Кваліфікаційна робота на тему:

Створення гібридної NLP-системи морфемного аналізу українськомовних слів

Доповідач: Кирилін Єгор, МП «Комп'ютерні науки» 2 р.н.

Керівник: Смиш Олег Русланович, ст. викладач, PhD

2026

Мета та завдання кваліфікаційної роботи

Мета: автоматизувати процес морфемної сегментації українськомовних слів через розроблення системи для визначення структури слова.

Завдання:

- дослідити наявні методи морфемного аналізу та їхні обмеження;
- сформувати навчальну вибірку з BIO-розміткою;
- навчити базові та нейромережеві моделі морфемної сегментації;
- порівняти результати трьох архітектурних підходів;
- розробити гібридну каскадну систему обробки слова;
- створити вебзастосунок для наочної візуалізації морфемного розбору;
- оцінити точність системи на тестовій вибірці.

Специфіка та проблематика морфемної сегментації

Автоматизацію морфемного розбору української мови ускладнено низкою лінгвістичних чинників, що визначають її граматичну структуру:

- використання префіксально-суфіксальних способів словотворення та складних ланцюгів афіксів;
- фонетичні чергування голосних і приголосних звуків у коренях під час словозміни;
- структурні аномалії, як-от нульові закінчення чи наявність постфіксів після закінчень;
- морфемна омонімія, за якої однакові буквосполучення мають різне значення залежно від контексту;
- поява неологізмів та термінів, яких немає у словникових реєстрах.

Аналіз наявних рішень

Морфосинтаксичні аналізатори
UDPipe, Stanza, spaCy

Лематизують і розмічують частини мови, але розглядають слово як неподільну атомарну одиницю.

Лексикографічні ресурси
ВЕСУМ, Словник афіксальних морфем

Надають точні дані про склад словоформ, проте обмежені принципом прямого пошуку за статичною базою.

Великі мовні моделі
GPT, Gemini, Sonnet

Визначають контекст, але через алгоритми підслівного кодування втрачають доступ до літерного складу слова.

Статистичні методи
Morfessor, SIGMORPHON

Автоматично визначають межі поділу слів на основі частотності символів без можливості класифікації типів морфем.

Підготовка та структурування набору даних

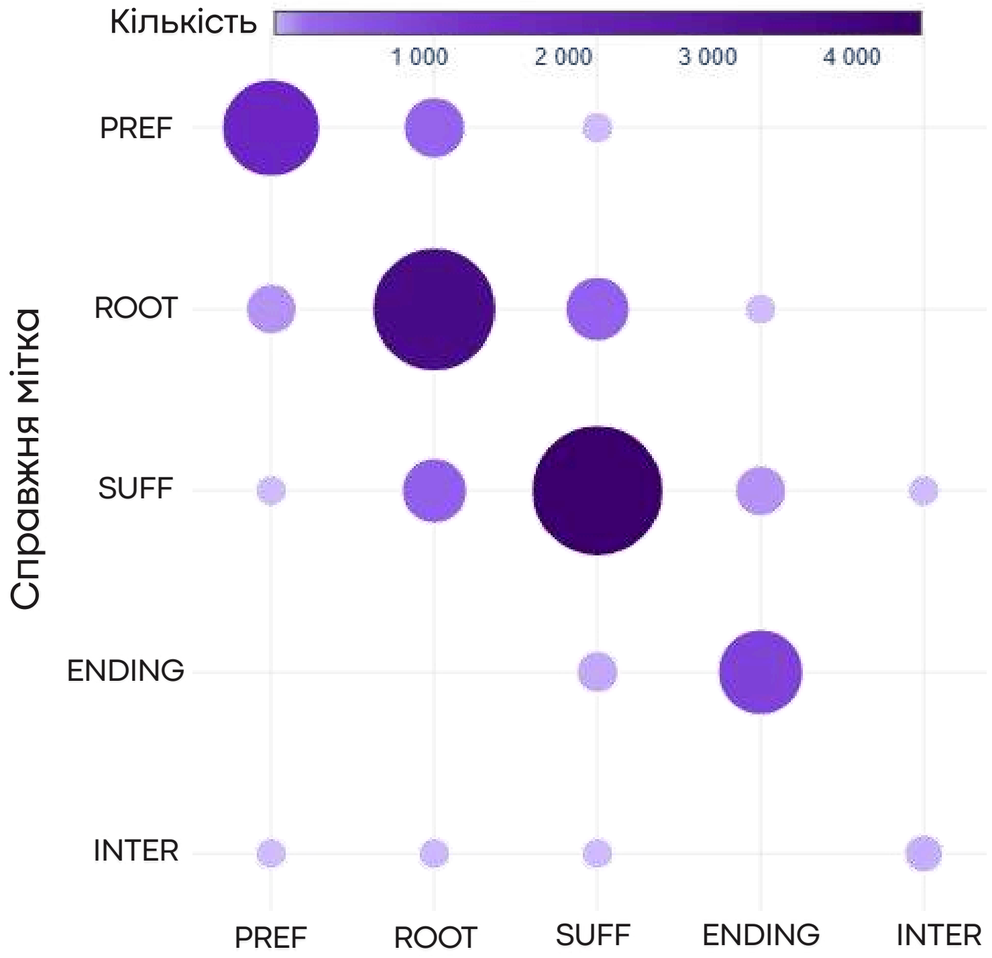
Навчальний корпус сформовано через інтеграцію даних зі Словника афіксальних морфем та ВЕСУМ. Морфемні межі перетворено у формат BIO-маркування для посимвольної класифікації кожної літери.

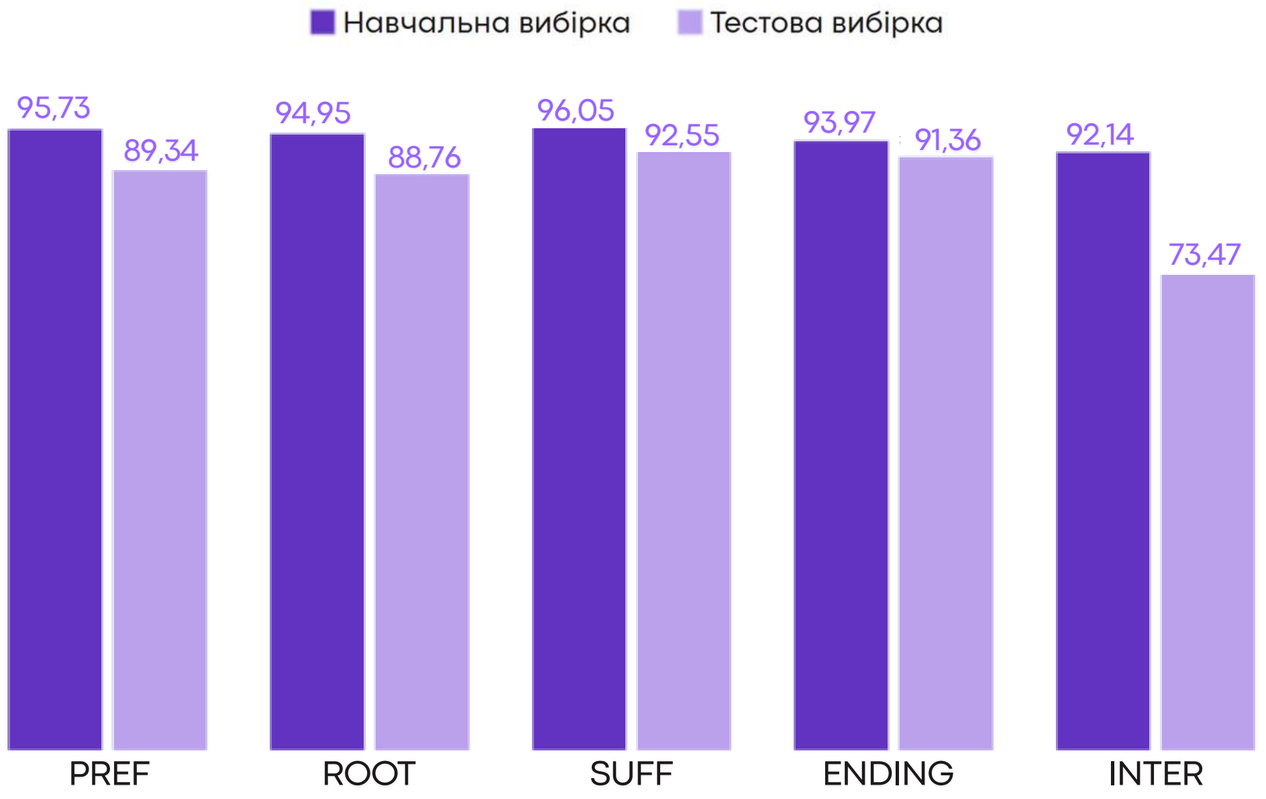
Приклад представлення морфемної структури слова у форматі BIO-маркування

Літера	Тип морфеми	BIO-мітка
Н	префікс	B-PREF
А	префікс	I-PREF
Д	префікс	I-PREF
Т	корінь	ROOT
О	корінь	ROOT
Ч	корінь	ROOT
Н	суфікс	B-SUFF
И	закінчення	ENDING
Й	закінчення	ENDING

Оцінка базового класифікатора

Точність повнослівного розбору — 44,70 %

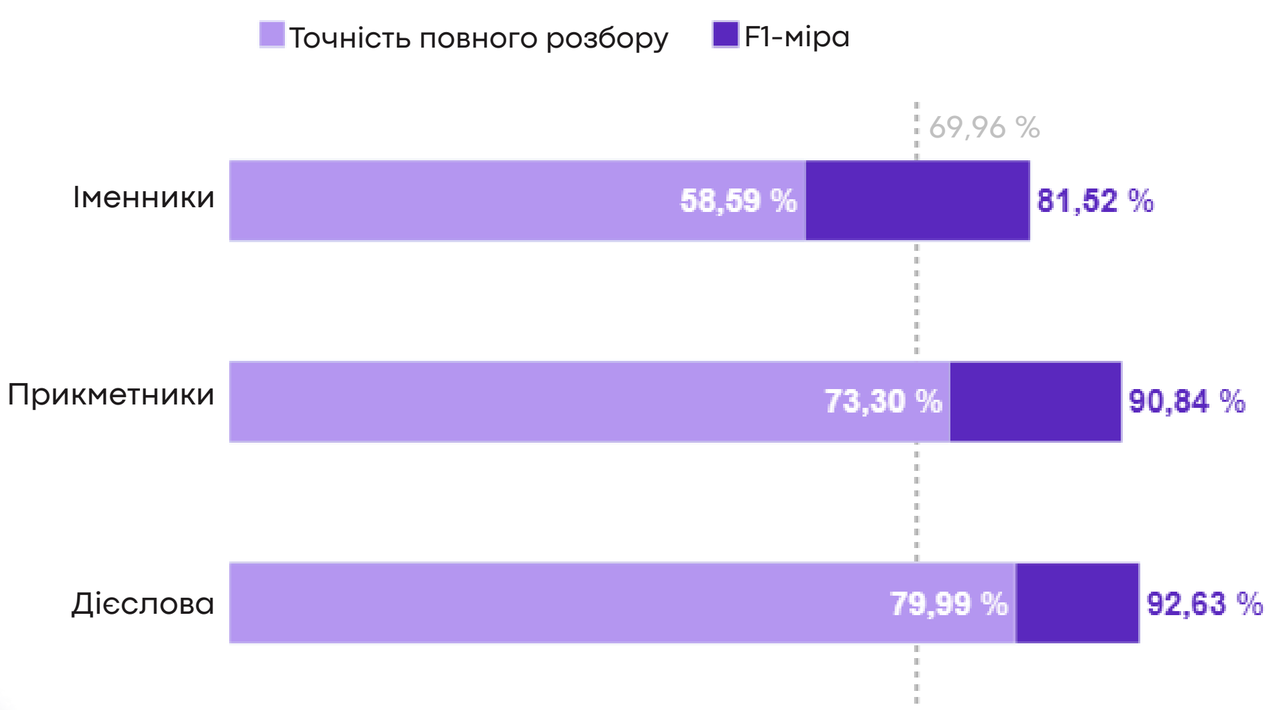


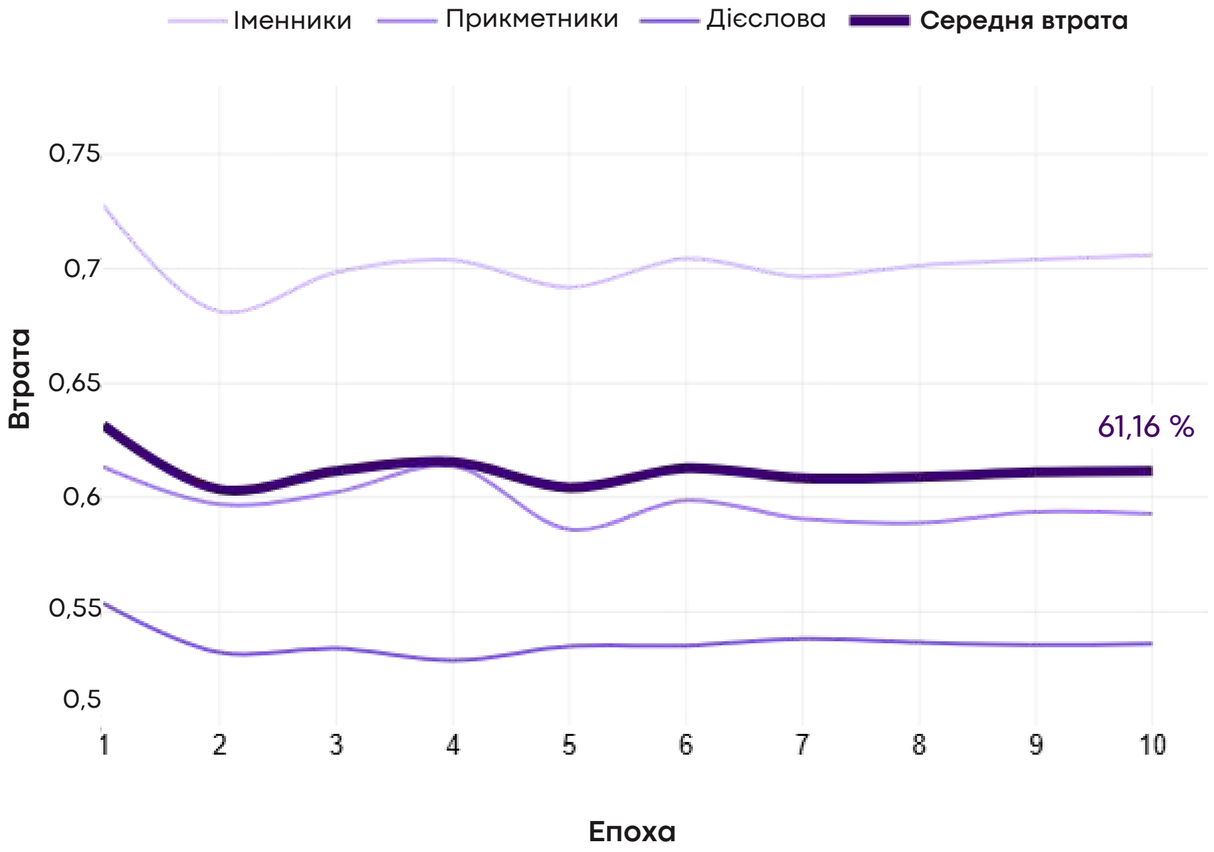


Порівняння F1-міри за типами морфем, %

Оцінка спеціалізованих моделей за частинами мови

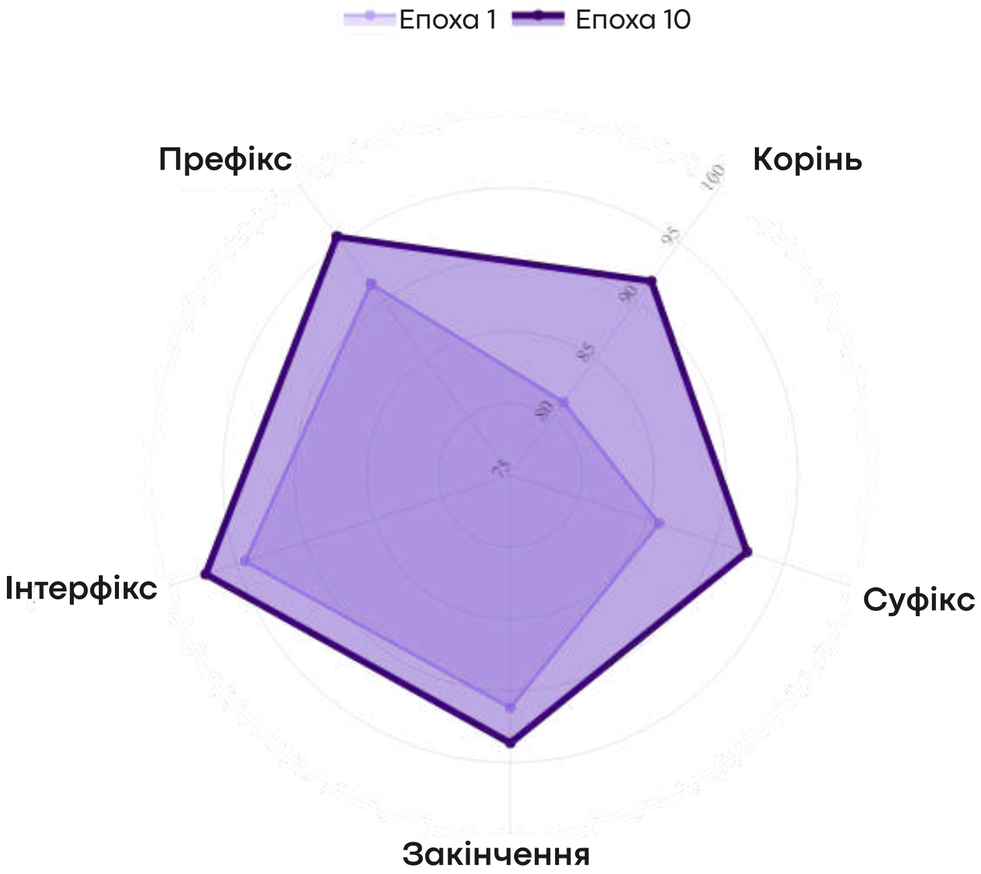
Точність повнослівного розбору — 69,96 %



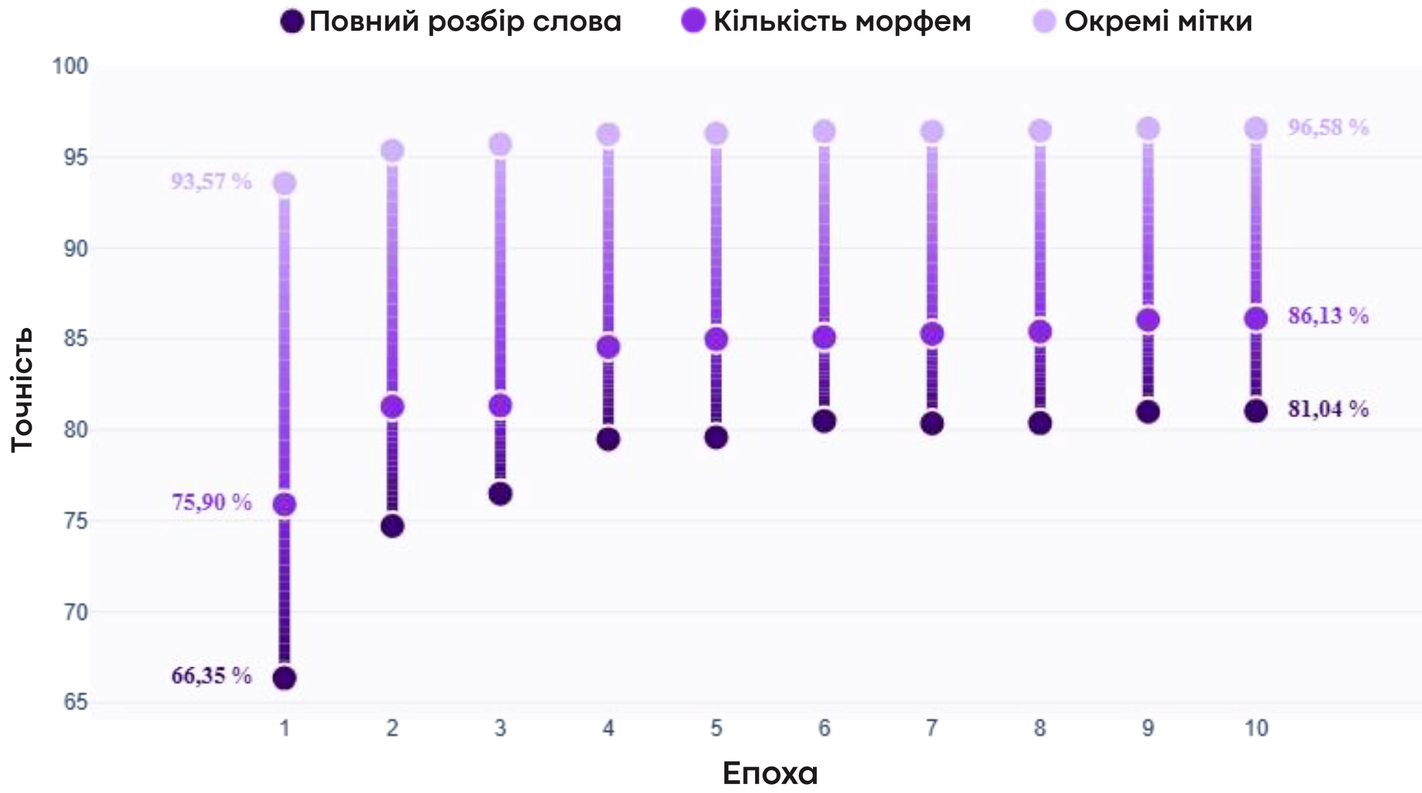


Оцінка універсальної моделі

Точність повнослівного розбору — 81,04 %



Порівняння F1-міри за типами морфем



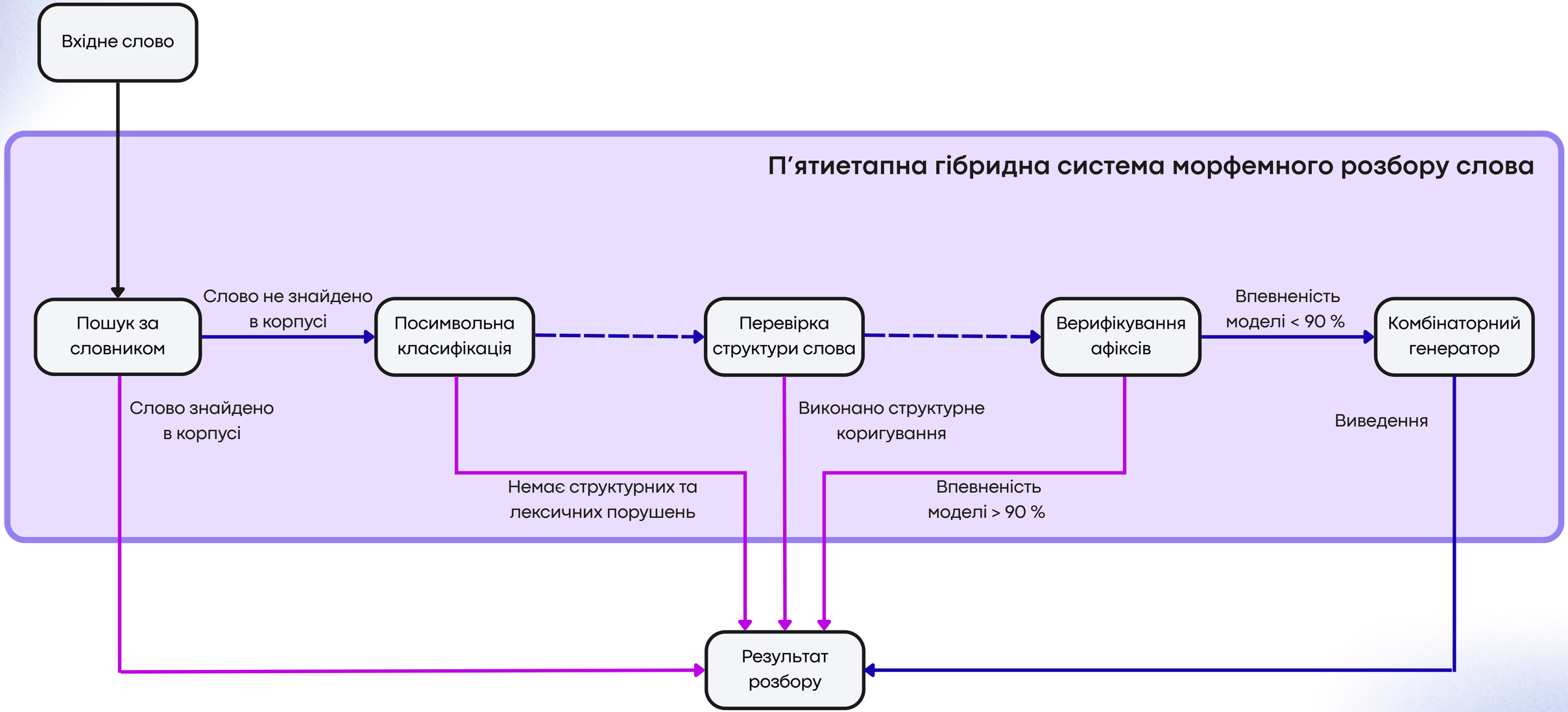
Динаміка метрик якості

Порівняння метрик розроблених методів

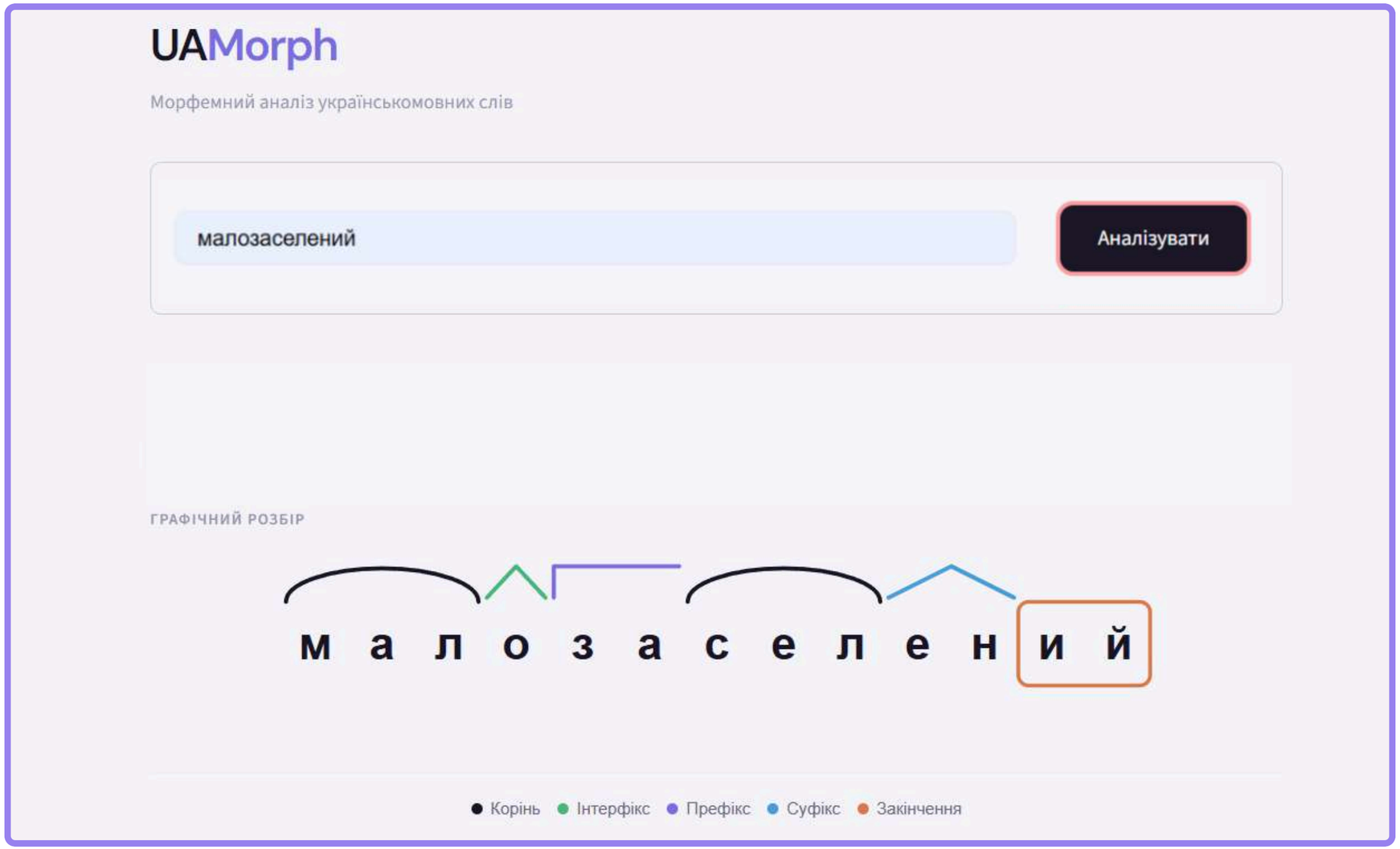
Метод аналізу	Архітектура	Повнослівний розбір	Середньозважена F1-міра
Базовий класифікатор	Random Forest	44,70 %	90,79 %
Ансамбль частиномовних моделей	Transformer (POS-Split)	69,96 %	91,62 %
Універсальна модель	**Transformer (End-to-End)**	**81,04 %**	**92,91 %**

Універсальна модель демонструє приріст точності повного розбору на **+36,33 %** порівняно з базовим класифікатором

Архітектура гібридної системи



Програмна реалізація та візуалізація результатів системи



Апробація отриманих результатів

Висновки

Досліджено проблематику автоматичного морфемного аналізування.

Спроєктовано гібридну систему, що поєднує нейромережевий підхід із лінгвістичними правилами.

Сформовано українськомовний набір даних із морфемною розміткою.

Розроблено три архітектурні підходи: базову модель лінійного типу **(44,70 %)**, ансамбль частиномовних моделей **(69,96 %)** та універсальну трансформерну нейромережу **(81,03 %)**.

Доведено перевагу єдиної трансформерної архітектури над спеціалізованими моделями.

Точність системи **перевищує** результати великих мовних моделей у середньому на **85,75 %.**

Дякую за увагу