Вінницький національний технічний університет
Факультет комп’ютерних систем та автоматики

Магістерська кваліфікаційна робота на тему::

Розробка методів та дослідження моделей обробки текстових документів в інтелектуальних системах автоматизації

Керівник дипломного проекту:
к.т.н., професор Биков М.М.

Розробив: студент групи
2АКІТ-18м Калінчук Р.С.

Актуальність

- На сьогоднішній день представлення текстів за допомогою графічних форматів (таких, наприклад, як *.pdf або *.djv) дозволяє розв’язати проблему підвищення швидкості їх електронізації, однак вимагає наявності людини для опрацювання з метою аналізу і розуміння. Використання традиційних технологій електронізації документів в текстових форматах, які дозволяють представити символи в ASCII кодах, і, таким чином, автоматизувати їх аналіз, передбачає посимвольне розпізнавання графічного зображення тексту за допомогою наявних програмних засобів (наприклад, FineReader). Однак такі технології в своїх історичних витоках орієнтовані на брак апаратних ресурсів (швидкодії і пам’яті), не враховують технічних можливостей сучасних обчислювальних систем і мікропроцесорних засобів, а також не використовують мовних складових в інформаційній ієрархії текстового документа. Тому тема магістерської роботи, присвячена розробці ефективних методів обробки текстових документів для електронізації , є актуальною. Запропонована технологія для підвищення швидкості і надійності введення і розпізнавання передбачає використання поряд з параметричними складовими графічного опису тексту також мовних складових: лексичних, морфологічних і синтаксичних

СТРУКТУРА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

UML-ДІАГРАМА КЛАСІВ

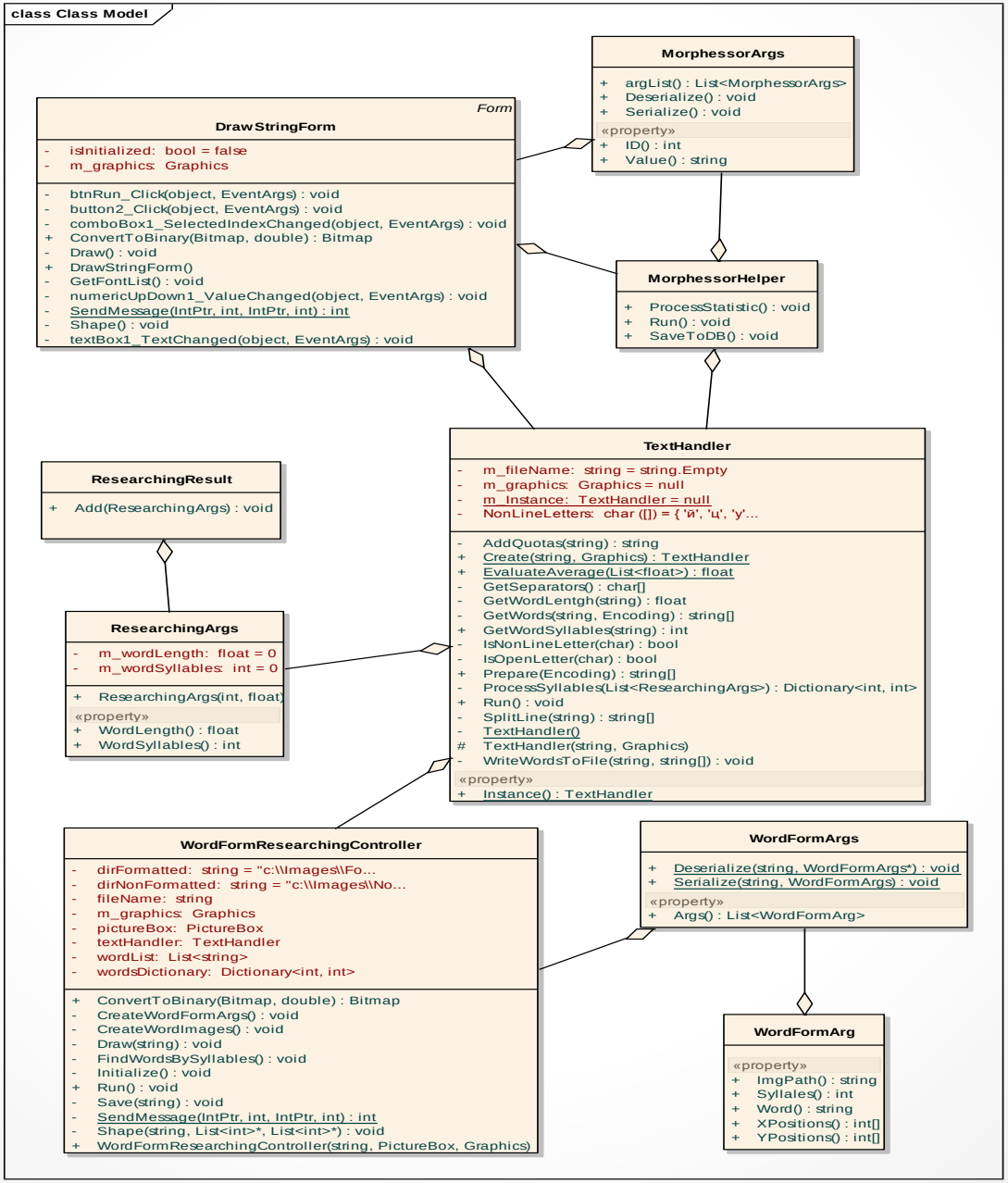


СХЕМА АЛГОРИТМУ ПІДГОТОВКИ КОРПУСУ ТЕКСТУ ДО ТЕСТУВАННЯ

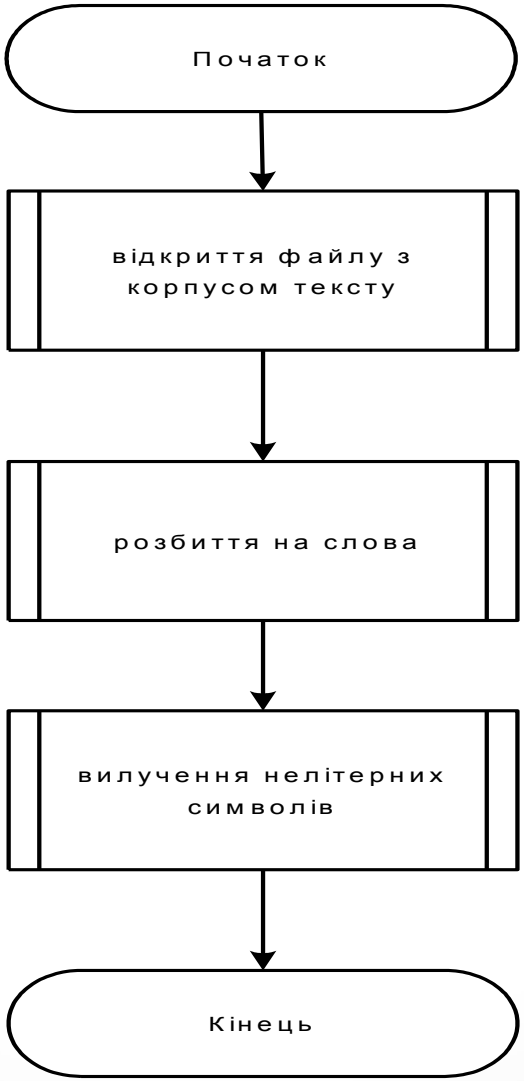


СХЕМА АЛГОРИТМУ ПОШУКУ МОРФЕМ УКРАЇНСЬКОЇ МОВИ

СХЕМА АЛГОРИТМУ ВИЗНАЧЕННЯ СТАТИСТИЧНИХ ЙМОВІРНОСТЕЙ ПОЯВИ МОРФЕМ

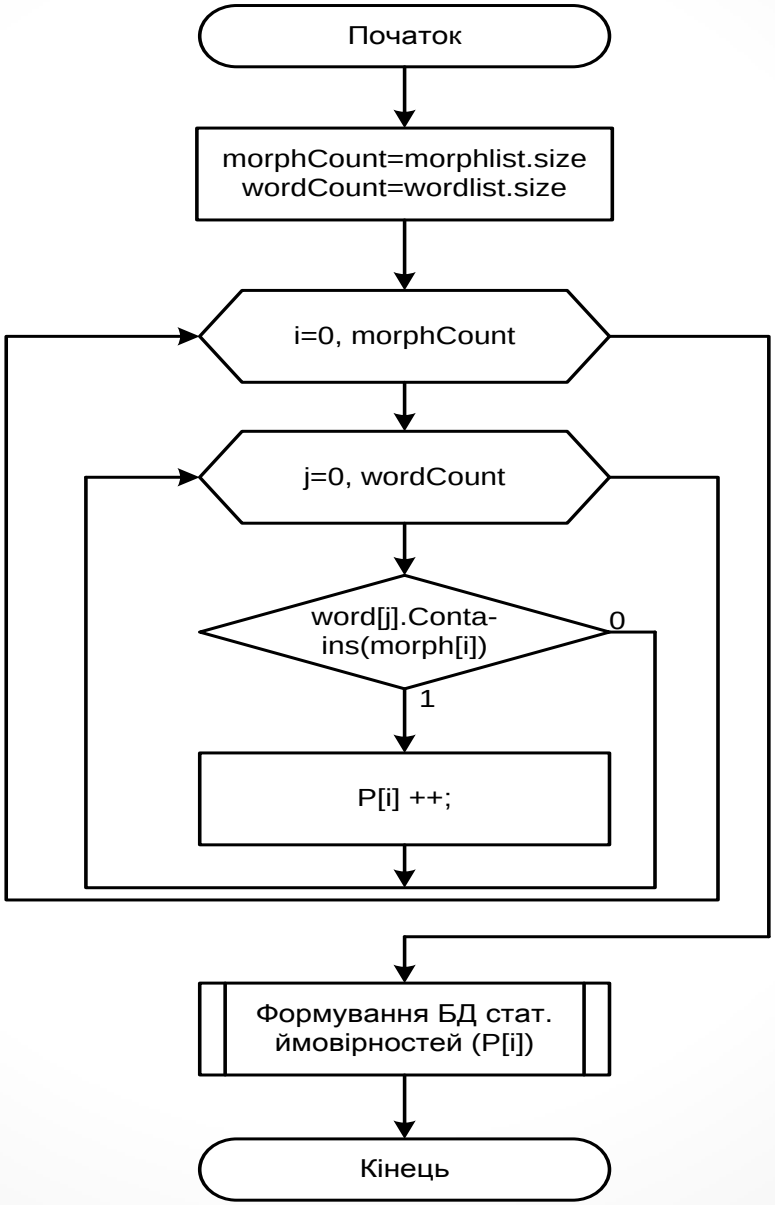


СХЕМА АЛГОРИТМУ ВИЗНАЧЕННЯ ЙМОВІРНОСТЕЙ ПЕРЕХОДІВ МІЖ МОРФЕМАМИ КОРПУСУ ТЕКСТУ

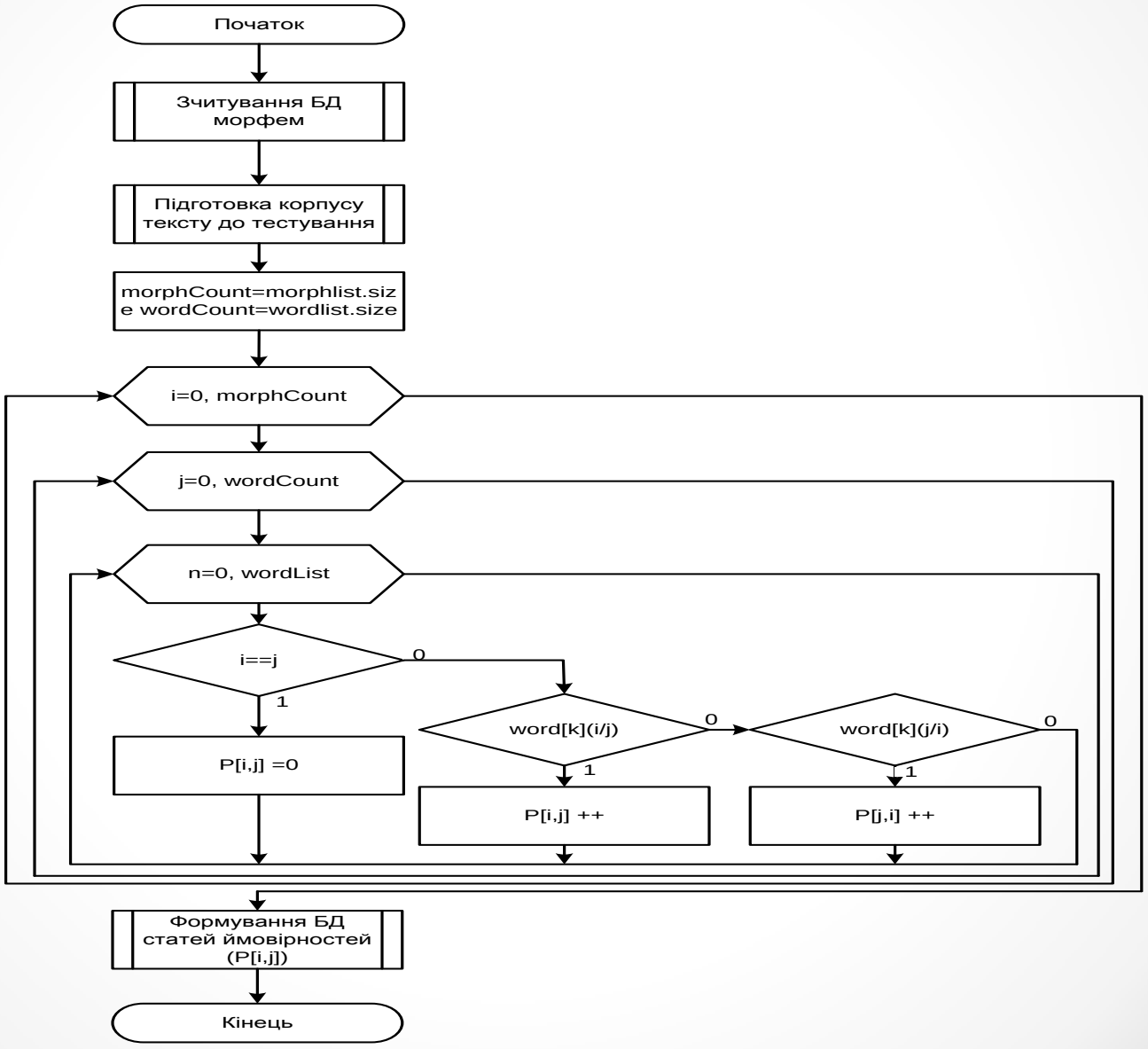


СХЕМА АЛГОРИТМУ ВИЗНАЧЕННЯ ЗАЛЕЖНОСТІ ДОВЖИНИ СЛОВА ВІД КІЛЬКОСТІ СКЛАДІВ

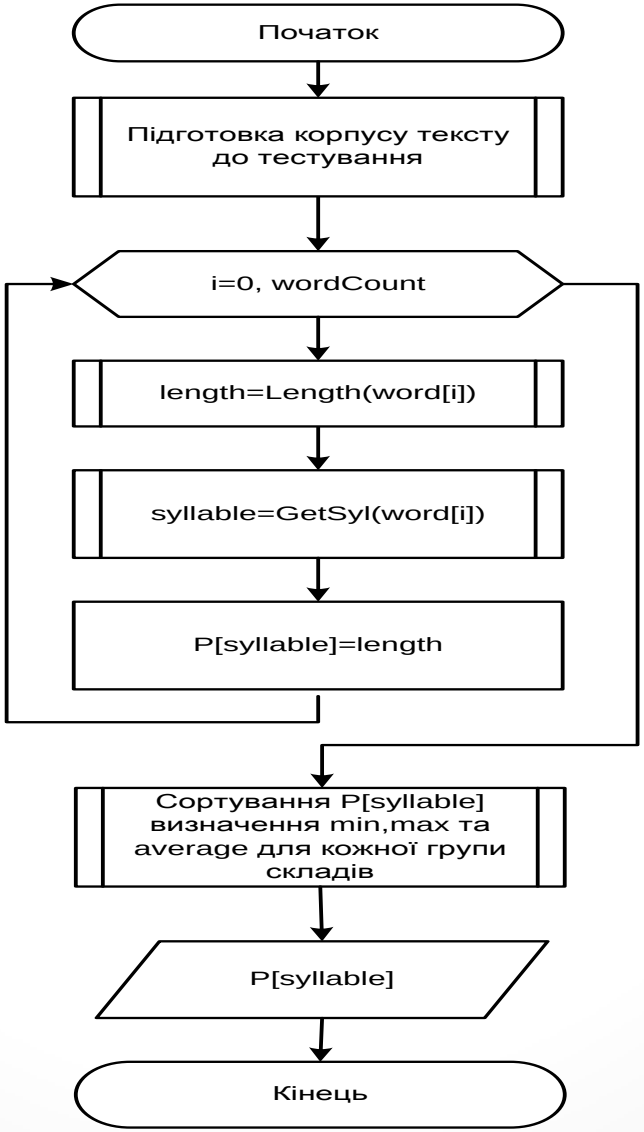


СХЕМА АЛГОРИТМУ ВИЗНАЧЕННЯ СТАТИСТИЧНИХ ЙМОВІРНОСТЕЙ ПОЯВИ ЛІТЕР, ЩО МАЮТЬ НАДРЯДКОВІ ТА ПІДРЯДКОВІ ОЗНАКИ

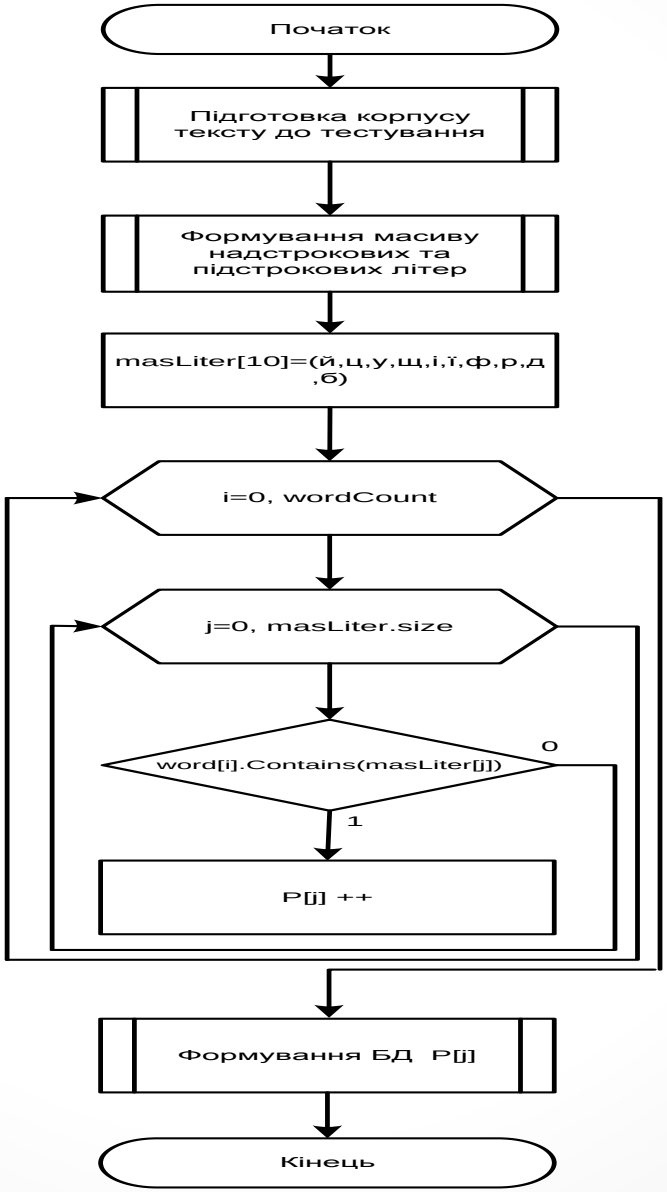
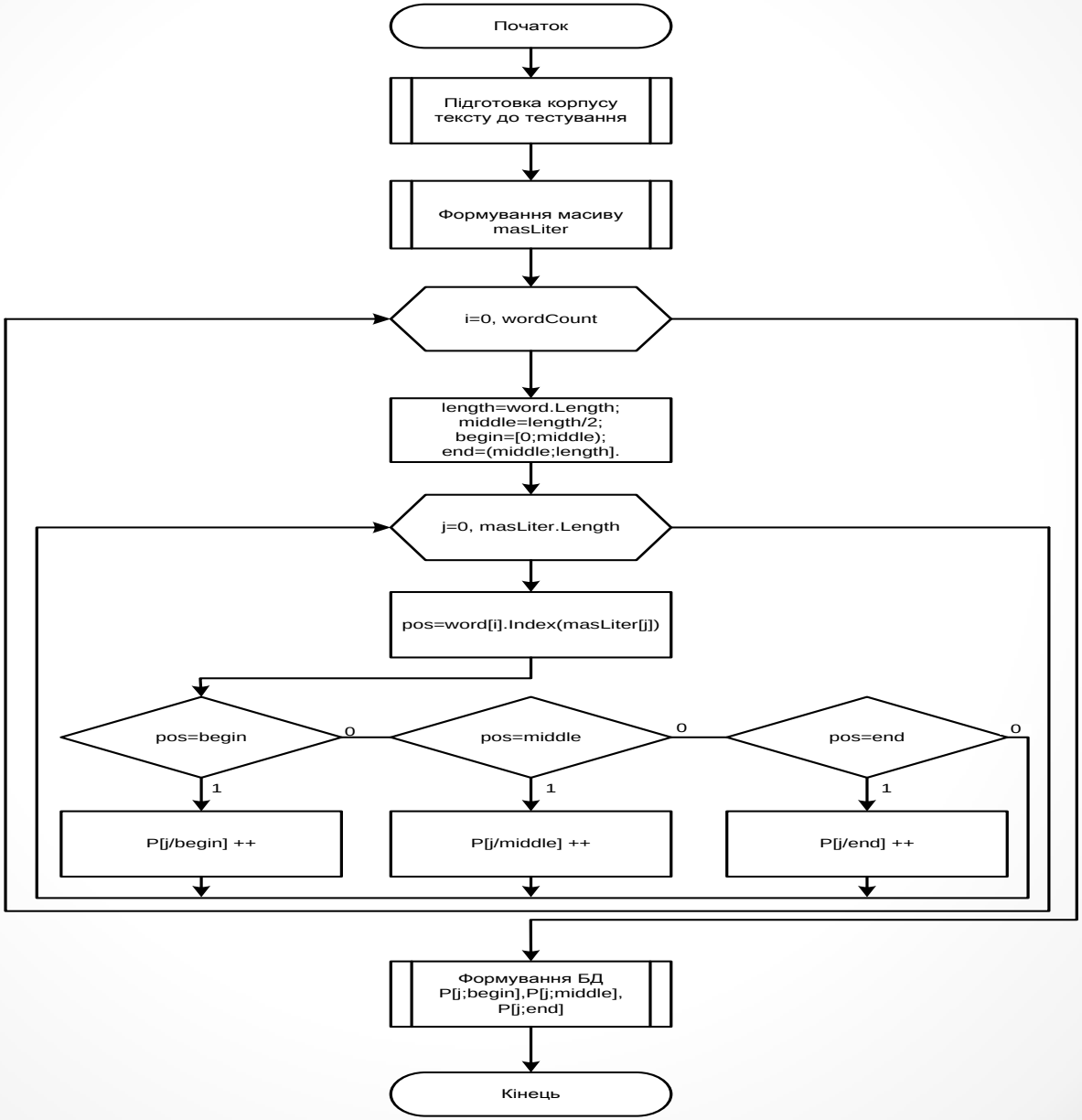
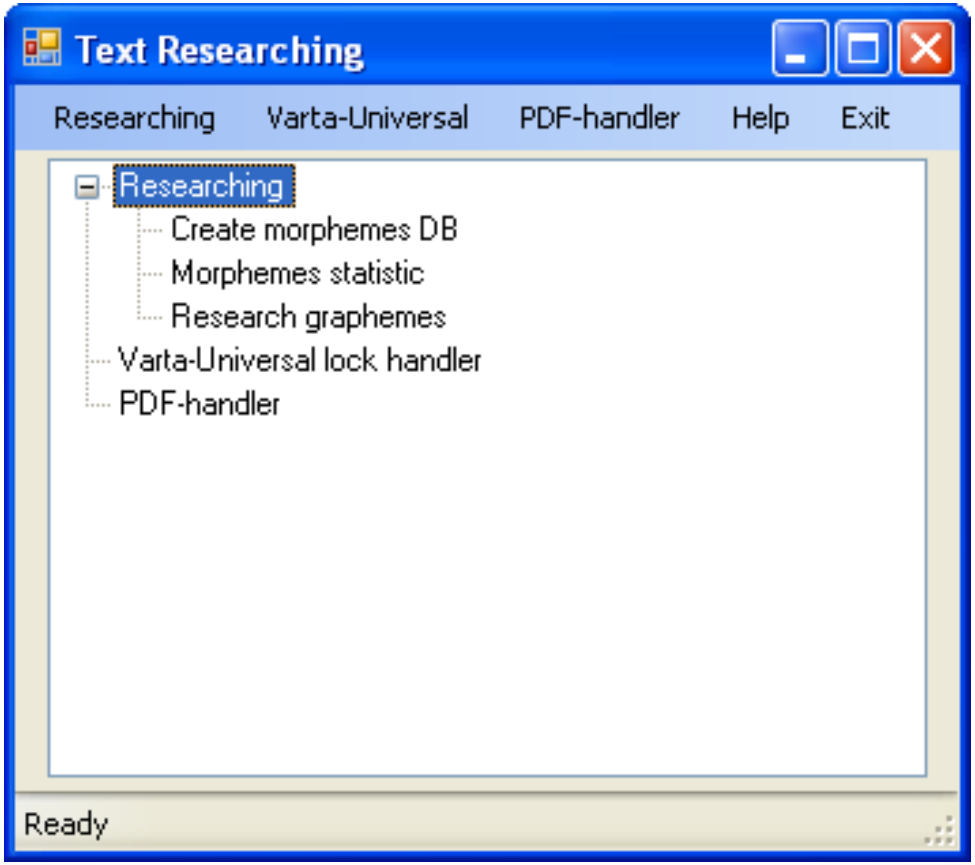


СХЕМА АЛГОРИТМУ ВИЗНАЧЕННЯ МІСЦЕЗНАХОДЖЕННЯ ЛІТЕРИ В СЛОВІ, ЯКА МАЄ НАДРЯДОВІ ТА ПІДРЯДКОВІ ОЗНАКИ

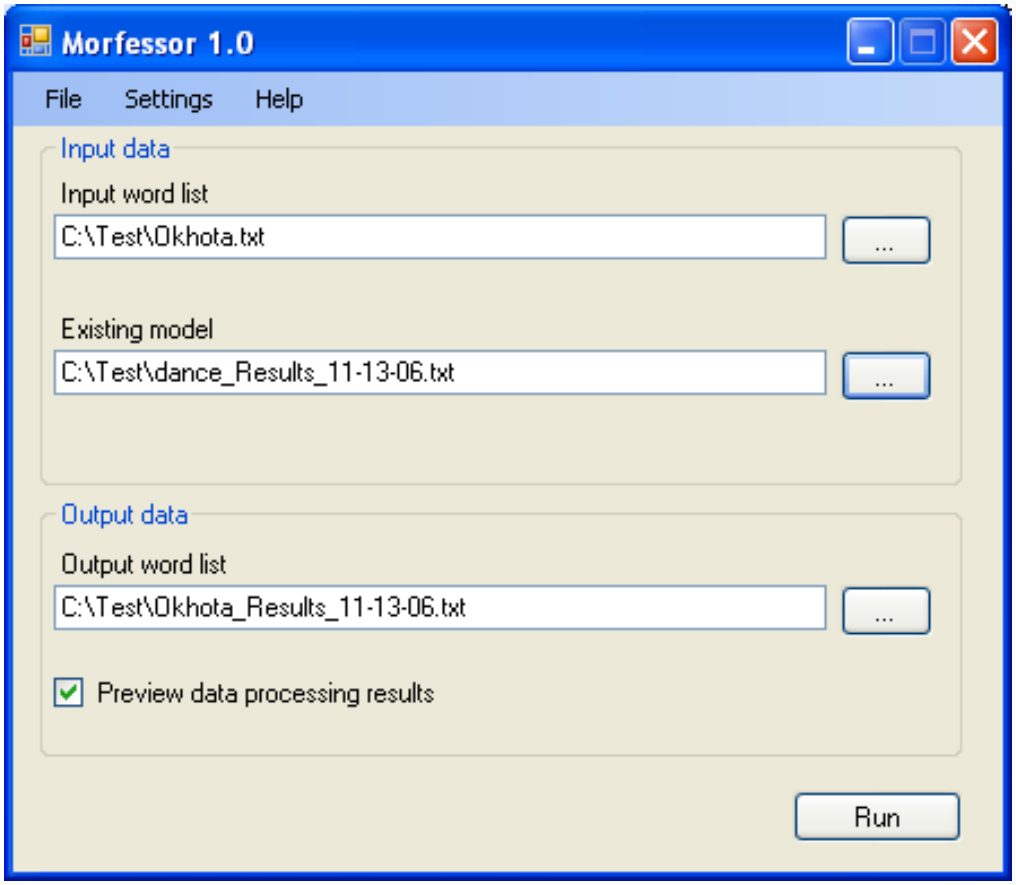


РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

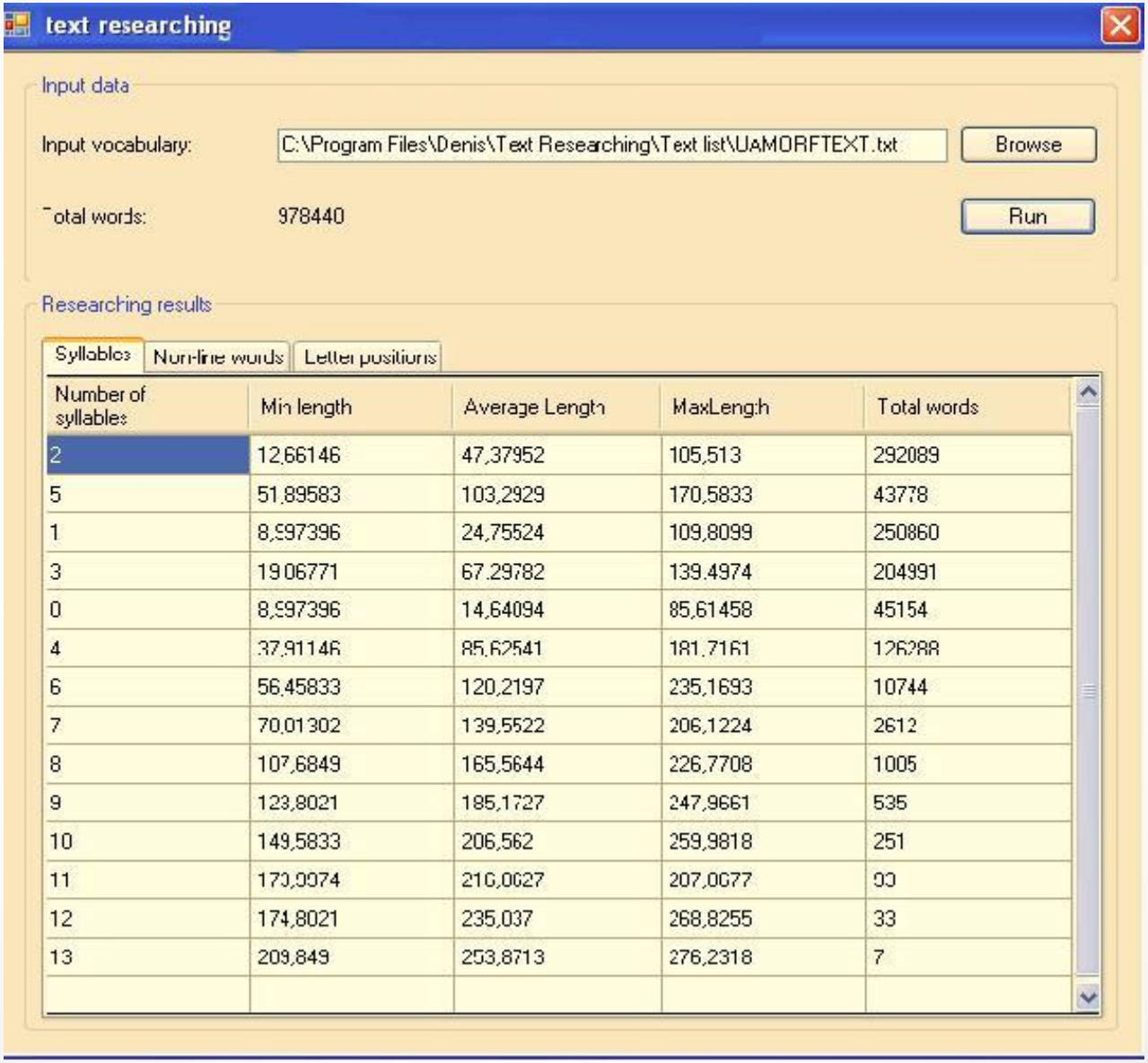
Головне вікно програми



Графічний інтерфейс програми “Морфесор 1.0”

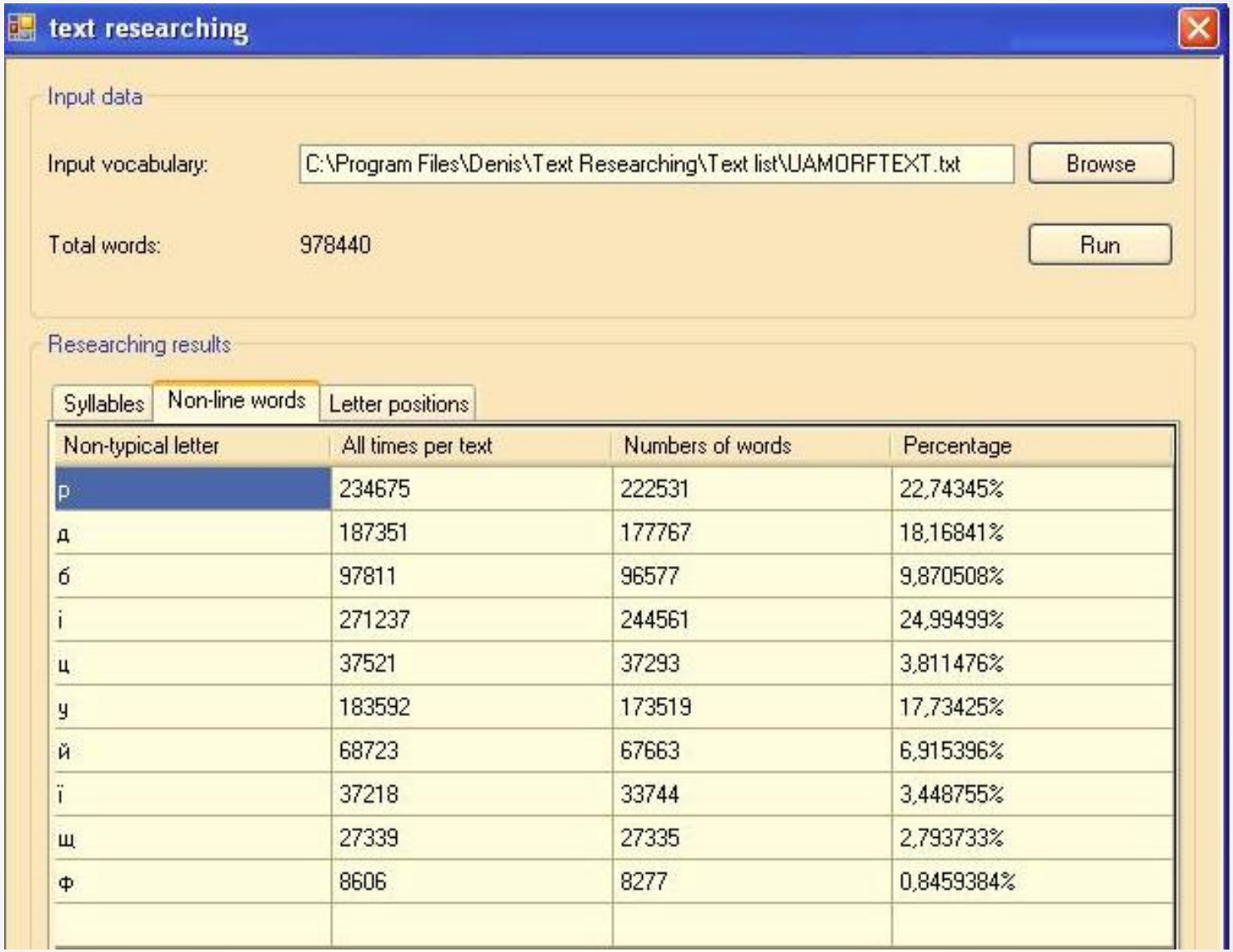


База даних залежності кількості складів слова від його довжини



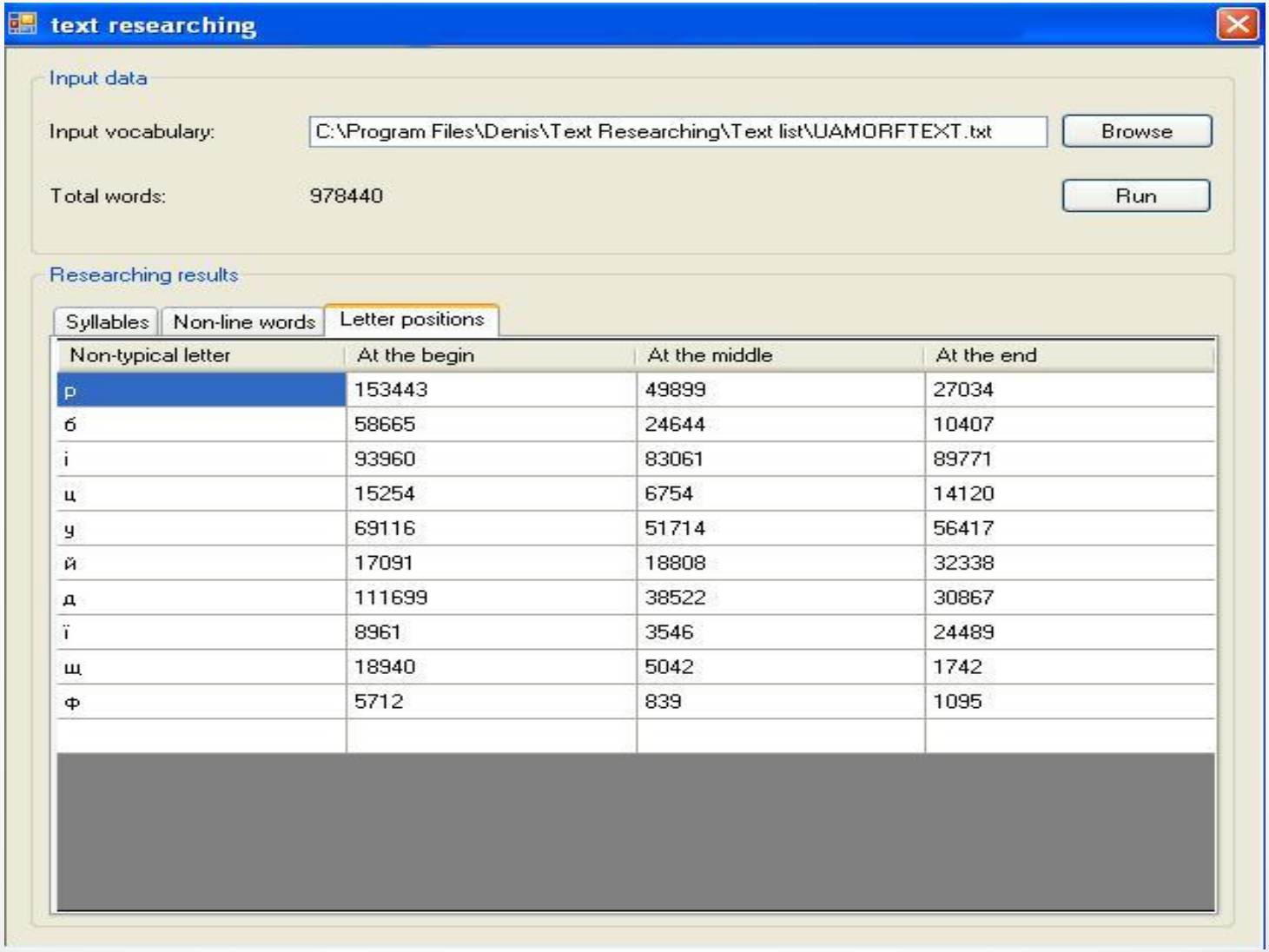
Number of syllables	Min length	Average Length	MaxLength	Total words
2	12,66146	47,37952	105,513	292089
5	51,89583	103,2929	170,5833	43778
1	8,997396	24,75524	109,8099	250860
3	19,06771	67,29782	139,4974	204991
0	8,997396	14,64094	85,61458	45154
4	37,91146	85,62541	181,7161	126288
6	56,45833	120,2197	235,1693	10744
7	70,01302	139,5522	206,1224	2612
8	107,6849	165,5644	226,7708	1005
9	123,8021	185,1727	247,9661	535
10	149,5833	206,562	259,9818	251
11	170,9074	216,0627	207,0677	99
12	174,8021	235,037	268,8255	33
13	209,849	253,8713	276,2318	7

База даних появи літер, що мають надрядкові та підрядкові ознаки



Non-typical letter	All times per text	Numbers of words	Percentage
р	234675	222531	22,74345%
д	187351	177767	18,16841%
б	97811	96577	9,870508%
і	271237	244561	24,99499%
ц	37521	37293	3,811476%
у	183592	173519	17,73425%
й	68723	67663	6,915396%
ї	37218	33744	3,448755%
щ	27339	27335	2,793733%
ф	8606	8277	0,8459384%

База даних знаходження літери на початку, середині чи кінця слова, яка має надрядкові чи підрядкові ознаки



Висновки

Запропонована нова інформаційна технологія введення і оброблення текстової інформації в автоматизованих інформаційно-пошукових системах, яка відрізняється від існуючих тим, що передбачає використання на етапі введення і розпізнавання не тільки графічного зображення тексту, а й низки мовних складових інформації (лексичної, морфологічної, синтаксичної та інш.), що містяться в цьому зображенні і дозволяють здійснити його часткове розуміння, а також оптимально розподіляє процес обробки документа між пристроєм введення і комп'ютерною системою.

Дякую за увагу!