Особливості частотного аналізу шифротексту на основі української абетки

Кармазіна Ю.В.
студентка 3 курсу
ПНПУ імені В.Г. Короленка

Протягом століть взлому шифрів допомагає частотний аналіз появи літер та їх поєднань. Частотний аналіз є розповсюдженим методом криптоаналітичної атаки. Ідея цього методу добре відома любителям детективів за оповіданням А. Конан Дойля «Танцюючі чоловічки». Даний метод криптоаналізу використовує той факт, що ймовірності появи окремих літер, а також їх порядок в словах і фразах природної мови підкоряються задокументованим статистичним закономірностям. Аналізуючи досить довгий текст, зашифрований методом заміни, можна за частотами появи символів зробити зворотну заміну і відновити вихідний текст. Це використовується при взломі шифрів Цезаря, Віженера, Вермана та інших моноалфавітних.

Очевидно, частотний аналіз вимагає насамперед еталонних частот повторюваності літер абетки, на якій написані відкриті тексти, і частот повторюваності n-грам ($n \geq 2$). Для російської, англійської та майже всіх європейських мов середньостатистичні частоти повторюваності літер, біграм, триграм можна знайти в літературних джерелах та Інтернеті. На жаль, для української мови в літературі наведені лише частоти повторюваності літер. Сушко С.О., Фомичова Л.Я. та Барсуков Є.С. у своєму дослідити частоти повторюваності літер і біграм української мови на основі вибраних випадково текстів українською мовою.

Метою даної роботи є дослідження частоти повторюваності літер української мови відповідно до стилістики вхідного тексту. Було проаналізовано тексти наукового, художнього та ділового стилів, об’єм яких становив близько 6 мб. За основу було взято дослідження Сушко С.О., Фомичова Л. Я., Барсукова Є. С. [1].

Результати проведеного аналізу середньостатистичних частот літер в україномовному тексті наводиться в таблиці 1 та на рис. 1.

Таблиця 1

Літери укр. абетки	Стилі тексту			Сер. знач.
	Діловий	Художній	Науковий	
А	0,095	0,091	0,088	0,091
Б	0,014	0,019	0,016	0,016
В	0,062	0,066	0,059	0,062
Г	0,015	0,023	0,015	0,018
Ґ	0,000	0,000	0,000	0,000
Д	0,044	0,036	0,038	0,039
Е	0,047	0,053	0,051	0,050
Є	0,003	0,007	0,006	0,005
Ж	0,009	0,010	0,008	0,009
З	0,031	0,027	0,026	0,028
И	0,055	0,067	0,070	0,064
І	0,045	0,050	0,054	0,050
Ї	0,009	0,011	0,010	0,010
Й	0,008	0,014	0,011	0,011
К	0,038	0,035	0,040	0,038
Л	0,026	0,043	0,034	0,034
М	0,031	0,036	0,040	0,036
Н	0,095	0,066	0,078	0,080
О	0,107	0,102	0,0101	0,073
П	0,036	0,031	0,034	0,034
Р	0,055	0,057	0,052	0,055
С	0,043	0,042	0,041	0,042
Т	0,057	0,048	0,059	0,055
У	0,035	0,038	0,036	0,036
Ф	0,004	0,003	0,005	0,004
Х	0,011	0,010	0,010	0,010
Ц	0,010	0,006	0,011	0,009
Ч	0,011	0,017	0,013	0,014
Ш	0,004	0,009	0,007	0,007
Щ	0,003	0,008	0,005	0,005
Ь	0,014	0,013	0,017	0,015
Ю	0,008	0,008	0,008	0,008
Я	0,030	0,025	0,023	0,026

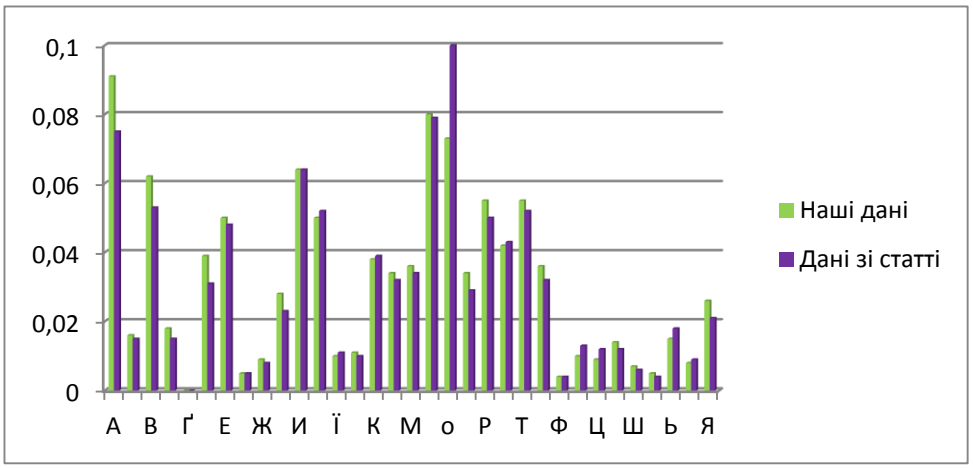

Рис. 1. Діаграма порівняння отриманих результатів із результатами в [1].

Проведений аналіз підтвердив дані, отримані вищезгаданими дослідниками. Враховано, що для української мови, як і для решти європейських мов, притаманне чергування голосних та приголосних. Якщо дослідити інші тексти, може бути присутня певна різниця в цифрах наведених частот літер, що пояснюється, по-перше, довжиною досліджуваного тексту, а по друге його тематикою. Наприклад, загалом мало вживана літера Ф може стати досить частою в технічних текстах, бо використовується в таких словах, як функція, диференціал, дифузія, коефіцієнт і т. п. Ще більші відхилення від традиційного вживання окремих літер спостерігаються в деяких художніх творах, особливо у віршах.

При практичному застосуванні частотного аналізу в криптоаналітичній атаці варто дотримуватися таких рекомендацій:

1. Спочатку необхідно визначитися із тим якою мовою написаний вхідний відкритий текст.
2. Дослідити слова, які містять подвоєння літер, оскільки не так багато подвоєнь літер властиве українській мові.
3. Якщо в шифротексті є пропуски між словами, то потрібно визначити слова, які складаються з однієї, двох або трьох літер.
4. Для полегшення аналізу підготувати таблицю частотності літер для повідомлення, яке піддається дешифруванню.

Результати проведеного аналізу можуть бути використані не лише при застосуванні частотного аналізу до україномовного шифротексту, а і при вивченні «Основ криптології» студентами 5 курсу спеціальності «Інформатика».

Список використаних джерел

1. Сушко С.О. Частоти повторюваності букв і біграм у відкритих текстах українською мовою [Електронний ресурс]. – Режим доступу: http://jrnl.nau.edu.ua/index.php/ZI/article/view/1968

2. Перебийніс В.І., Муравицька М.П., Дарчук Н.П. Частотні словники та їх використання. К.: Наукова думка, 1983.