УДК 004.02

ФОРМАЛЬНІ МЕТОДИ ВИЗНАЧЕННЯ АВТОРА ТЕКСТУ

М. В. Гвозденко, Н. А. Кошева, к.т.н., доцент
Національний університет «Юридична академія України імені Ярослава Мудрого»
Gvozdenko@rambler.ru, n_kosheva@mail.ru

Мережеві технології та поява великої кількості приватних видавництв призвело до стрімкого зростання несанкціонованого розповсюдження наукових та художніх текстових творів, що, в свою чергу, викликало масове порушення авторських прав.

Виявлення автора текстового твору на сьогоднішній день ускладнюється тим, що оригінальні твори підлягають суттєвому редагуванню досить обізнаних рерайтерів, які використовують широкий спектр прийомів та засобів для максимальної відмінності відредагованого тексту від оригінального.

То ж потреба в проведенні якісної, достовірної лінгвістичної експертизи набуває значної актуальності.

Типові випадки експертиз по визначенню автора описуються наступними ситуаціями:

- множинна невизначеність – на експертизу надані декілька текстів. Треба визначити, скільки авторів їх писали і яку частину кожного тексту написав кожен автор;
- порівняння за зразком – є текст або декілька текстів, створених конкретним автором. Треба визначити, чи є цей автор автором тексту, по якому проводиться експертиза;
- конкуренція зразків – є зразки текстів декількох

авторів. Треба визначити, хто з них є автором декількох запропонованих текстів.

Ідентифікація автора тексту – це набір методів встановлення автора за приватними особливостями тексту.

Ідентифікація може бути виконана експертними методами, які виконуються професійним лінгвістом-експертом, або формальними методами, які засновані на аналізі обчислюваних характеристик тексту.

Перевага формальних методів визначення автора тексту полягає в тому, що саме цей метод забезпечує високий ступінь об'єктивності результатів лінгвістичних експертиз, а також використовується для створення програм та інформаційних систем визначення автора тексту.

Основною проблемою використання формальних методів є саме визначення приватних ознак, причому ці ознаки повинні відповідати декільком умовам: ознака повинна відображати ті характеристики тексту, які автор використовує підсвідомо, зберігати постійне значення для одного автора і мати істотно різні значення для різних авторів.

На сьогоднішній день до таких ознак відносять:

– підрахунок частоти і природи лексичних, орфографічних, синтаксичних і граматичних помилок;

– дослідження стилістичних факторів письмової мови (довжина слів, довжина речень; кількість складів, приставок і суфіксів на 100 слів);

– метод опорних слів, (метод Фоменко) - підрахунок кількості появи союзів, частинок і приводів;

– метод розділових знаків - підрахунок тільки кількості внутрішніх і зовнішніх розділових знаків;

– метод слів - підрахунок тільки слів певної довжини;

– метод речень - підрахунок тільки речень певної довжини;

– синтаксичний метод - підрахунок розділових знаків, слів і речень певної довжини знаків;

– комбінований - об'єднання методу Фоменко і синтаксичного методу;

– підрахунок відсотка зустрічаємості в тексті частин мови - співвідношення дієслів до прикметників, дієслів - до іменників і т. п.,;

– показник TTR (Type Token Ratio) – представлення у формі десяткового дробу співвідношення кількості різних слів із загальною кількістю слів в тексті.

У доповіді розглянуті формальні методи, які покладені в основу розробки та функціонування комп'ютерних програм проведення лінгвістичної експертизи: розглянуті метод порівняння гістограм частот слів різної довжини, наївний байесовський (НБА) метод, метод розподілу параметра порівняння кількостей нових слів в суперечливому тексті, метод відносної ентропії, метод стійкості частот, індекс Флеша, FOG-індекс, підхід Колтарда, лінгво-статистичний аналіз неповнозначної лексики, розпізнавання автора тексту з використанням ланцюгів А.А. Маркова, атрибуція Мінімальної Умовної Складності Стиснення (МУСС — атрибуція), апарат машини опорних векторів (Support Vector Machines, SVM).

Наведені перспективи розвитку методів визначення авторства, зокрема атрибуція на основі не лише формальних характеристик письмової мови, а на основі їх суб'єктивних образів, які менш мінливі.