Міністерство освіти й науки України
Львівський національний університет імені Івана Франка
Факультет електроніки та комп'ютерних технологій
з предмета: Комп'ютерна лінгвістика

Звіт
про виконання лабораторної роботи № 4
«Закон Гіпса для слів в окремих текстах»

Виконав:
Студент групи
ФеС-32с
Бойко Кирило

Львів 2024

Завдання

Використовуючи програму +proj6stats&plots, на лінгвістичному рівневі слів дослідити виконання закону Гіпса для обраних Вами текстів у лабораторній роботі №1. Провести дослідження для однограм (n = 1) за різних опцій (умов).

Хід виконання роботи

Підготував тексти з першої лабораторної роботи

Clemencia Novela de costumbres by Fernán Caballero

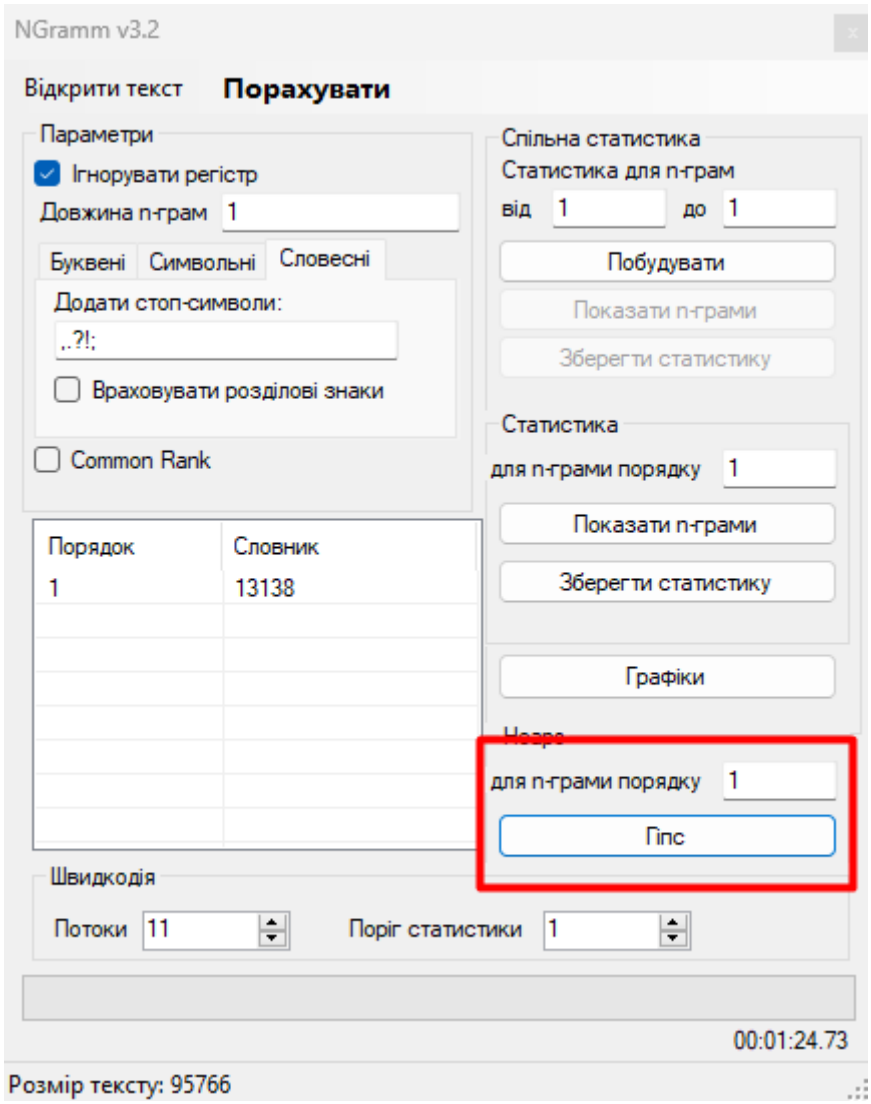
DONALD J. TRUMP January 20, 2017

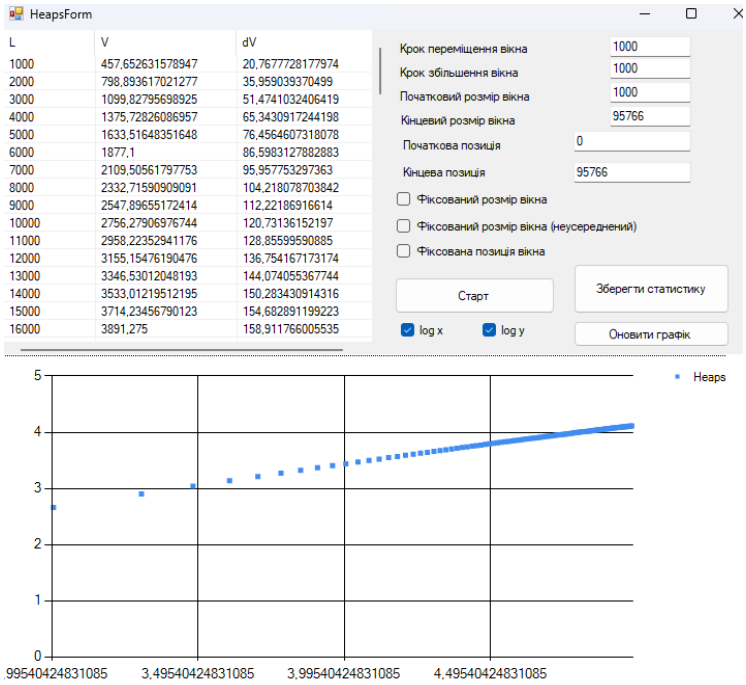
Написав код програми Python для визначення лінійної апроксимації зі збережених результатів

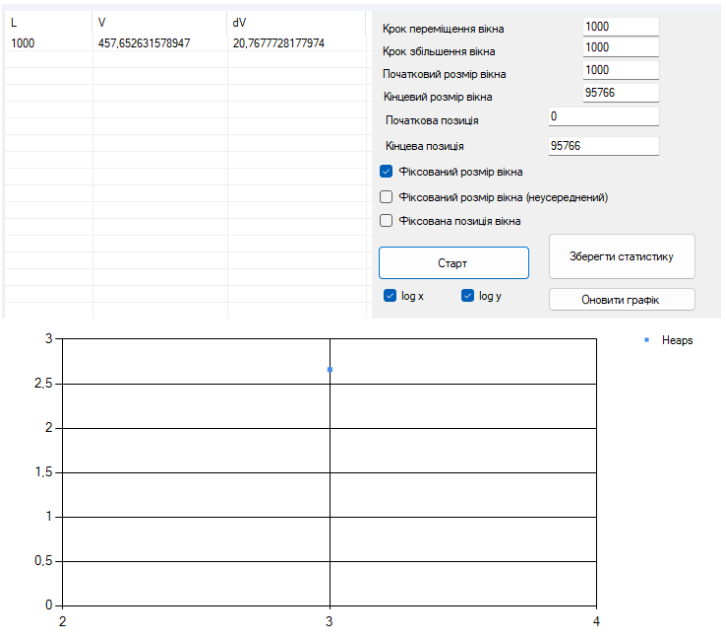
```python
import numpy as np
import matplotlib.pyplot as plt

file_path = "1.1.txt"
fixed_file_path = "1.1fix.txt"

with open(file_path, 'r', encoding='utf-8') as input_file, open(fixed_file_path, 'w', encoding='utf-8') as output_file:
    for line in input_file:
        output_file.write(line.replace(',', '.'))

data = np.loadtxt(fixed_file_path, delimiter='\t', skiprows=1)
length = data[:, 0]  # Довжина
volume = data[:, 1]  # Об'єм
delta_volume = data[:, 2]  # Різниця об'ємів

# Логарифмічні перетворення
log_length = np.log(length)
log_volume = np.log(volume)

# Лінійна апроксимація залежності V(L) ∝ L^p
coefficients = np.polyfit(log_length, log_volume, 1)
slope = coefficients[0]
intercept = coefficients[1]

plt.figure(figsize=(10, 6))
plt.scatter(log_length, log_volume, color='green', label='Дані (логарифмічні координати)')
plt.plot(log_length, np.polyval(coefficients, log_length), color='orange', label=f"Апроксимація: p ≈ {slope:.2f}")
plt.title("Лінійна апроксимація для закону Гіпса")
plt.xlabel("log(L)")
plt.ylabel("log(V)")
plt.legend()
plt.grid(True)
plt.show()

print(f"Коефіцієнт нахилу (p): {slope:.2f}")
print(f"Перетин з віссю (intercept): {intercept:.2f}")
```

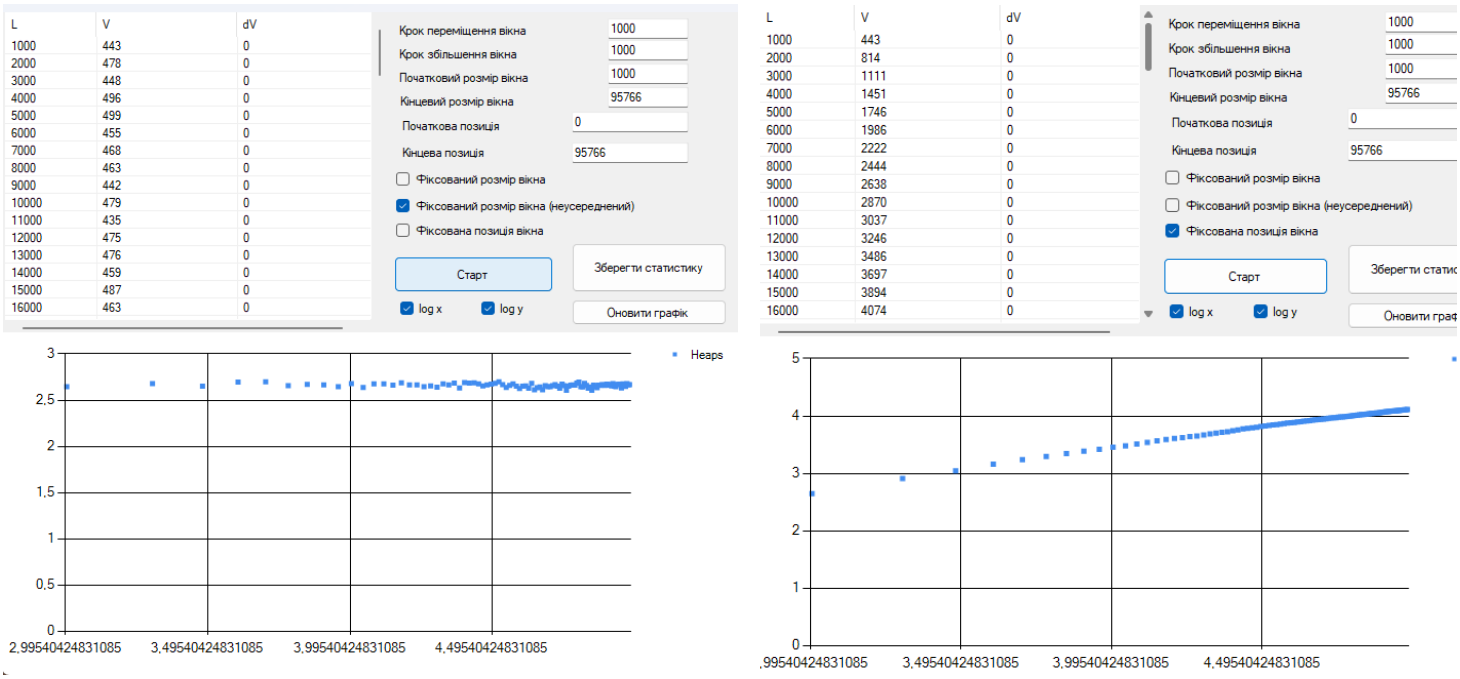
Запустив програму **+proj6stats&plots** і завантажив текстові файли для проведення дослідження закону Гіпса

Результати

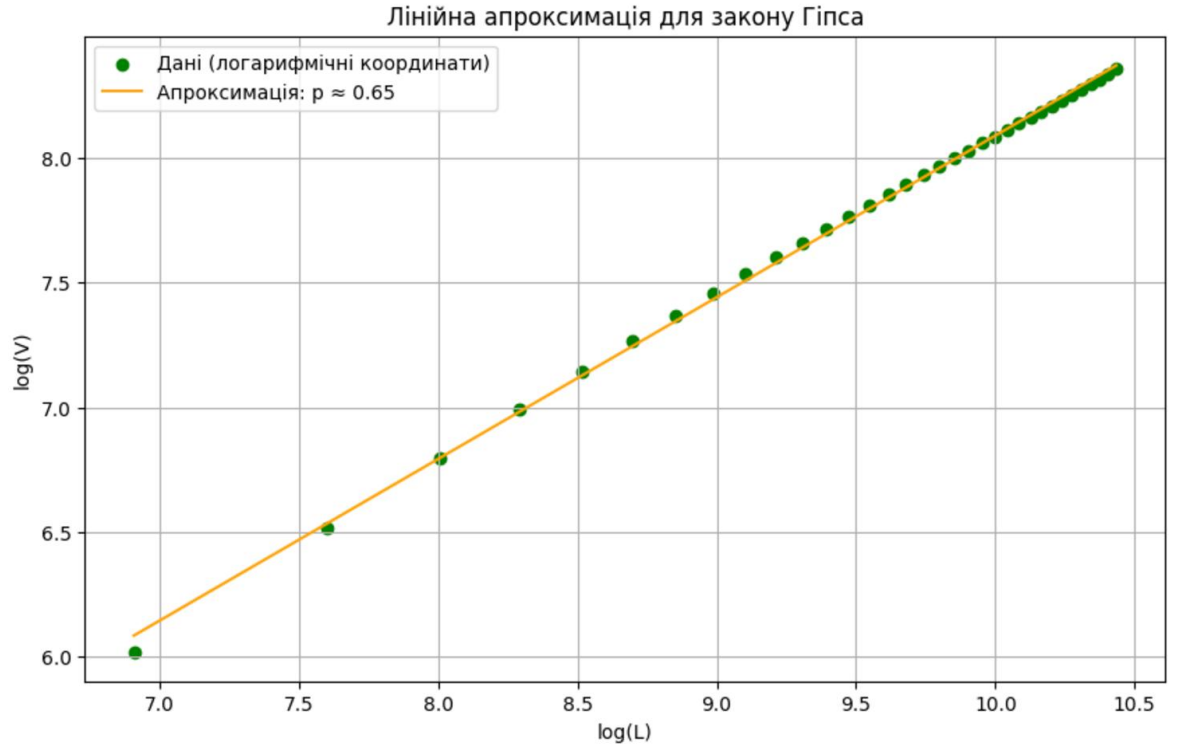

NGramm v3.2
Відкрити текст
Порахувати
Параметри
Ігнорувати регістр
Довжина n-грам 1
Буквені
Символьні
Словесні
Додати стоп-символи:
.,?!;
Враховувати розділові знаки
Common Rank
Спільна статистика
Статистика для n-грам
від 1
до 1
Побудувати
Показати n-грами
Зберегти статистику
Статистика
для n-грами порядку 1
Показати n-грами
Зберегти статистику
Графіки
Heaps
для n-грами порядку 1
Гіпс
Порядок
Словник
1
13138
Швидкодія
Потоки 11
Поріг статистики 1
00:01:24.73
Розмір тексту: 95766


HeapsForm
L
V
dV
1000 457,652631578947 20,7677728177974
2000 798,893617021277 35,959039370499
3000 1099,82795698925 51,4741032406419
4000 1375,72826086957 65,3430917244198
5000 1633,51648351648 76,4564607318078
6000 1877,1 86,5983127882883
7000 2109,50561797753 95,957753297363
8000 2332,71590909091 104,218078703842
9000 2547,89655172414 112,22186916614
10000 2756,27906976744 120,73136152197
11000 2958,22352941176 128,85599590885
12000 3155,15476190476 136,754167173174
13000 3346,53012048193 144,074055367744
14000 3533,01219512195 150,283430914316
15000 3714,23456790123 154,682891199223
16000 3891,275 158,911766005535
Крок переміщення вікна 1000
Крок збільшення вікна 1000
Початковий розмір вікна 1000
Кінцевий розмір вікна 95766
Початкова позиція 0
Кінцева позиція 95766
Фіксований розмір вікна
Фіксований розмір вікна (неусереднений)
Фіксована позиція вікна
Старт
Зберегти статистику
log x
log y
Оновити графік
Heaps
5
4
3
2
1
0
3,99540424831085
3,49540424831085
3,99540424831085
4,49540424831085


L
V
dV
1000 457,652631578947 20,7677728177974
Крок переміщення вікна 1000
Крок збільшення вікна 1000
Початковий розмір вікна 1000
Кінцевий розмір вікна 95766
Початкова позиція 0
Кінцева позиція 95766
Фіксований розмір вікна
Фіксований розмір вікна (неусереднений)
Фіксована позиція вікна
Старт
Зберегти статистику
log x
log y
Оновити графік
Heaps
3
2,5
2
1,5
1
0,5
0
2
3
4

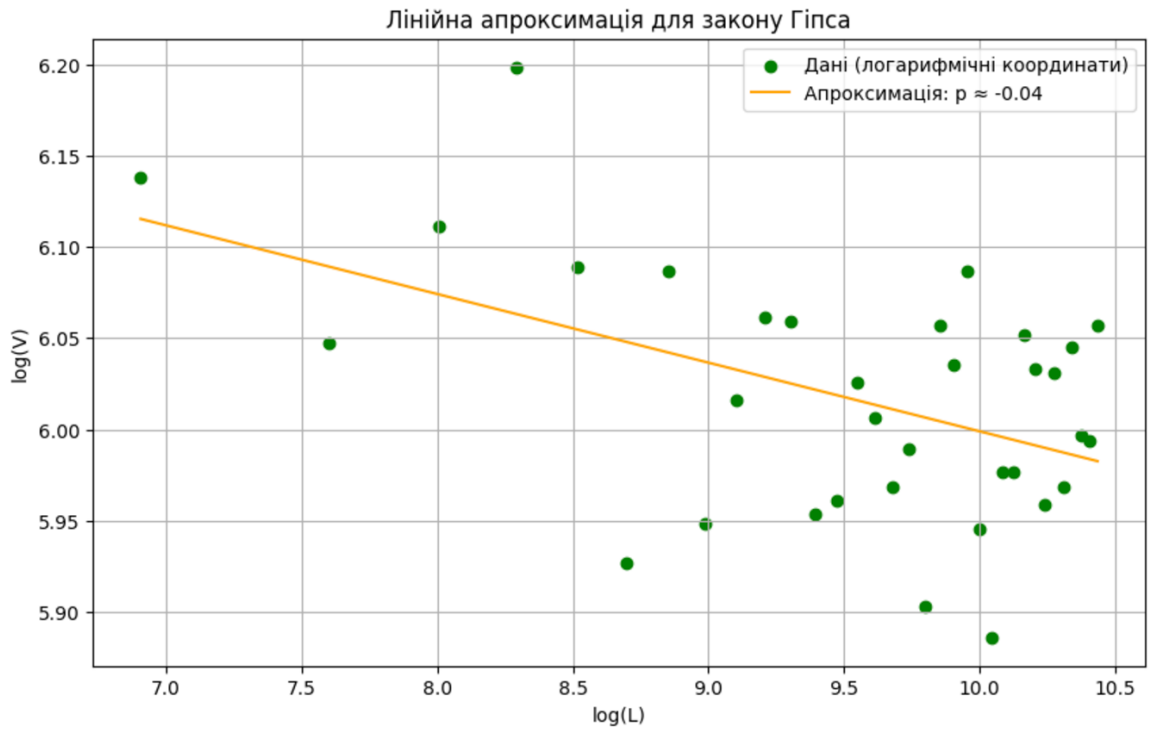


Результати виконання закону Гіпса для файлу ***Clemencia Novela de costumbres by Fernán Caballero***



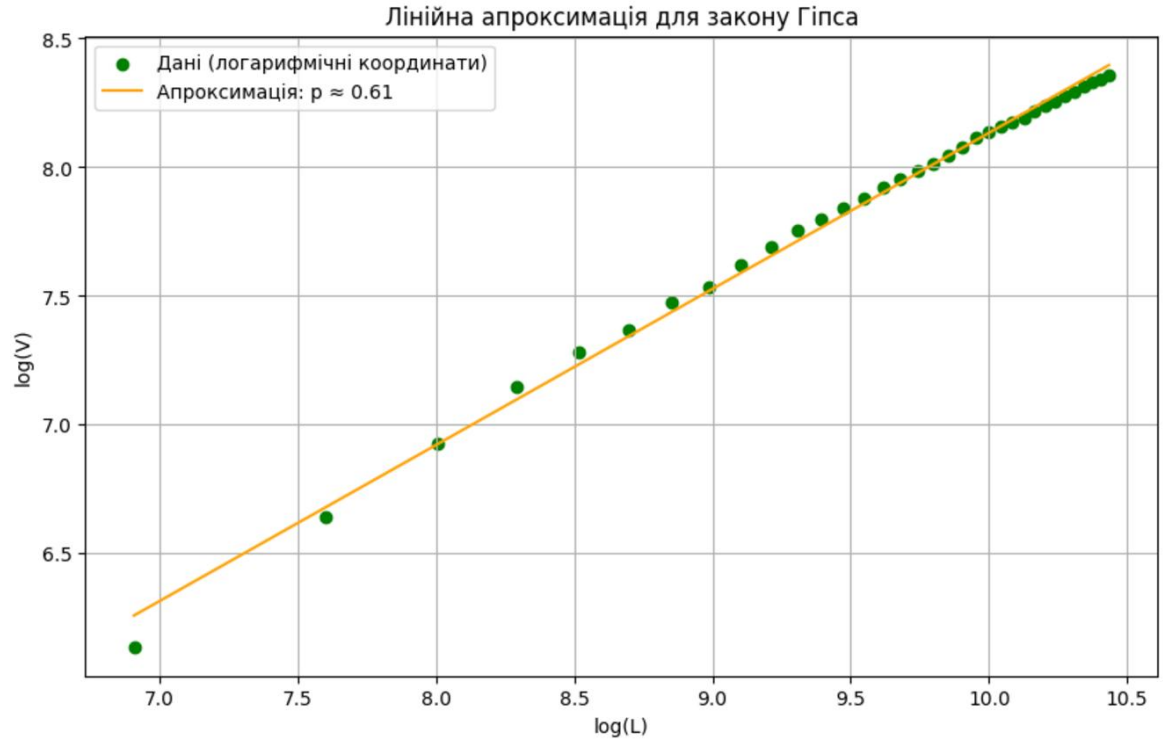

```
Коефіцієнт нахилу (p): 0.65
Перетин з віссю (intercept): 1.61
```

Лінійна апроксимація до результатів першого файлу першого тексту



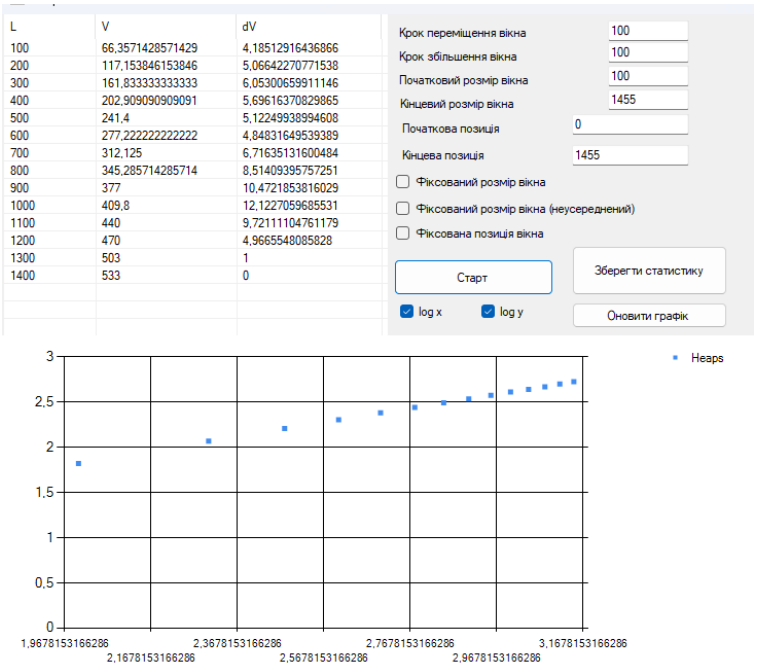

```
Коефіцієнт нахилу (p): -0.04
Перетин з віссю (intercept): 6.38
```

Лінійна апроксимація до результатів другого файлу першого тексту

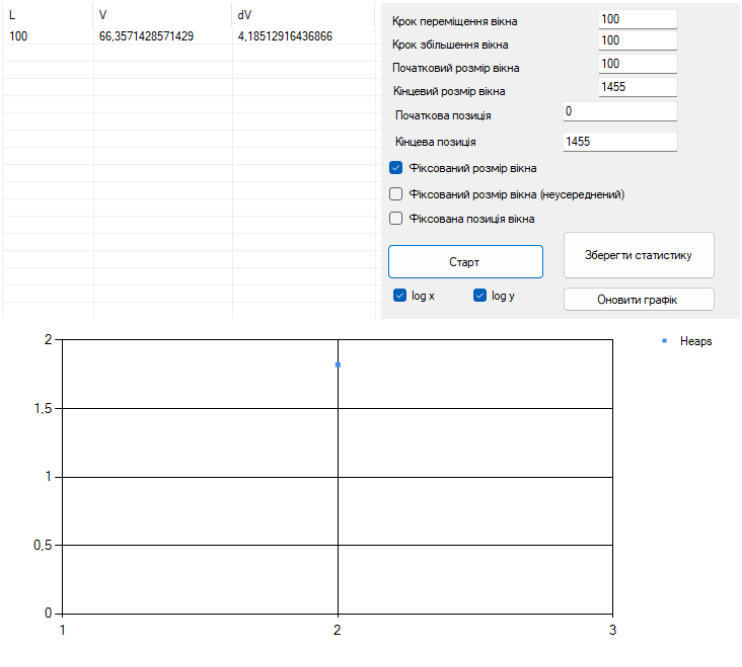



```
Коефіцієнт нахилу (p): 0.61
Перетин з віссю (intercept): 2.07
```

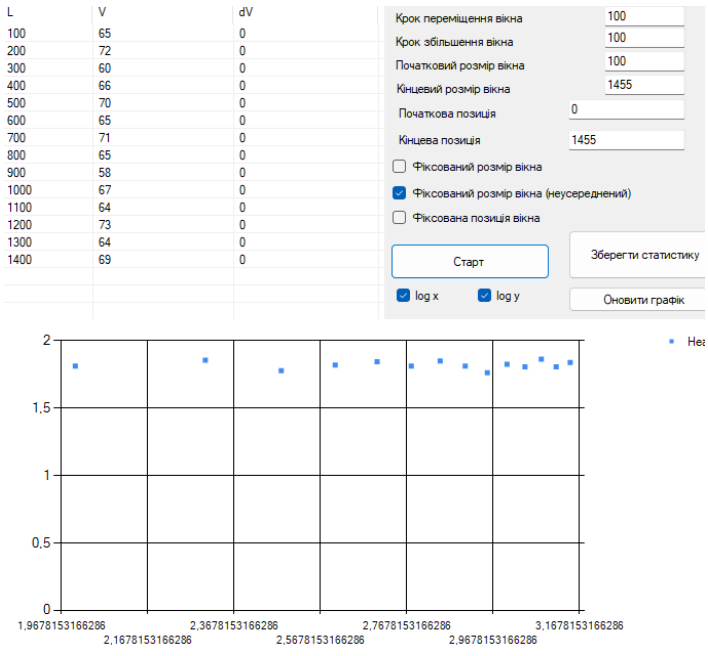
Лінійна апроксимація до результатів третього файлу першого тексту



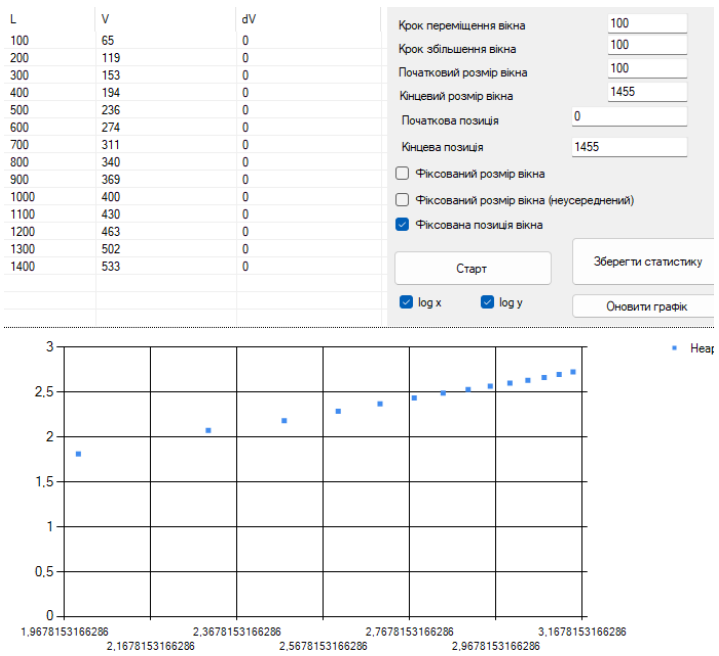
L	V	dV
100	66,3571428571429	4,18512916436866
200	117,153846153846	5,06642270771538
300	161,833333333333	6,05300659911146
400	202,909090909091	5,69616370829865
500	241,4	5,12249938994608
600	277,222222222222	4,84831649539389
700	312,125	6,71635131600484
800	345,285714285714	8,51409395757251
900	377	10,4721853816029
1000	409,8	12,1227059685531
1100	440	9,72111104761179
1200	470	4,9665548085828
1300	503	1
1400	533	0



L	V	dV
100	66,3571428571429	4,18512916436866

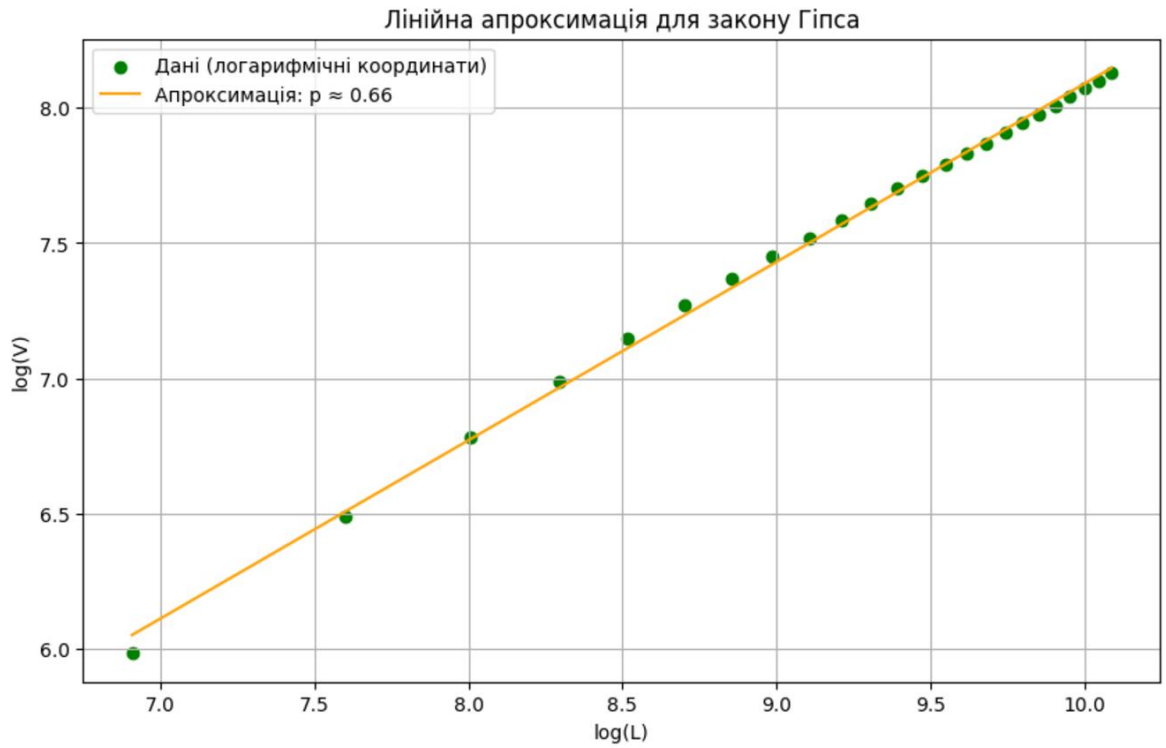


L	V	dV
100	65	0
200	72	0
300	60	0
400	66	0
500	70	0
600	65	0
700	71	0
800	65	0
900	58	0
1000	67	0
1100	64	0
1200	73	0
1300	64	0
1400	69	0



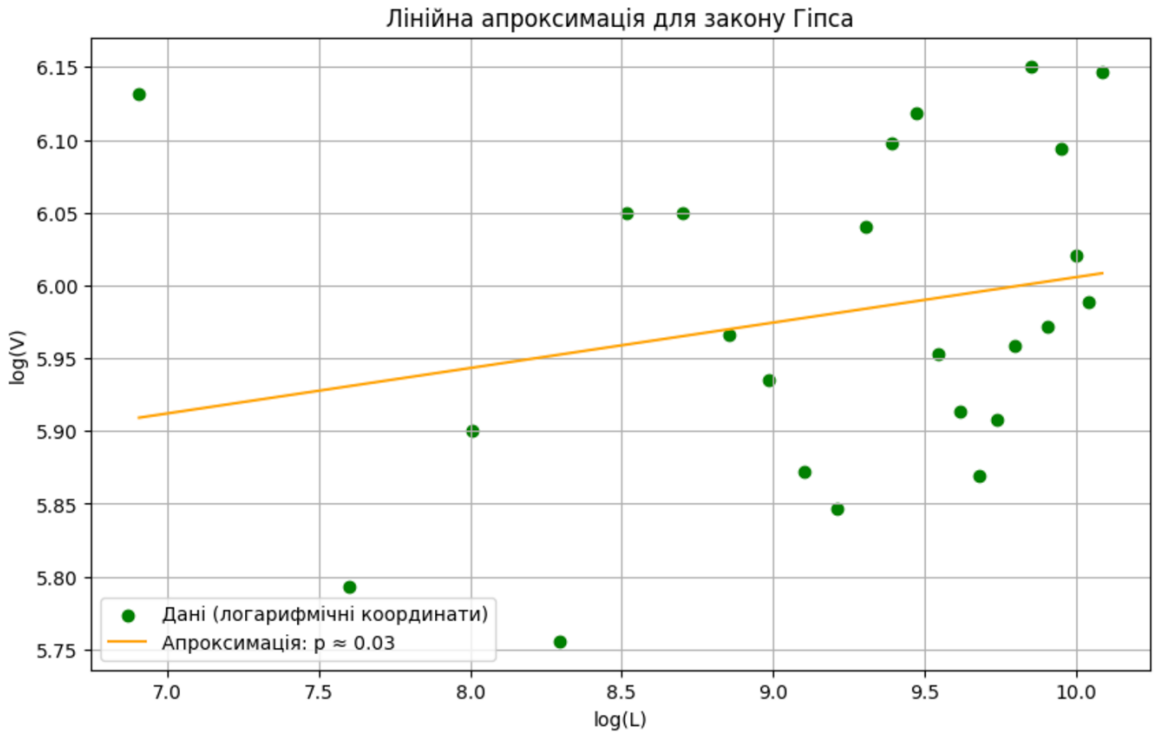
L	V	dV
100	65	0
200	119	0
300	153	0
400	194	0
500	236	0
600	274	0
700	311	0
800	340	0
900	369	0
1000	400	0
1100	430	0
1200	463	0
1300	502	0
1400	533	0

Результати виконання закону Гіпса для ***DONALD J. TRUMP January 20, 2017***



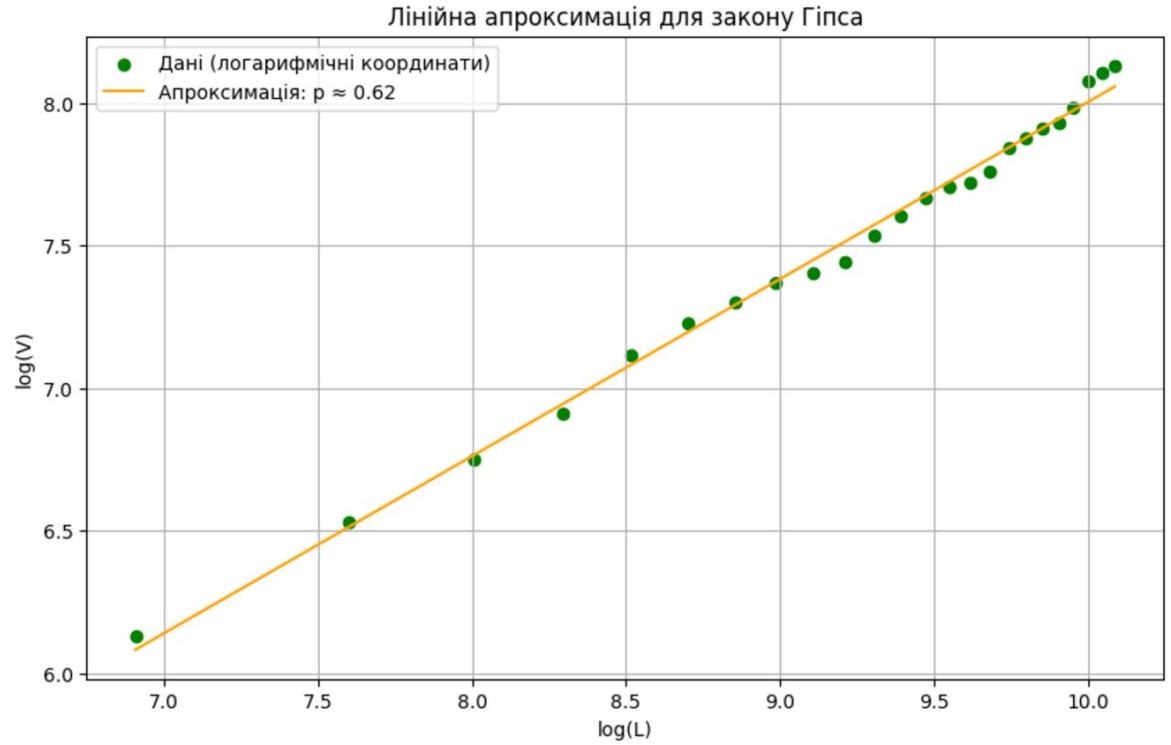

```
Коефіцієнт нахилу (p): 0.66
Перетин з віссю (intercept): 1.50
```

Лінійна апроксимація до результатів першого файлу другого тексту




```
Коефіцієнт нахилу (p): 0.03
Перетин з віссю (intercept): 5.69
```

Лінійна апроксимація до результатів другого файлу другого тексту




```
Коефіцієнт нахилу (p): 0.62
Перетин з віссю (intercept): 1.79
```

Лінійна апроксимація до результатів третього файлу другого тексту

Кожен з текстів має характерну експоненційну форму кривої, що підтверджує відповідність розподілу частот закону Гіпса. Це означає, що найбільш вживані слова мають значно вищу частоту, ніж менш вживані, що свідчить про ієрархічну структуру лексики.

Малі відхилення від ідеального експоненційного розподілу можуть бути зумовлені стилістичними особливостями текстів та авторським слововживанням. Однак загальна тенденція розподілу частот у обох текстах підтверджує відповідність закону Гіпса, що свідчить про те, що лексика кожного з текстів підпорядковується цьому закону.

Висновок: У цій лабораторній роботі я дослідив виконання закону Гіпса для текстів на рівні однограм. Результати показали, що частоти слів у текстах загалом слідують експоненційному розподілу, що підтверджує виконання закону Гіпса. Хоча на графіках можна помітити деякі відхилення, вони, ймовірно, зумовлені стилістичними особливостями авторів та структурою текстів. Отже, проведене дослідження підтверджує, що частоти слів у літературних текстах відповідають закономірностям, описаним законом Гіпса