МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕПАРТАМЕНТ ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ
ВИКОНАВЧОГО ОРГАНУ КИЇВСЬКОЇ МІСЬКОЇ РАДИ
(КИЇВСЬКОЇ МІСЬКОЇ ДЕРЖАВНОЇ АДМІНІСТРАЦІЇ)
КИЇВСЬКЕ ТЕРИТОРІАЛЬНЕ ВІДДІЛЕННЯ МАЛОЇ АКАДЕМІЇ НАУК УКРАЇНИ
КИЇВСЬКА МАЛА АКАДЕМІЯ НАУК УЧНІВСЬКОЇ МОЛОДІ

Відділення: математики
Секція: прикладна математика

ОЦІНКА ЕФЕКТИВНОСТІ КРИТЕРІЇВ ЗМІСТОВНОГО ТЕКСТУ У ЗАДАЧАХ КРИПТОГРАФІЇ

РОБОТУ ВИКОНАВ:
Шульженко Артем Борисович,
учень 11 класу
Політехнічного ліцею НТУУ "КПІ"
м. Києва

Науковий керівник:
Яковлєв Сергій Володимирович,
кандидат технічничних наук, доцент
кафедри ММЗІ ФТІ “КПІ” ім. Ігоря
Сікорського

КИЇВ – 2018

ОЦІНКА ЕФЕКТИВНОСТІ КРИТЕРІЇВ ЗМІСТОВНОГО ТЕКСТУ У ЗАДАЧАХ КРИПТОГРАФІЇ

Шульженко Артем Борисович; Київське територіальне відділення МАНУ; КПНЗ «Київська Мала академія наук учнівської молоді»; Політехнічний Ліцей НТУУ «КПІ»; 11 клас; м. Київ; Яковлєв Сергій Володимирович; кандидат технічних наук, доцент кафедри ММЗІ Фізико-технічного інституту КПІ ім. Ігоря Сікорського.

Дана робота присвячена шифрам підстановки та оцінюванню практичної ефективності критеріїв визначення змістовності текстів, які використовуються під час криптоаналізу таких шифрів.

Метою роботи аналіз та вдосконалення підходів до визначення змістовності текстів, що дозволить підвищити практичну ефективність методів криптоаналізу.

У першому розділі роботи наведено теоретичні відомості про найбільш уживані шифри підстановки (шифр Цезаря, шифр Віженера, шифр афінної заміни, шифр Хілла), методи атаки на них та історична довідка щодо їх використання.

У другому розділі на основі певних розрахункових параметрів сформульовано п’ять груп критеріїв змістовності текстів для англійської мови, після чого проведено експериментальну перевірку ефективності запропонованих критеріїв при криптоаналізі шифрів підстановки. Встановлено, що для шифротекстів довжини у 300-700 символів дані критерії дозволяють майже безпомилково розрізнити змістовний текст від незмістовного; найкращі результати як за довжиною текстів, так і за імовірністю помилки показав критерій на основі заборонених біграм.

Одержані результати свідчать про високу точність запропонованих критеріїв та можливість їх застосування для перевірки на змістовність текстів навіть невеликої довжини.

ЗМІСТ

ВСТУП

З плином часу змінюється та вдосконалюється все, до чого людина прикладає свою руку, і криптографія не стала винятков. Ще кілька десятиліть тому людство створювало моделі та методи захисту даних, які сьогодні вже вважаються ненадійними. Це ставить нові задачі для криптоаналітиків: якщо алгоритм не є надійним, але його злам вимагає певного великого перебору ключів, то потрібно автоматизувати процес зламу, оскільки проводити «ручну» атаку зазвичай надзвичайно незручно та потребує багато часу. Проте у процесі зламу шифру постає питання: як саме відрізняти випадкові тексти від змістовних? Для розв'язання цієї проблеми було запропоновано ряд методів, які грунтуються на статистичних властивостях змістовного тексту, оскільки останні зазвичай є дуже нерівномірними та мають велику надлишковість. Однак одні й ті самі методи при аналізі різних шифрів можуть давати зовсім різні результати, що зумовлюється особливостями алгоритмів шифрування; аналітичне передбачення поведінки методів є дуже складною задачею, яка досі повністю не розв'язана. Тому тема даного дослідження є актуальною та має практичну цінність.

Метою роботи є аналіз та вдосконалення підходів до визначення змістовності текстів, що дозволить підвищити практичну ефективність методів криптоаналізу.

Досягнення поставленої мети передбачає наступні **завдання дослідження**, які були виконані в роботі:

1) ознайомлення з основними принципами криптографії та основними типами шифрів підстановки;

2) ознайомлення з методами криптоаналізу шифрів підстановки;

3) аналіз характеристик англійської мови та формулювання на основі проведеного аналізу різних критеріїв змістовності англійських текстів;

4) експериментальна перевірка ефективності кожного з запропонованих критеріїв для різних шифрів пістановки, зокрема, визначення імовірності помилок критерію в залежності від довжини шифротексту.

Об’єктом дослідження стали інформаційні процеси в системах криптографічного захисту.

Предметом дослідження стали моделі змістовних текстів у криптографічних задачах.

При розв’язанні поставлених завдань використовувались такі **методи дослідження**: методи лінійної та абстрактної алгебри, теорії імовірностей, математичної статистики, методи комп’ютерного моделювання.

Наукова новизна роботи полягає у розвитку статистичних методів визначення змістовності текстів у задачах криптографії та емпіричному аналізу ефективності найбільш використовуваних критеріїв змістовності повідомлень.

Результати дослідження можуть бути використанні для покращення роботи програм розпізнавання текстів, а також при створенні програм автоматичного зламу шифрів підстановки.

РОЗДІЛ І
ОЗНАЙОМЛЕННЯ З ШИФРАМИ

1.1. Класичні шифри та їх розвиток

Криптографія — наука про математичні методи забезпечення конфіденційності, автентифікації та цілісності даних. Вона існує вже тисячі років, зародившись ще в стародавньому світі, за часів єгипетських фараонів, близько двох тисяч років до н.є. З тих пір наука невпинно розвивалася, видозмінюючись та рухаючись від одних методів до інших. Умовно можна поділити періоди еволюції криптографії (точніше, шифрів) на 4 етапи:

1) Моноалфавітний та перестановчий — перший і найдовший. Ознаменував початок розвитку криптографії як науки. Під час цього періоду домінували шифри простої підстановки та заміни. Характерна риса таких шифрів — створення нової абетки, де кожній літері шифротексту відповідала одна літера відкритого тексту. Почали з’являтися шифри перестановки, ідея яких полягала у перестановці літер всередині повідомлення.

2) Поліалфавітний — перехід до більш ускладнених варіантів (наприклад, застосування різних шифрів до різних частин повідомлення). До перших робіт цього періоду відносять праці арабського філософа та математика Аль-Кінді [1](IX ст. н.є.) на Близькому Сході та видатного вченого Леона Альберті [2](XV ст. н.є.) у Європі. Криптографія починає ставати невід’ємною частиною життя високопосадовців.

3) Механічний — перехідний етап в історії криптографії. Хоча поліалфавітні шифри й продовжували використовуватися, але на допомогу «ручному» шифруванню прийшли механічно-електронні прилади, найвідомішою з яких є роторна машина «Енігма», яка використовувалась Німеччиною до кінця Другої світової війни. Вдала атака на «Енігму» за допомогою машини «Бомба»,

сконструйована групою британських математиків на чолі з видатним криптографом Аланом Тюрінгом, допомогла значно пришвидшити перемогу сил Союзників.

4) Сучасний — перехід до математичної криптографії. З'являються блочні та поточні шифри. Саме шифрування виходить на абсолютно новий рівень та починає застосовуватись майже у всіх сферах людської діяльності. Початок можна віднести до 1949 року, коли було офіційно опубліковано фундаментальну працю Клода Шеннона «Теорія зв'язку у секретних системах» [3]. Ця робота є переломним моментом, оскільки саме в ній було вперше представлено криптографію як одну з математичних наук.

Основні принципи криптографії були сформульовані ще на кінці XIX ст. голландським вченим Огюстом Керкгоффзом [4]. Наведемо їх короткий опис.

1) Система має не піддаватися атакам зловмисників, якщо не математично, то хоча б практично.

2) Система не повинна потребувати секретності, потрапляння системи до рук зловмисників не має бути проблемою (максимум Шеннона: ворог знає систему).

3) Користувачі повинні мати змогу спілкуватися та пам'ятати ключ без допоміжних записів, при необхідності легко змінити ключ.

4) Система повинна бути такою, на її обслуговування повинно вистачати однієї людини.

5) Система повинна мати можливість бути використаною при телеграфному листуванні.

6) Користувачу повинно бути легкою у використанні, її користування не повинно потребувати списку правил.

Очевидний наслідок з цих принципів, що також інколи називають принципом Керкгоффза, звучить так: складність системи повинна залежати від складності ключа, а не від секретності алгоритмів, що можуть бути відкриті публічно, без втрати стійкості алгоритмів. Ці правила, хоч і видозмінені, до сього дня використовуються у алгоритмах передачі даних.

1.2. Класифікація шифрів перестановки

Шифри перестановки та заміни — одні з найперших спроб людства захистити повідомлення. Перший характеризується зміщенням символів чи групи символів у тексті згідно певних правил, при чому обидва випадки можна комбінувати, наприклад змішуючи біграми та переставляючи літери у біграмах. Одним з типових представників перестановочних шифрів є анаграми. Шифр заміни має в основі іншу ідею, замість переставляння літер у повідомленні він замінює їх на нові. Традиційно, ці шифри поділяють на 4 види.

1) Моноалфавітний шифр – у такому шифрі кожній літері присвоюється нове значення з того ж самого алфавіту. Типовим представником такого шифру є, наприклад, афінний шифр [5].

2) Поліграмний шифр – алгоритму цього типу замість одного символу починають оперувати одразу групами символів, багаторазово збільшуючи кількість можливих ключів у порівнянні з моноалфавітним. Прикладом такого шифру є шифр Плейфера [6].

3) Омофонічний шифр – такий шифр протидіє частотному аналізу шляхом створення множини допустимих символів для кожного символу відкритого тексту. Таким чином при повторенні знаку у шифротексті з’являється два різних символи, що допомагає скрити справжню частоту тієї чи іншої літери. Представником таких шифрів є квадрат Полібія 2×2 [7].

4) Поліалфавітний шифр – його суть полягає у циклічному застосуванні кількох моноалфавітних шифрів в одному повідомленні. Перший символ шифрується за допомогою першого алфавіту, другий – другого, і так допоки не закінчаться шифруючі алфавіти, після чого процедура повторюється до тих пір, коли усе повідомлення не буде зашифроване. Прикладом такого шифру є шифри Віженера та Бофора [8].

1.3. Шифр Цезаря

Шифр Цезаря [9] – один з найдавніших задокументованих шифрів, названий на честь римського диктатора Гая Юлія Цезаря, що одним з перших почав його використовувати. Являється одним з найпростіших моноалфавітних шифрів. Незважаючи на це, у стародавні часи його надійність можна оцінити як достатню, оскільки рівень обізнаності ворогів Цезаря був малим.

Математично цей алгоритм можна записати таким чином:

$$y = (x + k) \bmod N$$
$$x = (y - k + n) \bmod N,$$

де N — кількість символів у алфавіті, k — ключ шифру, x та y — символи відритого і закритого тексту відповідно. Для застосування шифру необхідно пронумерувати кожний символ алфавіту певним значенням, після чого робити криптографічні перетворення (зазвичай літери нумерують у алфавітному порядку, починаючи з нуля).

Оскільки існує всього $N - 1$ варіантів ключа, що в середньому складає від 25 до 35 символів, то шифр можна зламати шляхом повного перебору усіх можливих ключів. Як приклад можна навести такий спосіб: на аркуші у стовпчик виписується увесь алфавіт, кількість аркушів рівна кількості символів у повідомлені (насправді, їх кількість можна зменшити до “відстані унікальності”). Після цього стрічки складаються поруч, щоб одна з горизонтальних ліній літер утворили зашифроване повідомлення. Достатньо прочитати усі інші утворені варіанти, і якщо довжина розглянутого шифротексту більше відстані унікальності, то можна буде однозначно визначити зміст шифротексту (оскільки змістовний варіант буде єдиним). Відстань від шифротексту до відкритого тексту буде ключем шифру (за модулем довжини алфавіту).

Відстань однозначності [10] — мінімальна відстань, що необхідна для однозначного визначення ключа шифру перестановки при атаці шляхом повного перебору. Значення відстані однозначності є константним для кожної мови та кожного алгоритму шифрування. Приблизне значення обчислюється за формулою $U = H(k)/D$, де U – відстань однозначності, $H(k)$ – ентропія множини ключів (рівна бінарному логарифму від кількості ключів), а D – надлишковість мови, тобто різниця бінарного логарифму від кількості літер у абетці та кількістю інформації, що несе в собі один символ (для природніх мов знаходиться в границях від 0.6 до 1.5).

Слід зазначити, що метод лише постулює мінімальну відстань, на якій ключ є єдиним можливим, але не дає жодних вказівок щодо його знаходження.

1.4. Шифр Віженера

Шифр Віженера [11] — один з поліалфавітних шифрів, що отримав назву на честь французького дипломата та криптографа Блеза Віженера. Не зважаючи на досить легке формулювання та реалізацію, даний шифр досить довго (понад 300 років) не мав загального методу атаки, та навіть називався французами “незламним”.

Вперше був опублікований у 1553 році італійцем Джованом Баттістою [7]. Незважаючи на це, історія забула автора шифру та надала її французу. Цей момент згадав та осудив Девід Кан: “ … шифр був названий іменем Віженера, хоча той не приклав жодних зусиль до його створення” [7, с. 15].

У XIX ст. Фрідріх Касіскі довів [7], що шифр можна зламати та опублікував загальну атаку. Вже у XX ст. Гільберт Вернам провів роботу над шифром та представив новий шифр [3], що отримав назву одноразових блокнотів (інакше відомий як шифр Вернама), для якого пізніше було доведено абсолютно криптостійкість (тобто, що його не можливо зламати) при виконанні ряду правил (абсолютна випадковість ключів (їх розмір повинен бути рівний довжині тексту) та їх одноразовість).

Ключом шифру Віженера є слово, у випадку, коли довжина слова менше тексту, воно продовжується записуватися допоки не скінчиться текст. Тепер кожній літері відкритого тексту відповідає одна з букв ключового слова. Порядковий номер літери і є ключом для даного символу відкритого тексту.

$$c_i = (m_i + k_i) \, mod \, N$$

де N – розмір алфавіту, m_i – номер символу відкритого тексту, k_i – літери ключа.

Атака на шифр Віженера умовно поділяється на два етапи:

1) пошук довжини ключа;
2) послідовна атака на кожен з алфавітів.

Під час першого етапу відбувається пошук довжини ключа проводиться шляхом поступового розбиття тексту. Спочатку аналізується випадок з довжиною ключа 1. Рахується ентропія повідомлення, у випадку, коли значення дуже відрізняється від випадкового, робиться висновок що розглянута довжина ключа і є правильною. Якщо ж значення близьке до випадкового, то довжина ключа підвищується на 1, після чого знову відбувається розбиття шифротексту на різні підблоки, які повинні шифруватись окремими літерами ключа, для яких обчислюється ентропія. Даний процес повторюється до тих пір, допоки не буде знайдено довжину ключа, після чого атака переходить на наступний етап.

Оскільки шифрування відбувається за допомогою алгоритму, що нічим не відрізняється від шифру Цезаря, то атака на другому етапі зводиться до атаки на шифр Цезаря, що було розглянуто у підрозділі 1.4.

1.5. Біграмний афінний шифр

Біграмний афінний шифр [5] – представник поліграмних шифрів, що оперує біграмами відкритого тексту.

Даний алгоритм являє собою наступне перетворення. Нехай x_1, x_2, x_3, x_4 ... — символи відкритого тексту. Об'єднаємо символи у біграми $(\mathrm{x}_1, \mathrm{x}_2), (x_3, x_4), \ldots$ і пронумеруємо літери алфавіту від 0 до $N-1$. Таким чином кожній біграмі ми зможемо співставити число в інтервалі від 0 до $N^2 - 1$:

$$(x_{2i-1}, x_{2i+1}) \leftrightarrow X_i = x_{2i-1} * N + x_{2i}$$

Шифрування відбувається незалежно для кожної біграми за такою формулою:

$$Y_i = (a * X_i + b)\ mod\ N^2$$

де Y_i — порядковий номер біграми шифротексту, а X_i — порядковий номер біграми відкритого тексту, $0 < a < N^2$ – взаємнопросте з N^2 число, $0 \leq b < N^2$ - довільне число у цих границях.

Атака на біграмний афінний шифр проводиться за допомогою частотного аналізу біграм тексту та мови. Нехай після своїх спостережень криптоаналітик знайшов, що:

$$X^* \rightarrow Y^*,$$
$$X^{**} \rightarrow Y^{**},$$

тоді, для невідомих a, b можна скласти таку систему рівнянь:

$$\begin{cases} Y^* \equiv a * X^* + b\ mod\ N^2 \\ Y^{**} \equiv a * X^{**} + b\ mod\ N^2 \end{cases}$$

Звідки маємо рівняння для визначення a:

$$Y^{**} - Y^* \equiv a * (X^{**} - X^*)\ mod\ N^2$$

Дане рівняння може мати кілька коренів, для кожного з яких необхідно визначити коефіцієнт b:

$$b = Y^* - a * X^* \, mod \, N^2$$

Остаточно ключ шифрування знаходиться шляхом застосування пари ключів до решти тексту.

Міркування щодо значень біграм Y^*, Y^{**} можемо отримати, наприклад, порівнюючи частоти біграм у мові та шифротексті. Треба зауважити, що отримані дані не завжди є вірними, цілком можлива ситуація коли найбільш часті біграми у шифротексті не є найчастішими у мові. У такому випадку обирається інші, найбільш вірогідніші кандидати, для яких знову застосовується даний алгоритм.

1.6. Шифр Хілла

Шифр Хілла [12] – поліграмний шифр, оснований на лінійній алгебрі та модульній арифметиці, що був винайдений американським математиком Лестором Хіллом 1929 року. Це був перший шифр, що на практиці дозволив оперувати більше ніж трьома символами. Алгоритм не знайшов великої популярності у зв'язку з відсутністю алгоритмів генерації прямих та обернених матриць великого розміру. Девід Кан так описав Хілла та його шифр: "Хілл самотужки розробив потужний метод та вперше зробив поліграмну криптографію практичною" [7, с. 65].

Ідея шифру Хілла полягає у представленні n-грами у вигляді n-вимірного вектора. Саме ж шифрування відбувається шляхом множення отриманого вектора на матрицю розміру $n \times n$. Матриця, яка є ключем шифру, повинна бути оборотною у Z_{26}^n, у противному випадку шифротекст неможливо буде дешифрувати.

Для випадку $n = 2$ шифрування виглядає наступним чином:

$$\begin{pmatrix} c_1 \\ c_1 \end{pmatrix} = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} * \begin{matrix} p_1 \\ p_2 \end{matrix} \, (mod \, N)$$

Дане рівняння можна записати у спрощеному вигляді:

$$\begin{cases} c_1 = k_{11}p_1 + k_{12}p_2 \ mod\ N \\ c_2 = k_{21}p_1 + k_{22}p_2 \ mod\ N \end{cases}$$

Перші два рівняння описують саме шифрування, третя ж рівність задає умову для ключа-матриці, що буде оборотною тоді і тільки тоді, коли її детермінант буде взаємнопростим з модулем.

Атака на шифр Хілла подібна до атаки на афінний шифр і проводиться з тих самих міркувань про перехід однієї біграми у іншу. Єдине суттєве розходження виникає лише при зіставленні системи рівнянь – замість двох ми отримуємо чотири (по два на кожну біграму).

РОЗДІЛ ІІ
ФОРМУЛЮВАННЯ ТА ОЦІНКА КРИТЕРІЇВ

2.1. Групові властивості шифрів

На сучасному етапі розвитку технологій багато шифрів більш не потребують ручної роботи, а сам алгоритм атаки автоматизують. В результаті його роботи криптоаналітик отримує певний текст (змістовний при вгаданому ключі, чи незмістовний при неправильному ключі). Проблема автоматизації полягає не в написанні алгоритму атаки, а в перевірці вже отриманого результату на вірність[16]. Саме для цієї задачі і використовують критерії змістовності тексту.

Визначимо групові властивості таким чином [14]:

1) $\forall\, k_1, k_2\colon \exists\, k_3\colon\ E_{k_1}\left(E_{k_2}(x)\right) = E_{k_3}(x)$ – тобто при шифруванні вже зашифрованого тексту ми не підвищуємо складність шифру.

2) $\exists\, k_0\colon E_{k_0}(x) = x$ – тобто, для шифра існує такий ключ, що виконує тотожне перетворення.

При виконанні цих умов маємо, що функція розшифрування – це шифрування на іншому ключі, а тому статистичну поведінку текстів після розшифрування на неправильному ключі можна оцінювати на шифротекстах.

Розглянемо шифр Цезаря. Нехай маємо символи відкритого тексту $x_1 x_2 x_3 x_4$ ..., тоді:

$$y_i = (x_i + k_1)\ mod\ N$$

При повторному шифруванні отримаємо

$$y_i^* = (y_i + k_2)\ mod\ N = (x_i + k_1 + k_2)\ mod\ N$$

Покладемо $k_1 = k_2 + k_3$, тоді $y_i^* = (x_i + k_3)\ mod\ N$, що і треба було довести.

Покладемо $k_0 = 0$. Тоді $y_i = (x_i + 0)\, mod\ N = x_i$. Таким чином, шифри Цезаря задовольняють груповим властивостям.

Очевидно, що шифр Віженера також задовільняє груповим властивостям, оскільки в його основі лежить застосування шифра Цезаря.

Перейдемо до шифра афінної біграмної заміни. Нехай маємо символи відкритого тексту $x_1x_2x_3x_4$ ..., тоді $y_i = (a_1 * x_i + b_1) mod\ N$. При повторному шифруванні отримаємо:

$$y_i^* = (a_2 * y_i + b_2) mod\ N = (a_1a_2 * x_i + a_2b_1 + b_2) mod\ N.$$

Тоді $a_3 = a_1 * a_2;\ b_3 = a_2 * b_1 + b_2$. Крім того, якщо a_1 та a_2 взаємнопрості з модулем n, то їх добуток також буде взаємнопростим з n, тому визначена таким чином пара (a_3, b_3) є коректним ключем шифру біграмної заміни.

Нехай $a_1 = 1; b_1 = 0$, тоді

$$y_i = (a_1 * x_i + b_1)\, mod\ N = (1 * x_i + 0)\, mod\ N = x_i,$$

таким чином, шифр афінної біграмної заміни задовольняє груповим властивостям.

Розглянемо шифр Хілла. Оскільки у шифруванні кожен символ n-грами впливає на результат шифрування, то $k_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ (одинична матриця).

Покладемо $k_1 = \begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}$, тоді $\begin{matrix} y_1 = a_1 * x_1 + b_1 * x_2 \\ y_2 = c_1 * x_1 + d_1 * x_2 \end{matrix}$. Проведемо шифрування іншим ключем $k_2 = \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix}$ та розкриємо дужки:

$$y_1^* = a_2 * a_1 * x_1 + a_2 * b_1 * x_2 + b_2 * a_1 * x_1 + b_2 * b_1 * x_2$$
$$y_2^* = c_2 * c_1 * x_1 + c_2 * d_1 * x_2 + d_2 * c_1 * x_1 + d_2 * d_1 * x_2$$

Проведемо групування доданків:

$$y_1^* = x_1(a_2a_1 + b_2a_1) + x_2(a_2b_1 + b_2b_1)$$
$$y_2^* = x_1(c_2c_1 + d_2c_1) + x_2(c_2d_1 + d_2d_1)$$

Легко бачити, що це шифрування аналогічне шифруванню ключем

$$k_3 = \begin{bmatrix} a_2a_1 + b_2a_1 & b_2b_1 + a_2b_1 \\ c_2c_1 + d_2c_1 & c_2d_1 + d_2d_1 \end{bmatrix}.$$

Таким чином, усі розглянуті шифри задовольняють групові властивості. Отже, при невдалій спробі дешифрувати повідомлення, зашифроване одним з розглянутіх ключів, ми отримуємо той самий шифр, але з іншим ключем.

2.2 Формулювання критеріїв

Усі шифри підставноки замінюють одні символи іншими, що призводить то зміни частот літер у повідомленні. У будь-якій мові існують більш- та менш-уживані символи, що відображається у надлишковості мов. Було проведено аналіз 26 художніх творів (перелік наведено у Додатку А), виконаних англійською мовою, на основі якого було встановлено частоти англійських літер у текстах.

Таблиця 2.1

Таблиця експериментально визначених частот англійської мови

A – 8%	E –12.16%	I – 7.1%	M – 2.72%	Q – 0.07%	U – 3.01%	Y – 2.4%
B – 1.54%	F – 1.91%	J – 0.22%	N – 6.63%	R – 5.33%	V – 0.87%	
C – 2.28%	G – 2.37%	K – 1.26%	O – 7.66%	S – 6.19%	W – 2.5%	Z – 0.09%
D – 4.55%	H – 6.16%	L – 4.2%	P – 1.65%	T – 8.97%	X – 0.15%	

Легко бачити, що частота п’яти найбільш уживаних літер становить близько 45%. Імовірність відсутності їх у повідомленні довжиною навіть на ста символах буде становити близько $\left(\frac{11}{20}\right)^{100} \cong \frac{1}{2^{100}}$. З цього робимо припущення, що відсутність літер E, T, N, I, A з великою імовірністю свідчить про незмістовність тексту.

Також зазначимо, що надмірна частота цих символів (більше 80%) чи недостатня частота (менше 20%) також свідчиме про імовірну незмістовність тексту.

Як зазначалося раніше, усі мови мають надлишковість, що проявляється у збільшенні появи певних літер. Тому, однією з ознак випадкового тексту буде згладження частот та їх наближення до $1 / N$, де N – кількість літер у алфавіті. Очікувана частота трьох найбільш уживаних літер становить близько 30%, отже якщо їх сумма буде становити менше 18% (3/26 * 1,5), це в багатьох випадках буде свідчити про незмістовність тексту. Не слід забувати і про наявність хоча б однієї голосної серед розглянутих літер.

Тепер проаналізуємо частоти голосних літер. Серед усіх триграм переважна більшість має структуру з однією голосною, на другому місці – з двома голосними, що дає нам змогу зробити припущення щодо вірогідності їх появи у текстах. Так, якщо сумарна частота усіх голосних буде більше за порогове значення у 67% (тобто текст переважно складається з триграм з двома чи трьома голосними) або менше за порогове значення у 30% (тобто наявна велика кількість триграм, що складаються лише з приголосних), текст буде незмістовним.

Очевидно, що й найменшуживані літери не повинні складати значну частину тексту. Очікувана імовірність появи п’яти найменш частих символи складає менше 3%. Враховуючи невелику довжину досліджуваних повідомлень, визначимо порогове значення для цих літер у 13%, що приблизно відповідє кожному восьмому символу тексту.

Іншим, не менш важливим елементом текстів є біграми. Незважаючи на різноманітність мов та їхніх біграм, існують і заборонені комбінації символів. Для англійської мови такий показник складає близько 10% від загальної кількості. Очевидно, що змістовний текст не матиме заборонених біграм, але шифротекст, на відміну від відкритого, такі біграми мати може. Поява такої біграми одразу вкаже на незмістовність розглянутого тексту.

Таблиця 2.2

Перелік заборонених біграм англійської мови

bx	cj	cv	cx	dx	fq	fx	gq	gx	hx
jc	jf	jg	jq	js	jv	jw	jx	jz	kq
kx	mx	px	pz	qb	qc	qd	qf	qg	qh
qj	qk	ql	qm	qn	qp	qs	qt	qv	qw
qx	qy	qz	sx	vb	vf	vh	vj	vm	vp
vq	vt	vw	vx	wx	xk	xx	zj	zq	zx

Розглянемо поняття ентропії на символ джерела текстів [13]:

$$H_1 = -\sum_{i=1}^{n} p(a_i) * \log_2 p(a_i)$$

де $p(a_i)$ – імовірність появи букви алфавіту. Ця величина харатеризує надлишковість мови шляхом підрахунку частот літер для досліджуваного тексту. Для абсолютно випадкового тексту з алфавітом розміру N отримуємо $H_0 = \log_2 N$ (для англійської мови $H_0 = 4.7$). Провівши аналіз текстів художньої літератури отримуємо $H_1 = 4.2$.

Зрозуміло, що при наближенні тексту до випадкового (частотою символів) значення H_1 буде прямувати до H_0. Враховуючи це, можна стверджувати наступне - при абсолютному відхиленні від H_0 менше, ніж на 0.2 текст не є змістовним, оскільки не має місця надлишковість мови.

Біграмна ентропія [13] – аналог ентропії тексту, але на відміну від останньої оперує частотами біграм:

$$H_2 = -\sum_{i=1}^{n}\sum_{j=1}^{n} p(a_i a_j) * \log_2 p(a_i a_j)$$

де $p(a_i a_j)$ – імовірність появи біграми $a_i a_j$ у тексті.

Також існує величина $H^{(2)}$ – вона враховує частоту появи біграми і її відношення до очікуваної частоти (імовірності появи другої літери незалежно від першої):

$$H^{(2)} = -\sum_{i=1}^{n}\sum_{j=1}^{n} p(a_i a_j) * \log_2 \frac{p(a_i a_j)}{p(a_j)}$$

У розглянутих текстах величина $H^{(2)}$ = 3.63; H_2= 3.97. Аналогічно H_1 сформулюємо критерій $H^{(2)}$ – при відхиленні від H_1 менше ніж на 0.4, текст вважатимемо незмістовним.

Оцінка ефективності критеріїв змістовності залежить від трьох параметрів: довжини шифротексту для аналізу та імовірностей помилкок першого та другого роду, тобто імовірності, з якою критерій визначить змістовний текст як незмістовний та навпаки. Для деяких простих критеріїв змістовності існують аналітичні оцінки імовірностей помилок [15], але для більш складних критеріїв їх теоретичний аналіз зіткається з великими складнощами. До того ж, помилки другого роду залежать від структури множини помилкових текстів, яка визначається алгоритмом шифрування. Тому ефективність запропонованих критеріїв необхідно визначати експериментально.

2.3. Експериментальна перевірка запропонованих критеріїв

Згрупуємо критерії у 5 груп.

1) Усі критерії стосовно частих символів та надмірності – наявність п’яти найбільш частих символів абетки, надмірна частота цих символів, надмірна частота голосних, частота найбільш уживаних літер, наявність серед них хоча б однієї голосної.

2) Усі критерії стосовно рідкісних символів або недостатньої частоти – недостатня частота голосних, надмірна частота рідких символів, недостастня частота найбільш уживаних символів.

3) Наявність заборонених біграм.

4) Значення H_1.

5) Значення H_2 та $H^{(2)}$.

Перевіримо критерії на практиці – застосуємо їх на текстах художньої літератури, що раніше були використані для знаходження значень H_1, H_2 та $H^{(2)}$.

Опишемо експеримент: випадково оберемо текст загальною довжиною $n \times 3000$ символів, де n – кількість символів в одному повідомленні та перевіримо отримане повідомлення на змістовність. Після цього зашифруємо його за допомогою одного з розглянутих шифрів (для кожного повідомлення буде використано 25 різних ключів) та повторно перевіримо критерії. Оберемо довжину повідомлень у межах від 100 до 1000 символів, що найбільш частіше зустрічається на практиці. Таким чином, для кожної довжини повідомлення для кожного критерія буде проведено 75000 дослідів.

У таблиці 2.3 наведено результати перевірки критеріїв на відкритих текстах.

Таблиця 2.3

Результати перевірки відкритих текстів

Кількість символів у одному повідомленні	Відсоток вірних відповідей				
	1-го методу	2-го методу	3-го методу	4-го методу	5-го методу
100	100%	100%	99,7%	100%	100%
200	100%	100%	99,5%	100%	100%
300	100%	100%	99,4%	100%	100%
400	100%	100%	99,2%	100%	100%
500	100%	100%	99,1%	100%	100%

Експеримент було перервано після довжини у 500 символів, оскільки чітко прослідковувалася поведінка критеріїв. Отримані критерії дають позитивні відповіді приймаючи тексти художньої літератури, за виключенням критерія заборонених

біграм. Такий результат отримано в результаті особлистей програми переліку. Оскільки будь-якого виду розділові знаки у переважній більшості не шифруються у шифрах підстановки, то і шифротексти їх не містять. Щоб створити умови не відмінні від умов шифротекстів, усі розділові знаки було усунуто з розглянутих текстів, що призвело до невеликої кількості заборонених біграм (вони утворилися на границі двох слів, але якщо раніше між ними стояв розділовий знак, то зараз його там не має).

Отримані результати свідчать про те, що сформульовані критерії вдало виконують свою роботу та можуть розпізнати змістовний текст.

Тепер проведемо експеримент з визначення точності отриманих критеріїв шляхом аналізу зашифрованих текстів. Розглядатимемо довжини починаючи з сотні та роблячи поступове збільшення на сотню, допоки не дійдемо до тисячі символів.

З отриманих даних зробимо висновок про ефективність критеріїв.

2.3.1. Шифр Цезаря

Шифр Цезаря зсуває літери відкритого тексту на певну відстань по алфавіту. Таким чином множина частот біграм та літер зберігається, змінюється лише носії цих частот. З цього робимо припущення, що критерії H_1, H_2 та $H^{(2)}$ будуть давати позитивні відповіді, а критерії, розраховані на частоти конкретних літер будуть давати негативні відповіді.

Результати експерименту представленні у таблиці 2.4.

Як і очікувалося, ентропія залишається сталою і не вказує на незмістовність тексту.

Перші три методи продемонстрували достаньо високу точність. На малих розмірах повідомлень заборонені біграми зустрічалися не завжди, у зв’язку з малою кількістю біграм. Натомість, метод рідкісних символів показав найбільшу точність на малих довжинах, що пояснюється алгоритмом шифра Цезаря – зміщення алфавіту.

Таблиця 2.4

Результати експерименту на шифрі Цезаря

Кількість символів у одному повідомленні	Відсоток неправильних відповідей				
	1-го методу	2-го методу	3-го методу	4-го методу	5-го методу
100	4,04%	0,57%	2,42%	100%	100%
200	2,58%	0,25%	0,16%	100%	100%
300	2,13%	0,2%	0,008%	100%	100%
400	1,91%	0,17%	0%	100%	100%
500	1,73%	0,1%	0%	100%	100%
600	1,5%	0,07%	0%	100%	100%
700	1,28%	0,04%	0%	100%	100%
800	1,27%	0%	0%	100%	100%
900	1,15%	0,02%	0%	100%	100%
1000	1,1%	0%	0%	100%	100%

Таким чином рідкісні літери отримували невластиву їм частоту, що й призводило до правильних відповідей щодо змістовності тексту. Перший метод показав порівняно невелику точність, що пов’язано з його формулюванням – його відповідь спирається на надмірну частоту, що при шифрі Цезаря зустрічається досить рідко.

Шифр Віженера є застосуванням кількох різних шифрів Цезаря до різних літер, тому результати дослідження будуть співпадати з отриманими для звичайного шифру Цезаря.

2.3.2 Афінний шифр біграмної заміни

Ідея афінного біграмного шифру полягає в заміні одних біграм іншими. Таким чином, значення H_2 та $H^{(2)}$ залишаються сталими, оскільки множина частот біграм зберігається, змінюються лише носії цих частот. Але частоти самих літер змінюються без узагальнюючого правила, тому значення H_1 зміниться.

Результати експерименту наведено у таблиці 2.5.

Як і передбачалося, останній критерій не може відрізнити шифротексти від змістовних. H_1 продемонстрував значний ріст точності з ростом числа символів у

повідомленнях, проте не досягає рівня перших трьох методів, навіть у найгірших для них умовах. Вони в свою чергу знову продемонстрували величезну точність, вже на 400 символах даючи стовідсоткову негативну відповідь. Такий швидкий ріст зумовлений поступовим збільшенням псевдовипадковості тексту.

Таблиця 2.5

Результати експерименту на афінному шифрі біграмної заміни

Кількість символів у одному повідомленні	Відсоток неправильних відповідей				
	1-го методу	2-го методу	3-го методу	4-го методу	5-го методу
100	2,04%	0,22%	2,07%	96,6%	100%
200	0,44%	0,024%	0,14%	72,6%	100%
300	0,14%	0,0044%	0,017%	50,9%	100%
400	0,054%	0%	0%	37,54%	100%
500	0,022%	0%	0%	27,37%	100%
600	0,044%	0%	0%	21,45%	100%
700	0,031%	0%	0%	18,04%	100%
800	0%	0%	0%	15,43%	100%
900	0%	0%	0%	13,24%	100%
1000	0%	0%	0%	11,82%	100%

2.3.3. Шифр Хілла

Шифр Хілла є прототипом перший блочних шифрів, де на кожен наступний елемент впливає попередній. Хоча Хіллу було далеко до сучасних алгоритмів, все ж таки його метод мав достатню криптостійкість для свого часу.

Для випадку $n = 2$ шифр мало чим відрязняється від афінного біграмного щифру підстановки. В основі алгоритму лежать тіж самі ідеї, хоча результат розподілу біграм і є більш випадковим, сама множина імовірностей біграм залишається сталою. Отже, значення H_2 та $H^{(2)}$ знову нічим не допоможуть у визначенні змістовного тексту, а інші методи повині показати схожі результати.

Результати експерименту наведені в таблиці 2.6

Таблиця 2.6

Результати експерименту на шифрі Хілла

Кількість символів у одному повідомленні	Відсоток неправильних відповідей				
	1-го методу	2-го методу	3-го методу	4-го методу	5-го методу
100	0,85%	0,16%	1,72%	76,48%	100%
200	0,035%	0,017%	0,08%	10,34%	100%
300	0%	0%	0%	0,62%	100%
400	0%	0%	0%	0,10%	100%
500	0%	0%	0%	0%	100%
600	0%	0%	0%	0%	100%

Результати аналізу текстів, зашифрованих шифром Хілла, виявилися найбільш точними з усіх. Це пояснюється достатньо великою випадковістю окремих символів, про що свідчить стрімке падіння похибки H_1 до нуля. Проте слід зазначити, що частота самих біграм не змінилася, що видно з результатів критеріїв біграмної ентропії.

ВИСНОВКИ

У ході даної роботи було виконано ознайомлення з різними типами шифрів, що виникали впродовж історії людства, зокрема, шифрами підстановки та методами їх криптографічного аналізу. Визначено, що задачі автоматизованого зламу шифрів підстановки вимагають перевірки текстів після розшифрування на змістовність.

На основі статистичного аналізу англомовних текстів художньої літератури було сформульовано критерії, що вказуватимуть на змістовність тексту. Запропоновані критерії розбито в п’ять груп за основними властивостями, які використовувались при побудові: часті літери, малоімовірні літери, заборонені біграми, значення ентропії на символ тексту та на біграму тексту.

Для перевірки можливості використання запропонованих критеріїв у задачах криптоаналізу було проведено експериментальну оцінку імовірностей помилок критеріїв для шифру Цезаря, шифру афінної біграмної заміни та шифру Хілла. Встановлено, що для шифротекстів довжини у 300-700 символів дані критерії дозволяють майже безпомилково розрізнити змістовний текст від незмістовного. Для всіх трьох шифрів експерименти показали високу точність критеріїв, що спиралися на недостатню вживаність частих символів та надмірну вживаність рідкісних символів. Найменшу кількість помилок при фіксованій довжині шифротекстів дає пошук заборонених біграм; однак даний критерій робить більше помилок для змістовних текстів (що пояснюється відсутністю розділових знаків). Критерій на основі ентропії символі виявився непридатним для аналізу шифрів простої підстановки, але проявив себе у більш складних шифрах.

Одержані результати свідчать про високу точність запропонованих критеріїв та можливість їх застосування для перевірки змістовності текстів навіть невеликої довжини. Однак поза увагою залишились омофонічні шифри та шифри перестановок, а також сучасні алгоритми шифрування. Задачі, пов’язані з аналізом зазначених шифрів, формують напрямки подальших досліджень за даною тематикою.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Аль-Кінді Про дешифрування криптографічних повідомлень, близько 850 р.
2. Батіста Л. А. Трактат про шифри, 1466 р.
3. Shannon Claude E. Communication Theory of Secrecy Systems // Bell System Technical Journal. – 1949. – vol. 28(4), page 656–715.
4. Auguste Kerckhoffs La Cryptographie Militaire // Journal des sciences militaires – January 1883 vol. IX, pp. 5–83, February 1883, pp. 161–191.
5. J. Buchmann Introduction to Cryptography, New York: Springer, 2004 – p. 95-96.
6. Friedrich L.Bauer Decrypted Secrets: Methods and Maxims of Cryptology, New York: Springer, 1997 — page 61-63.
7. Kahn D. The Codebreakers; The Comprehensive History of Secret Communication from Ancient Times to the Internet, N- Y: Macmillan Publ. Co. 1996. page 1 – 89.
8. Franksen, Ole Immanuel Babbage and cryptography. Or, the mystery of Admiral Beaufort's cipher//Mathematics and Computers in Simulation.–1993.–vol.35 p.327-367.
9. Reinke E. C. Classical Cryptography//The Classical Journal.–1962,Vol.58p.113–121.
10. Alfred J. Menezes, Paul C. van Oorschot, Scott A. Vanstone Handbook of Applied Cryptography, Florida: CRC Press – 1996,. page 246.
11. Martin, Keith M. Everyday Cryptography // Oxford University Press. – 2012, p. 142.
12. Lester S. Hill Cryptography in Algebraic Alphabet // The American Mathematical Monthly, June 1929, page 306-312.
13. Shannon, Claude E. A Mathematical Theory of Communication // Bell System Technical Journal. –1948 – vol. 27 (3): page 419.
14. Вербіцький О.В. Вступ до криптології. — Львів: ВНТЛ, 1998. — 248 стор.
15. Бабаш А.В., Шанкин Г.П. Криптография. — М.: СОЛОН-ПРЕСС, 2007. — 512 с. — (Серия книг «Аспекты защиты»).
16. Konheim A. G. Computer security & cryptography. — Hoboken: John Wiley & Sons, Inc., 2007. — 521 pp.

ДОДАТОК А
ПЕРЕЛІК ЗМІСТОВНИХ ТЕКСТІВ АНГЛІЙСЬКОЮ МОВОЮ, ВИКОРИСТАНИХ ДЛЯ ЕКСПЕРИМЕНТАЛЬНОЇ ЧАСТИНИ

Список літератури, використанної для статистичного аналізу у роботі. Усі книжки перебувають у вільному доступі і були взяті з онлайн-бібліотеки BookRix.

1. William Shakespeare, Romeo and Juliet
2. Sir Arthur Conan Doyle, The hound of the Baskervilles
3. Sir Arthur Conan Doyle, The adventures of Sherlock Holmes
4. Sammantha Lewis, Haunted
5. Oscar Wilde, The picture of Dorian Grey
6. Mf Harris – Awakening Summer
7. Lorelei Sutton – A beatiful terrible love
8. Lucy Lucero – Renesmee
9. Lorelei Sutton – A howl in the night
10. Kelly Joan – Pack Gems
11. Kaylyn Kahle – Enemies with Benefits
12. Katy Wormald – When summer ends
13. Jess Wygle – Hush
14. J-a Booker – Slaves of the night
15. J-a Booker – Mark of the moon
16. Eric Johnson – Star wars infinite darkness
17. Emily Zimmermann – Instant enemies
18. Eftos Ent – Kingdom of a thousand
19. Danielle D. – Rosalina’s hope
20. David Burgess – The Samsara project
21. Cory Doctorow – Makers
22. Cory Doctorow – Little brother
23. C.B. Cooper – The Daughter
24. Bob Moats – Classmate murders
25. Alyza Slaton – Boarding schools sercrets and jerks
26. Ally Carter – Cross my heart and hope to spy