалавр.