Автоматичний морфологічний аналіз слів української мови

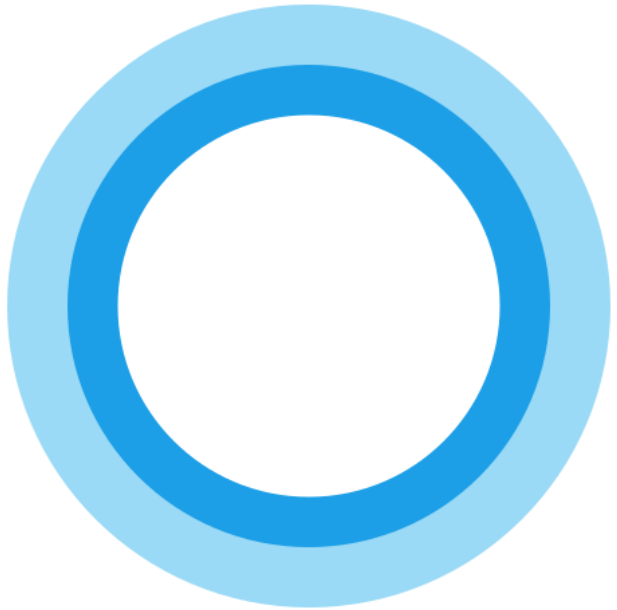
Виконав: студент

Групи ТІ-61м

Олійник Ю. Р.

Керівник: к.т.н., доцент

Стативка Ю. І.



YAHOO!®

Bing

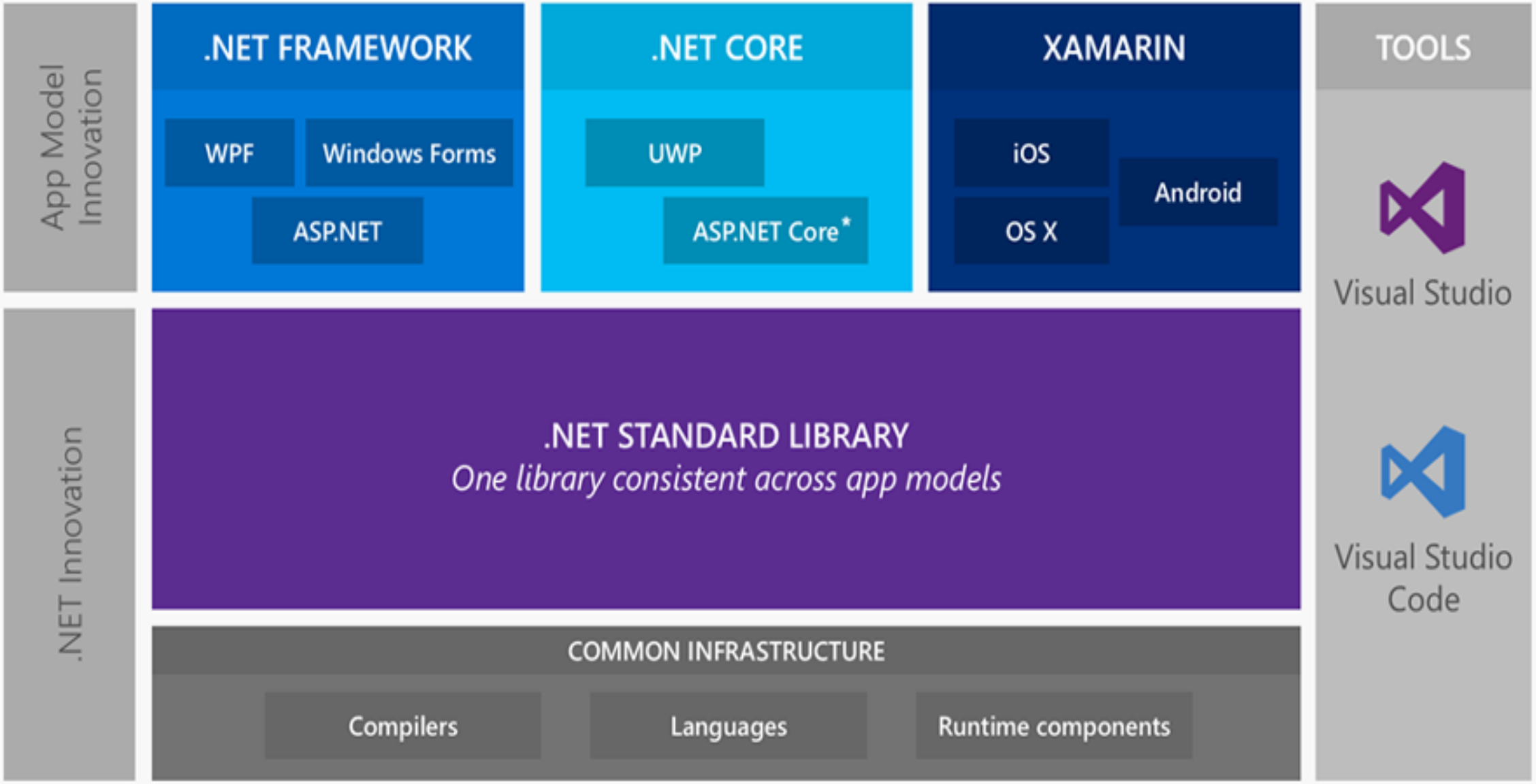
Мета дослідження

Полягає в розробці морфологічного аналізатора української мови для відкритого вільного доступу.

Завдання дослідження

- проаналізувати існуючі підходи та алгоритми автоматичного морфологічного аналізу;
- дослідити можливість застосування існуючих алгоритмів для української мови;
- забезпечити морфологічний аналіз слів, що відсутні у словнику;
- забезпечити архітектуру програмного забезпечення та розробити автоматичний морфологічний аналізатор української мови.

.NET future innovation

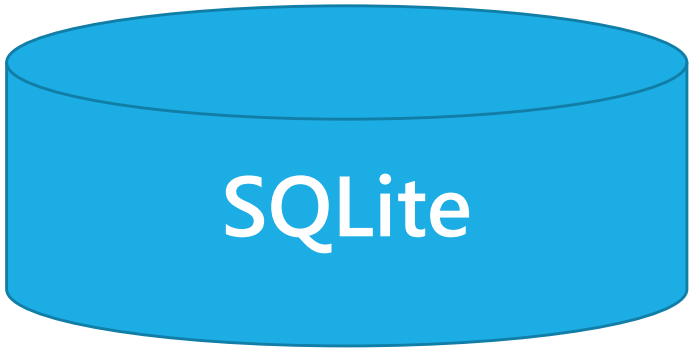


Entity Framework Core

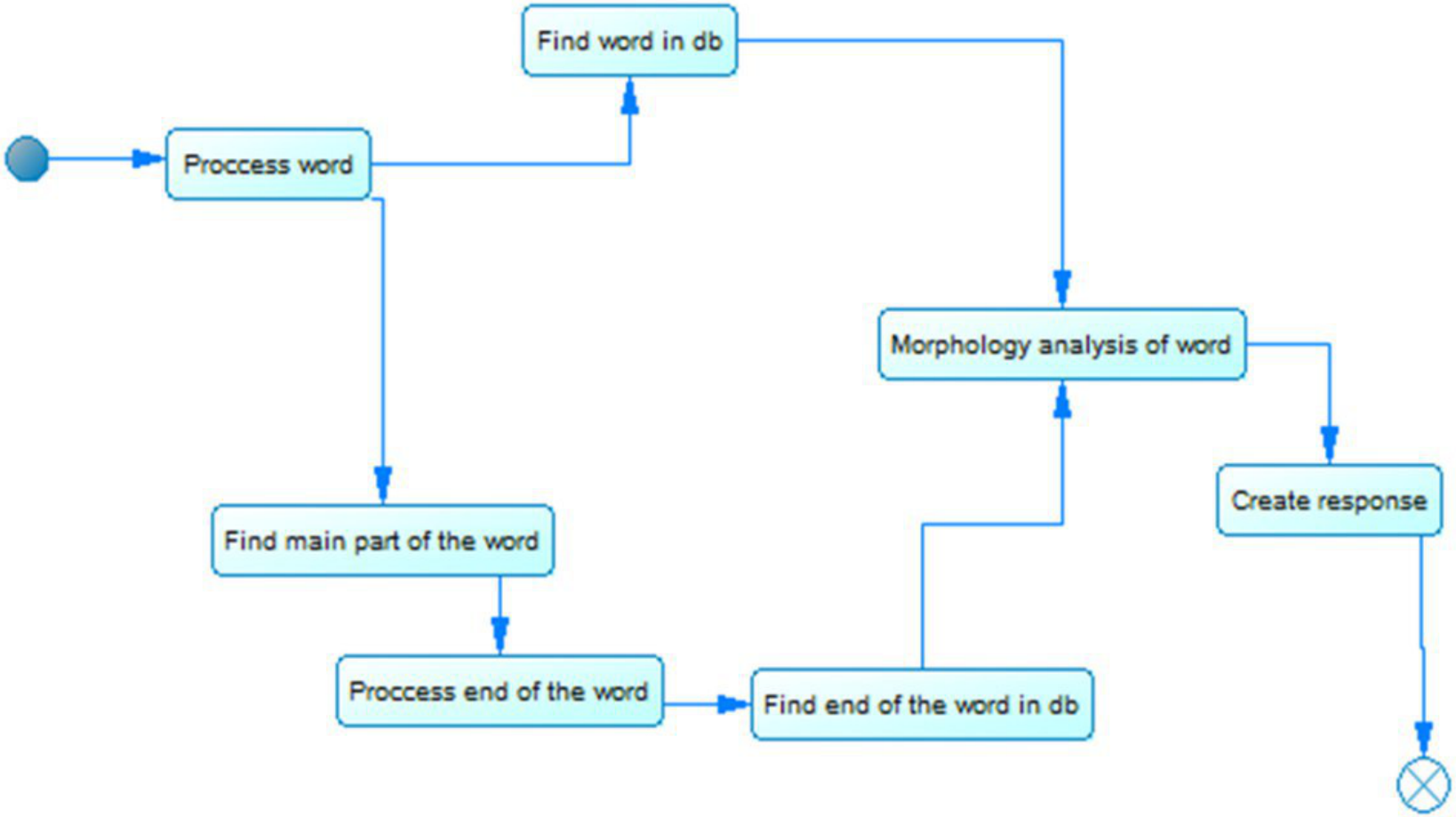
Application Layer

Analyser.Core Layer	Data Access Layer	Services Layer

Redis Cache



Алгоритм пошуку слова

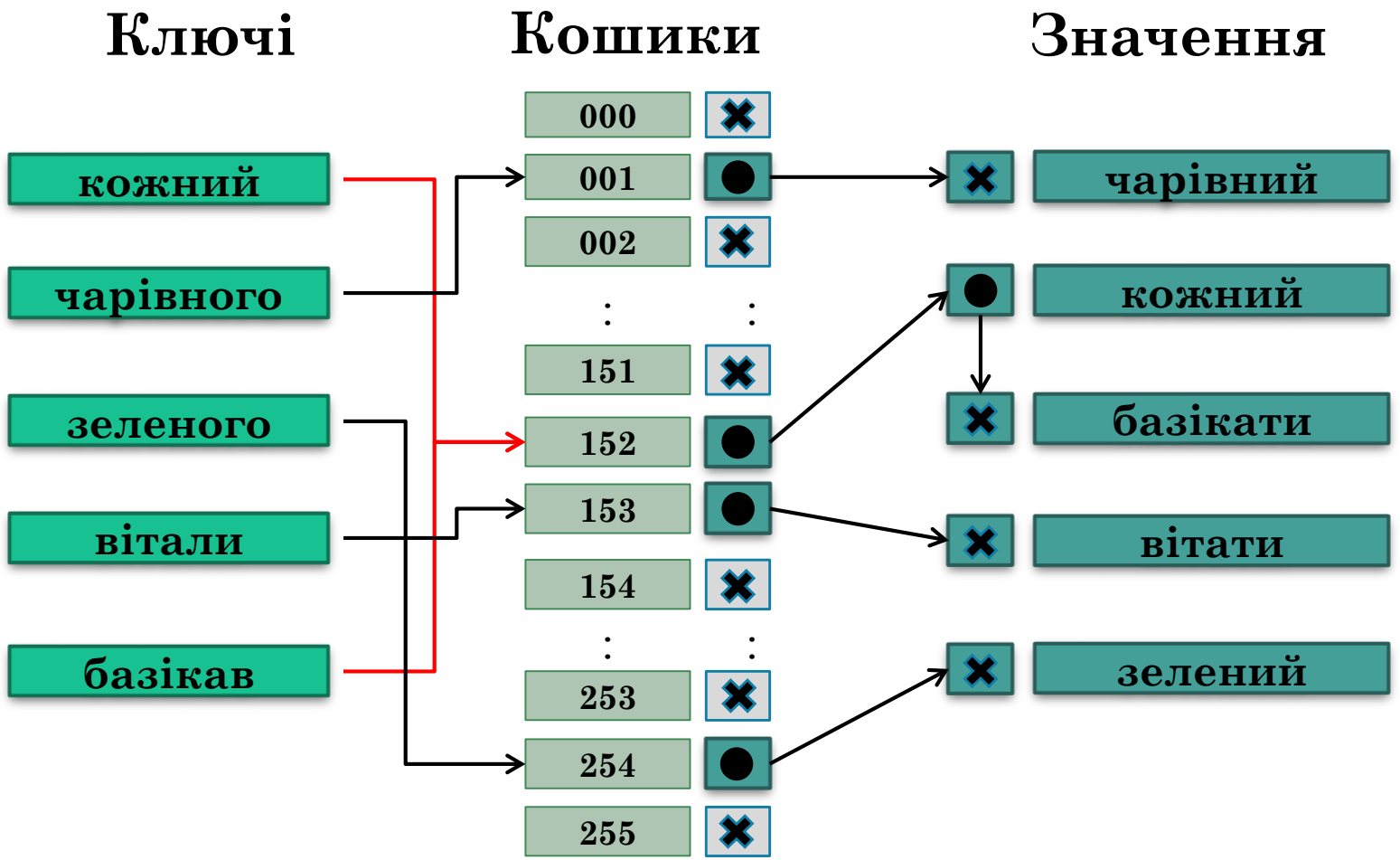


Пошук методом аналогії

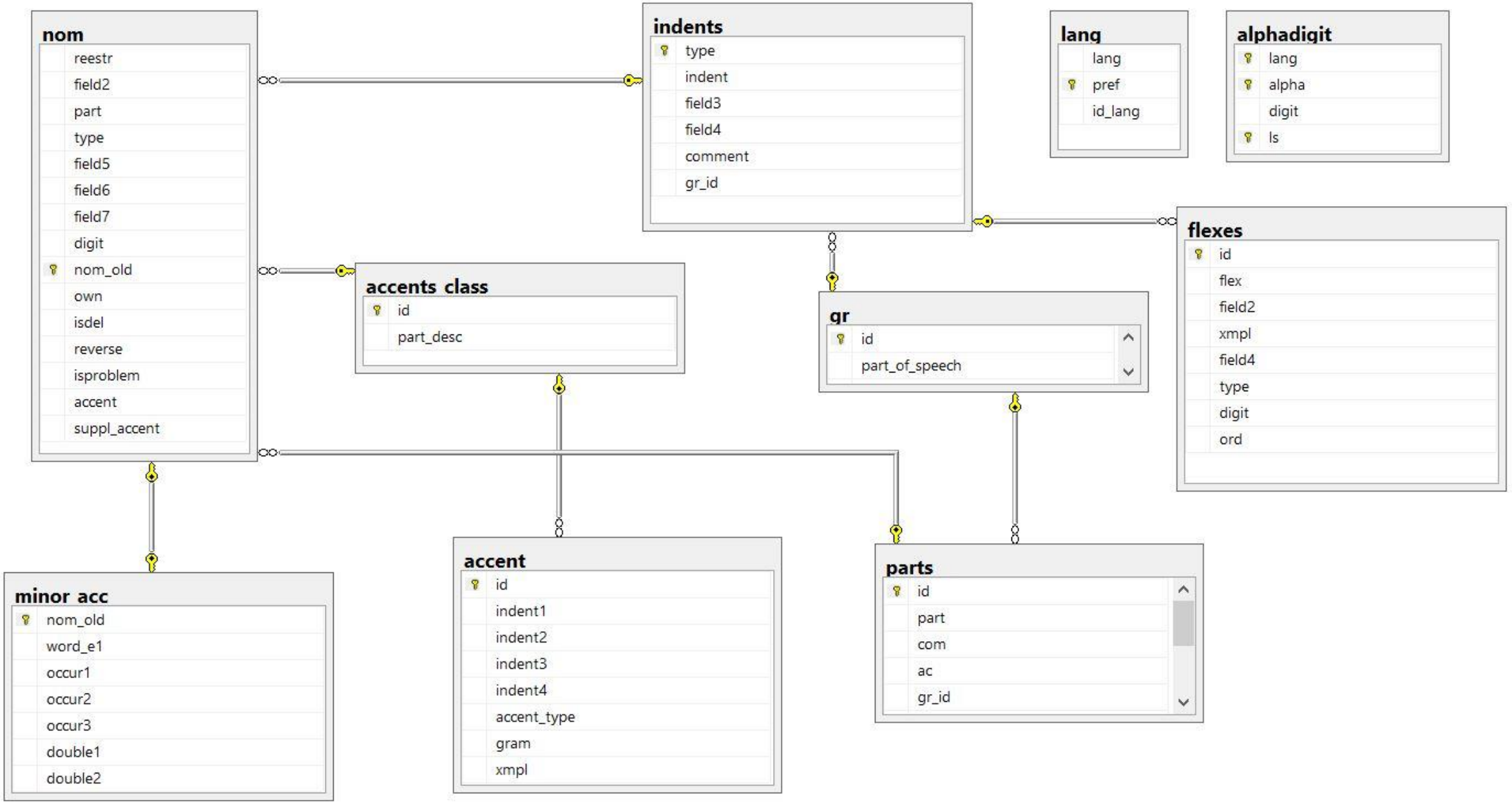
1. В аналізованого виділяється та відсікається квазізакінчення довжиною 4-1.
2. Визначаються слова, які можуть мати таке закінчення.
3. Визначаємо граматичну категорію та частину мови.
4. По граматичним категорія та частинам мов відбувається пошук квазізакінчень із словника.
5. До слова приписуються встановлені квазізакінчення.

```
{
  "data": {
    "inputWord": "пообговорювати",
    "words": [
      {
        "wordBase": "агітувати",
        "lemma": "пообговорю",
        "identifier": "дієслово недоконаного виду",
        "wordForms": [
          {
            "word": "пообговорювати",
            "field2": 1,
            "grId": 0,
            "grammarCategoryDesc": "Інфінітив, "
          },
          {
            "word": "пообговорюй",
            "field2": 2,
            "grId": 0,
            "grammarCategoryDesc": "Наказовий спосіб, Друга особа, Однина, "
          },
          {
            "word": "пообговорюймо",
            "field2": 3,
            "grId": 0,
            "grammarCategoryDesc": "Наказовий спосіб, Перша особа, Множина, "
          },
```

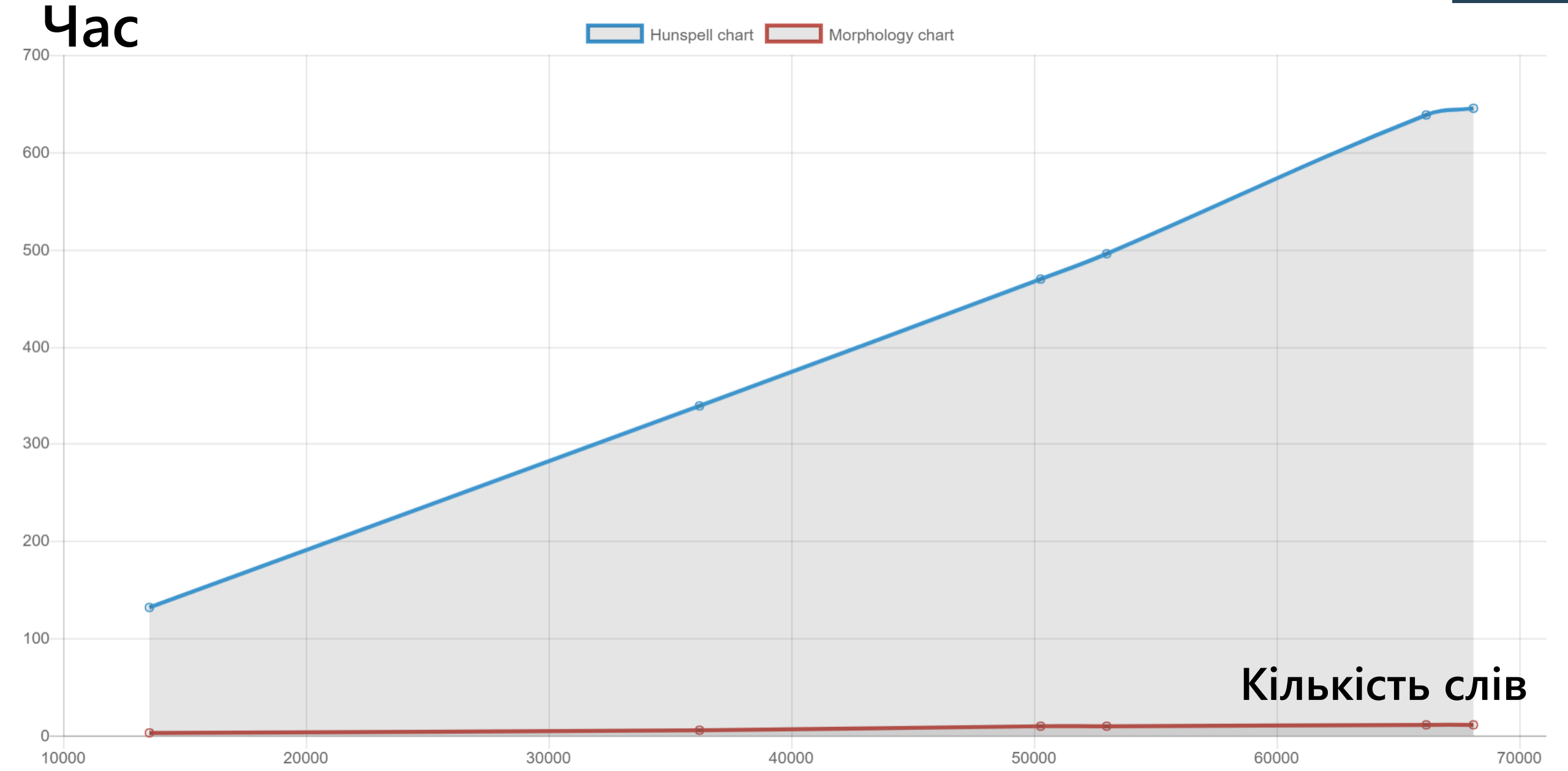
Структура хеш-таблиць по словоформам



Структура Бази Данних



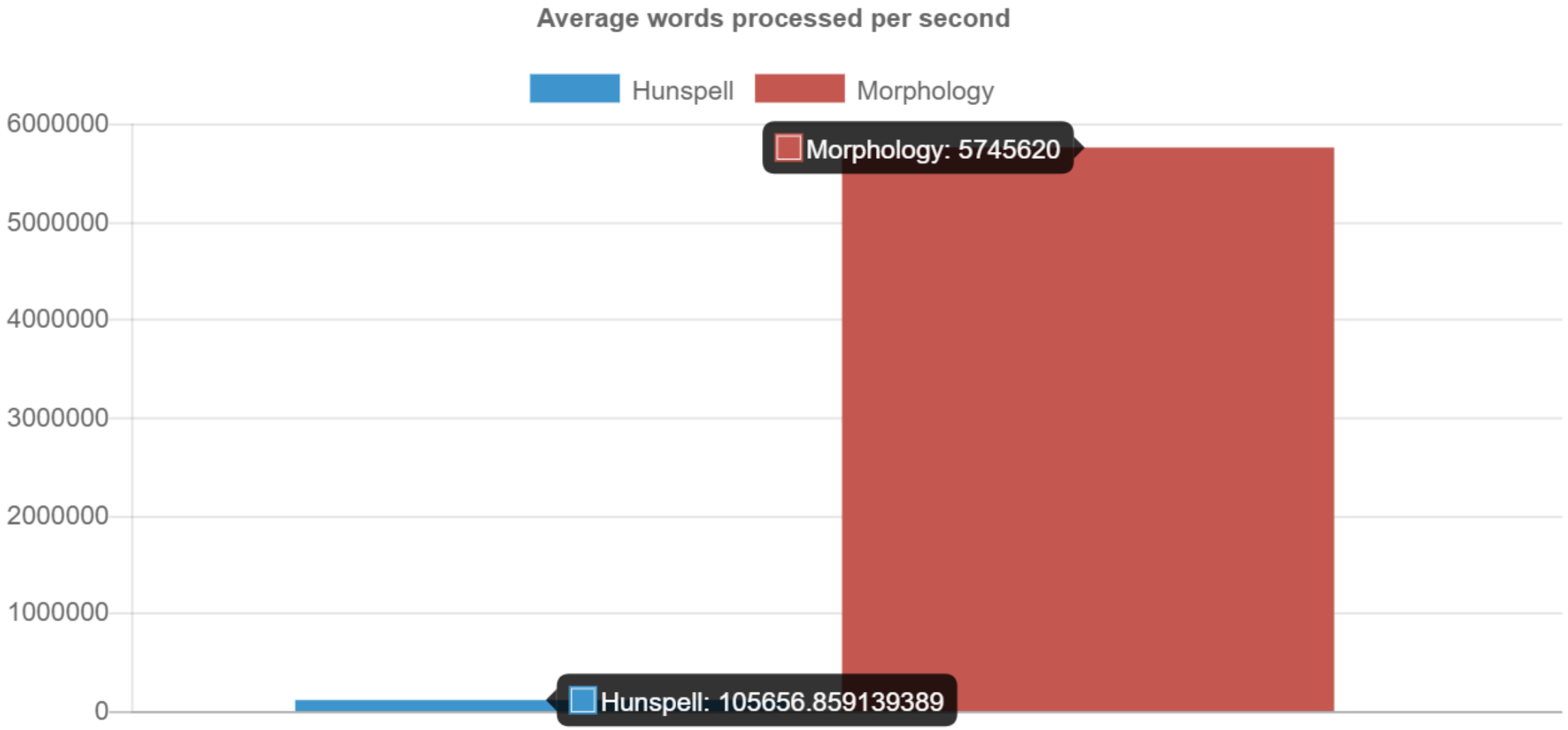
Результати порівняння



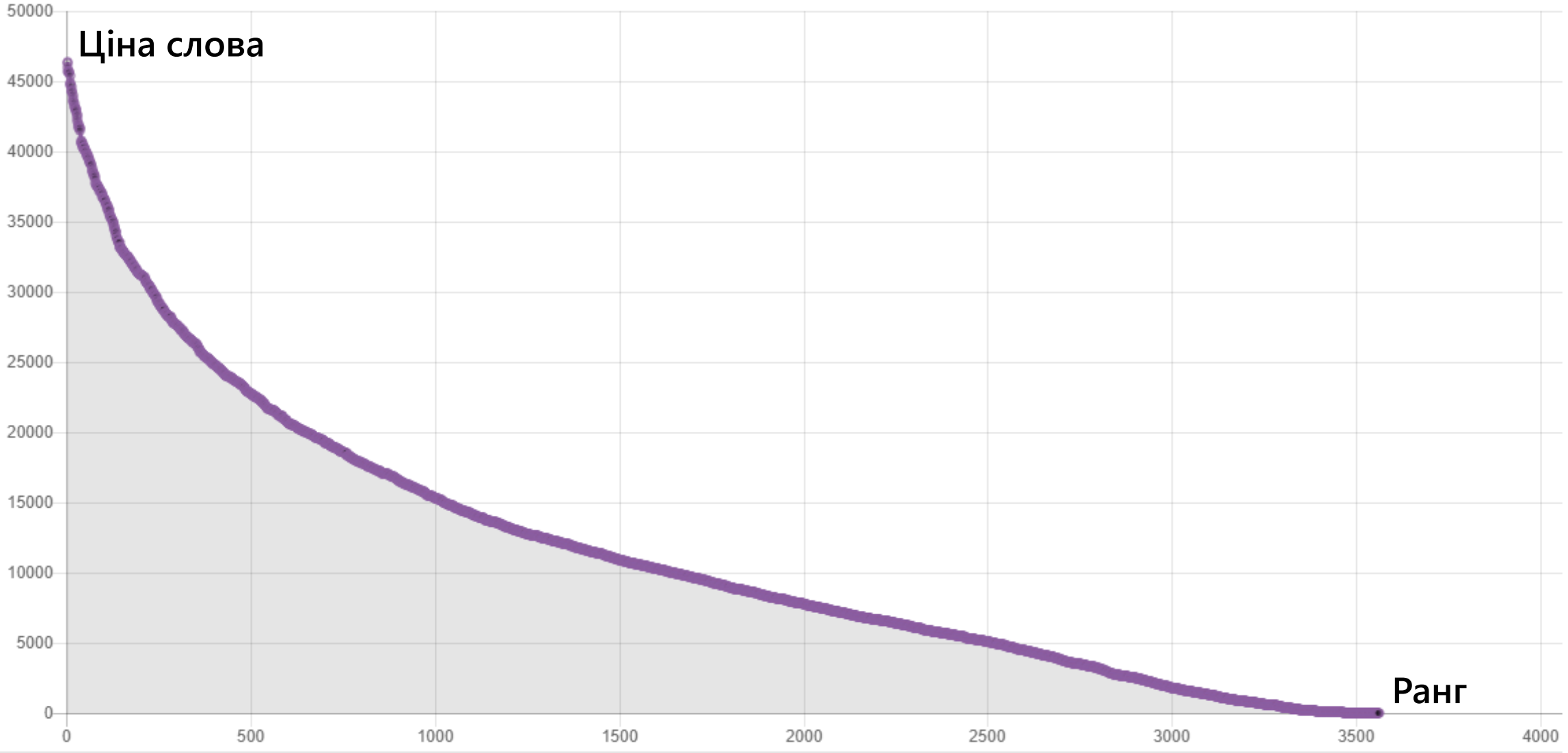
Таблиця порівняння швидкості

Кількість слів в тексті	Час затрачений для обробки тексту морфологічним аналізатором Hunspell (мс)	Час затрачений для обробки тексту морфологічним аналізатором Morphology (мс)
13580	132	3
36214	339	6
50255	470	9
53006	495	10
66125	638	11
68101	645	11

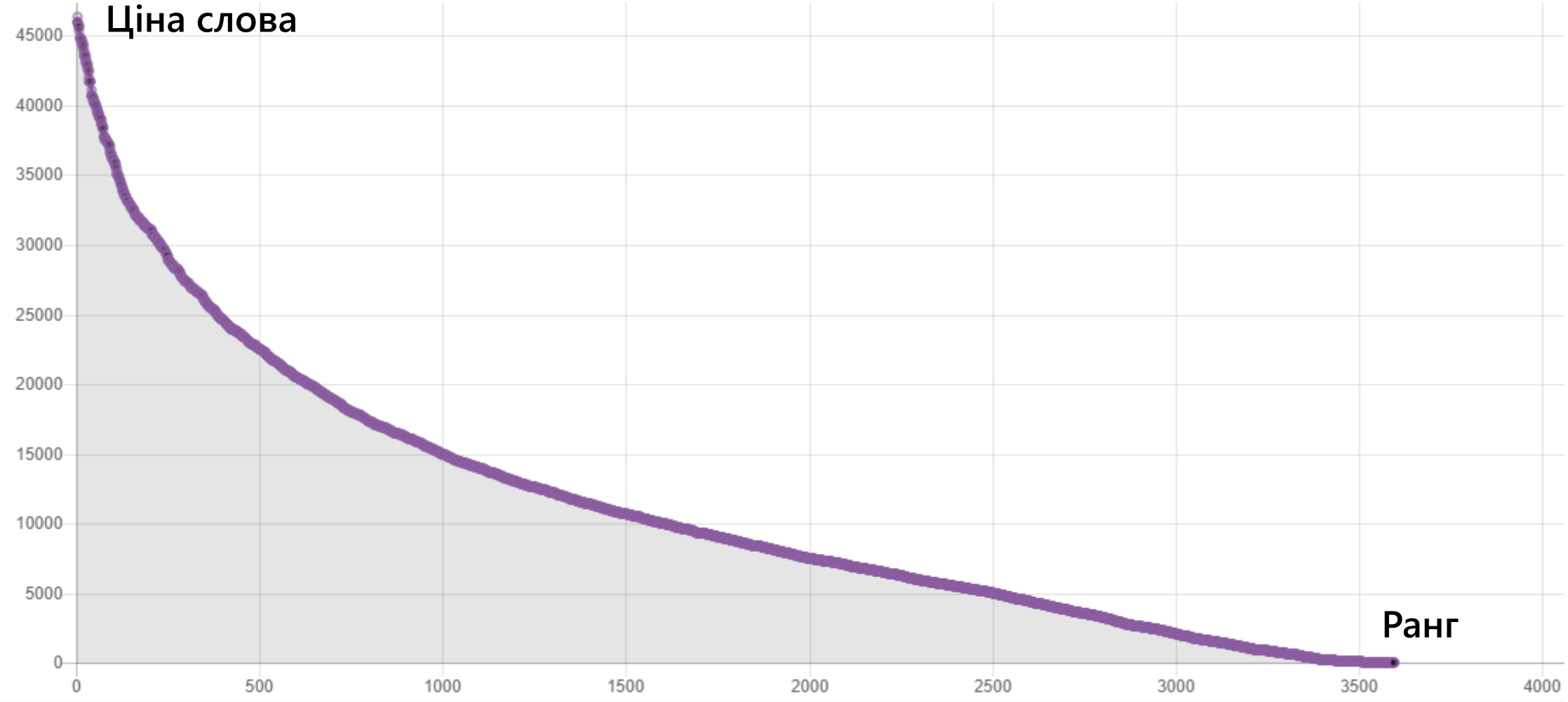
Кількість оброблених слів за секунду



Аналіз тексту методом LSA



Аналіз тексту методом LSA Hunspell



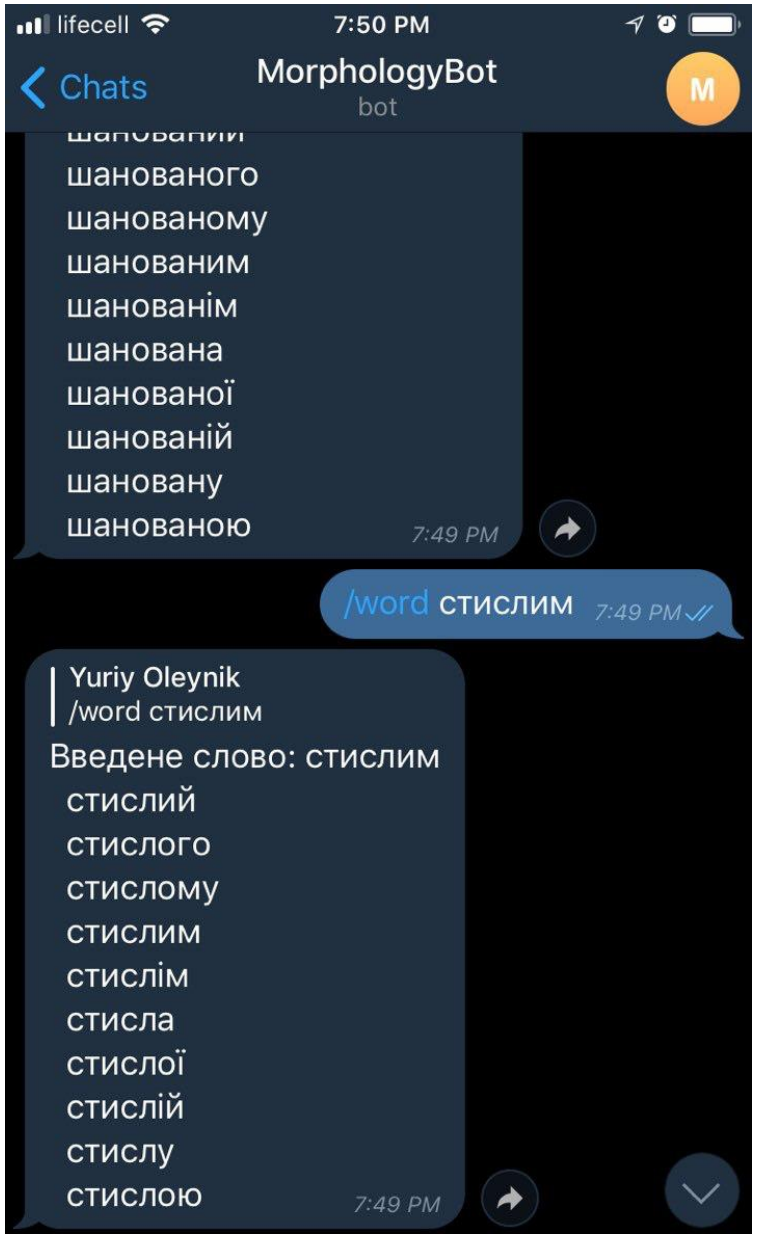
Аналіз тексту методом LSA (Таблиця)

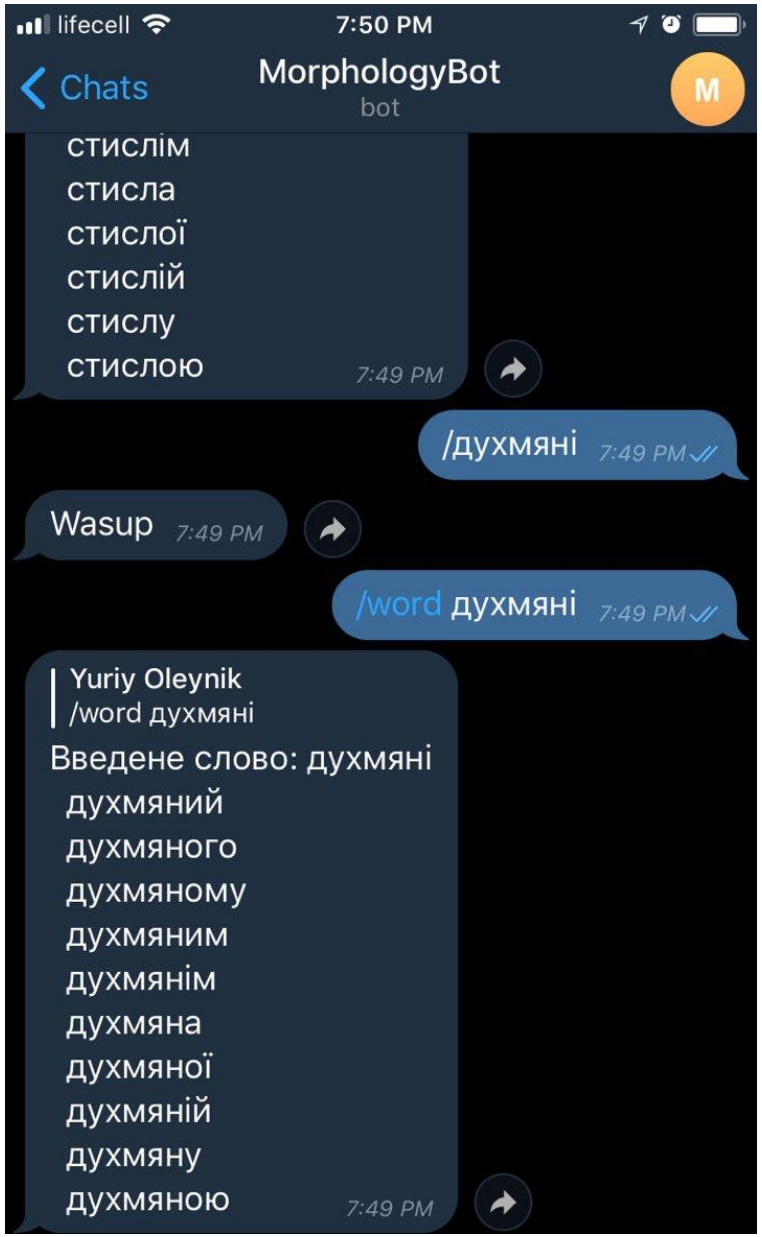
Порядок	Слово Morphology	Слово Hunspell	Ціна Morphology	Ціна Hunspell
1	кар’єр	кар’єр	46323	46323
2	розчервонілий	вигукуючи	46282	45976
3	звал	звал	46017	45975
4	вимкнутися	мою	45714	45912
5	процитувати	процитувати	45708	45708
6	обвинувачення	обвинувачення	45699	45699

Зазимую тут і залітую, В цій великій хаті не своїй, У кутку відтихну, відлютую, Намовчусь у темряві німій

Слово з речення	Лема від Hunspell	Лема від Morphology
Зазимую	зазимувати	зазимувати
тут	тут	тут
залітую	-	залітувати
цій	цій	цій
великій	великий	великий
хаті	хата	хата
своїй	своїй	свій
кутку	куток	куток
відтихну	-	відтихнути
відлютую	-	відлютувати
Намовчусь	-	Намовчатися
темряві	темрява	темрява
німій	німіти, німій	німіти, німій

Приклад роботи телеграм боту через Api





Висновок:

В результаті роботи було:

- Проаналізовано існуючі методи автоматичного морфологічного аналізу української мови;
- Під час дослідження було виявлено, що частота розпізнавання слів значно покращилась по відношенню до існуючого аналогу Hunspell. Швидкість збільшилась приблизно в 50 разів.
- Розроблено відкритий автоматичний морфологічний аналізатор української мови.