СТРУКТУРНА, ПРИКЛАДНА ТА МАТЕМАТИЧНА ЛІНГВІСТИКА

УДК 81
DOI https://doi.org/10.32838/2710-4656/2021.6-2/20

Строкань А. В.
Національний університет «Львівська політехніка»

МЕТОДОЛОГІЯ ПРОЄКТУВАННЯ ПАРАЛЕЛЬНИХ КОРПУСІВ АКАДЕМІЧНИХ ТЕКСТІВ

У статті проаналізовано методологію побудови паралельного корпусу академічних текстів за допомогою інструменту для корпусних досліджень SketchEngine. Варто відзначити, що SketchEngine є корисним не тільки для лінгвістів і філологів, але і лексикографів, перекладачів і тих, хто вивчає і викладає мову. Цей корпусний інструмент дозволяє лінгвістам досліджувати великі корпуси текстів і створювати складні запити задля витягування нетривіальної інформації із цих корпусів. Окрім того, для аналізу текстів користувачеві доступний великий спектр інструментів, починаючи від звичайного пошуку слова у тексті та закінчуючи спеціальними фільтрами для пошуку пропозицій за певною схемою. Система має також свою власну регулярну мову, яка дозволяє користувачеві?. У цій роботі ми представляємо процес розроблення паралельного корпусу англійської та української мов, можливі напрямки досліджень на його основі. Це перший паралельний корпус англійської та української академічних мов в Україні. Метою роботи є дослідження та аналіз методів створення паралельного корпусу академічних текстів, визначення ключового напрямку дослідження і подальших перспектив лінгвістичних корпусних досліджень. В основу дослідження покладені методи дескриптивного і корпусного аналізу, інтерпретації. Паралельні корпуси вже аналізувала досить велика кількість дослідників, оскільки ця проблема є дуже актуальною. Невирішеною частиною цього питання є групування отриманих даних на категорії, тому у дослідженні ми поетапно зобразили кроки створення паралельного корпусу і виведення термінів за допомогою інструменту вилучення ключових слів. Приклади із корпусу також можуть використовуватися під час навчання мови, оскільки дають студентам практичний матеріал, із яким вони зіткнуться у разі використання мови у реальних ситуаціях міжкультурної комунікації. Корпуси можуть використовуватися для аналізу і виявлення недоліків наявних матеріалів для викладання іноземних мов. Перспективою досліджень є те, що за допомогою паралельного корпусу академічних текстів можна вилучити термінологічну лексику, тобто виділити академічні терміни та укласти словник цих термінів. Варто підкреслити, що за термінологічними складнощами перекладу ховаються культурологічні причини двоякого роду. Одна з них криється у відмінностях мовної свідомості англійців та українців, що зумовлює відмінності у способі об'єктивації пов'язаних між собою понять.

Ключові слова: *корпусна лінгвістика, паралельний корпус, SketchEngine, академічний дискурс, термін.*

Постановка проблеми. Останніми роками створення корпусів і корпусно-орієнтовані дослідження стали невід'ємною частиною діяльності лінгвістів. Корпусна методологія стає частиною лінгвістичної науки і всі лінгвісти, які працюють у різних областях, зазвичай проводять свої дослідження на базі корпусів.

Один із напрямків корпусної лінгвістики – створення і використання паралельних корпусів, що застосовуються для вирішення різноманітних завдань, таких як створення і налаштування систем машинного перекладу, порівняльне вивчення мов, розвиток теорії перекладознавства, навчання мов [1-3]. Корпуси і конкорданси до них надають лінгвістам, перекладачам, перекладознавцям і студентам безцінний і раніше недоступний лінгвістичний матеріал, що характеризується великим обсягом, різноманітністю стилів і жанрів із можливістю швидкого знаходження прикладів на аналізовані слова і конструкції.

У цій роботі ми представляємо процес розроблення паралельного корпусу англійської та україн-

ської мов та можливі напрямки досліджень на його основі. Це перший паралельний корпус англійської та української академічних мов в Україні.

Паралельні корпуси відкривають можливості для компаративістських досліджень, надають нову інформацію порівняно із дослідженнями на базі одномовних корпусів [1, с. 12], розширюють наші знання про мови, їхні універсальні особливості поряд із типологічними і культурними відмінностями.

Постановка завдання. Метою роботи є дослідження детальної методології створення паралельного корпусу академічних текстів, визначення ключового напрямку дослідження і подальших перспектив лінгвістичних корпусних досліджень. В основу дослідження покладені методи дескриптивного і корпусного аналізу, інтерпретації.

Аналіз останніх досліджень і публікацій. Це питання вже аналізувала досить велика кількість дослідників, оскільки проблема є дуже актуальною. Невирішеною частиною цього питання є групування отриманих даних на категорії. Тому у цьому дослідженні ми успішно сгрупували отримані матеріали, зокрема приклади різних способів творення академічної термінології, до кожної групи прикріпили достатню кількість прикладів для підтвердження поданої теорії.

Виклад основного матеріалу. Sketch Engine – інструмент для корпусних досліджень, таких, які виконуються на матеріалі корпусів і великих електронних колекцій текстів. Sketch Engine може бути корисним не тільки для дослідників-лінгвістів і філологів, але і для лексикографів, перекладачів і тих, хто вивчає і викладає мову.

За В. П. Захаровим [18], формування корпусів відбувається за таким алгоритмом: проєктування; забезпечення надходження текстів відповідно до зазначених джерел; підготовка технологічного опису; перетворення у зчитувану машиною форму; конвертування і попереднє оброблення текстів; графематичний аналіз (токенізація); метарозмітка; лінгвістична розмітка (виділення наше, оскільки саме наявність розмітки різних типів уможливлює оперування корпусу як інформаційно-пошукової системи для вирішення практичних завдань); коригування результатів автоматичної розмітки; завантаження розмічених текстів у структуру корпус-менеджера; забезпечення доступу до корпусу (пошук); створення документального забезпечення.

Sketch Engine дозволяє створити «скетч», начерк, образ окремого слова, тексту або навіть цілого корпусу. Із його допомогою ми можемо, наприклад, зрозуміти, в яких контекстах зустрічається слово, що цікавить нас; які ключові слова є цікавими тексту або корпусу, а потім уже інтерпретувати і використовувати отримані результати.

Інструменти Sketch Engine. Наприклад, інструмент «Word Sketch» шукає слова і словосполучення, котрі задає користувач, оцінює частоту їх появи і показує, в яких контекстах зустрічаються задані користувачем слова у корпусі. Інструмент «Concordance» (конкорданс) дозволяє побачити розширений контекст слова/ терміна, тобто не просто поєднання слів, але і цілі речення, в яких зустрічається зазначене слово.

На основі морфологічно розміченого корпусу ця система формує списки слів, в яких міститься інформація про їхню «лінгвістичну структуру». Sketch Engine може видавати список колокацій на потрібному лексичному рівні. Крім того, висвітлюється список із зазначенням частоти кожної колокації у корпусі та значення зв'язку між ключовим словом і колокацією. У системі Sketch Engine є спеціальні інструменти, що визначають рівень синтагматичних і парадигматичних зв'язків на основі дистрибуції лексем у корпусі: тезаурус (thesaurus), кластеризація (clustering) і диференціація (differences) [8].

Система Sketch Engine є веб-системою, яка дозволяє лінгвістам досліджувати великі корпуси текстів і створювати складні запити для того, щоб витягувати нетривіальну інформацію із цих корпусів. Система містить 292 готових текстових корпусів, які користувач може використовувати для своїх досліджень. Якщо розглядати кількість корпусів за мовами, то використовується 70 мов.

Для аналізу текстів користувачеві доступний великий спектр інструментів, починаючи від звичайного пошуку слова у тексті та закінчуючи спеціальними фільтрами для пошуку пропозицій за певною схемою. Крім того, система має свою власну регулярну мову, яка дозволяє користувачеві знаходити певні типи пропозицій і створювати різні спеціалізовані запити. Велика перевага корпусних менеджерів у тому, що порівняно з окремими самостійно зробленими корпусами текстів працювати з ними набагато простіше, адже не потрібно опановувати специфічну символьну мову довільного корпусу. Engine – потужний інструмент для створення свого власного корпусу текстів (підкорпусу) або для завантаження наявних масивів даних. Система дає можливість сформувати частотний словник і згрупувати лексичні одиниці у лексико-семантичні поля.

Під час роботи над корпусом потрібно було виконати такі завдання:

а) провести відбір і початкове введення текстів;

в) створити чи адаптувати модуль пошукової системи (корпусний менеджер);

г) завантажити тексти у корпус;

д) провести статистичний аналіз корпусних даних;

з) проаналізувати отримані результати.

Для того, щоб дослідити методи і прийоми перекладу академічних термінів на основі паралельного корпусу академічних текстів, ми створили власний англійсько-український та українсько-англійський корпуси академічних текстів і досліджували їх за допомогою лінгвістичної програми SketchEngine (https://www.sketchengine.eu/).

Веб-сайт ERASMUS+ (https://erasmusplus.org.ua/), а саме магістерські програми «2020 Erasmus+ Programme Guide», «ProgrammeCountry_benefitsrisks», та «IMPLEMENTATION OVERVIEW_infobox» і журнал «Вісник Маріупольського державного університету» (http://visnyk-pravo.mdu.in.ua/) використані для створення англо-українського паралельного корпусу академічних текстів, а для побудови українсько-англійського корпусу використані різноманітні анотації статей, реферати, наукові автореферати до дисертацій (див. список джерел ілюстративного матеріалу). Обсяг академічних корпусів, зокрема англо-українського, становить 128750 слів, а українсько-англійського – 92884 слова (табл. 1).

Першим етапом створення паралельного корпусу було перенесення оригінально тексту англійською мовою та його перекладу українською мовою за абзацами у документ формату XLSX. У перший стовбець таблиці ми помістили тексти англійською мовою, а у другий – відповідні їм переклади українською мовою. Кожна клітинка першого стовбця створеної таблиці містить один абзац оригінального тексту, кожна клітинка другого стовбця – відповідний абзац перекладу (рис. 3). Загальний обсяг текстів корпусу становить 200 абзаців оригінального англійського тексту і паралельні до нього переклади українською мовою.

Другим етапом створення англійсько-українського паралельного академічного корпусу було завантаження файлу у форматі XLSX у лінгвістичну програму Sketch Engine (рис. 3-10).

English	Ukrainian
he discussion of the changing the status of Ukraine in EU-funded Erasmus+ Programme to Programme ntry (Associated to the Programme), the National Erasmus+ Office in Ukraine (EU-funded project) have pared the description of benefits, risks, status and recommendations, including comparison of the licipation of 5 countries in Erasmus+ for 2014-2020.	Щодо обговорення доцільності ініціації переходу України з країни-партнера в статус країни-члена Програми Еразмус+ (асоційованого), Національний Еразмус+ офіс в Україні (проект ЄС) підготував короткий опис переваг, ризиків, стану та рекомендацій, включаючи порівняння участі 5ти країн-учасників Програми ЄС Еразмус+ 2014-2020 рр. у різних статусах.
efits:	Переваги:
entially, Ukraine in the status of Erasmus+ Programme Country (Associated to the Programme) will have access to all both as applicants and/or partners to all open calls. The Programme Countries organisations re their expertise, exchange the best practices and innovations with other countries in the world, create tegic partnerships and alliances to boost innovations based on the developed innovative capacities. ong other things the Programme Countries organisations have access to school, VET, adult education in ition to other opportunities that are currently open to Ukraine. Additional ones (i.e. mobility in school adult education, sport, small scale partnerships, Teachers academy, policy support under Action 3).	У статусі країни-члена Програми Україна отримує доступ до всіх напрямів, відкритих для країн-членів (асоційованих до) Програми. Доступ полягає у можливості для українських організацій як визнаних в світі експертних центрів, що засвідченого успішними практиками та інноваціями, бути аплікантами та/чи партнери конкурсів проектів. Такі організації з розвинутим інноваційним потенціалом зможуть обмінюватися успішними практиками та інноваціями з іншими країнами світу, створюючи стратегічні партнерства та альянси, також реалізовувати проекти у сферах шкільної, професійної (професійно-технічної) й фахової передвищої освіти на додаток до вже відкритих для України можливостей у сферах освіти, молоді та спорту. Додаткові можливості включають, для прикладу: мобільність у сфері шкільної освіти, освіти дорослих, спорту, малі партнерства, підтримка реформ напряму КАЗ.
aine in the Country Associated to the Programme, will have to meet all rules and implement all onsibilities in cooperation with Participating Countries recognized by the European Union by rnational Law.	Отримуючи статус країни асоційованої до Програми, Україна повинна буде виконувати всі функції та зобов'язання, що виконують інші країни в такому статусі, включаючи співпрацю з усіма країнами, які визнає ЄС як країни-учасниці Програми.
ks:	Ризики:
access to the opportunities will be open to the limited number of Ukrainian organisations. Only the anisations that could present and confirm the existing innovative and inclusive infrastructure, recognized t practices in the education (school, VET, higher education, adult education), training, youth and sports ready to share them with other countries in the world will be able to initiate/join the projects.	Обмеженість доступу українських організацій до участі в проектах. У проектах зможуть брати участь тільки ті організації, які доведуть та продемонструють наявність розвиненої інноваційної інфраструктури та успішних визнаних практик у сферах освіти (шкільної, професійної (професійно-технічної), фахової передвищої, вищої освіти, освіти дорослих), професійної підготовки, молоді та спорту, щоб ділитись успішним досвідом з іншими країнами світу, у якості експертів, які мають інфраструктуру для рівного доступу всіх груп населення (інклюзивність).

Рис. 1. Підготовка до побудови паралельного корпусу з англійської на українську мову у середовищі Microsoft Excel

Таблиця 1

Обсяг академічного англо-українського та українсько-англійського корпусів

Назви текстів у корпусі	Загальна кількість слів у корпусі	Кількість слів англ. мовою	Кількість слів укр. мовою
2020 Erasmus+ Programme Guide	6660	3637	3023
ProgrammeCountry_benefitsrisks	5081	2716	2365
IMPLEMENTATION OVERVIEW_infobox	23114	12093	11021
BULLETIN of Mariupol State University	93865	57091	36774
Загальна к-сть слів в англо-українському академічному корпусі =			128750 слів
Українсько-англійський корпус академічних текстів (на основі анотацій, рефератів, авторефератів наукових статей і дисертацій)	92884	52222	40662

Ukrainian	English
Дисертацію присвячено зіставному дослідженню привативних дієслів (ПД) у сучасній англійській та українській мовах (to cut off / відрізати, to steal / красти). На основі опрацьованих критеріїв добору матеріалу укладено корпус ПД в обох досліджуваних мовах. Створено модель зіставного опису семантики ПД в неблизькоспоріднених мовах, яка зокрема враховує актантну структуру, словотвірну будову і стилістичну маркованість досліджуваних лексичних одиниць.	The thesis focuses on the contrastive semantic analysis of privative verbs (e.g. *to cut off / відрізати*, *to steal / красти*) in modern English and Ukrainian. Using the system of the criteria worked out for the selection of the units for analysis, the corpus of privative verbs in both contrasted languages was compiled. A pattern for contrastive description of the semantics of these verbs in distantly related languages, which involves their actantial structure, word-building and stylistic peculiarities, was developed.
Запропонована класифікація ПД виконана з урахуванням предикатної та аргументної інформації. Простежено взаємозв'язок між стилістичною маркованістю ПД у зіставлюваних мовах та їх належністю до певних лексико-семантичних груп. Виявлено найбільш продуктивні моделі словотворення ПД англійської та української мов та типові смисли, які вони виражають. Проаналізовано особливості актантної структури ПД, установлено співвідношення між їхньою семантикою і морфологічною будовою та заповненням позицій суб'єкта й об'єкта посесивності в реченні.	The classification of privative verbs considering the predicative and argumentative information has been made. The relationship between the stylistic peculiarities of English and Ukrainian privative verbs and belonging of the latter to certain lexico-semantic groups has been traced. The most productive word-building patterns and the typical meanings they convey have been described. The peculiarities of the actantial structure of the verbs under analysis have been discussed. The interdependence between the semantics and the morphological structure of these verbs, and filling up of the positions of the possessive subject and the possessive object in the sentence has been established.
Виявлено спільне та відмінне в семантиці ПД, їхній словотвірній будові, стилістичній маркованості та функціонуванні у мовленні. Привативна семантика в дієсловах англійської та української мов демонструє більше спільного, ніж відмінного (базові семантичні типи, підтипи, групи та підгрупи значною мірою збігаються), що свідчить про ізоморфізм двох зіставлюваних мов у плані вираження привативних значень. Диференційні риси ПД в англійській та українській мовах стосуються переважно продуктивності семантичних типів (підтипів, груп, підгруп), лакуни зафіксовано виняткова в межах окремих підгруп.	Privative semantics in English and Ukrainian verbs demonstrates more common than distinctive features (the basic types, subtypes, groups and subgroups largely coincide), which is indicative of the fact that the contrasted languages are isomorphic in terms of expressing privative meanings. The allomorphic features of English and Ukrainian privative verbs involve the productivity of the semantic types (subtypes, groups, subgroups); the semantic lacunae have been observed exclusively within single subgroups.
Магістерську дисертацію присвячено розгляду можливостей корпусної лінгвістики, аналізу основних проблем, з якими стикаються перекладачі технічної літератури під час перекладу спеціалізованих текстів, а також пропозицій щодо усунення перекладацьких помилок, покращення якості та	This Master's Thesis is devoted to the consideration of corpus linguistics capabilities, investigation of the main problems encountered by technical translators while translating engineering texts and provision of an alternative solution for them to eliminate translation weaknesses, improve quality, and optimize translation itself.

Рис. 2. Підготовка до побудови паралельного корпусу з української на англійську мову у середовищі Microsoft Excel

Рис. 3. Загальний вигляд програми SketchEngine

Рис. 4. Завантаження файлу англо-українського корпусу у форматі XLSX у лінгвістичну програму SketchEngine

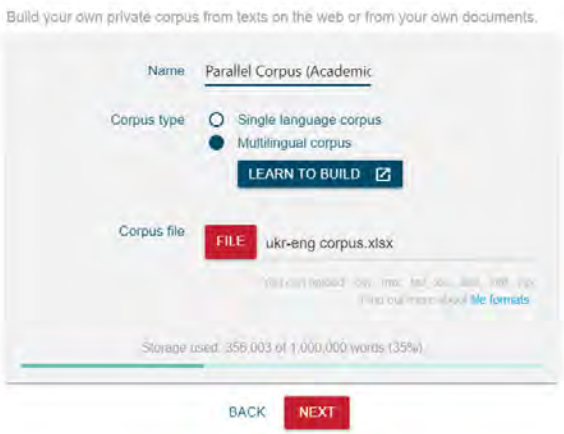

Рис. 5. Завантаження файлу українсько-англійського корпусу у форматі XLS у лінгвістичну програму SketchEngine

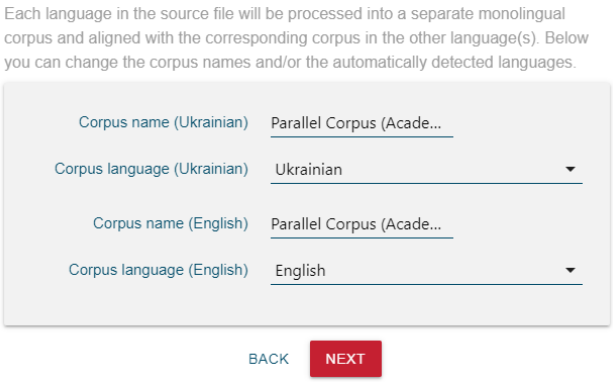

Рис. 6. Налаштування мов в англійсько-українському паралельному корпусі

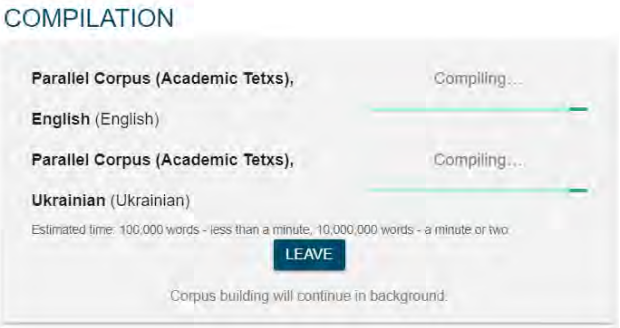

Рис. 7. Завершальний етап (компіляція) створення англо-українського паралельного корпусу

Наступним етапом було вилучення термінів зі створеного паралельного корпусу академічних текстів. Для цього було застосовано функцію ключових слів/вилучення термінів (рис. 8, 9).

KEYWORD у SketchEngine. Цей інструмент порівнює корпуси і визначає, що є унікальним або типовим. Вибраний корпус порівнюють із довідковим корпусом для визначення таких ключових показників:

1) ключові слова, окремі слова (можна залучати будь-який токен);

2) терміни, ключові багатослівні вирази у форматі, типовому для термінології у мові корпусу;

3) N-грамові ключові багатослівні вирази (будь-які послідовності лексем). Залучаються лише елементи, що з'являються у вибраному корпусі частіше, ніж у довідковому.

Результати показують, що є типовим для обраного корпусу порівняно з еталонним корпусом (рис. 8). Вилучення ключових слів і термінів використовується для:

- вилучення термінології для використання у перекладі;
- виділення одиниці слів і більшості слів, які є типовими для корпусу/документа/тексту або визначають його зміст чи тему;
- порівняння двох корпусів/документів/текстів, визначивши, що є унікальним у першому корпусі порівняно із другим.

Результат поділено на ключові слова (одиничні слова) і терміни (багатослівні елементи) (рис. 9) і зображаються разом із посиланнями на речення як у фокусі, так і у довідковому корпусі. Ключові слова і терміни, виділені із корпусу текстів про цифрову фотографію. Інструмент вилучення термінів (Terminology extraction) виділяє слова, типові для теми документа або корпусу, тобто вони з'являються у корпусі частіше, ніж у загальній мові. Для представлення загальної мови використовується великий неспеціалізований корпус у мові. Налаштувань за замовчуванням зазвичай досить для отримання високоякісних результатів.

Граматика термінів – це набір правил, написаних на CQL, які визначають лексичні структури, зазвичай іменникові фрази, які мають залучатися під час виділення термінів. Термін «граматика» використовує POS-теги. Використання граматики термінів забезпечує чистий результат вилучення терміна, який вимагає дуже малого редагування.

Фактичні правила є набагато складнішими і дозволяють артиклі, необов'язкові лексеми. Вони також перевіряють узгодження прикметників та іменників у числі, роду чи відмінку.

Терміни – це поняття, яке використовується через інструмент Keywords & Terms. Термін – це багатослівний вираз, що складається із кількох лексем, який частіше зустрічається в одному корпусі (фокусному корпусі) порівняно з іншим корпусом (довідковим корпусом); водночас вираз має формат терміна у мові. Формат визначається у граматиці термінів, яка є специфічною для кожної мови. Термін «граматика» зазвичай зосереджується на ідентифікації іменникових фраз.

Вилучені терміни є типовими для змісту корпусу і можуть використовуватися для визначення теми корпусу (рис. 9). Інструмент OneClick Terms – це потужний онлайн-вибірник термінів із одномовними і двомовними можливостями їх вилучення. Він працює на основі унікальної технології вилучення термінів від Sketch Engine.

Вилучення ключових слів і термінів. Ключові слова і терміни – це слова і фрази, типові для вашого корпусу, оскільки вони з'являються у вашому корпусі частіше, ніж у загальній мові. Їх можна використовувати для визначення або розуміння основної теми корпусу.

Sketch Engine поєднує статистичні дані із лінгвістичними критеріями для вилучення ключових слів і термінів. Це простий інтерфейс вилучення

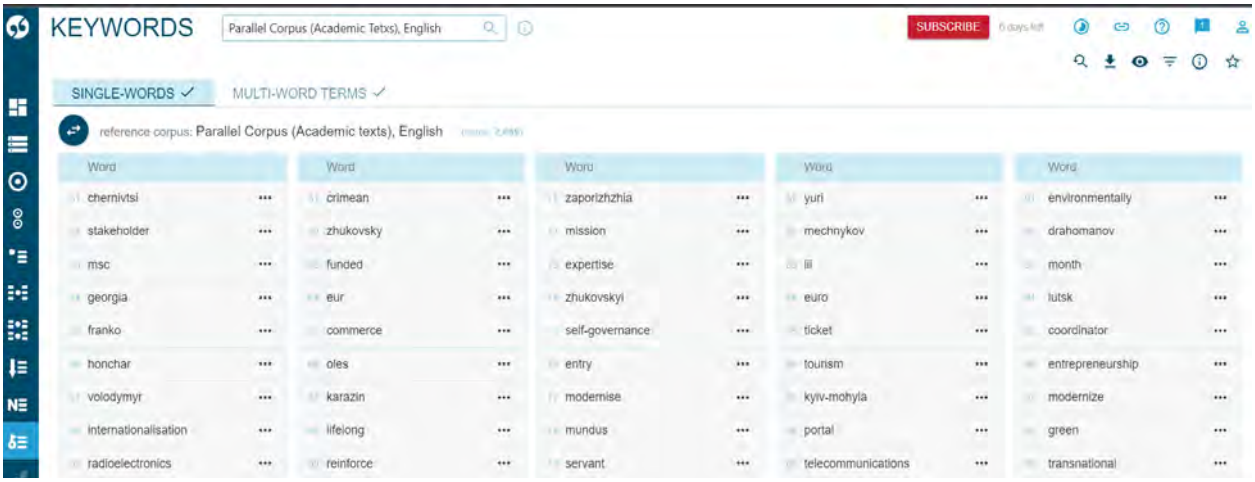

Рис. 8. Ключові слова, знайдені в англійських текстах паралельного корпусу за допомогою функції KEYWORDS

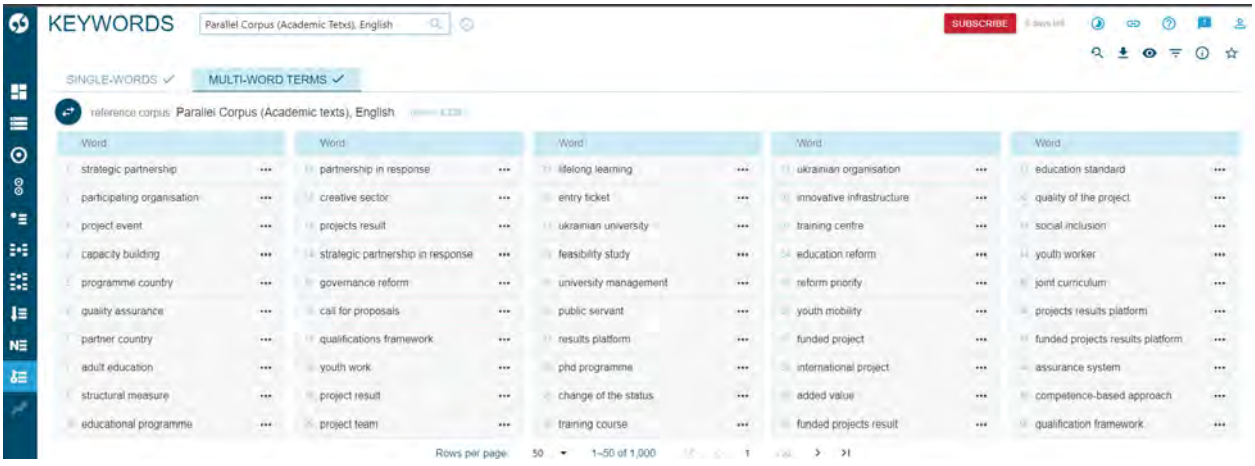

Рис. 9. Фрази (multiword terms), знайдені в англійських текстах паралельного корпусу за допомогою функції KEYWORDS

термінів, що надає легкий доступ до функцій вилучення термінології. Ключові слова – це окремі слова (лексеми), які з'являються у фокусному корпусі частіше, ніж у загальній мові.

Терміни є багатослівними одиницями (фразами), які відповідають двом умовам:

1) вони з'являються у фокусному корпусі частіше, ніж у загальній мові (або у довідковому корпусі);

2) у них є структура, дозволена для термінів у мові (встановлена у граматиці термінів).

Як видно із рис. 8-9, багато слів не є термінами, а лише ключовими словами, тобто вручну було відібрано саме академічні терміни.

Вилучення термінів зазвичай має сенс лише для корпусів користувачів. Ви можете створити корпус зі своїх власних текстів або, якщо у вас його немає, ви можете попросити Sketch Engine знайти відповідні тексти для вас.

Останнім етапом було знаходження українських відповідників до знайдених термінів англійською мовою за допомогою функції паралельного конкордансу (рис. 10). Ця функція дозволила простежити вживання термінів у контексті, визначити метод перекладу кожної знайденої термінологічної одиниці.

Висновки із дослідження і перспективи у цьому напрямку. У багатьох наукових, технологічних чи політичних галузях не вистачає термінологічних словників і довідкової літератури, що створює проблеми перекладачам і призводить до непослідовних і неправильних перекладів.

Корпус паралельних текстів дозволяє проводити порівняння не тільки тексту оригіналу і тексту перекладу, але і, навпаки, порівнювати текст перекладу із текстом оригіналу. Цій стороні процесу перекладу завжди приділялося дуже мало уваги, тоді як глибоке вивчення цих питань дозволить

Таблиця 2

Вилучення ключових слів із англо-українського паралельного корпусу академічних текстів

method name: extract_keywords					
corpus: user/NastiaStrokan/parallel_corpus__academic_tetxs___english					
Item	**Frequency (focus)**	**Frequency (reference)**	**Relative frequency (focus)**	**Relative frequency (reference)**	**Score**
participating organisation	19	220	895,1286	0,00489	891,766
project event	19	272	895,1286	0,00605	890,741
education institution	23	10809	1083,5768	0,24037	874,401
programme country	15	186	706,68048	0,00414	704,765
strategic partnership	22	33572	1036,4647	0,74656	594,005
partner country	12	3211	565,34436	0,0714	528,6
structural measure	10	97	471,12033	0,00216	471,104
educational institution	16	32208	753,79254	0,71623	439,798
educational programme	9	4423	424,0083	0,09836	386,949
projects result	8	0	376,89627	0	377,896
strategic partnership in response	8	0	376,89627	0	377,896
partnership in response	8	0	376,89627	0	377,896
field of education	8	0	376,89627	0	377,896
governance reform	8	2484	376,89627	0,05524	358,115
creative sector	8	4418	376,89627	0,09825	344,091
call for proposals	7	0	329,78424	0	330,784
qualifications framework	7	0	329,78424	0	330,784
young people	7	2	329,78424	0,00004	330,77
adult education	12	43790	565,34436	0,97378	286,934
ukrainian university	6	147	282,67218	0,00327	282,748
project result	6	289	282,67218	0,00643	281,861
entry ticket	6	2437	282,67218	0,05419	269,089
project activity	6	2885	282,67218	0,06416	266,57
capacity building	17	93991	800,90454	2,09013	259,505
change of the status	5	0	235,56017	0	236,56
results platform	5	0	235,56017	0	236,56
ukrainian organisation	5	29	235,56017	0,00064	236,408
reform priority	5	66	235,56017	0,00147	236,213
youth mobility	5	574	235,56017	0,01276	233,579
innovative infrastructure	5	616	235,56017	0,0137	233,363
university management	5	1443	235,56017	0,03209	229,205

краще зрозуміти процеси перекладу із погляду на психологію, когнітивістику, лінгвістику, кібернетику тощо. Водночас потрібно пам'ятати, що у таких дослідженнях важливу роль відіграє напрямок перекладу у паралельному корпусі.

Паралельні корпуси вже перекладених текстів можуть використовуватись як ресурс для автоматичного вилучення стилістично забарвленої лексики, словосполучень та їхніх перекладів.

У цій роботі описано методологію створення паралельного корпусу академічних текстів і способи вилучення термінологічної лексики зі створеного корпусу текстів. Ми використали вирівнювання за абзацами задля створення англійсько-українського паралельного корпусу, а для вилучення термінологічної лексики було застосовано два методи: ключових слів і конкордансів.

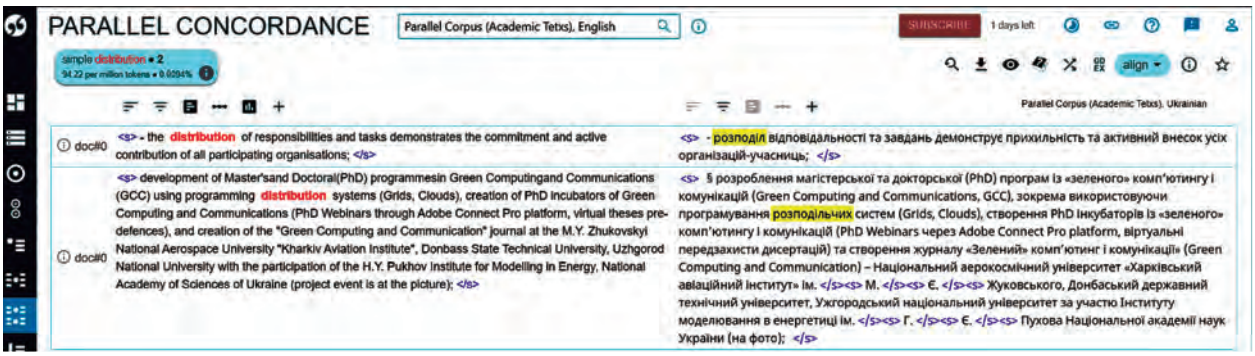

Рис. 10. Результат пошуку українського відповідника до академічного терміна distribution у паралельному корпусі

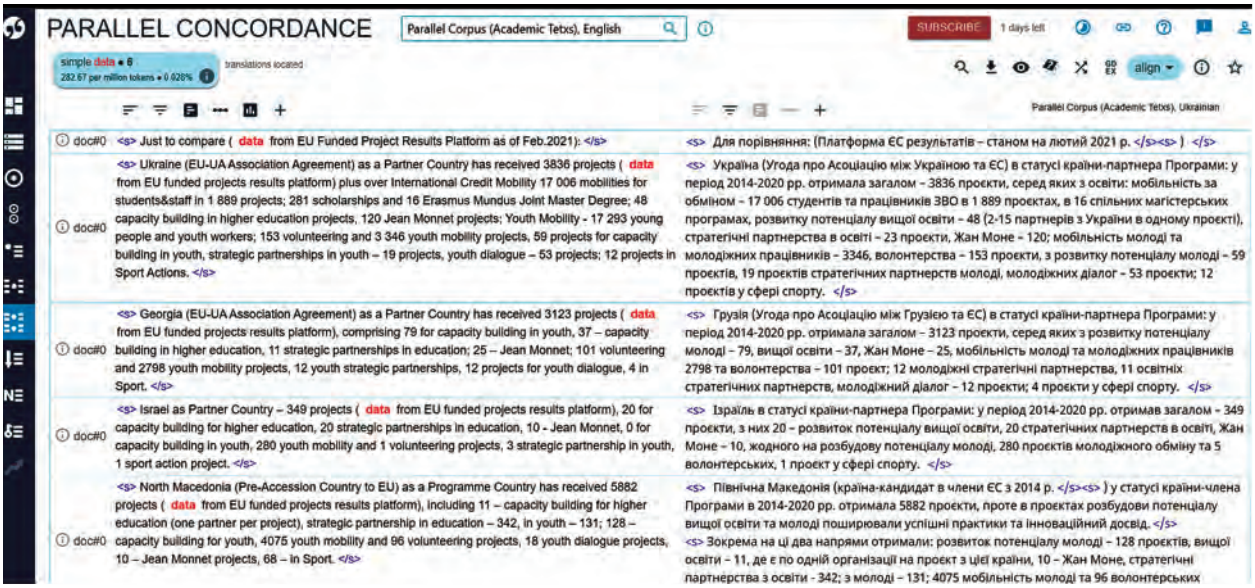

Рис 11. Результат пошуку українського відповідника до терміна data у паралельному корпусі

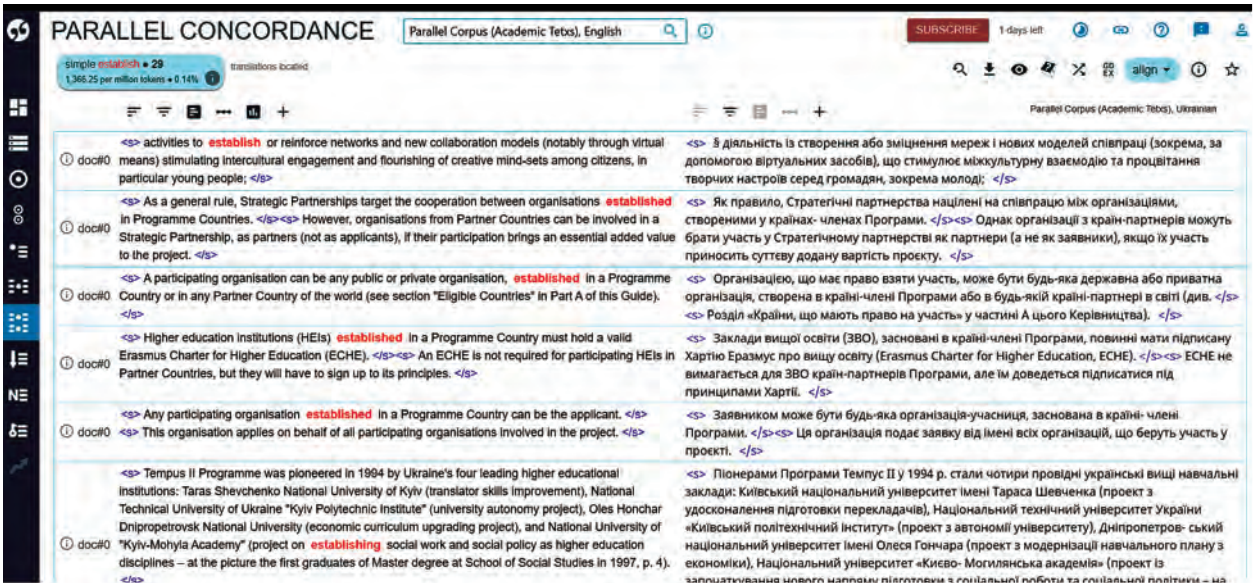

Рис 12. Результат пошуку українського відповідника до терміна establish у паралельному корпусі

У цій роботі ми описали цілі, значення і процес створення паралельного корпусу англійської та української мов, а також способи його використання. Нині у теорії прийняття рішень, зокрема у наукових дослідженнях, існує два підходи: нормативний і дескриптивний. Створення паралельного корпусу дозволяє аналізувати переклад із англійської на українську в аспекті дескриптивного підходу на основі реального мовного матеріалу.

Спеціально створений для цього дослідження англійсько-український корпус паралельних академічних текстів та аналіз методів перекладу комп'ютерних термінів може бути корисним для перекладачів, які стикаються із проблемами під час перекладу текстів у галузі інформаційних технологій. Такий корпус може мати перспективу дослідження методів перекладу академічних текстів, особливо академічних термінів. Варто зауважити, що у SketchEngine можна провести не лише якісне, але і кількісне дослідження; не лише оцінку перекладених текстів, їхніх переваг і недоліків, але і дослідження природи та універсальності перекладеної мови.

Список літератури:

1. Вахтерова Е.В. Понятие академического дискурса в англоязычной лингвокультуре. *Язык и национальное сознание.* 2019. Вып. 25. С. 43-48.

2. Дарчук Н. П. Корпусна лінгвістика: проблеми, методи, перспективи (робоча навчальна програма для аспірантів) Київ : КНУ імені Тараса Шевченка, 2013. С. 11.

3. Жуковська В. В. Вступ до корпусної лінгвістики: навч. посіб. Житомир: Вид-во ЖДУ імені І. Франка, 2013. С. 142.

4. Данчевська Ю. О., Кульчицький І. М., Ліхнякевич І. О. Деякі аспекти створення та використання паралельних корпусів. *Науковий вісник ВНУ ім. Лесі Українки. Серія: Філологічні науки.* 2013. С. 48-52.

5. Кротова Е.Б. Sketch Engine для лингвистических исследований. Германистика сегодня : материалы Международной научно-практической конференции, Казань, 16–17 октября 2018 г. Казанский (Приволжский) федеральный университет. Казань, 2019. С. 107-112.

6. Січінава Д. В., Тищенко-Монастирська О.О., Шведова М.О. Паралельні українсько-російський та російсько-український корпуси. *Лексикографічний бюлетень.* 2011. Вип. 20. С. 35-38.

7. Sketch Engine. URL: https://www.sketchengine.eu/ (дата звернення: 12.11.2021).

8. Stubbs M. British traditions in text analysis: From Firth to Sinclair.In M. Baker, F. Francis and E. Tognini-Bonelli (eds.). Text and technology: In honor of John Sinclair, 1–36. Amsterdam: John Benjamins, 1993.

Strokan A. V. METHODOLOGY OF DESIGNING PARALLEL CORPS OF ACADEMIC TEXTS

The article analyzes the methodology of constructing a parallel corpus of academic texts using the tool for corpus research - SketchEngine. It is worth noting that SketchEngine is useful not only for linguists and philologists but also for lexicographers, translators, and those who study and teach the language. This corpus tool allows linguists to explore large corpora of texts and create complex queries to extract non-trivial information from these corpora. Also for text analysis, the user has a wide range of tools available, from the usual word search in the text, ending with special filters to search for sentences according to a certain scheme, and the system has its regular language that allows the user. In this paper, we present the process of developing a parallel corpus of English and Ukrainian languages and possible areas of research based on it. This is the first parallel corpus of English and Ukrainian academic languages in Ukraine. The article aims to study and analyze the methods of creating a parallel corpus of academic texts, to determine the key direction of research and further prospects of linguistic corpus research. The research is based on methods of descriptive and corpus analysis, interpretation. A large number of researchers have already analyzed parallel buildings, as this problem is very relevant. And the unresolved part of this issue was the grouping of the data into categories. Therefore, in this study, we step-by-step outlined the steps for creating a parallel body and deleting terms, using the keyword extraction tool. Examples from the corpus can also be used in language teaching, as they give students practical material that they will encounter when using language in real situations of intercultural communication. Corpora can be used to analyze and identify shortcomings of available materials for teaching foreign languages. As a perspective for research is that with the help of a parallel body of academic texts it is possible to remove terminological vocabulary, ie to select academic terms and compile a dictionary of these terms. It should be emphasized that the terminological difficulties of translation hide culturological reasons of two kinds. One of them lies in the differences in the linguistic consciousness of the British and Ukrainians, which causes differences in the way of objectification of related concepts.

Key words: *corpus linguistics, parallel corpus, SketchEngine, academic discourse, term.*