УДК 004.9

С.В. Голуб, О.В. Константиновська, М.С. Голуб

*Черкаський національний університет імені Богдана Хмельницького, Черкаси*

# ФОРМУВАННЯ ПОКАЗНИКІВ МАСИВУ ВХІДНИХ ДАНИХ ДЛЯ ІДЕНТИФІКАЦІЇ АВТОРСТВА ТЕКСТОВИХ ПОВІДОМЛЕНЬ

*Запропоновано результати досліджень процесу формування показників масиву вхідних даних для ідентифікації авторства друкованого тексту. Виявлено, що інформативність ознак та їх вибір залежить від конкретного завдання ідентифікації.*

***Ключову слова***: *ознака, текст, ідентифікація, масив вхідних даних.*

## Вступ

Проблема ідентифікації автора текстового повідомлення актуальна та багатогранна. Дослідження процесів розв'язання комплексу задач, що дозволяють усунути цю проблему, містять одну із спільних фундаментальних складових – формування переліку ознак масиву вхідних даних, що дозволяють розпізнати особу автора, його стать, вік, психологічний та фізичний стани та інше. Результати подібних досліджень знаходять своє використання в літературознавстві (зокрема при атрибуції давніх літературних текстів невідомих авторів або при перевірці належності твору до вірогідного автора), у криміналістиці (при проведенні оперативно-пошукових заходів, при атрибуції анонімів, псевдонімів, встановлення підробок та плагіатів), в різних історико-філологічних дисциплінах, що займаються дослідженням давніх та сучасних текстів будь-якої стилістичної та жанрово-тематичної приналежності а також в інших предметних областях, що прямо або опосередковано використовують відображення властивостей об'єкта в формі текстового повідомлення.

**Аналіз останніх досліджень і публікацій.** Сама проблема ідентифікації авторства виникла дуже давно, але спроби використати обчислювальну техніку для її вирішення почали робити тільки з 70-х років XX ст.

Давні ("домашинні") методи, наприклад, початку XX ст., відрізнялися наявністю великої кількості суб'єктивних оцінок: весь обсяг роботи виконувався вручну. Як приклад такої роботи можна привести працю Н.В. Морозова "Лінгвістичні спектри..." [1]. Дуже цікаві ідеї визначити автора за спектрами вживання службових частин мови перевірялися на практиці вручну з олівцем в руках.

В наш час нових обчислювальних технологій з'явилися нові методи, що зумовлюють високу оцінку тих критеріїв та ознак, які визначають належність тексту тому чи іншому автору.

Прикладом вживання таких технологій можуть бути методика Л.І. Бородкіна та Л.В. Мілова, в основі якої лежить побудова графа сильних зв'язків за матрицею частот парної повторюваності граматичних класів слів [2], методика Захарова В.Н. та ін., яка основана на діалоговій комп'ютерній обробці літературних творів та використовує багато граматичних характеристик [3].

Ці методи вимагають великої витрати часу та залучення в експеримент багатьох дослідників і вибірок великої кількості текстового матеріалу, крім того, ці праці характеризуються великим набором ознак, за допомогою яких можна досягти кінцевої мети – атрибуції тексту, але пошук та етапну обробку цих ознак потребує значно кращої автоматизації.

**Мета дослідження.** Які б ми не взяли ознаки для ідентифікації, (неважливо, чи це літературний твір, чи це політична стаття), всі вони перекриваються, тобто вживаються і тим, і іншим автором, але в різному обсязі. Це обумовлено тим, що люди в межах одного регіону використовують одну мову, один словниковий набір, але при цьому деякі слова використовують частіше, а деякі – рідше. Для того, щоб ідентифікувати автора тексту, потрібно перебрати дуже велику кількість цих ознак. Тому метою даного дослідження є виявлення переліку ознак, що дозволять розв'язувати задачі ідентифікації автора текстового повідомлення, та забезпечать можливість автоматизації робіт із визначення чисельних характеристик цих ознак.

## Виклад основного матеріалу

Для досягнення цієї мети ми пропонуємо використати *технологію розпізнавання образів*. Для цього ми поставили перед собою завдання пошуку інформативних ознак, котрі відповідали б наступним вимогам:

- достатньо високий рівень повторюваності;
- високий рівень інформативності;
- можливість автоматизації визначення характеристик.

Як було зазначено раніше, кожна людина має свій словниковий запас, який вона використовує стосовно різних призначень: на засіданнях, в розмо-

 

ві з співробітниками, друзями, в магазині, вдома тощо. Для кожного з цих випадків вона використовує лексику, притаманну тому чи іншому стилю мови. Але більшість слів складають так званий загальнолексичний шар, тому немає потреби в розробці методів, в яких враховується приналежність твору якомусь із стилів мови. Те ж саме стосується і жанрово-тематичної спрямованості тексту. Перед нами стоїть завдання створити універсальну методику ідентифікації автора будь-якого тексту.

Об'єктом нашого дослідження є авторський текст. Предметом – процес пошуку індивідуальних ознак тексту, притаманних окремим авторам.

Основними складниками тексту виступають речення. Речення є "основною синтаксичною одиницею-конструкцією,... являє собою граматично організоване поєднання слів... та має певну смислову й інтонаційну завершеність" [4]. Речення, в свою чергу, є поєднанням словосполучень та слів або є окремими словосполученнями та словами. Речення має складну структуру. Для визначення інформативних параметрів або ознак авторства треба визначити властивості його складників та семантико-синтаксичних зв'язків в ньому. Якщо взяти за одиницю виміру словосполучення чи слово – треба брати до уваги їх зв'язки як елементів речення, не забути про морфологічні властивості слів та їх словозміни. При дослідженні інформативності найменших частин слів – морфем, на перешкоді встає їх полісемія: наприклад, морфема **–и** може означати закінчення іменника називного відмінка жіночого роду множини (ве́рб**и**), родового відмінка жіночого роду однини (вербо́**й**), називного відмінка чоловічого роду множини (сино́**й**). Тому ці параметри не можна вважати інформативними. Залишається найменша мовна одиниця – літера. На перший погляд цей параметр не несе інформації: вживання кожної літери в тексті досить велике. Але саме відмінність у цьому вживанні кожного з авторів і лежить в основі інформативності.

Наші попередні дослідження виявили, що найбільшим ступенем вжитку у тексті визначаються літери, поєднання літер у пари (графічні частини слова), вживання інших графічних пар знаків (за участю проміжку між словами, пунктуаційних знаків, що графічно позначають початок чи кінець слова або речення), найменшим – поєднання на письмі трьох літер та більше. Останній критерій несе найменшу інформативність, оскільки величина повторення цих ознак суттєво незначна. Таким чином, на початку дослідження було виявлено такі основні групи ознак, що попадають під перший критерій:

1. Вживання літер українського алфавіту – 33 ознаки.

2. Вживання кожної пари графічного сполучення літер – 1024 ознаки.

3. Вживання початкових літер слова (пара "проміжок, літера") – 32 ознаки.

4. Вживання кінцевих літер слова (пара "літера, проміжок") – 33 ознаки.

5. Вживання пунктуаційних знаків (тільки знак, знак з одним проміжком або двома навколо, сполучення знаків) – 23 ознаки.

6. Довжина слова (кількість літер у слові) – 16 ознак.

7. Вживання прийменників, що складаються з двох або трьох літер – 17 ознак.

8. Вживання сполучників, що складаються з двох або трьох літер – 12 ознак.

9. Кількість слів у реченні (середнє значення).

10. Кількість літер у слові (середнє значення).

Загальна їх сума становить 1190 ознак. Всі ознаки вибиралися з усіх текстів вручну, використовуючи можливості програми Microsoft Word 2000.

Початкове дослідження інформативності параметрів проведено на наступному матеріалі.

1. Вибірка 1. Гончар О.Т. "Прапороносці", Ч.1, 1-3 сторінки (Гончар О.Т. Соч. в 6-ти томах.- Том 1. ./ за ред. С.А. Захарової.- К.: Дніпро, 1978).

2. Вибірка 2. Франко І.Я. "Борислав сміється", Ч.1, 1-3 сторінки (Франко І.Я. Борислав сміється. Boa Constrictor: Повісті.- Упоряд. М. Гончарука.- К.: Дніпро, 1981).

3. Вибірка 3. Коцюбинський М.М. "Невідомий", 1-3 сторінки (Коцюбинський М.М. Вибр. твори./ за ред. Л.М. Кирильця.- К.: Дніпро, 1977).

З кожного із вказаних творів взято три сторінки друкованого авторського тексту різних форматів, та запропонована автоматична вибірка в об'ємі 1000 знаків тексту. Вибірки здійснювались з розрахунком на досить невелику кількість знаків можливого досліджуваного твору (анонiма, поетичного стовпчика тощо).

Підбір авторів та творів зроблений випадково, не враховуючи ідентичність жанрово-тематичної та історичної приналежності творів.

Розрахована повторюваність всіх ознак на 1000 знаків тексту по кожному з вибраних авторів, яка коливалася від 0 до 170. За нижчий (пороговий) рівень було взято значення повторюваності "3". Всі ознаки, що мали повторюваність нижче цього рівня, відкидалися, тому що не вважалися за інформативними – можливість їх повторюваності надалі дуже сумнівна.

Всі ознаки, що мають повторюваність $\geq 3$ за сумою усіх авторських текстів на 1000 знаків з проміжками, занесені в табл. 1.

Також окремо були взяті ознаки, що становлять собою середнє значення довжини речення, виражене у кількості слів (СДР) та середнє значення довжини слова, виражене у кількості букв (СДС) по кожному автору.

Таблиця 1

Перелік інформативних ознак

| Групи | Ознаки |
|---|---|
| Літери | а б в г д е ж з и і ї й к л м н о п р с т у х ц ч ш щ ю я ь |
| Графічні пари літер | ав ал ам ан ат бу ва ви ві во го ди до ен ер за ив ий им ин ис ит ід ін ка ки ко ла ли ло ма ми мо на не ни ні но ну ов ог од ол ом ор пе по пр ра ре ри ро се ст. ся сь та те ти то ть чи що як |
| Початкова літера слова | _а (проміжок, літера) _б _в _г _д _з _і _к _м _н _о _п _р _с _т _у _ч _щ _я |
| Кінцева літера слова | а_ (літера, проміжок) в_ е_ з_ и_ і_ й_ м_ о_ у_ х_ я_ ь_ |
| Пунктуаційні знаки | . (крапка)<br>._ (крапка, проміжок)<br>, (кома)<br>,_ (кома, проміжок) |
| Довжина слова (Д) | 1-літерні, 2-х, 3-х, 4-х, 5-ти, 6-ти, 7-ми, 8-ми, 9-ти, 10-ти |
| Прийменники | на |
| Сполучники | і |

В результаті досліджень було отримано 144 ознаки, що мають різну інформативність. Необхідно було виділити ті ознаки, які мають суттєве значення для конкретного завдання – ідентифікації автора. За міру інформативності ознаки використовувався критерій (1):

$$КА = ДСПА / ДСПС, \quad (1)$$

де КА – критерій інформативності ознаки творів автора, що ідентифікується;

ДСПА – це дисперсія (розкид) значень ознаки творів автора, що ідентифікується;

ДСПС – це дисперсія суми ознак всіх авторів, що розглядаються.

Подальше дослідження інформативності параметрів проведено на наступному матеріалі.

1. Вибірка 1. Гончар О.Т. “Прапороносці”, Ч.1, 1-3 сторінки (Гончар О.Т. Соч. в 6-ти томах.- Том 1. ./ за ред. С.А. Захарової.- К.: Дніпро, 1978).
2. Вибірка 2. Гончар О.Т. “Гори співають”, 1-3 сторінки (Гончар О.Т. Соч. в 6-ти томах.- Том 1. ./ за ред. С.А. Захарової.- К.: Дніпро, 1978).
3. Вибірка 3. Гончар О.Т. “Весна за Моравою”, 1-3 сторінки (Гончар О.Т. Соч. в 6-ти томах.- Том 1. ./ за ред. С.А. Захарової.- К.: Дніпро, 1978).
4. Вибірка 4. Франко І.Я. ”Борислав сміється”, Ч.1, 1-3 сторінки (Франко І.Я. Борислав сміється. Boa Constrictor: Повісті.- Упоряд. М. Гончарука.- К.: Дніпро, 1981).
5. Вибірка 5. Франко І.Я. ”Boa Constrictor”, Ч.1, 1-3 сторінки (Франко І.Я. Борислав сміється. Boa Constrictor: Повісті.- Упоряд. М. Гончарука.- К.: Дніпро, 1981).
6. Вибірка 6. Франко І.Я. ”Борислав сміється”, 3 останні сторінки (Франко І.Я. Борислав сміється. Boa Constrictor: Повісті.- Упоряд. М. Гончарука.- К.: Дніпро, 1981).
7. Вибірка 7. Коцюбинський М.М. “Невідомий", 1-3 сторінки (Коцюбинський М.М. Вибр. твори./ за ред. Л.М. Кирильця.- К.: Дніпро, 1977).
8. Вибірка 8. Коцюбинський М.М. “Fata Morgana", 1-3 сторінки (Коцюбинський М.М. Вибр. твори./ за ред. Л.М. Кирильця.- К.: Дніпро, 1977).
9. Вибірка 9. Коцюбинський М.М. “Intermezzo", 1-3 сторінки (Коцюбинський М.М. Вибр. твори./ за ред. Л.М. Кирильця.- К.: Дніпро, 1977).

Були проведені розрахунки кількості повторюваності ознак, вказаних в таблиці 1, та виділені з них найбільш інформативні серед них, що мають значення показника КА, що не перевищує значення 0,5. Результати досліджень приведені в таблиці 2.

Як видно з таблиці 2, інформативність ознаки залежить від автора, якого ми ідентифікуємо, тобто всі подані параметри практично не перекриваються.

## Висновки

Таким чином, показана можливість використання як ознак таких малих фрагментів тексту, як літери та їх графічні пари, а також пари інших текстуальних знаків на прикладі авторських творів О.Гончара, І.Франка, М.Коцюбинського. З загальної кількості ознак (1190) за критерієм повторюваності вибрано 144 ознаки. За міру інформативності визначено відношення дисперсії значень ознаки творів автора, що ідентифікується та дисперсії суми ознак всіх авторів, що розглядаються.

Виявлено, що інформативність ознаки та їх вибір залежить від конкретного завдання ідентифікації.

Проведення подальших досліджень полягає в необхідності розробити програму автоматичної вибірки цих 144 інформативних ознак із тексту та обрати метод розпізнавання образів для ідентифікації авторів текстового документа.

## Список літератури

*1. Морозов Н.А. Лингвистические спектры: средство для отличия плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд // Известия отд. русского языка и словесности Имп.Акад.наук. Т.XX, кн. 4. – 1915.*

*2. От Нестора до Фонвизина. Новые методы определения авторства / Под ред. Л.В. Милова. – М.: Прогресс, 1994.*

*3. Программная система поддержки атрибуции текстов статей Ф.М. Достоевского / В.Н. Захаров, А.А. Леонтьев, А.А. Рогов, Ю.В. Сидоров.*

*4. Вихованець І.Р. Граматика української мови. Синтаксис: Підр. / І.Р. Вихованець. – К., Либідь, 1993.*

*Надійшла до редколегії 25.12.2013*

**Рецензент:** д-р техн. наук, проф. В.М. Рудницький, Черкаський державний технологічний інститут, Черкаси.

Таблиця 2

Повторюваність ознак

| Гончар | | Франко | | Коцюбинський | |
|---|---|---|---|---|---|
| Ознаки | КА | Ознаки | КА | Ознаки | КА |
| Д7 | 0,002 | з_ | 0,004 | СДР | 0,007 |
| г | 0,004 | . | 0,004 | ни | 0,007 |
| те | 0,009 | ін | 0,017 | Пна | 0,015 |
| ,_ | 0,012 | д | 0,040 | ві | 0,016 |
| _с | 0,015 | й_ | 0,042 | м_ | 0,019 |
| ._ | 0,015 | _м | 0,044 | ер | 0,024 |
| ід | 0,022 | та | 0,054 | ь_ | 0,028 |
| Д9 | 0,023 | ер | 0,061 | ал | 0,035 |
| ли | 0,030 | им | 0,062 | е_ | 0,038 |
| ог | 0,031 | ис | 0,067 | на | 0,044 |
| _я | 0,031 | до | 0,072 | _н | 0,047 |
| , | 0,034 | ._ | 0,076 | . | 0,054 |
| й | 0,035 | р | 0,079 | ко | 0,070 |
| о_ | 0,037 | ся | 0,079 | ре | 0,090 |
| д | 0,041 | и_ | 0,087 | те | 0,092 |
| п | 0,044 | _а | 0,087 | з | 0,105 |
| _г | 0,049 | я | 0,094 | СДС | 0,107 |
| _б | 0,050 | _б | 0,094 | ва | 0,109 |
| СДР | 0,052 | по | 0,100 | о_ | 0,111 |
| ов | 0,063 | но | 0,102 | Д8 | 0,117 |

## ФОРМИРОВАНИЕ ПОКАЗАТЕЛЕЙ МАССИВА ВХОДЯЩИХ ДАННЫХ ДЛЯ ИДЕНТИФИКАЦИИ АВТОРСТВА ТЕКСТОВЫХ СООБЩЕНИЙ

С.В. Голуб, А.В. Константиновская, М.С. Голуб

*Предложены результаты исследований процесса формирования показателей массива входящих данных для идентификации авторства печатного текста. Выявлено, что информативность показателей та их выбор зависит от конкретной задачи идентификации*

***Ключевые слова:*** *показатель, текст, идентификация, массив входящих данных.*

## FORMING OF INDEXES OF ARRAY OF ENTRANCE DATA IS FOR AUTHENTICATION OF AUTHORSHIP OF TEXT MESSAGES

S.V. Holub, A.V. Constantinovscaya, M.S. Holub

*The results of researches of process of forming of indexes of array of entrance data offer for authentication of authorship of the printed text. It is educed that informing of signs and their choice depends on a certain task to authentication.*

***Keywords :*** *sign, text, authentication, array of entrance data.*