**Теория**

Атрибуция текста – исследование текста с целью установления авторства или получения

каких-либо сведений об авторе и условиях создания текстового документа. Задачи атрибуции

можно разделить на идентификационные и диагностические

Методы атрибуции позволяют исследовать текст на пяти уровнях: пунктуационном, орфографическом, синтаксическом, лексико-фразеологическом, стилистическом.

Пунктуационный уровень помогает выявить особенности употребления автором знаков

препинания, характерные ошибки.

Орфографический уровень выявляет характерные ошибки в написании слов.

Синтаксический уровень позволяет определить особенности построения предложений,

предпочтение тех или иных языковых конструкций, употребление времен, активного или

пассивного залога, порядок слов, характерные синтаксические ошибки.

Лексико-фразеологический уровень определяет словарный запас автора, особенности использования слов и выражений, склонность к употреблению редких и иностранных слов,

диалектизмов, архаизмов, неологизмов, профессионализмов, арготизмов, навыки употребления фразеологизмов, пословиц, поговорок, «крылатых выражений» и т. д.

Стилистический уровень позволяет определить жанр, общую структуру текста, для литературных произведений – сюжет, характерные изобразительные средства (метафора, ирония,

аллегория, гипербола, сравнение), стилистические фигуры (градация, антитеза, риторический

вопрос и т. д.), другие характерные речевые приемы.

Под «авторским стилем» обычно понимаются последние три уровня. Анализ именно синтаксического, лексико-фразеологического и стилистического уровней представляет наибольший интерес и наибольшую сложность

A diagram of a company

Description automatically generated

В качестве критерия близости двух текстов вводится так или иначе вычисляемое «рас-

стояние» между соответствующими векторами. В простейшем случае можно представить

наборы параметров как обычные векторы в n-мерном декартовом пространстве, выходящие

из начала координат, и считать расстоянием между текстами обычное декартово расстояние

между концами соответствующих им векторов. Есть множество других вариантов. Именно

«расстояние» является в итоге интегральной характеристикой различия текстов. Оно определенным образом нормируется, и тексты, для которых расстояние велико, считаются с высокой вероятностью относящимися к разным авторам. Таким образом, чтобы сопоставить авторство двух текстов, достаточно вычислить для них параметры и определить расстояние.

Чтобы сопоставить текст с автором, сравниваются векторы параметров автора и данного тек-

ста, т. е. фактически снова сравниваются два текста – текст с заведомо известным автором

(эталонный текст) и текст, авторство которого требуется установить, подтвердить или опровергнуть (спорный текст). Можно также составить векторы формальных параметров, различающие не конкретных авторов (или их группы), а выделяющие определенные характеристики авторов (например, образовательный уровень).

В большинстве случаев в качестве характеризующих параметров текста выбираются те

или иные его статистические характеристики: количество использования определенных частей речи, некоторых конкретных слов, знаков препинания, фразеологизмов, архаизмов, редких и иностранных слов, количество и длина предложений (измеренная в словах, слогах,

знаках), объем словаря, количество полнозначных и служебных слов, средняя длина предложения, отношение числа глаголов к общему количеству словоупотреблений в тексте и т. д.

Одной из последних разработок в этой сфере является методика, предложенная О.

Хрулевым [1], определяющая автора текста на основе Евклидова расстояния. Суть метода

заключается в следующем:

 в распоряжении исследователя имеется текст неизвестного автора, и принимается

гипотеза о некотором списке возможных претендентов на роль автора данного тек-

ста

на основе анализа всех известных произведений (содержащихся в базе данных)

создается частотный словарь автора – словарь, содержащий лексемы, употребляемые в анализируемых текстах, с частотами употребления и частотный словарь тек-

ста, авторство которого необходимо определить;

 в качестве нормирующего параметра используется частота употребления данной

лексемы в русском языке в соответствии с данными, представленными в частотном

словаре русского языка С.А. Шарова [2], при этом данные, представленные в частотном словаре автора делятся на среднюю частоту употребления в русском языке;

 определяется расстояние между словарями авторов и словарем анализируемого

текста как

A black text on a white background

Description automatically generated

где xi – частота употребления i-ой лексемы в тексте, автор которого неизвестен,

нормированная по средней частоте употребления в русском языке, ai – частота

употребления i-ой лексемы в частотном словаре предполагаемого автора, нормированная по средней частоте употребления в русском языке, N – количество анализируемых лексем. Учитываются только те лексемы, которые присутствуют в обоих

словарях. При этом анализу подвергаются только первые 5 000 – 10 000 наиболее

употребляемых в русском языке слов.

 В качестве автора анализируемого текста выбирается тот автор, чей словарь находится на меньшем расстоянии до словаря анализируемого текста