УДК 004.02

М.В. Гвозденко, В.В. Карасюк

Національний університет «Юридична академія України ім. Ярослава Мудрого», Харків

АВТОМАТИЗАЦІЯ ЛІНГВИСТИЧНОЇ ЕКСПЕРТИЗИ ЕЛЕКТРОННИХ ДОКУМЕНТІВ

В статті розглянуті актуальні питання створення методики, програм та інформаційних систем визначення автора текстового і, зокрема, електронного документа. Визначено, що в розробці комп'ютерних програм визначення авторства використовуються формальні методи ідентифікації автора тексту, які засновані на порівнянні обчислюваних характеристик текстів. Наведені умови, яким повинні відповідати особливості тексту певного автора. Наданий перелік сучасних формальних методів визначення авторства тексту. Наведений перелік і опис сучасних програм визначення автора тексту.

***Ключові слова**: лінгвістична експертиза, ідентифікація автора, формальні методи, аналіз тексту*

Вступ

Стрімкий розвиток мережевих технологій призвів до значного зростання кількості електронних текстових документів. Одночасно з цим зросла і кількість порушень авторських прав авторів текстових творів, виникли проблеми масового розповсюдження конфіденційної інформації та інформації, що порушує гідність та ділову репутацію окремих осіб та установ, що вимагає розробки надійних, достовірних і швидких засобів визначення авторства.

Одним із таких засобів є проведення лінгвістичної експертизи електронних документів.

Наукові статті, дисертаційні роботи, підручники, твори художньої літератури і т. ін. безконтрольно розповсюджуються в мережі Інтернет, іноді без будь-яких редакторських змін, іноді після незначного редагування. І якщо пряме копіювання, так званий прийом Copy-Paste, не вимагає проведення експертного дослідження для виявлення автора тексту, то текст, відредагований штатом досить обізнаних рерайтерів, потребує ретельного і достовірного експертного дослідження. Адже рерайтери використовують широкий арсенал засобів перетворення унікального тексту в інші тексти: переклад унікального тексту на іншу мову і редагування одержаного перекладу, використання синонімів (навіть з використанням спеціальних програм – синонімайзерів), заміна прямої мови на непряму, додавання в унікальний текст або видалення з нього слів та словосполучень, які не несуть смислового навантаження тощо.

В наш час відзначається зростання числа цивільних справ у судах загальної юрисдикції, пов'язаних із захистом авторських прав на твори науки, літератури і мистецтва, патентних прав, прав на товарні знаки, фірмові найменування та інші об'єкти інтелектуальної власності.

Якщо у справах, пов'язаних із захистом честі, гідності та ділової репутації, вже склалася стійка практика проведення лінгвістичних експертиз спірних текстів ЗМІ, відпрацьовані експертні методики, то у відношенні об'єктів інтелектуальної власності відповідні напрями експертно-лінгвістичної діяльності ще тільки формуються, експертна та судова практика ще напрацьовується.

Оскільки підвищення якості та швидкості виконання експертних досліджень значною мірою залежить від використання засобів інформаційних технологій, то автоматизація лінгвістичної експертизи електронних документів на сьогодні набуває великої актуальності.

Постановка проблеми у загальному вигляді

Ідентифікаційні задачі лінгвістичної експертизи повинні дати відповіді на такі питання: підтвердження авторства певної особи, виключення авторства певної особи, перевірка того, що автором всього тексту є одна і та ж людина, перевірка того, що виконавець тексту є одночасно його автором.

Ідентифікація автора тексту – це набір методів встановлення автора за характеристиками стилю, що мають відображення в особливостях тексту.

Основною проблемою встановлення авторства є саме визначення особистих ознак, причому ці ознаки повинні відповідати декільком умовам: ознака повинна відображати такі характеристики тексту, які автор відтворює підсвідомо, зберігати постійне значення для одного автора і мати істотно різні значення для різних авторів.

Чим більший об'єм тексту пропонується для експертного дослідження, тим простіше виявити стабільні частоти деяких мовних елементів, притаманних саме даному автору і стає можливим виділення частотних показників.

В розробці комп'ютерних програм визначення авторства використовуються формальні методи іде-

нтифікації автора тексту, які засновані на порівнянні обчислюваних характеристик текстів, таких як 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); підрахунок відсотка наявних в тексті частин мови: співвідношення дієслів до прикметників, дієслів - до іменників і т. п., а також показник TTR (TypeTokenRatio) - представлення у формі десяткового дробу співвідношення кількості різних слів із загальною кількістю слів в тексті, 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, лінгвостатистичний аналіз неповнозначної лексики, розпізнавання автора тексту з використанням ланцюгів А.А. Маркова, атрибуція Мінімальної Умовної Складності Стиснення (МУСС — атрибуція), апарат машини опорних векторів (SupportVectorMachines, SVM) тощо.

Таким чином, постановка проблеми ідентифікації автора тексту набуває більшої актуальності у зв'язку з розширенням електронного обігу документів у мережі Internet. Рішення цієї проблеми є особливо важливим для визначення авторів, причетних до фактів деструктивної маніпуляції з текстами у засобах масової інформації.

Аналіз досліджень і публікацій з проблематики лінгвістичної експертизи

Останнім часом питання лінгвістичної експертизи текстових документів вирішується широким колом фахівців – математиків, програмістів, вебмайстрів, системних адміністраторів, тощо.

Одним із найстаріших методів ідентифікації автора тексту є метод частотних словників, запропонований А.А. Марковом в 1913.

Серед сучасних науковців, фахівців в галузі атрибуції тексту, які внесли значний внесок у розвиток формальних методів визначення авторства, слід відзначити таких дослідників, як Хмельов Д. В., Баранов А.Н., Романов А.С., Галяшина Е. І., Белянин В.П., Апресян В.Ю., Будко Т.В., Голев Н. Д., В.В. Піддубний, О.Г. Шевелев, Фоменко В.П. [1-13] та ін., а також розробників автоматизованих систем визначення автора тексту, таких, як Хмелев Д. В., Деліцин Л. Л., Шевелев О. Г., Белянина В. П., Батов В. И. та ін. В цілому напрацьований статистичний апарат досліджень з даної проблематики, але ефективність пропонованих методів є різною відповідно до типів текстових документів та їх об'єму.

Мета статті. Провести огляд і виконати порівняльний аналіз сучасних автоматизованих засобів проведення лінгвістичної експертизи. Зробити висновки про ефективність засобів та надати рекомендації щодо їх використання та подальшого вдосконалення.

Огляд програм автоматизованого аналізу текстів

Формалізовані методики визначення автора тексту надають можливість виявлення та підрахунку ідентифікаційних ознак тексту, що, в свою чергу, дає можливість створення автоматизованих систем лінгвістичної експертизи.

На сьогодні, на жаль, відсутні досконалі методики і, відповідно, автоматизовані системи, які дають достовірний та стабільний результат, особливо на навеличких фрагментах тексту, але кожна з нижче розглянутих програм показує досить високий відсоток достовірних результатів лінгвістичних експертиз.

Програма Prostyle (США)

Програма здійснює аналіз будь-якого тексту, що вводиться, і виводить в порядку номерів фактори, що дозволяють провести статистичний аналіз значення в будь-яких розбіжностях в двох досліджуваних текстах. Серед факторів, що враховуються програмою Prostyle, знаходяться:

–граничний індекс чіткості (наскільки даний текст легкий або важкий для розуміння);

–індекси FOG і Флеша - Кінкейда;

–показник частотності страждальних конструкцій, що дозволяє достатньо точно виявити індивідуальні особливості автора;

–кількість використовуваних лексичних одиниць, яка при обчисленні відсотка співвідношення з загальною кількістю слів в тексті дає показник словарного запасу автора;

–відсоток складних слів по префіксам, суфіксам, кількості складів (у Prostyle - тільки по останньому фактору);

–середня довжина речення, що прямо корелює з рівнем освіти автора;

–"читацький вік", представлений даним текстом;

–кількість погрішностей письмового стилю в тексті (можливі помилки: неправильне використання абстрактних іменників; неправильне вживання дієслівних форми прийменників; опущення дієслова; недоречне вживання сленгу і жаргону; використання застарілих слів; порушення пасивних конструкцій; грубі і непристойні слова; слабке знання мови).

Програма "E'RIDAtextvisor"

Програма для аналізу тексту веб-сервером сторінок сайту "E'RIDAtextvisor" проводить загальний

аналіз тексту сторінок сайту, аналіз тексту мета-тегів. Аналіз тексту сторінки:

–сканує текст сторінки;

–сканує текст в метатегах сторінки;

–сканує текст анкора (anchor) посилань на сторінці;

–санує опис посилань (Alt).

Отримані результати представлені у вигляді:

–загальна кількість слів на сторінці, а так само окремо в "меті", "тексті" і "посиланнях";

–загальна кількість символів на сторінці, а так само окремо в "мета", "тексті" і "посиланнях";

–демонстрація кожного слова з вказівкою кількості однокорінних слів, а так само де і скільки кожне із слів розташовується (у структурних елементах: текст, мета, посилання);

–відсоткове співвідношення кожного слова до загальної кількості слів, а так само роздільно "% в мета", і "% в тексті" (зручно для контролю ключових слів).

Аналіз мета-тегів:

–сканує текст title;

–сканує текст description;

–сканує текст keywords.

Отримані результати представлені у вигляді:

– кількість слів в кожному з пунктів;

– кількість символів в кожному з пунктів;

– кількість кожного слова, з урахуванням однокорінних слів;

– відсоткове співвідношення кожного слова в кожному з пунктів (окремо % у title, description і keywords);

– визначення скільки кожного із слів знаходиться в кожному з пунктів.

Аналіз посилань:

–визначення тексту кожного посилання (anchor);

–визначення опису кожного посилання (alt).

Отримані результати представлені у вигляді:

–окремо опис і окремо текст посилання;

–розбір за словами опису і тексту посилання;

–визначення кількості слів і кількості символів, як в описі так і в тексті кожного посилання.

Програма «АНТИПЛАГІАТ»

eTXT Антиплагіат - програма перевірки унікальності тексту.

Програма дозволяє провести докладний аналіз унікальності тексту і визначити оригінальність статті у відсотковому співвідношенні. У програмі враховані особливості роботи копірайтера. Програма має 2 версії: установка на комп'ютер користувача і режим on-line .

При роботі зі встановленою програмою користувач може:

–знаходити і виділяти не унікальні фрагменти тексту безпосередньо на відтвореній копії вебсторінки, що значно полегшує визначення унікальності тексту;

–створювати докладні звіти перевірки унікальності контенту з можливістю налаштування різних параметрів пошуку - числа вибірок з тексту, кількості слів в шинглі і ін.;

–перевіряти на унікальність всі сторінки сайту, видаючи докладний звіт по сайту;

–вести пакетну перевірку всіх файлів з теки.

On-line версія програми eTXT Антиплагіат дозволяє:

–перевірити текст на унікальність незалежно від зовнішніх факторів, таких як швидкість інтернет-з'єднання або встановлена на вашому ПК операційна система;

–не боятися блокування пошуковими системами;

–зберігати результати перевірки на сервері і мати можливість надати їх постійну адресу при необхідності;

–економити трафік.

Програма ШТАМПОМЕР

Програма ШТАМПОМЕР виконує статистичний аналіз тексту і порівняльний аналіз текстів.

При статистичному аналіз тексту програма збирає різні статистичні дані і записує їх у файл результатів у вигляді таблиць відношень або відсоткового змісту:

–загальні дані;

–зміст розділових знаків;

–зміст завершуючих розділових знаків;

–вміст речень в абзаці;

–вміст слів в реченні;

–вміст розділових знаків в реченні, а також таблиць аналізу штампів:

–повторення штампів n-го рівня;

–повторення штампів n-го рівня в одному абзаці;

–повторення штампів n-го рівня в одному реченні.

У цій програмі під штампом n-го рівня розуміємо словосполучення із n слів. Тобто штамп 1-го рівня - це одне слово, а 5-го рівня – словсполучення із 5 слів.

При порівняльному аналізі текстів програма обчислює виражені у відсотках різниці відповідних таблиць даних, отриманих на етапі статистичного аналізу текстів. Такі дані характеризують відмінність таблиць, і, чим вище їх значення, тим нижче вірогідність ідентичності авторства початкових текстів.

Іформаційна система "Статистичні методи аналізу літературного тексту" (ІС "СМАЛТ").

ІС складається з двох основних блоків:

–функціонального блоку, призначеного для морфологічного і синтаксичного аналізу текстів, поповнення БД літературних творів, а також внесення виправлень;

– аналітичного блоку, що складається з модулів, що реалізують різноманітні методики статистичного аналізу текстів.

Як початкове джерело даних для клієнтського застосування використовується текстовий файл в кодуванні Unicode, що дозволяє уникнути проблем, пов'язаних з використанням в окремих текстах символів, специфічних як для окремих мов, так і для орфографії різних періодів однієї мови. Обробка текстів в інформаційній системі проводиться у декілька етапів. На першому кроці виконується автоматизоване розбиття початкового тексту на лексичні одиниці, серед яких виділяються частина (або розділ), абзац, речення, слово. Розбиття здійснюється на основі апарату регулярних виразів. Шаблон розбиття можна змінювати від тексту до тексту, при цьому він зберігається разом з даними про розбиття. На другому етапі здійснюється автоматична обробка тексту і його морфологічний розбір. При морфологічному розборі для окремих частин мови виділяється до 20 морфологічних ознак. На базі побудованого морфологічного розбору проводиться третя стадія обробки тексту – синтаксичний аналіз. На цій стадії для кожнго речення початкового тексту визначаються в середньому близько 15 ознак. Після здійснення обробки вхідного тексту, її результати поміщаються в централізоване сховище (репозиторій текстів, готових для статистичного аналізу).

На наступному етапі користувач може виконувати операції по аналізу текстів, що знаходяться в репозиторії як з використанням клієнтського програмного забезпечення, так і частково через web, використовуючи інтерфейс, що надається web-вузлом. Окрім цього, користувачам ІС СМАЛТ надається можливість внесення змін і виправлень в опубліковані дані. Таким чином, можна проглянути одні і ті ж дані в редакції різних фахівців, а також порівняти результати, що отримуються при статистичній обробці різних редакцій.

Система «Плагиат-информ»

Система Плагиат-інформ дозволяє визначити факт плагіату.

Система Плагиат-інформ, розроблена на основі унікальної технології пошуку документів, схожих за змістом, дозволяє легко виявити плагіат, насамперед, в студентських роботах.

Масштабування дає можливість об'єднувати декілька вузів в один інформаційний простір, а також факультети, або філії вузу, розташовані на далині один від одного.

Для початку повноцінної роботи з програмою Плагиат-інформ потрібна наявність хоч би одного пошукового індексу.

Створення різних індексів в програмі Плагиат-інформ дає можливість використовувати декілька варіантів пошуку плагіату для рефератів і здійснювати пошук плагіату окремо по кожному індексу.

Спершу пошук плагіату йде по індексу, де реферат, що перевіряється, цілком порівнюється зі всіма документами в індексі. Якщо при перевірці запозичень у файлі не знайдено, то запускається пошук по іншому індексу, в якому документи розбиті на абзаци, і документ, що перевіряється, теж перевіряється по кожному абзацу. Цей пошук виконується повільніше попереднього, зате набагато точніше визначає плагіат і його відсоток. Часто текст, що не був плагіатом після перевірки по першому індексу, стає плагіатом при другому виді пошуку, причому відсоток змісту запозиченої інформації буде достатньо високим.

Прграмма дозволяє виконати:

–тестування цілого файлу на наявність запозичень – можна визначити не тільки рівень плагіату, але також можна знайти файл-першоджерело і його зміст. Запозичений з такого документа текст буде відмічений;

–тестування файлу при перенесенні абзаців усередині тексту;

–тестування файлу при додаванні нового тексту, видаленні фрагмента, переміщенні речень;

–тестування файлу, складеного з фрагментів інших документів;

–тестування файлу, складеного з фрагментів інших документів з перестановкою абзаців.

Програма АТРІБУТОР

Программа атрібутор є лінгвістичним процесором для автоматичного порівняння і класифікації текстів по параметрах індивідуального авторського стилю. Мета роботи програми – розпізнавання автора тексту або видача списку найбільш близьких до нього по стилістиці авторів з числа вхідних в деякий заздалегідь заданий перелік "еталонних" авторів.

При роботі програми передбачено 3 ситуації:

–найбільш вірогідним автором є Х. Цей висновок означає, що в нашій виборці є тексти наданого на дослідження письменника;

–автора цього тексту в нашій базі немає. Цей висновок означає, що присланий текст містить особливості індивідуального стилю, по яких він достатньо різко відрізняється від наявних у вибірці письменників. Цей текст, мабуть, не містить індивідуальних стилістичних рис;

–список найбільш близьких авторів (в порядку убування вірогідності). Цей висновок означає, що досліджуваний текст по стилістиці не збігається ясно ні з одним з наявних у вибірці письменників і, в той же час, не має різких відмінностей відразу від декількох з них.

Як ознаки для аналізу і оцінки індивідуального авторського стилю використовуються трьохлітерні поєднання – тріади. Обробку проходять всі слова тексту, причому початок і кінець слова доповнюються пропусками, які також враховуються в тріадах. Однакові тріади підсумовуються, із зібраних по

тексту тріад виходить профіль, який є пошуковим образом, що характеризує стиль.

У обробку потрапляють всі слова тексту за винятком власних назв. У лінгвістичному сенсі трьохлітерні поєднання є інтегральною характеристикою, що об'єднує відразу декілька різнорідних стильових ознак. При такій методиці дослідження окремими тріадами в підрахунок потрапляють розподіли однолітерних і пар тріад – двохлітерних службових слів, а це значна частина найбільш частотних прийменників, союзів, частинок і вигуків, які традиційно вважаються за значущі стилеметричні показники. З цієї причини двохлітерні, чотирьох- і більш літерні ланцюжки менш показові, що і було виявлено в процесі перевірки їх розрізняльної сили.

Решта літеросполучень так чи інакше відображає і граматичні явища (частоту граматичних частин, вжитих в тексті слів), і лексичні (літеросполучення з основи слова), причому нерозчленовано. Хоча розрізняльна сила окремих літеросполучень очевидно неоднакова, в даній версії атрибутора при оцінці і зважуванні це поки не враховується.

Програма ЛІНГВОАНАЛІЗАТОР

Програма виконує читання і обробку тексту невідомого походження з метою визначення близькості до одного з авторських еталонів, визначених заздалегідь.

"Лінгвоаналізатор" розбирає текст на складові, використовуючи математичну модель, в якій враховані такі характеристики тексту, як:

–число службових слів (прийменників, союзів і інших частинок);

–морфеми (префіксальні, кореневі, суфіксальні, флективні) і їх послідовності;

–складність граматичних конструкцій;

–власне словник, використаний автором.

Програма вимірює всі ці параметри і зводить в таблиці, що містять сотні змінних, які характеризують письменника. У кожного автора з бази даних є своя таблиця, яка є авторським еталоном. Початкові тексти "Лінгвоаналізатор" у себе не зберігає.

При введенні аналізованого тексту відбувається побудова ще однієї таблиці по вхідному тексту. Після цього вхідна таблиця зіставляється з **X** таблицями по кожному авторові і виводиться **X** інтегральних величин для оцінки близькості даного тексту до кожного з **X** письменників. Кожна з цих **X** інтегральних величин називається відносною ентропією. Програма повідомить імена трьох авторів, для яких відносна ентропія по даному тексту мінімальна. В більшості випадків програма правильно називає автора, навіть якщо пропонувати їй твори, що не містяться в базі даних. Це можливо лише, оскільки алгоритм роботи програми не зводиться до повнотекстового пошуку по всій базі даних - використовуються тільки інтегральні характеристики текстів.

Нормативне забезпечення авторознавчої експертизи

Проведення авторознавчих експертиз призначається і проводиться відповідно до нормативних документів України:

Закону України «Про авторське право і суміжні права» { Із змінами, внесеними згідно із Законами N 850-IV (850-15) від 22.05.2003, ВВР, 2003, N 35, ст.271 N 1294-IV (1294-15) від 20.11.2003, ВВР, 2004, N 13, ст.181 N 2939-VI (2939-17) від 13.01.2011, ВВР, 2011, N 32, ст.314 };

Постанови 04.06.2010 N 5 «Про застосування судами норм законодавства у справах про захист авторського права і суміжних прав»;

Закону України «Про наукову і науково-технічну експертизу» (із змінами і доповненнями, внесеними Законом Україні від 21 вересня 1999 року N 1069-XIV).

Висновки

Таким чином, дослідивши методи і засоби лінгвістичної експертизи, можна зробити декілька висновків:

–проблема визначення авторства тексту є актуальною і увага до неї збільшується у відповідності до збільшення кількості порушень авторських прав у електронному обігу документів;

–методика виявлення авторства не є тривіальною, тому у програмних засобах використовується велика кількість оригінальних евристичних методів;

–отримали розповсюдження методи, основані на фільтрації тексту, стеммінгу, перетворенні символів, що дає змогу системам знаходити запозичені тексти навіть при їх незначній модифікації;

–якість дослідження суттєво зростає при наявності значної бази досліджуваних текстів;

–проаналізовані види статистичного аналізу документа - індекс Флеша, FOG-індекс, стилеметрія, підхід Колтарда, лінгво-статистичний аналіз неповнозначної лексики і топографічний аналіз;

–найточніші дані про автора документа надають статистичні методи авторознавчої експертизи, зокрема стилеметрія;

–експертна лінгвістика, як складова юридичної лінгвістики, перехрещується, але не дублює теорію судово-лінгвістичної експертизи, яка розвивається в рамках судової експертології.

Деякі з наведених висновків можуть лягти в основу подальших досліджень. Тобто, окрім огляду методів і засобів лінгвістичної експертизи, підготована база для наступних робіт і визначений план перспективних досліджень.

Список літератури

1. Хмелев Д. В. Распознавание автора текста с использованием цепей А. А. Маркова. // Вестник МГУ: Серия 9, Филология. – 2000. – № 2. С. 115 – 126.

2. Кукушкина О.В. Определение авторства текста с использованием буквенной и грамматической информации / О. В. Кукушкина, А. А. Поликарпов, Д.

В. Хмелев // Проблемы передачи информации. – 2001. – Т. 37. № 2.- С. 96 – 109.

3. Баранов А. Введение в прикладную лингвистику: Учебное пособие / А. Н. Баранов – М.: Эдиториал УРСС, 2003. – 360 с.

4. Галяшина Е. И. Возможности судебных речеведческих экспертиз по делам о защите прав интеллектуальной собственности / Е. И. Галяшина // Интеллектуальная собственность: Авторское право и смежные права. – 2005. – № 9. – С. 50 – 59.

5. Будко Т.В. Щодо розробки методики встановлення факту деструктивної маніпуляції за текстами засобів масової інформації / Т.В. Будко // Матеріали "круглого столу" на тему: "Проблеми та пріоритети розвитку правової науки в інформаційній сфері", 11 листопада 2010 р. / Національна академія правових наук України, Науково-дослідний центр правової інформатики. – К.: НДЦПІ, 2010. – С. 32 – 34.

6. Богославська М. О. Лінгвістична термінологічна експертиза / М.О. Богославська // Актуальні питання державотворення в Україні очима молодих вчених: міжн. наук.-практ. конф., 23-24 квітня 2009 р. – К.: Київськ. націон. ун-т ім. Т.Шевченка, 2009. – С. 128-129.

7. Голев Н. Д. Экспертиза конфликтных текстов в современной лингвистической и юридической парадигмах / Н. Д. Голев // Теория и практика лингвистического анализа текстов СМИ в судебных экспертизах и информационных спорах: Сб. материалов научно-практического семинара. Ч.2. – М.: Галерия, 2003. – С.64—73.

8. Фоменко А.Т. Новая хронология Греции: Античность в средневековье. Т. 2. – М.: Изд-во МГУ, 1996. - 916 с.

9. Поддубный В. В. Сравнительный анализ эффективности алгоритмов распознавания авторства текстов по частотам переходов / В.В. Поддубный, О.Г. Шевелев, А.А. Фатыхов // Вестник Томского государственного университета. Серия "Математика. Кибернетика. Информатика". – 2006. – № 290. – С. 232 – 234.

10. Кукушкина О. В. Построение таблиц стилей текстовых произведений с использованием алгоритмов классификации на основе деревьев решений / О.В. Кукушкина, В.В. Поддубный, О.Г. Шевелев, А.И. Кубарев // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2012. – № 4 (21). – С. 79-88

11. Вавіленкова А. І. Виявлення логічних протиріч у текстовій інформації / А. І. Вавіленкова //Вісник НТУ «ХПІ». Серія: Інформатика і моделювання. – Харків: НТУ «ХПІ». – 2012. – № 38. – С. 32 – 37.

12. Сєдов А. В. Анализ неоднородностей в тексте на основе последовательностей частей речи [Электронный ресурс] / А.В. Седов, А.А. Рогов // Современные проблемы науки и образования. Электронный научный журнал. – 2013. – № 1. Режим доступа: http://www.science-education.ru/pdf/2013/1/271.pdf

13. Шарапов Р. В. Система проверки текстов на заимствования из других источников / Р.В. Шарапов, Е.В. Шарапова // Труды 13й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2011, 19 – 22 октября 2011 г. – ВГУ, Воронеж, Россия. – 2011. – С. 121– 126.

Надійшла до редколегії 6.03.2013

Рецензент: д-р техн. наук, проф. М.Г. Любарський, Національний університет «Юридична академія України імені Ярослава Мудрого», Харків.

АВТОМАТИЗАЦИЯ ЛИНГВИСТИЧЕСКОЙ ЭКСПЕРТИЗЫ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ

Гвозденко М.В, Карасюк В,В.

В статье рассмотрены актуальные вопросы создания методик, программ и информационных систем определения автора текстового и, в частности, электронного документа. Определено, что в разработке компьютерных программ определения авторства используются формальные методы идентификации автора текста, основанные на сравнении вычисляемых характеристик текстов. Приведены условия, которым должны соответствовать особенности текста определенного автора. Предоставлен перечень современных формальных методов определения авторства текста. Приведен перечень и описание современных программ определения автора текста.

***Ключевые слова**: лингвистическая экспертиза, идентификация автора, формальные методы, анализ текста*

AUTOMATION OF THE LINGUISTIC EXPERTISE OF ELECTRONIC DOCUMENTS

Gvozdenko MV, Karasyuk V.V.

The article is devoted to topical questions of creation of methods, programs and information systems for identify the author of the text and, in particular, the electronic documents. It is determined that in the development of computer programs for determine the authorship are used formal methods of identification of the author of the text, based on the comparison of the calculated characteristics of the texts. In article are given the conditions, for which the features of the text must correspond to particular author. There is provided the list of modern formal methods for determining the authorship of the text. A list of modern programs for determination of the author's text and their description are given.

***Keywords**: linguistic expertise, identification of the author, formal methods, text analysis.*