



Григор'єва Єлизавета Романівна,
студентка III курсу
(групи УМЛб 1-23 40д.), спеціальності 035
«Філологія»
Факультету української філології,
культури і мистецтв
Київського столичного університету імені
Бориса Грінченка.

Науковий керівник:

М.О. Вінтонів, доктор філологічних наук, професор, завідувач кафедри української мови Факультету української філології, культури і мистецтв Київського столичного університету імені Бориса Грінченка.

РОЗПІЗНАВАННЯ УКРАЇНСЬКИХ ФРАЗЕМНИХ КОНСТРУКЦІЙ У ВЕЛИКИХ МОВНИХ МОДЕЛЯХ: ПОРІВНЯЛЬНЕ ДОСЛІДЖЕННЯ CHATGPT, GEMINI ТА CLAUDE

Стаття присвячена дослідженню здатності сучасних великих мовних моделей (LLM) — ChatGPT, Gemini та Claude — розпізнавати українські фраземні конструкції в текстах корпусу ГРАК. Проведено кількісний аналіз частотності вибраних фразеологізмів та емпіричне тестування моделей за трьома критеріями: ідентифікація меж фраземи, інтерпретація її значення та семантично коректне перефразування. Результати свідчать про неоднорідність лінгвістичних стратегій і різну чутливість моделей до частотності фраземних одиниць, що

має важливі наслідки для подальшого удосконалення алгоритмів обробки природної мови в українському мовному середовищі.

Ключові слова: фраземна конструкція, фразеологізм, великі мовні моделі, українська мова, корпус ГРАК, розпізнавання фразем, ChatGPT, Gemini, Claude.

Постановка проблеми в контексті сучасної філологічної науки.

Автоматизоване розпізнавання фраземних конструкцій є однією з ключових проблем сучасної прикладної філології, оскільки стійкі вирази мають складну структурно-семантичну природу, що ускладнює їхню комп'ютерну обробку. Прогрес у галузі великих мовних моделей відкриває нові методологічні можливості для аналізу і розпізнавання ідіоматичних виразів, проте залишається питання щодо впливу частотності вживання фразем на ефективність моделей. Це ставить за мету вивчення взаємозв'язку між корпусною частотністю та точністю автоматичного розпізнавання, що є актуальним для лінгвістичних досліджень і розробки інноваційних систем обробки природної мови.

Метою статті є визначення впливу частотності вживання фразеологічних конструкцій у корпусі ГРАК на ефективність їх розпізнавання великими мовними моделями. Для досягнення цієї мети передбачено:

- провести кількісний аналіз частоти появи обраних фразеологічних одиниць у корпусі ГРАК;
- здійснити тестування продуктивності кількох сучасних LLM у завданнях ідентифікації та інтерпретації цих фразем;
- встановити кореляційний зв'язок між частотними параметрами та успішністю моделі у фразеологічному розпізнаванні;
- узагальнити потенційні шляхи удосконалення алгоритмів обробки природної мови з урахуванням результатів дослідження.

Аналіз останніх досліджень і публікацій, у яких започатковано розв'язання означеної проблеми і на які опирається автор. Теоретико-методологічною основою дослідження стали праці Г. Ситар, зокрема дисертація «Синтаксичні фразеологізми в українській мові: структурно-семантичний, прагматичний і прикладний виміри» (2017), а також дисертація «Типологія синтаксично нечленованих одиниць» (2017) М. Личук, документація Sketch Engine «Мова корпусних запитів – CQL (Corpus Query Language)», стаття В. Степанова, А. Могилюк «Потенціал корпусних технологій у вивченні фразеологізмів та сленгізмів» (2025).

Виклад основного матеріалу дослідження з обґрунтуванням авторських положень. На основі словника фразеологізмів української мови Національної академії наук України (Київ, 2003) обрано шість фраземних конструкцій: *козяча ніжка, обголити лоба, важити головою, куди вітер дме, водити за чуприну, баш на баш*. А також запропоновано три найуживаніші великі мовні моделі: Claude sonnet4, Gemini, GPT 4o-mini.

Промт:

Вам нададуть речення.

Ваше завдання – виконати такі дії:

Крок 1: Визначте ідіоматичний вираз (фразеологізм), наявний у поданому реченні.

Крок 2: Визначте його значення та дайте пояснення (значення).

Крок 3: Перефразуйте речення, усуваючи ідіому, але зберігаючи початкове значення (перефразоване речення).

Дотримуйтеся наведеного нижче формату виводу:

Вхідне речення: [Вхідне речення]

Фразеологізм: [Ідентифікована ідіома]

Значення: [Значення та пояснення]

Перефразоване речення: [Перефразоване речення]

Ось речення для аналізу [Речення тут]

Речення для аналізу в промти взяті з корпусу ГРАК (1), на кожен фраземну конструкцію було по два речення з корпусу для забезпечення репрезентативності роботи моделі та уникнення зарахування галюцинації моделі за системну коректну відповідь. У подальшому для покращення якості дослідження має бути використано мінімум п'ять речень із різних джерел, які не є частиною корпусу ГРАК.

За результатами дослідження отримано такі показники від моделей:

Критерії	Claude sonnet4	Gemini	GPT 4o-mini
Визначення меж фраземної конструкції	100%	81.8%	81.8%
Значення фраземної конструкції	63.6%	72.7%	45.5%

Перефразоване речення без використання фраземної конструкції та втрати змісту	63.6%	45.5%	18.2%
---	-------	-------	-------

Де 100% – це 12 з 12 правильних відповідей від великої мовної моделі. Зіставлено частотності вживання фраземних конструкцій у корпусі ГРАК: баш на баш – 41 вживання – 41 фраземне; важити головою – 65 уживань – 55 фраземних; куди вітер дме – 33 вживання – 20 фраземних; козяча ніжка – 7 уживань, 3 фраземних; обголити лоба – 1, фраземне; водити за чуприну – 1, фраземне – у словнику.

Для пошуку фраземних конструкцій у корпусі використано простий пошук, у перспективі для покращення дослідження застосовано розширений пошук, до прикладу, фраземну конструкцію “баш на баш” шукати за такою моделлю (1:[tag=»noun.*»] []{0,3} 2:[] within < s/>) & 1.lemma=2.lemma (6).

Висновки й перспективи подальших розвідок у визначеному напрямі.

Під час дослідження виявлено різну чутливість сучасних мовних моделей до частотності вживання фразеологічних конструкцій у корпусі ГРАК. Для Claude встановлено, що ефективність розпізнавання фразеологізмів не має статистично значущого зв'язку із частотою їх уживання: низько- та високочастотні одиниці розпізнаються зі схожою результативністю. Натомість для моделі Gemini підтвердилася гіпотеза про те, що менша частотність фразеологізму в корпусі корелює з гіршим розумінням ідентифікаційних завдань, тобто моделі складніше розпізнавати рідкісні фраземні конструкції. Цікаво, що для ChatGPT, попри подібний розподіл результатів із моделлю Gemini, статистично значущої кореляції між частотністю вживання та точністю розпізнавання також не зафіксовано. Це свідчить про ймовірно відмінні принципи обробки фразеологічного матеріалу між модельованими архітектурами, а також висвітлює різні лінгвістичні стратегії засвоєння стереотипізованих мовних одиниць.

Список використаних джерел

- Генеральний регіонально анотований корпус української мови (ГРАК) / М. Шведова, Р. фон Вальденфельс, С. Яригін, А. Рісін, В. Старко, Т. Ніколаєнко, А. Лукашевський та ін. Київ, Львів, Єна, 2017–2025. uacorporus.org.
- Ситар Г.В. Синтаксичні фразеологізми в українській мові: структурно-семантичний, прагматичний і прикладний виміри : дис. ... д. філол. наук : 10.02.01 – українська мова; 10.02.21 – структурна, прикладна та математична лінгвістика. Київ, 2017. 516 с.
- Личук М.І. Типологія синтаксично нечленованих одиниць: дис. ... д. філол. наук : 10.02.01 – українська мова. Київ, 2017. 475 с.

- Степанов В., Могилюк А. Потенціал корпусних технологій у вивченні фразеологізмів та сленгізмів. Філологічні трактати 17 (1) 2025. С.63–77.
- Білоноженко В. М., Гнатюк І. С., Дятчук В. В., Неровня В. Н., Федоренко Т. О. Словник фразеологізмів української мови. Київ : Наукова думка, 2003. 1104 с.
- SketchEngine. 2025. <https://www.sketchengine.eu/documentation/>

References

- Bilonozenko V. M., Hnatiuk I. S., Diatchuk V. V., Nerovnia V. N., Fedorenko T. O. Slovnyk frazeolohizmiv ukrainskoi movy. Kyiv : Naukova dumka, 2003. 1104 s.
- Generalnyi rehionalno anotovanyi korpus ukrainskoi movy (HRAK) / M. Shvedova, R. fon Valdenfels, S. Yaryhin, A. Rysin, V. Starko, T. Nikolaienko, A. Lukashevskiy ta in. Kyiv, Lviv, Yena, 2017–2025. uacorus.org
- Lychuk M.I. Typolohiia syntaksyschno nechlenovanykh odynts: dys. ... d. filol. nauk : 10.02.01 – ukrainska mova. Kyiv, 2017. 475 s.
- SketchEngine. 2025. <https://www.sketchengine.eu/documentation/>
- Stepanov V., Mohylnyk A. Potentsial korpusnykh tekhnohii u vyvchenni frazeolohizmiv ta slenkhizmiv. Filolohichni traktaty 17 (1) 2025. S.63–77.
- Sytar H.V. Syntaksyschni frazeolohizmy v ukrainskii movi: strukturno-semantychnyi, prahmatychnyi i prykladnyi vymiry : dys. ... d. filol. nauk : 10.02.01 – ukrainska mova; 10.02.21 – strukturna, prykladna ta matematychna linhvistyka. Kyiv, 2017. 516 s.

RECOGNITION OF UKRAINIAN PHRASAL CONSTRUCTIONS IN LARGE LANGUAGE MODELS: A COMPARATIVE STUDY OF CHATGPT, GEMINI AND CLAUDE

Yelyzaveta Grygorieva, third-year student (bachelor's degree), majoring in 'Philology' at the Faculty of Ukrainian Philology, Culture and Arts of the Borys Grinchenko Kyiv Metropolitan University.

The article is devoted to researching the ability of modern large language models (LLMs) – ChatGPT, Gemini, and Claude – to recognise Ukrainian phraseological constructions in texts from the GRAK corpus. A quantitative analysis of the frequency of selected phraseological units and empirical testing of models were conducted according to three criteria: identification of phrase boundaries, interpretation of its meaning, and semantically correct paraphrasing. The results indicate the heterogeneity of linguistic strategies and the different sensitivity of models to the frequency of phraseological units, which has important implications for the further improvement of natural language processing algorithms in the Ukrainian language environment.

Keywords: phraseological construction, phraseological unit, large language models, Ukrainian language, GRAK corpus, phrase recognition, ChatGPT, Gemini, Claude.