

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Київський національний університет імені Тараса Шевченка

Навчально-науковий інститут філології
Катедра української мови та прикладної лінгвістики

Ідеографічна база даних фразеологізмів української мови

Кваліфікаційна робота

освітнього ступеня «бакалавр»
за спеціальністю 035 «Філологія»,
спеціалізацією 035.10 «Прикладна
лінгвістика»,
галузі знань 03 «гуманітарні науки»
ОПП «Прикладна (комп'ютерна)
лінгвістика та англійська мова»
студентки IV курсу

Богдани ІВАХНЕНКО

Науковий керівник:

доцент кафедри української мови
та прикладної лінгвістики,
кандидат філологічних наук,
Оксана ЗУБАНЬ

«Допущено до захисту»

Протокол № 11 засідання кафедри
української мови та прикладної лінгвістики
ННІФ від 01.06.2023

Завідувач кафедри _____ **Сергій Різник**

Київ — 2023

ЗМІСТ

ВСТУП.....	3
РОЗДІЛ 1. ТЕОРЕТИЧНІ ЗАСАДИ ТА СПОСОБИ СИСТЕМАТИЗАЦІЇ ФРАЗЕОЛОГІЗМІВ У КОМП'ЮТЕРНІЙ ЛЕКСИКОГРАФІЇ.....	6
1.1. Термінологічна визначеність та типологія ідіоматичних синтагм.....	6
1.2. Фразеологізми у сучасній тлумачній та семантичній лексикографії.....	19
1.2.1. Фразеологізми у тлумачних словниках.....	21
1.2.2. Семантичні словники.....	24
1.2.3. Ідеографічне моделювання семантики фразеологізмів.....	30
1.3. Комп'ютерна фразеографія та автоматичне оброблення фразеологізмів.....	40
1.3.1. Корпусні дослідження фразеологізмів.....	40
1.3.2. Укладання комп'ютерних фразеологічних словників.....	41
1.3.3. Застосування фразеографічних баз даних в машинному перекладі.....	49
1.4. Висновки до розділу і робочі поняття.....	61
РОЗДІЛ 2. УКЛАДАННЯ ІДЕОГРАФІЧНОЇ БАЗИ ДАНИХ ФРАЗЕОЛОГІЗМІВ.....	63
2.1. Інфологічна модель БД фразеологізмів.....	63
2.2. Даталогічна модель БД фразеологізмів.....	67
2.3. Програмне забезпечення.....	75
2.3.1. Програма для завантаження словника.....	75
2.3.2. Програма для парсингу статей словника і укладання бази даних.....	78
2.3.2. Програма інтерфейсу користувача.....	83
2.4. Інтерфейс користувача: пошуковий та класифікаційний потенціал.....	94
2.5. Семантико-ідеографічний опис фразеологізмів у БД.....	100
2.6. Перспективи розвитку бази даних.....	103
ВИСНОВКИ.....	105
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	107
СПИСОК ДЖЕРЕЛ ЛЕКСИКОГРАФІЧНОГО МАТЕРІАЛУ.....	118
ДОДАТКИ.....	123
Додаток 1. Код програми-скрейпера.....	123
Додаток 2. Код програми для парсингу словника.....	123
Додаток 3. Ідеографічна база даних фразеологізмів.....	123
Додаток 4. Код інтерфейсу.....	123

ВСТУП

Що таке фразеологізми? Найбільш простою, інтуїтивно зрозумілою відповіддю на це питання було б: *“вирази, значення яких не можна логічно вивести зі значень окремих слів, які його утворюють”* (тут і далі переклад мій — Б.І.) [54]. Подібне визначення повсюдно зустрічається в наукових статтях і популярних текстах. І власне це неспівпадіння форми і значення робить обробку фразеологізмів у комп’ютерній лінгвістиці викликом для дослідників.

Дані є важливою частиною будь-яких проєктів, що покликані розв’язувати сучасні задачі комп’ютерної лінгвістики. Вони, насправді, є визначальними для якості продукту часом більше, ніж алгоритми [97]. Обробка фразеологізмів потребує великої кількості лінгвістичного матеріалу. У завданні укладання словників комп’ютерна лінгвістика значною мірою спирається на корпуси. Використання корпусів дозволяє уникнути суб’єктивності при складанні дефініцій, це спосіб зібрати реальні контексти використання того чи іншого слова чи фразеологізму.

За версією сайту Statista [Statista] та з розподілу мов із останнього збору даних Common Crawl [Common Crawl] українська мова є низькоресурсною, займаючи 18 і 28 місця відповідно. Це наочно демонструє потребу у збільшенні репрезентації української мови в мережі та сприянні відритості даних.

Структура семантики фразеологізмів є основною проблемою, що розкривається в роботі. Раніше її досліджували Петро Дудик [8], Жанна Краснобаєва-Чорна [14; 15; 16; 17], Наталія Сніжко [27; 28; 29; 30], Анна Середницька [25], Віра Білоноженко [Білоноженко-1, Білоноженко-2], Григорій Удовиченко [32], Леонід Авксентьєв [1], Микола Алефіренко [2], Віктор Ужченко [33; 34], Дмитро Ужченко [34], Сергій Фокін [12], Леся Коцюк [13], Тарас Береза [13, Береза], Володимир Калашник і Жанна Колоїз [Калашник], Микола Коломієць

і Євген Регушевський [Коломієць], Федір Медведєв [19], Ірина Патен [20], Ігор Мельчук [89], Віктор Виноградов [5], Розамунд Мун [92; 93], Ентоні Кові [49; 50], Наталія Амосова [3], Борис Ларін [18], Хосе Луїс Рохас Діас [55], Петро Редін [24], Юрій Прадід [22; 23], Шарль Баллі [41], Пітер Марк Роже [Roget Original, Roget Gutenberg], Герардо Сьєрра [115], Алістаір Кеннеді [76; 77], Стен Шпакович [77], Кристіан Феллбаум [58; 59], Франтішек Чермак [47], Елізабет Пірайнен [100] та інші.

Об'єктом вивчення є фразеологізми української мови, а предметом — ідеографічна структура фразеологічної системи української мови семантики фразеологізмів.

Мета — укласти ідеографічну базу даних фразеологізмів української мови.

Завдання:

1. проаналізувати літературу й визначити термінологічну та класифікаційну базу дослідження фразем;
2. вивчити існуючі лексикографічні та текстові матеріали, в яких систематизовано фразеологізми;
3. оглянути способи обробки фразеологізмів у комп'ютерній лінгвістиці;
4. визначити лексикографічну концепцію та практичне використання ідеографічного словника фразеологізмів;
5. створити електронний ідеографічну базу даних фразеологізмів;
6. проанотувати фразеологізми в базі даних за семантично-ідеографічним принципом;
7. проаналізувати анотовану базу даних;
8. окреслити можливості розвитку бази даних та її застосування у системах автоматичного оброблення природної мови.

Методами дослідження є метод комп'ютерного моделювання, прийом семантичної класифікації, метод ідеографічного опису семантики.

Новизна полягає у створенні вільно доступної бази даних фразеологізмів української мови ідеографічного характеру, яку можна буде використовувати для реалізації задач автоматичної обробки природної мови.

Результат роботи у вигляді ідеографічної бази даних фразеологізмів має перспективи використання в актуальних аспектах автоматичного оброблення природної мови, де постає завдання визначення фразеологізмів.

Матеріал дослідження: фразеологізми електронної версії словника української мови в 11 томах [СУМ-11] і синоптична схема тезауруса Роже [Roget Gutenberg].

Структура кваліфікаційної роботи бакалавра.

У [першому розділі](#) досліджено термінологічну варіативність фразеологізмів, їхню семантичну та ідеографічну типологію (і зокрема здійснено спробу адаптувати синоптичну схему тезауруса Роже [Roget Gutenberg] для української мови), оглянуто семасіологічні та ономасіологічні словники на наявність фразеологізмів і спосіб їх представлення, способи автоматичної обробки фразеологізмів у комп'ютерній лінгвістиці і можливе застосування фразеологічних баз даних у машинному перекладі.

У [другому розділі](#) викладено етапи укладання ідеографічної бази даних фразеологізмів української мови, детально описано роботу програм, які було створено для реалізації цього завдання, а також процес семантичного та ідеографічного опису фразеологізмів, проаналізовано отриману базу даних і окреслено можливості її розвитку й застосування.

Загалом у роботі опрацьовано більше 130 джерел і більше 40 джерел лексикографічного матеріалу.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ЗАСАДИ ТА СПОСОБИ СИСТЕМАТИЗАЦІЇ ФРАЗЕОЛОГІЗМІВ У КОМП'ЮТЕРНІЙ ЛЕКСИКОГРАФІЇ

1.1. Термінологічна визначеність та типологія ідіоматичних синтагм

Українською найширше вживаним є термін «фразеологізм», він зустрічається і в словниках [СУМ-11, 638 (т. 10); Єрмоленко, 194], і в наукових працях різними мовами [8; 17; 5, 62; 119]. Він лаконічний та зрозумілий. Тим не менш, зустрічаються й інші терміни. Багато науковців надають перевагу «фразеологічній одиниці». Його вживають Наталія Амосова, Віктор Виноградов, Сильвія Моліна Плаза, Тетяна Бушнак та інші [3; 5; 102; 44; 16; 8]. Жанна Краснобаєва-Чорна стверджує, що він походить із «locution phraséologique» — терміна, що його ввів Шарль Баллі у 1905 році, започаткувавши фразеологію як науку [16], хоча мені вдалося знайти лише «locution phraséologique», яке скоріше перекладається як «фразеологічний вираз» (якщо дослівно) [41].

Також можна зустріти терміни «фразеологічний зворот» [Єрмоленко], «фразема» /«*phraseme*»/ [8; 17; 112; 18; 89], «ідіома» /«*idiom*»/ [17; 87; 65; 84; 88] і «фраземозв'язане словосполучення» [32].

Українсько-англійський словник лінгвістичної термінології подає такі відповідники «фразеологізму» для англійської мови: «*idiom*», «*phraseologism*», «*phraseological unit*». «*Idiom*» є найбільш розповсюдженим серед мовців, що видно зі словників, у яких присутній лише цей варіант з усіх можливих [Cambridge; Merriam-Webster]. Окрім цих термінів, зустрічаються ще «*idiomatic expression*» [56; 91; 87; 61; 84; 88], «*fixed expression*» [37], «*multi-word expression*» [62], «*multiword unit*» [48], «*periphrase*» [52], «*isolation*» [99; 74], і навіть «*collocation*» [38; 51; 74]. Буває, автори вживають декілька термінів в одному тексті взаємозамінно (наприклад, Ельмар Шафрот вживає одночасно «*idiom*», «*idiomatic phrase*», «*idiomatic expression*», «*phraseological unit*» і «*phraseme*» [72]).

У цій роботі у загальному випадку будемо користуватися терміном «фразеологізм».

Із розмаїття термінів для загальної категорії можна зробити висновок, що у фразеології вони ще не усталені. Навіть яскравіше це видно у роботах на межі фразеології та інших наук. Наприклад, у дослідженні Керол Торіс (психологія), де розбирається швидкість сприйняття сенсорних фразеологізмів, фразеологічні єдності ототожнено із загальною категорією «idioms» [125]. Також подібні випадки наводить Тетяна Бушнак [44].

З багатьох властивостей фразеологізмів найчастіше виокремлюють такі:

- ідіоматичність /idiomaticity/ — власне незалежність значення фразеологізму відносно значення його частин;
- відтворюваність /repetition/ — поширене вживання сталої форми;
- незмінність /fixedness/ — вирази мають певну варіативність, але вона обмежена, а одна необережна зміна може призвести до втрати ідіоматичності (Хосе Луїс Рохас Діас цитує Леонор Руїс Гурільо і Ентоні Кові [109; 49], а Жанна Краснобаєва-Чорна — Тею Шиппан [111]).

За Борисом Ларіним, якого цитує Жанна Краснобаєва-Чорна, ці властивості можуть з'являтися у вільних словосполучень завдяки втраті реалії, яку вони позначають, спрощенню через втрату подробиць та фіксації граматичної форми, за якою закріплюється додаткове значення або метафоризації [11; 17].

Фразеологізми, очевидно, містять більше, ніж одне слово (майже всі з них є словосполученнями і лише до фразеологічних виразів прийнято відносити речення), тож до них можна застосовувати і синтаксичний аналіз. Навіть якщо значення фразеологізму залишається прихованим для комп'ютерних програм-парсерів, то синтаксична структура є зрозумілою, за винятком унікальних слів (як *байдики*).

В англійськомовній літературі також зустрічаються дослідження multiword units (MWUs) — вони включають у себе фразеологізми, але зазвичай не лише їх, наприклад, туди входить і ідіома *kick the bucket*, і частина prepositional phrase *in the middle of* [48]. Це більш подібно на збірку n-грамів, ніж на чітко окреслену групу лексики.

Серед науковців, що розробляли класифікацію за будовою, Жанна Краснобаєва-Чорна згадує Віру Білоноженко, Ірину Гнатюк, Федора Медведєва, Ніколая Шанського [16].

Щодо інших класифікацій, то їх розлогий набір описано у Жанни Краснобаєвої-Чорної: семантично-граматична, експресивно-стилістична, класифікація за сферою вживання, структурно-семантична, (відома в англійськомовній літературі як «тематична композиційність» /thematic composition/ [66]), генетична, з погляду переосмислення компонентів, за дериваційною базою, за лексичним складом, функціональна класифікація, за оточенням і на основі методу фразеологічної аплікації [16]. Тут ми їх усі розглядати не будемо, але розберімо деякі підходи до аналізу, які важливі для цієї роботи.

Семантична спаяність фразеологізмів

Лексична неподільність є характерною ознакою фразеологізмів. За ступенем її вираженості лінгвісти виокремлюють від двох до п'яти типів, зокрема можна виділити класифікації Віктора Виноградова для російської мови [5] та Ігоря Мельчука для англійської [89]. За словами Хосе Луїса Рохаса Діаса, Леонор Руїс Гурільо і Ентоні Кові стверджують, що у класифікаціях фразеологізмів за семантичною спаяністю різного авторства майже завжди присутні такі елементи: повністю фразеологізовані компоненти, нефразеологізовані стійкі вирази і вислови, які є реченнями — прислів'я і приказки [109; 49; 54].

Отже, розберімося детальніше:

- найбільш злиті за значенням фразеологізми:

Їхнє значення взагалі не має нічого спільного зі значеннями складників у них. Вони часто містять унікальні слова, які не зустрічаються в іншому оточенні — архаїзми або okazionalizmi. Набір слів, а часом і їхня форма, є сталими. Приклади: *бити байдики* /байдикувати/, *сон рябої кобили* /нісенітниці/, *зостатися на цідильці* /попастися на брехні/ [СУМ-11].

Петро Дудик називає їх ідіомами [8]. В аналізі Віктора Виноградова для російської мови їх названо «фразеологічними зрощеннями» /фразеологические сращения, phraseological fusion/ [5; 55]. Лінгвістичний словник подає обидва варіанти [Єрмоленко].

Англійською мовою зустрічається переважно термін «idiom» [93], також є «pure idiom» [49], «full idiom», яку Ігор Мельчук визначає формалізовано так: “*An idiom AB is a full idiom iff its meaning does not include the meaning of any of its lexical components: ‘AB’ ⊄ ‘A’ and ‘AB’ ⊄ ‘B’*” /Ідіома АВ є повною ідіомою тоді і тільки тоді, коли її значення не включає значень жодного з її лексичних складників ‘АВ’ ⊄ ‘А’ and ‘АВ’ ⊄ ‘В’/ [89].

У цій роботі на позначення таких семантично зрощених висловів буде вжито термін «ідіома».

- фразеологізми, значення яких частково мотивоване значенням їхніх складників:

Вони також стійкі, але не до кінця злиті семантично. Важливо, що можуть вживатися і в прямому значенні, хоча образ часто абсурдний (див. Малюнок 1.1). Приклади: *довгий язик* /про балакливу людину, що говорить зайве, або про її вдачу/, *на роздоріжжі* /у стані нерішучості, важких роздумів, вагань/, *побувати у бувальцях* /багато в житті побачити, зазнати/ [СУМ-11].

Часто вживано термін «фразеологічна єдність» [8; Єрмоленко]. Російською, за Віктором Виноградовим, — «фразеологическое единство» /фразеологічна

єдність, phraseological unity/ [5]. Не всі науковці виокремлюють цей тип фразеологізмів: Шарль Баллі, Григорій Удовиченко, Леонід Авксентьев та Ігор Мельчук цього не роблять, об'єднуючи їх із ідіомами [41; 33; 1; 89].



Малюнок 1.1. Ілюстрація фразеологічної єдності з англійської мови «hold your horses» [Nahessy]

Зустрічаються також терміни «ідіоматичний вираз» /idiomatic locution/ [93], «частково предикативний фразеологізм» /частично предикативный фразеологизм/ [3], «метафоричне словосполучення» [11]. Ентоні Кові використовує «образна ідіома» або «метафорична ідіома» /figurative idiom/ [49].

У роботі буде використано термін «фразеологічна єдність».

- фразеологізми, у яких одне слово вжито у прямому значенні:

Через вільне значення одного зі складників ці фразеологізми доволі легко розгадати, особливо в контексті. Приклади: *вовчий **апетит*** /дуже великий **апетит**/, *застригти **вухами*** /почати рухати **вухами** (про тварин)/, *понакривати **столи*** /поставити на **столи** їжу/ [СУМ-11]. Можна помітити, що в дефініціях вжито слова-складники фразеологізмів, що наочно демонструє їхнє вільне

значення. У деяких із них цей складник може змінюватися без втрати висловом ідіоматичності: *голова репається|лопається|лускає* /сильно болить голова/ [СУМ-11]. Також до цієї категорії належать книжні вислови, які є доволі розповсюдженими, як-от *ім [же] ім'я — легион* /когось надзвичайно, дуже багато/ [СУМ-11].

Щодо термінів, то найчастіше використовують «фразеологічне сполучення»: і в словнику лінгвістичних термінів [Єрмоленко], і в роботах [8; 1; 32]. Шарль Баллі вживав «*les series phraseologiques*» /фразеологічний ряд/ [41]. Для російської Віктор Виноградов впровадив «фразеологическое сочетание» /фразеологічне сполучення, *phraseological combination*/ [5]. Також побутують терміни «змінне словосполучення» [18], «фразема» /*phraseme*/ і «фразеолоїд» /*phraseoloid*/ [3], «обмежене словосполучення» /*restricted collocation*/ [49] і «*semi-idiom*»: “*an idiom AB is a semi-idiom iff its meaning 1) includes the meaning of one of its lexical components, but not as its semantic pivot, 2) does not include the meaning of any other component and 3) includes an additional meaning ‘C’ as its semantic pivot: ‘AB’ ⊃ ‘A’, and ‘AB’ ⊄ ‘B’, and ‘AB’ ⊃ ‘C’*” /Ідіома АВ є напів-ідіомою тоді і тільки тоді, якщо її значення: 1) містить значення одного з своїх лексичних складників, що не є *semantic pivot*; 2) не містить значення жодного іншого складника; 3) містить додаткове значення ‘С’ як свій *semantic pivot*: ‘АВ’ ⊃ ‘А’, and ‘АВ’ ⊄ ‘В’, and ‘АВ’ ⊃ ‘С’/ [89].

У роботі буде використано термін «фразеологічне сполучення».

- вислови, компоненти яких поєднані у прямому значенні:

Ці одиниці знаходяться на межі фразеологізмів і колокацій, бо значення їх є доволі прозорим. Вони більшою мірою відображають мовні звички носіїв, ніж переносне значення. Через це деякі дослідники проти зарахування їх до фразеологізмів [18]. Приклади: *набратися хоробрости* /осмілитися діяти певним чином; рішуче взятися за якусь справу/, *всіма кольорами райдуги* /багатьма

різними барвами/, *нехитре діло* /щось нескладне, що легко зробити, здійснити і т. ін/ [СУМ-11].

Петро Дудик вживає термін фразеологізоване словосполучення [8], воно також є в термінологічному словнику [Єрмоленко]. Шарль Баллі назвав їх «*les groupements usuels*» /звичні групи/ [40]. В англomовній літературі можна зустріти термін «*collocation*» [102, 48] та «*open collocation*» [49].

Ігор Мельчук вживає термін «*quasi-idiom*» або «*weak idiom*»: “*An idiom AB is a quasi-idiom, or weak idiom, iff its meaning 1) includes the meaning of both of its lexical components, neither as the semantic pivot, and 2) includes an additional meaning ‘C’ as its semantic pivot: ‘AB’ ⊃ ‘A’, and ‘AB’ ⊃ ‘B’, and ‘AB’ ⊃ ‘C’*” /Ідіома є квазі-ідіомою або слабкою ідіомою тоді й тільки тоді, коли її значення 1) включає значення обидвох (або усіх — Б.І.) її лексичних складників, за умови, що жоден із них не є її *semantic pivot*, і 2) включає додаткове значення ‘C’, що є її *semantic pivot*/ [89].

У цій роботі буде використано термін «фразеологізоване словосполучення».

- вислови-речення:

Вони є реченнями, але словосполученнями і складаються зі слів у прямому значенні. Вони мають стійку традицію вживання у конкретних ситуаціях, тому їх легко розпізнати. Приклади: *за моє жито мене й бито, що легко нажить — легко біжить* [8], *здрастуйте вам у хату* [СУМ-11].

Рік	Автор	Загальна категорія	Злиті незмінні одиниці	Частково мотивовані одиниці	Напів вільні одиниці	Неметафоризовані одиниці	Прислів'я і приказки
1905	Шарль Баллі	locution phraséologique	les unites phraseologiques		les series phraseologiques	les groupements usuels	<i>невідомо</i>
1963	Наталія Амосова	phraseological unit	idiom	частично предикативный фразеологизм	фразема phraseme phraseoloid	<i>немає даних</i>	<i>не визнає</i>
1965	Григорій Удовиченко	фраземозв'язане словосполучення	ідіома		фразеологічне сполучення	<i>немає даних</i>	фразеологічний вираз
1973	Петро Дудик	фразеологізм, фразема	ідіома	фразеологічна єдність	фразеологічне сполучення	фразеологізоване словосполучення	фразеологічний вираз
1977	Борис Ларін	фразема	ідіома	метафоричне словосполучення	змінне словосполучення	<i>проти їх виокремлення</i>	<i>немає даних</i>
1977	Віктор Виноградов	фразеологическая единица /phraseological unit/	фразеологическое сращение /phraseological fusion/	фразеологическое единство /phraseological unity/	фразеологическое сочетание /phraseological combination/	<i>немає даних</i>	фразеологическое единство /phraseological unity/

Рік	Автор	Загальна категорія	Злиті незмінні одиниці	Частково мотивовані одиниці	Напів вільні одиниці	Неметафоризовані одиниці	Прислів'я і приказки
		фразеологізм					
1983	Леонід Авксентьєв	<i>немає даних</i>	ідіоми		фразеологічні сполучення	<i>немає даних</i>	<i>проти їх виокремлення</i>
1998	Розамунд Мун	phrasal lexemes	idiom	idiomatic locution	<i>немає даних</i>	<i>немає даних</i>	<i>немає даних</i>
2001	словник лінгвістичних термінів [Єрмоленко]	фразеологізм, фразеологічна одиниця, фразеологічний зворот	ідіома, фразеологічне зрощення	фразеологічна єдність	фразеологічне сполучення	фразеологізоване словосполучення	фразеологічний вираз
2001	Ентоні Кові	word combinations	pure idiom	figurative idiom	restricted collocation	open collocation	<i>немає даних</i>
2006	Сильвія Моліна Плаза	phraseological unit idiom /expresión	<i>немає даних</i>	<i>немає даних</i>	<i>немає даних</i>	collocation /colocación/	<i>немає даних</i>

Рік	Автор	Загальна категорія	Злиті незмінні одиниці	Частково мотивовані одиниці	Напів вільні одиниці	Неметафоризовані одиниці	Прислів'я і приказки
		idiomática/					
2012	Ігор Мельчук	phraseme	full idiom		semi-idiom	quasi-idiom, weak idiom	<i>немає даних</i>
2013	Джорджі Колумбус	multiword unit	idiom			(restricted) collocation	idiom

Таблиця 1.1 Порівняння термінів на позначення різних рівнів семантичної спаяності компонентів у хронологічному порядку (частково використано таблицю з Ентоні Кові [49])

Загалом як українською, так і англійською, терміни «прислів'я» /*proverb*/, «приказка» /*saying*/ є розповсюдженими. Петро Дудик і Григорій Удовиченко називають їх «фразеологічними виразами» [8; 32]. Перший до них також зараховує крилаті вислови. Віктор Виноградов вважає їх частиною «фразеологічних єдностей» /*фразеологическое единство*/. Джорджі Колумбус не відокремлює їх із ідіом («*idioms*») [48]. Леонід Авксентьєв проти того, щоб виокремлювати їх як фразеологізми.

У роботі буде використано термін «фразеологічний вираз».

Також цією проблемою, згідно з Жанною Краснобаєвою-Чорною [16], займалися Дмитро Баранник [4], Федір Медведєв [19], Микола Алефіренко [2], Лариса Скрипник [26], Мар'ян Демський [7], а Елахех Рафатбахш і Аліреза Ахмаді [105] наводять таких авторів: Ричард Александр [36], Ентоні Кові з іншими [50], Читра Фернандо [61], Мері МакГі Вуд [127].

У таблиці 1.1 зібрано і узагальнено дані про класифікацію фразеологізмів за семантичною спаяністю компонентів від різних авторів. Її організовано у хронологічному порядку.

Ідеографічний опис фразеологізмів

Також вживається термін «ономасіологічна класифікація» /*onomasiological classification*/, «ідеологічна класифікація» /*ideological classification*/ і часом «тематична класифікація» /*thematic classification*/, але остання переважно не передбачає ієрархічності [12]. Ідеографічний опис полягає у спробі організації лексики відповідно до картини світу. Він узагальнює значення, використовуючи гіпо- і гіперонімічні відношення між словами або вкладання їх значень одне в одного (шляхом визначення диференційних сем у компонентному аналізі), використовуючи спеціальну метамову дескрипторів. Навіть це уже створює підґрунтя для складнощів із узгодженням підходів різних авторів, бо світобачення

у кожного своє, починаючи хоча б із визначення того, що знаходиться в основі всесвіту — Бог, космос чи щось іще. Самі слова теж можуть підпадати під кілька категорій або тем (Сергій Фокін наводить приклад зі словом *каюта*, яке має стосунок до кораблів, але водночас є житлом), або через багатозначність (*совок* як інструмент для прибирання і *совок* як тюрма народів). Останнє якраз актуальне у контексті фразеологізмів.

В українській ідеографії використовується багато різних термінів: фразеологічне поле, фразеосемантична група, фразеосемантична підгрупа, фразеосемантичне поле, фразеотематична, група варіантно-синонімічна група, фразеоідеографічна парадигма, фразеологічна група тощо. Кожен автор використовує свої і всі вони миготять перед очима своєю схожістю. За Жанною Краснобаєвою-Чорною [14], в Україні на 2011 рік існують такі класифікації: Миколи Алефіренка [2], Анатолія Івченка [10], Жанни Краснобаєвої-Чорної [15], Петра Редіна [24], а також Юрія Прадіда [22; 23] та Віктора й Дмитра Ужченків [33; 34], основи яких наведено у таблиці 1.2 і 1.3 відповідно. Спостережено декілька розбіжностей між авторами:

- Рівень узагальнення: у Віктора й Дмитра Ужченків немає одного узагальнюючого поняття, класифікація починається одразу трьох входів, а в Юрія Прадіда ідеографічні поля зібрані в архіполе «Всесвіт».
- Терміни: у Юрія Прадіда ідеографічне *поле* є загальнішим за ідеографічну *групу*, натомість у Віктора й Дмитра Ужченків фразеотематичне *поле* входить у фразеотематичну *групу*.
- Визначення місця людини: класифікація Віктора й Дмитра Ужченків є виразно антропоцентричною, бо не тільки є окрема, не підпорядкована нічому, фразеотематична група «Людина», а й фразеотематичні поля є лише в цієї групи і загалом фразеосемантичних полів у них значно більше ніж в

інших групах; натомість у Юрія Прадіда ідеографічна група «Людина» є частиною ідеографічного поля «Жива природа».

Архіполе	Ідеографічне поле	Ідеографічна група
Всесвіт	Жива природа	Флора
		Фауна
		Людина
	Нежива природа	Предмет
		Явище

Таблиця 1.2. Ідеографічна класифікація фразеологізмів Юрія Прадіда [22; 23].

Фразеотематична група	Фразеотематичне поле	Фразеосемантичне поле
Людина	Людина як жива істота	Зовнішній вигляд
		Фізичний стан
		Психічний стан
		Одяг
		Рух
	Людина як розумна істота	Розум
		Емоції, почуття, стан
		Характер, поведінка
	Людина як суспільна істота	Стосунки між людьми
	Абстрактні відношення і поняття	
Простір		
Кількість		
Якість		

Природа		Час
		Предмет

Таблиця 1.3. Ідеографічна класифікація фразеологізмів Віктора й Дмитра Ужченків [33; 34].

У цій роботі для ідеографічного опису буде використано синоптичну схему тезауруса Роже. Вона є загальномовною, а не лише фразеологічною як вищенаведені, представляючи значно ширший набір дескрипторів. У [пункті 1.2.3](#) наведено переклад її перших рівнів на українську на наведені деякі загальні спостереження.

1.2. Фразеологізми у сучасній тлумачній та семантичній лексикографії

Є дві антонімічні пари термінів:

- семасіологічний словник /semasiological dictionary/ — ономасіологічний словник /onomasiological dictionary/;
- тлумачний словник /forward dictionary/ — обернений словник /reverse dictionary/ [131].

Перша пара є загальноприйнятною, а друга є поширеною в англійській літературі, але викликає непорозуміння в українській традиції, бо в нас поняття «оберненого словника» /reverse dictionary/ (також побутує термін «зворотний словник» та інколи «інверсійний» [12]) перш за все асоціюються першим із трьох розумінь цього терміна:

1. слова у словнику сортовані за абеткою, але не з початку слова, а з кінця [11, 12];
2. “словнику якому змінено місцями глосу і лему” [12];
3. словник, у якому пошук виконувало від поняття до слова, тобто від плану змісту до плану вираження [115; 114].

Як на мене, у другому розумінні, “інверсійний” словник усе ще є тлумачним. У контексті електронних словників, форма бази даних залишається такою самою,

змінюється лише зона для пошуку, а тому не вважаю доцільним виокремлювати цей тип. У цій роботі нас цікавить останнє розуміння терміна «обернений словник», але задля уникнення непорозумінь будемо говорити «ономасіологічний словник» та «ідеографічний словник».

Згідно з Сергієм Фокіним, тематичний словник, ідеологічний словник, тезаурус, лінгвістичний атлас — це також синонімічні терміни [12, 44] (хоча Елізабет Пірайнен зауважує, що зарахування тематичних словників до ономасіологічних є помилковим [100]). Герардо Сьєрра наводить ще такі: «семантичний словник» /semantic dictionary/ (Malkiel), «концептуальний слоник» /conceptual dictionary/ (Rey), «словник з точки зору мовця» /speaker-oriented lexicon/ (Mallinson), «тематичний словник» /thematic wordbook/ (McArthur), «класифікатор» /nomenclator/ (Riggs), а також зараховує до них словники синонімів та антонімів [115].

Одне із досліджень присвячене процесу укладання ономасіологічних словників [114] подає типологію «обернених словників» (див. схему 1.1), розглядаючи сучасні підходи до укладання словників, оцінювання їх якості і можливі майбутні підходи. З сучасних підходів укладання словників наведено такі:

- інформаційний пошук;
- побудова графів;
- відтворення ментального лексикону;
- використання векторів (вкладення слів /word embeddings/);
- тренування нейронних моделей.

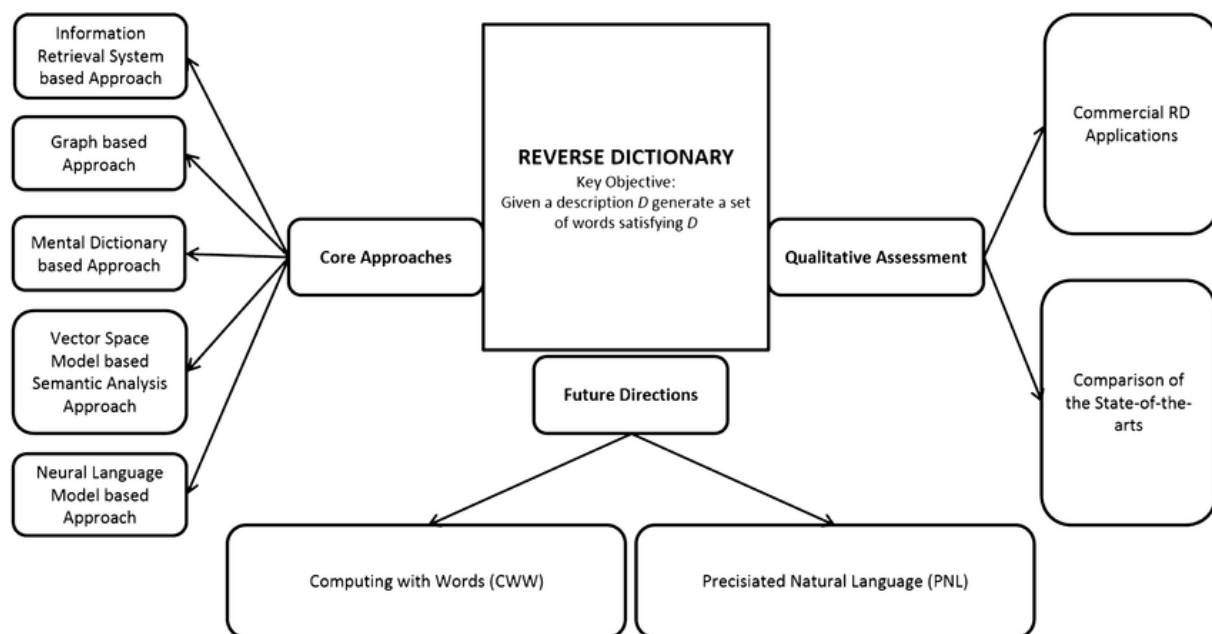


Схема 1.1. Типологія «обернених словників» за Бушрою Сіддіке та Суф'яном Бегом [114].

Ономасіологічні словники полегшують життя читачам, бо побудовані з урахуванням відношень між одиницями, а це перший крок до створення словників-мереж, які за структурою наближались би до структури людського мислення [123]. Більше того, уже існують проекти штучного інтелекту, що базовані на ідеї зв'язків і використання їх для відтворення мисленнєвого процесу, наприклад, Сус Project [120]. Саме такий спосіб є значно вигіднішим для організації фразеологізмів та обходить проблему невідповідності форми і значення [64].

1.2.1. Фразеологізми у тлумачних словниках

Найрозповсюдженішими словниками є загальні тлумачні. Серед звичайної лексики у них трапляються і фразеологізми. Наприклад, у Словнику української мови у 11 та 20 томах [СУМ-11, СУМ-20], їх можна знайти за статтями їхніх слів-складників. Це ускладнює пошук фразеологізму, бо потрібно не тільки конче

знати його форму вираження, а ще й угадати, за яким словом він записаний у словнику. На малюнку 1.2 наведено приклад першої статті зі словника.

A¹, *невідм., с.* Перша літера українського алфавіту на позначення голосного звука “а”.

◇ **(1) Від а до я (до зет)** – від початку до кінця; все повністю. *Асоціація міст України видала практичний посібник “Від влади – до громади, від громади – до влади”, в якому дано конкретні інструкції та пояснення усіх тонкощів від а до я, як створити інформаційно-консультативні центри (з газ.); Прочитати книжку від а до я.*

Малюнок 1.2. Перша стаття Словника сучасної української мови у 11 томах [СУМ-11].

Порівняймо із Академічним словником сучасної чеської мови, який укладає департамент сучасної лексикології та лексикографії при інституті чеської мови Чеської академії наук [Čeština]. На малюнку 1.3 подано сторінку першої статті у словнику (останньою представлено одиницю, яка є відповідником вказаного вище «від а до я»).

Також існує багато тлумачних словників, що присвячені винятково фразеологізмам. Приклади для української мови:

- Фразеологічний словник української мови Віри Білоноженко та колеги, що претендує на першість в академічному описі українських фразеологізмів. Виданий 1993 року у двох книгах [Білоноженко-1, Білоноженко-2].
- Фразеологічний словник української мови з близько 2500 фразеологізмами, виданий 1998 року [Ужченки].
- Словник фразеологізмів української мови 2003 року. Він містить “7922 найуживаніші фразеологічні одиниці сучасної української мови” [Винник].

Обсяги таких словників дозволяють уявити приблизну кількість фразеологізмів у мові загалом і орієнтуватися на неї при зборі даних для автоматичної обробки.

Приклади фразеологічних словників для англійської мови:

Малюнок 1.3. Перша стаття Академічного словника сучасної чеської мови [Čeština].

a I

[ʔa:]

(neskl.)

podstatné jméno rodu středního

1. i A, á, Ā

první písmeno abecedy • (krátká, dlouhá) samohláska reprezentovaná tímto písmenem:

*malé psací a
velké tiskací A
vyslovit dlouhé á*

2. častěji A

co je označeno písmenem a, A, zprav. první místo v pořadí • zkratka nebo značka něčeho, co začíná na písmeno a, A:

*tým / družstvo A
vitamin A / A vitamín
první A (1. A) třída ve škole
krevní skupina A
pavilon A
viz bod a)
a ar
adm. ā. ā. ad acta
fyz. A ampér*

3. častěji A

co patří do kategorie A, představuje nejvyšší úroveň, kvalitu:

*Z matematiky dostal známku A.
Špičkové lednice mívají energetické hodnocení A nebo B.*

4. i A, hud.

šestý tón základní sedmitónové stupnice • příslušná nota:

*stupnice / tónina A dur
stupnice / tónina a moll
nota A*

□ **komorní A / a** hud.

1. základní tón v hudební akustice

2. určující tón k ladění hudebních nástrojů v orchestru

◇ **když se řekne a / á, musí se / mělo by se říct i b / bé**

co se začne říkat, naznačí, to by se mělo také dokončit, doříct:

Když se řekne a, musí se říct i b – popisovanými opatřeními to nemůže skončit, těmito to teprve začne.

Když se řekne á, mělo by se říct bé. Jakým způsobem stát vymůže od dlužníků peníze?

◇ **neříct ani a / á, ani b / bé**

neříct na nějaký návrh ap. nic určitého, vůbec se k něčemu nevyjádřit:

*Už jsem to nadhodila před mužem, ale zatím neřekl ani á, ani bé.
Kouč zatím neřekl ani a, ani b, protože by to bylo předčasné.*

◇ **od a / á (až) do z / zet**

◇ **od a / á (až) po z / zet**

od začátku do konce, celé • důkladně, do všech podrobností:

*Noviny jsem přečetl od A do Z.
O zákazníka se postaráme od A až do Z.
Táta mě zná od A do Z.
Je to od á do zet vyřhané.
Měl podnik organizačně na starosti od A po Z.
Je to od á až po zet ruční práce.
Můj muž byl velký houbař, znal houby od a až po zet.*

Poznámka Méně často se používá i pravopisná podoba á, která vychází z výslovnosti.

● «Oxford Dictionary of Idioms» четвертого видання 2020 року містить близько 10 тисяч одиниць [Oxford-4]. Для порівняння, друге видання 2004 року містило близько 5 тисяч фразеологізмів [Oxford-2].

● «Collins Cobuild Dictionary Of Idioms», що видається з 1995 року (останнє видання 2020 року) [Collins]. Він позиціонується як корпусо-орієнтований та міжнародний, покриваючи різні діалекти англійської.

Перекладні словники фразеологізмів є манною небесною для перекладачів. Саме для перекладу фразеологізми становлять найбільшу проблему, оскільки значно вповільнюють процес та можуть призвести до часткової втрати смислу або авторської ідеї. Такі словники навряд чи стануть в нагоді у синхронному усному перекладі

без підготовки, але у письмовому — так. Прикладом такого словника може бути Німецько-італійський ідіоматичний словник на 35 тисяч одиниць, виданий 2009 року [Fenati], що також є й італійсько-німецьким. Ще один зразок: грузинсько-новогрецький Онлайн словник фразеологізмів [86]. Він містить 12 тисяч одиниць і подає тлумачення, ілюстрації та іншомовні відповідники. Також Леся Коцюк і Тарас Береза згадують про «Англо-український фразеологічний словник» Костянтина Баранцева [Баранцев]. Вони ж самі уклали тематичний «Двомовний тезаурус фразеологічних одиниць», зібравши фразеологізми навколо неієрархічної системи понять [13].

1.2.2. Семантичні словники

Словники синонімів та антонімів

Словники синонімів та антонімів є перехідною ланкою до ідеографічних словників. У цій категорії можна згадати «Словник фразеологічних антонімів української мови» вийшов друком 2004 року і містить 585 протиставлень [Калашник].

Визначення у цьому словнику сформовано лаконічно, вони чітко виражають протиставлювані поняття, натомість прикладів вживання “як у kota сліз”. На малюнку 1.4 зображено приклад розгорнутої статті з нього [Калашник].

Не оминімо увагою і «Сучасний англо-український словник живої мови» /The Contemporary English-Ukrainian Idiom Activator/ Тараса Берези 2012 року — це фразеологічний словник, що містить 3000 «чистих ідіом» (термінологія автора) — ідіом і фразеологічних єдностей [Береза]. Цей словник текстоорієнтований і включає тільки ті одиниці, які зустрілися в онлайн медіа, що були поширеними під час укладання словника. Приклади вживання взято із «Британського національного корпусу» /British National Corpus/ [BNC], «Корпусу сучасної американської англійської мови» /Corpus of Contemporary American English/ [COCA] та тих же

медіа. У кожній статті є посилання на синоніми (“*див. також*” /*see also*/) та антоніми (“*антоніми*” /*antonyms*/), але через те, що фразеологічні варіанти не згруповано, ці посилання часто ведуть на дуже подібні статті. Так само і з міжмовними відповідниками: вони подані навпроти українських ідіом 1:1 (навіть якщо їх могло бути 3:1, це все одно три різні статті формату 1:1), а отже часто зустрічаються фруструючі посилання з дублями, як у прикладі на малюнку 1.5.

531. ХОЧ ЛОПАТОЮ ГОРНІ // КОТОВІ НА СЛЬОЗИ НЕМА

дуже багато

аж кишить
до біса
кури не клюють
на волівій шкурі
не списати
ні проїхати ні пройти
хоч греблю гати
як за гріш маку
як піску морського

Отоді Клим і приглянувся
до Онисьчиного зали-
цяльника. Дуже приваб-
ливий, працьовитий... А
культури – хоч лопатою
греби.
(М. Рудь)

дуже мало

не дуже густо
кіт наплакав
із заячий хвіст
на заячий скік
з горобину душу
як у ката сліз
на макове зерно
крапля в морі

В них там орної землі і котові
на сльози нема, єдиний за-
робіток біля колії.
(І. Муратов)

Малюнок 1.4. Приклад розгорнутої статті Словника фразеологічних антонімів української мови [Калашник].

See also: Do a bad turn for sb/sth; Do sb/sth a bad turn.

Див. також: Зробити ведмежу послугу; Зробити ведмежу послугу.

Antonyms: Bring grist to the mill of sb/sth; Do sb a great turn; Play into sb's hands; Play into the hands of sb/sth.

Антоніми: Лити воду на млин; Грати на руку; Грати на руку; Грати на руку.

Малюнок 1.5. Фрагмент зі статті-прикладу Сучасного англо-українського словника живої мови [Береза].

Статті у словнику подано виключно в алфавітному порядку. До кожної дописано «поняття» /*notion*/, яке фразеологізм виражає (наприклад, *рукою подати* — «близькість»).

Тим не менш, ні у словнику Володимира Калашника і Жанни Колоїз, ні у словнику Тараса Берези немає поняттєвого покажчика, що унеможлиблює пошук від ідеї до слова (при цьому маючи алфавітні покажчики для фразеологізмів). Значно логічнішою видається макроструктура «Словника фразеологічних синонімів» Миколи Коломійця та Євгена Регушевського 1988 року, де одиниці згруповані за поняттями у синонімічні ряди [Коломієць]. Зважаючи на відсутність подальшого згрупування одиниць, цей словник можна назвати тематичним. Ілюстративну статтю зі словника подано на малюнку 1.6.

Наприкінці додано алфавітний покажчик для самих фразеологізмів, але він не дублює існуючу структуру словника, а дає додатковий спосіб входу. Самі поняття також подані алфавітно.

У словника Миколи Коломійця та Євгена Регушевського існує електронна версія на проєкті Розум [Розум], щодо якого мені не вдалося знайти жодної інформації про розробників і майже весь час роботи на цим текстом сайт був недоступний.

Вражаючим прикладом є «Чеський фразеологічний словник» 1994 року, який створив Франтішек Чермак [47]. Він не тільки містить близько 16000 одиниць, а ще й подає відповідники декількома мовами: англійською, німецькою, французькою та російською (має синоніми, антоніми, пов'язані вислови і власне відповідники іншими мовами). Детальніше про словник написано у статті Алеся Клегра [80]. На превеликий жаль, не знайдено ні його електронної версії, ні електронної копії.

3

ЗАБУТИ — (перестати пам'ятати, згадувати кого-, що-небудь) вибити [собі] з голови кого, що; викинути з голови (з думки, з пам'яті) кого, що; покинути думку про кого, що; викреслити з свого життя (з свого серця) кого, що; вирвати з [свого] серця (з [своєї] душі) кого, що.

— *О, я хочу мерщій вибить з голови інститутський романтизм.* — сказала Катерина (Нечуй-Левицький, 2, 1965, с. 112); — *Невже ж ти знов хотіла б відходити? — питала вдовиця з тривогою, а zarazом любо.* — *І то в таку непогоду? Вибий собі лиш те з голови, я тебе не пущу* (Кобилянська, 1, 1962, с. 73); [Матн:] — *Раджу, Степо, викинути з голови кавалерів, візьмися ліпше за книжки* (Гуріненко, Днів твоїх небагато, 1984, с. 154); *Літ, може, п'ятнадцять, як пішов [Остап] на Дін. Уже його й з ревізії викинули, не тільки з думки* (Панас Мирний, 2, 1954, с. 38); [Квятковська:] *Я б і з пам'яті її викинула, коли втекла і проміняла не на знать кого* (Старицький, 3, 1964, с. 530); *Треба було покинути думку про якийсь справжнє діло серед свого люду, у рідній країні* (Коцюбинський, 1, 1955, с. 223); *Вона [Раїса] навіть рахувала день свій з тої хвилі, коли могла врешті бути з о. Василем. Решту ж вона охоче викреслювала з свого життя* (Коцюбинський, 1, 1955, с. 328); *Може, Ольга забула про нього, викреслила з свого серця?* (Рибак, Час сподівань і звершень, 1960, с. 57); — *В цю тяжку для мене хвилину я знайду в собі сили, щоб вирвати тебе, осоружного, назавжди з свого серця* (Цюпа, Вічний вогонь, 1960, с. 16).

Малюнок 1.6. Приклад статті Словника фразеологічних синонімів [Коломієць].

Ідеографічні словники

Анна Середницька згадує [25], що першими ідеографічними словниками були такі:

- грецький «Attikai Lexeis» Істроса (I століття н.е.) [75];
- санскритський словник «Амаракоша» /Amarakosha/ або /Namalinganushasanam/ (II-III століття н.е.), який досі досліджують [94].

Євгенія Карпіловська згадує про декілька ідеографічних словників (тобто тезаурусів) [11], тож поговоримо про них.

Першим ідеографічним словником новішого часу є славнозвісний «Тезаурус англійських слів і фраз Роже» /Roget's Thesaurus of English Words and Phrases/ [Roget Gutenberg], що визначив собою сутність сучасного тезауруса. Він набув широкої популярности і вже невдовзі після публікації у 1852 році його почали поширювати і перекладати іншими мовами, як-от французькою 1859 року, що зробив Теодор Робертсон [70], та німецькою 1873 року — Даніел Сандерс [69]. Укладати словники на його манір продовжують і досі, наприклад проєкт 2008 року для чеської мови від Алеся Клегра [80]. Також, оригінальний тезаурус Роже неодноразово перевидавали та доповнювали сам Пітер Марк Роже, його син та онук (видання 1936 року), а далі Роберт Датч (1962), С'юзан Ллойд (1982), Бетті Киркпатрик (1987) [108]. Надалі тезаурус Роже активно використовували в обробці природної мови. Серед таких проєктів особливо важливим видається робота Алістаїра Кеннеді і Стена Шпаковича з оцифрування та автоматичного розширення тезауруса на базі університету Торонто [76; 77], що завершилася створенням проєкту «Open Roget» [Open Roget].

Наступним опублікованим ідеографічним словником став Аналоговий словник французької мови Жана-Батиста-Пруденса Буасьєра 1862 року [Boissière], що містить одночасно і алфавітну та ідеографічну систематизацію. Ролі словника Роже він не повторив і мало вплинув на лексикографію [69]. Зразок структури сторінки зі словника Буасьєра наведено на малюнку 1.7.

<i>ruine.</i>	Achoppement, sm. Achopper, v. V. choc, obstacle, scandale.	chromatisation, sf. Achromatiser, v. Achromatisme, sm. V. couleur, 2 optique.	ACI, ACY, ASCI, ASI, ASSI, ASY.
Acheveur, sm. V. tréfilerie, 2 or.	Achore, sm. V. 2 peau.	Achronique ou Acronyque, a. V. astronomie.	Acicalaire, a. V. aiguille, pointe.
Achevoir, sm. V. 2 or.	Achoa, Achourou, sm. V. 3 bois.	Achthéographe ou Achthographie (ak), sm. Achthéographie, sf. V. poids.	ACIDE, a. sm. Acidifère, a. Acidifiable, a. Acidifiant, a. Acidification, sf. Acidifier, v. Acidité, sf. Acidule, a. Aciduler, v. (?).
Achille (<i>chi-le</i>), sm. V. brave, pieul.	Achras (<i>a-krasse</i>), sm. V. poire.		
Achillée (<i>chite-lée</i>), sf. V. plante.	Achromatique (<i>kro</i>), a. A-		
Achit (<i>chite</i>), sm. V. vigne.			

2 **Client**, acheteur habituel; — Clientèle, ensemble des clients.
Coopération, achat réciproque, une des formes du mariage roumain.
Command, celui qui charge d'acheter.
Commande, proposition d'acheter, ou ordre d'expédier.
Commence; commerçant; négociant; etc.
Commission, mission d'acheter ou de vendre; — Commissionnaire, celui qui vend ou achète par commission.
Comptant (au), en payant sur-le-champ.
Conquêt, acquêt fait en communauté.
Corrompre quelqu'un, acheter son suffrage, sa partialité; — Corruption; etc. V. séduire.
Crédit (à), à terme, sans payer comptant.
2 **Désachalandier**, — Désachalandier un marchand, lui faire perdre ses chalands.
Donner (se) quelque chose, se le procurer en l'achetant.
Emplette, chose achetée : faire emplette de, acheter.

(*) ACIDE

Se dit d'un grand nombre de substances d'une saveur aigre ou tirant sur l'aigre, qui ont la vertu chimique de décomposer d'autres substances.

Acerbe, âpre et acide; — Acerbité, état d'acerbe.
Acescence, état d'— Acescent, qui tourne à l'acide.
Acétate, sel formé d'acide acétique.
Acéteux, qui a le goût du vinaigre.
Acétique (acide), qui est la base du vinaigre.
Acétite, mot remplacé aujourd'hui par *acétate*.
Acétoité, état d'acéteux.
Acidifère, qui contient un acide.
Acidifiable, qui peut être acidifié.
Acidifiant, qui rend acide.
Acidification, action d' — Acidifier, changer en acide.
Acidité, état d'acide.
Acidule, légèrement acide; — Aciduler, rendre légèrement acide.

Малюнок 1.7. Частина сторінки з Аналогового словника французької мови [Boissière].

За Анною Середницькою, українська ідеографія почалася із роботи Наталії Сніжко, що розробила автоматизований ідеографічний тезаурус іменників української мови у 1995-1996 роки на 54 тисячі одиниць [25; 28; 29; 30]. Дослідниця досі займається цією темою, наприклад, 2018 року вона представила ідеографічну систематизацію лексики, яку придумали чи популяризували письменники [28].

2004 року Анна Середницька уклала «Ідеографічний словник дієслів переміщення сучасної української мови» [25]. Використавши СУМ-11 як джерело, вона назбирала і класифікувала 7 тисяч одиниць. Ономасіологічний опис включає синонімічні ряди, словотвірні групи, антонімічні пари і родо-видові зв'язки (архісеми з диференційними ознаками), що дозволяє зробити висновок про комплексний опис дієслів у її словнику.

Також не зайве буде згадати іншомовні ідеографічні словники:

- «Ідеологічний словник іспанської мови» /Diccionario Ideológico de la lengua española/ 1942 року Хуліо Касареса, який є визнаним науковцем в іспанській лексикографії [55; 46].
- «Німецький словник за предметними групами» /Der deutsche Wortschatz nach Sachgruppen/, виданий 1959 року [53].
- Згаданий Сергієм Фокіним [12] «Словник пов'язаних ідей» /Diccionario de ideas afines/ Фернандо Корріпіо на 400 тисяч одиниць, що був уперше виданий 1983 року [Corripio].
- Електронний «Російський семантичний словник» /Русский семантический словарь/ Караулова 1982 року [Шведова];
- WASPS — тезаурус від 2003 року, автоматично укладений на ґрунті однакових залежностей між словами у реченнях; містить 17844 іменника, 4033 дієслова та 7274 прикметника [78].

Ідеографічних словників, які містять у собі фразеологізми, надзвичайно мало. Як приклад можу навести лише «ІдеоФраз. Багатомовний словник фразеологічних синонімів» — це електронний ідеографічний словник обсягом 6 тисяч одиниць, що був розроблений Сергієм Фокіним [ІдеоФраз]. Він містить фразеологізми іспанською, українською, англійською, італійською, російською, французькою мовами, івритом та латиною.

1.2.3. Ідеографічне моделювання семантики фразеологізмів

Проблема із синоптичними схемами ідеографічних словників, яка, ймовірно, частково докладається до їхньої низької популярності, полягає у тому, що вона є завжди суб'єктивною [115]. Вище, у пункті 1.1, я подала дві синоптичні схеми, розроблені спеціально для фразеологізмів — Юрія Прадіда [22; 23] та Віктора і Дмитра Ужченків [33; 34]. У порівнянні вони виявилися засадничо несхожими. Із

тезаурусом Роже ситуація така сама, але також вона посилюється тим, що каркас словника розроблявся у першій половині XIX століття [Roget Original]. Скажімо, у тезаурусі Роже є цілий окремий клас, що детально описує лексику, пов'язану з мораллю та релігією.

Щодо метаструктури словника, то Пітер Марк Роже не подає її опису, тому використаємо для ідеографічних рівнів назви з роботи Алістаіра Кеннеді: Class /Клас/ → Section /Секція/ → Subsection /Підсекція/ → Head Group(s) /Керівна група(и)/ → Head /Керівне слово/ [76].

Загальних класів у синоптичної схеми тезауруса Роже всього 6:

1. «words expressing abstract relations» /слова, що виражають абстрактні відношення/;
2. «words relating to space» /слова, що стосуються простору/;
3. «words relating to matter» /слова, що стосуються матерії/;
4. «words relating to the intellectual faculties» /слова, що стосуються інтелектуальних здібностей/;
5. «words relating to the voluntary powers» /слова, що стосуються волевиявлення/;
6. «words relating to the sentient and moral powers» /слова, що стосуються емоцій та моральних відчуттів/ [Roget Gutenberg].

Кожен із класів розбивається на секції та підсекції, що уточнюють значення, за гіпонімічним принципом. Загальна кількість найспецифічніших значень, «керівних слів», складає 1000.

Якщо візуалізувати ці ієрархічно організовані дескриптори, то ми отримаємо граф у формі дерева, що має корінь у найзагальніших поняттях і розгалужується на кожному рівні, поки не дороходить до “листя” — кінцевої мети використання тезауруса Роже, тобто самих слів і фраз англійської мови. Пропоную на позначення повного набору ієрархічно впорядкованих дескрипторів слова (від

найзагальнішого до найспецифічнішого) адаптувати термін з теорії графів і називати його «ономасіологічний шлях».

Наприклад, ідіома «через пень колоду» буде мати такий ономасіологічний шлях: *volition > individual > voluntary action > complex voluntary action > unskillfulness*.

У таблиці 1.4 подано перші три рівні синоптичної схеми тезауруса Роже: класи, секції та підсекції, а також їхні відповідники англійською. Для віднаходження найближчих еквівалентів використано електронну версію словника Вебстера 1913 року видання (використаний в роботі тезаурус Роже виданий 1911 року) [Webster's 1913].

Деякі спостереження:

- Неозброєним оком помітно, що ця класифікація є значно об'ємнішою за попередні.
- Дескриптори застаріли і часом вживаються у значенні, яке вже навіть у [Webster's 1913] було маргінальним, наприклад, визначення «*affection*», що найкраще пасує до схеми, є третім ЛСВ: “*3. Bent of mind; a feeling or natural impulse or natural impulse acting upon and swaying the mind; any emotion; as, the benevolent affections, esteem, gratitude, etc.; the malevolent affections, hatred, envy, etc.; inclination; disposition; propensity; tendency.*” [Webster's 1913].
- В семантичній схемі помічено ознаки лінгвістичного детермінізму: під парасолькою того ж дескриптора «*affections*» надано дескриптори, які найдоцільніше перекладаються українською як «судження» і «відчуття». Тому чи іншому варіанту перекладу надано перевагу залежно від вкладених дескрипторів і слів, тоді як автор(и) схоже, що не розрізняли їх, вживаючи як одне поняття, упереміш.

Class /Клас/	Section /Секція/	Subsection /Підсекція/
<p>Words expressing abstract relations</p> <p>/слова, що виражають абстрактні відношення/</p>	Existence /існування/	Being, in the abstract /абстрактне буття/
		Being, in the concrete /конкретне буття/
		Formal existence /формальне існування/
		Modal existence /модальне існування/
	Relation /відношення/	Absolute relation /безпосередньо відношення/
		Continuous relation /неперервний зв'язок/
		Partial relation /подібність/
		General relation /загально про стосунки/
	Quantity /кількість/	Simple quantity /проста кількість/
		Comparative quantity /кількість у порівнянні/
		Conjunctive quantity /об'єднання кількостей/
		Concrete quantity /кількість у масі/
	Order /порядок/	Order in general /порядок загалом/
		Consecutive order /послідовний порядок/

Class /Клас/	Section /Секція/	Subsection /Підсекція/
		Collective order /колективний порядок/
		Distributive order /розподільний порядок/
		Order as regards categories /порядок відносно категорій/
	Number /число/	Number, in the abstract /число в абстракції/
		Determinate number /визначене число/
		Indeterminate number /невизначене число/
	Time /час/	Absolute time /абсолютний час/
		Relative time /відносний час/
		Recurrent time /повторюваний час/
	Change /зміна/	Simple change /проста зміна/
		Complex change /складна зміна/
	Causation /причиновість/	Constancy of sequence in events /стала послідовність подій/
		Connection between cause and effect /причинно-наслідковий зв'язок/
		Power in operation /причина дії/

Class /Клас/	Section /Секція/	Subsection /Підсекція/
		Indirect power /непряма причина/
		Combinations of causes /поєднання причин/
Words relating to space /слова, що стосуються простору/	Space in general /простір загалом/	Abstract space /абстрактний простір/
		Relative space /відносний простір/
		Existence in space /існування у просторі/
	Dimensions /розміри/	General dimensions /загальні розміри/
		Linear dimensions /лінійні розміри/
		Central dimensions /виміри навколо центру/
	Form /форма/	General form /загальна форма/
		Special form /спеціальна форма/
		Superficial form /форма поверхні/
	Motion /рух/	Motion in general /рух загалом/
		Degrees of motion /ступені руху/
		Motion conjoined with force /рух із силою/

Class /Клас/	Section /Секція/	Subsection /Підсекція/
		Motion with reference to direction /рух у напрямку/
Words relating to matter /слова, що стосуються матерії/	Matter in general /матерія загалом/	—
	Inorganic /неорганічна/	Solid matter /тверда матерія/
		Fluid matter /рідка матерія/
		Imperfect fluids /неідеальні рідини/
	Organic /органічна/	Vitality /життя/
		Sensation /сенсорні відчуття/
Words relating to the intellectual faculties /слова, що стосуються інтелектуальних здібностей/	Formation of ideas /формування ідей/	Operations of intellect in general /діяльність інтелекту загалом/
		Precursory conditions and operations /попередні стани і дії/
		Materials for reasoning /матеріали для міркування/
		Reasoning processes /процес міркування/
		Results of reasoning /результат міркування/
		Extension of thought /мандри думки/
		Creative thought

Class /Клас/	Section /Секція/	Subsection /Підсекція/
		/творча думка/
	Communication of ideas /висловлювання ідей/	Nature of ideas communicated /природа висловлюваних ідей/
		Modes of communication /спосіб висловлювання/
		Means of communicating ideas /засіб висловлювання/
Words relating to the voluntary powers /слова, що стосуються волевиявлення/	Individual volition /індивідуальне волевиявлення/	Volition in general /волевиявлення загалом/
		Prospective volition /майбутнє волевиявлення/
		Voluntary action /добровільна дія/
		Antagonism /протистояння/
		Results of voluntary action /результат добровільної дії/
	Intersocial volition /соціальне волевиявлення/	General intersocial volition /соціальне волевиявлення загалом/
		Special intersocial volition /спеціальне соціальне волевиявлення/
		Conditional intersocial volition /умовне соціальне волевиявлення/
		Possessive relations /відношення володіння/

Class /Клас/	Section /Секція/	Subsection /Підсекція/
Words relating to the sentient and moral powers /слова, що стосуються емоцій та моральних відчуттів (нераціональне)/	Affections in general /відчуття загалом/	—
	Personal affections /особистісні відчуття/	Passive affections /пасивні відчуття/
		Discriminative affections /диференційні відчуття (судження)/
		Prospective affections /відчуття щодо майбутнього/
		Contemplative affections /відповідність очікуванням/
		Extrinsic affections /судження і відчуття назовні/
	Sympathetic /відчуття до інших/	Social affections /соціальні відчуття/
		Diffusive sympathetic affections /вираження відчуттів до інших/
		Special sympathetic affections /специфічні відчуття до інших/
		Retrospective sympathetic affections /ретроспективні відчуття до інших/
	Moral /мораль/	Moral obligations /моральні обов'язки/
		Moral sentiments /моральні судження/
		Moral conditions

Class /Клас/	Section /Секція/	Subsection /Підсекція/
		/рівень моралі/
		Moral practice /моральні практики/
		Moral institutions /моральний кодекс/
	Religious /релігія/	Superhuman beings and regions /надлюдські істоти і місця/
		Religious doctrines /релігійні доктрини/
		Religious sentiments /рівень релігійности/
		Acts of religion /релігійні практики/
		Religious institutions /інститут церкви/

Таблиця 1.4. Ідеографічна класифікація тезауруса Роже [Roget Gutenberg]

В фразеологічній базі даних цієї роботи для ідеографічного опису фразеологізмів буде використано синоптичну схему тезауруса Роже, оскільки цей словник досі активно застосовують в обробці природної мови і не лише англійської [Open Roget; 130], тож це дозволяє в майбутньому інтегрувати цю роботу з іншими проєктами.

1.3. Комп'ютерна фразеографія та автоматичне оброблення фразеологізмів

1.3.1. Корпусні дослідження фразеологізмів

Складність автоматичної обробки фразеологізмів полягає у знаходженні їх у тексті [93; 80]. Складнощі виникають насамперед через можливі трансформаційні варіанти, лексичний склад і порядок слів, що можуть відрізнятися від словникового. У фразеологізмів є змінні та незмінні частини (наприклад, у *дати дуба* «дати» буде відмінюватися, а «дуба» має завжди лишатися в цій же формі). Також можуть траплятися okazionalni variaції (скажімо не *Пилип з конопель*, а *Василь*), які відслідкувати іще важче, бо вони непередбачувані.

Навіть уже маючи методику пошуку фразеологізмів у тексті, складно зрозуміти, коли вона не працює. Таких фразеологізмів просто не буде в зібраних даних, тобто ми матимемо справу з хибно негативними результатами /false negative (FN)/ [94]. Для вирішення цієї проблеми систему пошуку принаймні на початку можна зробити менш строгою і вручну вичищати хибно позитивні /false positive (FP)/ кейси. Також для перевірки її повноти /recall/ можна зібрати невеликий тестувальний корпус, де фразеологізми будуть розмічені вручну. У такий спосіб можна забезпечити створення якісної текстоорієнтованої лексикографічної бази даних словника, яка буде відображати реальний стан справ. Для цього і проводять *корпусні дослідження фразеологізмів*.

За твердженням Василя Старка, Розамунд Мун започаткувала вивчення фразеології, базоване на корпусних дослідженнях [92; 94; 31]. Серед тих, хто вивчали фразеологізми, використовуючи корпуси, Елахех Рафатбахш і Аліреза Ахмаді [105] називають Д. С. Баддорф та М. В. Евенс [40], Ділінь Ліу [85], Риту Симпсон і Душ'янті Мендіс [116], також можна згадати Ніколь Грегуар, що зібрала

5 тисяч «багатослівних виразів» /multi-word expressions (MWEs)/ для нідерландської мови з корпусу на 500 мільйонів слововживань [63].

NLP проєкти широко застосовують корпуси і від їхньої якості насамперед залежатиме якість продукту. Тому відкриті корпуси, як-от COCA (Corpus of Contemporary American English) [COCA], часто можна зустріти в роботах [79; 105].

Василь Старко присвятив статтю визначенню обсягу корпусу, необхідного для повноцінного опису фразеологізмів. Хоч він і не дав чіткої рекомендації, але визначив достатнім для їхнього вивчення корпус «LexisNexis Academic» (тепер «Nexis UNI») [Nexis] на 9 мільярдів слововживань, розмір якого він розрахував з огляду на частотність фразеологізмів у ньому [31].

Чи існують корпуси, що мають анотацію на фразеологізми? Таких знайти не вдалося, але [63] згадує про програму-парсер. Створення фразеологічної розмітки для корпусів або відбирання їх із текстів вручну є важкодосяжний завданням у часи автоматичної обробки природної мови, тому також існують проєкти, що намагаються автоматизувати виявлення фразеологізмів у тексті.

1.3.2. Укладання комп'ютерних фразеологічних словників

Створення словників за допомогою NLP

Як і створення корпусів, укладання словників вручну — це надзвичайно складна і копітка задача, що займає роки, а у випадку із ідеографічними словниками і десятиліття [25]. Наприклад, на створення першої версії тезауруса Роже пішло 47 років (1805-1852) [Roget Original]. Отже, із розповсюдженням комп'ютерів, їх усе ширше застосовували для спрощення одноманітних процесів на різних етапах укладання словників, а тепер буває створюють їх і повністю автоматично. Існує відповідна класифікація словників, яку наводять Валентина Перебийніс і Віктор Сорокін [21]:

- традиційний словник — словник, укладений з використанням комп'ютера або без і виданий у друкованому форматі.
 - електронна копія словника — відсканований традиційний словник у форматі електронної книги.
 - електронна версія словника — словник із інтерфейсом (програма чи сайт), пошуковою системою, фільтруванням результатів тощо, який укладено шляхом конвертації традиційного словника.
- електронний словник — словник без паперового варіанту, що має інтерфейс для взаємодії з ним і різноманітні додаткові функції пошуку, фільтрації тощо.
- автоматичний словник — різноманітні бази даних, які використовуються у проєктах, кінцевим продуктом яких є не словник, тобто користувачі не мають прямого доступу до нього. Також додатково зазначимо, що такі словники можуть мати незвичну структуру: не мати визначень, прикладів і стилістичних міток, але мати якусь іншу розмітку, яка актуальна для задачі. Наприклад, як Японський словник словосполучень /Japanese Dictionary of Multiword Expressions (JDMWE)/ на 100,000 одиниць, що має інформацію про синтаксичні варіанти та дерево, морфологічні дані, але нічого з того, що можна очікувати від тлумачного словника [113].

Наступним кроком у словникарській справі стали багатомовні ономасіологічні словники створені за допомогою машинного навчання. Група дослідників із університету Цінхуа розробила ономасіологічний словник [WantWords], використавши мультиканальну модель [131]. Під мультиканальністю мається на увазі поєднання декількох підходів, що дозволяє забезпечити результат вищої якості. Їхній словник має не тільки англійську та китайську версії, а й англійсько-китайську та китайсько-англійську [104]. Особливістю новіших

автоматичних словників є спосіб взаємодії користувача з ними: не потрібно ходити синоптичною схемою, щоб знайти шукане слово, а достатньо описати його природною мовою. Це досягається за допомогою різноманітних способів порівняння тексту, починаючи з індексів і завершуючи використанням моделей машинного навчання. До останнього способу вдалися й Бушра Сіддіке і Суф'ян Бег у проєкті i-RD — вони використали для тренування невелику базу даних з 753 найпопулярніших англійських фразеологізмів і отримали 70% точности системи [114].

Підхід до вирішення задачі за допомогою правил останнім часом застосовується доволі точно, коли цього вимагає конкретна задача. Наведемо три приклади:

- Проєкт 1998-2009 років із додавання фразеологізмів до бази даних WordNet авторства Кристіан Феллбаум [59; 58]. Він описує додавання до системи 200 фразеологізмів.
- Проєкт 2009 року, що зосередився на пошуку дієслівних фразеологізмів англійської мови. Автори застосували як критерій рівень сталости словосполучень і досягли фінальної точности у 74% [59].
- Бібліотека Idiomatch 2021 року автоматично шукає фразеологізми у різних трансформаційних варіантах у текстах [Idiomatch]. Цей проєкт є модулем для роботи з бібліотекою SpaCy, яка вже працює на моделі [SpaCy], а також частиною проєкту «Idiomify», про який буде йти мова пізніше. Автор надихався розробкою того року для автоматичного пошуку фразеологізмів у BNC [BNC] або будь-яких інших на основі пошукового сервера «Elasticsearch» [68].

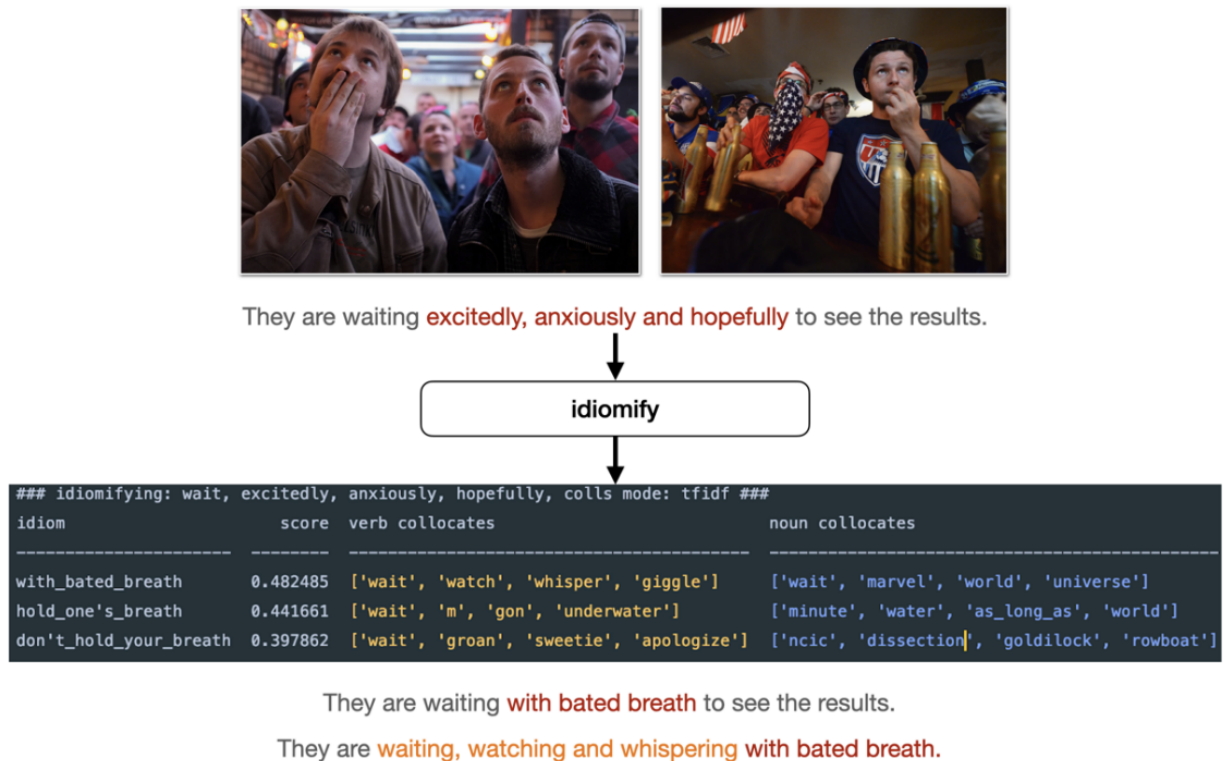
Раніше моделі рідко застосовували для фразеологізмів. На це було щонайменше дві причини:

- Короткі залежності. До 2016 розповсюдженими були статистичні моделі, які працювали на n-грамах [81], що дозволяло враховувати лише дуже близький контекст слів і, відповідно, давало поганий результат як через низьку частотність фразеологізмів, так і через їхню головну особливість — невідповідність змісту і форми. Ситуація почала покращуватися із появою «механізмів уваги» /attention mechanisms/ у рекурентних нейронних мережах /recurrent neural networks (RNN)/ — вони підвищили і якість перекладу загалом, і обробку фразеологізмів. Приклади статей: Джанкарло Солтон, Роберт Росс і Джон Келлегер [110], Ашиш Васвані, Ноам Шазір, Нікі Пармар та ін. [126].
- Частотність фразеологізмів досі важлива. Усі моделі є за своєю суттю статистичними, тобто вимагають великої кількості репрезентативних та зважених даних. Фразеологізми ж переважно є доволі нечастотними, а отже, для їх дослідження необхідно значно більше даних (як-от той корпус LexisNexis на 9 мільярдів слововживань в оцінці, який згадує Василь Старко [30]). Для прикладу, Елахех Рафатбахш і Аліреза Ахмаді дослідили частотність фразеологізмів і виявили, що 10 найчастотніших зустрічаються 1,38-4,71 разів на мільйон слововживань (і при цьому найвища частота з великим відривом від інших — *behind the scenes*) [105].

Прості моделі

Моделі машинного навчання, зокрема, використовують для анотування даних. Створені автоматично датасети дозволяють зберегти час людей-анотаторів, їх можна використовувати і для створення фразеологічної розмітки корпусів, і для швидшого укладання словників, і для швидшого створення і апробації ідеї data-driven проєктів. Приклади проєктів, що займалися фразеологізмами:

1. контекстуальні вкладання слів /contextual word embeddings/ з оновленим підходом, що враховує багатозначність, який дав хороші результати точности класифікації — на рівні 86-98% залежно від використаної моделі (2021 рік), авторства Тадей Шкворц, Полона Гантар, Марко Робнік-Шиконя [117];
2. ще один приклад використання векторів — проєкт «Idiomify» 2021 року. Автор поєднує правила (у препроцесингу) із моделлю Word2Vec (для забезпечення функції ономасіологічного словника) [79].



Діаграма 1.1. Демонстрація принципу роботи системи «Idiomify» [79].

3. ансамблева модель /ensemble model/ Дж. Брискілал і С. Н. Субалаліта для класифікації фразеологічних і вільних словосполучень, що навчалася на тих же вкладаннях слів і має точність класифікації у 90% (2022 рік) [43];
4. інша ансамблева модель 2022 року для тої ж класифікації від С. Абарана, Дж. І. Шиба, С. Прадип Деванаян. Вона поєднала у собі три інші моделі

(BERT, RoBERTa і K-BERT), досягнувши за рахунок цього точности /accuracy/ у 95,86%, що перевищує початкову, без використання цього підходу, на 8% [35];

5. модель-трансформер натренована Сімоне Тедескі, Фуесто Мартеллі, Роберто Навільї для 10 мов (китайська, нідерландська, англійська, французька, німецька, італійська, японська, польська, португальська, іспанська) аотувала датасети найвищої якості з точністю 82.85% порівняно з людьми-експертами (2022 рік) [121].


LLM (великі мовні моделі)

Вище було зазначено, що для того, щоб навчити моделі визначати фразеологізми у тексті, потрібна величезна кількість даних. Великі мовні моделі мають якраз таку особливість — поза різноманітними архітектурами з мільярдами параметрів, вони також навчаються “на всьому інтернеті”, а отже отримують велику кількість зразків вживання фразеологізмів.

Візьмімо для прикладу ChatGPT [ChatGPT]. Я спробувала використати його для пошуку міжмовних еквівалентів для фразеологізмів. Два різні приклади запитів і отриманих результатів наведено на малюнках 1.8 і 1.9. Обидва рази я запитувала про міжмовні фразеологічні відповідники англійською, українською, польською та іспанською мовою. У першому випадку мій запит був обмежений значенням: я вказала, що мені потрібні фразеологізми зі значенням *проспати*, а в другому такого обмеження не було. Для української мови ці результати в найкращому випадку виглядають перекладеними з інших мов. Натомість усі приклади англійською є справжніми.



Equivalent idioms in English, Ukrainian, Polish, and Spanish on the topic of oversleeping in a form of a table:

 Certainly! Here's a table with equivalent idioms about oversleeping in English, Ukrainian, Polish, and Spanish:   

English	Ukrainian	Polish	Spanish
To sleep like a log	Спати як сурмач	Spać jak suseł	Dormir como un tronco
To sleep in	Вистояти у ліжку	Wylegiwać się w łóżku	Dormir hasta tarde
To hit the snooze button	Торкнути сніданку	Przycisnąć przycisk drzemki	Presionar el botón de repetición
To sleep through the alarm	Проспати будильник	Przespać budzik	Dormirse con el despertador
To be a late riser	Бути довго вставатим	Być późnym ptaszkiem	Ser un trasnochador

Please note that idioms can vary between languages and cultures, and these examples aim to capture the general idea rather than provide exact translations.

Малюнок 1.8. Результат запиту до моделі ChatGPT з обмеженням теми [ChatGPT].



Equivalent idioms in English, Ukrainian, Polish, and Spanish in a form of a table:

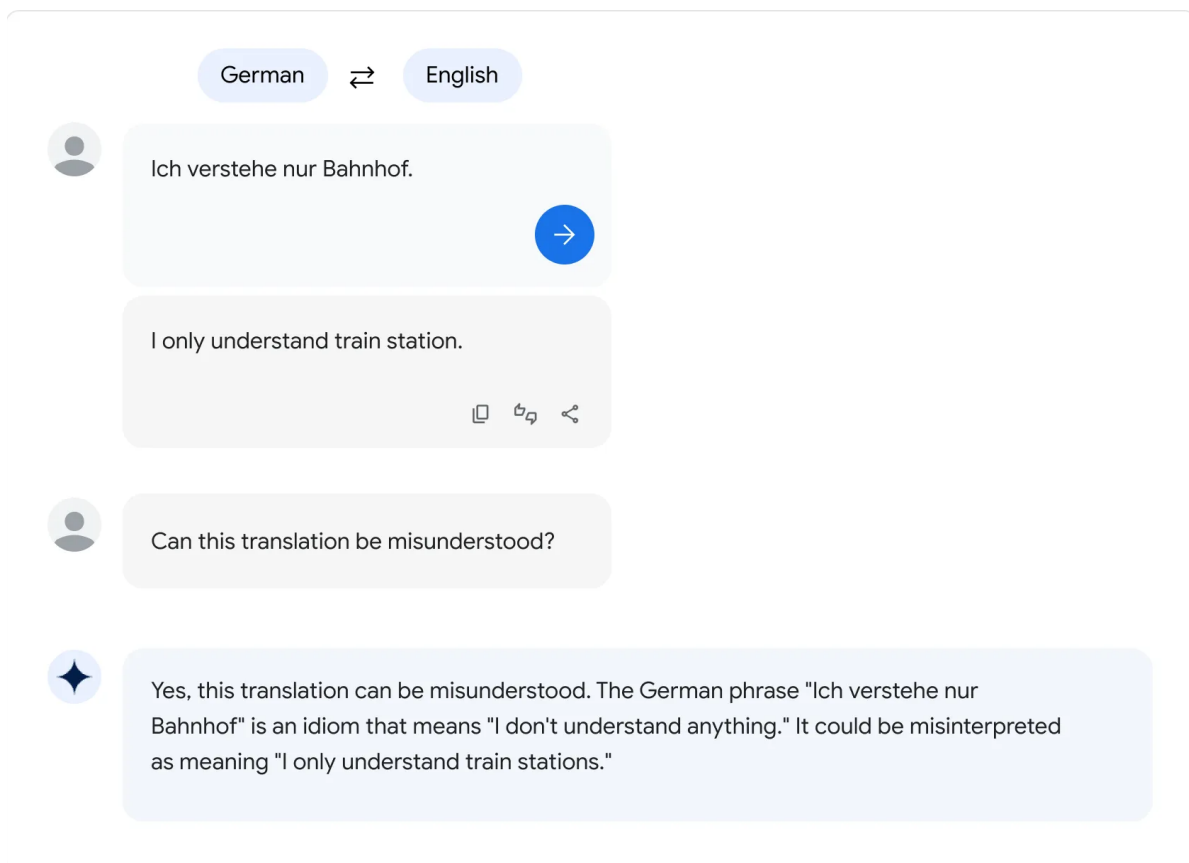
 Certainly! Here's a table with equivalent idioms in English, Ukrainian, Polish, and Spanish on various topics:   

English	Ukrainian	Polish	Spanish
To let the cat out of the bag	Випустити кота з мішка	Puścić kota z worka	Revelar el secreto
To kill two birds with one stone	Зарити двох зайців одним пострілом	Upiec dwie pieczenie na jednym ogniu	Matar dos pájaros de un tiro
To beat around the bush	Ходити навколо гарячої каші	Owijać w bawełnę	Andarse con rodeos
To have a frog in one's throat	Мати ляльку в горлі	Mieć klucha w gardle	Tener la voz ronca
To let the grass grow under one's feet	Не сидіти на місці	Nie trzymać się kurczowo ziemi	No quedarse de brazos cruzados

Please note that idioms can vary between languages and cultures, and these examples aim to capture the general idea rather than provide exact translations.

Малюнок 1.9. Результат запиту до моделі ChatGPT без обмеження теми [ChatGPT].

Із надзвичайно швидким розвитком нейронних моделей ситуація змінюється також швидко. Ефективність нейронних моделей, кількість даних, на яких вони навчаються, і їхні аналізаторські здібності покращуються. 10 травня 2023 року Google оприлюднив свою нову модель — PaLM 2 [PaLM 2], що має 540 мільярдів параметрів [95]. Найважливішою її відмінністю у контексті фразеологізмів є те, що *“попередні великі мовні моделі, як-от PaLM, використовували переважно англомовні датасети для попереднього тренування моделі. PaLM 2 має більш багатомовний і різноманітний корпус, що включає сотні природних мов та мов програмування, математичних рівнянь, наукових статей і веб-сторінок”* [PaLM 2]. Така зміна надала моделі можливість розуміти фразеологізми, що чудово демонструє малюнок 1.10.



Малюнок 1.10. Демонстрація роботи моделі PaLM 2 із фразеологізмами [PaLM 2].

Серед багатьох категорій запитів, які демонстровано, є й використання моделі як ономасіологічного словника [106]. Цю модель не випробувано через відсутність доступу до неї.

1.3.3. Застосування фразеографічних баз даних в машинному перекладі

Якщо база даних електронна і має структуру даних, не заточену під конкретну задачу, ніщо не обмежує способів її застосування. Тобто вона, у найкращому сценарії, є окремим модулем, розробленим з перспективою на розвиток, розширення, а також легку і швидку інтеграцію у різнопланові проекти; вона є модулем, що має свої інструменти обробки тексту, специфічні для моделювання фразеологізмів. Припустимо, що вже маємо таку базу. Що можна з нею робити із задач автоматичної обробки природної мови /natural language processing (NLP)/? Килгарифф наводить такі способи використання тезаурусів (особливо актуальні для англійської) [78]:

1. Для парсингу речень:

- У випадку багатозначності синтаксичної структури у корпусі текстів шукають подібну, щоб вона допомогла вирішити проблему за аналогією. Тезаурус може узагальнити такі структури за рахунок слів із тих самих категорій, що і шукані, а це збільшить імовірність знайти те, що потрібно.

Приклад для англійської:

eat fish with a fork

eat fish with bones [78]

Приклад для української:

гуляти із задоволенням

гуляти із подругою

- Для визначення меж сполучення слів: що більш подібні значенням компоненти, то більша вірогідність, що означення першого слова розповсюджується і на друге. Для української мови така проблема актуальна хіба що якщо компоненти у множині.

Приклад для англійської:

old boots and shoes

old boots and apples [79]

Приклад для української:

старанні студенти і викладач

старанні студенти і викладачі

2. Для визначення зв'язків у тексті:

- Часткове розв'язання проблеми кореференції — у випадках, коли слово замінено родовим поняттям.

Приклад:

*На столі лежав **гарбуз**. Тато розрізали **овоч** і вчинили каші.*

- Для автоматичного виявлення когерентних і некогерентних уривків тексту за допомогою відслідковування тем у них.

3. Для розведення омонімії:

Знову мова про контекст вживання: можна порівнювати частотність поєднання слів зі словами, які близькі до омонімів у тезаурусі.

Ще одним зі способів застосувати базу даних фразеологізмів є автоматичне насичення текстів фразеологізмами або навпаки заміна ідіоматичних виразів прямим значенням. Таких проєктів небагато, можливо вони існують як одна з особливостей більшого проєкту для редагування тексту. Наведу як приклад «Idiomify» має на меті допомагати користувачам вивчати фразеологізми, підказуючи, коли їх можна використати замість висловів із прямим значенням [79]. Для цього автор використовує попередньо створену з корпусів базу даних.

Використання ідеографічних фразеологічних словників

Письмо, копірайтерство, редагування. Написання та редагування текстів вимагають пошуку влучних форм вираження ідей, хоча прагматична функція використання фразеологізмів у мовленні, на відміну від інших стилістичних фігур, залишається менш очевидною. У дослідженні Roberts і Kreuz наведено доволі невпевнені показники опитування:

- щоб додати гумору /“to be humorous”/ (44%);
- щоб бути звичними /“to be conventional”/ (38%);
- щоб бути ввічливими /“to be polite”/ (38%) [107].

Тим не менш, автори застосовують їх постійно, особливо часто фразеологізми зустрічаються в медіа [Берега]. Складнощі з пошуком підхожих до ситуації фразеологізмів можуть виникати зокрема і через рівень знання мови, що потребує перекладних або ономасіологічних словників. З іншого боку, залежно від частоти виконання подібного завдання та віку людина може мати складнощі зі згадуванням фразеологізмів, які вона вже знає. Останнє явище в англійській літературі відоме за терміном *tip-of-the-tongue phenomenon* [124], тобто «крутиться на язичку». У цій ситуації ономасіологічні словники (з легким пошуковим алгоритмом!) також є надзвичайно помічними [115].

Вивчення іноземної мови. Є багато досліджень, присвячених важливості вивчення фразеологізмів для досягнення вільного рівня володіння мовою [67; 72; 82; 96; 128; 42; 45; 83; 122].

Як стверджує Сергій Фокін, поширеною практикою під час навчання є використання тематичних підбірок фразеологізмів [12]. З особистих спостережень, зазвичай використовують спеціально підготовлені матеріали, а не словники. Причин вбачаю щонайменше дві: використання словників передбачає певну підготовленість студентів; словник занадто гомогенний для цієї задачі — під час вивчення мови важливий розподіл лексики за рівнем володіння нею. Для

визначення цього можна спиратися на частотність фразеологізмів у корпусах [111; 94], бо нечастотна лексика апіорі належить до вищого рівня. Досі частотність використовували, щоб з'ясувати які фразеологізми викладати під час вивчення мови як іноземної взагалі. Елахех Рафатбахш і Аліреза Ахмаді, скажімо, дослідили для цього фразеологізми за темами з класифікації Оксфордського словника ідіом /Oxford Dictionary of Idioms/ [Oxford] у Корпусі сучасної американської англійської мови /Corpus of Contemporary American English/ [COCA]. Тематична організація сприяє їх легшому засвоєнню при вивченні, а визначення частотності у корпусі також дозволила їм виключити із підбору матеріалів для студентів фактор суб'єктивності [105]. Детальніше тематичний підхід у контексті двомовних словників для мовців-початківців розглянув Василь Старко, пропонуючи розробляти гібридні словники, що будуть і двомовними, і тематичними і навчальними [118].

Міжкультурна комунікація. Фразеологізми, поза невеличкою купкою інтернаціоналізмів (вони також часто обмежені так званим “західним світом”), є унікальними для своєї мови. Вивчаючи іноземну мову, учні вивчають і фразеологізми цієї мови. Це може допомогти у спілкуванні з іноземцями з відповідної країни, але є дві проблеми: 1) ці частотні фразеологізми і вислови часто сприймаються як кліше, яких варто уникати; 2) вони часом бувають неінклюзивними: через брак контексту студенти цього не помічають, але на співрозмовників це може справити негативне враження; 3) якщо комунікація у групі людей відбувається певною іноземною для всіх мовою (часто це англійська), то рекомендовано взагалі уникати будь-яких фразеологізмів, щоб забезпечити вищу якість розуміння та прозорості комунікації [101; 62; 98; 129]. Така ж рекомендація існує і для академічних текстів [90].

Перекладацтво. Найбільш проблематичними фразеологізми є у перекладі, а особливо, якщо на нього обмежений час. Пошук достатньо близького відповідника

для конкретного контексту, який ще й передавав би авторську ідею, гру слів чи пасував би до побудованого на ідіомі жарту, — задача не з легких, а інколи й неможлива взагалі. Більше того, як і будь-яка лексика, фразеологізми не є постійними, вони існують у соціолектному, діалектному і часовому контекстах мови, а отже з'являються і виходять із ужитку, не завжди стають поширеними серед усього суспільства. Тому для перекладачів важливо добре орієнтуватися в різноманітних виразах і мати зручні та вичерпні словники для полювання на відповідники. Саме ідеографічні словники у цьому випадку є найкращим рішенням.

Традиційний переклад

Найбільше уваги фразеологізмам приділяють під час перекладу художніх текстів. Дослівний переклад у цьому випадку неприйнятний за невеликим винятком інтернаціоналізмів, що зустрічаються у дуже схожій формі в різних мовах, але вони є скоріше випадковістю, ніж правилом. Елізабет Пірайнен постановила систематизувати цей пласт фразеології серед європейських мов. На позначення загальної категорії фразеологізмів, що зустрічаються у групах мов, вона вживає «поширені ідіоми» /widespread idiom/ і стверджує, що більшість із них походять із Біблії [100]. Також займалися класифікацією фразеологізмів у контексті перекладу Роксоляна Зорівчак [9], Сергій Влахов [6], Моніка Бейкер [39], Сильвія Моліна Плаза [102].

Способи перекладу фразеологізмів за Сергієм Влаховим і Монікою Бейкер [6; 39]:

- фразеологізм => фразеологічний еквівалент (повний і частковий):

«Поширені ідіоми» Елізабет Пірайнен, що мають повністю або майже повністю однаковий лексичний склад і структуру. Вони є доволі рідкісними.

Приклад: хапати бика за роги => take the bull by the horns

- фразеологізм => фразеологічний аналог:

Фразеологізм перекладено фразеологізмом такого ж значення, але іншої форми (пряме застосування ідеографії!).

Приклад: *to be as busy as a beaver* => *крутиться, як білка в колесі*

- фразеологізм => фразеологічна калька:

Часом інтернаціоналізми виникають за рахунок запозичення, стаючи фразеологічними кальками — “дослівним перекладом фразеологічної одиниці за частинами. Напр.: *жереб кинуте* (лат. *alea jacta est*); *поставити крапки над «і»* (фр. *mettre les point sur les «i»*)” [Єрмоленко].

- фразеологізм => дефініція:

Використання визначення оригінального фразеологізму в перекладі, що застосовується у випадку його непрозорої образності і відсутності аналогів, тобто дефразеологізація.

- фразеологізм => фразеологізм, що є контекстуальним відповідником:

Переклад здійснюється фразеологізмом, але це має сенс лише в наявному контексті, оскільки він відрізняється за значенням і образністю від оригінального.

Висновок щодо того, чи є фразеологізм добрим аналогом, залежить від аспектів їх значення та функціонування в мові, які ми враховуємо. Комп’ютерна обробка природної мови передбачає спрощення та допускає відхід від теоретичної інтерпретації термінів. Натомість, якщо враховувати усе, то повних еквівалентів практично не існує [73]. Ельмар Шафрот стверджує, що необхідно враховувати такі деталі:

- “усі формальні варіанти, враховуючи морфологічні, лексичні, синтаксичні й інколи навіть просодичні аспекти ... ;
- частотність, що спирається на дані з ряду великих лінгвістичних корпусів;
- загальноприйнятий вжиток і сильна когнітивна усталеність ідіоми зі сторони мовця;

- формальність, починаючи від книжних чи поетичних виразів і закінчуючи табуйованими;
- значення: семантична диференціація багатозначних ідіом, точний опис загального значення і контекстно-залежного;
- семантичне поле, наприклад, дружба, любов, їжа, людське тіло, обмеженість, відчай, радість;
- прагматичний ефект (ввічлива чи ні, іронічна, саркастична, гумористична тощо);
- іллокуційний вплив (мовний акт): вибачення, вдячність, комплімент, лестощі, критика, докір, погроза тощо;
- синтаксис:
 - внутрішній синтаксис (опціональне дієслово чи необхідне, опціональні словосполучення чи базові, частотні означення);
 - зовнішній синтаксис (більші структури, у яких може траплятися ідіома, опціональні зв'язки з *that-clauses*, *-ing-clauses*, *to-infinitive clauses* чи *prepositional objects* і аналогічними конструкціями в інших мовах)” [72].

Автоматичний переклад

Ця сфера науки є молодшою за саму фразеологію — вона бере початок від Джорджтаунського експерименту у співпраці з ІВМ 1954 року, що став результатом довгої роботи на опрацюванням ідеї та підготовкою у 1947-1954 роки [71]. Переклад — це перше, що спадає на думку, із багатомовними бази даних фразеологізмів. Ця задача є нелегкою навіть для двох мов, починаючи від укладання такої бази даних, закінчуючи способом організації одиниць та пошуку еквівалентів.

Спеціалізованих проєктів з перекладу фразеологізмів не знайшлося, але наводжу приклад використання тезауруса Роже для загального автоматичного перекладу. Віолет Принс і Жак Шош використали тезаурус Роже і французький тезаурус «Larousse» (словник синонімів і антонімів) для побудови вкладень слів у них. Словник Роже, що складається з 1000 одиниць концептів, став основою векторної моделі. Французький словник містить менше концептів — 873, тому авторам потрібно було також укласти систему міжсловникових відповідників. Під час тестування готової системи досягнуто 86% повноти /recall/ і 71% влучності /precision/ 71% [103].

Автоматичний переклад фразеологізмів різного рівня семантичної спаяності компонентів продемонстровано на роботі онлайн-перекладачів, що використовують нейронні моделі. Цитати для контекстного перекладу запозичено з СУМ-11. Використано переклади систем Google Translate, DeepL і Microsoft Bing, результати подано у таблицях 1.5, 1.6, 1.7, 1.8 та 1.9.

У таблиці 1.5 подано порівняння перекладу для ідіом *міняти шило на швайку* і *спряжити на схаб*. Жоден із перекладів і не наблизився до адекватного перекладу. *Швайка* виявилось більш прозорим словом для моделей, тож переклад лишається приблизно у тій же сфері, хоч і не є правильним навіть у розумінні дослівного. Для слова ж *схаб* усі варіанти повністю нерелевантні, жодна із систем не впоралася.

О Ц І Н К А	Фразеологізм	<i>міняти шило на швайку</i> /невигідно змінювати що-небудь на щось/	<i>спряжити на схаб</i> /знищити повністю кого-небудь/
	Оригінал	Не на те втекли ми від пана Бжеського, щоб шило на швайку міняти.	Постійте, хаами! Чекай ти, попе! Буде з вами не те! Я спряжу вас на схаб!

0	Google Translate	We did not run away from Mr. Brzeski to exchange a needle for a seamstress.	Hold on, you bastards! Wait, dad! It will not be the same with you! I will tie you to the shab!
0	DeepL	We didn't run away from Mr. Brzeski to exchange an awl for a maid.	Wait, you bastards! Wait a minute, priest! I'll put you in the hayloft!
0	Bing Microsoft Translator	That's not why we ran away from Mr. Brzeski to change the awl for a sewing.	Wait, hams! Wait for you, pop! It will not be with you! I'll conjugate you to the suck!

Таблиця 1.5. Порівняння автоматичного перекладу ідіом у контексті.

У таблиці 1.6 подано порівняння перекладу для фразеологічних єдностей *промацати ґрунт* і *довгий язик*. Для першої DeepL надав задовільний переклад, *test the waters* є адекватним відповідником, тоді як переклад від Bing, можливо, є зрозумілим, але не хорошим: в англійських словниках присутній лише вираз *testing ground*, який вживається на позначення місця. Щодо перекладів для другої, в англійській мові є слово, яке дуже схоже на фразеологізм *довгий язик*, — *long-tongued* [long-tongued], тож імовірно переклад *long tongue* буде зрозумілим.

О Ц І Н К А	Фразеологізм	<i>промацати ґрунт</i> /з'ясувати, уточнити, перевірити можливість чого-небудь/	<i>довгий язик</i> /про балакливу людину, що говорить зайве, або про її вдачу/
	Оригінал	Мамій вирішив ретельно промацати ґрунт щодо можливості одруження з Тетяною.	Я бачив, що моя розповідь схвилювала її, і жорстоко картав себе за довгий язик.
1	Google Translate	Mamii decided to carefully feel the soil about the possibility of	I saw that my story excited her, and I severely reprimanded myself

		marrying Tatyana.	for my long tongue .
2	DeepL	Mamii decided to thoroughly test the waters about the possibility of marrying Tetiana.	I could see that my story had excited her, and I was severely chastising myself for my long tongue .
1,5	Bing Microsoft Translator	Mom decided to carefully test the ground regarding the possibility of marrying Tatiana.	I saw that my story excited her, and severely reproached myself for her long tongue .

Таблиця 1.6. Порівняння автоматичного перекладу фразеологічних єдностей у контексті.

У таблиці 1.7 подано порівняння перекладу для фразеологічних сполучень *лускати зі (зо) сміху* і *наковтатися сліз*. Для першого непогані результати показали Google Translate та Microsoft Bing (*squel with laughter, burst out laughing*), але з другим найкраще впорався DeepL (*cry one's eyes out*).

О Ц І Н К А	Фразеологізм	<i>лускати зі (зо) сміху</i> /дуже весело, голосно сміятися/	<i>наковтатися сліз</i> /багато поплакати/
	Оригінал	Дівчата лускали зо сміху, штовхаючи одна одну ліктем.	Сама Яресьчиха сліз наковталася вволю на доччиному весіллі.
1	Google Translate	The girls squealed with laughter , nudging each other with their elbows.	Yareschikha herself swallowed tears at her daughter's wedding.
1	DeepL	The girls burst out laughing, elbowing each other.	Jareschikha herself cried her eyes out at her daughter's wedding.
1	Bing Microsoft Translator	The girls burst out laughing , pushing each other with their	Yareschikha herself swallowed tears willy-nilly at her daughter's

		elbows.	wedding.
--	--	---------	----------

Таблиця 1.7. Порівняння автоматичного перекладу фразеологічних сполучень у контексті.

У таблиці 1.8 подано порівняння перекладу для фразеологізованих сполучень *на рідкість* і *видне місце*. Пропозиції від Google Translate (*rare gift of*) і DeepL (*unusually*) для першого сполучення є хорошими контекстуально прийнятними варіантами перекладу. Друге переклав у переносному значенні лише DeepL (*be prominent*).

О Ц І Н К А	Фразеологізм	<i>на рідкість</i> /дуже, надзвичайно/	<i>видне місце</i> /помітне, видатне становище кого-небудь/
	Оригінал	Природа його наділила на рідкість розкішним красномовством.	Ще в гімназії Борис займав видне місце серед товаришів.
1	Google Translate	Nature endowed him with a rare gift of luxurious eloquence.	Back in the gymnasium, Boris occupied a prominent place among his friends.
2	DeepL	Nature has endowed him with an unusually luxurious eloquence.	Back at the gymnasium, Borys was prominent among his classmates.
0	Bing Microsoft Translator	Nature endowed him with extremely luxurious eloquence.	Even in the gymnasium, Boris occupied a prominent place among his comrades.

Таблиця 1.8. Порівняння автоматичного перекладу фразеологізованих словосполучень у контексті.

У таблиці 1.9 подано порівняння перекладу для фразеологічних виразів *бодай вас злидні побили* і *нехай він спухне*. У випадку з першим виразом лише

DeepL зрозумів, що *бідність* як явище є актором, а також правильно сформулював прокльон. У другому випадку жодна із систем не впоралася із перекладом.

О Ц І Н К А	Фразеологізм	<i>Бодай вас злидні побили</i> /побажання нещастя, бідності; прокляття/	<i>Нехай він спухне</i> /уживається як лайка, проклін/
	Оригінал	Бодай же вас, цокотухи, та злидні побили.	Картоплю носила, бо мій дармоїд вилежується, щоб він спух...
0	Google Translate	Perhaps you, tsokotukhs, were beaten by those poor people.	I wore potatoes, because my worm is lying down so that it can dry out...
0,5	DeepL	<i>May poverty beat you, you tzokotuks.</i>	I was carrying potatoes because my freeloader is lying down to swell them...
0	Bing Microsoft Translator	At least you, clatters, and poverty have been beaten.	I carried potatoes because my parasite lies down to swell...

Таблиця 1.9. Порівняння автоматичного перекладу фразеологічних виразів у контексті.

Протестовані системи перекладу працюють не ідеально, кожна по-різному, але все ж таки певні, хоч і не рівномірні, результати є. Підрахувавши умовні оцінки, можемо зробити висновок, що DeepL зараз працює з фразеологізмами найкраще: він переклав 5,5 речень із 10. Google Translate набрав 3 переклади із 10, Bing Microsoft отримав найменші 2,5 із 10. Загалом, моделі найкраще впоралися із фразеологічними єдностями і добре себе показали із фразеологічними сполученням і фразеологізованими словосполученнями. Виявилось, що вони поки ще не можуть обробляти фразеологічні вирази та ідіоми.

1.4. Висновки до розділу і робочі поняття

Фразеологія досі є новою наукою і її терміносистема не до кінця усталена, тож необхідно мати на увазі, що кожен автор може мати на увазі щось інше за наше власне розуміння котрогось із термінів.

Фразеологізми широко досліджені в російському мовознавстві через тренд, який задав Віктор Виноградов [5]. З останніх років ХХ століття почало рости зацікавлення фразеологізмами і в англійській літературі, особливо з огляду на те, що машинне навчання дозволяє автоматизувати надзвичайно часозатратні етапи укладання словника.

В літературі спостережена узгоджена думка стосовно того, що ідеографічний спосіб опису фразеологізмів пасує найкраще до їхньої природи, але самих фразеологічних ідеографічних словників було знайдено небагато, зокрема через складність їх укладання.

У якості робочих понять використовуємо такі:

- «фразеологізм» на позначення загальної категорії;
- класифікація за семантичною спаяністю компонентів Петра Дудика [8], уточнена визначеннями Ігоря Мельчука [89] і Віктора Виноградова [5]:
 - «ідіома» (семантично зрощені вислови),
 - «фразеологічна єдність»,
 - «фразеологічне сполучення»,
 - «фразеологізоване словосполучення»,
 - «фразеологічний вираз» [8];
- «ідеографічний словник», «ономасіологічний словник», останній протиставляється терміну «семасіологічний словник»;
- «ономасіологічних шлях» — запропонований мною термін, що позначає шлях, який треба пройти від найзагальнішої до найспеціальнішої категорії в

графові синоптичної схеми тезауруса Роже, щоб отримати повне нашарування сенсів.

РОЗДІЛ 2. УКЛАДАННЯ ІДЕОГРАФІЧНОЇ БАЗИ ДАНИХ ФРАЗЕОЛОГІЗМІВ

Ідея практичної частини полягає у тому, щоб за допомогою ідеографічної синоптичної схеми описати значення фразеологізмів різних мов. Це перший крок до автоматичного знаходження одномовних та міжмовних фразеологічних синонімів. Для якісного встановлення міжмовних відповідників можна також додатково визначати диференційні семи фразеологізмів, їхню тематику, сферу вживання, продуктивність вживання у текстах тощо. Натомість класифікація фразеологізмів за семантичною спаяністю компонентів є актуальною для задач перекладу, бо дозволяє обрати напрямок пошуку доброго відповідника.

2.1. Інфологічна модель БД фразеологізмів

Для цієї роботи вирішено обмежитися укладанням одномовної бази даних фразеологізмів. Це вплинуло на *вибір словника-основи*, бо більше не було вимоги багатомовності словника. Натомість визначено, що це джерело має бути доступним, загальномовним і структурованим так, що уможливлуватиме формальне моделювання статті. Після перегляду декількох словників джерелом для бази даних визначено електронну версію Словника української мови (СУМ) в 11 томах [СУМ-11], а саме сайт sum.in.ua. Саме він відповідає усім перерахованим критеріям.

Для ідеографічного опису фразеологізмів обрано синоптичну схему тезауруса Роже у версії проєкту Гутенберг [Roget Gutenberg].

Відбір та структурування об'єктів опису. СУМ-11 є тлумачним словником, тож структурування інформації полягало у розшиванні його статей і вибиранні з них тільки інформації, що присвячена фразеологізмам:

- фразеологізм;
- його дефініція;

- (за наявності) ілюстрації: цитати і їх джерела.

На малюнку 2.1 можна побачити приклад статті із СУМ–11, де в одному ЛСВ зібрано декілька визначень фразеологізмів.

КУРКА, и, *жін.*

1. Свійська птиця, яку розводять на м'ясо і для одержання яєць; самка півня. *В баби було на хазяйстві четверо овець, три курки та півень* (Нечуй-Левицький, III, 1956, 323); *Курка рябенька, мов знести хотіла, бочком заглядала в шапку* (Павло Тичина, I, 1957, 248); * *У порівняннях. Невістка заслала стіл білою скатеркою.., бігає по хаті, як курка з яйцем* (Нечуй-Левицький, II, 1956, 25);

// М'ясо цієї птиці. *Мати постаралася найдкамі: вона зготувала юшку з куркою* (Панас Мирний, IV, 1955, 144).

◆ **Курка лапою загребе** — те саме, що [І] кури загребуть (див. *кури*); **Курці на сміх** — те саме, що Курям на сміх (див. *кури*); *Співцям про це сказати я мушу. Щоб курці не співать на сміх, заглянь, поет, в свою ти душу* (Володимир Сосюра, Поезія., 1961, 6); **Курці ніде клюнути** — дуже тісно, немає вільного місця. *Збив [пан] мене — курці ніде клюнути!..* (Панас Мирний, III, 1954, 164); **Мокра курка** — про жалюгідну на вигляд або безвольну, нерішучу людину. [Катерина:] *Така дівчина була [Джульєтта], а ти мокра курка* (Олександр Корнійчук, I, 1955, 307); **Сліпа курка** — про людину, що погано бачить або короткозору (у 1 знач.). **Як курка лапою** — про що-небудь виконане, виконане незграбно, неакуратно. *Оце книжки, а оце я так пишу — дивіться, гарно? Як курка лапою все одно* (Юрій Яновський, V, 1959, 160); *Взяла Катерина граблі з залізними зубками, латку озимого жита заволочила, загребла, як курка лапою...* (Степан Чорнобривець, Визволена земля, 1959, 206).

2. Страва з курячого м'яса. *Сіли за полудень... Печена курка й пироги пощезали з тарілок* (Нечуй-Левицький, III, 1956, 26).

3. *спец.* Загальна назва птахів родини курячих. *Звертає на себе увагу своїм чудовим синьо-фіолетовим забарвленням характерний птах Нової Зеландії султанська курка.., що також належить до нелітаючих птахів* (Посібник з зоогеографії, 1956, 13).

Словник української мови: в 11 томах. — Том 4, 1973. — Стор. 410.

Коментарі (0)

Малюнок 2.1. Приклад статті із СУМ-11, що містить фразеологізми.

Визначення властивостей і відношень між об'єктами. Формальним маркером, що позначає початок фразеологічної частини статті, є «◆». Кожен фразеологізм виділено жирним і відділено від визначення за допомогою тире. Ілюстрації й джерела подано сірим кольором, а самі ілюстрації ще й курсивом. На рівні HTML розмітки маркери жирності і курсиву зникають, але з'являються нові формальні ознаки:

- `◆` — початок фразеологічної частини статті;
- `...` — визначення фразеологізм;
- “ — ...” — визначення фразеологізму;
- `<i class="illus">...</i>` — приклад вживання фразеологізму;
- `(Володимир Сосюра, Поезія., 1961, 6)`

Інформація з СУМ-11 у базі даних доповнена семантичною і ономазіологічною анотацією.

Класифікація фразеологізмів за семантичною спаяністю компонентів. За основу семантичної класифікації слугує класифікація за семантичною спаяністю компонентів Петра Дудика [8], її вточнено визначенням Ігоря Мельчука [89] й Віктора Виноградова [5]. Кожному фразеологізму присвоєно один із типів:

- ідіома;
- фразеологічна єдність;
- фразеологічне сполучення;
- фразеологічний вираз;
- фразеологізоване словосполучення (сюди також віднесено терміни).

Анотування фразеологізмів за семантичною спаяністю компонентів відбувалося за таким алгоритмом:

1. *Вислів може бути самостійним реченням? або Словосполучення є відомою цитатою?*
 - a. так => «фразеологічний вираз»
 - b. ні => наступний пункт
2. *Чи значення повністю відірване від значення компонентів? або Чи вислів має архаїчні слова, які більше ніде не зустрічаються?*
 - a. так => «ідіома»
 - b. ні => наступний пункт
3. *Чи це словосполучення можна вжити у прямому значенні (навіть якщо картинка абсурдна)?*
 - a. так => «фразеологічна єдність»
 - b. ні => наступний пункт

4. Чи до залежного слова у реченні можна підібрати квазі-синоніми зі збереженням ідіоматичності? або Чи одне із повнозначних слів вжито в прямому значенні? (+ допоміжне Чи має вираз книжне походження?)
- a. так => «фразеологічне сполучення»
 - b. ні => наступний пункт
5. Чи словосполучення є терміном/назвою? або Чи всі компоненти вжиті у прямому значенні?
- a. так => «фразеологізоване словосполучення»
 - b. ні => подумати ще раз

Синоптичну схему ідеографічного опису запозичено із тезаурусу Роже [Roget Gutenberg]. Ідеографічне анотування фразеологізмів відбувалося шляхом визначення ядра значення і пошуку відповідного керівного слова у тезаурусі Роже. Перевірити правильність вибору керівного слова допомагав огляд англійських слів, наявних у знайдений статті. Якщо воно не підходило, то пошук продовжувався із використанням фразеологізмів. Після визначення підходячого керівного слова, від нього вибудовувався повний ономасіологічний шлях до найбільш загальної категорії — класу, за допомогою додаткової технічної таблиці із заготовками ономасіологічних шляхів.

Приклад опису фразеологізму:

1. Візьмімо для прикладу такий фразеологізм:

idiom	definition	citation	source
На перепутті	у стані сумнівів, хитань при виборі подальшого шляху	Не стать, не ждять в путі, на перепутті, А прокладать, торить шляхи, стежки	Петро Дорошко, Літа..., 1957, 58

Таблиця 2.1. Словникова стаття фразеологізму із таблиці «sum-idioms» для демонстрації опису.

2. Проаналізуймо його значення: “у стані сумнівів, хитань при виборі подальшого шляху”. Мова йде про невизначеність майбутнього.

3. Використаймо алгоритм опису фразеологізмів за семантичною спаяністю компонентів:

- a. *Вислів може бути самостійним реченням?* — ні;
- b. *Чи значення повністю відірване від значення компонентів?* — не зовсім;
- c. *Чи це словосполучення можна вжити у прямому значенні?* — так, а отже визначаємо, що тим фразеологізму — це фразеологічна єдність.

4. Використаймо алгоритм ідеографічного опису фразеологізмів:

- a. Нам потрібно знайти значення фразеологізму в синоптичній схемі тезауруса Роже. Отже, шукаємо одне з ключових слів, “невизначеність”, англійською серед керівних слів у схемі. Знаходимо, що “uncertainty” там присутнє і слова під ним відповідають шуканому. Впевнившись у правильності вибору керівного слова, вибудуємо весь ономасіологічний шлях у зворотному порядку до найзагальнішої категорії “інтелект”. Також додаємо номер керівного слова (475) в окремий атрибут.

5. Отже, результат опису фразеологізму такий:

idiom	class>section>subsecti on>head group>head	#	semant ic-type	definition	citation	source
На перепутті	intellect>formation of ideas>materials for reasoning>uncertainty	475	фраз. єдність	у стані сумнівів, хитань при виборі подальшого шляху	Не стай, не ждять в путі, на перепутті, А прокладать, торить шляхи, стежки	Петро Дорошко, Літа..., 1957, 58

Таблиця 2.2. Описана словникова стаття фразеологізму із таблиці «sum-idioms».

2.2. Даталогічна модель БД фразеологізмів

Етапами укладання фразеологічної бази даних були:

1. добір лексикографічного матеріалу;
2. структурування матеріалу в машино- й людиночитаний формат;

3. описати фразеологізми за семантико-ідеографічним принципом;
4. розробка способів взаємодії з базою даних.

Ці етапи передбачали виконання наступних завдань:

1. завантажити статті словника з веб-сайту;
2. відібрати ті статті та ЛСВ слів, які містять опис фразеологізмів у абзаці;
3. здобути необхідні дані з відібраних статей;
4. укласти з цих даних таблицю;
5. описати фразеологізми ідеографічно;
6. анотувати фразеологізми за семантичною спаяністю компонентів;
7. розробити користувацький інтерфейс для роботи з базою даних через командний рядок.

Добір лексикографічного матеріалу відбувався у декілька етапів, кожен із яких мав свої проміжні результати (детальніший опис див. у [пункт 2.3.1](#)). Перед початком завантаження статей, програма-скрейпер знайшла і зберегла посилання на всі слова у словнику у файлі `word_links_sum` формату JSON. Він виглядає так:

```
{
  "а": {
    "а": "/s/a",
    "абажур": "/s/abazhur",
    "абажурний": "/s/abazhurnyj",
    "абажурчик": "/s/abazhurchyk",
    "абак": "/s/abak",
    "абака": "/s/abaka",
    "абат": "/s/abat",
    "абатиса": "/s/abatysa",
```

Використовуючи посилання з файлу, програма автоматично завантажила всі статті зі словника, що загалом складає 129397 одиниць. Словникові статті збережено у JSON файлі `words_sum`, фрагмент якого виглядає так:

```
{
  "a": {
    "a": [
      "<div itemp=\\\"articleBody\\\"><p><strong class=\\\"title\\\" itemp=\\\"headline\\\">A</strong> 1, <abbr class=\\\"mark\\\" ti
      <div itemp=\\\"articleBody\\\"><p> <strong class=\\\"title\\\" itemp=\\\"headline\\\">A</strong> 2, <abbr class=\\\"mark\\\"
      <div itemp=\\\"articleBody\\\"><p> <strong class=\\\"title\\\" itemp=\\\"headline\\\">A</strong> 3 (нерідко вимовляється
      <div itemp=\\\"articleBody\\\"><p> <strong class=\\\"title\\\" itemp=\\\"headline\\\">A</strong> 4, <abbr class=\\\"mark\\\"
    ],
    "абажур": [
      "<div itemp=\\\"articleBody\\\"><p><strong class=\\\"title\\\" itemp=\\\"headline\\\">АБАЖ<span class=\\\"stressed\\\">У</span
    ],
    "абажурний": [
      "<div itemp=\\\"articleBody\\\"><p><strong class=\\\"title\\\" itemp=\\\"headline\\\">АБАЖ<span class=\\\"stressed\\\">У</span
    ],
  },
}
```

З них відібрано ті, що містять фразеологізми (їхня присутність визначена за формальним маркером — «♦»). Вони складають 9531 фразеологізм. Їх збережено у JSON файлі idioms_articles_sum, фрагмент якого виглядає так:

```
{
  "a": {
    "a": "<div itemp=\\\"articleBody\\\"><p><strong class=\\\"title\\\" itemp=\\\"headline\\\">A</strong> 1, <abbr class=\\\"mark\\\" ti
    "абордаж": "<div itemp=\\\"articleBody\\\"><p><strong class=\\\"title\\\" itemp=\\\"headline\\\">АБОРД<span class=\\\"stressed\\\">А
    "аврал": "<div itemp=\\\"articleBody\\\"><p><strong class=\\\"title\\\" itemp=\\\"headline\\\">АВР<span class=\\\"stressed\\\">А</span
    "адамів": "<div itemp=\\\"articleBody\\\"><p><strong class=\\\"title\\\" itemp=\\\"headline\\\">АД<span class=\\\"stressed\\\">А</span
    "адреса": "<div itemp=\\\"articleBody\\\"><p><strong class=\\\"title\\\" itemp=\\\"headline\\\">АДР<span class=\\\"stressed\\\">Е</span
    "аз": "<div itemp=\\\"articleBody\\\"><p><strong class=\\\"title\\\" itemp=\\\"headline\\\">АЗ</strong>, а, <abbr class=\\\"mark\\\"
  },
}
```

Останнє фільтрування статей відбувається під час їх парсингу. Залежно від повноти HTML розмітки, а саме наявності розмітки для ілюстрацій, структуровані дані про фразеологізми збережено до таблиці idioms_sum_table або idioms_sum_table_to_clean формату TSV (tab-separated values). Обидві мають такі атрибути — усі текстового типу:

- idiom /фразеологізм/;
- definition /визначення/;
- citation /ілюстрація/;
- source /джерело ілюстрації/.

Статей із повною HTML розміткою у таблиці idioms_sum_table вийшло 1613.

Приклад рядка-статті:

idiom	definition	citation	source
Довга пісня	про те, що займає великий відрізок часу, чого не можна швидко зробити, виконати, розповісти	Ага, що це у вас тут з Ганною скоїлося? — дивлячись на сестру, запитав Данько. — Довга пісня, — жваво заговорила Вутанька	Олесь Гончар, II, 1959, 142

Таблиця 2.3. Рядок із таблиці idioms_sum_table.

Усі інші статті (7918 одиниць) опинилися у таблиці idioms_sum_table_to_clean того ж TSV формату. Вона потребує додаткової ручної або автоматизованої обробки з відділення ілюстрацій від визначень. Приклад рядка із неї:

idiom	definition	citation	source
У (в) добрі руки	,із сл дістатися, потрапити і т. ін. — до розумної, дбайливої, надійної людини. Вони так зрадили, одібравши книжки, ..і я ще більше запевнився, що книжки дістались у добрі руки		(Михайло Коцюбинський, III, 1956, 122)

Таблиця 2.4. Рядок із таблиці idioms_sum_table_to_clean.

Для подальшої обробки й укладання бази даних використано лише таблицю «idioms_sum_table», бо вона не потребує цієї обробки. Тим не менш, велика частина з утворених статей містила дублети — фразеологізми без опису, що містять лише посилання на інші статті.

Таблицю «idioms_sum_table» з формату TSV перетворено у таблицю Google Sheets (аркуш «sum-idioms» таблиці «idioms»). Після цього виявлено формальні фразеологізми-дублети двох типів:

- статті, що мають лише фразеологізм і посилання винесено до аркушу «див»; приклад статті з таблиці «див»:

idiom	definition
Авгісві конюшні	дивавгіїв

Таблиця 2.5. Приклад статті, що містить у дефініції лише посилання.

- статті, що мають не тільки посилання на дублет, а й ілюстрацію, винесено до аркушу «те саме»; приклад статті з таблиці «те саме»:

idiom	definition	citation	source
-------	------------	----------	--------

Влетіти в [добру] копійчину	те саме, що Влетіти в [добру] копійку (дивкопійка)	Ну й понавиписували ж отаких машин [екскаваторів]! Мабуть, у добру копійчину влетіло	Гордій Коцюба, Нові береги, 1959, 85
-----------------------------	--	--	--------------------------------------

Таблиця 2.6. Приклад статті, що містить дефініцію-посилання, але має й ілюстрацію.

Наведені дублети не мають визначень, які необхідні для опису семантики фразеологізмів, тому їх відкинуто. У підготовленій таблиці «sum-idioms» створено додаткові атрибути для ідеографічного опису фразеологізмів та анотації за семантичною спаяністю компонентів:

- *class>section>subsection>head group>head* (ономасіологічний шлях із тезауруса Роже; текстові дані);
- # (номер останнього head із тезауруса Роже; теж текстові дані, бо трапляються номери на кшталт 864a);
- *semantic type* (тип фразеологізму за семантичною спаяністю компонентів; текстові дані).

Синоптична схема ідеографічного опису фразеологізмів має ієрархічну структуру, але в самій таблиці фразеологізмів це ніяк не виражено. База даних міститься в одній таблиці і не має ані внутрішніх, ні зовнішніх реляційних зв'язків. Приклад рядка з анотованої бази даних:

idiom	class>section>subsection>head group>head	#	semantic -type	definition	citation	source
Стріляний (старий) горобець	volition>individual>voluntary action>complex voluntary action>proficient	700	фраз. єдність	про того, хто мав великий досвід, кого важко обдурити	Адже ми з вами, Мар'я Олексіївна, старі горобці, нас на полову не піддуриш	Микола Чернявський, Що робити?, перекл. Дукина, 1936, 135

Таблиця 2.7. Рядок із анотованої таблиці «sum-idioms».

Після завершення ідеографічного опису фразеологізмів, таблицю було відсортовано за номерами понять у тезаурусі Роже. Це дозволило виявити у даних складніші дублети за такими ознаками:

- однакові фразеологізми, ідентичні за змістом визначення (несуттєво розходяться по формі);
- однакові фразеологізми, визначення одного з яких відсутнє і натомість є посиланням на іншу статтю;
- фразеологізми і їхні визначення відрізняються винятково морфологічно;
- змішаного типу.

Усі знайдені типи дублетів видалено. Після цього у таблиці «sum-idioms» залишилося 840 статей. Наводжу декілька прикладів, включаючи дублети, винесені в окремі таблиці «див» і «те саме» [СУМ-11]:

1. однакові фразеологізми, визначення яких несуттєво розходяться по формі і є ідентичними за змістом:

№	Фразеологізм	Визначення	Ілюстрація	Цитування
1.	Тягти (терти) лямку	взятися за важку, виснажливу довготривалу роботу	Старий Серебренников посідав на дільниці місце робочого вола. Неприємна й брудна робота завжди потрапляла до нього, і він покірливо тяг лямку	Юрій Яновський, II, 1954, 101
	Тягти (терти) лямку	виконувати важку, виснажливу роботу, жити у важких умовах	[Харько:] Бачу, що моему горю Вже не допомогти, Певно, до кінця прийдеться Лямку цю тягти	Марко Кропивницький, V, 1959, 609
2.	Ударити в закаблуки	пуститися в танець; затанцювати	Як сонні, ходять, сплівши руки, а тут би — вихором пройтись, ударить шквально в закаблуки	Іван Гончаренко, Вибр., 1959, 361
	Ударити в закаблуки	танцювати з запалом		
3.	Сидіти (сісти, жити) хутором	жити на відокремленій земельній ділянці	Саме проти Черкас, нижче Домонтова, сидів хутором над Дніпром Максим Тримач	Марко Вовчок, I, 1955, 89
	Сидіти (сісти,	оселятися на	На північ від Богуслава, в	Петро

№	Фразеологізм	Визначення	Ілюстрація	Цитування
	жити) хутором	відокремленій земельній ділянці	урочищі Маслів Став, жив хутором старий козак Добридень	Панч, Гомон. Україна, 1954, 5

Таблиця 2.8. Приклади варіативних дублів фразеологізмів.

2. однакові фразеологізми, визначення одного з яких відсутнє і натомість є посиланням на іншу статтю:

Фразеологізм	Визначення	Ілюстрація	Цитування
Упливати (упливти) за водою	зменшуватися, слабшати, зникати	Поволі вся лють Рідкодуба танула, за водою кудись упливала	Іван Кириленко, Вибр., 1960, 231
Упливати (упливти) за водою	те саме, що Багато (чимало) води утекло (утече) (диввода)	Упливло за водою його здоров'я	Леонід Смілянський, Сад, 1952, 248

Таблиця 2.9. Приклад дублюючих статей із посиланнями.

3. морфологічні дублети:

Фразеологізм	Визначення	Ілюстрація	Цитування
Зметаний на живу нитку	зшитий як-небудь, аби трималося	Увечері тільки Василь сам про себе лаяв одержу, що на його вже не приходилася та чимало і зносилася: рукава трохи не доходили до ліктів, стан насеред спини, одна пола розірвана зметана на живу нитку	Панас Мирний, IV, 1955, 117
Зметати на живу нитку	зшити як-небудь, аби трималося		

Таблиця 2.10. Приклад морфологічних дублетів.

4. дублети змішаних типів:

№	Фразеологізм	Визначення	Ілюстрація	Цитування
1.	Натискати (натискувати, натиснути) [на] кнопку (на всі кнопки)	дивнатискати		
	Натискати (натискувати, натиснути) [на] кнопку (на всі кнопки)	докладати всіх зусиль для швидкого виконання чого-небудь	Куцевич досконало знав, коли і де саме треба натиснути потрібну кнопку, щоб злагоджено зарухались всі великі і малі механізми складного господарства інституту, який він очолював	Натан Рибак, Час, 1960, 488
	Натискати (натискувати, натиснути) [на] кнопку (на всі кнопки)	уживати необхідних заходів для досягнення своєї мети		
	Натискати (натискувати, натиснути) на всі педалі	докладати всіх зусиль для швидкого виконання чого-небудь	Вони поїхали додому. Ліза натискала на всі педалі	Микола Руденко, Вітер., 1958, 163
2.	Ставати (стати) цапа	зводитися на задні ноги; здіблюватися	Велетенський жеребець не зносив людського духу. Він іржав і ставав цапа, як тільки бачив коло себе живу людину	Павло Загребельний, Шепіт, 1966, 58
	Ставати (стати) цапки	те саме, що Ставати (стати) цапа (дивцапа)	Осадив [Наливайко] білокопитого коня, що той аж цапки став	Іван Ле, Наливайко, 1957, 100
	Становитися цапа	дивстановитися		
	Становитися цапа	рішуче протидіяти чому-небудь, заперечувати щось	Кінь змилений став цапа	Андрій Головка, І, 1957, 72
	Становитися цапки	дивстановитися		
	Становитися цапки	те саме, що Ставати (стати) цапа (дивцапа)	Вона [зима] не хотіла скоритися: ставала цапки, мела лютою хурделицею..,	Григорій Епик, Тв., 1958, 164

№	Фразеологізм	Визначення	Ілюстрація	Цитування
			передчуваючи свій неминучий кінець	

Таблиця 2.11. Приклади змішаних дублетів.

2.3. Програмне забезпечення

Створено сторінку проєкту на GitHub [посилання: <https://github.com/bohdana-ivakhnenko/ukrainian-idioms>], де можна отримати вільний доступ до кодів та анотованої бази даних.

2.3.1. Програма для завантаження словника

Для реалізації програми-скрейпера словника використано чотири бібліотеки:

- `collections` — дозволяє використовувати просунутий тип даних `dict`, що автоматизує додавання нових елементів;
- `requests` — здійснює запити до сторінок сайту;
- `bs4` — парсить HTML розмітку веб-сторінок, що дозволяє автоматизовано в них орієнтуватися;
- `json` — дозволяє читати й зберігати файли у форматі JSON.

На початку програми ці бібліотеки імпортовано:

```
from collections import defaultdict
import requests
from bs4 import BeautifulSoup
import json
```

Програма завантажує словник у декілька етапів. Спершу вона знаходить посилання на всі слова у словнику, використовуючи для цього літерний покажчик, і зберігає їх організованими за літерами у JSON файл `word_links_sum`. Цей функціонал реалізовано у програмі цим кодом:

```
def get_links_dict(source_link, tag="vkazivnyk"):
    page = BeautifulSoup(requests.get(source_link).content,
features="html.parser")
    parts = page.find("div", {"id": tag}).find_all("a")
```

```

        links = {parts[num].text: parts[num]["href"]} for num in
range(len(parts))}
    return links

dict_link = "http://sum.in.ua"
root_link = "http://sum.in.ua/vkazivnyk"
letters_links = get_links_dict(root_link)
letter_comb_links = {}
words_links = defaultdict(dict)

for letter, link_letter in letters_links.items():
    letter_comb_links[letter] = get_links_dict(dict_link +
link_letter)
    for letter_comb, link_comb in letter_comb_links[letter].items():
        if link_comb.startswith("/s"):
            words_links[letter][letter_comb] = link_comb
        else:
            structured_word_links = get_links_dict(dict_link +
link_comb)
            words_links[letter].update(structured_word_links)

words_links["dict_link"]["main"] = dict_link
print(words_links)

with open("word_links_sum.json", "w", encoding="utf-8") as file:
    json.dump(words_links, file, ensure_ascii=False, indent=4)

```

Далі програма відкриває щойно збережений `word_links_sum` і створює на його основі два нових файли такої ж структури. У першому файлі посилання на статтю замінене на саму словникову статтю у формі списку лексико-семантичних варіантів (ЛСВ) слова. Текст словника подано у HTML розмітці. Ці дані збережено у JSON файлі `words_sum`. У другому файлі програма замість посилань на статті зберігає тільки ті ЛСВ, які містять у собі фразеологізми. Словникові статті, де не описано фразеологізмів, відсіяно повністю. Ці дані збережено у JSON файлі `idioms_articles_sum`.

Завантаження, фільтрування і збереження словникових статей у файлах `words_sum` і `idioms_articles_sum` реалізовано за допомогою цього коду:

```

with open("words_sum.json", "w", encoding="utf-8") as file:

```

```

word_articles = json.loads(file.read())

with open("word_links_sum.json", "r", encoding="utf-8") as file:
    words_links = json.loads(file.read())

with open("idioms_articles_sum.json", "r", encoding="utf-8") as file:
    articles_with_idioms = json.loads(file.read())

for letter, words in list(words_links.items()):
    if letter in word_articles.keys():
        continue
    word_articles[letter] = defaultdict(dict)
    articles_with_idioms[letter] = defaultdict(dict)

    print("START", letter)

    try:
        for word, link in list(words.items()):
            print(word)
            response = requests.get(dict_link + link).content
            page = BeautifulSoup(response, features="html.parser")
            article = page.find_all("div", {"itemprop":
"articleBody"})

            word_articles[letter][word] = [str(part) for part in
article]
            idiom = [str(part) for part in article if "♦" in
str(part)]
            if idiom:
                articles_with_idioms[letter][word] = idiom

    except ConnectionResetError:
        print("ERROR:", letter)

    with open("words_sum.json", "w", encoding="utf-8") as file:
        json.dump(word_articles, file, ensure_ascii=False, indent=4)

    with open("idioms_articles_sum.json", "w", encoding="utf-8") as
file:
        json.dump(articles_with_idioms, file, ensure_ascii=False,
indent=4)

```

Посилання на повний код програми на порталі GitHub подано у додатку 1.

2.3.2. Програма для парсингу статей словника і укладання бази даних

Програма парсить HTML розмітку статей із електронної версії СУМ-11, здобуваючи з неї тільки ті дані, що визначають фразеологізми (сам фразеологізм, його визначення й ілюстрації). Матеріалом для програми є JSON файл `idioms_articles_sum`, створений на попередньому етапі. Він містить статті або окремі ЛСВ, де описано фразеологізми. HTML розмітка словника не в усіх випадках є повною — у деяких статтях вона не описує ілюстрацій, що унеможлиблює їх відділення від визначень. З цієї причини програма укладає дві таблиці:

- `idioms_sum_table`, що містить статті з усіма необхідними складовими і є основою робочої бази даних;
- `idioms_sum_table_to_clean`, що містить статті без розмічених ілюстрацій, а тому потребує виправлення парсингу перед використанням у базі даних.

Для реалізації задачі програми застосовано три зовнішні бібліотеки:

- `json` — дозволяє читати й зберігати файли у форматі JSON;
- `csv` — автоматично читає та зберігає файли у форматі TSV (tab separated values);
- `bs4` — парсить HTML розмітку веб-сторінок, що дозволяє автоматизовано в них орієнтуватися.

На початку програми бібліотеки імпортовано:

```
import json
import csv
from bs4 import BeautifulSoup
```

Основні дії програми з читання файлу з матеріалами, запуском парсингу та зберіганням зібраних структурованих даних у нових файлах втілені цим кодом:

```
with open("idioms_articles_sum.json", "r", encoding="utf-8") as file:
    dictionary = json.loads(file.read())
```

```

lines = [("idiom", "definition", "citation", "source")]

for letter, articles in tuple(dictionary.items()):
    for word, definition in tuple(articles.items()):
        print(word)
        idioms = parse(definition)
        if idioms:
            for idiom in idioms:
                if idiom['illustrations']:
                    lines.append((idiom['idiom'], idiom['definition'],
idiom['illustrations'][0]['citation'],
idiom['illustrations'][0]['source']))
                else:
                    lines.append((idiom['idiom'], idiom['definition'],
None, None))
            print()

with open("idioms_sum_table.tsv", "w", encoding="utf-8") as file:
    writer = csv.writer(file, delimiter='\t', lineterminator='\n')

    for line in lines:
        writer.writerow(line)

with open("idioms_sum_table_to_clean.tsv", "w", encoding="utf-8") as
file:
    writer = csv.writer(file, delimiter='\t', lineterminator='\n')

    for line in lines:
        writer.writerow(line)

```

Цей код оперує абстрактною функцією `parse`, що приховує усю логіку. Ця функція визначає тип статті за наявністю розмітки для ілюстрацій, знаходить частину ЛСВ чи всієї статті, де описано фразеологізм, та викликає ще одну функцію (`collect_article`), яка досліджує знайдений абзац із фразеологізмом і збирає необхідні дані у типову структуру. Код функції `parse` виглядає так:

```

def parse(text):
    article = BeautifulSoup(text, features="html.parser")

    if article.find("i", {"class", "illus"}):
        dirty = False
    else:
        dirty = True
    print("dirty:", dirty)

```

```

if article.find("p", {"class", "znach"}):
    paragraphs = [article.find("p", {"class", "znach"})]
    paragraphs.extend(paragraphs[0].findAllNext("p", {"class",
"znach"}))
elif article.find("p"):
    paragraphs = [article.find("p")]
else:
    print("Other problem")
    return

idioms = []
for paragraph in paragraphs:
    if paragraph.find("span", {"class", "diam"}) and not dirty:
        idioms_ = collect_article(paragraph, dirty)
        idioms.extend(idioms_)
return idioms

```

Функція `collect_article` отримує на вхід шматочки необхідного абзацу й ітерується на них, повертаючи на виході список структурованих фразеологізмів.

Визначено чотири булеві змінні, що визначають стани функції:

`diam`, `illus`, `citation` та `explanation_next`. Всі вони на початку виклику функції мають значення хиби. Поки програма не доходить до шматочка, що містить знак "◆", який у словнику є міткою фразеологізмів, нічого не відбувається. Коли знак "◆" виявлено, змінна `diam` набуває істинного значення. Під час наступної ітерації зберігається текст фразеологізму і надається істинне значення змінній `explanation_next`, яка означає, що далі зберігатиметься визначення фразеологізму. Для дефініції немає спеціальної розмітки, тому програма вважає кожен шматочок тексту частиною визначення, поки не натрапить на розмітку для ілюстрації. Як тільки знайдено ілюстрацію, значення змінної `explanation_next` змінюється на хибу, змінна `illus` набуває істинного значення, а сама ілюстрація зберігається. Після цього програма знаходить цитування джерела, зберігає його прив'язаним до попередньої ілюстрації у список ілюстрацій фразеологізму і змінює значення змінної `citation` на істинне. Якщо фразеологізм має не одну ілюстрацію, то наступні збережуться без помилок, бо програма знову знайде відповідну розмітку.

Якщо словникова стаття надає тлумачення декількох фразеологізмів, то програма виявить це за розміткою і перед аналізом наступного фразеологізму збереже попередній у список, а також оновить усі технічні змінні: `illus`, `citation` та `explanation_next` знову отримають значення хибі, а список із частинами визначення і словник із ілюстрацією буде очищено. Далі процес буде таким самим, як і для першого фразеологізму, і буде повторюватися поки не дійде до кінця абзацу. Як тільки це відбудеться, останній знайдений фразеологізм буде додано до списку і повернуто.

Якщо стаття не містить розмітки для ілюстрацій, то процес обробки відрізняється тим, що ілюстрація потрапляє до визначення, а в ілюстраціях зберігається лише цитування джерела. У словнику може траплятися по декілька ілюстрацій на фразеологізм, але незалежно від того, чи є розмітка для ілюстрацій, в таблиці збережено лише по одній заради надання даним одноманістости.

Код функції `collect_article` виглядає так:

```
def collect_article(paragraph, dirty):
    diam = False
    illus = False
    citation = False
    explanation_next = False

    idiom = ""
    pieces = []
    illustration = {"citation": "",
                   "source": ""}
    illustrations = []
    phrases = []

    if not dirty:
        for index, child in enumerate(paragraph.children):

            string = child.text.strip(" \n-.").replace("\n", " ")
            if diam:
                if '<span class="s">' in str(child):
                    citation = True
                    illustration["source"] = string
                    illustrations.append(illustration)
                    illustration = {"citation": "",
```

```

        "source": ""}
    if '<i class="illus">' in str(child):
        illus = True
        illustration["citation"] = string
        explanation_next = False
    if explanation_next and not (illus or citation or
                                '<span class="z">' in
str(list(paragraph.children)[index])):
        pieces.append(string)
    if '<span class="z">' in str(child):
        if pieces and ((illus and citation)
                        or '<span class="z">' in
str(list(paragraph.children)[index])
                                or index + 1 ==
len(list(paragraph.children))):
        phrase = {"idiom": idiom,
                  "definition": "".join(pieces),
                  "illustrations":
illustrations.copy()}
        phrases.append(phrase)
        idiom = ""
        pieces.clear()
        illustrations.clear()
        illus = False
        citation = False
        explanation_next = False

        idiom = string
        explanation_next = True
    elif not diam and child.text == "◆":
        diam = True

    phrase = {"idiom": idiom,
              "definition": "".join(pieces),
              "illustrations": illustrations.copy()}
    phrases.append(phrase)

    return phrases

else:
    second_def = False

    for index, child in enumerate(paragraph.children):
        string = child.text.strip("\n-").replace("\n", " ")

        if diam:
            if '<span class="s">' in str(child) or \

```

```

        index in [len(list(paragraph.children))-1,
len(list(paragraph.children))-2] \
            or (pieces and '<span class="z">' in
str(list(paragraph.children)[index+1]))):
        illustration["source"] = string
        illustrations.append(illustration)
        explanation_next = False

        if second_def or index in
[len(list(paragraph.children))-1, len(list(paragraph.children))-2]:
        phrase = {"idiom": idiom,
                "definition": "".join(pieces),
                "illustrations": illustrations}
        phrases.append(phrase)
        break

    if illustrations and '<span class="z">' in str(child):
        second_def = True
        phrase = {"idiom": idiom,
                "definition": "".join(pieces),
                "illustrations": illustrations}
        phrases.append(phrase)
        idiom = str
        pieces = []
        illustration = {"citation": "",
                "source": ""}
        illustrations = []

    if explanation_next and not citation:
        pieces.append(string)

    if '<span class="z">' in str(child):
        idiom = string
        explanation_next = True

    elif not diam and child.text == "◆":
        diam = True

    return phrases

```

Посилання на повний код програми на порталі GitHub подано у додатку 2.

2.3.2. Програма інтерфейсу користувача

Програма інтерфейсу користувача забезпечує пошук у базі даних за класами і секціями Роже, виведення результатів пошуку на екран і, за потреби, збереження

цих результатів. Для реалізації цієї задачі використано декілька зовнішніх бібліотек:

- `os` — дозволяє переглядати назви файлів, що лежать у заданій теці;
- `csv` — автоматично читає та зберігає файли у форматі TSV (tab separated values);
- `re` — дозволяє використовувати регулярні вирази;
- `tabulate` — форматує дані у таблицю для естетичного виведення їх на екран;
- `collections` — дозволяє використовувати просунутий тип даних `dict`, що автоматизує додавання нових елементів;
- `alphabet_detector` — автоматично визначає алфавіт, яким написана стрічка тексту.

На початку програми найперше імпортуються бібліотеки:

```
import os
import csv
import re
from tabulate import tabulate
import collections
from alphabet_detector import AlphabetDetector
alphabet = AlphabetDetector()
```

Після цього визначається клас `IdiomThesaurus`, а в кінці його виклик, що запускає інтерфейс відразу після початку роботи програми:

```
if __name__ == '__main__':
    IdiomThesaurus()
```

Принцип роботи інтерфейсу полягає в тому, щоб здобути від користувача клас і секцію, у яких необхідно здійснити пошук, дізнатися чи необхідно зберегти результат у файл, а тоді власне виконати пошук, згрупувати результати і показати їх користувачеві. Усі деталі роботи інкапсульовані у класі.

У базі даних ідеографічний опис фразеологізмів виконано англійською мовою. Класи і секції збережено так:

```

_classes_sections = {
    'abstract relations': ('existence', 'relation', 'quantity',
'number', 'time', 'change',
'causation'),
    'affections': ('affections in general', 'personal',
'sympathetic', 'moral', 'religious'),
    'intellect': ('formation of ideas', 'communication of ideas'),
    'matter': ('generally', 'inorganic', 'organic'),
    'space': ('generally', 'dimensions', 'form', 'motion'),
    'volition': ('individual', 'intersocial')
}

```

Оскільки інтерфейс виконано українською мовою, то додатково задано українськомовні відповідники:

```

_en_uk = {
    'abstract relations': 'абстрактні відношення',
    'existence': 'існування',
    'relation': 'відношення',
    'quantity': 'кількість',
    'order': 'порядок',
    'number': 'число',
    'time': 'час',
    'change': 'зміна',
    'causation': 'причиновість',
    'affections': 'емоції та моральні відчуття',
    'affections in general': 'відчуття загалом',
    'personal': 'особистісні відчуття',
    'sympathetic': 'відчуття до інших',
    'moral': 'мораль',
    'religious': 'релігія',
    'intellect': 'інтелектуальні здібності',
    'formation of ideas': 'формування ідей',
    'communication of ideas': 'висловлювання ідей',
    'matter': 'матерія',
    'generally': 'загалом',
    'inorganic': 'неорганічна',
    'organic': 'органічна',
    'space': 'простір',
    'dimensions': 'розміри',
    'form': 'форма',
    'motion': 'рух',
    'volition': 'волевиявлення',
    'individual': 'індивідуальне',
    'intersocial': 'соціальне'
}
_uk_en = {value: key for key, value in _en_uk.items()}

```

Робота інтерфейсу організована у методи `__init__` класу `IdiomThesaurus`. Він визначає файл бази даних, основні змінні, якими оперують методи, і запускає всі процеси. Метод `__init__` виглядає так:

```
def __init__(self, file='sum-idioms-annotated.tsv'):
    self.file = file
    self.database = []
    self.columns = {}
    self.initialize()

    self.class_ = ''
    self.section_ = ''
    self.save = False
    self.search_tag = ''
    self.get_query()

    self.idioms = []
    self.get_idioms()

    self.idiom_groups = collections.defaultdict(list)
    self.group_idioms()

    self.provide_results()

    if self.next_round():
        self.__init__(file=self.file)
```

Спершу метод `__init__` викликає метод `initialize`, який відкриває файл бази даних, зчитує його, автоматично визначає атрибути бази даних:

```
def initialize(self):
    with open(self.file, 'r', encoding='utf-8') as file:
        reader = csv.reader(file, delimiter='\t')
        self.database = [line for line in reader]
        self.columns = {column: index for index, column in
            enumerate(self.database.pop(0))}
```

Далі метод `__init__` визначає змінні, що використовуються для налаштування пошуку: `class`, `section`, `save`, `search_tag`. `Class` і `section` є класом і секцією фразеологізмів у синоптичній схемі тезауруса Роже. `Save` є булевою змінною, що визначає чи зберігати результат пошуку у файл. `Search_tag` використовується за умови істинного значення змінної `save` і використовується для генерації назви

файлу. Після визначення змінних викликається метод `get_query`. Цей метод містить пряму рекурсію для забезпечення стабільної перевірки введеної користувачем інформації (метод буде викликати себе знову і знову доти, доки користувач не введе підтримувані значення). Спершу метод друкує доступні класи і просить користувача обрати якийсь із них або всі. Це можна зробити, написавши назву класу українською чи англійською, його номер чи щось схоже на «шукати в усіх класах». Якщо відповідь відповідає одній із можливих, то метод переходить до визначення секції за тим же принципом, а якщо ні, то запускається знову і ще раз запитує про клас. Якщо на рівні класу обрано пошук в усіх класах, то секція також автоматично набуває цього ж значення. Коли і клас, і секцію для пошуку визначено, то метод запитує користувача про тег для пошуку. Якщо користувач його не вводить, то результат не буде збережено, а тільки виведено на екран. Метод `get_query` виглядає так:

```
def get_query(self):
    if not self.class_:
        classes = self._classes_sections.keys()
        print('Виберіть номер класу або напишіть сам клас:')
        [print(f'{index + 1}. {IdiomThesaurus._en_uk[class_]}')
         for index, class_ in enumerate(classes)]
        print(f'{len(classes) + 1}. шукати в усіх класах')
        answer_class = input().strip()
        if self.answer_not_appropriate(answer_class, classes):
            print('\nВи ввели неправильне числове значення або
вказали неіснуючий клас.')
            print('Спробуйте ще раз!', end='\n\n')
            self.get_query()
        else:
            if answer_class.isdigit():
                self.class_ = list(classes)[int(answer_class) - 1]
            elif alphabet.only_alphabet_chars(answer_class,
"CYRILLIC") \
                and answer_class in
IdiomThesaurus._uk_en.keys():
                self.class_ = IdiomThesaurus._uk_en[answer_class]
            elif alphabet.only_alphabet_chars(answer_class,
"LATIN") \
```

```

                                                    and answer_class in
IdiomThesaurus._en_uk.keys():
    self.class_ = answer_class
    else:
        self.class_ = "ysi"
        self.section_ = "ysi"

    if self.class_ and not self.section_:
        sections = IdiomThesaurus._classes_sections[self.class_]
        print('Можете також обрати секцію:')
        [print(f'{index + 1}. {IdiomThesaurus._en_uk[section]}')]
for index, section_ in enumerate(sections)
    print(f'{len(sections) + 1}. шукати в усіх секціях')
    answer_section = input().strip()
    if self.answer_not_appropriate(answer_section, sections):
        print('\nВи ввели неправильне числове значення або
вказали неіснуючу секцію.')
        print('Спробуйте ще раз!', end='\n\n')
        self.get_query()
    else:
        if answer_section.isdigit():
            self.section_ = list(sections)[int(answer_section)
- 1]
            elif alphabet.only_alphabet_chars(answer_section,
"CYRILLIC") \
                                                    and answer_section in
IdiomThesaurus._uk_en.keys():
                                                    self.section_ =
IdiomThesaurus._uk_en[answer_section]
            elif alphabet.only_alphabet_chars(answer_section,
"LATIN") \
                                                    and answer_section in
IdiomThesaurus._en_uk.keys():
                self.section_ = answer_section
            else:
                self.section_ = "ysi"

    print("Напишіть тег для пошуку, якщо хочете зберегти результат
у файл:")
    answer_tag = input().strip()
    self.search_tag = answer_tag

    if not (self.class_ and self.section_):
        print('Упссс, щось пішло не так, спробуйте ще раз!')
        self.get_query()
return

```

Для перевірки відповіді користувача метод `get_query` викликає інший метод — `answer_not_appropriate`, що повертає істину, коли відповідь є не відповідає вимогам. `Answer_not_appropriate` перевіряє щоб відповідь кирилицею належала до множини можливих українськомовних відповідників назв класів і секцій синоптичної схеми тезауруса Роже, а якщо вона латинкою, то множини назв класів і секцій синоптичної схеми тезауруса Роже в оригіналі (англійською). Якщо це не відповідає дійсності, то перевіряється чи користувач обрав шукати всюди, а наостанок — чи не обрав щось із перерахованого вище цифрою. Метод виглядає так:

```
@staticmethod
def answer_not_appropriate(answer, categories):
    all_options = ['шукати в усіх класах', 'в усіх класах',
                  'усі', 'всі', 'шукати в усіх секціях',
                  'в усіх секціях', 'усюди', 'шукати в усіх',
                  'шукати всюди', 'в усіх', 'всюди']

    words = answer.isalpha() and
            not (answer.isdigit() or
                (answer.lower() in
                 list(IdiomThesaurus._uk_en.keys()) +
                 list(IdiomThesaurus._en_uk.keys()) and
                 IdiomThesaurus._uk_en[answer].lower() in
                 categories) or
                answer in categories or
                answer.lower() in all_options)
    number = (answer.startswith('-') or
              (answer.isdigit() and
               not int(answer) <= len(categories)+1))
    unknown = not (answer.isalpha() or answer.isdigit())
    return words or number or unknown
```

Після визначення необхідних користувачеві класу і секції метод `__init__` створює пустий список для фразеологізмів і викликає метод `get_idioms`, який виймає з бази даних всі фразеологізми, що відповідають запиту користувача, тобто їхні атрибути `class` і `section` відповідають тим, що були задані раніше. Користувачеві надаються не всі атрибути з бази даних, бо деякі з них є технічними. Цей метод обирає які атрибути обрати для демонстрації:

- head group>head (кінець ономасіологічного шляху, що уточнює загальні значення, за яким відбувається пошук);
- idiom (сам фразеологізм);
- definition (визначення фразеологізму);
- semantic-type (тип фразеологізму за семантичною спаяністю компонентів);
- citation (ілюстрація, що демонструє приклад вживання фразеологізму);
- source (джерело, звідки наведено ілюстрацію).

Атрибути class, section і subsection є більш технічними, оскільки також застосовуються для сортування фразеологізмів.

Метод get_idioms виглядає так:

```
def get_idioms(self):
    full_path_id = self.columns['head group>head']
    idiom_id = self.columns['idiom']
    definition_id = self.columns['definition']
    type_id = self.columns['semantic-type']
    citation_id = self.columns['citation']
    source_id = self.columns['source']

    class_id = self.columns['class']
    section_id = self.columns['section']
    subsection_id = self.columns['subsection']

    if self.class_ == 'yci':
        self.idioms = [(line[full_path_id],
                        line[idiom_id],
                        line[definition_id],
                        line[type_id],
                        line[citation_id],
                        line[source_id],
                        line[class_id],
                        line[section_id],
                        line[subsubsection_id])
                       for line in self.database]
    elif self.section_ == 'yci':
        self.idioms = [(line[full_path_id],
                        line[idiom_id],
                        line[definition_id],
                        line[type_id],
                        line[citation_id],
                        line[source_id],
```

```

        line[class_id],
        line[section_id],
        line[subsection_id])
    for line in self.database
        if line[self.columns['class']] ==
self.class_]
    else:
        self.idioms = [(line[full_path_id],
            line[idiom_id],
            line[definition_id],
            line[type_id],
            line[citation_id],
            line[source_id],
            line[class_id],
            line[section_id],
            line[subsection_id])
            for line in self.database
                if line[self.columns['class']] ==
self.class_ and line[self.columns['section']] == self.section_]

```

Після отримання фразеологізмів, Оскільки в класифікації Роже існує ще декілька рівнів ієрархії, то було вирішено групувати результати за підсекціями, щоб забезпечити легше сприйняття інформації. Цим займається метод `group_idioms`. Фразеологізми групувано спершу за класами (важливо, коли пошук найбільш загальний), тоді за секціями (важливо, коли користувач хоче отримати всі фразеологізми у класі), а врешті за підсекціями (актуально завжди). Код методу `group_idioms` виглядає так:

```

def group_idioms(self):
    self.idiom_groups = collections.defaultdict(list)
    for idiom_line in self.idioms:
        class_ = idiom_line[-3]
        self.idiom_groups[class_].append(idiom_line)

    for class_, idiom_lines in self.idiom_groups.items():
        sections = collections.defaultdict(list)
        for idiom_line in idiom_lines:
            section_ = idiom_line[-2]
            sections[section_].append(idiom_line)
        self.idiom_groups[class_] = sections

    for class_, sections in self.idiom_groups.items():
        for section_, idiom_lines in sections.items():

```

```

subsections = collections.defaultdict(list)
for idiom_line in idiom_lines:
    subsection_ = idiom_line[-1]
    subsections[subsection_].append(idiom_line)
self.idiom_groups[class_][section_] = subsections

```

На останньому етапі роботи метод `__init__` запускає метод `provide_results`, що організовує передачу зібраних і згрупованих результатів пошуку користувачеві. Цей метод ітерується за всіма наявними у відібраних фразеологізмах класами, секціями та підсекціями, збираючи всі наявні у кожній підсекції фразеологізми у таблицю, і виводить її на екран, попередньо підписуючи за якої підсекцією вони згруповані. Якщо користувач надав тег для змінної `search_tag`, то ту саму інформацію також збережено у файлі формату TSV. Метод `provide_results` виглядає так:

```

def provide_results(self):
    if self.search_tag:
        headers = ('id', 'class', 'section',
                  'subsection', 'head group>head',
                  'idiom', 'definition', 'semantic type',
                  'citation', 'source')
        file = open(f"queries/{self.get_file_name()}.tsv", "w")
        writer = csv.writer(file,
                            delimiter='\t',
                            lineterminator='\n')
        writer.writerow(headers)

    for class_, sections in self.idiom_groups.items():
        for section_, subsections in sections.items():
            for subsection_, idioms in subsections.items():
                if subsection_:
                    headers = ('id', 'head group>head',
                              'idiom', 'definition',
                              'semantic type',
                              'citation', 'source')
                    print(f"{class_} > {section_} >
{subsection_}").upper())
                else:
                    headers = ('id', 'head',
                              'idiom', 'definition',
                              'semantic type',
                              'citation', 'source')
                    print(f"{class_} > {section_}").upper())

```

```

lines_to_show = [[index+1] + list(idiom[:6])
                  for index, idiom
                  in enumerate(idioms)]
print(tabulate(lines_to_show,
               headers=headers,
               tablefmt="fancy_grid",
               stralign="left",
               maxcolwidths=[None, None, 30,
                              40, 20, 50, 34]))

print()

if self.search_tag:
    lines_to_write = [[index + 1, class_,
                      section_, subsection_]
                     + list(idiom[:6])
                     for index, idiom
                     in enumerate(idioms)]

    for line in lines_to_write:
        writer.writerow(line)

if self.search_tag:
    file.close()

```

Кожен результат зберігається в файл із унікальною назвою. Назва має такі елементи, об'єднані нижнім підкресленням:

- idioms;
- пошуковий тег користувача;
- порядковий номер файлу з таким пошуковим тегом.

Для генерації унікальної назви в одноманітному форматі `provide_results` викликає метод `get_file_name`, який у свою чергу використовує метод `get_doubling_files`. `Get_doubling_files` шукає серед файлів у теці `queries`, де зберігаються результати пошуку, всі файли з таким самим пошуковим тегом, що його ввів користувач, і повертає всі знайдені порядкові номери. Цей метод виглядає так:

```

def get_doubling_files(self):
    paths = os.listdir("queries/")

    file_name_pattern =
re.compile(r'(?<=idioms_' + self.search_tag + '_)[0-9]+(?=.tsv)')
    doubling_files = file_name_pattern.findall('\n'.join(paths))

```

```
return [int(file) for file in doubling_files]
```

Метод `get_file_name` отримує ці номери і якщо подібні файли присутні, то він генерує назву файлу з порядковим номером на один більшим ніж останній. Якщо таких файлів ще немає, то він використовує одиницю. Код цього методу виглядає так:

```
def get_file_name(self):
    doubling_files = self.get_doubling_files()
    if doubling_files:
        return f"idioms_{self.search_tag}_{max(doubling_files)+1}"
    return f"idioms_{self.search_tag}_{1}"
```

Після надання результатів користувачу метод `__init__` викликає останній метод — `next_round`, що запитує в користувача, чи хоче той виконати іще один пошук. Якщо користувач відповідає, що хоче, то весь процес починається з початку, а якщо ні, то програма припиняє роботу. Метод `next_round` виглядає так:

```
@staticmethod
def next_round():
    print('Виконати ще один пошук? (так/ні)')
    answer = input().strip().lower()
    answers = {True: ('так', 'yes', '1'),
               False: ('ні', 'no', '0')}
    if answer in answers[True]:
        return True
```

Посилання на повний код програми на порталі GitHub подано у додатку 4.

2.4. Інтерфейс користувача: пошуковий та класифікаційний потенціал

Для взаємодії з базою даних розроблено інтерфейс командного рядка, що дозволяє здійснювати пошук за класами і підкласами синоптичної схеми тезауруса Роже та за потреби зберігати результат пошуку у файлі формату TSV.

Приклад роботи з інтерфейсу

1. Формування запиту: обирання класу
 - a. за допомогою цифри (відповідь — 2)

Виберіть номер класу або напишіть сам клас:

1. абстрактні відношення
2. емоції
3. інтелектуальні здібності
4. матерія
5. простір
6. волевиявлення
7. шукати в усіх класах

2

b. за допомогою слова (відповідь — *емоції*)

Виберіть номер класу або напишіть сам клас:

1. абстрактні відношення
2. емоції
3. інтелектуальні здібності
4. матерія
5. простір
6. волевиявлення
7. шукати в усіх класах

емоції

2. Формування запиту: обирання секції

a. за допомогою цифри (відповідь — 3)

Можете також обрати секцію:

1. відчуття загалом
2. особистісні відчуття
3. відчуття до інших
4. мораль
5. релігія
6. шукати в усіх секціях

3

b. словами (відповідь — *відчуття до інших*)

Можете також обрати секцію:

1. відчуття загалом
2. особистісні відчуття
3. відчуття до інших
4. мораль
5. релігія
6. шукати в усіх секціях

відчуття до інших

Напишіть тег для пошуку, якщо хочете зберегти результат у файл:

3. Якщо ввести щось незрозуміле для програми, то вона увійде в режим прямої рекурсії і буде в ньому, поки не отримає правильно записаного запиту.

Виберіть номер класу або напишіть сам клас:

1. абстрактні відношення
2. емоції та моральні відчуття
3. інтелектуальні здібності
4. матерія
5. простір
6. волевиявлення
7. шукати в усіх класах

дуллеуерт

Ви ввели неправильне числове значення або вказали неіснуючий клас.

Спробуйте ще раз!

4. У випадку ж успішного формування запиту програма запитає пошуковий тег. Якщо його ввести, то результат пошуку буде збережений у файлі з цим ім'ям, а якщо затиснути клавішу введення, то результат буде лише виведено на екран.
5. Виведення результату запиту “емоції>відчуття до інших” на консоль (знімок екрану не включає всіх рядків і стовпчиків із ілюстрацією та цитуванням заради збереження місця) — малюнки 2.2, 2.3, 2.4.

AFFECTIONS > SYMPATHETIC > SOCIAL AFFECTIONS

id	head group>head	idiom	definition	semantic type
1	resentment	У хату наплювати	зайти, навідатися до кого-небудь	ідіома
2	congratulation	Любов та совіт	побажання щасливого, дружного, мирного життя тим, хто одружується	фраз. вираз
3	sociality	Моє [вам] шанування	усталена формула вітання з ким-небудь	фраз. вираз
4	friendship	Переважувати (переважити) на свій бік	робити своїм прихильником, спільником у чому-небудь	фраз. єдність
5	resentment	Скрегіт зубовний	страшенна лють, злість	фраз. єдність
6	resentment	Вціляти (вцілити) в [саме] серце	боляче дошкуляти, висловлюючи образливу або небажану думку	фраз. єдність
7	resentment	Вражати (вразити) [в саме] серце	завдавати душевного болю	фраз. єдність
8	resentment	Кусати губи	закусуючи губи, виявляти стримувані почуття досади, образи і т. ін	фраз. єдність

Малюнок 2.2. Приклад результату пошукового запиту до бази даних, що виводиться на екран (наведено уривок початку результату).

AFFECTIONS > SYMPATHETIC > DIFFUSIVE SYMPATHETIC AFFECTIONS

id	head group>head	idiom	definition	semantic type
1	malevolence	Шинкувати кров'ю	з корисливих міркувань нещадно губити когось	ідіома
2	malevolence	Чужим рукам на потирання	на знущання комусь	ідіома
3	malevolence	Як сльота	невідступно, надокучливо, набридливо	ідіома
4	malevolence	Троюдити серце	посилювати чий-небудь душевний біль, завдавати комусь ще більших моральних страждань	ідіома
5	malevolence	Оскілками дивитися	дивитися на кого-небудь неприязно, вороже	фраз. єдність
6	malevolence	Кров повипивати	дуже знесилити, змучити	фраз. єдність
7	malevolence	Насисатися (нассатися) крові (поту, сліз)	жорстоко експлуатувати кого-небудь	фраз. єдність
8	malevolence	Сколочувати (сколотити) спокій	порушувати спокій	фраз. єдність

Малюнок 2.3. Приклад результату пошукового запиту до бази даних, що виводиться на екран (наведено уривок кінця результату).

6. Збережений у таблиці результат пошуку (у файлі під назвою *idioms_sympathetic_1.tsv*):

id ...	class (2)	section (3)	subsection (4)	head group>head (5)	idiom (6)	definition (7)	semantic type (8)	citation (9)	source (10)
id	class	section	subsection	head group>head	idiom	definition	semantic type	citation	source
1									
1	affections	sympathetic	diffusive sympathetic affections	malevolence	Шинкувати кров'ю	з корисливих міркувань нещадно губити когось	ідіома	Кругом неправда і неволя, Народ замучений мовчить. І на апостольським престолі Чернець годований сидить. Людською кровію шинкує І рай у найми	Тарас Шевченко, I, 1963, 262
2	affections	sympathetic	diffusive sympathetic affections	malevolence	Чужим рукам на потирання	на знуцання комусь	ідіома		
3	affections	sympathetic	diffusive sympathetic affections	malevolence	Як сльота	невідступно, надокучливо, набридливо	ідіома	Вона [біда] й за ним ходила змалку, як сльота, поки не відкрився перед ним чарівний світ книжок, повний незаних чудес	Петро Колесник, Терен..., 1959, 78
4	affections	sympathetic	diffusive sympathetic affections	malevolence	Троядити серце	посилювати чий-небудь душевний біль, завдавати комусь ще більших моральних страждань	ідіома	Івась лежав, мовчав. Зло, досада давили його за горло, троядили серце, точили сльози з очей	Панас Мирний, I, 1954, 213
5	affections	sympathetic	diffusive sympathetic affections	special sympathetic affections>pity	Мати зглядь	на кого – виявляти увагу, співчуття до кого-небудь	ідіома		
6	affections	sympathetic	diffusive sympathetic affections	malediction	Хай (нехай) йому (їй, їм) грець	виражає сильне незадоволення, несхвалення чого-небудь	фраз. вираз	Ці круті гори, хай їм грець, можуть ухекати найкращого бігуна	Василь Кучер, Чорноморці, 1956, 237
7	affections	sympathetic	diffusive sympathetic affections	malediction	Бодай вас злидні побили (посіли)	побажання нещастя, бідності; прокляття	фраз. вираз	Бодай же вас, цокотухи, Та злидні побили	Тарас Шевченко, I, 1963, 23
8	affections	sympathetic	diffusive sympathetic affections	malediction	Нехай він (ти, вона, воно, вони) спухне (спухнеш, спухнуть)	уживається як лайка, проклін.	фраз. вираз	[Каська:] Картоплю носила, бо мій дармоїд вилежується, щоб він спух...	Мирослав Ірчан, I, 1958, 77

Малюнок 2.4. Приклад результату пошукового запиту до бази даних, що зберігається в унікальний файл (наведено початку результату).

7. Після закінчення пошуку програма пропонує провести ще один. Якщо погодитися, то вона перезапуститься і почне з початку, а якщо відмовитися, то зупиниться:

Виконати ще один пошук? (так/ні)

Подивитися базу даних повністю можна за посиланням у [додатку 3](#).

2.5. Семантико-ідеографічний опис фразеологізмів у БД

У базі даних проанотовано 840 фразеологізмів прийомом семантичної класифікації та методом ідеографічного опису семантики. Ідеографічний опис дозволяє аналізувати різні рівні узагальнення значень слів. Аналіз узагальнень дозволяє зробити висновок про мовну картину світу, представлену у словнику, а також порівняти наповненість різних словників.

Загалом вжито 331 унікальний ономасіологічний шлях, тобто керівні слова. З цього випливає, що на кожен шлях в таблиці припадає в середньому 2,5 фразеологізми. А також це означає, що керівні слова тезауруса Роже представлені на третину, оскільки він їх має 1000.

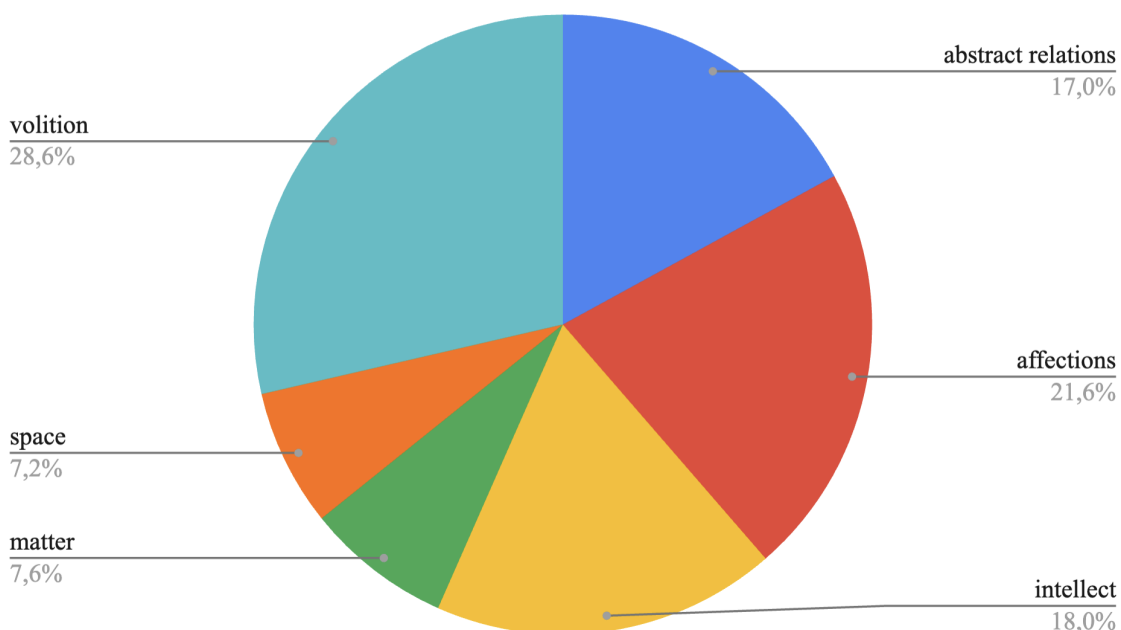


Діаграма 2.1. Варіаційний ряд ономасіологічних шляхів у базі даних.

На діаграмі 2.1 можна побачити варіаційний ряд ономасіологічних шляхів. Його слід читати так: є 156 ономасіологічних шляхів, які зустрілися лише по одному разу, є 12 ономасіологічних шляхів, що зустрілися по 7 разів і, скажімо, один ономасіологічний шлях, що зустрівся 14 разів. Найчастотнішим є шлях *intellect>communication of ideas>modes of communication>>falsehood*, що зустрічається 16 разів. Наприклад, *напускати дурману*.

Синоптична схема тезауруса Роже має 6 класів, що є найзагальнішими категоріями. Відсортувавши їх продуктивність у порядку спадання, отримаємо, що найчастотнішим є клас слів, що пов'язані з волевиявленням /volition/ — їх виявилось 28,6%. Наступним у рейтингу є клас слів, що стосуються емоцій та моральних відчуттів з часткою у 21,6%. На третьому місці з невеликим відривом стоять слова, пов'язані з інтелектуальних здібностями (18,0%) і абстрактними відношеннями (17,0%). Найменш частотними є класи, що стосуються матерії і простору (7,2%). Співвідношення загальних класів синоптичної схеми можна добре узагальнити діаграмою 2.2.

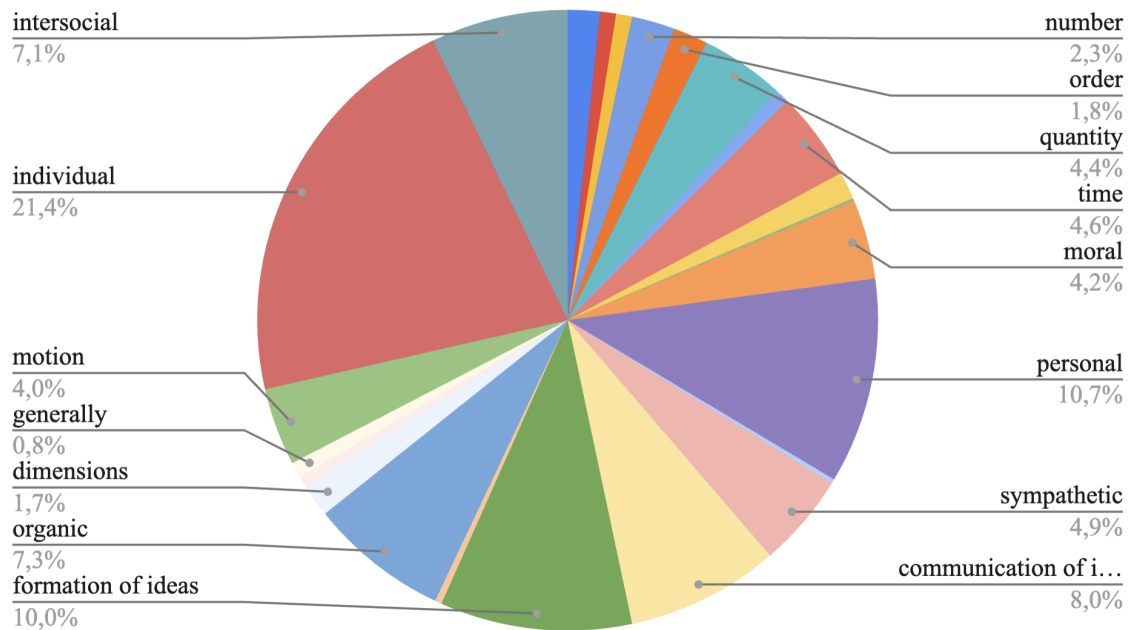
Кількість значень для "roget/class"



Діаграма 2.2. Співвідношення класів з ономасіологічних шляхів у базі даних.

Якщо спроектувати отримане спостереження на діаграму із секціями, то до того ж класу *volition* належить і найчисленніша секція — індивідуальне волевиявлення (*individual*):

Кількість значень для "section"



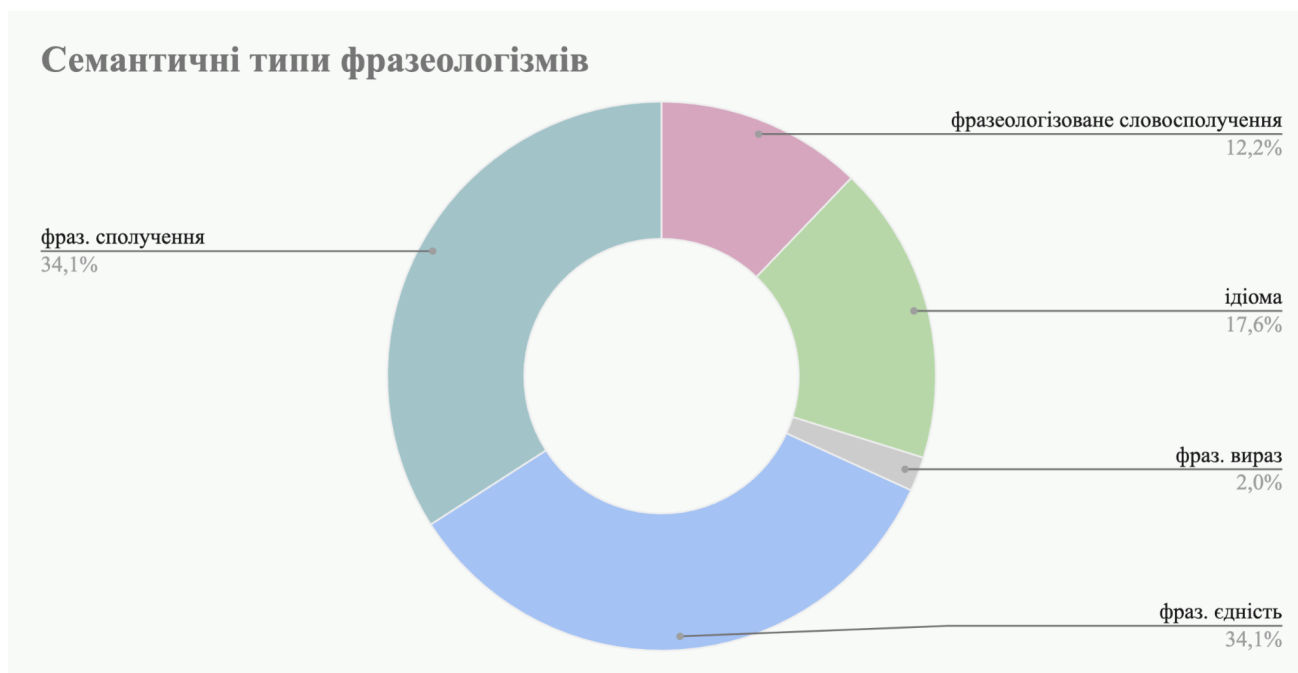
Діаграма 2.3. Співвідношення секцій з ономасіологічних шляхів у базі даних.

У результаті анотування фразеологізмів за семантичною спаяністю компонентів з'явилася можливість дізнатися розподіл значень семантичних типів фразеологізмів. Ідіом виявилось 148, що складає 17,6% від усіх наявних у базі даних. Фразеологічних єдностей і сполучень однакова кількість: по 286 одиниць (34,1% кожні). Фразеологічні вирази фактично не представлені: їх усього 17 одиниць (2,0%). Фразеологізованих сполучень 102 одиниці (12,2%).

Якщо проектувати на ці дані ту приблизну оцінку якості автоматичного перекладу, то можемо вирахувати, що очікувано моделі найкраще впораються з перекладом текстів з фразеологічними єдностями (34,1%), добре себе покажуть у 46,3% випадків (перекладаючи фразеологічні сполучення та фразеологізовані

словосполучення) і не зможуть перекласти 19,6% текстів, бо ті матимуть ідіоми й фразеологічні вирази.

Звісно, ця оцінка дуже приблизна, не базується на достатній кількості даних і скоріше за все результат буде також залежати від частотности конкретних одиниць (що більш нечастотні, то складніше перекласти), та все ж наочно показує, що є сенс експериментувати з використанням моделей машинного навчання для перекладу фразеологізмів.



Діаграма 2.4. Типи фразеологізмів за семантичною спаяністю компонентів у базі даних.

2.6. Перспективи розвитку бази даних

Проект можна розвивати в таких напрямках:

- перевести його на використання системи Open Roget's, що є більш сучасною та повною [Open Roget];
- розширити базу даних за рахунок додавання фразеологізмів із тої частини словника, яка ще потребує обробки, щоб надати базі даних репрезентативности;

- додати паралельні ономаціологічні шляхи, щоб виражати декілька сем та/або додати рівень анотації для диференційних сем;
- додати трансформаційний аналіз, втілити автоматичну систему пошуку фразеологізмів у тексті та виконати корпусне дослідження: це дозволило б не тільки зібрати дані для створення тестів, а й уточнити реальне значення фразеологізмів;
- уточнити переклад понять синоптичної схеми тезауруса Роже українською мовою;
- додати фразеологізми з інших мов.

ВИСНОВКИ

Мета цієї кваліфікаційної роботи полягала в укладанні ідеографічної бази даних фразеологізмів української мови. Цю мету досягнуто. База даних складається із 840 одиниць, її фразеологізми описано ідеографічно і анотовано на семантичну спаяність компонентів. На сторінці проекту на GitHub можна отримати вільний доступ до кодів, розроблених для укладання бази даних, виконання пошукових запитів до неї та самої анотованої бази даних [посилання: <https://github.com/bohdana-ivakhnenko/ukrainian-idioms>].

Загалом у роботі проаналізовано більше 130 джерел і 40 джерел лексикографічного матеріалу. Досліджено термінологію на позначення фразеологізмів та їх типів за семантичною спаяністю компонентів, способи класифікувати фразеологізми, оглянуто різні типи словників, де вони присутні, й аргументовано переваги саме ономасіологічної класифікації. Також описано світовий досвід автоматичного створення ономасіологічних словників за допомогою моделей машинного навчання та загалом обробки фразеологізмів в NLP й можливі способи застосування таких словників у комп'ютерній лінгвістиці.

У практичній частині описано укладання ідеографічної бази даних фразеологізмів української мови. Описано поетапне перетворення електронної версії тлумачного Словника української мови в 11 томах [СУМ-11] у базу даних фразеологізмів, а також спосіб її анотування: ідеографічний опис та аналіз фразеологізмів за семантичною спаяністю компонентів. Опісля детально розібрано роботу програм і наведено спостереження щодо бази даних.

Ідеографічний опис фразеологізмів показав, що їх значення представлені 331 ономасіологічним шляхом, хоча водночас ~47% ономасіологічних шляхів репрезентовано лише одним фразеологізмом. У середньому ж кожен шлях зустрічається в таблиці 2,5 рази. Найчастотніший шлях *intellect > communication of*

ideas>modes of communication>>falsehood зустрівся в базі даних 16 разів. Натомість найчастотнішим класом (найзагальніше значення у синоптичній схемі тезауруса Роже) виявився той, що пов'язаний із волевиявленням (28,6%). Найчастотніша секція (перший рівень уточнення значення) також виявилася у цьому класі — вона пов'язана з індивідуальним волевиявленням (21,4% від усіх речень). Це видається доволі символічним.

Тим не менш, ці дані не можна вважати надійними: вони не є репрезентативними навіть в контексті одного словника — більше половини фразеологізмів, записаних у СУМ-11, не ввійшли до анотованої бази даних, оскільки статті, у яких вони містилися, мають недостатньо деталізовану розмітку і потребують значного постпроцесингу після парсингу.

Виявилось, що використання повних ономасіологічних шляхів не має особливого сенсу для бази даних у контексті програмування, де достатньо знайти одне ключове слово, але має сенс для універсальної, розробленої для застосування у багатьох задачах. Отже, укладена база даних має такі перспективи застосування:

- створити зручний у використанні електронного словника (одно- чи багатомовного);
- фразеологізувати тексти: пропонувати фразеологізм замість прямого значення, де це доречно;
- дефразеологізувати текстів: пропонувати замінити фразеологізми на пряме значення;
- стилістично покращувати тексти: урізноманітнювати їх, пропонуючи цікавіші альтернативи фразеологізмам-кліше;
- створити застосунок з вивчення фразеологізмів;
- автоматично перекладати тексти з фразеологізмами тощо.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Авксентьев Л. Сучасна українська мова. Фразеологія. – Харків : Вища школа, 1983. – С. 42-53.
2. Алефіренко М.Ф. Теоретичні питання фразеології. – Харків : Вища школа, 1987. – 136 с.
3. Амосова Н.Н. Основы английской фразеологии. – М. : Просвещение, 1963. – 359 с.
4. Баранник Д.Х. Фразеологія в усному монологічному мовленні // Питання мовної культури. – К.: Наукова думка, 1968. – Вип. 2. – С. 34-38.
5. Виноградов В. В. Избранные труды. Лексикология и лексикография. - М., 1977. - С. 140-161.
6. Влахов С. И. Непереваемое в переводе / С. И. Влахов, С. Флорин. – М. : Международные отношения, 1980. – 342 с.
7. Демський, М.Т. Українські фраземи й особливості їх творення [Текст] / М. Т. Демський. – Львів : Просвіта, 1994. – 64 с.
8. Дудик П. С. Синтаксис сучасного українського розмовного літературного мовлення : просте речення; еквівалент речення / П. С. Дудик; АН УРСР, Ін-т мовознав. ім О. О. Потебні. – Київ: Наук. думка, 1973. – 284 с.
9. Зорівчак, Р.П. Фразеологічна одиниця як перекладознавча категорія / Р. П. Зорівчак. – Львів : Вища школа, 1983. – С. 32-46.
10. Івченко А. Українська народна фразеологія: ономазіологія, ареали, етимологія / А. Івченко. – Харків : ФОЛІО, 1999. – 304 с.
11. Карпіловська Є.А. Вступ до прикладної лінгвістики: комп'ютерна лінгвістика: Підручник.— Донецьк: ТОВ «Юго-Восток, Лтд», 2006.— 188 с.
12. Комп'ютерні інноваційні технології в перекладацькій та перекладознавчій діяльності / Навчальний посібник. С.Б. Фокін. – К.: КНУ імені Тараса Шевченка, 2019 – 254 с. URL:
http://postup.zzz.com.ua/Fokin_S_B_Computational_Innovative_Tools_for_Translators_And_Traductologists.pdf
13. Коцюк Л., Береза Т. Проект створення двомовного тезауруса фразеологічних одиниць. // Актуальні проблеми прикладної лінгвістики: Матеріали міжнародної Інтернет-конференції (23-24 червня 2011 р.). – Умань, 2011. – С. 36-39.
14. Краснобаєва-Чорна Ж. Ідеографічна класифікація фразеологічних одиниць у

- словнику фразеологічних термінів сучасної української мови // Лінгвістичні студії : зб. наук. праць / Донецький нац. ун-т; наук. ред. А. П. Загнітко. – Л 59
Донецьк : ДонНУ, 2011. – Вип. 22. – 339 с. URL:
<https://philology.lnu.edu.ua/wp-content/uploads/2015/05/studii22.pdf>
15. Краснобаєва-Чорна Ж. Концептуальний аналіз як метод концептивістики (на матеріалі концепту ЖИТТЯ в українській фраземіці) / Ж. Краснобаєва-Чорна // Українська мова. – 2009. – №1. – С. 41-52
 16. Краснобаєва-Чорна Ж. Терміносистема фразеології: структура та складники термінологічної мікросистеми "фразеокласифікація" // Лінгвістичні студії: зб. наук. праць / Донецький нац. ун-т; наук. ред. А. П. Загнітко. – Л 59
Донецьк : ДонНУ, 2013. — Вип. 26. — С. 156-163. — URL:
http://nbuv.gov.ua/UJRN/lingst_2013_26_33
 17. Краснобаєва-Чорна Ж. Фраземіка та фраземографія в сучасній лінгвопарадигмі : підручник для студентів філологічних факультетів вищих навчальних закладів / [за ред. А.П. Загнітка]. – Вінниця: Вид-во, 2018. – 200 с.
 18. Ларин Б.А. О народной фразеологии // Ларин Б.А. История русского языка и общее языкознание. – М., 1977. – С. 149-162.
 19. Медведєв, Ф.П. Українська фразеологія : Чому ми так говоримо [Текст] / Ф. П. Медведєв. – Харків : Вища школа, 1977. – 230 с.
 20. Патен І. М. Ідеографічний та структурно-семантичний методи моделювання фразем з дієсловами руху у зіставному мовознавстві / І. М. Патен // Молодий вчений. — 2017. — № 4.3. — С. 191-195. URL:
http://nbuv.gov.ua/UJRN/molv_2017_4.3_48
 21. Перебийніс В., Сорокін В. Традиційна та комп'ютерна лексикографія: навч. посібник. Київ: Вид. центр КНЛУ, 2009. 218 с.
 22. Прадід Ю. Ф. Засади укладання фразеологічного словника ідеографічного типу / Ю. Ф. Прадід // Мовознавство. – 1995. – №4-5. – С. 35-40
 23. Прадід Ю. Ф. Фразеологічна ідеографія (проблематика досліджень) / Ю. Ф. Прадід. – К.; Сімферополь, 1997. – С. 16, 36-39.
 24. Редин П. Фразеологизмы с пространственным и временным значением в современном украинском языке: автореф. дис. ... канд. филол. наук : 10.02.02 / Харьковский государственный университет им. А.М. Горького. – Харьков, 1989. – 23 с.
 25. Середницька А. Я. Ідеографічний словник дієслів переміщення як засіб

- представлення системності лексики / А. Я. Середницька // Науковий вісник Міжнародного гуманітарного університету. Серія : Філологія. - 2015. - Вип. 15(1). - С. 64-66. URL:
http://nbuv.gov.ua/UJRN/Nvmgu_filol_2015_15%281%29__21
26. Скрипник, Л.Г. Фразеологія української мови / Л. Г. Скрипник. – К. : Наук. думка, 1973. – 280 с. – Бібліогр. : с. 275-279.
 27. Сніжко Н. Идеографічне систематизування українського неолексикону (за матеріалами нових тлумачних словників) / Н. Сніжко // Лексикографічний бюлетень: Зб. наук. пр. — К.: Ін-т української мови НАН України, 2018. — Вип. 27. — С. 30-42. — Бібліогр.: 8 назв. — укр. URL:
<http://dspace.nbuv.gov.ua/handle/123456789/185525>
 28. Сніжко Н.В. Идеографічний тезаурус як модель лексико-семантичної системи (за наслідками автоматизованого аналізу українських іменників) // Мовознавство. - 1995. - № 6. - С. 28-35.
 29. Сніжко Н.В., Сніжко М.Д. Моделі структурно-функціональних властивостей лексики (за наслідками автоматизованого аналізу українських іменників) // Матеріали 4-ї та 5-ї Міжнародних науково-практичних конференцій “УКРСОФТ”. - Львів.: СП “БаК”, 1995. - С. 173-178.
 30. Сніжко Н.В., Сніжко М.Д. “Идеографічний тезаурус” як інформаційно-довідкова система при вивченні закономірностей структурно-функціональної організації лексики // Мовознавство. - 1996. - № 4-5. - С. 23-28.
 31. Старко В. Фразеологія, лексикографія і корпус / Лексикографічний бюлетень: Зб. наук. пр. — К.: Ін-т української мови НАН України, 2006. — Вип. 13. — С. 43-47. URL: <http://dspace.nbuv.gov.ua/handle/123456789/72852>
 32. Удовиченко Г.М. Предмет фразеології та семантичні типи фразеологізмів // Українська мова і література в школі. – 1965. – №2. – С. 20-24.
 33. Ужченко, В.Д. Східноукраїнська фразеологія: монографія / В. Д. Ужченко. – Луганськ : Альма-матер, 2003. – 362 с. – Бібліогр.: с. 330-352.
 34. Ужченко, В.Д., Ужченко, Д.В. Фразеологія сучасної української мови : навч. посіб. / В. Д. Ужченко, Д. В. Ужченко. – Луганськ : Альма-матер, 2007. – С. 52-56. – Бібліогр.: с. 441-464.
 35. Abarna S., Sheeba J.I., Devaneyan P. S. An ensemble model for idioms and literal text classification using knowledge-enabled BERT in deep learning // Measurement: Sensors, 24, 2022. DOI:

- <https://doi.org/10.1016/j.measen.2022.100434>.
36. Alexander R. J. (1987). Problems in understanding and teaching idiomaticity in English. *Anglistik und englischunterricht*, 32, 105–122.
 37. Anastasiadi-Symeonidi, A., Efthymiou, A. (2006): *Οι Στερεότυπες Εκφράσεις και η Διδακτική της Νέας Ελληνικής ως Δεύτερης Γλώσσας*, Athens: Patakis.
 38. Avram M. (1986): *Gramatica pentru toȚi („Die Grammatik für Alle“)*. BucureȚti.
 39. Baker, M. 1992. *In other words: a coursebook on translation*. London and New York: Routledge
 40. Baddorf, D. S., & Evens, M. W. (1998). Finding phrases rather than discovering collocations: Searching corpora for dictionary phrases. In *Proc. of the 9th Midwest Artificial Intelligence and Cognitive Science Conference (MAICS-98)* (pp. 110–116).
 41. Bally Ch. *Précis de stylistique. Esquisse d’une méthode fondée sur l’étude du français moderne*, Genève, Eggimann. 1905.
 42. Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic expressions and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, 10(3), 245–261.
 43. Briskilal J., Subalalitha C.N. An ensemble model for classifying idioms and literal texts using BERT and RoBERTa // *Information Processing & Management*, 59(1), 2022. DOI: <https://doi.org/10.1016/j.ipm.2021.102756>.
 44. Bushnaq T. A Retrospective Analysis of the Term Phraseological Unit, *The Proceedings of the International Conference Globalization, Intercultural Dialogue and National Identity. Section: Language and Discourse*, 2. Tîrgu-Mureș: Arhipelag XXI Press. 2015, p. 167-176. URL: <https://www.diacronia.ro/en/indexing/details/V1591>
 45. Cain, K., Towse, A. S., & Knight, R. S. (2009). The development of idiom comprehension: An investigation of semantic and contextual processing skills. *Journal of Experimental Child Psychology*, 102(3), 280–298.
 46. Casares J. Nuevo Concepto del Diccionario de la Lengua: Discurso leído el el acto de su recepción // *Real Academia Española*. — Madrid: G. Koehler - Pacífico, 35, 1921. URL: https://juliocasares.es/wp-content/uploads/2017/03/Discurso_ingreso_Julio_Casares.pdf
 47. Čermák F. *Czech Idiom Dictionary* // Euralex. — Prague: 1994. pp. 426-431. URL:

- https://www.euralex.org/elx_proceedings/Euralex1994/47_Euralex_F.%20Cermak%20-%20Czech%20Idiom%20Dictionary.pdf
48. Columbus, G. (2013). In support of multiword unit classifications: Corpus and human rating data validate phraseological classifications of three different multiword unit types. *Yearbook of Phraseology*, 4(1), 23-44.
<https://doi.org/10.1515/phras-2013-0003>.
 49. Cowie A. (2001) "Introduction" In: Cowie, Anthony (ed.) 2001. *Phraseology: Theory, Analysis, and Application*. Oxford: Oxford University Press, pp. 1-22.
 50. Cowie A. P., Mackin R., McCaig I. R. (1983). *Oxford dictionary of current idiomatic English: Phrase, clause and sentence idioms (Vol. 2)*. Oxford: Oxford University Press.
 51. Dimitrescu F., *Locuțiunile verbale în limba română (Verb Phrases in the Romanian language)*, București, Editura Academiei Republicii Populare Române, 1958.
 52. Dimitriu C., *Gramatica limbii române explicată. Morfologia*. Iași: Editura Junimea, 1979, 398p.
 53. Dornseiff F. *Der deutsche Wortschatz nach Sachgruppen*. Berlin, Boston: De Gruyter; 1959. DOI: <https://doi.org/10.1515/9783111712116>
 54. Díaz J. L. R. (2020) "From head to toe: A lexical, semantic, and morphosyntactic study of idioms in phraseological dictionaries in English and Spanish." DOI: <https://doi.org/10.6035/MonTI.2020.ne6.9>
 55. *El Diccionario Ideológico de la lengua española - DILE - 1942 // Julio Casares. Legado Familiar*. URL: <https://juliocasares.es/diccionario-ideologico/>
 56. Evseev, I. (1974) *Semantica verbului, categoriile de acțiune, devenire și stare*. Timișoara: Editura Facla.
 57. Fazly A., Cook P., Stevenson S. Unsupervised Type and Token Identification of Idiomatic Expressions // *Computational Linguistics* 35(1), 2009. pp. 61-103.
<https://doi.org/10.1162/coli.08-010-R1-07-048>
 58. Fellbaum C. The Representation of Idioms in WordNet // *Computer Science*, 2009. URL: <https://aclanthology.org/W98-0707.pdf>
 59. Fellbaum Ch. Towards a Representation of Idioms in WordNet. In *Usage of WordNet in Natural Language Processing Systems*. 1998. URL: <https://aclanthology.org/W98-0707>
 60. Fernando C. (1996). *Idioms and idiomaticity*. Oxford: Oxford University Press.
 61. Gonner S. How to Speak to a Multicultural Audience. *CultureWiz*, 2022. URL: <https://www.theculturewiz.com/post/how-to-speak-to-a-multicultural-audience>

62. Granger S., Meunier F. *Phraseology. An Interdisciplinary Perspective*. Amsterdam – Philadelphia: John Benjamins. xxviii, 2008. 422 pp. URL: https://www.academia.edu/47249543/Sylviane_Granger_and_Fanny_Meunier_ed_s_2008_Phraseology_An_Interdisciplinary_Perspective_Amsterdam_Philadelphia_John_Benjamins&nav_from=1379137b-c8fb-4d01-97da-f95225adac74&rw_pos=0
63. Gregoire N. 2007. Design and Implementation of a Lexicon of Dutch Multiword Expressions. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 17–24, Prague, Czech Republic. Association for Computational Linguistics. URL: <https://aclanthology.org/W07-1103/>
64. Guławska-Gawkowska M. Semantische Aspekte im Vergleich von deutschen und polnischen Phraseologismen // *Zeitschrift des Verbandes Polnischer Germanisten*, 1(1), 2012. pp. 74-88. URL: <http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.ojs-nameId-4f32a095-f755-35dc-9a43-deac5c460d86-year-2012-volume-1-issue-1-article-5072>
65. Hockett C. F. (1958). *A course of modern linguistics*. New York, NY: Macmillan. DOI: <https://doi.org/10.1111/j.1467-1770.1958.tb00870.x>
66. Horn G. M. 2003. “Idioms, metaphors and syntactic mobility” in *Journal of Linguistics* 39(2):245–273.
67. Howarth P. *Phraseology and Second Language Proficiency* // *Applied Linguistics*, 19(1), 1998, pp. 24–44, DOI: <https://doi.org/10.1093/applin/19.1.24>
68. Hughes, C., Filimonov, M., Wray, A., & Spasić, I. (2021). Leaving No Stone Unturned: Flexible Retrieval of Idiomatic Expressions from a Large Text Corpus. *Machine Learning and Knowledge Extraction*, 3(1), 263-283. DOI: <https://doi.org/10.3390/make3010013>
69. Hüllen W. '5 Roget in Germany: The Adaptation of the Thesaurus (1852) in Daniel Sanders' *Sprachschatz* (1873)', *Networks and Knowledge in Roget's Thesaurus* (Oxford, 2009; online edn, Oxford Academic, 1 Jan. 2009), pp. 60-75. DOI: <https://doi.org/10.1093/acprof:oso/9780199553235.003.0005>.
70. Hüllen W. '6 Roget in the Romance World: Théodore Robertson's Translation of Roget's Thesaurus (1852) into French as *Le Dictionnaire Idéologique* (1859)', *Networks and Knowledge in Roget's Thesaurus* (Oxford, 2009; online edn, Oxford Academic, 1 Jan. 2009), pp. 76-91. DOI: <https://doi.org/10.1093/acprof:oso/9780199553235.003.0006>.
71. Hutchins, J. *From First Conception to First Demonstration: the Nascent Years of*

- Machine Translation, 1947–1954. A Chronology. *Machine Translation* 12, 195–252 (1997). <https://doi.org/10.1023/A:1007969630568>
72. Imperiale, R., & Schafroth, E. (2016). Italian Phraseology for Learners and Specialists: The Constructionist Approach in Practice. *RiCOGNIZIONI. Rivista Di Lingue E Letterature Straniere E Culture Moderne*, 3(6), 105-132. <https://doi.org/10.13135/2384-8987/1851>
 73. Iordan I. *Stilistica limbii române*, Ed. Științifică, 1975, 405 p.
 74. Iordan I., Robu V. *Limba română contemporană*. Editura Didactică și Pedagogică, 1978, 684 p.
 75. Istros the Callimachean, Wikiwand. URL: https://www.wikiwand.com/en/Istros_the_Callimachean
 76. Kennedy A. (2010). Automatically Expanding the Lexicon of Roget's Thesaurus. In *Advances in Artificial Intelligence, 23rd Canadian Conference on Artificial Intelligence, Canadian, AI 2010, Ottawa, Canada, May 31 - June 2, 2010*, 410-411. URL: http://www.cs.toronto.edu/~akennedy/publications/roget_cai_2010.pdf
 77. Kennedy A., Szpakowicz S. (2014). Evaluation of Automatic Updates of Roget's Thesaurus. *Journal of Language Modelling* 2 (1), 1–49. URL: https://www.cs.toronto.edu/~akennedy/publications/roget_jlm_2014.pdf
 78. Kilgarriff A., "Thesauruses for natural language processing," *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003, Beijing, China, 2003*, pp. 5-13, DOI: <https://doi.org/10.1109/NLPKE.2003.1275859>.
 79. Kim Eu-Bin. *Idiomify - Building a Collocation-supplemented Reverse Dictionary of English Idioms with Word2Vec for non-native learners*. — Manchester: 2021, URL: <https://arxiv.org/pdf/2204.05634.pdf>
 80. Klegr A. (2008). Turning Roget's Thesaurus into a Czech Thesaurus. URL: https://www.researchgate.net/publication/309733595_Turning_Roget's_Thesaurus_into_a_Czech_Thesaurus
 81. Koehn Ph., Och F. J., Marcu D. *Statistical Phrase-Based Translation*. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2003*. pp. 127–133. URL: <https://aclanthology.org/N03-1017/>
 82. Laufer B., Goldstein Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54, 399–436.

83. Lim E. A. C., Ang S. H., Lee Y. H., Leong S. M. (2009). Processing idioms in advertising discourse: Effects of familiarity, literality, and compositionality on consumer ad response. *Journal of Pragmatics*, 41(9), 1778–1793.
84. Lipka L. *English Lexicology. Lexical Structure, Word Semantics, and Word-Formation*, Tübingen, Germany: Gunter Narr, 2002. *Cognitive Linguistics Bibliography (CogBib)*, URL: <https://www.degruyter.com/database/COGBIB/entry/cogbib.7761/html>
85. Liu, D. (2008). *Idioms: Description, comprehension, acquisition, and pedagogy*. New York: Routledge.
86. Lobzhanidze I. *Online Dictionary of Idioms*. // *Proceedings of the 17th EURALEX International Congress — Tbilisi: Ivane Javakhishvili Tbilisi University Press*, 2016. pp. 710-716. URL: <https://euralex.org/publications/online-dictionary-of-idioms/>
87. Makkai A. *Idiom Structure in English*. London: Mouton, 1972.
88. McCarthy, M. (1998). *Spoken Language & Applied Linguistics*. Cambridge: Cambridge University Press.
89. Mel'cuk I. (2012). Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology*. 3(1). <https://doi.org/10.1515/phras-2012-0003>.
90. *Misuse of Idioms in Academic Writing // Capstone Editing*, accessed in June 2023. URL: <https://www.capstoneediting.com.au/blog/misuse-of-idioms-in-academic-writing>
91. Mitsis. (2004). *Language teaching through the communication approach*. Athens: Gutenberg.
92. Moon R. (1998) *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. New York: Oxford University Press.
93. Moon, R. 1999. Needles and haystacks, idioms and corpora: Gaining insights into idioms, using corpus analysis. In: Herbst, T. and Popp, K. ed. *The Perfect Learners' Dictionary (?)*. Berlin, Boston: De Gruyter, pp. 265-282. <https://doi.org/10.1515/9783110947021.265>.
94. Nair S. S. *The Knowledge Structure in Amarakośa // Doctor thesis*, 2011. URL: <https://sanskrit.uohyd.ac.in/scl/amarakosha/index.html>
95. Narang Sh., Chowdhery A. *Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance // GoogleBlog*, 2022. URL: <https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.htm>

- l.
96. Nation I. S. P., Meara P. (2002). Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 35–54). New York: Oxford University Press.
 97. Norvig P. The Unreasonable Effectiveness of Data // UBC Computer Science, 2011. URL: <https://www.youtube.com/watch?v=yvDCzhbjYWs>
 98. Núñez M. Save Time and Money by Avoiding Idioms and Slang. SimulTrans, 2021. URL: <https://www.simultrans.com/blog/save-time-money-by-avoiding-idioms-and-slang>
 99. Philippide Al. Gramatică elementară a limbii române'. Iași, Editura Librăriei Isr. Kuppermann, 1897
 100. Piirainen E. Europeanism, Internationalism or Something else? Proposal for a Cross-linguistic and Cross-cultural Research Project on Widespread Idioms in Europe and Beyond // *Hermes, Journal of Linguistics*. – 2005. – No35. – P. 45-75. DOI: <https://doi.org/10.7146/hjlc.v18i35.25816>
 101. Plain language and word choice / Australian Government. 2022 URL: <https://www.stylemanual.gov.au/writing-and-designing-content/clear-language-and-writing-style/plain-language-and-word-choice>
 102. Plaza S. M. The Making of a Bilingual Dictionary of Phraseological Units English-Spanish/ Spanish-English with Corpora Examples — Panace@ – Vol. VII, N.º 23. Junio de 2006
 103. Prince V., Chauché J. Building a Bilingual Representation of the Roget Thesaurus for French to English Machine Translation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA), 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/626_paper.pdf
 104. Qi F., Zhang L., Yang Y., Liu Z., Sun M. WantWords: An Open-source Online Reverse Dictionary System. *Conference on Empirical Methods in Natural Language Processing*, 2020. DOI: <https://doi.org/10.18653/v1/2020.emnlp-demos.23>
 105. Rafatbakhsh E., Ahmadi A. A thematic corpus-based study of idioms in the Corpus of Contemporary American English. *Asian. J. Second. Foreign. Lang. Educ.* 4, 11 (2019). DOI: <https://doi.org/10.1186/s40862-019-0076-4>
 106. Reverse dictionary // Prompt gallery, GenerativeAI Google. URL: <https://developers.generativeai.google/prompts/reverse-dictionary>
 107. Roberts R. M., Kreuz R. J. 1994. Why do people use figurative language?

- Psychological Science, 5: 159-163.
108. "ROGETS THESAURUS" Concise Oxford Companion to the English Language. Retrieved May 25, 2023. URL:
<https://www.encyclopedia.com/humanities/encyclopedias-almanacs-transcripts-and-maps/rogets-thesaurus>
 109. Ruiz Gurillo L. (2001) *Las locuciones en el español actual*. Madrid: Arco Libros.
 110. Salton G. D., Ross R. J., Kelleher J. D. Exploring the Use of Attention within an Neural Machine Translation Decoder States to Translate Idioms, 2018. URL:
<https://arxiv.org/pdf/1810.06695v1.pdf>
 111. Schafroth E. (2020). Why equivalence of idioms in different languages is the exception. Arguments from a constructional perspective. *Contrastive Phraseology*. pp. 129-150. URL:
https://www.academia.edu/43814361/Why_equivalence_of_idioms_in_different_languages_is_the_exception_Arguments_from_a_constructional_perspective
 112. Schippan T. *Lexikologie der deutschen Gegenwartssprache*. – Tübingen : Verlag Max Niemeyer, 2002. – 316 p.
 113. Shudo K., Kurahone A. Tanabe T. “A Comprehensive Dictionary of Multiword Expressions.” Annual Meeting of the Association for Computational Linguistics (2011). URL:
<https://www.semanticscholar.org/paper/A-Comprehensive-Dictionary-of-Multiword-Expressions-Shudo-Kurahone/fee85729f6edfca382ce747b877bda307b6145e9>
 114. Siddique B., Sufyan Beg M. M. Reverse Dictionary Formation: State of the Art and Future Directions // *SN Computer Science* 4(2), article # 168, 2023. DOI:
<https://doi.org/10.1007/s42979-022-01495-1>
 115. Sierra G. *The onomasiological dictionary: a gap in lexicography*. — México, Euralex, 2000, pp. 223-235.
 116. Simpson R., Mendis D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly*, 37(3), 419-441.
 117. Škvorc T., Gantar P., Robnik-Šikonja M. MICE: Mining Idioms with Contextual Embeddings // *Knowledge-Based Systems*, 235, 2022. DOI:
<https://doi.org/10.1016/j.knosys.2021.107606>
 118. Stark M. *Bilingual Thematic Dictionaries*. Berlin, Boston: De Gruyter; 2011.
<https://doi.org/10.1515/9783110258899>
 119. Symeonidis, Ch. (2000): *Εισαγωγή στην Ελληνική Φρασεολογία*, Thessaloniki: Kodikas.

120. Technology Overview // Cyc, 2021. p. 22. URL:
<https://cyc.com/wp-content/uploads/2021/04/Cyc-Technology-Overview.pdf>
121. Tedeschi S., Martelli F., Navigli R. ID10M: Idiom Identification in 10 Languages // Conference: 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2022) — Seattle, Washington, Online: 2022. DOI:
<https://doi.org/10.18653/v1/2022.findings-naacl.208>.
122. Teodorescu A. (2015). Mobile learning and its impact on business English learning. *Procedia-Social and Behavioral Sciences*, 180, 1535–1540.
123. Theves S., Fernandez G., Doeller Ch. The Hippocampus Encodes Distances in Multidimensional Feature Space // *Current Biology* 29(7), 2019. pp. 1226-1231. DOI:<https://doi.org/10.1016/j.cub.2019.02.035>
124. Tip-of-the-Tongue Phenomenon. Learning and Memory. Encyclopedia.com. 25 May. 2023. URL: <https://www.encyclopedia.com>.
125. Toris C. (1). Language at Its Word: A Lexical Analysis of Idiomatic Speech. *Jefferson Journal of Science and Culture*, (1). URL:
<https://journals.sfu.ca/jjsc/index.php/journal/article/view/5>.
126. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention Is All You Need // 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. URL:
<https://arxiv.org/pdf/1706.03762.pdf>
127. Wood M. M. (1981). A definition of idiom. Bloomington: Indiana University Linguistics Club.
128. Wray A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, 21(4), 463–489.
<https://doi.org/10.1093/applin/21.4.463>
129. Zaraysky Susanna. Writing for global audiences. Medium, 2020. URL:
<https://medium.com/google-design/writing-for-global-audiences-d339d23e9612>
130. Zechner N., Borin L. Towards a Swedish Roget-Style Thesaurus for NLP // Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020// Proceedings of the Globalex Workshop on Linked Lexicography, pp. 53–60. URL: <https://aclanthology.org/2020.globalex-1.9.pdf>
131. Zhang L., Qi F., Liu Z., Wang Y., Liu Q., Sun M. (2020). Multi-Channel Reverse Dictionary Model. Proceedings of the AAAI Conference on Artificial Intelligence, 34(01), 312-319. <https://doi.org/10.1609/aaai.v34i01.5365>

СПИСОК ДЖЕРЕЛ ЛЕКСИКОГРАФІЧНОГО МАТЕРІАЛУ

- Баранцев: Англо-український фразеологічний словник / Уклад. К.Т. Баранцев. – 3-те вид., випр. – К.: Т-во "Знання", КОО, 2006. – 1056 с.
- Береза: Береза Т. А. Сучасний англо-український словник живої мови / Т. А. Береза. – Львів : Апріорі, 2012. – 400 с.
- Білоноженко-1: Фразеологічний словник української мови / АН України, Ін-т укр. мови ; [уклад.: В. М. Білоноженко та ін., редкол.: Л. С. Паламарчук (голова) та ін.]. - Київ : Наук. думка, 1993. Кн. 1 . – 1993. – 528 с. URL: <http://irbis-nbuv.gov.ua/ulib/item/UKR0002726>
- Білоноженко-2: Фразеологічний словник української мови / АН України, Ін-т укр. мови ; [уклад.: В. М. Білоноженко та ін., редкол.: Л. С. Паламарчук (голова) та ін.]. - Київ : Наук. думка, 1993. Кн. 2 . – 1993. – [2] с., с. 529-980. URL: <http://irbis-nbuv.gov.ua/ulib/item/UKR0002727>
- Винник: Словник фразеологізмів української мови / відповідальний ред. Винник В. О. - К.: Наукова думка, 2003. - 788 с. URL: <https://archive.org/details/slov557/mode/2up>
- Єрмоленко: Українська мова. Короткий тлумачний словник лінгвістичних термінів / За ред. С.Я. Єрмоленко. — К.: Либідь, 2001. — 224 с. - ISBN 966-06-0177-8. URL: <https://archive.org/details/terminy2001>
- ІдеоФраз: ІдеоФраз: Багатомовний словник фразеологічних синонімів / Фокін С.Б., 2021. URL: <http://postup.zzz.com.ua/IdeoPhrase.html>
- Калашник: Словник фразеологічних антонімів української мови / В. С. Калашник, Ж. В. Колоїз; НАН України, Харк. нац. ун-т ім. В. Н. Каразіна, Укр. мовно-інформ. фонд. – Київ: Довіра, 2004. – 284 с. – (Словники України). URL: <http://irbis-nbuv.gov.ua/ulib/item/UKR0001518>

- Коломієць: Коломієць М.П., Регушевський Є.С. Словник фразеологічних синонімів / Під ред. В.А. Винника, — К.: Рад. шк., 1988. — 200 с. URL: <http://irbis-nbuv.gov.ua/ulib/item/UKR0001519>
- Розум: Проєкт «Розум» // Коломієць М.П., Регушевський Є.С. Словник фразеологічних синонімів. URL: <http://www.rozum.org.ua/index.php?a=index&d=24>
- СУМ-11: Словник української мови: в 11 томах. — 1979. — URL: <http://sum.in.ua/>
- СУМ-20: Словник української мови: в 20 томах. — 2015-2023. — URL: <https://sum20ua.com/>
- Удовиченко: Удовиченко Г.М. Фразеологічний словник української мови : В 2 т. — К. : Вища школа, 1984. — Т.1: А-М. — 304 с.; Т.2: Н-Я. — 384 с.
- Ужченки: Ужченко В.Д., Ужченко Д.В. УЗЗ Фразеологічний словник української мови. — К.: Освіта, 1998. — 224 с. ISBN 966-04-0027-6. URL: https://schoolplusnet.com/up_portfolio/2016-3-2-21-42-22_file_20.pdf
- Шведова: Русский семантический словарь: Опыт автоматического построения тезауруса: от понятия к слову / Под ред. Н. Ю. Шведовой - М., 1982
- BNC: The British National Corpus, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: <http://www.natcorp.ox.ac.uk/>.
- Boissière: Dictionnaire analogique de la langue française : répertoire complet des mots par les idées et les idées par les mots (4e édition) / par P. Boissière,... 1862. URL: <https://gallica.bnf.fr/view3if/ga/ark:/12148/bpt6k22361w/fl>
- Cambridge: Cambridge Advanced Learner's Dictionary & Thesaurus. Cambridge: Cambridge University Press. ABSTRACT: This article explores the impact of global culture on our lifestyle.
- Čeština: Akademický slovník současné češtiny [online] (2017–2023). Praha: Ústav pro jazyk český AV ČR, v. v. i. Cit. 02.06.2023. URL: <https://slovníkcestiny.cz>

ChatGPT: OpenAI. (2023). ChatGPT (May 13 version) [Large language model]. URL:
<https://chat.openai.com>

COCA: Davies M. (2008-) The Corpus of Contemporary American English (COCA).
Available online at <https://www.english-corpora.org/coca/>.

Collins: Collins Cobuild Dictionary Of Idioms / edited by Penny Hands — Italy: Grafica
Veneta S.p.a., 4th ed., 2020.

Common Crawl: Distribution of Languages // Statistics of Common Crawl Monthly
Archives, 2023. URL:
<https://commoncrawl.github.io/cc-crawl-statistics/plots/languages>

Corripio: Corripio F. Diccionario De Ideas Afines Herder — Barcelona: Editorial
Herder, 2a ed., 1985. URL:
<https://archive.org/details/corripio-fernando-diccionario-de-ideas-afines-herder-1994/page/n1/mode/2up>

DeepL: DeepL. Translator. URL: <https://www.deepl.com/translator>

Fenati: Dizionario Idiomatico TEDESCO-ITALIANO / Fenati B., Rovere G., Schemann
H. et al. — Bologna: Zanichelli editore s.p.a., 2009.

Google Translate: Google. Translate. URL:
<https://translate.google.de/?hl=uk&tab=TT&sl=en&tl=uk&op=translate>

Hahessy: Hahessy R. Funny English Idioms And Their Meanings. Bored Panda, 2016.
URL:
<https://www.boredpanda.com/funny-english-idioms-meanings-illustrations-roisin-hahessy/>

Merriam-Webster: Merriam-Webster.com Dictionary, URL:
<https://www.merriam-webster.com/dictionary/idiom>.

Jochim: Jochim C. et al. (2018). SLIDE: a Sentiment Lexicon of Common Idioms.
Proceedings of the Eleventh International Conference on Language Resources and
Evaluation (LREC'2018). European Languages Resources Association (ELRA)

Open Roget: Kennedy A., Szpakowicz S.. The Open Roget's Project. URL:
<https://sites.google.com/site/openrogets/>

Idiomatch: Kim Eu-Bin. Idiomatch. URL: <https://github.com/eubincto/idiomatch>

Larousse: Larousse [thesaurus] / Larousse, accessed in June 2023. URL:
<https://www.larousse.fr/>

long-tongued: Collins Dictionary, s.v. “long-tongued,” accessed June 9, 2023, URL:
<https://www.collinsdictionary.com/dictionary/english/long-tongued>

Microsoft Bing: Microsoft. Bing Translator. URL: <https://www.deepl.com/translator>

Nexis: Nexis UNI Corpus, LexisNexis, 2023 URL: www.nexisuni.com

Oxford-2: The Oxford dictionary of idioms / edited by Judith Siefring — Oxford; New York : Oxford University Press, 2nd ed., 2004. 340 p.

Oxford-4: The Oxford Dictionary of Idioms / edited by Ayto J. — Oxford University Press, 4th ed., 2020. DOI: <https://doi.org/10.1093/acref/9780198845621.001.0001>

PaLM 2: Google. AI ACROSS GOOGLE: PaLM 2. URL:
<https://ai.google/discover/palm2>

Roget Original: Roget P. M. Roget's Thesaurus. The Original Manuscript. URL:
<https://web.archive.org/web/20130502153549/http://www.rain.org/~karpeles/rogetfrm.html>

Roget Gutenberg: Roget P. M. Roget's Thesaurus of English Words and Phrases, 1911 // The Project Gutenberg eBook / L. John Old, 2004. URL:
<https://www.gutenberg.org/cache/epub/10681/pg10681-images.html>

SpaCy: Explosion. SpaCy v3.0 (Accessed in June 2023). URL: <https://spacy.io/usage/v3>

Statista: Languages most frequently used for web content as of January 2023, by share of websites // Statista, 2023. URL:
<https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/>

WantWords: WantWords: Reverse Dictionary. — Tsinghua: THUNLP, 2019. URL:
<https://wantwords.net/>

Webster's 1913: Webster's New International Dictionary of the English Language. —
Springfield: G. & C. Merriam Company. URL: <https://www.websters1913.com/>

ДОДАТКИ

Додаток 1. Код програми-скрейпера

Посилання на код програми:

https://github.com/bohdana-ivakhnenko/ukrainian-idioms/blob/main/sum_scraper.py

Додаток 2. Код програми для парсингу словника

Посилання на код програми:

https://github.com/bohdana-ivakhnenko/ukrainian-idioms/blob/main/parse_sum_article.py

Додаток 3. Ідеографічна база даних фразеологізмів

Посилання на базу даних:

<https://github.com/bohdana-ivakhnenko/ukrainian-idioms/blob/main/sum-idioms-annotated.tsv>

Додаток 4. Код інтерфейсу

Посилання на код програми:

<https://github.com/bohdana-ivakhnenko/ukrainian-idioms/blob/main/sum-idioms-interface.py>