Магістерська кваліфікаційна дипломна робота з теми «Комп'ютерна підсистема для статистичного аналізу прочитаних текстів»

Виконав: Вікулов Руслан Миколайович

Керівник: Снігур Анатолій Васильович

Мета та наукова новизна роботи

Метою магістерської кваліфікаційній роботі є є розширення функціональних можливостей засобів комп'ютерного аналізу статистичних параметрів текстових матеріалів.

Об'єктом дослідження є є процес взаємодії користувача з програмним забезпеченням для читання текстових матеріалів із врахуванням перерв у такому читанні.

Предметом дослідження є є математичні моделі та програмне забезпечення для статистичного аналізу прочитаних текстових матеріалів.

Наукова новизна полягає у

1. Дістали подальшого розвитку математичні моделі як усього процесу читання так і читання окремих фрагментів тексту
2. Дістали подальшого розвитку підходи визначення меж загального часу усього процесу читання, а отже і відповідної часової активності людини для різних способів читання.

Практичне значення одержаних результатів.

1. Розроблено програмне забезпечення для статистичного аналізу текстів та прочитаного матеріалу.
2. Розроблено програмне забезпечення для визначення поля зору людини.
3. Розроблено програмне забезпечення для аналізу процесу читання.

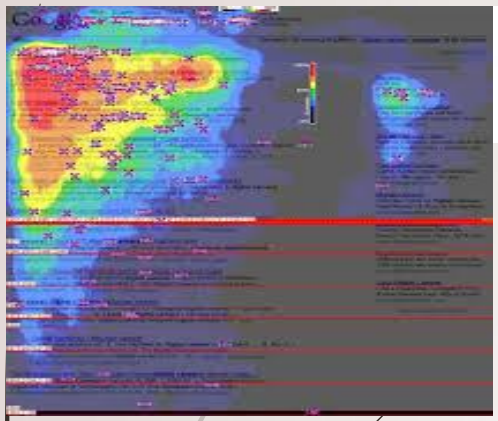
Актуальність теми:

- Апарат математичної статистики є одним з основних методів дослідження в сучасній науці. Статистичні методи дослідження є основою визначних досягнень в таких науках як біологія, медицина, економіка, соціальні науки. Принциповою особливістю статистичних методів дослідження є їх висока трудомісткість, великий обсяг розрахунків при відносно простому аналітичному апараті. В наш час існує безліч програм для статистичного аналізу тексту, які відрізняються між собою функціями, інтерфейсами, та багатьма іншими параметрами. Кожна з таких програм має ряд своїх переваг та недоліків і при цьому загалом не забезпечує аналізу фрагментів тексту, зокрема отриманих на основі розбиття тексту лекційного матеріалу за методами Б.Ф. Скінера, А. Краудера. Актуальним є завдання комп'ютерного статистичного аналізу текстів для їх застосування під час навчання, наприклад студентів та прогнозування. Оскільки це дає можливість виявити яким чином навчаються студенти, яку кількість матеріалу вони засвоюють та відповідно до цього формувати навчальний матеріал.

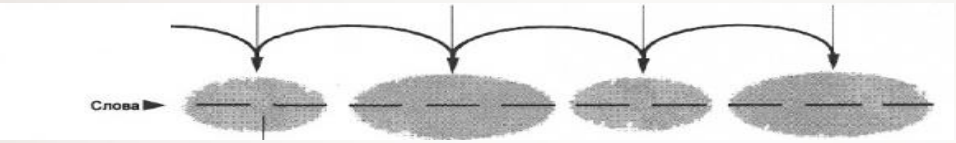
Програмне забезпечення для статистичного аналізу тексту

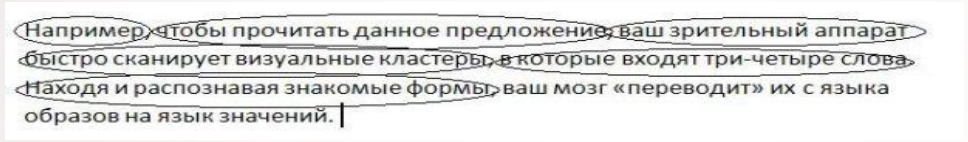
Назва програмного продукту	Переваги	Недоліки
WordStat	Простий інтерфейс, зручна в використанні	Недостатньо гнучка, з малою кількістю параметрів, відсутня можливість створення діаграм
Concordance	Реалізує пошук веб відповідностей, конвертуючи після цього результат аналізу в HTML файли	Помилка аналізу - неправильний поділ на склади
TextAnalyst	Великий набір параметрів для аналізу	Складний інтерфейс, вимагає складних налаштувань, платна (велика ціна)
Текст по правилам	Простий інтерфейс, зручна в використанні	Не працюють із поширеними текстовими форматами документів

Підходи аналізу прочитаного матеріалу

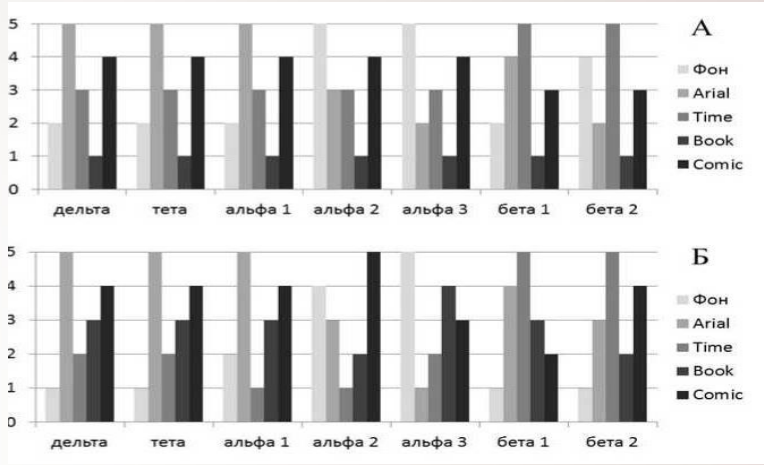


Приклад використання ай-трекінгу на сторінці сайту

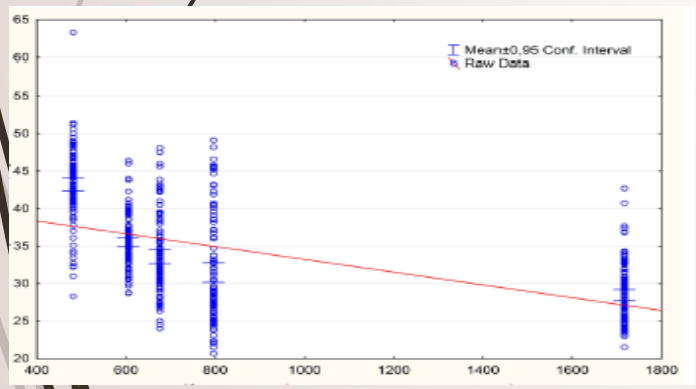




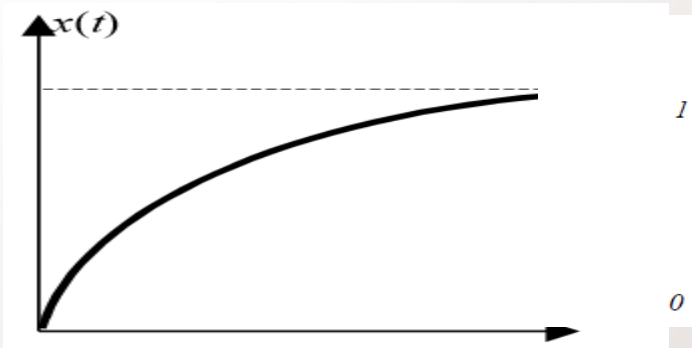
Аналіз параметрів стрибків ока та виділення візуальних кластерів при читанні текстів

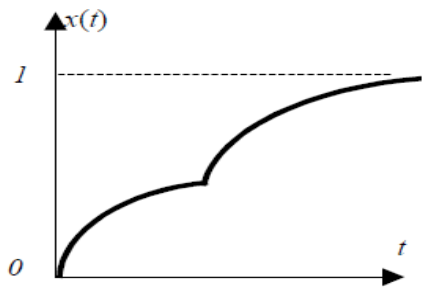


Значення відносних параметрів спектрів потужності електроенцефалограми при читанні різних гарнітур: А - у задніх областях кори головного мозку, Б - у правій півкулі



Залежності швидкості читання текстів від ізрізаності шрифту



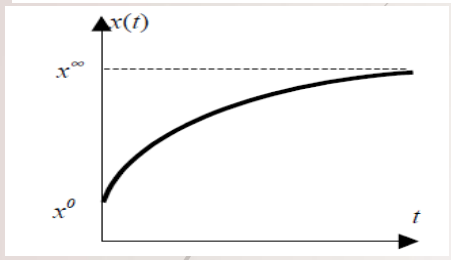


Криві навчання

Залежності, що характеризують читання:

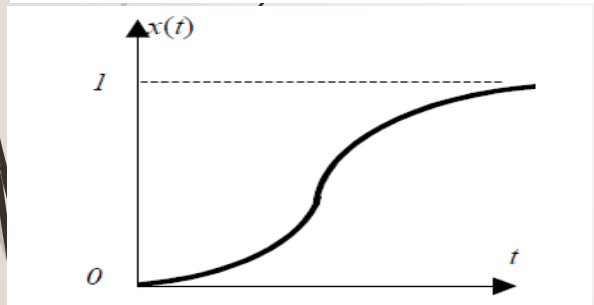
Експоненційна

$$x(t) = x^{\infty} + (x^{0} - x^{\infty})\, e^{-\gamma t},\ t > 0,$$



Логістична

$$x(t) = x^{0}\, x^{\infty} / (x^{0} + (x^{\infty} - x^{0})\, e^{-\gamma t})$$



Факториний аналіз

$$y_j = a_{1j} F_1 + a_{2j} F_2 + \ldots + a_{jp} F_p + d_j U_j,$$

Формула визначення легкості читання тексту, розроблена Р. Флешем

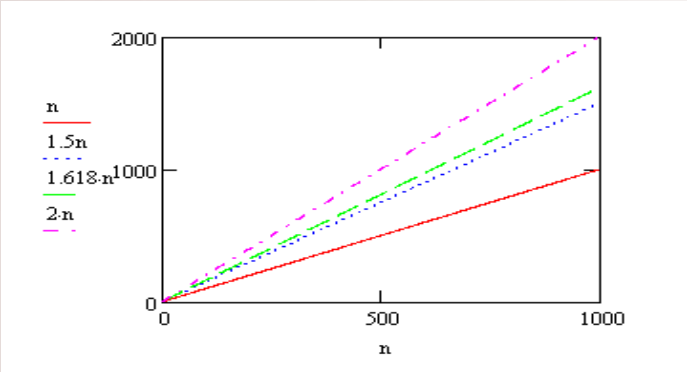
$$F_{Flesch} = 206{,}835 - (\frac{k}{s} * 1{,}015 + \frac{f}{k} * 84{,}6) =$$
$$= 206{,}835 - (w * 1{,}015 + p * 84{,}6).$$

Не дивлячись на велику кількість досліджень у даному напрямку, вони не дають можливості оцінювати різні способи читання, засвоєння матеріалу, не визначають їхні границі “зверху” та ”знизу”, не узагальнюють їх не класифікують їх, а отже не дають можливості формувати новий начальний матеріал відповідно до потреб тих, хто навчається.

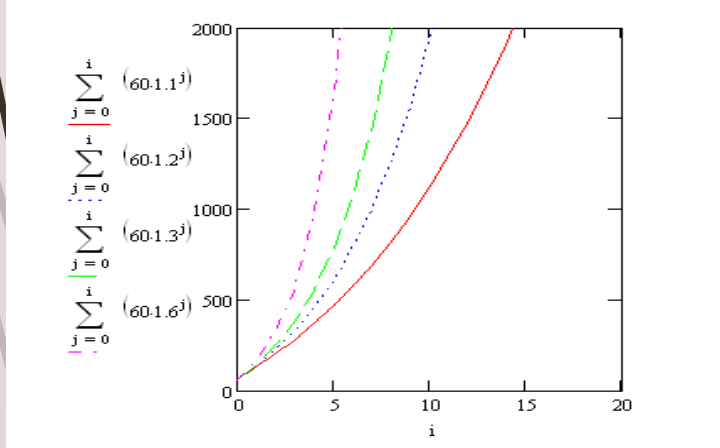
Побудова часових моделей оцінювання процесу читання

Тип ММ	Часова модель	Значення T_{n-1}	Значення T_n
Однаковий час читання	$T_{fc}(n) = nT_0$	$T_{fc}(n-2)$	$T_{fc}(n-1)$
Наступний час однаково більше попереднього	$T_{fc}(n) = \sum_{i=0}^{n-1} T_0 \cdot K^i$	$T_{fc}(0) \times \times K^{n-2}$	$T_{fc}(0) \times \times K^{n-1}$
Наступний час однаково менший попереднього	$T_{fc}(n) = \sum_{i=0}^{n-1} T_0 \cdot K^i$	$T_{fc}(0) \times \times K^{n-2}$	$T_{fc}(0) \times \times K^{n-1}$
Нерівномірний час читання	$T_{fc}(n) = \sum_{j=0}^{n} (T_0 \cdot \prod_{i=0}^{n-1} K_i$	$T_{fc}(n-2) \times \times K_{n-2}$	$T_{fc}(n-1) \times \times K_{n-1}$

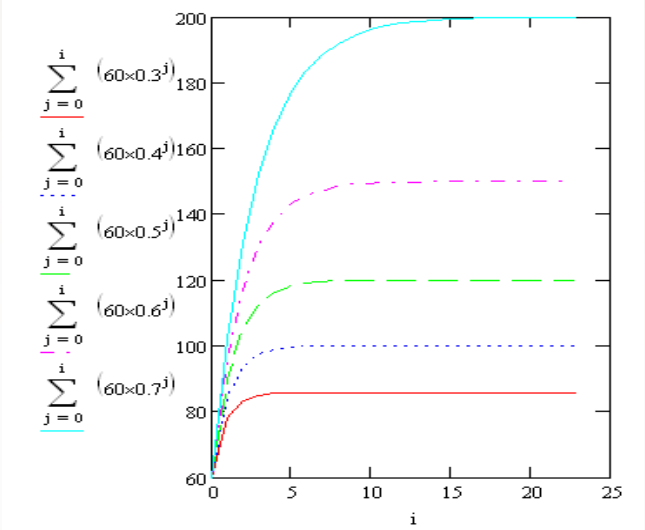
Отримані теоретичні залежності кількості прочитаного матеріалу від часу



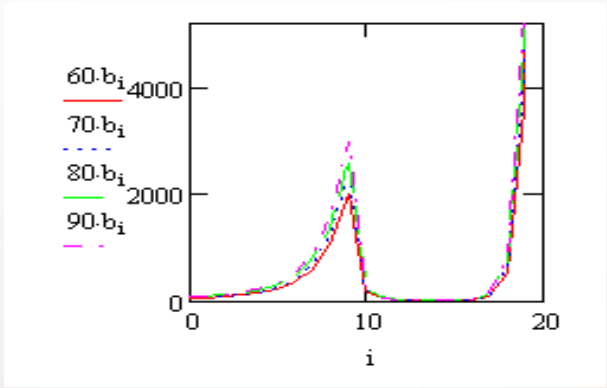
для рівномірної моделі читання

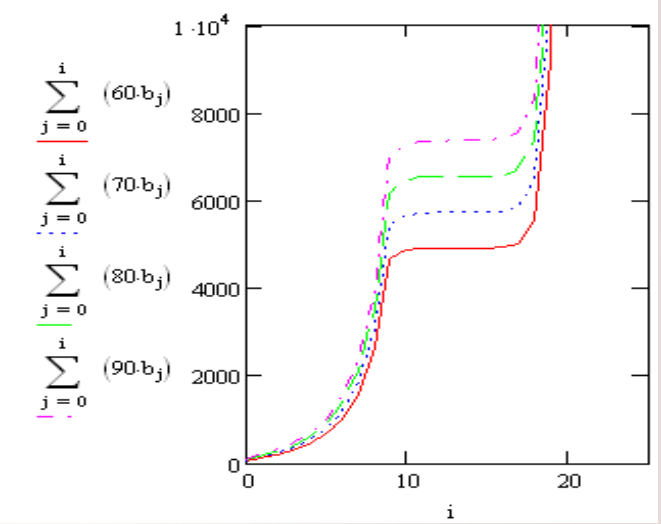


Наступний час однаково більший за попередній



Наступний час однаково менший за попередній





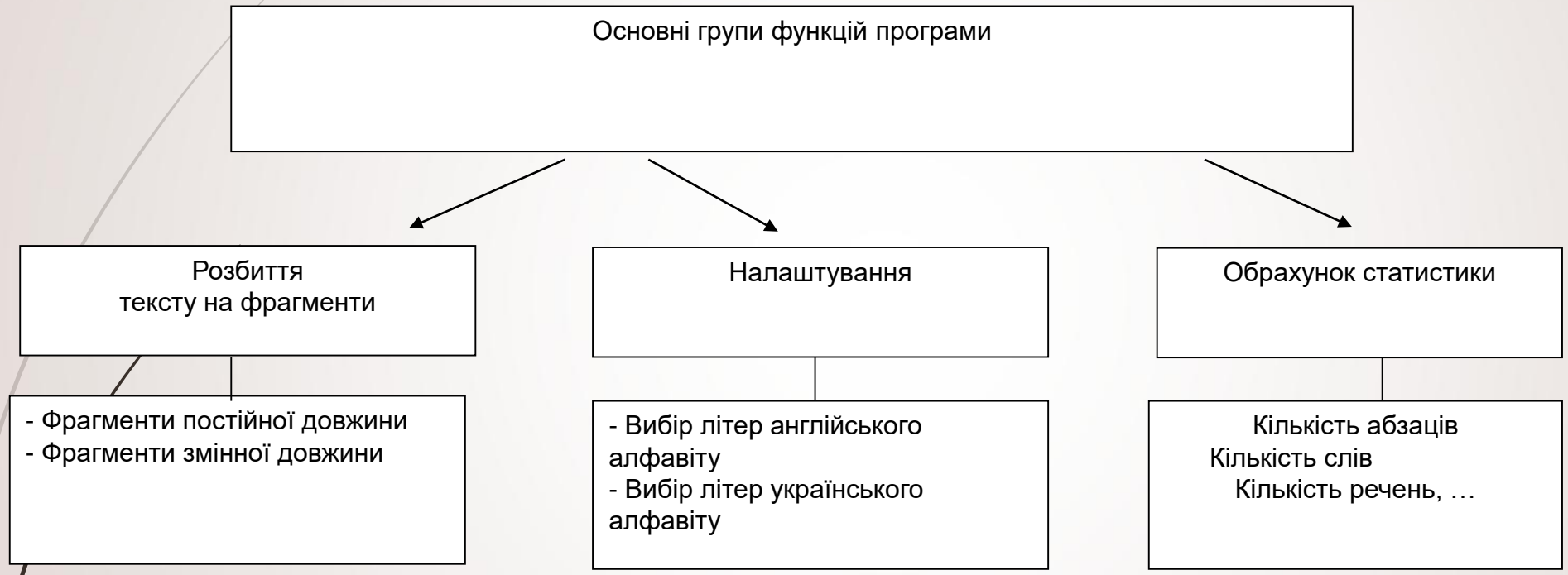
Нерівномірний процес читання

Коливальний процес читання

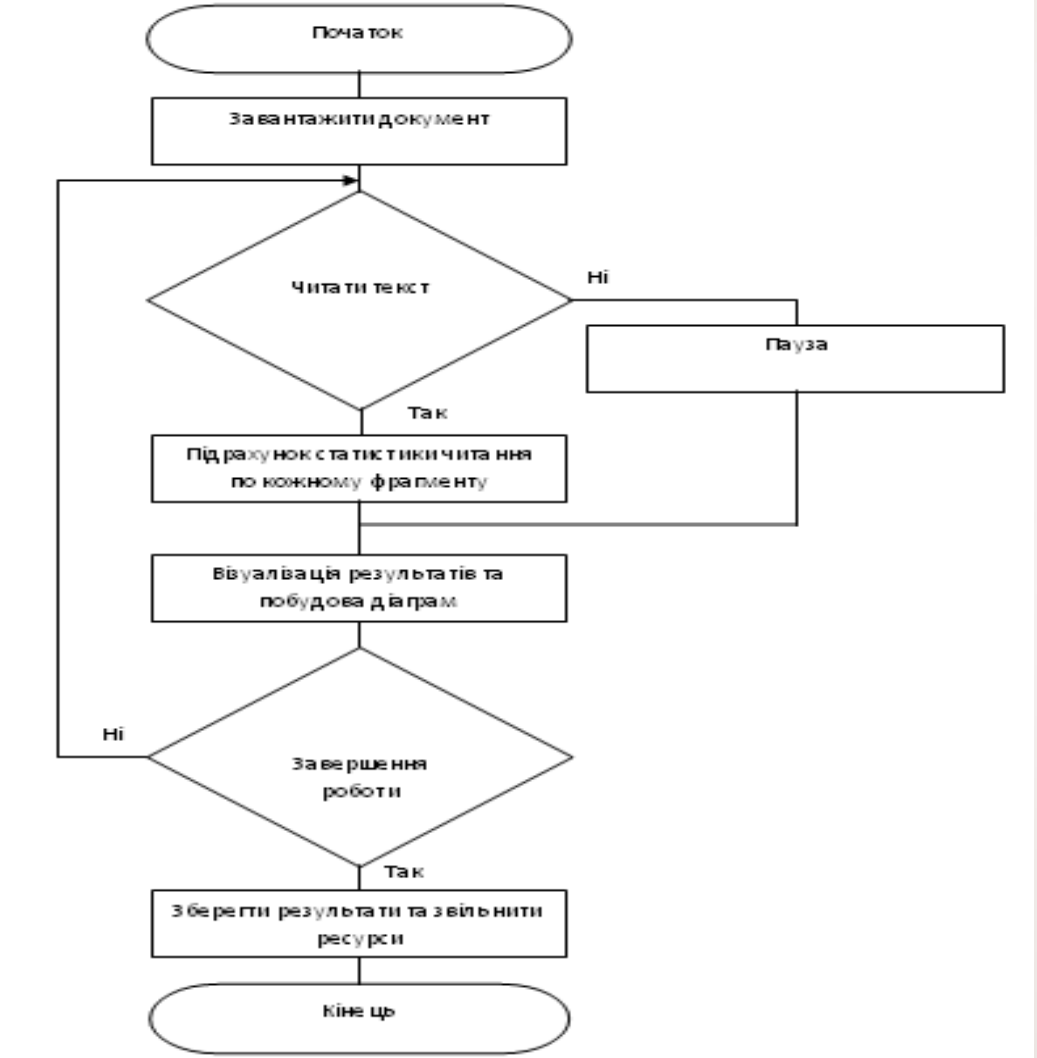
Верхні та нижні теоретичні часові границі процесу читання

Часова модель	Нижня межа $T_{fc\ min}$	Верхня межа $T_{fc\ max}$
$T_{fc}(n) = nT_0$	-	-
$T_{fc}(n) = \sum_{i=0}^{n-1} T_0 \cdot K^i$, де $K^{n-1} > 1$	$T_{fc\ min} = T_{fc}(0) \cdot K^0 = T_{fc}(0)$	$T_{fc\ max} = T_{fc}(0) \cdot K^{n-1}$
$T_{fc}(n) = \sum_{i=0}^{n-1} T_0 \cdot K^i$, де $K^{n-1} < 1$	$T_{fc\ min} = T_{fc}(0) \cdot K^{n-1}$	$T_{fc\ max} = T_{fc}(0) \cdot K^0 = T_{fc}(0)$
$T_{fc}(n) = \sum_{j=0}^{n} (T_0 \cdot \prod_{i=0}^{n-1} K_i)$	$T_{fc\ min} = \min\{T_{fc}(i) \cdot K_i\}$, де i=1…n.	$T_{fc\ max} = \max\{T_{fc}(i) \cdot K_i\}$, де i=1…n.

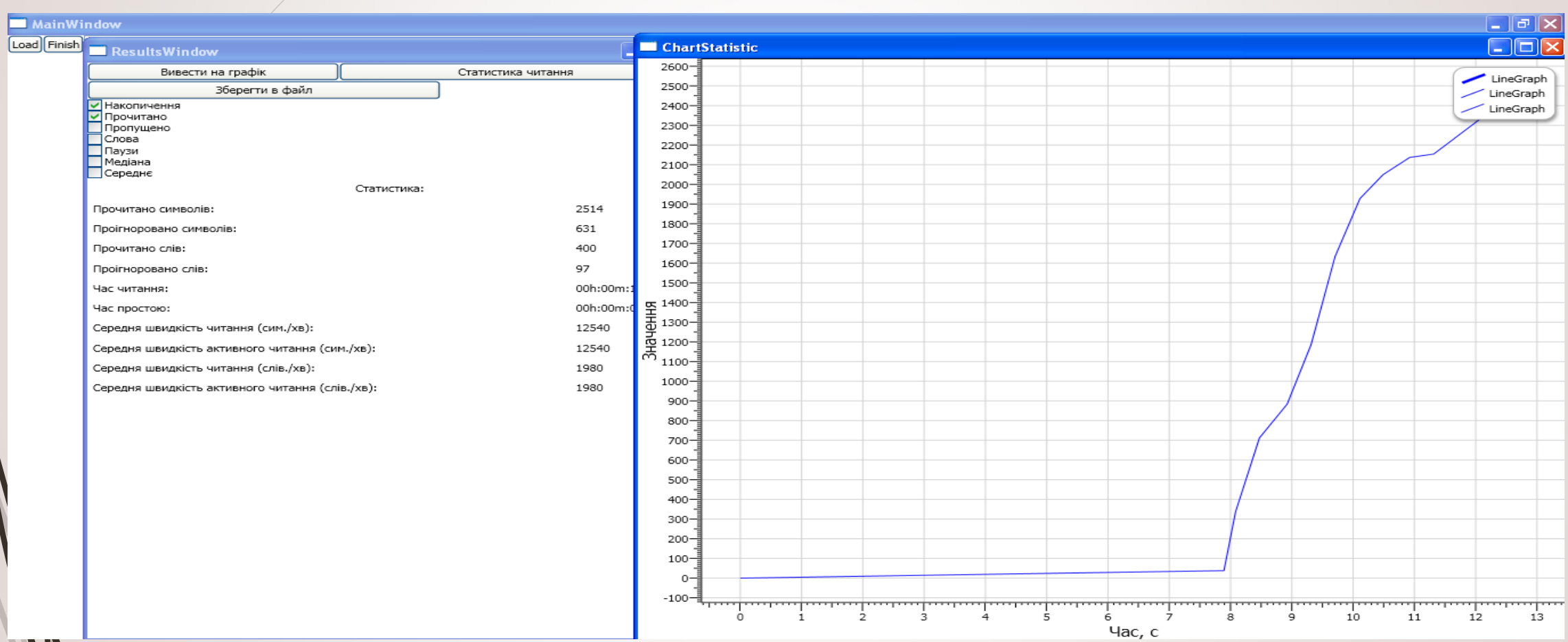
Основні функції розроблюваної програми статистичного аналізу



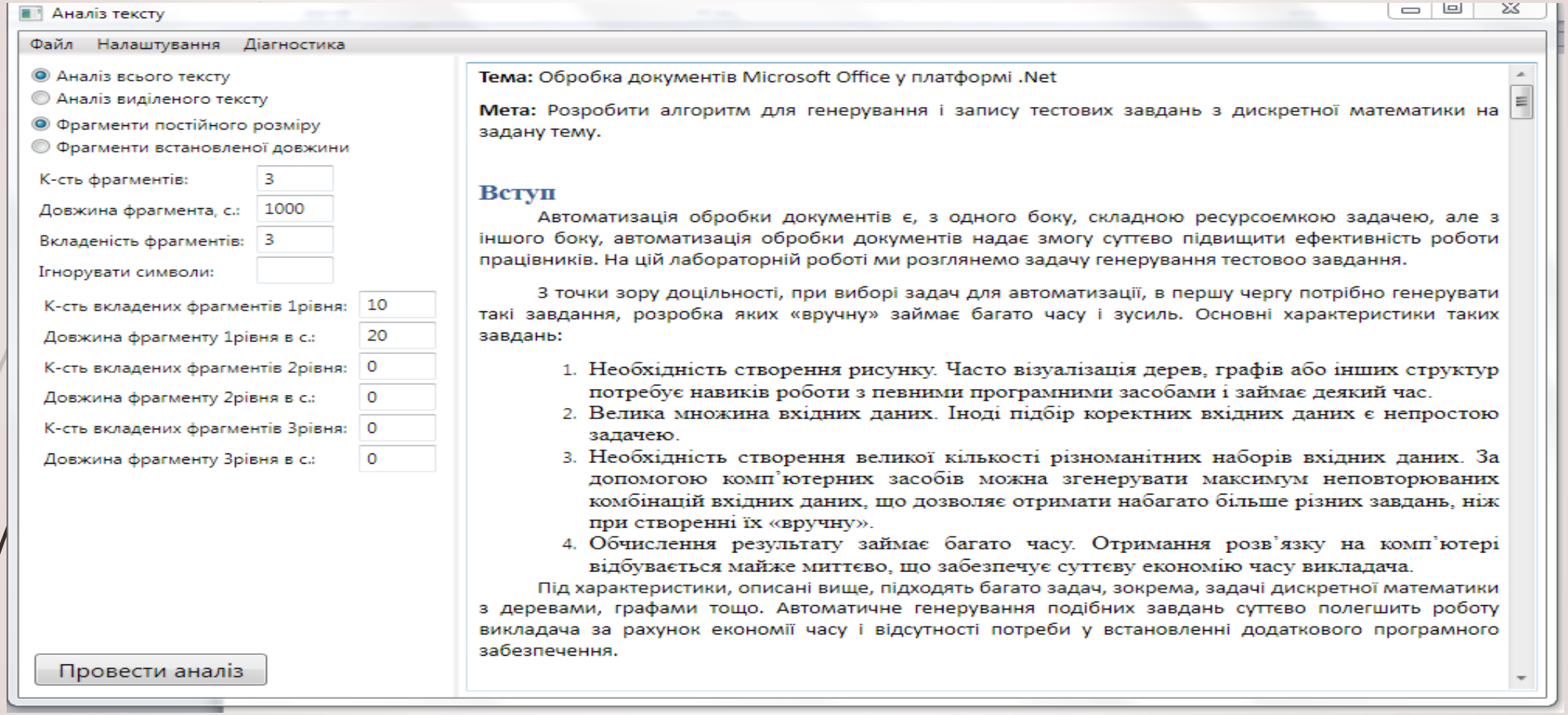
Алгоритм роботи програми аналізу процесу читання



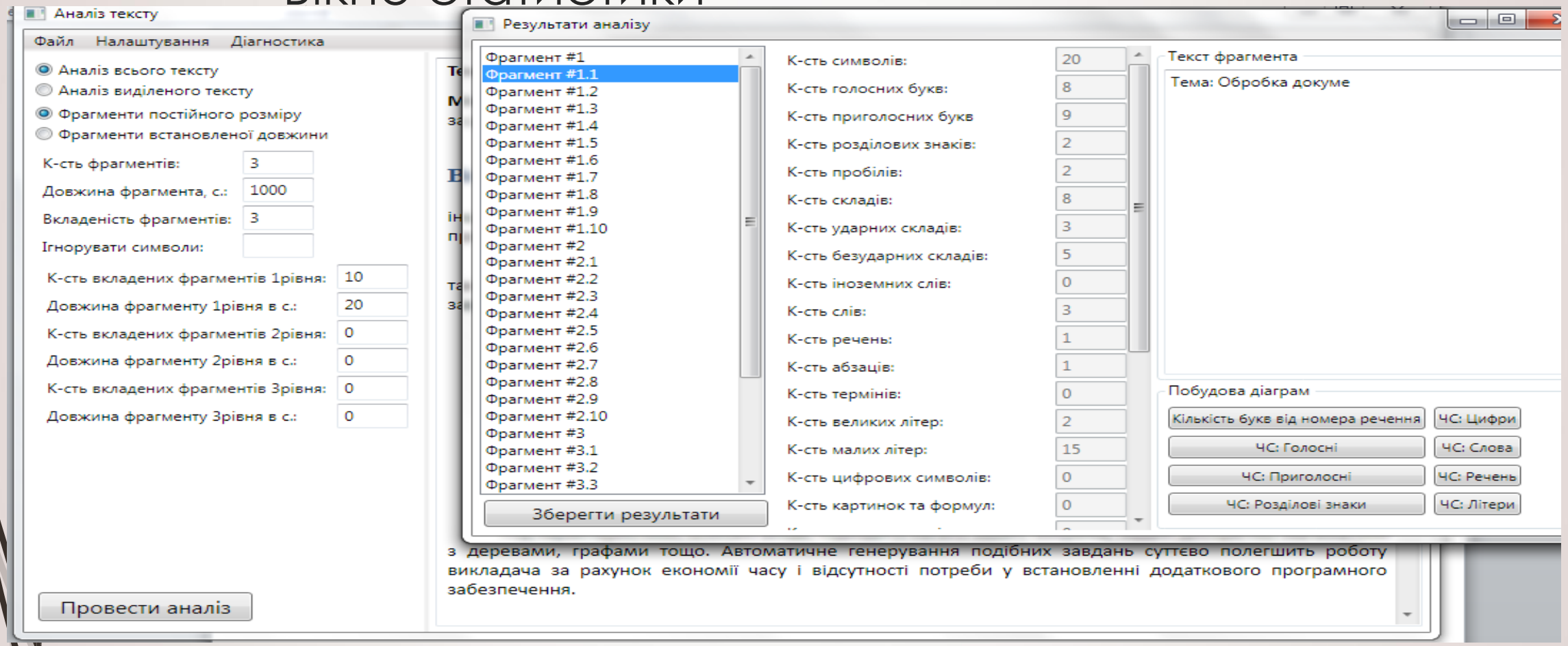
Головне вікно програми аналізу статистики процесу читання



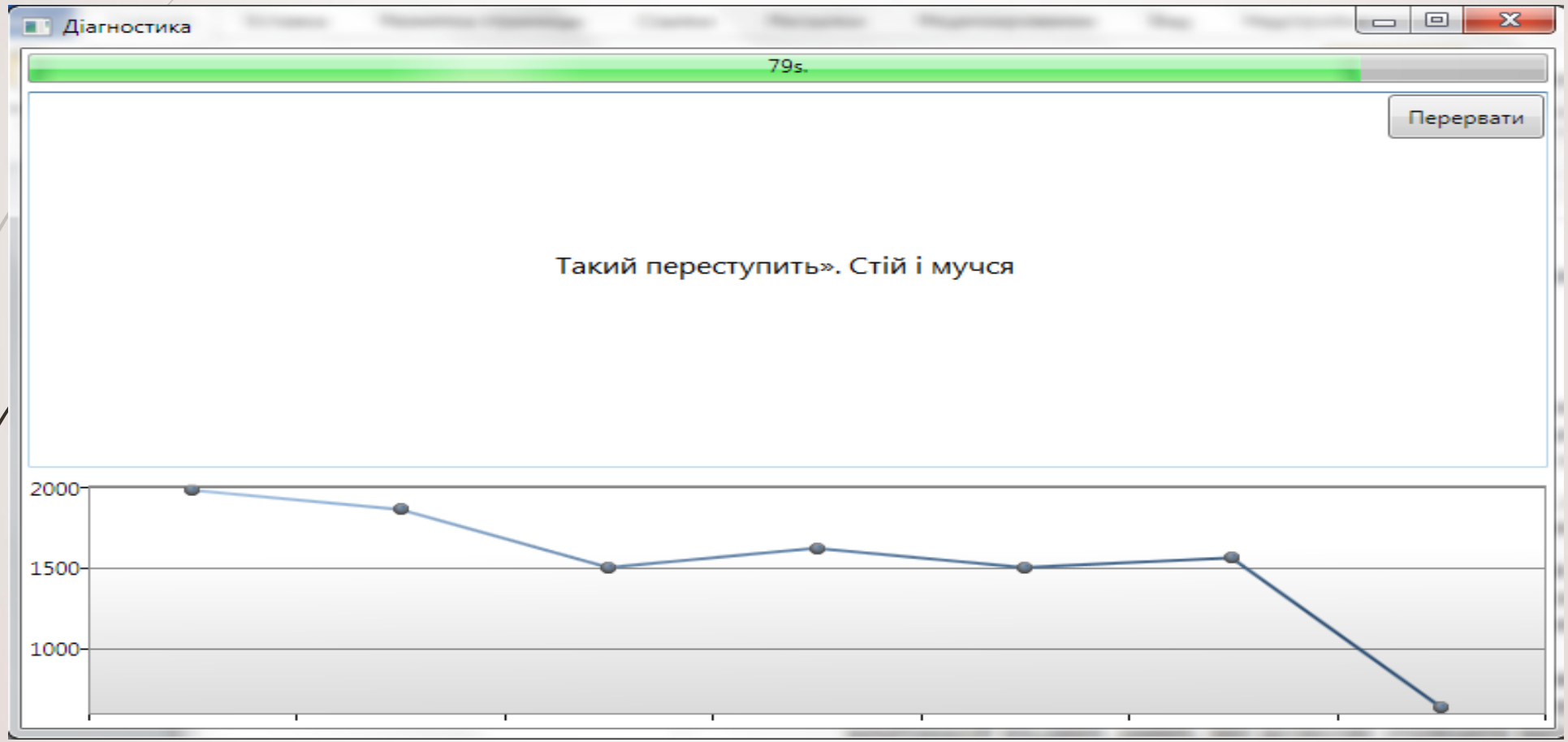
Головне вікно програми аналізу статистики



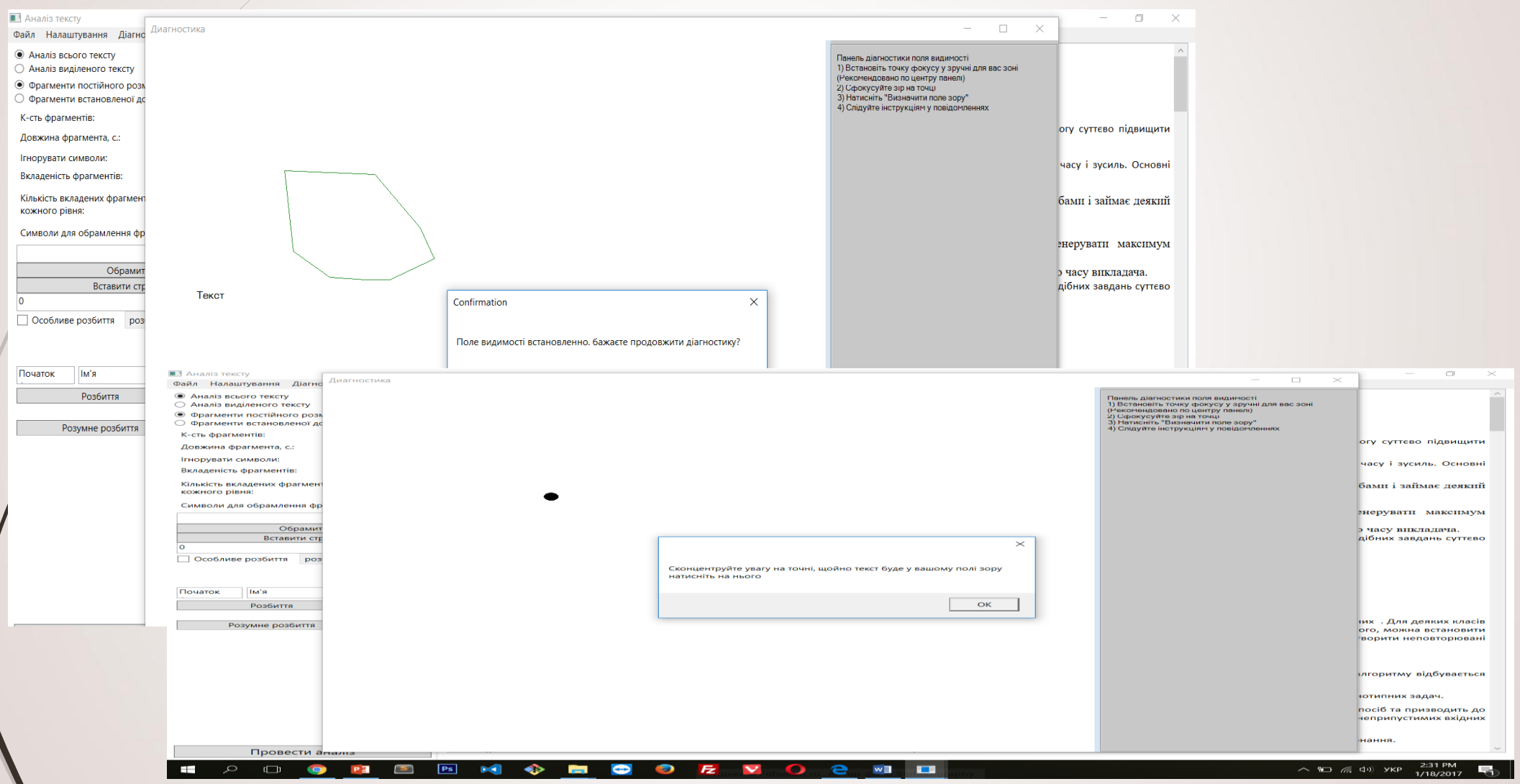
Вікно статистики



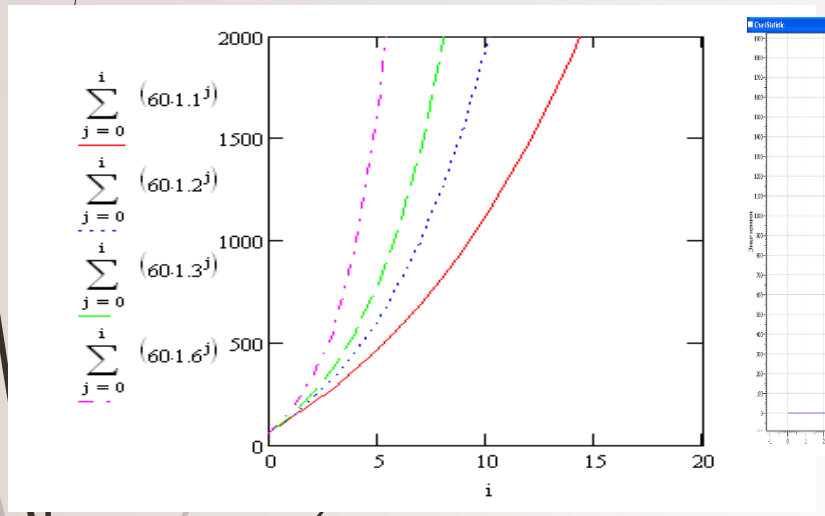
Вікно діагностування

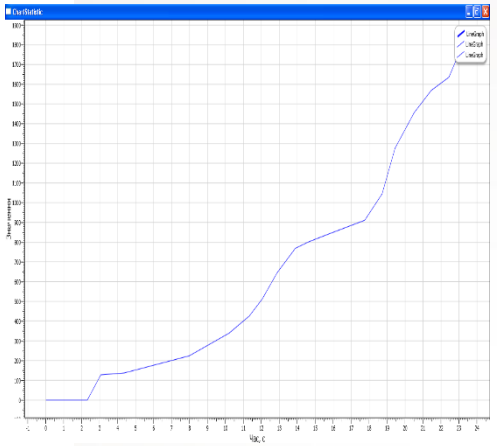


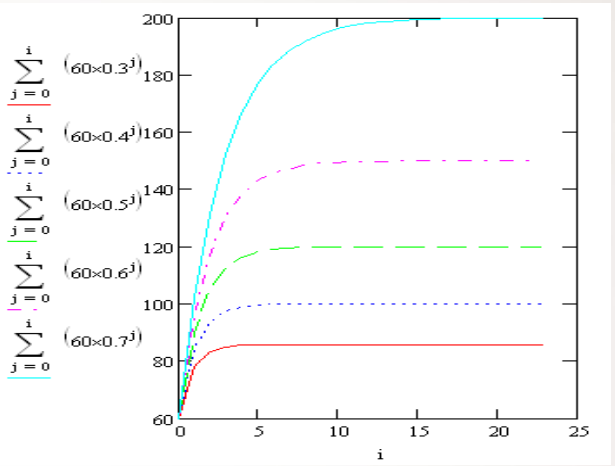
Визначення поля зору

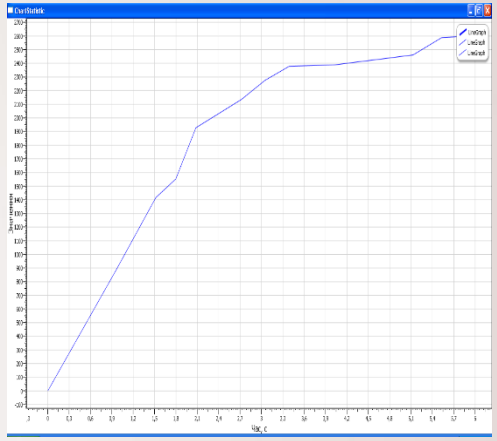


Порівняння теоретичних результатів із практичними результатами

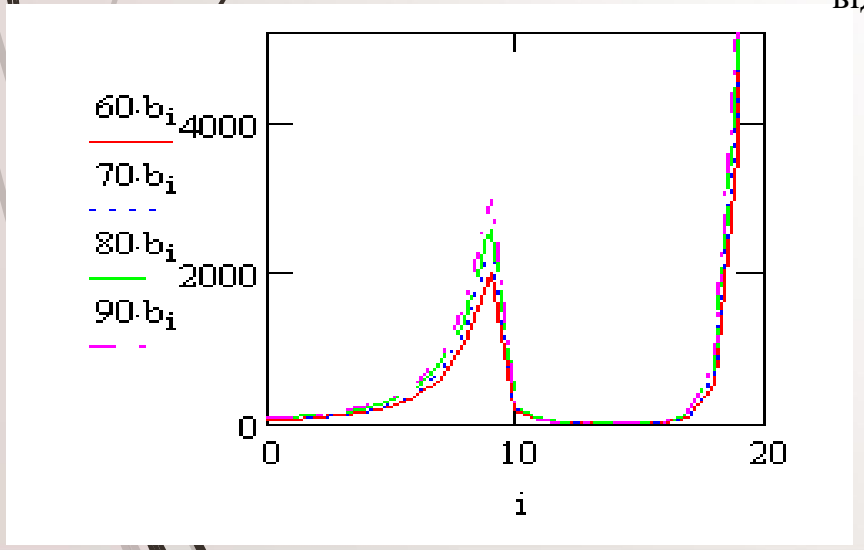


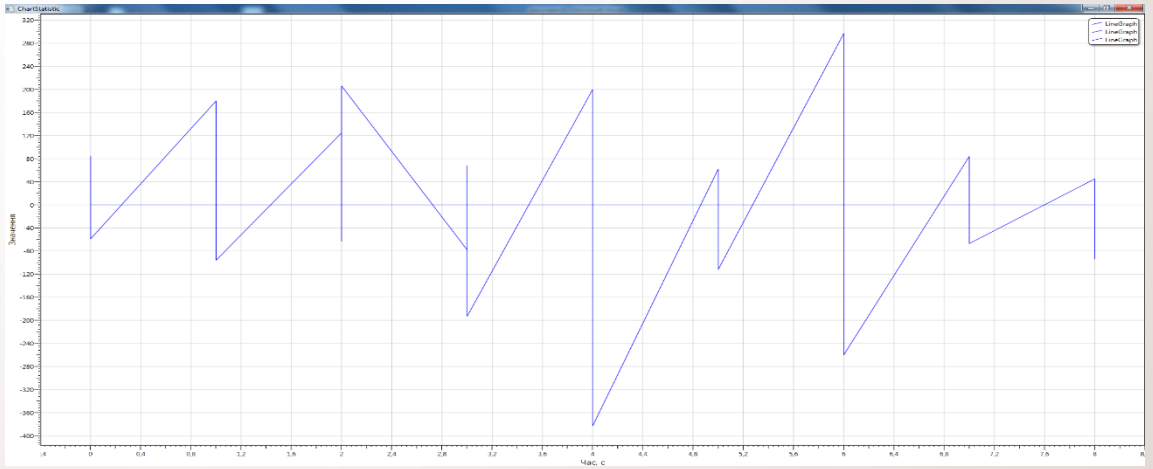






Коливальний процес читання окремих фрагментів тексту відносно умовного нульового рівня (від'ємні значення відповідають пропущеним фрагментам у читанні)





Дякую за увагу!