

# NLP: створення парсера для аналізування українськомовних віршів

**Доповідач:** Швець Дмитро Віталійович, ШЗ-3

**Керівник:** Смиш Олег Русланович, ст. викладач, PhD

# Мета та завдання курсової роботи

**Мета** курсової роботи: автоматизувати процес аналізування українськомовних віршів.

**Завдання** курсової роботи:

- дослідити наявні інструменти для оброблення українськомовного тексту;
- оцінити можливості аналізування українськомовних віршів великими мовними моделями;
- розробити парсер для визначення метра, рими, способів римування та художніх засобів;
- визначити точність парсера шляхом тестування на підготовленій збірці тестів;
- застосувати парсер для аналізування збірок віршів і візуалізувати отримані результати.

# Об'єкт і предмет дослідження

**Об'єкт** дослідження — автоматизація оброблення текстів, написаних українською мовою.

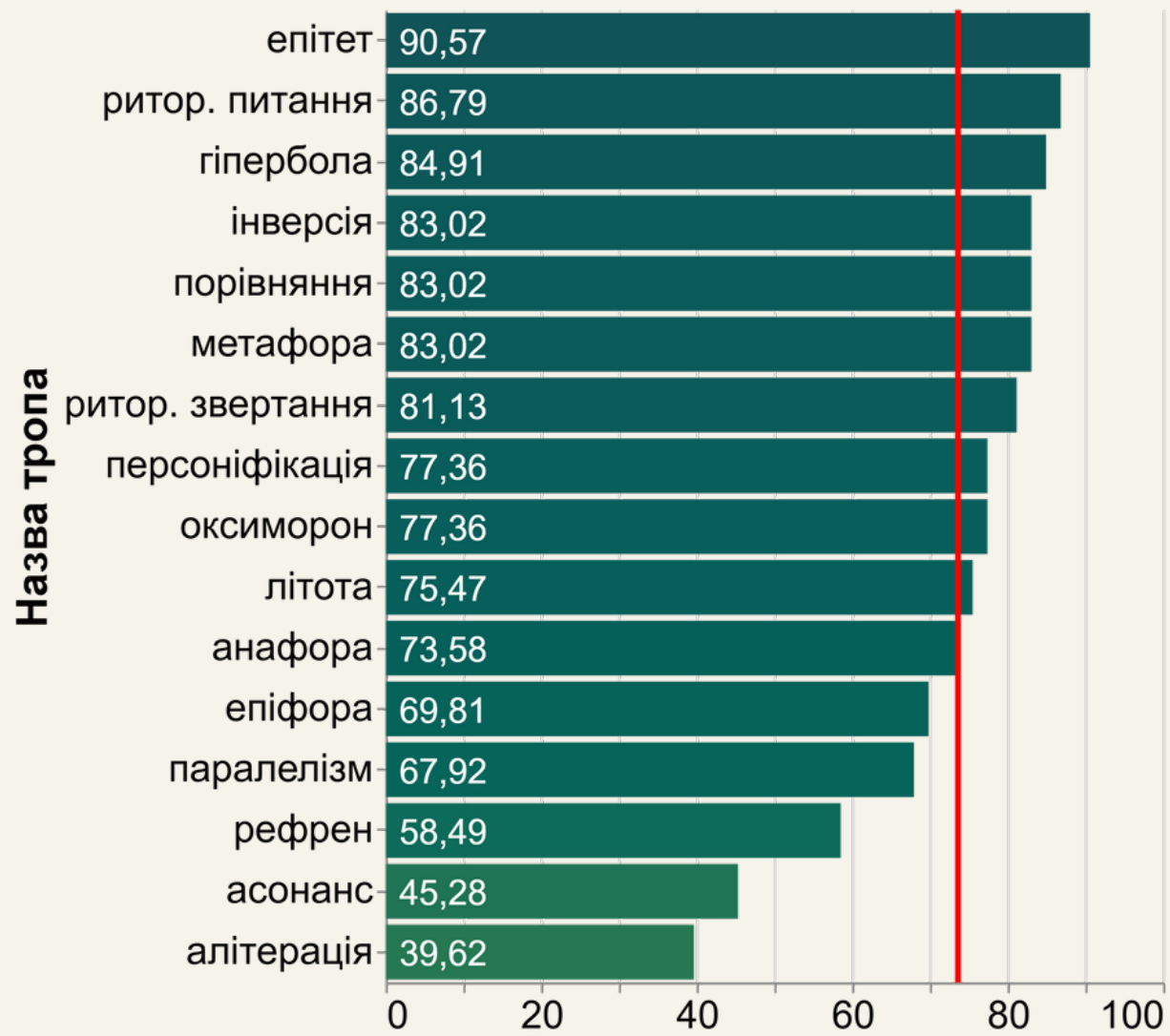
**Предмет** дослідження — використання методів обробки природної української мови в контексті різнобічного аналізування українськомовних віршів.

# Наукова новизна

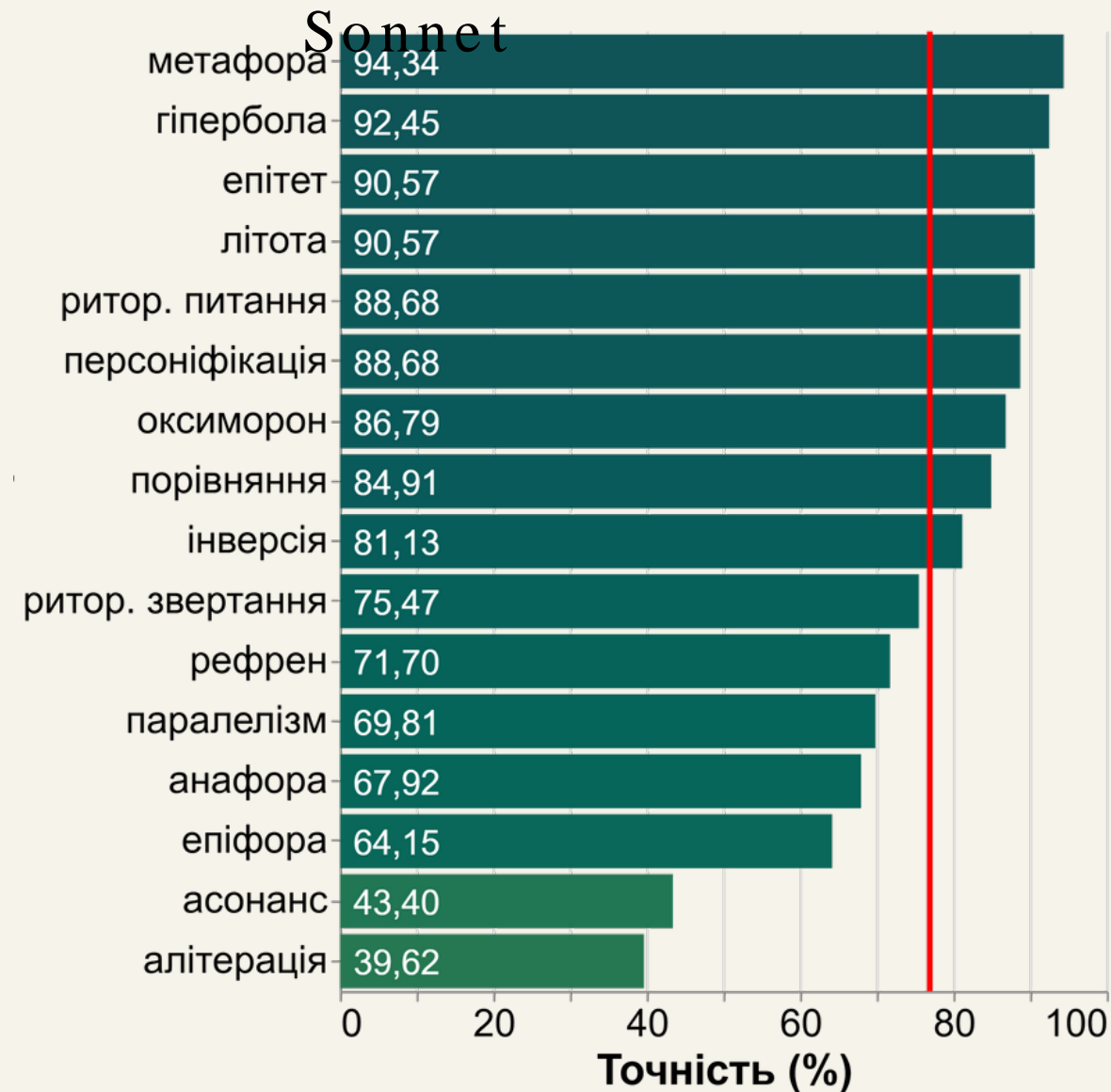
- Уперше створено **метод для комплексного аналізування віршового розміру, рими та способів римування в українськомовних віршах** засобами обробки природної мови, на основі побудови схем наголошування та фонетичного зіставлення рядків.
- Уперше запропоновано **метод трирівневого виявлення 16 художніх засобів (тропів) в українськомовних поетичних текстах**, що ґрунтується на обробці природної мови: поєднання підходів на основі правил і машинного навчання.

# Великі мовні моделі (LLM)

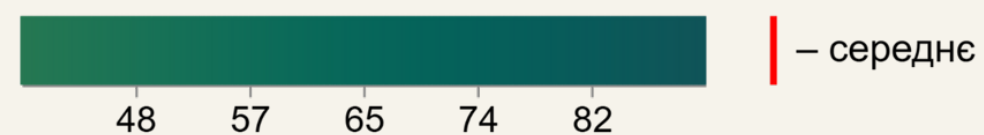
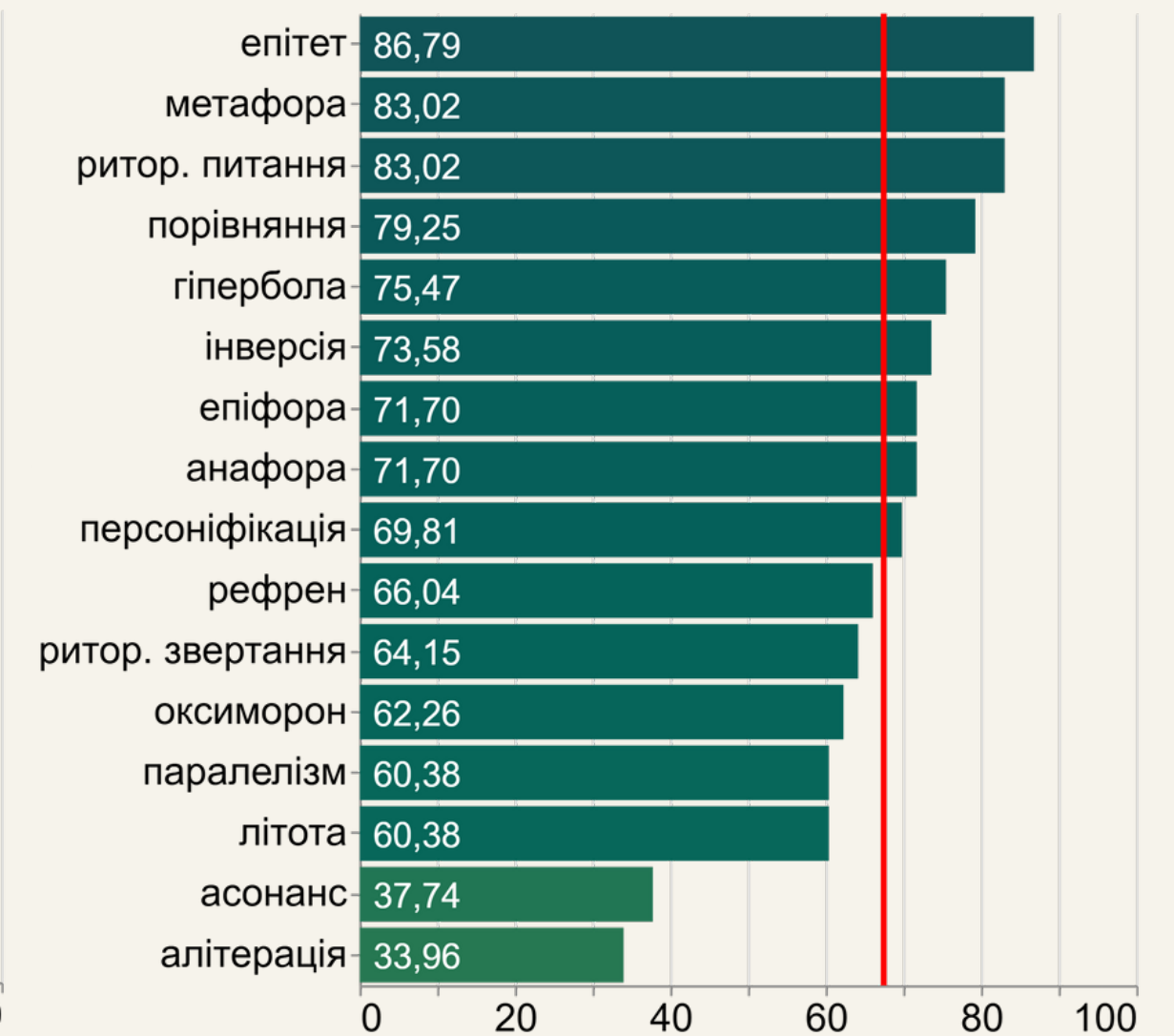
## ChatGPT-4



## Claude 3.5



## Gemini 2.0 Flash



# Мовні аналізатори

**UDPipe, Stanza, spaCy** — інструменти для токенізації, лематизації, розмічування частин мови та визначення морфосинтаксичних зв'язків у текстах.

№	слово	лема	ЧМ	морфологічні риси	голова	зв'язок
1	Вітаю	вітати	VERB	Aspect=Imp Mood=Ind Number=Sing Person=1 Tense=Pres VerbForm=Fin	0	root
2	!	!	PUNCT		1	punct
№	слово	лема	ЧМ	морфологічні риси	голова	зв'язок
1	Мене	я	PRON	Animacy=Anim Case=Acc Number=Sing Person=1 PronType=Prs	2	obj
2	звати	звати	VERB	Aspect=Imp VerbForm=Inf	0	root
3	Дмитро	Дмитро	PROPN	Animacy=Anim Case=Nom Gender=Masc NameType=Giv Number=Sing	2	nsubj
4	.	.	PUNCT		2	punct

# Визначення метра (віршового розміру)

Віршовий розмір складається з **назви** стопи та її **кількості**: чотиристопний ямб, тристопний дактиль тощо.

«-» — наголошений склад, «U» — ненаголошений склад

Стопа	Схема	Наголошено
Хорей	- U	перший склад із двох
Ямб	U -	другий склад із двох
Дактиль	- U U	перший склад із трьох
Амфібрахій	U - U	другий склад із трьох
Анапест	U U -	третій склад із трьох

# Визначення метра

$\xrightarrow{1}$  При́смеркові су́тінки опа́ли,  
 со́нну зе́млю й ду́шу оплели́.  
 Самоти́ згорьо́вані хорáли  
 геть мені́ дорóгу́ замели́.  
*В. Стус*

$\xrightarrow{2}$

$\xrightarrow{3}$

$[[\widehat{1}, 0, 0, 0, \widehat{1}, 0, 0, 0, \widehat{1}, 0],$   
 $[1, 0, 1, 0, 1, 0, 0, 0, 1],$   
 $[0, 0, 1, 0, 1, 0, 0, 0, 1, 0],$   
 $[0, 0, 1, 0, 1, 1, 0, 0, 1]]$

$[(5, \text{"ch"}),$   
 $(5, \text{"ch"}),$   
 $(5, \text{"ch"}),$   
 $(0, \text{None})]$

$\xrightarrow{4}$  Найчастіше:  
 "ch" — (1, 0)

$\xrightarrow{5}$  [3]  $[\widehat{0}, 0, \widehat{1}, 0, \widehat{1}, \widehat{1}, 0, 0, \widehat{1}]$

$\xrightarrow{6}$   $[\widehat{0}, 0, \widehat{1}, 0, \widehat{1}, 0, 0, 0, \widehat{1}]$

$\xrightarrow{7}$

$[(5, \text{"ch"}),$   
 $(5, \text{"ch"}),$   
 $(5, \text{"ch"}),$   
 $(5, \text{"ch"})]$

# Визначення рими

(a, b, c, d)

**a** – від’ємний індекс останнього наголошеного складу

**b** – голосний звук останнього наголошеного складу

**c** – решта голосних звуків після останнього наголошеного

**d** – усі голосні звуки рядка

$\xrightarrow{1}$  При́смеркові су́тінки опа́ли,  
со́нну зе́млю й ду́шу оплели́.  
Самоти́ згорьо́вані хора́ли  
геть мені́ доро́гу замели́.

$\xrightarrow{2}$  [(-2, 'а', "и", "иеоіуіоаи"),  
(-1, 'и', "", "оуеуууоеи"),  
(-2, 'а', "и", "аоіооаіоаи"),  
(-1, 'и', "", "ееіооуаеи")]

$\xrightarrow{3}$  [(0, 2),  
(1, 3)]

# Виявлення тропів (художніх засобів)

Рівень	Тропи	Підходи та засоби виявлення
Фонологічний	алітерація, анафора, асонанс, епіфора, рефрен, риторичне питання	rule-based approaches: кількісний аналіз, регулярні вирази
Морфосинтаксичний	епітет, інверсія, паралелізм, порівняння, риторичне звертання	rule-based approaches + machine learning approaches: мовні аналізатори UDPipe, Stanza, spaCy
Семантичний	гіпербола, літота, метафора, оксиморон, персоніфікація	rule-based approaches + machine learning approach: Claude LLM

## Тропи

## Метр і рима

### Текст вірша:

Задивляюсь у твої зіниці  
Голубі й тривожні, ніби рань,  
Крешуть з них червоні блискавиці  
Революцій, бунтів і повстань.

### Оберіть тропи:

- Обрати всі
- інверсія
- епітет
- паралелізм
- порівняння
- риторичне звертання
- алітерація
- асонанс
- рефрен
- анафора
- епіфора
- риторичне питання
- метафора
- персоніфікація
- гіпербола
- літота
- оксиморон

Знайти тропи

### Тропи:

інверсія епітет паралелізм

порівняння риторичне звертання

алітерація асонанс рефрен

анафора епіфора

риторичне питання метафора

персоніфікація гіпербола

літота оксиморон

### Маршрути:

- «/tropes», метод POST — виявлення тропів;
- «/metre-rhyme», метод POST — визначення метра, рими та способу римування.

## Тропи

## Метр і рима

### Текст вірша:

Реве, люте Візантія, A  
Руками берег достає, B  
Достали, зикнула, встає... B  
І на списках в крові німіє. A

### Знайдені види римування:

- кільцеве (сині)

Аналізувати риму та метр

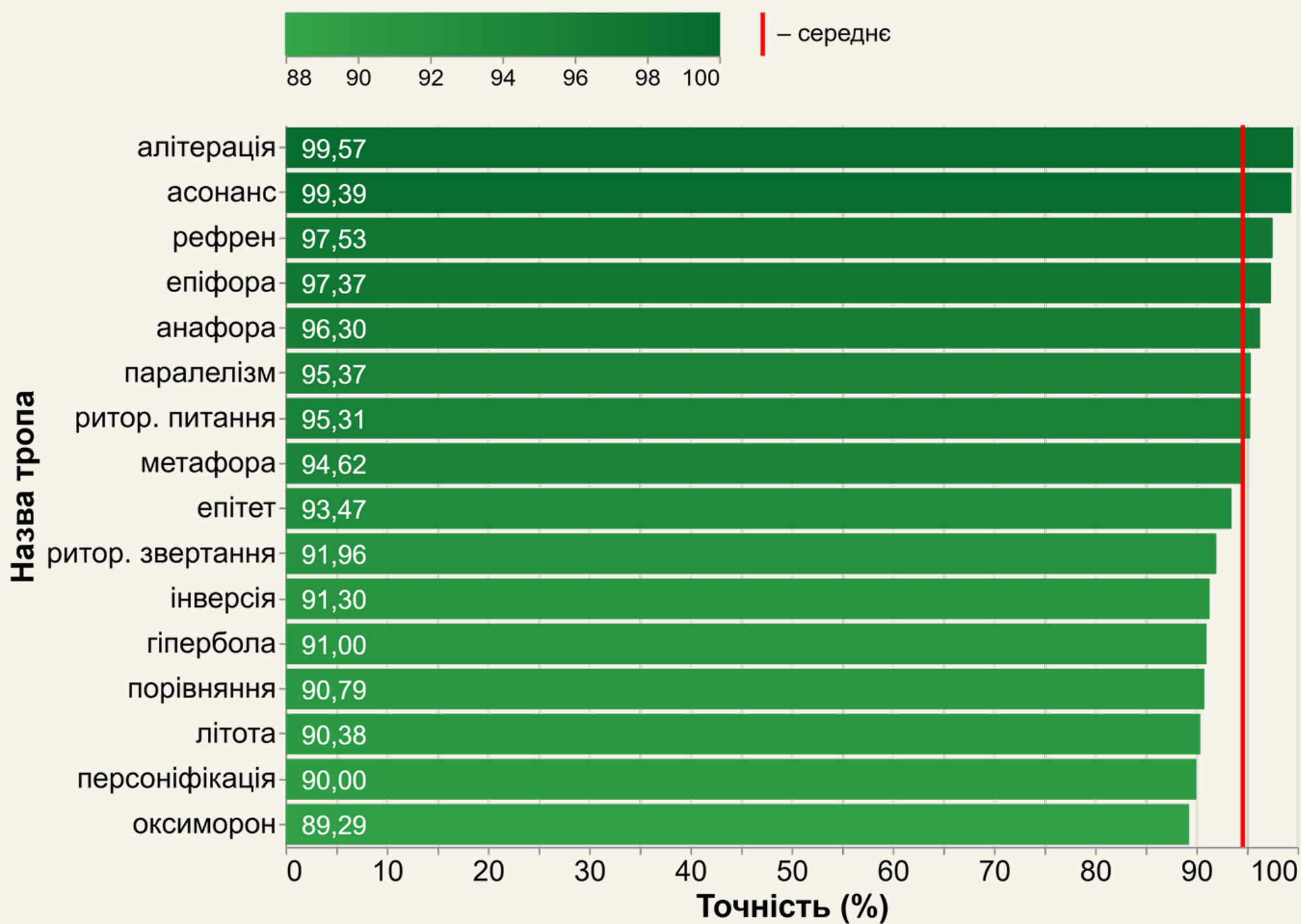
### Схема та метри:

U -|U -|U U|U -|U 4-ст. ямб  
U -|U -|U U|U - 4-ст. ямб  
U -|U -|U U|U - 4-ст. ямб  
U U|U -|U U|U -|U 4-ст. ямб

### Основний віршовий розмір:

4-стопний ямб

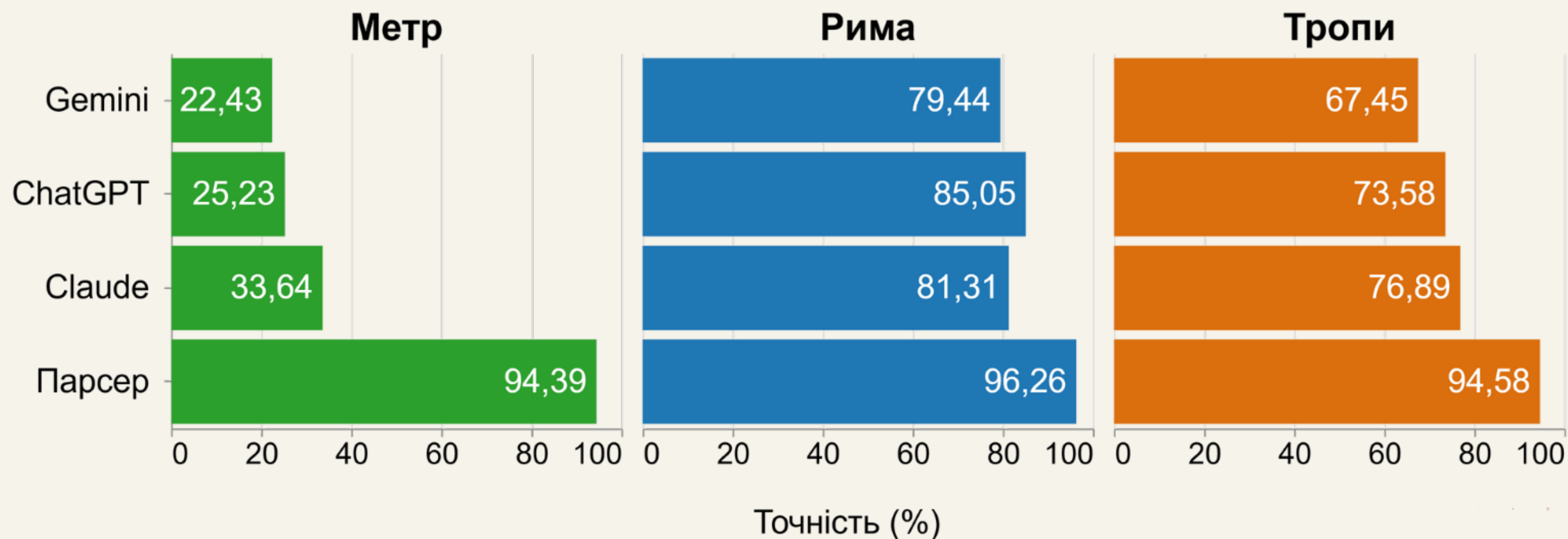
# Оцінювання точності парсера



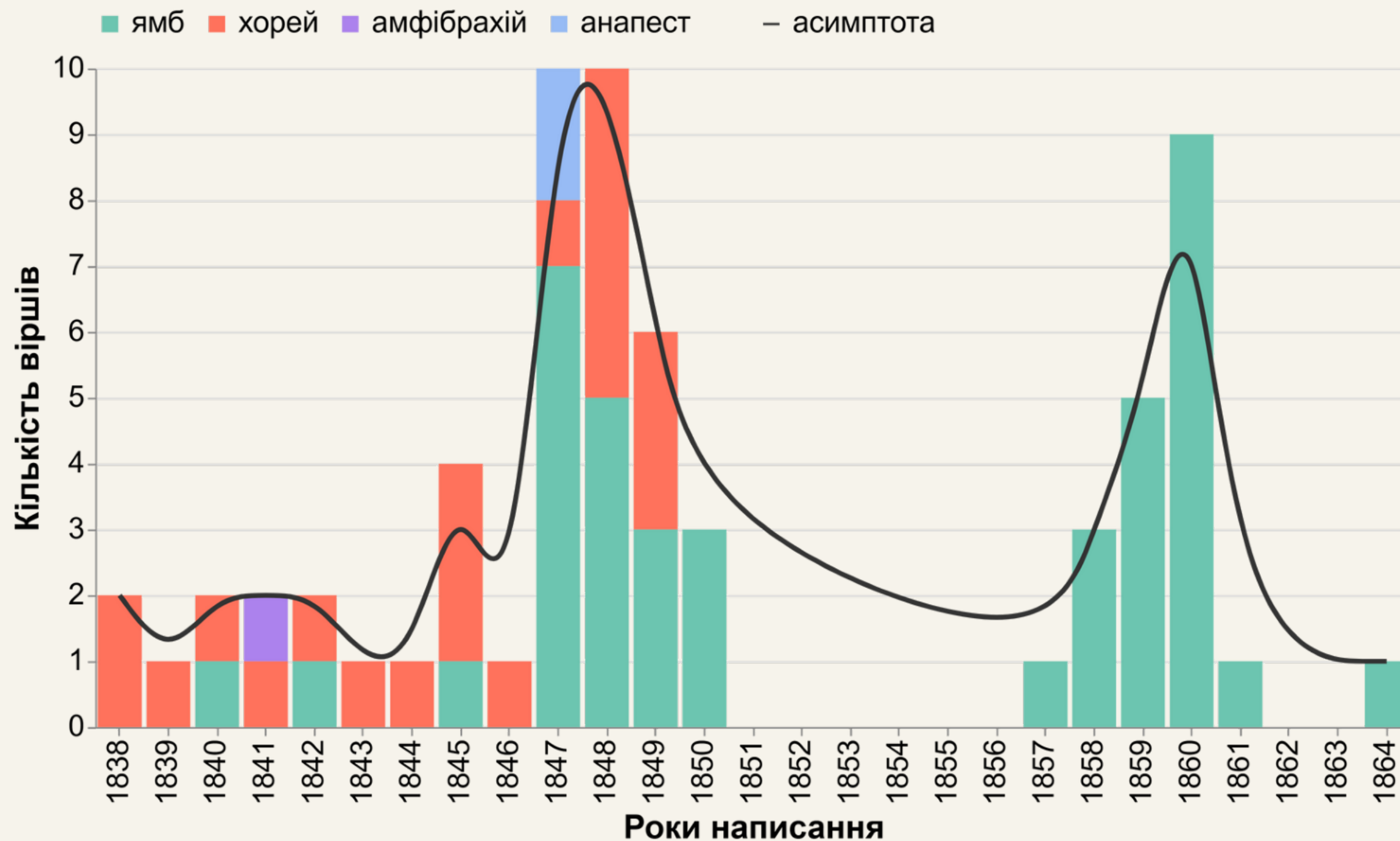
Середня точність виявлення тропів — **94,58 %**

# Порівняння точності парсера з LLM

Парсер має точність визначення метра вірша на **67,26 %** краще за середнє значення для LLM, рими – на **14,33 %**, а тропів – на **21,94 %**.

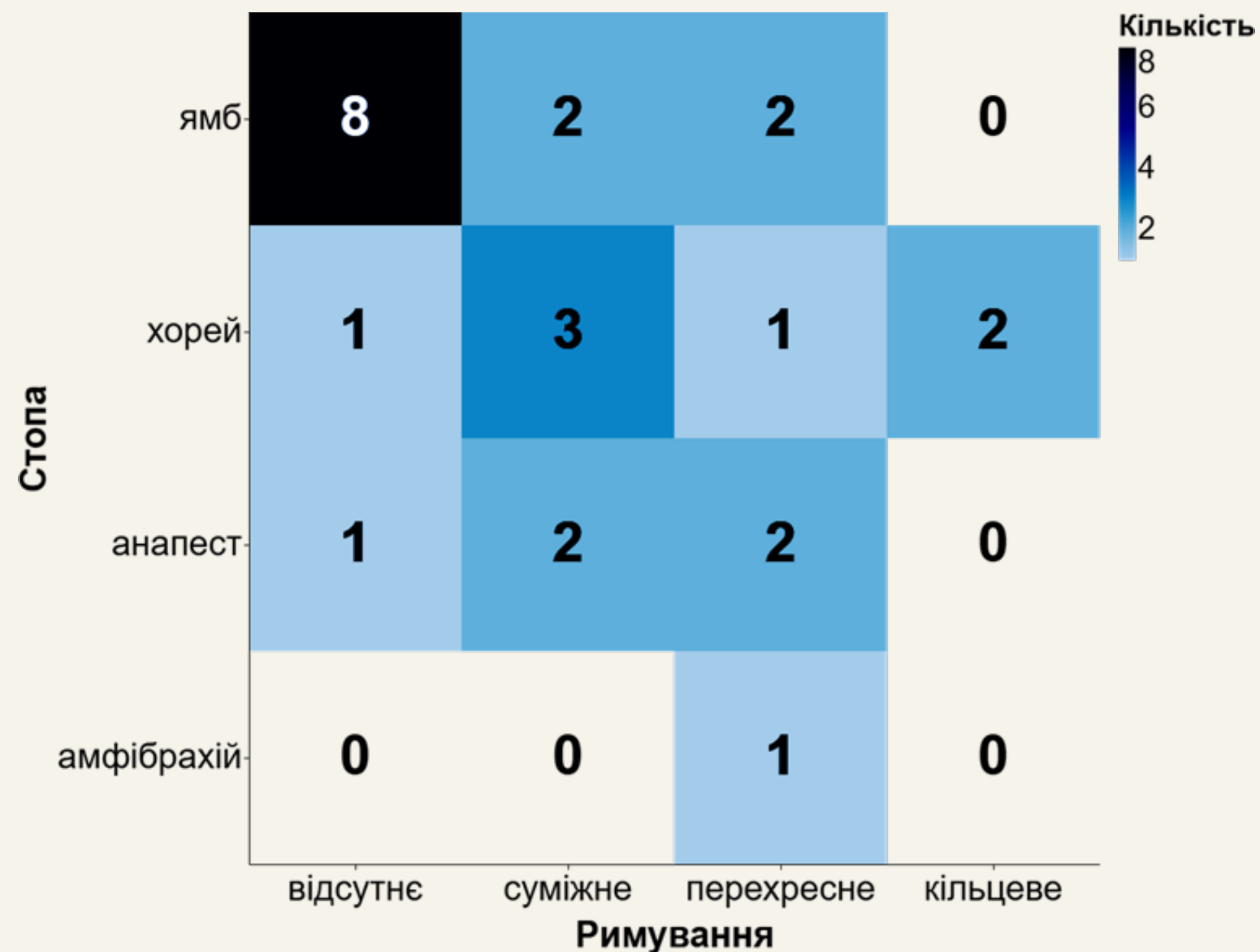


# Прикладне застосування парсера (1)



Розподіл віршів зі збірки «Кобзар» за роками написання з урахуванням стоп

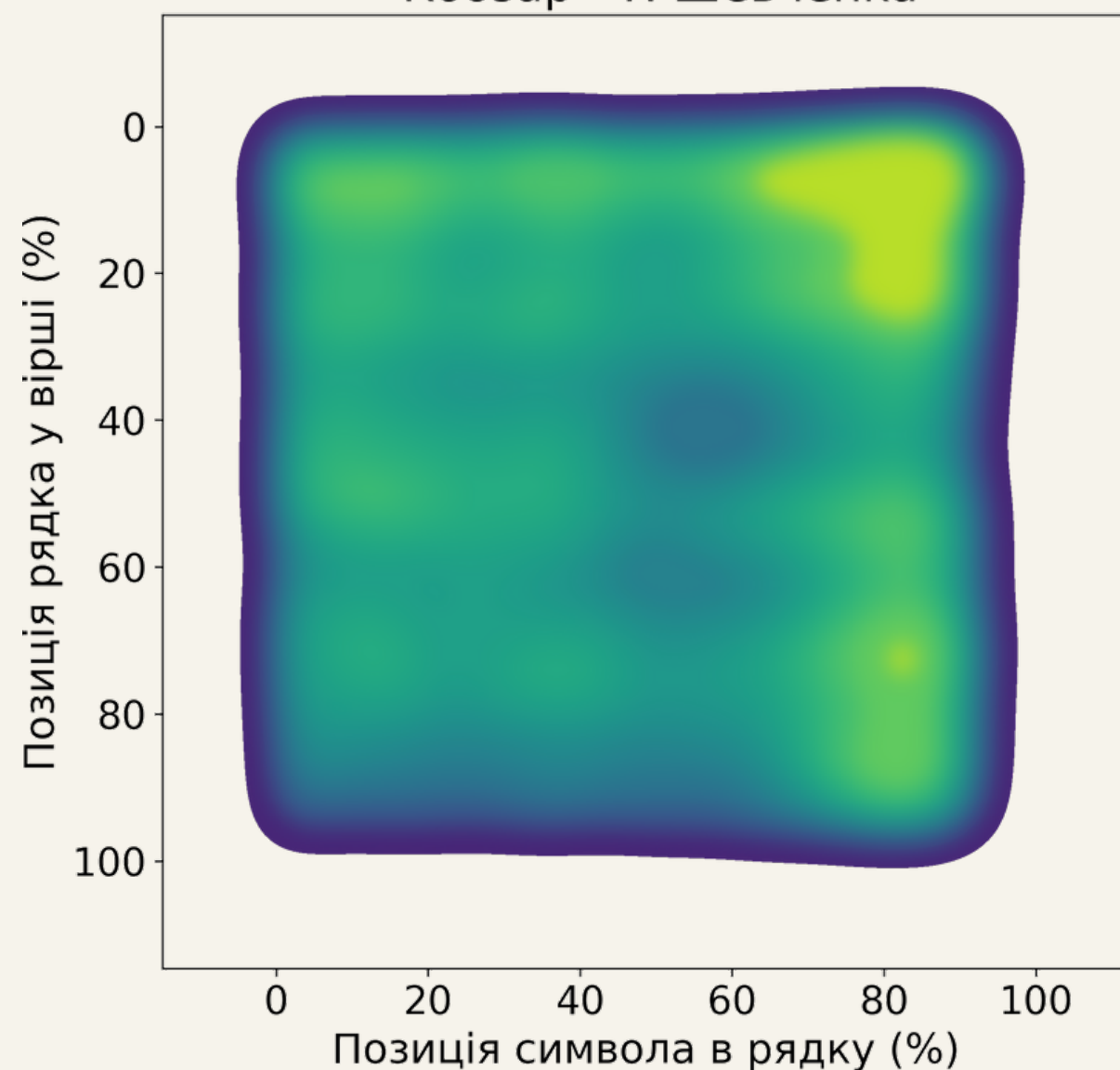
# Прикладне застосування парсера (2)



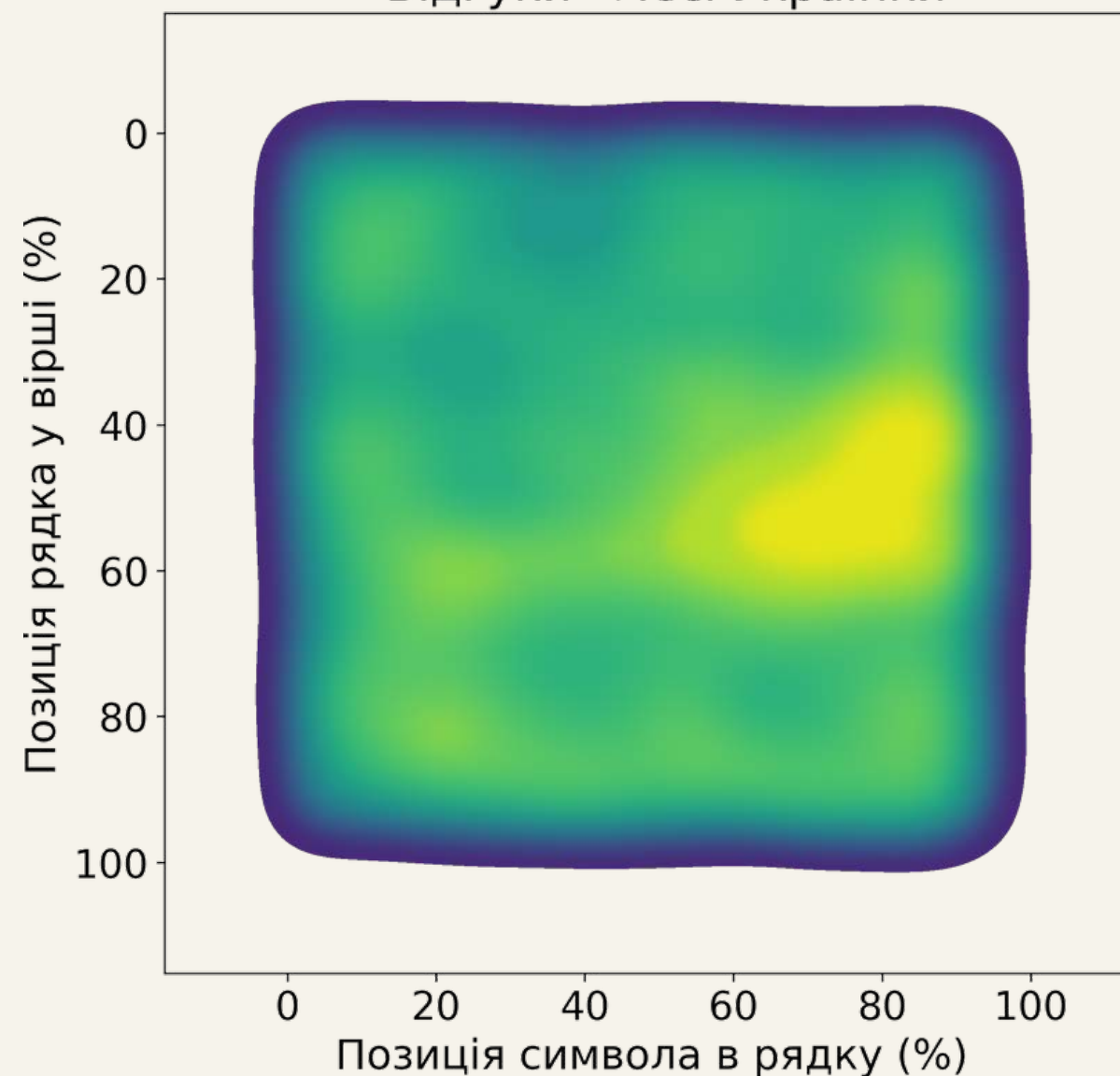
Згруповані за використаною стопою та способом римування вірші зі збірки «Відгуки»

# Прикладне застосування парсера (3)

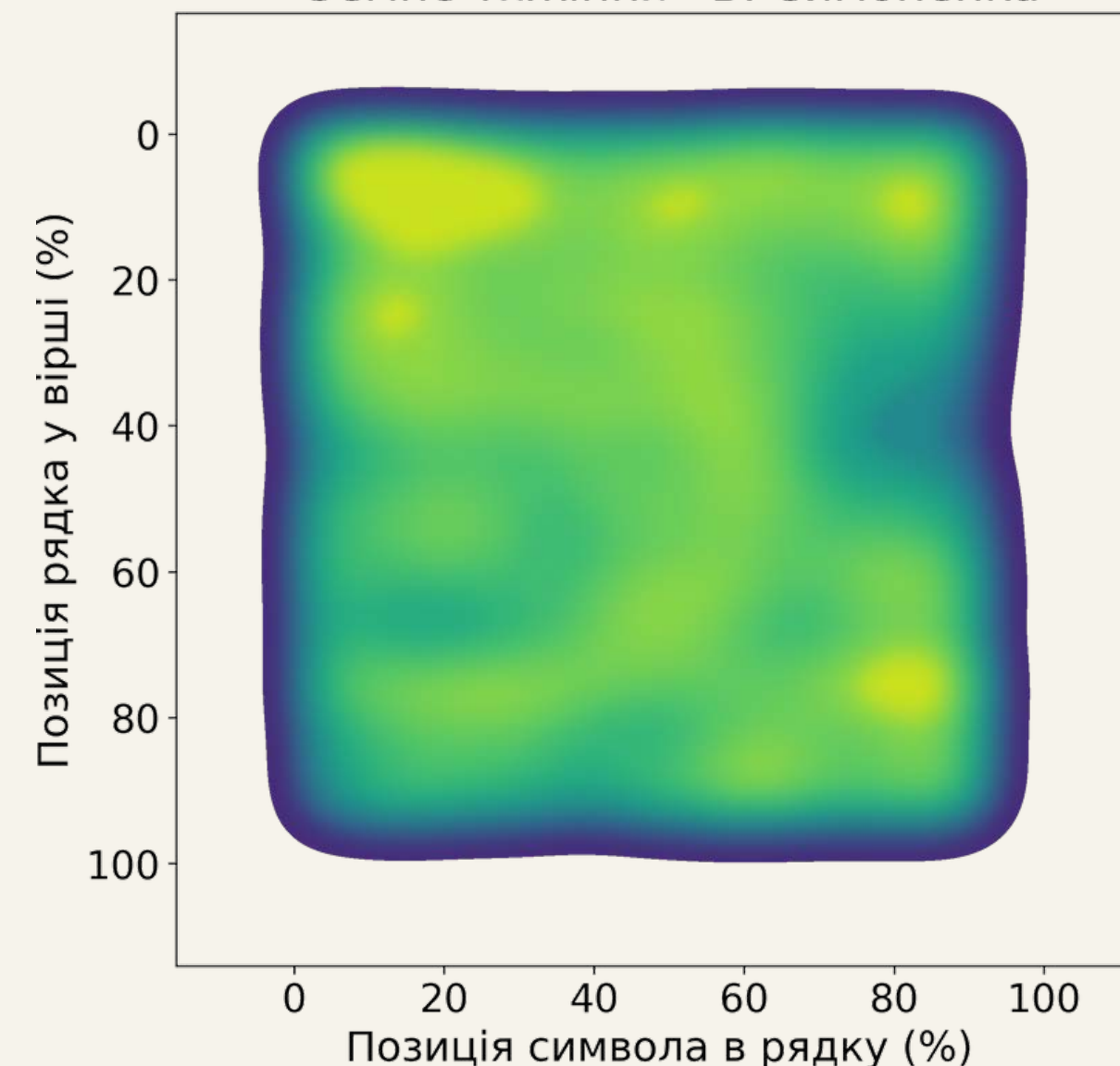
«Кобзар» Т. Шевченка



«Відгуки» Лесі Українки



«Земне тяжіння» В. Симоненка



КDE-графіки щільності розподілу тропів у збірках «Кобзар», «Відгуки», «Земне тяжіння»

# Висновки

- Оглянуто наявні інструменти для оброблення українськомовного тексту.
- Оцінено можливості аналізування українськомовних віршів **великими мовними моделями**.
- **Розроблено парсер** для визначення метра, рими, способів римування та художніх засобів.
- **Визначено точність парсера** шляхом тестування на підготовленій збірці тестів.
- **Застосовано парсер** для аналізування трьох збірок віршів.

The background features a light beige gradient. On the left side, there are three vertical bars: a wide light pink bar, a narrower teal bar, and a thin light beige bar. In the top right and bottom right corners, there is a pattern of small, light pink dots arranged in a grid that fades out towards the center.

**Дякую за  
увагу!**