### 3.2. Основні принципи пошуку в корпусах текстів

З огляду на широке розмаїття корпусних знарядь правила пошуку в різних корпусах можуть суттєво варіюватися, і все ж можна узагальнити його більш-менш універсальні принципи. Окрім простих елементарних запитів слова або сполучення слів, що можна виконувати у пошуковому рядку комп'ютерних оглядачів (браузерів) таких систем пошуку як *Google*, у корпусних ресурсах існують додаткові можливості, на яких зупинимось детальніше.

1. ***Однослівні запити пошуку.*** Можливо, найпоширенішим запитом пошуку є однослівні фрагменти, що їх користувачі вводять у пошуковий рядок. Під час такого пошуку необхідно чітко розмежовувати пошук словоформи та пошук лексеми. Пошук лексеми передбачає вибірку словоформ з усієї парадигми заданої лексеми. Скажімо, у корпусі *mova.info* відповідні опції називаються “словоформа/лексема” [Корпус текстів української мови]; у Корпусі іспанської мови XXI століття ці опції називаються “lema/forma” [Corpes XXI]. Пошук лексеми можливий завдяки у тому випадку, коли корпус пройшов відповідну розмітку або ж оснащений лематизатором. Однак часто можна проводити пошук лексеми й за відсутності заданих характеристик, якщо пошуковий рядок сприймає регулярні вирази.

Під регулярним виразом в інформатиці розуміють такий вираз, що в узагальненому вигляді описує цілу множину сполучень знаків, графічних слів, словосполучень. Він дозволяє охопити пошуком усю множину за одну операцію. Прикладом регулярного виразу, сумісного з більшістю мов програмування та запитів, який описує усі графічні слова, складені виключно з малих латинських літер від ***a*** до **z**, може бути такий:

**·[a-z]+·**

Знак “+” означає, що набір символів, що міститься у квадратних дужках, може бути різної довжини (від 1 символу)[1]; вертикальні крапки на початку і в кінці виразу відражають пробіли, які необхідно чітко позначувати як роздільник графічних слів (насправді пробіл – далеко не єдиний можливий роздільник, і при опрацюванні масивів текстів необхідно враховувати також пунктуаційні знаки, символи закінчення рядку, символи закінчення файлу чи потоку тощо). Прикладом пошуку всіх графічних слів, що починаються зі сполучення “брунатн” і містять після нього будь-які інші кирилічні символи в кількості від 0 до 3, що власне відповідатиме пошуку усієї граматичної парадигми прикметника “брунатний”, буде такий:

**·брунатн[а-яіїґє’]{0-3}·**

Літери “ї”, “ґ”, “є” і “ ’ ”додано до списку окремо, оскільки вони містяться у міжнародній таблиці символів *ASCII (American Standard Code for Information Interchange)* поза проміжком від “а” до “я”. Однак для спрощення запитів у різних системах пошуку знак “зірочка” “*” часто заміщує будь-яке сполучення літер; а знак питання – “?” – будь-який символ. Сполучення {0-3} означає, що кількість знаків у квадратних дужках має становити від 0 до 3, що відповідатиме довжині так званої квазіфлексії. Багато корпусів підтримують пошук з використанням обох символів: “*” і “?”. Ці символи також називають “байдужі символи”, “джокери”, англійською – “wildcard characters”, іспанською – “carácter comodín”, російською - “символы подстановки”.

У флективних мовах у більшості випадків заміна останніх трьох літер зірочкою (*) дозволяє здійснювати пошук усієї парадигми. Такий підхід, позбавлений критеріїв наукової строгості, може виявитися дуже корисним і виправданим при розв’язанні багатьох практичних і теоретичних завдань, коли достовірність висновків не залежить від високої точності підрахунків; однак при такому пошуку до результатів можуть випадково потрапляти й омоніми, пароніми; натомість втрачаються суплетивні форми, форми неправильних дієслів, словоформи з чергуванням у корені.

***Багатослівні запити пошуку.*** Користувачів корпусу може цікавити не лише пошук оточення певної словоформи чи лексеми, а й ширші парадигматичні аспекти. Для перевірки сполучуваності слова, варіантності тих чи інших виразів, фразеологізмів, нюансів дієслівного, іменникового, прикметникового керування доцільно вводити декілька слів. одразу Більшість сучасних корпусів такий пошук підтримують. У рамках багатослівного пошуку можна використовувати “байдужі символи”, що можуть заміщувати будь-яку літеру, сполучення літер, або регулярні вирази. Однак зі збільшенням обсягу корпусу запити з узагальненими виразами, масками слів, а надто – словосполучень, суттєво уповільнюють роботу програм. До прикладу, у корпусі сучасної іспанської мови *CREA* можна використовувати зірочку у сполученнях лише від чотирьох літер і більше, і не можна використовувати декілька байдужих символів в одному запиті. Адже запит, що складається з трьох літер і менше, або той, що містить більше одного байдужого символу, передбачає перебір величезної кількості варіантів.

---

1Інформатикам часто необхідно враховувати діапазон і від нуля знаків, і в такому разі замість символу “+” використовується зірочка “*”.

У малих корпусах, як свідчить успішний досвід пошуку у конкордансері *TexPeer*, цілком можливо проводити складний пошук з використанням більше, ніж одного байдужого символу. Скажімо, якщо потрібно зробити вибірку словосполучення “співати пісень/пісні/пісню” в усіх можливих парадигматичних формах (особових, видових формах дієслова, числа іменника), запит виглядатиме так:

**співа* піс***

У подібних запитах зірочка відповідає будь-якому сполученню символів до пробілу. Оскільки зірочка і знак питання можуть відповідати будь-якому символу, машина перетворює цей користувацький запит на оригінальну мову регулярних виразів:

**співа\w*(\W+\w+)\W+піс**

Так, наприклад, виглядатиме цей самий запит з додатковою умовою: якщо між “співа*” і “піс* допускається наявність до десяти будь-яких слів.

**співа\w*(\W+\w+){0,10}\W+піс**

Це дозволяє не втратити такі результати як “співаємо весь час пісень” або “він співав, коли траплялася така нагода, одну й ту саму веселу пісню”.

Окрім регулярних виразів, у багатослівному пошуку доцільно скористатися синтаксисом спеціальної мови запитів *CQL* – (англ. *Contextual Query Language).* Ось деякі приклади використання цієї мови:

За назвою:

**title = "synesthetic metaphor"**
**title exact = "the synesthetic metaphor"**

З використанням логічних операцій “та”, “або”:

**metaphor or stylistic device**
**metaphor and stylistic device**

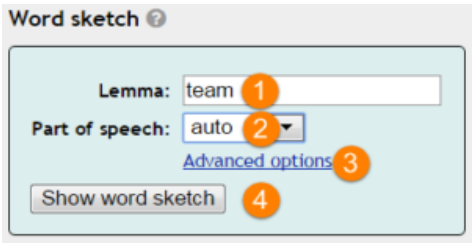
За заздалегідь відомими специфічними параметрами:

**date within "2002 2005"**
**subject any/relevant "metaphor"**

Звичайно, ми не маємо змоги проілюструвати навіть мінімальну базову частину таких запитів, інструкції зі здійснення таких запитів займають сотні сторінок. *CQL* первинно розроблено з метою пошуку видань в електронному каталозі Бібліотеки Конгресу США. Чіткої уніфікації оформлення цих запитів у різних системах пошуку немає. З метою адаптації *CQL* до застосування у конкордансерах розроблено систему пошуку *Word sketch engine*[2]. Пошук ведеться переважно за двома критеріями: вибір леми (початкової форми слова) та вибір частиномовної характеристики:

**[lemma = “try” & tag = “N.”]**

Цей запит означає: шукати усі словоформи “try”, коли це слово є іменником. Такі запити можна реалізувати лише в анотованих корпусах текстів. Інколи для уникнення синтаксичних помилок, користувачеві пропонується спеціальне діалогове вікно для введення леми та частиномовної характеристики:




(https://www.sketchengine.eu/user-guide/user-manual/word-sketch/)

2Готові бази даних для різних мов, у тому числі й української, містяться за посиланнями: https://www.sketchengine.eu/user-guide/user-manual/corpora/by-language/ukrainian-text-corpora/

У корпусах текстів, розроблених в Університеті Бріґама Янґа (США), маска пошуку може оформлюватися в такий спосіб:

- пошук словосполучення, що складається з прикметника і іменника:

* ADJ NOUN

пошук словосполучення з прикметника "gorgeous" та іменника

"gorgeous" NOUN

[The iWeb Corpus].

Отже, у цьому випадку ідеться про частиномовну маску словосполучення. Однак можуть існувати й інші маски: морфологічна маска слова, семантична маска словосполучення, речення тощо.

Перекладачам-практикам, що використовують комп'ютерну програму машинного перекладу на основі комп'ютерної перекладацької пам'яті, корисно знати, що деякі регулярні вирази націлені на автоматичну вибірку термінологічних одиниць з комп'ютерної перекладацької пам'яті. Звичайно, машина не здатна розуміти, чи є сполучення певних літер терміном, чи ні. Однак розробники програми *Déjà Vu* дійшли висновку, що якщо обирати слова, які містяться між артиклем і дієсловом-зв'язкою або між артиклем і модальним дієсловом, між вказівним займенником і дієсловом-зв'язкою або модальним дієсловом, існує висока ймовірність того, що слова, які міститься всередині, є термінами. І для вибірки термінів у такий спосіб можна скористатися виразами:

**\b(the|The)\b.*?\b(?=\W?\b(is|are|was|can|shall| must|that| which|about|by|at|if|when|should|among|above|under|$)\b)**

**\b(a|an|A|An)\b.*?\b(?=\W?\b(is|are|was|can|shall|must |that|which|about|by|at|if|when|among|above| under|$)\b)**

**\b(this|these|This|These)\b.*?\b(?=\W?\b(is|are|was|can| shall|must |that|which|about|by|at|if|when|among|above| under|$)\b)**

[Atril solutions]

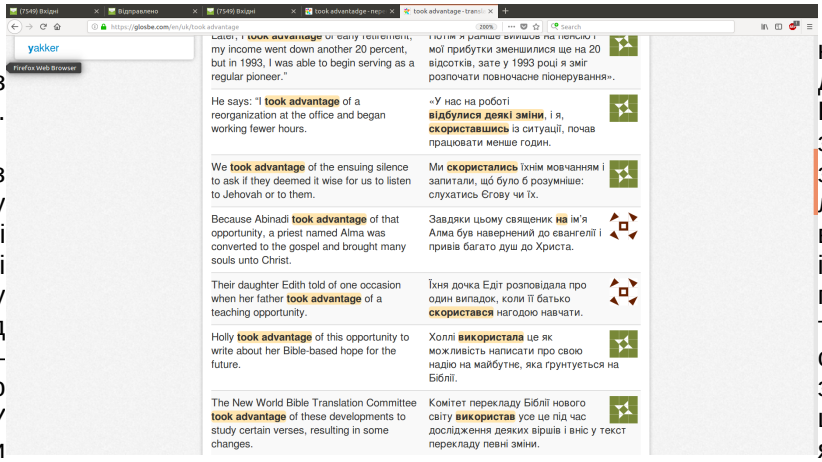
У відповідних іншомовних фрагментах перекладацької пам'яті, цілком імовірно, міститиметься переклад терміна (звичайно, якщо у відповідному фрагменті перекладу не було вилучень, перестановок, додавань, замін).

Отже, у сучасних корпусах пошук можна здійснювати за допомогою різноманітних запитів: простих запитів, що складаються з одного або декількох графічних слів, а також метамовних засобів, з використанням спеціальних дескрипторів, тегів граматичної, частиномовної, семантичної, екстралінгвістичної характеристики слів, виразів, що підлягають пошуку; широкі можливості пошуку надає використання регулярних виразів, які в узагальненому вигляді описують певну множину текстових реалізацій, які підходять під певну маску і можуть бути знайдені за одну операцію. За окремими формальними показниками можна робити висновки й щодо семантики, функцій словоформ. Перевага корпусів текстів над пошуком у загальному масиві інтернетних текстів полягає у тому, що корпуси достовірніше репрезентують як мову в цілому (нормалізовані корпуси), так і окрему підмову, стиль, жанр, дискурс.

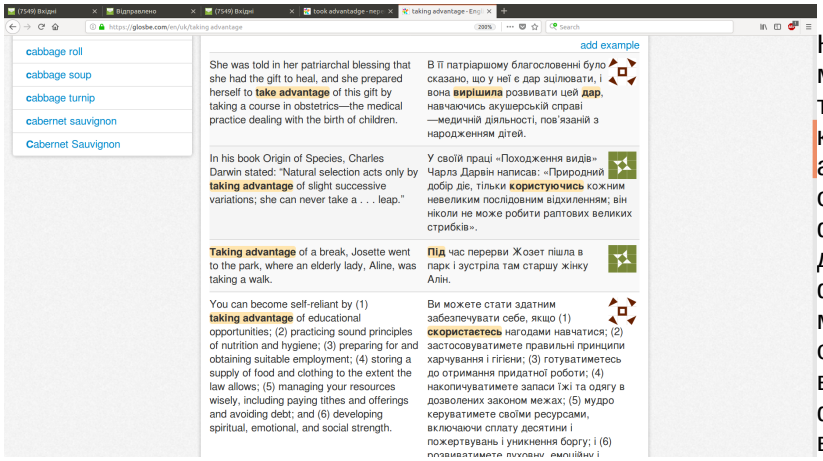
### 3.3. Перекладацькі корпуси і корпуси паралельних текстів

У своїй більшості сучасні корпуси текстів одномовні, адже в цифровому форматі перекладів існує значно менше, ніж оригінальних текстів. Переклади, виконані на особисті замовлення, не публікуються і не підлягають розголошенню. Отже, далеко не завжди існує можливість укласти корпус паралельних текстів двома або більше мовами. Тим не менш, корпуси паралельних текстів теж активно розробляються, і варто розглянути їхні характеристики та специфіку використання.

Про успіхи в галузі укладання корпусів перекладів текстів свідчать такі популярні ресурси як *Glosbe* [Glosbe] або *Linguee* [Contextual Search Engine Linguee], *Reverso* [Reverso], відомі усім сучасним перекладачам. Звичайно, у цих корпусах містяться мовні пари, у рамках яких здійснюється велика кількість перекладів.




Зазначені корпуси складаються з двох колонок і пошукового рядка. Під час пошуку програма знаходить повні та часткові збіги з запитом користувача у лівій колонці, а також імовірні відповідники у правій колонці іншою мовою. Завдяки такому пошуку можна знайти переклад тих слів (а що іще корисніше – словосполучень), які часто зустрічаються і перекладаються. У цих корпусах можна знаходити як фразеологічні одиниці, усталені та термінологічні сполучення, так і просто механічне поєднання слів, що ідуть поспіль і можуть мати відповідник в іншій мові. Найціннішим у таких корпусах є можливість пошуку варіантів перекладу для лексем і словосполучень не лише в початковій словниковій формі, а в особовій/відмінковій формах (так званій "інтегрованій формі"), у складі конструкцій, в оточенні інших лексем. Такий підхід став справжнім проривом в укладанні довідкових перекладацьких ресурсів: класичні словники не здатні вмістити усі можливі сполучення слів, до того ж, у різних парадигматичних формах. Тим паче, що на вибір перекладацького відповідника впливає не лише семантика лексем, а й їхні граматичні характеристики. Скажімо, герундіальна форма "taking advantage" дає підстави для деяких додаткових семантичних модуляцій, що були б неможливими або менш імовірними в разі перекладу особової форми дієслова "took advantage":




Кількість словосполучень певної мови практично неосяжна. Окрім того, що теоретично можлива кількість сполучень слів рахується астрономічно непідйомними обсягами, які обчислюються функцією факторіалу, про вичерпне додавання різних парадигматичних форм до двомовних словників не може бути і мови. На практиці ж сполучаються між собою далеко не всі лексеми й не всі парадигматичні форми, і саме в цьому сенсі величезна перевага корпусів з перекладеними текстами – наявність багатьох або навіть більшості узуальних сполучень слів та варіантів їхнього перекладу. Механічні поєднання тих чи інших графічних слів (що іменують біграмами, триграмами, *n*-грамами) програма виокремлює з текстів автоматично й так само автоматично проводить пошук сполучень слів у правій колонці, що, імовірно, відповідно до розташування у контексті відповідають за значенням сполученню слів у лівій колонці.

Описані вище корпуси, в яких можна шукати перекладні відповідники для слів і словосполучень, слід назвати перекладацькими паралельними корпусами. Вони зазвичай менш відомі, ніж перекладацькі корпуси, далеко не всі з них є у відкритому доступі. Приміром, на сайті "Комп'ютерна лінгвістика в Польщі" ("Computational Linguistics in Poland") наявний Польсько-український врівноважений корпус [Polsko-Ukraiński Korpus Równoległy]. Як підсумовує Н.П. Дарчук, *"*серед наявних нині ресурсів можна згадати паралельний російсько-український підкорпус Національного корпусу російської мови (Режим доступу: http://ruscorpora.ru/s earch-para-uk.html), польсько – український корпус (Режим доступу: http://domeczek.pl/~polukr), багатомовні корпуси *ParaSol* (Режим доступу: http://parasolcorpus.org/ParaVoz/) та Intercorp (Режим доступу: https://kontext.korpus.cz/first_form?corpname=intercorp_v9_uk). Також ведеться робота над болгарсько -українським паралельним корпусом. Відкритий доступ на сьогодні є тільки до російського та польського ресурсів" [Дарчук 2017, с. 29].

Окремо відзначимо Генеральний регіонально анотований корпус української мови [ГРАК]: до його складу входить підкорпус паралельних текстів з 38 іноземних мов у парі з українською. На відміну від звичайних перекладацьких корпусів, ГРАК є анотованим корпусом, у ньому є широкі можливості спеціалізованого пошуку, зокрема, за такими фільтрами: автор тексту, регіон, жанр тексту, місце

публікації, рік публікації, мова оригіналу документа; поза тим, пошук можна здійснювати і за лемою, і за словоформою.

На відміну від перекладацьких корпусів, дослідницькі паралельні корпуси чіткіше структуровані: окрім того, що в них має бути точно вказано, чи є текст оригіналом, чи перекладом, першоджерело фрагментів тексту, наявний зручний інтерфейс з різного роду фільтрами пошуку: за хронологічним, ареальним, соціолінгвістичним, тематичним та іншими критеріями. До прикладу, описуючи корпус паралельних текстів, що розробляється нині у лабораторії комп'ютерної лінгвістики Інституту філології Київського національного університету імені Тараса Шевченка, в якому передбачено наповнення в парі з українською таких іноземних мов: турецької, польської, болгарської, німецької, білоруської, Н.П. Дарчук зазначає: "УпарКУМ реалізовано три види розмітки: метатекстову, структурну і власне лінгвістичну. Метатекстова розмітка передбачає опис завантажених текстів (інформація про час створення тексту (точна дата або приблизний діапазон, якщо точна дата невідома)); бібліографічний опис видання тексту ( якщо це відомо) – ці дані вносяться як стосовно оригіналу, так і для перекладного тексту. Так само зазначаються дані про особу автора та перекладача (ім'я та прізвище, рік народження). Надалі планується доповнити метатекстовий опис жанрово-стильовими характеристиками. Метатекстова розмітка є важливим блоком корпусної розмітки, що виконує низку функцій, зокрема сприяє вибудові архітектури корпусу; дозволяє контролювати його наповнення, стежити за збалансованістю складу; надає користувачеві можливість добирати та групувати текстовий матеріал за різноманітними параметрами, закодованими розміткою: за хронологією, мовою оригіналу чи перекладу, за ім'ям автора чи навіть за його статтю. Це допоможе здійснювати пошук. Мовних явищ у сфері перекладу, обмежених цими параметрами" [Дарчук 2017, с. 30]. Безумовно, це відкриває багатющі перспективи щодо дослідження в галузі теорії перекладу, контрастивістики, методики викладання іноземних мов, комп'ютерної лінгвістики; однак така деталізована розмітка справді зайва і обтяжлива для перекладачів-практиків.