УДК 81’33

ОБЗОР ПРОГРАММНЫХ ПРОДУКТОВ РАЗРАБОТАННЫХ ДЛЯ АТРИБУЦИИ ХУДОЖЕСТВЕННЫХ ТЕКСТОВ

Мощенкова Д.С., Кривицкая Д.А., Амосова Н.С.
научный руководитель преподаватель Амосова Н.С.
Сибирский федеральный университет

Проблема установления авторства литературных произведений связана с огромным количеством псевдонимных и анонимных текстов и беспокоит многих филологов, юристов, историков и других специалистов уже сотни лет.

Актуальность работы состоит в том, что, не смотря на существование различных методов и способов атрибуции текста, установление авторства текста требует дополнительных исследований, поэтому необходимо выделить наиболее точные автоматические методы и системы атрибуции.

Атрибуция давно и широко исследуется различными филологами и литературоведами, в частности Марусенко М.А., Берковым П.Н, Миловым Л.В, Родионовой Е.С. и многими другими.

Целью работы является обзор и сравнительный анализ программных систем, облегчающих работу экспертов в области атрибуции художественных текстов.

Данная работа является обзорной и кратко разъясняет суть и методы атрибуции, а также программные продукты, направленные на решение проблемы определения авторства художественных текстов.

Задачи работы:

1. Проведение краткого экскурса в историю атрибуции художественных текстов.
2. Обзор последних программных продуктов, с помощью которых производится установление авторства в литературе.
3. Анализ программного обеспечения по критериям: средства анализа текстов, необходимый объем текста, точность.

Определение атрибуции текстов. Установление авторства текста является одной из древнейших филологических задач. Эксперты в этой области разрабатывали различные методы и способы на протяжении многих лет. Бесспорно, их работа является очень трудоемкой и для ее упрощения в последнее время были применены попытки автоматизировать процесс по определению авторства.

Долгое время для атрибуции текстов применялись в основном историко-документальные и филологические методы исследования. Математические и статистические методы были применены лишь к концу XIX века.

Сам термин атрибуция (согласно Большой Советской Энциклопедии) происходит от лат. Attributio – приписывание и подразумевает собой установление авторства, подлинности, а также времени и места создания произведения.

Задачи и методы определения атрибуции текстов. Атрибуция художественного текста включает в себя следующие задачи:

1. Идентификационные:

- подтверждение и исключение авторства определённого человека;
- проверка текста на то, что написавший текст является его единственным и подлинным автором.

2. Диагностические (определение личностных характеристик автора):

- Уровень образованности;
- Познания в иностранных языках;
- Место рождения / проживания;
- Профессия, увлечения;
- Пол, возраст, национальность и т.п.;

- Навыки в определенных речевых стилях;
- Наличие сознательного искажения стиля и письменной речи;
- Атрибуция осуществляется в основном в трёх основных направлениях:
- Обнаружение документально – фактических доказательств;
- Раскрытие идейно – образного содержания текста;
- Анализ языка и стиля.

О последнем направлении расскажем подробнее. Анализ языка и стиля текста проводится в основном по следующим направлениям: пунктуация, орфография, синтаксис, лексика и фразеология, стилистика.

- Анализ *пунктуации* подразумевает собой поиск характерных для автора ошибок, а также выявление особенностей и частоты употребления тех или иных пунктуационных знаков.
- На *орфографическом* этапе проверки, соответственно, проверяется наличие свойственных автору ошибок в написании слов.
- Анализ *синтаксиса* в тексте – это выявление особенностей построения предложений, наличие речевых оборотов, конструкций, частота употребления определенных слов.
- *Лексико-фразеологический* этап также играет немаловажную роль: происходит установление частоты употребления автором фразеологизмов, неологизмов и прочих художественных средств выразительности, а также богатство словарного запаса.
- *Стилистический* анализ помогает определить жанр произведения, сюжет и различные присущие автору речевые приемы.

Для атрибуции используются экспертные и формальные методы.

Экспертные методы производятся посредством обработки данных специально обученными профессионалами в сфере лингвистики.

Формальные же методы предназначены в основном для вычислительной техники. Появление возможности реализации методов, требующих огромных вычислений, значительно расширило границы в области атрибуции. Существующие программные продукты позволяют применять новейшие методы и способы, учитывают различные характеризующие текст параметры.

Далее рассмотрим доступные программные системы:

Система «Лингвоанализатор». Метод, применяемый в этой системе для определения авторства текста, основан на формальной математической модели.

Программа учитывает следующие характеристики языка автора:

- число служебных слов;
- используемые морфемы;
- уровень сложности употребленных грамматических конструкций;
- словарный запас.

Последние исследования, проведенные на механико-математическом факультете МГУ им. М.В. Ломоносова показали, что совокупность вышеперечисленных характеристик очень хорошо выявляет авторский стиль.

Система «Атрибутор». Данная программа является он-лайн лингвистическим процессором для машинного сравнения текстов и их классификации по параметрам индивидуального авторского стиля. Произведения подбирались так, чтобы тексты разных писателей имели как можно больше различий, а тексты одного писателя имели максимальные сходства. На данный момент система обучена сравнивать только тексты романов. Для атрибуции достаточно примерно шесть печатных страниц.

Система «СМАЛТ». Система состоит из двух основных блоков: *функционального* (анализ, база данных) и *аналитического* (реализация методик статистического анализа текстов).

Проект еще не доработан до конца и предполагает разработку информационной системы, применяющую статистические методы анализа. В основе должна иметься база литературных произведений, состоящая из публицистики 60-70 гг. 19 века. Обработка текстов в данной системе производится поэтапно:

1. выполнение автоматизированного разбиения исходного текста на: раздел, абзац, предложение, слово;
2. осуществление автоматической обработки текста и его морфологический разбор;
3. синтаксический анализ;
4. выполнение пользователем операций из базы данных по анализу текстов.

Система «Автоворед». Программа, основанная на фоносемантическом анализе, составляет психологический портрет автора. Система содержит набор DLL-библиотек, которые подключаются к текстовому процессору Word for Windows и в главном меню появляется новый пункт. Таким образом, данная программная система позволяет пользователю работать в привычной для него среде.

Ниже приведена таблица, сравнивающая основные характеристики вышеописанных программных средств.

Таблица 1

Некоторые системы автоматической атрибуции текстов

Название системы	**Средства анализа текстов**	**Необходимый объем текста**	**Точность**	**База**
Лингвоанализа-тор	Цепь Маркова, энтропийный подход	400- 100 000 символов	84-89%	132 автора 1357 произведений
Атрибутор	Цепь Маркова	Более 20 000 символов	48-56%	103 автора ~1287 произведений
СМАЛТ	Кластерный анализ, критерии Стьюдента	500 слов	Не указано	300 произв. из публицистики 60-70 гг. 19 века
Авторовед	Метод опорных векторов, кластерный анализ	20 000-25 000 символов	95-98%	Не указано

В работе были рассмотрены различные системы автоматического становления авторства художественного произведения, и их характеристики, подчеркивается важность продолжения исследований направленных на совершенствование имеющихся методов атрибуции художественных текстов, а также на создание новых и рационализацию уже созданных программных систем. Не менее важен поиск характеристик, позволяющий при малых объемах выборки, четко разделить стили авторов.