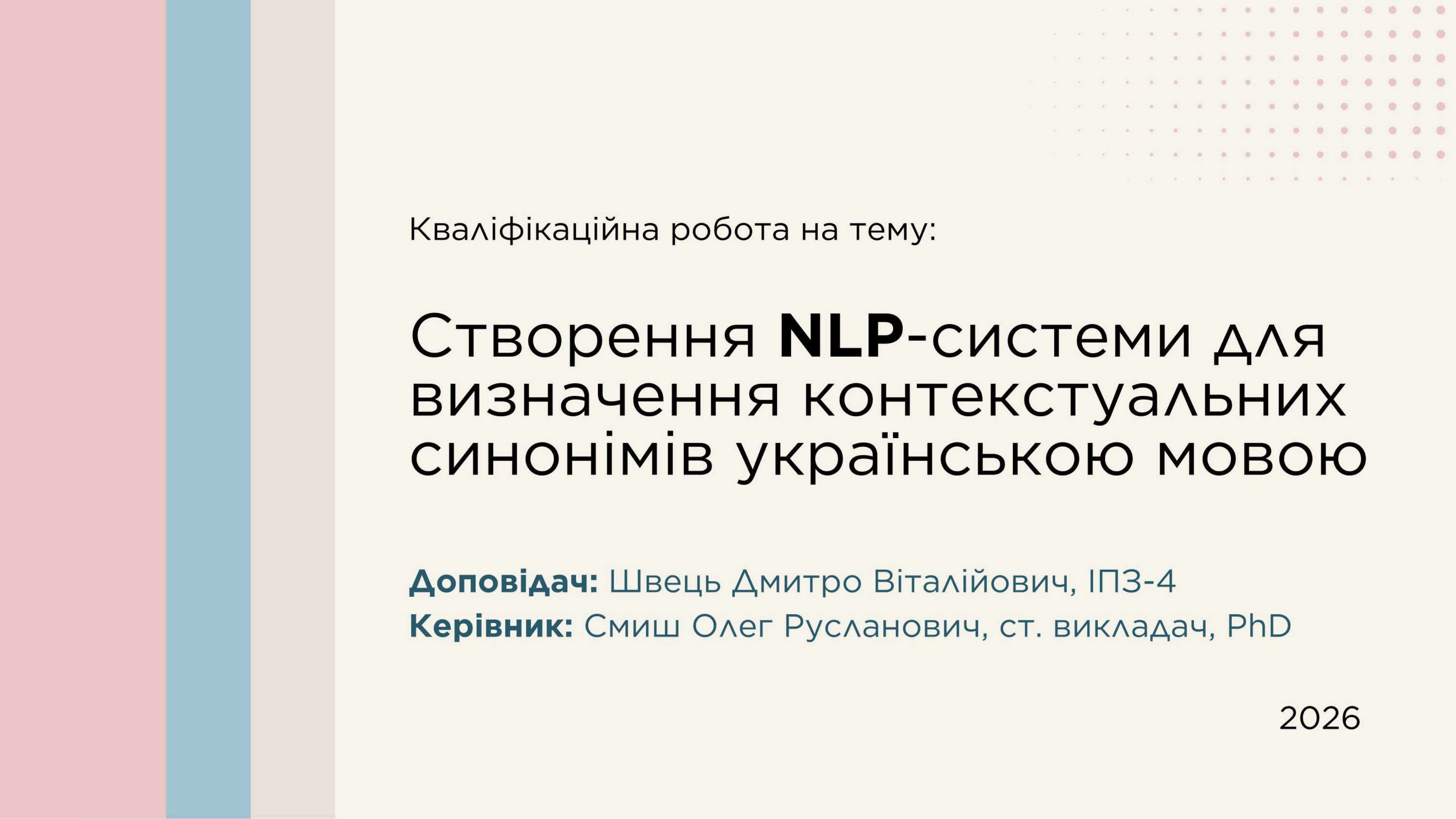




Кваліфікаційна робота на тему:

Створення **NLP**-системи для визначення контекстуальних синонімів українською мовою

Доповідач: Швець Дмитро Віталійович, ІПЗ-4

Керівник: Смиш Олег Русланович, ст. викладач, PhD

Мета та завдання кваліфікаційної роботи

Мета: автоматизувати процес визначення контекстуальних синонімів у текстах, написаних українською мовою.

Завдання:

- дослідити наявні інструменти та рішення для оброблення текстів;
- сформувати збірку контекстуальних векторних представлень слів з українськомовних корпусів;
- виокремити гіпероніми лексем з їхніх словникових тлумачень, долучивши до сформованої збірки;
- розробити методи добору контекстуальних синонімів до обраного слова в реченні;
- реалізувати модуль морфологічного узгодження синонімів;
- оцінити роботу системи та візуалізувати отримані результати.

Наявні інструменти

Корпуси:

- ГРАК — збалансований корпус з текстами різних стилів
- CNC-UA — новинний корпус публіцистичного стилю
- UD-treebanks — корпус для тренування UDPipe

Словники:

- ВЕСУМ — словник з морфологічною інформацією
- СУМ — тлумачний словник
- Словник синонімів uSofTrod

Мовні аналізатори: UDPipe, Stanza, spaCy

Наявні рішення

Дослідження: Contextual Embeddings for Ukrainian: A Large Language Model Approach to Word Sense Disambiguation / Y. Laba et al. *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*. 2023. P. 11–19.

Модель: paraphrase-multilingual-mpnet-base-v2 (PMMB2)

- Sentence Transformers (SBERT)
- XLM-RoBERTa
- mean pooling, розмірність 768
- найвища точність WSD серед аналогів: **73,5 %**

Дотренування на основі дослідження:

- мета: краще розрізнення омонімів
- набір трійок (anchor – positive – negative) з UberText 2.0
- TripletMarginLoss
- точність WSD після дотренування: **77,9 %**

Обчислення контекстуальних векторів

Вхідне речення з **обраним** словом

«Думки **формують** реальність»

Речення з **лематизованим** словом

«Думки **формувати** реальність»

Сформовані **субтокени**

_ду, м, ки, **_форм**, **увати**, _реальн, ість

Обчислення **вектора**

`mean_pooling([hidden[4], hidden[5]])`

Нормалізований **вектор**
розмірністю 768

`[0.2354, ..., 0.5426]`

Вилучення гіперонімів із тлумачень

1. Слово: «**атом**» (NOUN).

Тлумачення: «**найдрібніша частинка хімічного елемента**».

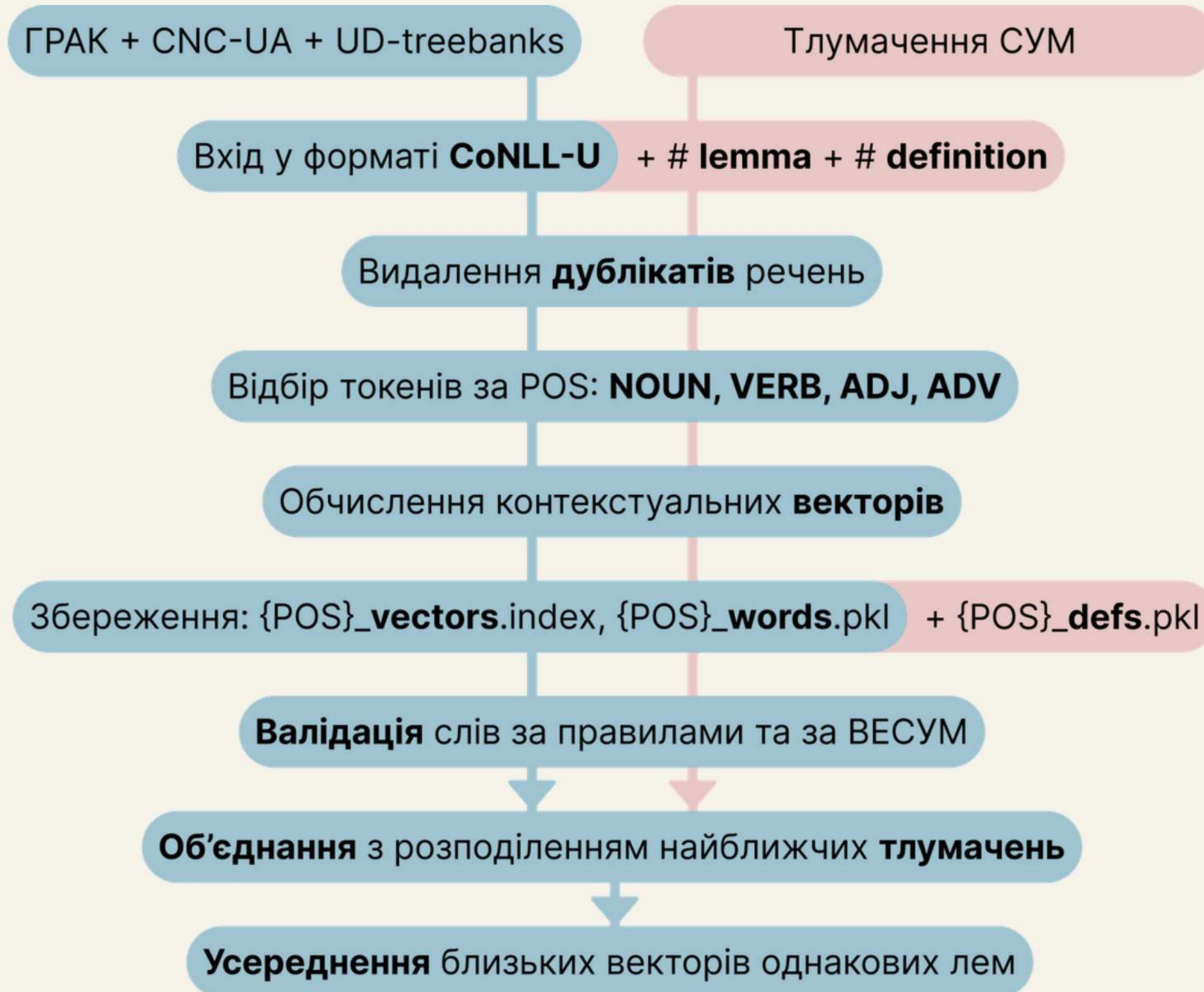
2. Результат роботи мовних аналізаторів:

ID	Form	Lemma	UPOS	Head	Deprel
1	найдрібніша	найдрібніший	ADJ	2	amod
2	частинка	частинка	NOUN	0	root
3	хімічного	хімічний	ADJ	4	amod
4	елемента	елемент	NOUN	2	nmod

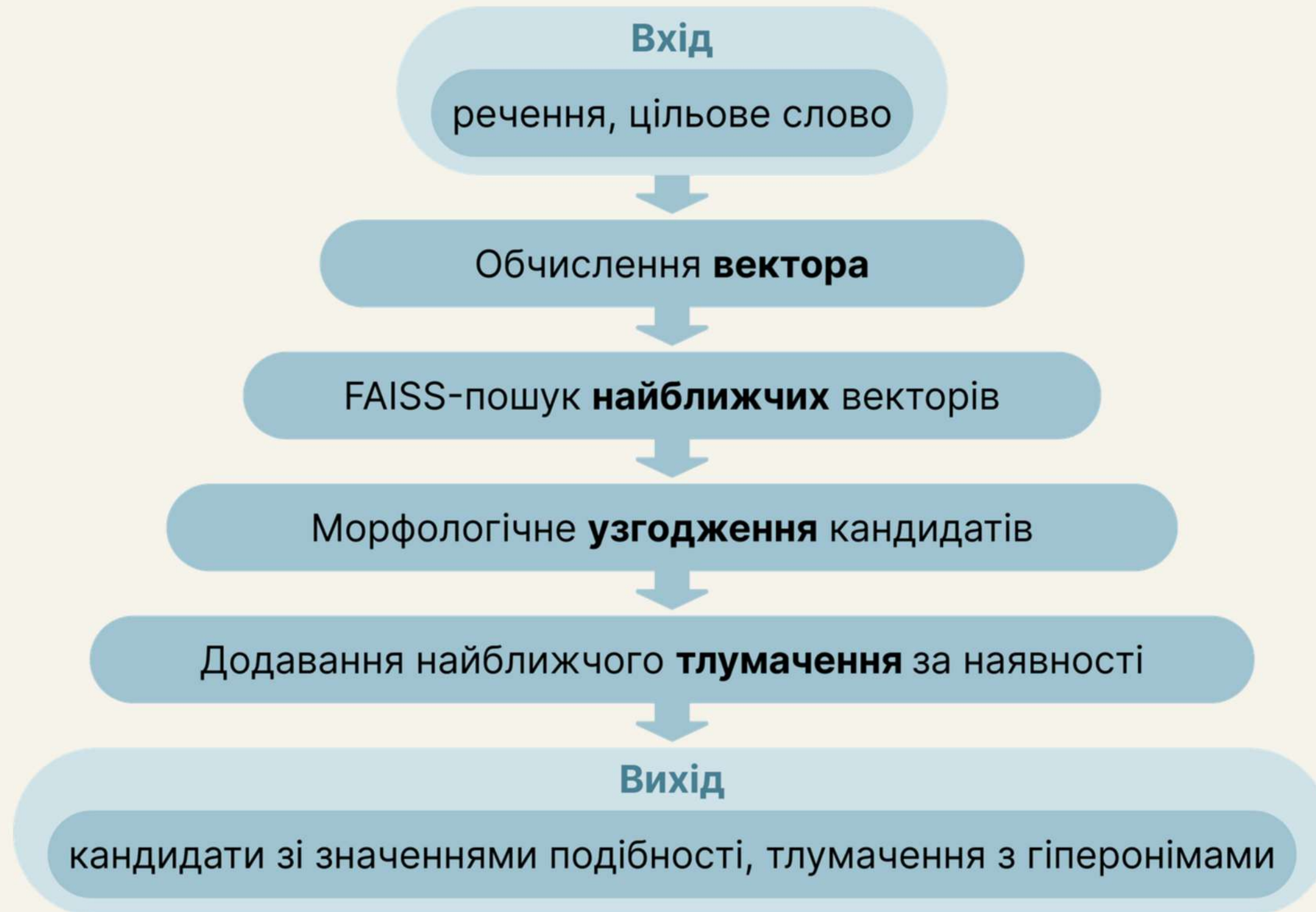
3. root = «**частинка**» (NOUN)

4. Результат: «найдрібніша **[частинка]** хімічного елемента»

Формування збірки ембедингів



Механізм добору контекстуальних синонімів



Створення вебзастосунку

Приклад визначення контекстуальних синонімів для слова «усвідомлював»:

ТЕКСТ

редагувати

Студент поступово осягав розумом складність поставленого завдання.

ОБРАНЕ СЛОВО

усвідомлював

Знайти синоніми

СИНОНІМИ

VERB

1 розумів	розуміти	98%
2 осмислював	осмислювати	95%
3 уявляв	уявляти	91%
4 тямив	тіяти	91%
5 виявляв	виявляти	88%
6 сприймав	сприймати	87%

ЗНАЧЕННЯ

усвідомлював

[осягав] розумом, [сприймав] свідомо, [розумів] значення, сенс

Удосконалення розробленої системи

Удосконалення	Проблема	Рішення
Перше	повне речення вносить семантичний шум у вектор слова	синтаксично орієнтований контекст
Друге	ранжування за схожістю двох слів не враховує загального змісту речення	переранжування результатів на рівні речень (sentence embeddings)

Методи оцінювання роботи системи

- w_0 — **цільове слово**
- $S(w_0)$ — **еталонна** множина: словникові синоніми
- $C(w_0)$ — **кандидатська** множина: слова зі збірки
- $f(w)$ — максимальна косинусна **подібність** w до w_0

$$S(w_0) = \left(\bigcup_{i=1}^n D_i \right) \setminus w_0$$

$$S(w_0) \subseteq C(w_0)$$

Верхній поріг подібності:

$$\tau_{up}(w_0) = \min\{f(w) : w \in C(w_0), \forall w' \in C(w_0) : f(w') \geq f(w) \Rightarrow w' \in S(w_0)\}$$

Нижній поріг подібності:

$$\tau_{low}(w_0) = \min\{f(w) : w \in C(w_0)\}$$

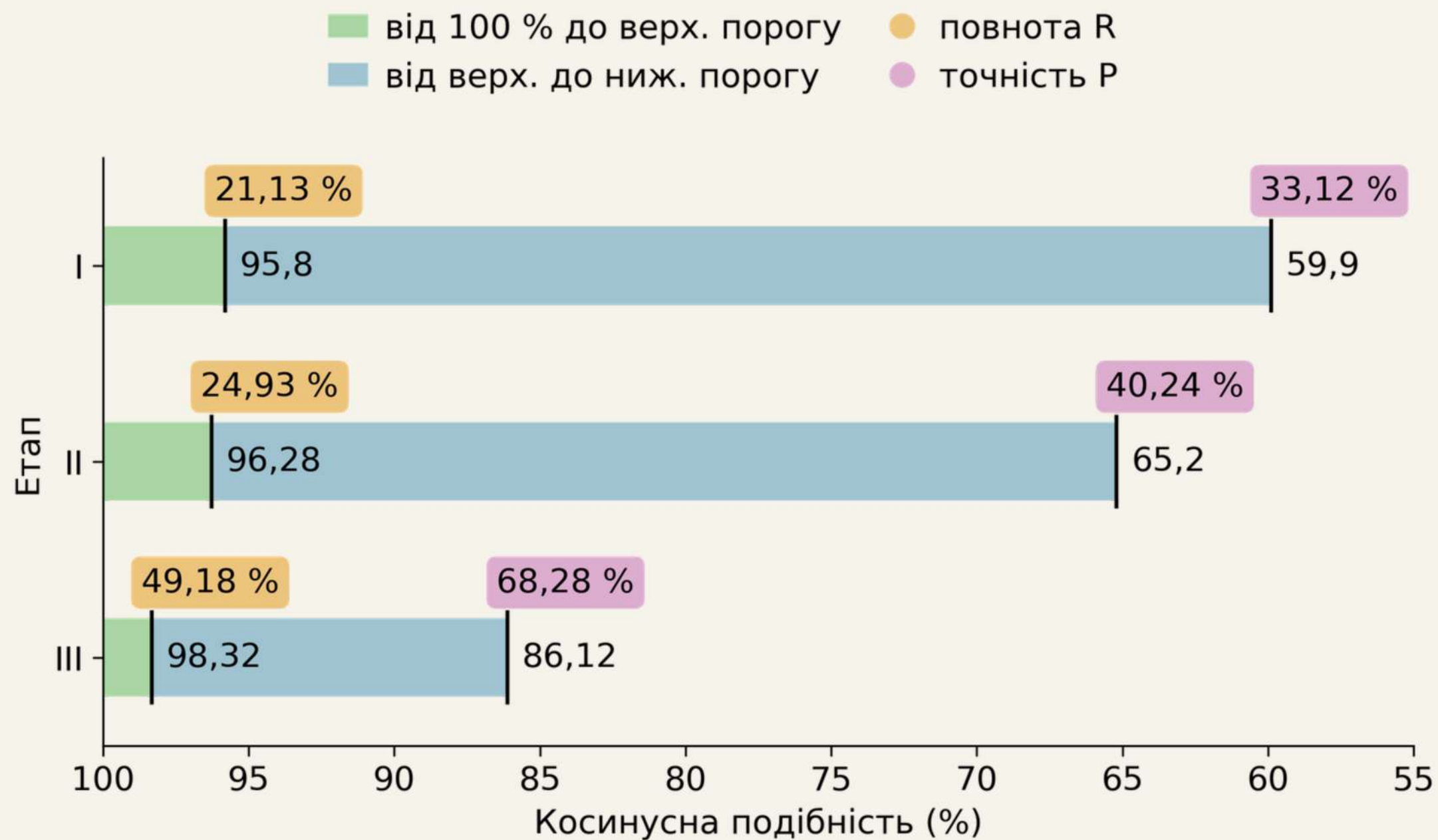
Повнота на верхньому порозі:

$$R(w_0) = \frac{|\{w \in S(w_0) : f(w) \geq \tau_{up}(w_0)\}|}{|S(w_0)|}$$

Точність на нижньому порозі:

$$P(w_0) = \frac{|S(w_0)|}{|C(w_0)|}$$

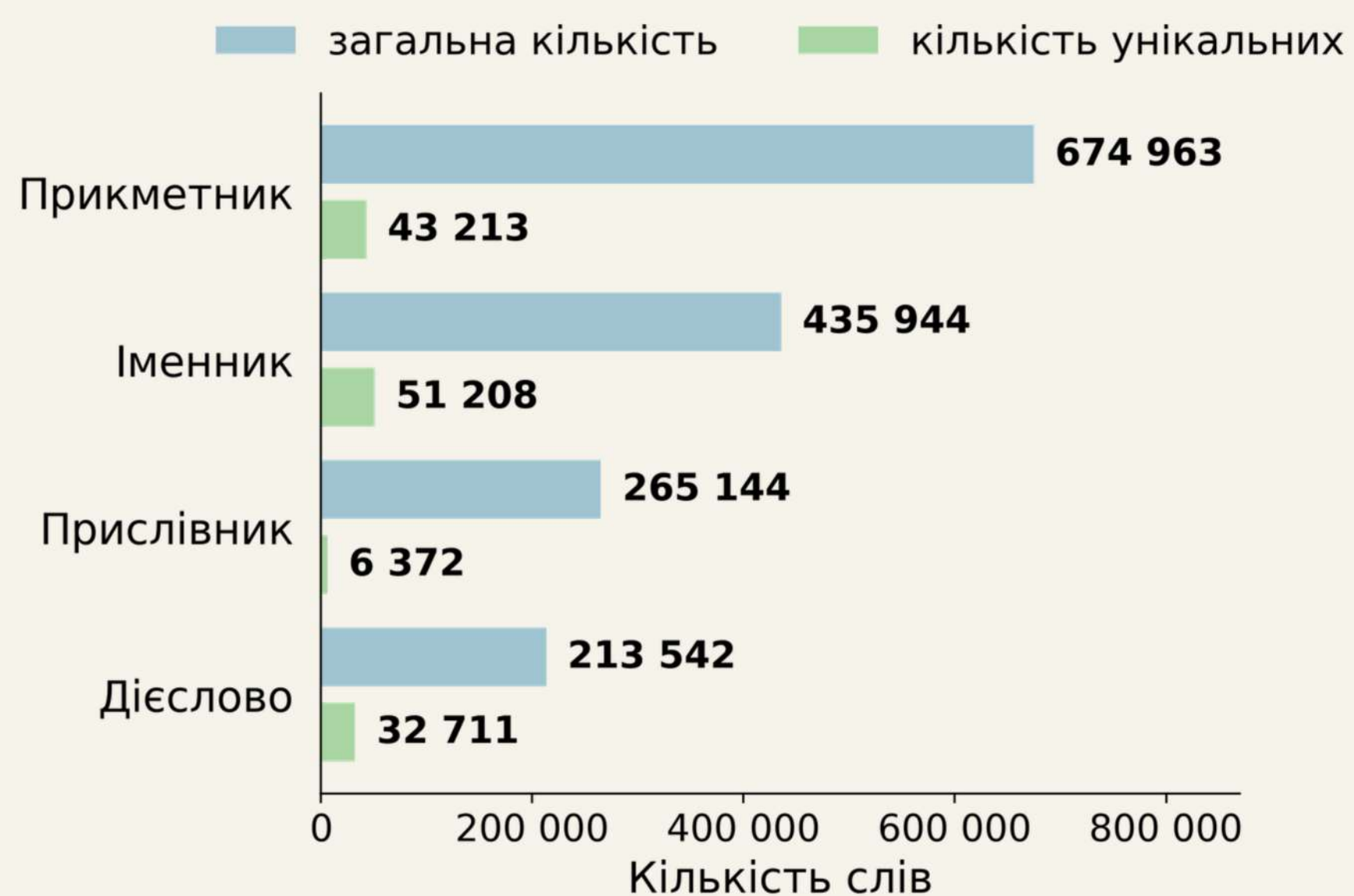
Результати оцінювання



Фінальні результати:

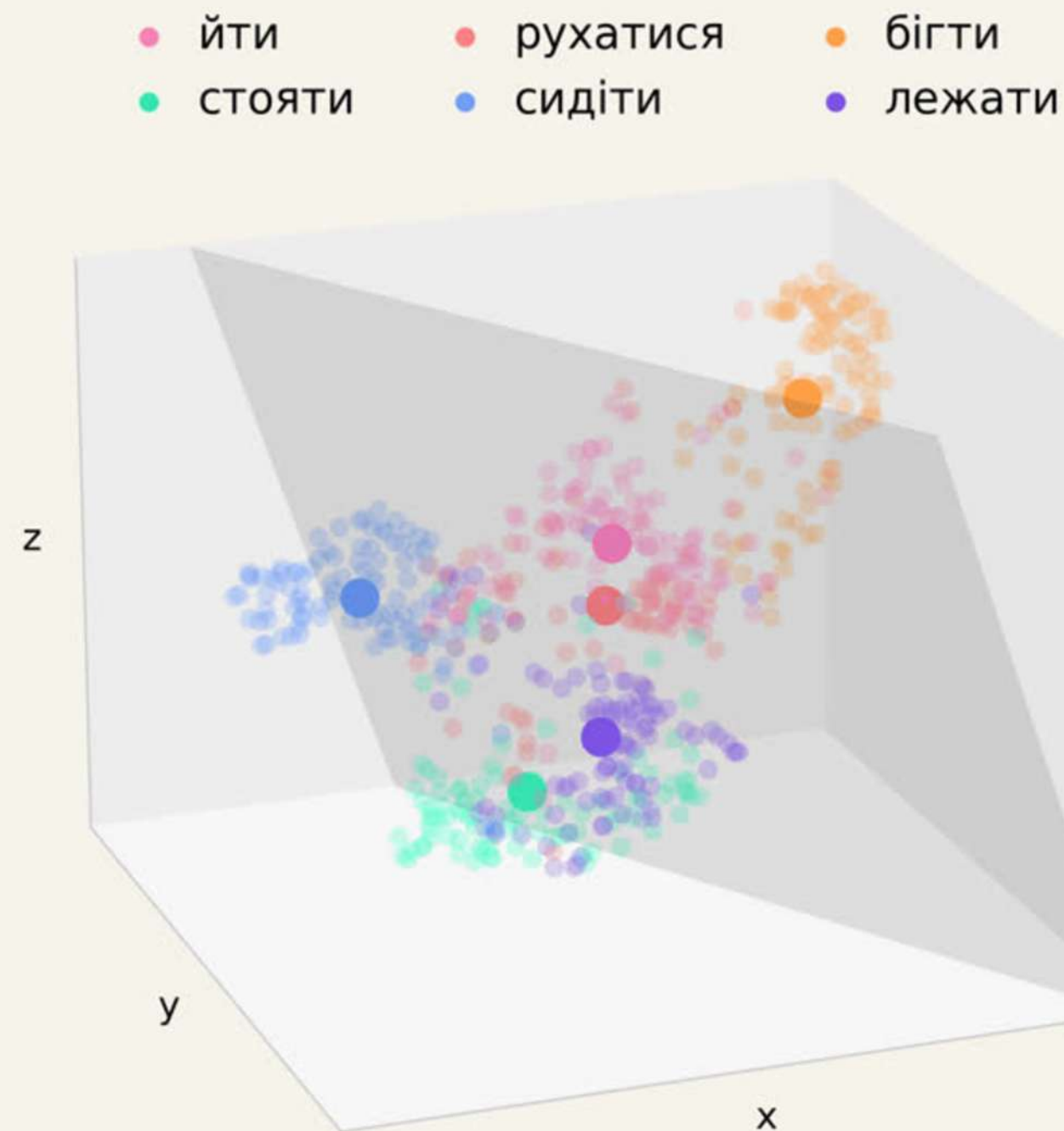
Повнота становить **49,18 %**, а **точність** — **68,28 %**

Апробація сформованої збірки векторів



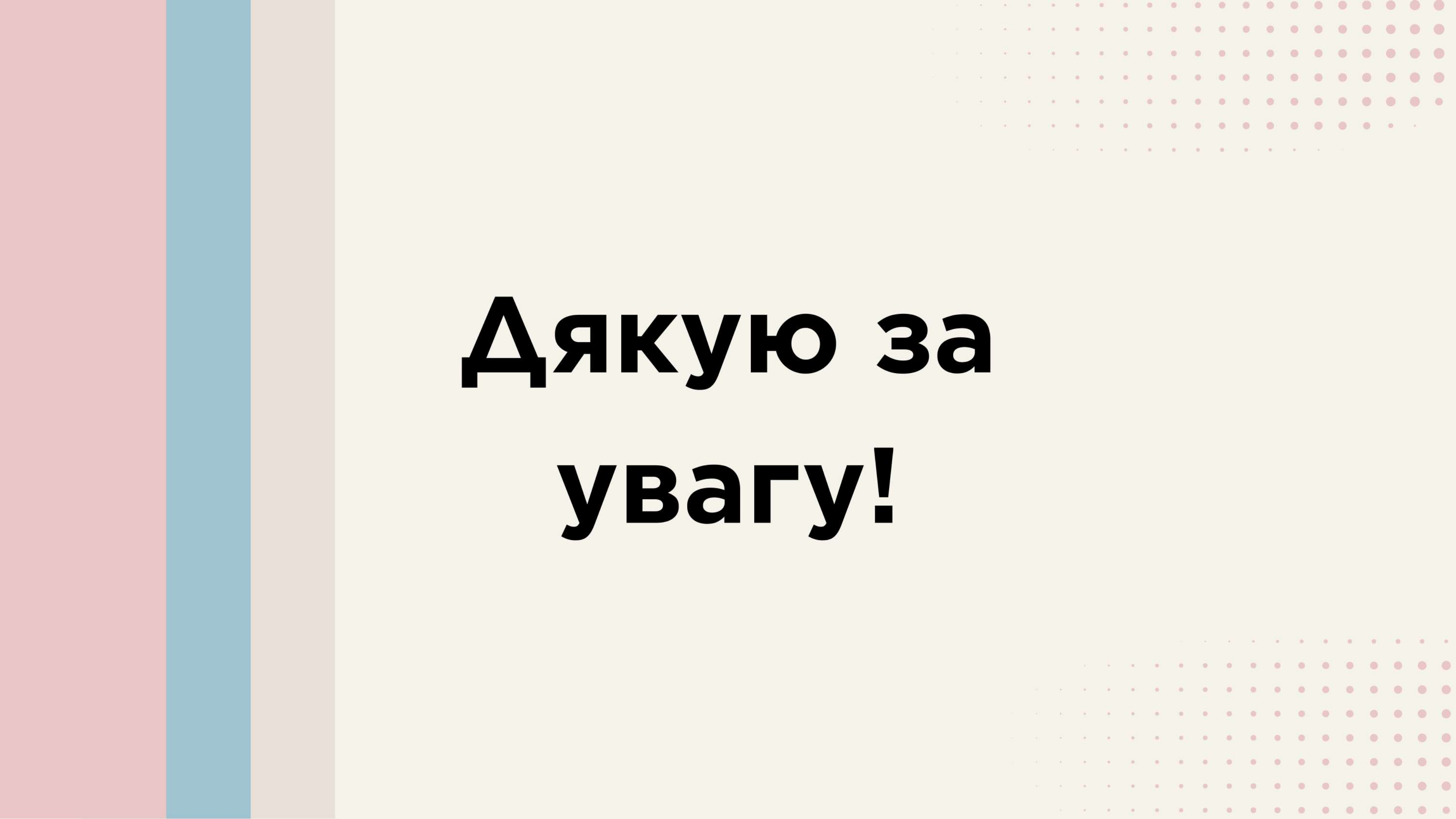
Загальна кількість слів: **1 589 593**

Унікальних слів: **94 604**



Висновки

- **Оглянуто** інструменти та наявні рішення
- **Сформовано збірку** контекстуальних ембедингів на основі текстових корпусів
- Здійснено **вилучення гіперонімів** зі словникових тлумачень
- **Реалізовано** модуль морфологічного **узгодження**
- Запропоновано два **вдосконалення** системи
- **Оцінено** роботу системи на основі словника синонімів: **точність — 68,28 %, повнота — 49,18 %**
- Отримані результати **візуалізовано** та розроблено **вебзастосунок**

The background features three vertical stripes on the left: a wide pink stripe, a narrower blue stripe, and a narrow beige stripe. The right side of the image is a light beige background with two rectangular areas of small, light pink dots in the top right and bottom right corners.

**Дякую за
увагу!**