КОМП'ЮТЕРНІ НАУКИ

УДК 004.58/934
DOI: https://doi.org/10.17721/ISTS.2024.7.45-51

Лариса МИРУТЕНКО, канд. техн. наук, доц.
ORCID ID: 0000-0001-8538-8996
e-mail: myrutenko.lara@gmail.com
Київський національний університет імені Тараса Шевченка, Київ, Україна

Яніна ШЕСТАК, канд. техн. наук, асист.
ORCID ID: 0000-0002-7291-1829
e-mail: luchenko@ukr.net
Київський національний університет імені Тараса Шевченка, Київ, Україна

Анастасія ЛОБАН, студ.
ORCID ID: 0009-0000-3828-6663,
e-mail: anastasialoban0912@gmail.com
Київський національний університет імені Тараса Шевченка, Київ, Україна

ДОСЛІДЖЕННЯ ІНДЕКСІВ ТЕОРЕТИЧНОЇ ВІДПОВІДНОСТІ ДЛЯ УКРАЇНСЬКОГО АЛФАВІТУ З ПРОБІЛОМ

В с т у п . *Важливу роль у виявленні слабкостей криптографічних систем і забезпеченні їхньої стійкості відіграють індекси теоретичної відповідності. Проаналізувавши доступні джерела інформації, не було виявлено вказаного показника для українського алфавіту з пробілом, що унеможливлює знаходження точних результатів у розрахунках, де використовують цей показник.*

М е т о д и . *Індекси теоретичної відповідності мають широке застосування, але найчастіше їх використовують у першому методі Фрідмана під час обчислення розміру ключа для шифру Віженера, який використовують у вивченні базових понять криптографічних систем. Розглянуто загальний алгоритм знаходження індексів теоретичної відповідності для українського алфавіту з пробілом на основі аналізу великої вибірки текстів.*

Для розрахунків проаналізовано 700 текстів, кількість елементів яких становить 1500. Тексти обрано за допомогою API із загальнодоступної вільної багатомовної онлайн-енциклопедії Wikipedia. Розрахунок виконували для ключів розміру від 2 до 7. Індекс теоретичної відповідності обчислено у два етапи. На першому етапі знайдено практичні індекси відповідності для всіх текстів, а далі за допомогою стандартного відхилення вибірки обчислено проміжок найповторюваніших показників. Програмний код для реалізації алгоритму обчислення індексу було представлено у вигляді компонентів проєкту, які відповідають за різні аспекти процесу.

Р е з у л ь т а т и . *В результаті дослідження виявлено варіацію значень індексу відповідності залежно від розміру тексту, що підкреслює необхідність враховувати довжину тексту під час аналізу та розроблення криптографічних систем, зокрема й у виборі ключів шифрування. Зроблено висновок, що розмір тексту може впливати на результати розрахунків індексів відповідності, але несуттєво впливає на загальну захищеність та ефективність шифрування.*

В и с н о в к и . *У процесі виконання роботи проведено аналіз українського алфавіту з пробілом та його властивостей. Розроблено та реалізовано алгоритм обчислення індексу теоретичної відповідності для цього алфавіту. Експериментально визначено індекс теоретичної відповідності для різних текстів українською мовою за допомогою розробленого алгоритму.*

К л ю ч о в і с л о в а : *індекси теоретичної відповідності, індекс збігу, перший метод Фрідмана, другий метод Фрідмана, комп'ютерна лінгвістика.*

Вступ

В сучасному світі, де кількість та обяг інформації стрімко зростають, виникає нагальна потреба в інструментах, спроможних забезпечити безпеку та конфіденційність обміну даними. Одним із важливих напрямів у цьому контексті є криптографія – наука, яка вивчає методи захисту інформації від несанкціонованого доступу. Серед

ISSN 2707-1758

криптографічних інструментів важливе місце належить індексам теоретичної відповідності (ІТВ), які є криптографічними метриками, що дозволяють оцінити ступінь відповідності між текстами. ІТВ використовують для аналізу й оцінювання якості шифрування, розпізнавання мови, визначення авторства тексту та багатьох інших завдань. У цьому контексті одним із найвідоміших індексів теоретичної відповідності є індекс Шеннона, який базується на концепції ентропії.

Вказані індекси відіграють важливу роль у виявленні слабкостей криптографічних систем і забезпеченні їхньої стійкості. Проте застосування індексів теоретичної відповідності може викликати труднощі, такі як складність обчислення та неоднозначність інтерпретації результатів. Для розв'язання цих питань, розроблення ефективних алгоритмів стає ключовою задачею, що сприятиме автоматизації та вдосконаленню процесу використання індексів теоретичної відповідності в криптографії. ІТВ дозволяють оцінити рівень випадковості та непередбачуваності згенерованих ключів. Вони можуть бути використані для виявлення атак на криптографічні системи, таких як атаки з використанням статистичного аналізу.

Індекси теоретичної відповідності допомагають забезпечити конфіденційність, цілісність і доступність передавання даних. Вони можуть бути використані для оцінювання якості різних криптографічних примітивів, наприклад, хеш-функції. ІТВ дозволяють здійснювати аналіз криптографічних систем із різними алфавітами й символьними системами (Boneh, & Shoup, 2017). Вони можуть застосовуватися для оцінювання якості генерування псевдовипадкових чисел у криптографії. Наприклад, шифр Віженера використовує індекси теоретичної відповідності для аналізу тексту та визначення ключа шифрування. Шифр Цезаря також може піддаватися аналізу за допомогою цих індексів, де вони застосовуються для визначення зсуву алфавіту й відновлення оригінального тексту. Криптосистема Ель-Гамаля також використовує індекси теоретичної відповідності для оцінювання стійкості до криптоаналізу.

Індекси теоретичної відповідності є корисним інструментом для аналізу й оцінювання безпеки криптографічних протоколів, таких як SSL/TLS. Вони допомагають виявити потенційні вразливості у протоколах, які використовують асиметричні шифри й обмін ключами.

Застосування індексів теоретичної відповідності дозволяє розробникам криптографічних систем покращувати їхню стійкість і надійність, аналізуючи оптимізацію параметрів та алгоритмів шифрування.

Проте виникають певні труднощі, наприклад, складність обчислення індексів для великих текстів або об'ємних даних. Також може виникнути неоднозначність в інтерпретації результатів через різні методології обчислення та відсутність єдиного стандарту для їхнього розрахунку. Проблема також полягає у визначенні оптимального порогового значення індексу, яке вказує на ступінь відповідності. Загалом, використання індексів теоретичної відповідності у криптографічних системах вимагає ретельного вивчення та розробок для забезпечення їхньої стійкості до криптоаналізу. Реалізація алгоритму сприятиме розвитку комп'ютерного аналізу мови та статистичного моделювання в текстових даних. Впровадження алгоритму дозволить зробити процес обчислення індексів теоретичної відповідності ефективнішим і швидшим, як показано у роботі A Graduate Course in Applied Cryptography (https://toc.cryptobook.us/).

Мета. Алгоритм знаходження індексів можна використати для покращення якості машинного перекладу й автоматичного оброблення текстів.

Наявність алгоритму спростить порівняння різних мов або діалектів на основі їхньої відповідності теоретичним моделям. Розроблення алгоритму викликає необхідність удосконалення методів оброблення тексту та статистичного аналізу даних (Rivest, Shamir, & Adleman, 1978, pp. 120–126).

Алгоритм знаходження індексів теоретичної відповідності є актуальним завданням у сучасних наукових і прикладних дисциплінах, що вивчають текстову інформацію.

Методи

Індекси теоретичної відповідності мають широке застосування, але найчастіше їх використовують у першому методі Фрідмана у процесі обчислення розміру ключа для шифру Віженера, який використовують у вивченні базових понять криптографічних систем. Розглянуто загальний алгоритм знаходження індексів теоретичної відповідності для українського алфавіту з пробілом (УАзП) на основі аналізу великої вибірки текстів.

Для розрахунків проаналізовано 700 текстів, кількість елементів яких становить 1500. Тексти обрано за допомогою API із загальнодоступної вільної багатомовної онлайн-енциклопедії Wikipedia. Розрахунок виконували для ключів розміру від 2 до 7. Індекс теоретичної відповідності розраховано у два етапи. На першому етапі знайдено практичні індекси відповідності для всіх текстів, а далі за допомогою стандартного відхилення вибірки обчислено проміжок найпов-

торюваніших показників. Програмний код для реалізації алгоритму обчислення індексу представлено у вигляді компонентів проєкту, які відповідають за різні аспекти процесу.

Результати

Загальний алгоритм знаходження індексів відповідності для українського алфавіту з пробілом передбачає такі кроки:

1. Зібрати достатньо велику вибірку текстів українською мовою, які містять пробіли. Це можуть бути літературні тексти, новини, наукові статті тощо. Важливо, щоб вибірка була репрезентативною і включала різноманітні тематики.

2. Знайти кількість літер УАзП. Український алфавіт складається з 33 літер, але з пробілом ця кількість становитиме 34.

3. Розбити кожен текст на окремі елементи, включно з пробілами. Кожен елемент може бути окремим символом або комбінацією символів, наприклад, літера з пробілом.

4. Порахувати частоту входження кожного елемента у вибірці текстів. Це можна зробити, складаючи таблицю, де стовпці відповідають символам, а рядки – кожному тексту у вибірці. Кожна комірка таблиці міститиме кількість входжень певного символу відповідно до тексту.

5. Обчислити частоту входження кожного символу у всій вибірці, підсумовуючи відповідні значення з усіх текстів.

6. Розрахувати загальну частоту входження символів із пробілом, яка дорівнюватиме частоті входження пробілу плюс сумі частот входження інших символів із пробілом.

7. Обчислити індекс відповідності для кожного символу, застосовуючи формулу

$$\text{Індекс} = \frac{\text{частота входження символу}}{\text{загальна частота входження символів з пробілом}}. \quad (1)$$

8. Порівняти отримані індекси відповідності для всіх символів і визначити ті, які мають найбільшу відповідність.

9. Виокремити проміжок найбільш повторюваних індексів відповідності. Цей проміжок вказуватиме на теоретичний індекс відповідності для УАзП.

10. За необхідності, виконати додаткову статистичне оброблення, наприклад, розрахувати середнє значення і стандартне відхилення для отриманих індексів відповідності.

11. Перевірити алгоритм на інших вибірках текстів для підтвердження його ефективності і стабільності результатів.

Цей загальний алгоритм дозволяє знаходити індекси відповідності для українського алфавіту з пробілом на основі аналізу великої вибірки текстів.

Отриманий теоретичний індекс відповідності можна використати для подальших розрахунків у криптографії та інших сферах, де необхідно враховувати особливості української мови з пробілом.

Програмний код для реалізації алгоритму обчислення індексу теоретичної відповідності (IBPr) представлено у вигляді компонентів проєкту, які відповідають за різні аспекти процесу:

1. Text.py: Цей клас відповідає за отримання випадкового українського тексту з Вікіпедії, його очищення й обчислення IBPr для заданого ключа з використанням шифру Віженера.

2. Settings.py: У цьому класі зберігаються всі конфігураційні параметри проєкту, такі як український алфавіт, мінімальна та максимальна довжина ключа, параметри API Вікіпедії тощо.

3. Cipher.py: Модуль, який надає функції для шифрування та дешифрування текстів за допомогою шифру Віженера. Включає функції для генерування випадкового ключа та виконання операцій шифрування та дешифрування.

4. IBpr.py: Модуль, який містить функцію для обчислення IBPr для заданого тексту. IBPr обчислюють як метрику для визначення ступеня відповідності тексту.

5. Main.py: Цей скрипт використовує згадані вище компоненти для генерування таблиці IBPr. Програма застосовує бібліотеки та модулі, такі як json, logging, numpy, tqdm, matplotlib, defaultdict та colorama. У функції main() відбувається організація виконання програми, яка отримує текст, генерує випадковий ключ певної довжини, обчислює індекс теоретичної відповідності для кожної довжини ключа й виводить графічне представлення результатів.

За основу для розрахунку взято аналіз 700 текстів, кількість елементів яких становить 1500, отриманих із Wikipedia за допомогою API. Розрахунок відбувається для ключів розміром від 2 до 7 (рис. 1, 2).

Обчислені значення індексу теоретичної відповідності зберігаються у відповідному форматі, а також виводяться графічні представлення результатів за допомогою matplotlib. Логування виконують за допомогою модуля logging, крім того, використовують бібліотеку colorama для кольорового форматування виводу. Результати також зберігають у JSON-файлі.

Ми знаходимо практичні індекси відповідності для всіх текстів. І за допомогою стандартного відхилення вибірки одержуємо проміжок найповторюваніших показників. Це і є наш індекс теоретичної відповідності (рис. 3, 4).

ISSN 2707-1758

```
# Set the amount of texts to analyze
self.amount_of_texts = 700

# Set the minimum size of the text to analyze
self.min_text_size = 1500

# Set the maximum and minimum lengths of the key to use in the Vigenère cipher
self.max_key_length = 7
self.min_key_length = 2
```

Рис. 1. Задання параметрів для аналізу текстів

```
/Users/admin/Desktop/IBPrTable/venv/bin/python /Users/admin/Desktop/IBPrTable/main.py
INFO:root:START...
  0%|          | 0/700 [00:00<?, ?it/s]INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/w/api.php
INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/w/api.php?action=query&prop=extracts&titles=Ігна
INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/w/api.php?action=query&prop=extracts&titles=Дюра
INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/w/api.php?action=query&prop=extracts&titles=Нор
INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/w/api.php?action=query&prop=extracts&titles=Гор
INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/w/api.php?action=query&prop=extracts&titles=Шу-
  0%|          | 1/700 [00:02<30:36,  2.63s/it]INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/
INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/w/api.php?action=query&prop=extracts&titles=Шир
INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/w/api.php?action=query&prop=extracts&titles=Гер
INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/w/api.php?action=query&prop=extracts&titles=Con
INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/w/api.php?action=query&prop=extracts&titles=Спи
  0%|          | 2/700 [00:04<25:33,  2.20s/it]INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/
INFO:wikipediaapi:Request URL: https://uk.wikipedia.org/w/api.php?action=query&prop=extracts&titles=Вік
Run  Python Packages  TODO  Python Console  Problems  Terminal  Services
```

Рис. 2. Виконання програмного коду

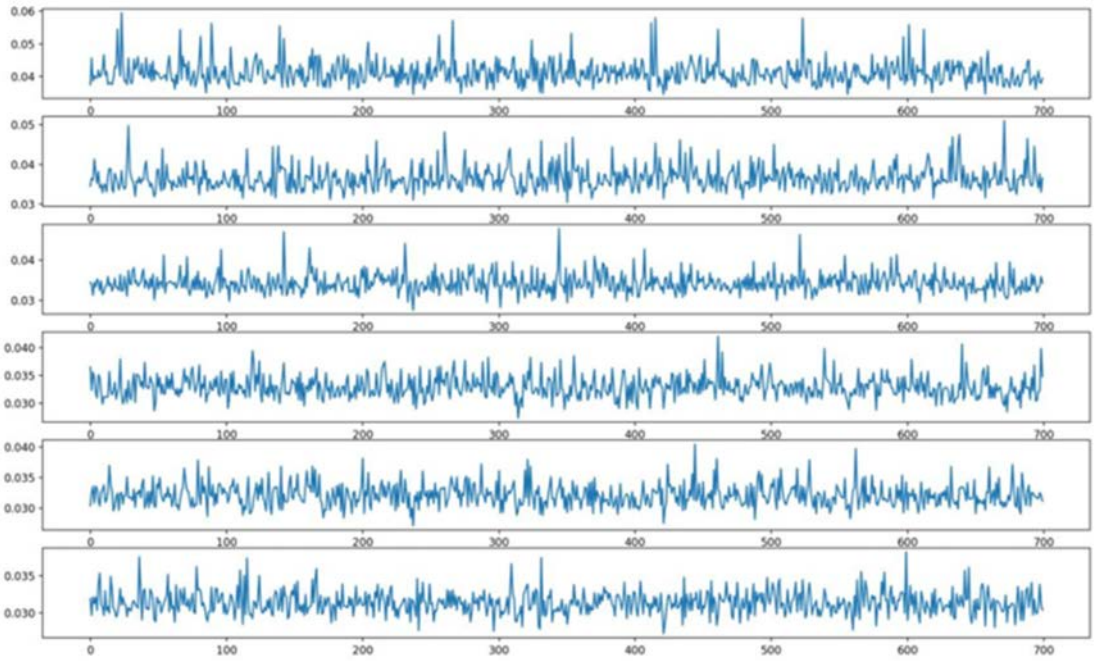

Рис. 3. Графік індексів практичної відповідності для всіх текстів розміром 1500 символів

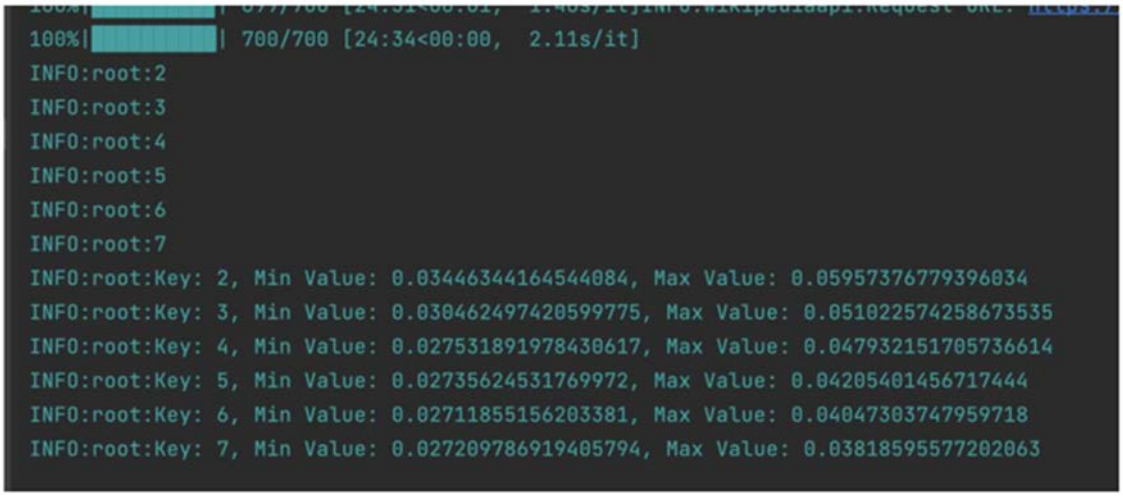

Рис. 4. Показники індексів теоретичної відповідності

Індекси теоретичної відповідності для українського алфавіту з пробілом відкривають широкі можливості для застосування в різних галузях. Наведемо деякі потенційні можливості і сфери застосування цих індексів:

1. Криптографія. ІТВ можуть бути використані для аналізу шифрованих текстів і встановлення можливості розшифрування. Вони допомагають визначити розподіл символів у шифрованому тексті та порівняти його з відомим розподілом символів у мові відкритого тексту.

2. Мовознавство. Індекси теоретичної відповідності застосовують для вивчення структури та властивостей української мови. Вони дозволяють оцінити ступінь відповідності між розподілом символів у тексті та типовим розподілом символів у мові.

3. Літературознавство. За допомогою ІТВ можна проводити аналіз текстів літературних творів. Вони дозволяють виявляти особливості авторського стилю, використання символіки та лексики.

4. Контентний аналіз. Індекси теоретичної відповідності можуть бути використані для оцінювання схожості або розбіжності між різними документами чи текстовими джерелами. Вони дозволяють проводити порівняльний аналіз, виявляти патерни й розробляти алгоритми автоматичного пошуку.

5. Аналіз соціальних мереж. В аналізі комунікаційних мереж, які складаються з повідомлень, коментарів та інших текстових даних, вони допомагають виявляти схожість між користувачами, групами або темами обговорень.

6. Виявлення плагіату. ІТВ допомагають порівнювати семантичну схожість між текстами та виявляти спільні фрази, що свідчать про можливе копіювання (Gorkavenko, Popova, & Tarasenko, 2019, p. 24).

7. Аналіз електронної пошти. ІТВ можна використати для аналізу текстових повідомлень, електронних листів та інших форм комунікації через електронну пошту.

8. Аналіз настрою та сентимент-аналіз. ІТВ застосовують для визначення настрою або емоційного відтінку тексту. Вони допомагають виявляти позитивний, негативний або нейтральний характер тексту, що має значення для аналізу відгуків користувачів і соціальних медіа.

9. Медичний аналіз. Можуть бути використані для аналізу медичних записів, пацієнтських оглядів та інших медичних документів. Вони допомагають виявляти зв'язки між симптомами, діагнозами й лікуванням.

10. Фінансовий аналіз. ІТВ можуть бути застосовані для аналізу фінансових звітів, економічних новин та інших фінансових документів. Допомагають виявляти тренди, патерни й ризики у фінансових ринках і підприємствах.

Наведені потенційні застосування індексів теоретичної відповідності показують, що вони можуть бути корисними інструментами для аналізу та розуміння текстових даних у різних галузях.

З розрахунків можна зробити такі висновки щодо різниці результатів залежно від розміру тексту. Збільшення розміру тексту з 1500 до 2000 символів призвело до деякого зниження значень індексу відповідності для всіх розглянутих ключів. Це може вказувати на більшу варіативність і розподіленість символів у більших текстах, що зменшує ступінь повторюваності та структурованості.

Хоча величина змін значень індексу відповідності не є значною, вона показує, що довжина тексту може впливати на ступінь подібності між українським алфавітом із пробілом і текстами, що

ISSN 2707-1758

аналізуються. Більші розміри текстів можуть призводити до меншої впевненості в точності індексів теоретичної відповідності.

Незважаючи на зниження значень індексу відповідності, результати залишаються в межах прийнятних значень для шифрування та забезпечення високого рівня захищеності текстів.

Варіація значень індексу відповідності залежно від розміру тексту підкреслює необхідність враховувати довжину тексту під час аналізу та розроблення криптографічних систем, зокрема й у виборі ключів шифрування.

Отже, розмір тексту може впливати на результати розрахунків індексів відповідності, але несуттєво впливає на загальну захищеність та ефективність шифрування (Rivest, Shamir, & Adleman 1978, pp. 120–126).

Дискусія і висновки

У процесі виконання роботи проведено аналіз українського алфавіту з пробілом і його властивостей. Розроблено й реалізовано алгоритм обчислення індексу теоретичної відповідності для цього алфавіту. Експериментально визначено індекс теоретичної відповідності для різних текстів українською мовою за допомогою розробленого алгоритму.

Отримані результати проаналізовано і зроблено висновки щодо їхньої значущості та застосування у криптографії та лінгвістиці. Ці результати дозволяють зрозуміти ступінь відповідності мовних текстів українській мові з використанням вказаного алфавіту. Вони можуть бути використані для побудови ефективних криптографічних систем, а також для дослідження мовних особливостей і стилістичних відмінностей української мови.

Внесок авторів: Лариса Мирутенко – концептуалізація; методологія; аналіз джерел; Яніна Шестак – підготування огляду літератури або теоретичних засад дослідження; Анастасія Лобан – збір емпіричних даних та їх валідація; емпіричне дослідження.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

Gorkavenko, V. M., Popova, T. G., & Tarasenko, O. V. (2019). *Mathematical Linguistics: a textbook*. Publishing House "Prosvita".

Rivest, R., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21(2), 120–126.

REFERENCES

Gorkavenko, V. M., Popova, T. G., & Tarasenko, O. V. (2019). *Mathematical Linguistics: a textbook*. Publishing House "Prosvita".

Rivest, R., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21(2), 120–126.

Отримано редакцією журналу / Received: 10.03.24
Прорецензовано / Revised: 29.03.24
Схвалено до друку / Accepted: 13.05.24

Larysa MYRUTENKO, PhD (Engin.), доц.
ORCID ID: 0000-0001-8538-8996
e-mail: myrutenko.lara@gmail.com
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Yanina SHESTAK, PhD (Engin.), Assist.
ORCID ID: 0000-0002-7291-1829
e-mail: luchenko@ukr.net
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Anastasiia LOBAN, Student
ORCID ID: 0009-0000-3828-6663
e-mail: anastasialoban0912@gmail.com
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

RESEARCH OF INDICES OF THEORETICAL CORRESPONDENCE FOR THE UKRAINIAN ALPHABET WITH SPACES

Background. *Theoretical compatibility indices play an important role in identifying the weaknesses of cryptographic systems and ensuring their stability. After analyzing the available sources of information, this indicator was not found for the Ukrainian alphabet with a space, which makes it impossible to find accurate results in calculations where this indicator is used.*

Methods. *Theoretical correspondence indices have a wide range of applications, but they are most often used in Friedman's First Method when calculating the key size for the Vigener cipher, which is used in the study of the basic concepts of cryptographic systems. The general algorithm for finding indices of theoretical correspondence for the Ukrainian alphabet with a space based on the analysis of a large sample of texts is considered.*

Results. *For calculations, 700 texts were analyzed, the size of which is 1500 elements. The texts were selected using an API from Wikipedia, a free and open multilingual online encyclopedia. The calculation took place for keys of size from 2 to 7. The index of theoretical correspondence was calculated in two stages. At the first stage, practical indices of correspondence were calculated for the whole range of texts, and then, using the standard deviation of the sample, the interval of the most repeated indicators was calculated. The program code for implementing the index calculation algorithm was presented in the form of project components responsible for various aspects of the process.*

As a result of the study, a variation of the values of the correspondence index was found depending on the size of the text, which emphasizes the need to consider the length of the text in the analysis and development of cryptographic systems, in particular, the selection of encryption keys. It was concluded that the size of the text can affect the results of the calculation of the compatibility indices, but it does not significantly affect the overall security and effectiveness of encryption.

Conclusions. *In the course of the work, an analysis of the Ukrainian alphabet with a space and its properties was carried out. An algorithm for calculating the index of theoretical correspondence for this alphabet was also developed and implemented. The index of theoretical correspondence was experimentally determined for various texts in the Ukrainian language using the developed algorithm.*

Keywords: *indices of theoretical correspondence, index of coincidence, First Friedman method, Second Friedman method; computational linguistics.*

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.