

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 кваліфікаційної роботи є розширення функціональних можливостей засобів комп'ютерного аналізу статистичних параметрів текстових матеріалів.

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

Предметом дослідження 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.

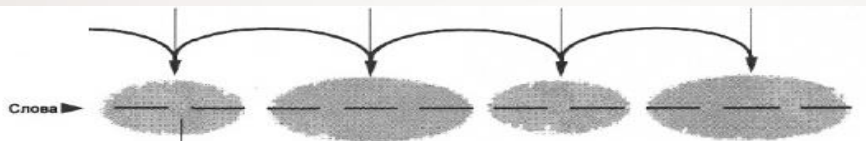
Програмне забезпечення для статистичного аналізу тексту

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

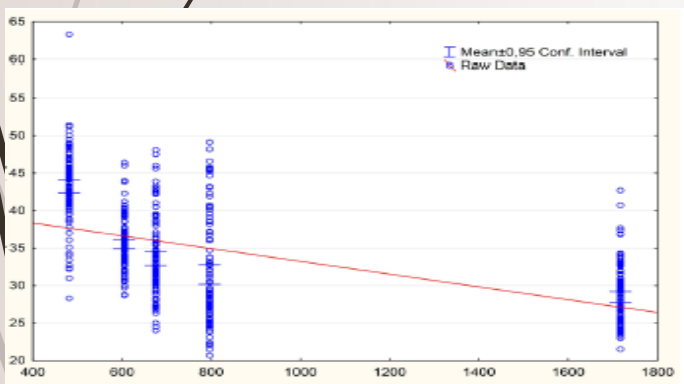


Приклад використання ай-трекінгу на сторінці сайту

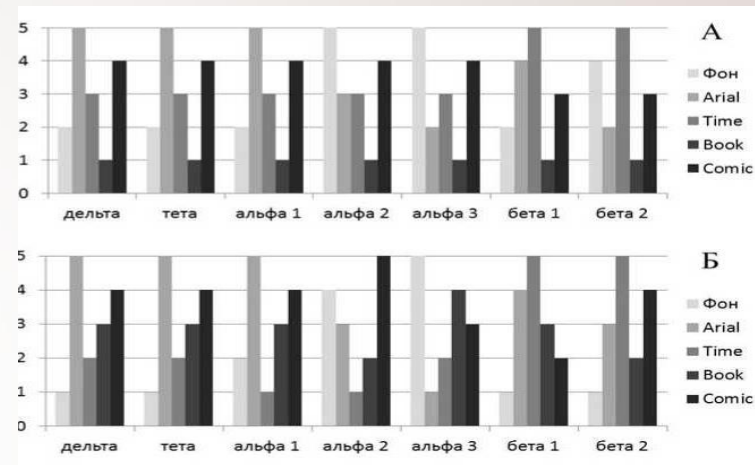


Наприклад, чтобы прочитать данное предложение, ваш зрительный аппарат быстро сканирует визуальные кластеры, в которые входят три-четыре слова. Находя и распознавая знакомые формы, ваш мозг «переводит» их с языка образов на язык значений.

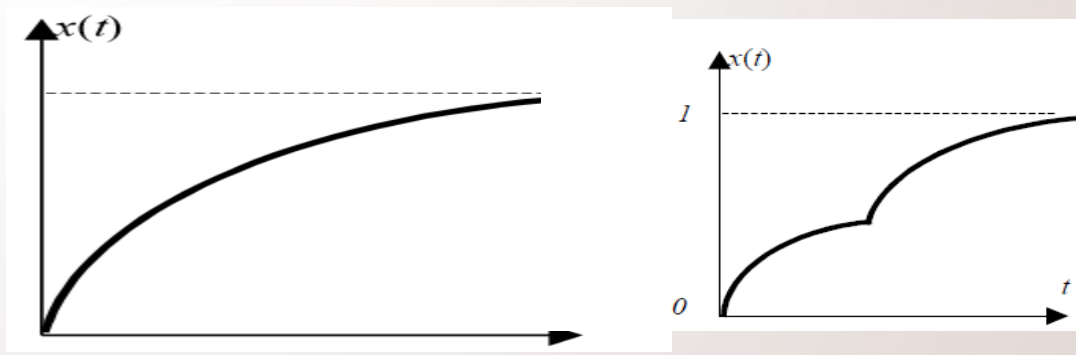
Аналіз параметрів стрибків ока та виділення візуальних кластерів при читанні текстів



Залежності швидкості читання текстів від ізрізаності шрифту



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

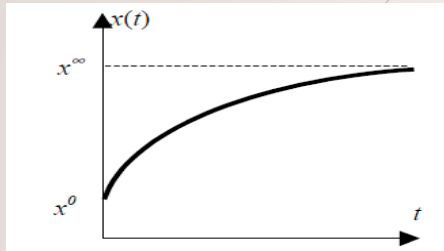


Криві навчання

Залежності, що характеризують читання:

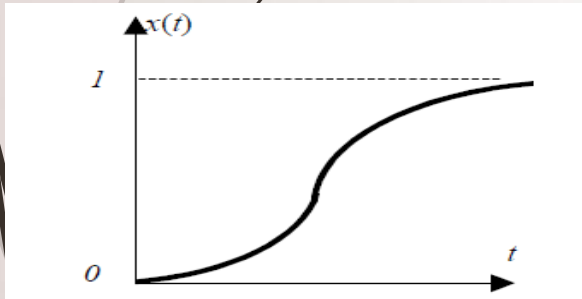
Експоненційна

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



Логістична

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



Факторний аналіз

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

Формула визначення легкості читання тексту, розроблена Р. Флешем

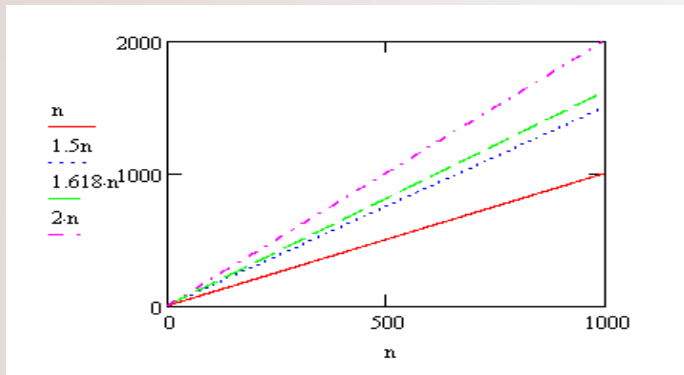
$$\begin{aligned} F_{Flesch} &= 206,835 - \left(\frac{k}{s} * 1,015 + \frac{f}{k} * 84,6 \right) = \\ &= 206,835 - (w * 1,015 + p * 84,6). \end{aligned}$$

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.

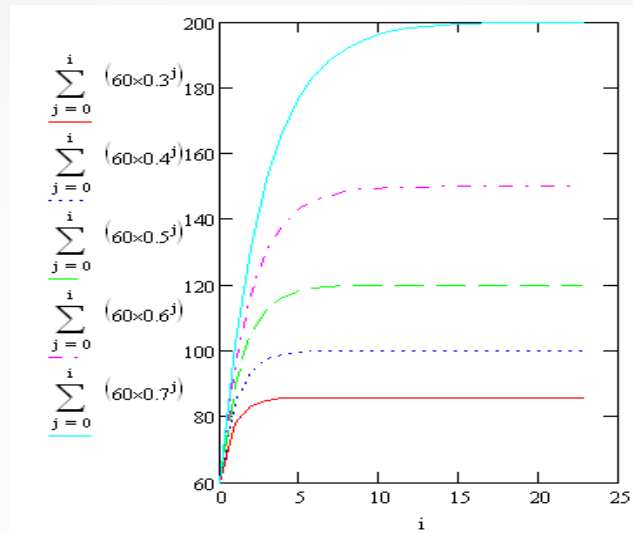
Побудова часових моделей оцінювання процесу читання

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 <u>більше</u> попереднього	$T_{fc}(n) = \sum_{i=0}^{n-1} T_0 \cdot K^i$	$T_{fc}(0) \times K^{n-2}$	$T_{fc}(0) \times K^{n-1}$
Наступний час однаково менший попереднього	$T_{fc}(n) = \sum_{i=0}^{n-1} T_0 \cdot K^i$	$T_{fc}(0) \times K^{n-2}$	$T_{fc}(0) \times K^{n-1}$
Нерівномірний час читання	$T_{fc}(n) = \sum_{j=0}^n (T_0 \cdot \prod_{i=0}^{n-1} K_i)$	$T_{fc}(n-2) \times K_{n-2}$	$T_{fc}(n-1) \times K_{n-1}$

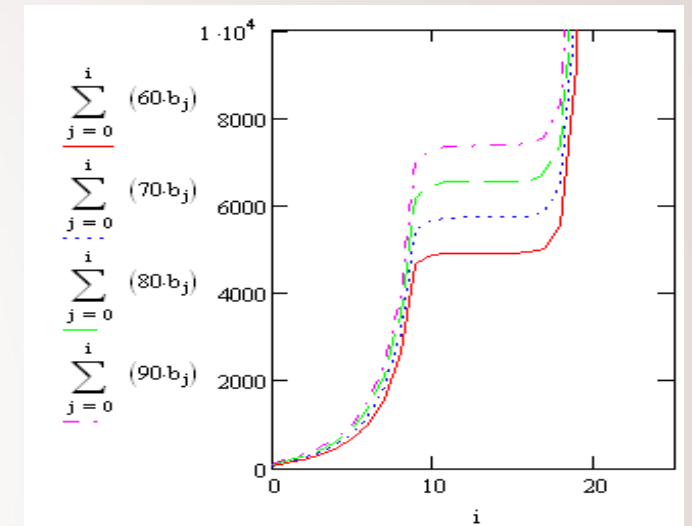
Отримані теоретичні залежності кількості прочитаного матеріалу від часу



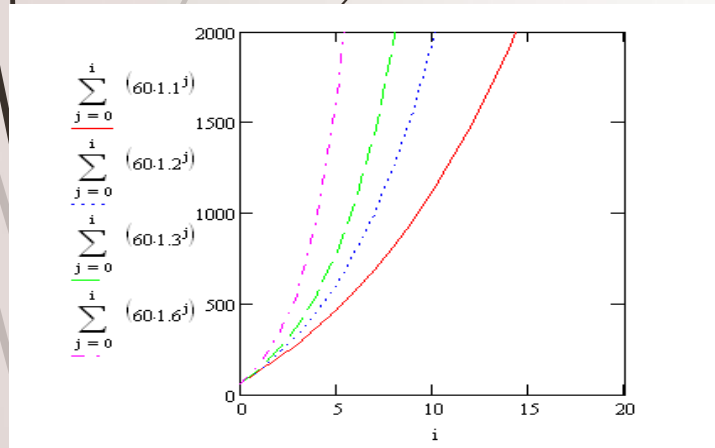
для рівномірної моделі читання



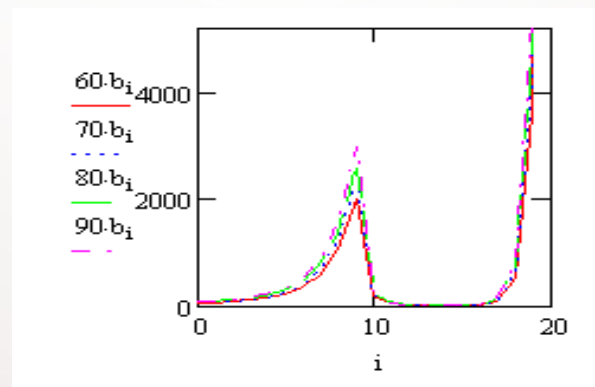
Наступний час однаково менший за попередній



Нерівномірний процес читання



Наступний час однаково більший за попередній

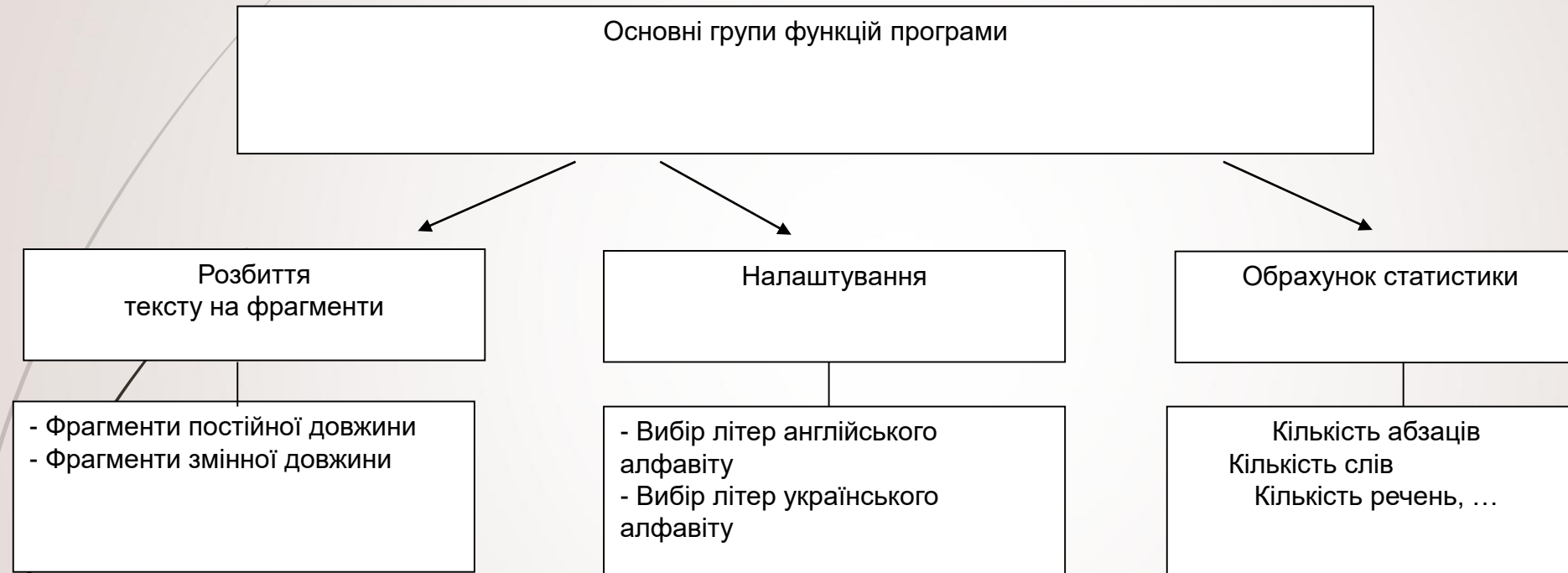


Коливальний процес читання

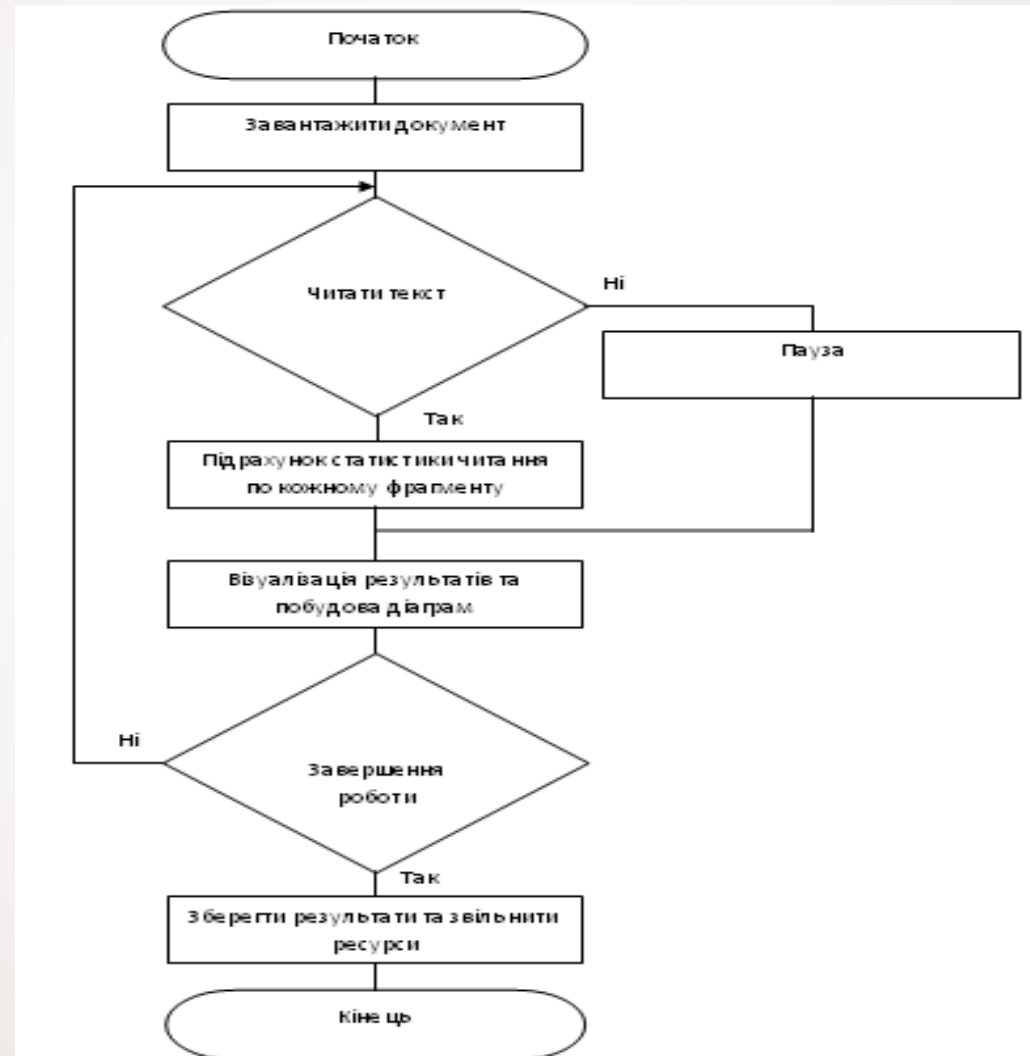
Верхні та нижні теоретичні часові границі процесу читання

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

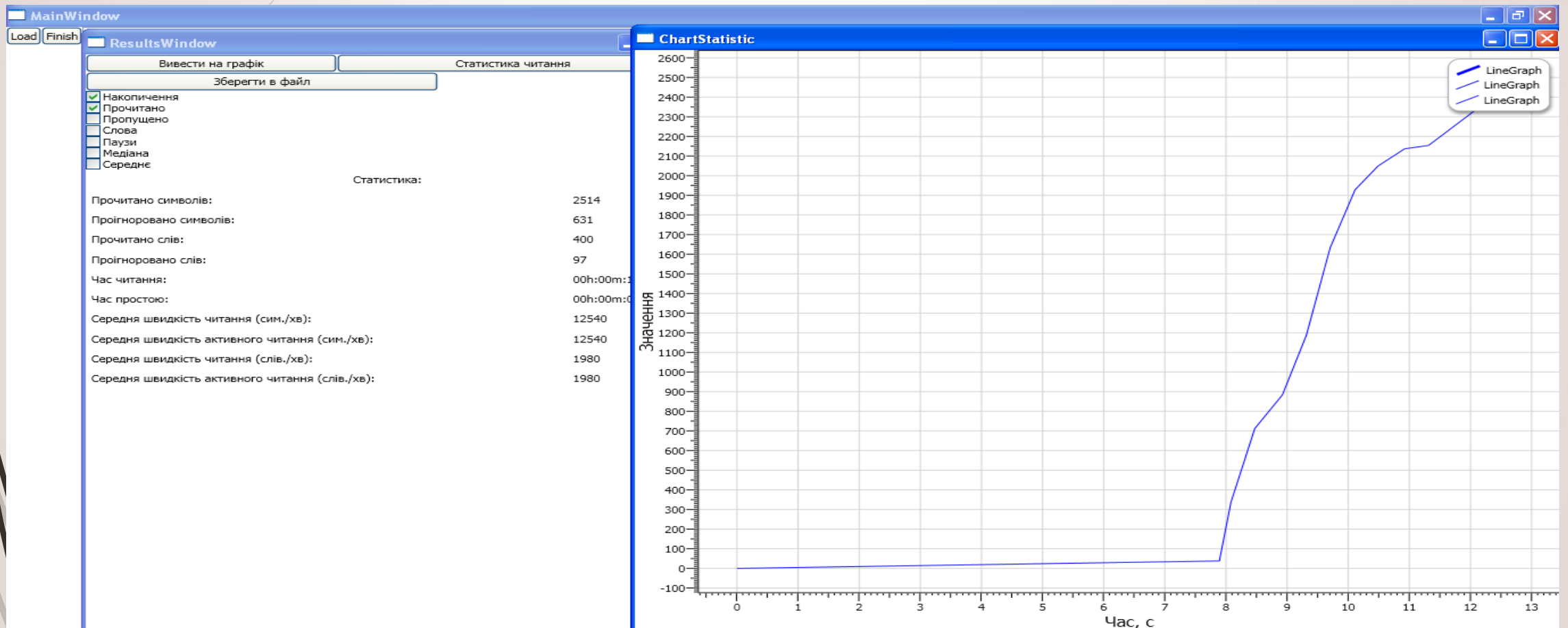
Основні функції розроблюваної програми статистичного аналізу



Алгоритм роботи програми аналізу процесу ЧИТАННЯ



Головне вікно програми аналізу статистики процесу читання



Головне вікно програми аналізу статистики

Аналіз тексту

Файл Налаштування Діагностика

Аналіз всього тексту
 Аналіз виділеного тексту
 Фрагменти постійного розміру
 Фрагменти встановленої довжини

К-сть фрагментів:

Довжина фрагмента, с.:

Вкладеність фрагментів:

Ігнорувати символи:

К-сть вкладених фрагментів 1рівня:

Довжина фрагменту 1рівня в с.:

К-сть вкладених фрагментів 2рівня:

Довжина фрагменту 2рівня в с.:

К-сть вкладених фрагментів 3рівня:

Довжина фрагменту 3рівня в с.:

Провести аналіз

Тема: Обробка документів Microsoft Office у платформі .Net

Мета: Розробити алгоритм для генерування і запису тестових завдань з дискретної математики на задану тему.

Вступ

Автоматизація обробки документів є, з одного боку, складною ресурсоємкою задачею, але з іншого боку, автоматизація обробки документів надає змогу суттєво підвищити ефективність роботи працівників. На цій лабораторній роботі ми розглянемо задачу генерування тестовою завдання.

З точки зору доцільності, при виборі задач для автоматизації, в першу чергу потрібно генерувати такі завдання, розробка яких «вручну» займає багато часу і зусиль. Основні характеристики таких завдань:

1. **Необхідність створення рисунку.** Часто візуалізація дерев, графів або інших структур потребує навиків роботи з певними програмними засобами і займає деякий час.
2. **Велика множина вхідних даних.** Іноді підбір коректних вхідних даних є непростою задачею.
3. **Необхідність створення великої кількості різноманітних наборів вхідних даних.** За допомогою комп'ютерних засобів можна згенерувати максимум неповторюваних комбінацій вхідних даних, що дозволяє отримати набагато більше різних завдань, ніж при створенні їх «вручну».
4. **Обчислення результату займає багато часу.** Отримання розв'язку на комп'ютері відбувається майже миттєво, що забезпечує суттєву економію часу викладача.

Під характеристики, описані вище, підходять багато задач, зокрема, задачі дискретної математики з деревами, графами тощо. Автоматичне генерування подібних завдань суттєво полегшить роботу викладача за рахунок економії часу і відсутності потреби у встановленні додаткового програмного забезпечення.

Вікно статистики

The image shows two overlapping windows from a text analysis application. The background window is titled "Аналіз тексту" (Text Analysis) and contains a menu bar with "Файл", "Налаштування", and "Діагностика". It has four radio buttons for analysis types: "Аналіз всього тексту" (selected), "Аналіз виділеного тексту", "Фрагменти постійного розміру", and "Фрагменти встановленої довжини". Below are input fields for "К-сть фрагментів:" (3), "Довжина фрагмента, с.:" (1000), "Вкладеність фрагментів:" (3), and "Ігнорувати символи:". There are also three rows of input fields for "К-сть вкладених фрагментів" at levels 1, 2, and 3, with corresponding "Довжина фрагменту" in characters. A "Провести аналіз" button is at the bottom.

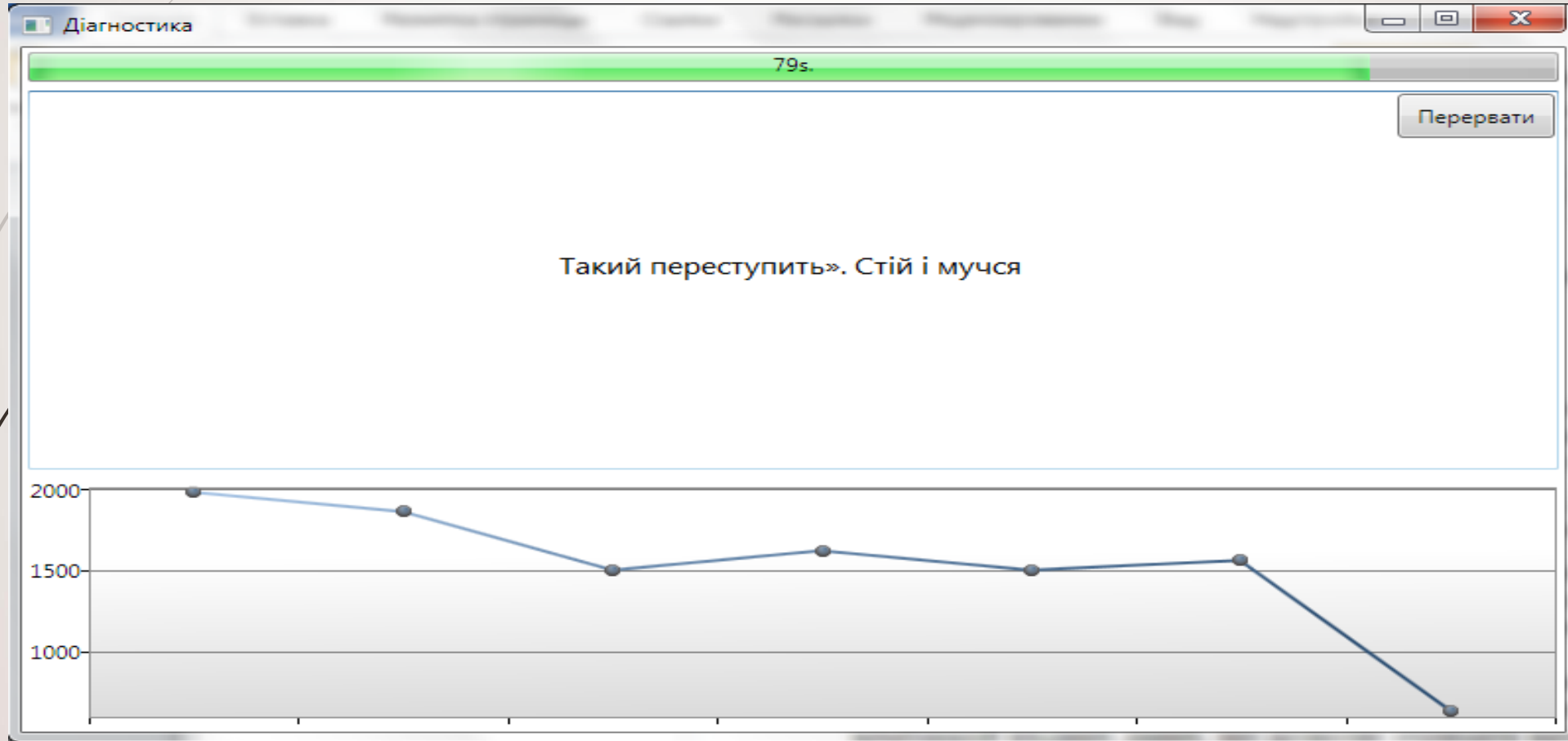
The foreground window is titled "Результати аналізу" (Analysis Results) and displays a list of fragments on the left, with "Фрагмент #1.1" selected. To the right is a table of statistics:

К-сть символів:	20
К-сть голосних букв:	8
К-сть приголосних букв:	9
К-сть розділових знаків:	2
К-сть пробілів:	2
К-сть складів:	8
К-сть ударних складів:	3
К-сть безударних складів:	5
К-сть іноземних слів:	0
К-сть слів:	3
К-сть речень:	1
К-сть абзаців:	1
К-сть термінів:	0
К-сть великих літер:	2
К-сть малих літер:	15
К-сть цифрових символів:	0
К-сть картинок та формул:	0

At the bottom of this window is a "Зберегти результати" button. On the right side of the "Результати аналізу" window, there is a "Текст фрагмента" section with the text "Тема: Обробка докуме" and a "Побудова діаграм" section with several buttons: "Кількість букв від номера речення", "ЧС: Цифри", "ЧС: Голосні", "ЧС: Слова", "ЧС: Приголосні", "ЧС: Речень", "ЧС: Розділові знаки", and "ЧС: Літери".

з деревами, графами тощо. Автоматичне генерування подібних завдань суттєво полегшить роботу викладача за рахунок економії часу і відсутності потреби у встановленні додаткового програмного забезпечення.

Вікно діагностування



Визначення поля зору

The image displays two overlapping screenshots of a software application titled "Аналіз тексту" (Text Analysis). The interface includes a menu bar with "Файл", "Налаштування", and "Діагностика". On the left, there are several configuration options:

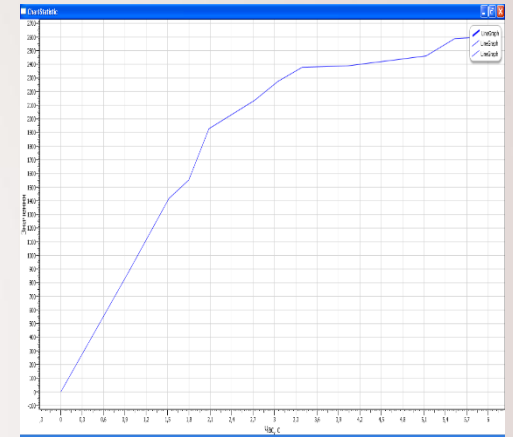
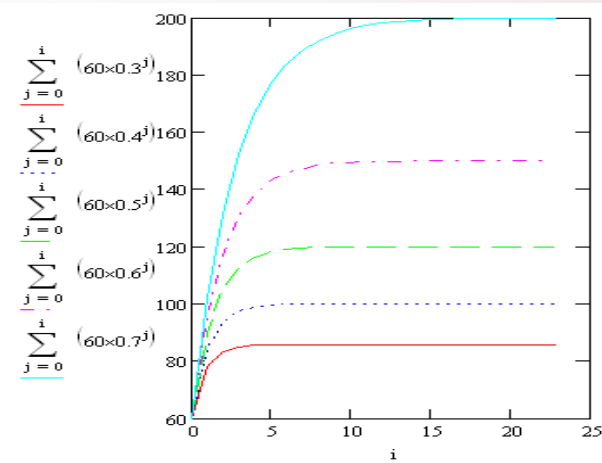
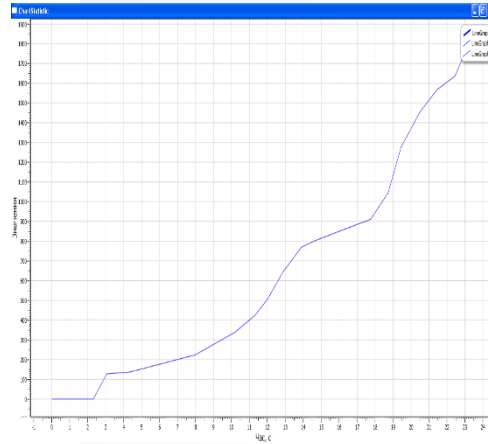
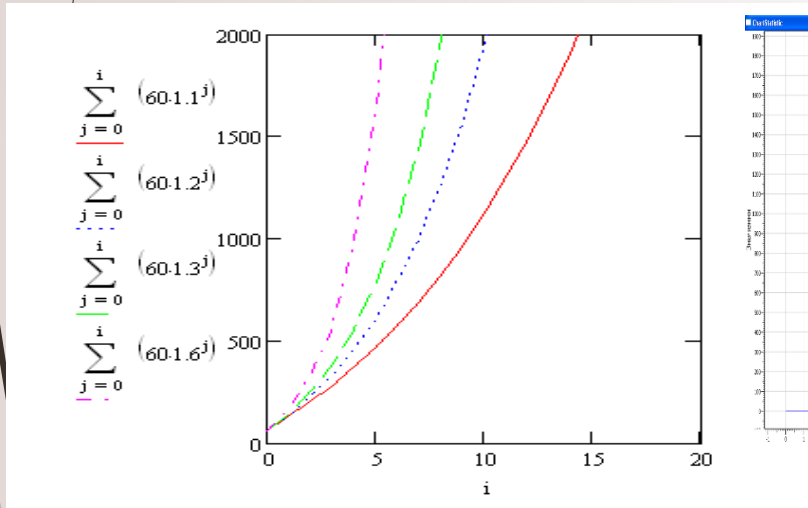
- Radio buttons for "Аналіз всього тексту" (selected), "Аналіз виділеного тексту", "Фрагменти постійного розк" (partially visible), and "Фрагменти встановленої д" (partially visible).
- Input fields for "К-сть фрагментів:", "Довжина фрагмента, с.:", "Ігнорувати символи:", and "Вкладеність фрагментів:".
- A section for "Символи для обрамлення фр" with a list of symbols and buttons for "Обрамит" and "Вставити стр" (partially visible).
- Buttons for "Початок", "Ім'я", "Розбиття", and "Розумне розбиття".

The main window is titled "Діагностика" (Diagnosis). The top screenshot shows a green polygon representing a visual field on a white background. A "Confirmation" dialog box is open, asking: "Поле видимості встановлено. бажаєте продовжити діагностику?" (Visual field set. Do you want to continue the diagnosis?). A "Панель діагностики поля видимості" (Visual field diagnosis panel) is also visible, containing instructions: "1) Встановіть точку фокусу у зручні для вас зоні (Рекомендовано по центру панелі)", "2) Сфокусуйте зір на точці", "3) Натисніть 'Визначити поле зору'", and "4) Слідуйте інструкціям у повідомленнях".

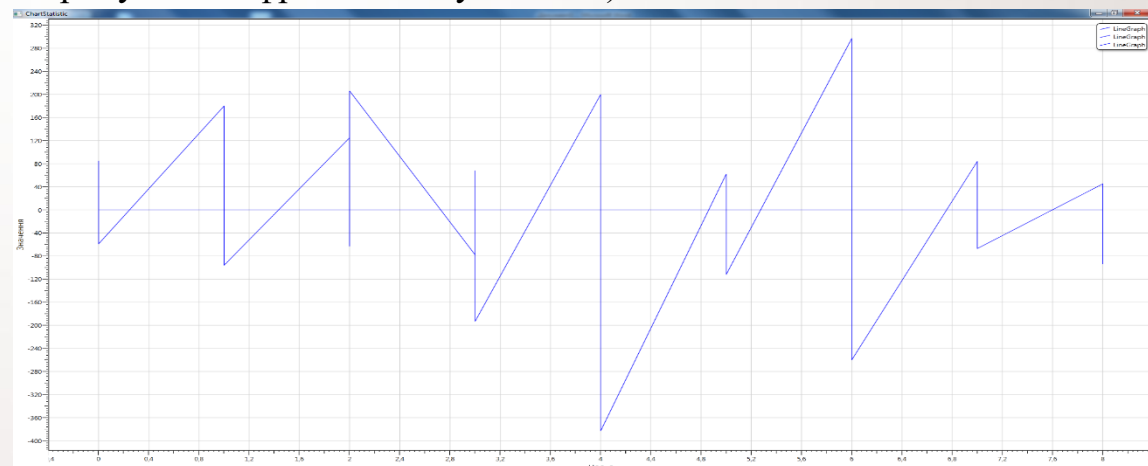
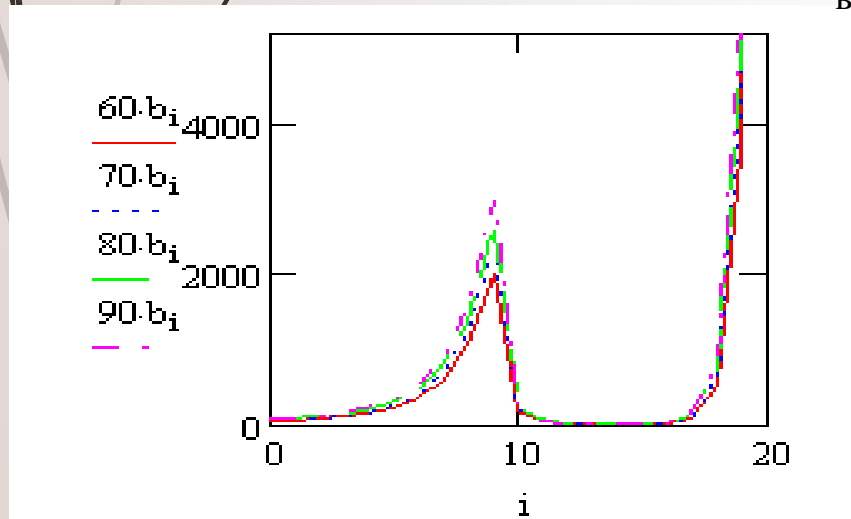
The bottom screenshot shows the same interface but with a black dot in the center of the visual field. A dialog box asks: "Сконцентруйте увагу на точці, щойно текст буде у вашому полі зору натисніть на нього." (Concentrate your attention on the point, as soon as the text is in your visual field, click on it). The "Панель діагностики" is also visible in this screenshot.

The Windows taskbar at the bottom shows the system tray with the date and time: "1/16/2017 2:31 PM".

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!