
ВИКОРИСТАННЯ КОРПУСНИХ ТЕХНОЛОГІЙ ДЛЯ ЛІНГВІСТИЧНОГО АНАЛІЗУ ТЕКСТУ НАУКОВО-ПОПУЛЯРНИХ СТАТЕЙ ВВС

Саламін Н. А.

DOI <https://doi.org/10.30525/978-9934-26-590-7-12>

ВСТУП

Досягнення науково-технічного прогресу виходять далеко за межі дослідницьких лабораторій і міцно закріплюються у сфері виробництва, житті та побуті звичайних людей. Вплив науки на життя пересічної людини настільки значний, що знати новини доводиться кожному освіченому читачеві. Джерела наукової інформації є досить різноманітними: радіопрограми, телевізійні передачі, фільми, книги, газети, журнали, Інтернет – портали. Їхньою відмінністю є подача специфічного матеріалу цікавою, доступною мовою, зрозумілою для читачів, які не є професіоналами у тій чи іншій науковій сфері.

Науково-популярні статті – це джерело, звідки отримує наукову інформацію більшість людей, які цікавляться наукою. «Саме науково-популярна література покликана виконати це відповідальне соціальне замовлення: залучити до сучасних досягнень науки і техніки широкі маси».¹ Цільова аудиторія науково-популярного тексту – читач, який не має спеціальних знань, не володіє основною термінологією в тій чи іншій галузі науки. При характерній науковому стилю суворості, логічності викладу в таких статтях присутня емоційно-експресивна складова, коли автор має право висловити своє ставлення до наукового відкриття, що обговорюється, або вдатися до інших способів, щоб зацікавити читача і утримати його увагу.

Інкорпорування новітніх технологій у науку дозволяє по-новому вирішувати її фундаментальні завдання. У теорії та практиці перекладу активно використовуються методи корпусної лінгвістики з метою виявлення закономірностей використання мовних одиниць у мовленні, що особливо актуально при перекладі нерідною мовою. Такий підхід

¹ Король Т. Г. Англомовна науково-популярна стаття в галузі управління як об'єкт перекладу. *Мова і міжкультурна комунікація*. 2019. № 2. С. 230–237.

має евристичну цінність для перекладу, особливо для практики перекладу та оцінки якості перекладу, оскільки при перекладі першорядну важливість представляє вміння визначати відповідності не стільки між мовними системами, скільки між вживаннями мовних одиниць у мові², а оцінка перекладу стає менш суб'єктивною.³

Використання корпусних методів у перекладі активно вивчається останнім часом як у вітчизняних⁴, так і в зарубіжних дослідженнях. Так, наприклад, корпусні методи дослідження перекладу були принесені в сферу аудіовізуального перекладу.

Корпусні дослідження також широко представлені в галузі юридичного перекладу.⁵ Було проведено корпусні дослідження у перекладознавстві розвідках в Угорщині.⁶ Особливу цінність представляють дослідження на стику кількох областей знання: перекладознавства, корпусної лінгвістики та когнітивної лінгвістики.⁷ Також розглядається застосування конкретних можливостей корпусних менеджерів у дескриптивному перекладознавстві.

Такий інтегративний підхід у сфері порівняльно-перекладацьких досліджень широко представлений у сучасних роботах закордоном, проте в раїні йому приділено увагу лише у кількох публікаціях. Широта та різноманітність робіт у галузі застосування корпусних методів у перекладознавстві можуть бути підтвердженням перспективності досліджень у цій сфері. Тим не менш, багато питань залишаються невирішеними, у тому числі: чи можуть корпусні дослідження зробити внесок у прескриптивне перекладознавство, наприклад з точки зору визначення можливих варіантних відповідностей у мові перекладу, а також надати дані про можливі відповідності з точки зору їх схожості та відмінності у системі мови, чи лише на рівні закономірностей їх використання у мовленні

Вибір корпусного методу для дослідження науково-популярних статей є доцільним, оскільки науково-популярний стиль поєднує особливості різних мовленнєвих стилів, та дозволяє виокремити закономірності, які реалізуються у всіх функціональних стилях незалежно

² Buts J., Jones H. From Text to Data: Mediality in Corpus-Based Translation Studies. *MonTI. Monografias De Traducción E Interpretación*. 2021

³ Bowker L. A Corpus-Based Approach to Evaluating Student Translations. *Translator*. 2000. Vol. 6. Iss. 2.

⁴ Дарчук Н. П., Алексієнко Л. А., Сорокін В. М. Термін у лінгвістичній інформатиці. *Інформаційні технології в освіті*. 2009. № 4. С. 191–199.

⁵ Corpus-Based Research in Legal and Institutional Translation: Special Issue / ed. by F. P. Ramos. *Translation Spaces*. 2019. Vol. 8. Iss. 1.

⁶ Edina R., Götz A., Pataky E. Translation Studies and Corpus Linguistics: Introducing the Pannonia Corpus. *Acta Universitatis Sapientiae. Philologica*. 2018. Vol. 9. Iss. 3.

⁷ Garcia A. F. Can a Corpus-Driven Lexical Analysis of Human and Machine Translation Unveil Discourse Features that Set Them Apart? 2021.

від тематики наукових текстів, що, у свою чергу, дозволяє залучити для аналізу весь матеріал корпусів, а не обмежитися будь-який тематикою чи функціональним стилем мовлення. Результати дослідження можуть застосовуватися в практиці перекладу, особливо на другому етапі перекладацького семіозису, тобто при перефразуванням мови оригіналу засобами мови перекладу, на етапі вибору мовних засобів для вираження інваріанту.

1. Особливості створення корпусу текстів за допомогою корпусного менеджера Voyant Tools

У представленому дослідженні висвітлено підхід до створення корпусу текстів на основі матеріалів інтернет-ресурсу BBC, а також описано методику аналізу науково-популярного контенту, що здійснюється на його основі. Наведено приклади аналітичних досліджень, які стали можливими завдяки сформованому корпусу. Для організації та обробки текстів було використано інструмент Voyant Tools.

Voyant Tools – це інтерактивне веб-середовище, створене з метою підтримки аналітичного читання та інтерпретації текстових матеріалів. Цей цифровий інструмент розроблено як частину наукового проєкту, який спрямований на вдосконалення методів роботи з текстами.

Використання Voyant Tools відкриває такі можливості:

- автоматизований аналіз текстових масивів;
- дослідження онлайн-контенту, зокрема інтернет-статей та публікацій;
- інтеграція аналітичних візуалізацій до веб-платформ, таких як блоги, сайти або онлайн-журнали;
- розробка власних функціональних рішень на базі наданого програмного коду.

Серед ключових характеристик Voyant Tools варто виділити:

- модульність – можливість комбінувати інструменти в різних конфігураціях;
- універсальність – придатність для аналізу текстів різних типів та форматів;
- доменно – орієнтована чутливість – здатність адаптуватися до різних способів цифрового представлення тексту;
- гнучкість у роботі з джерелами – підтримка як локальних, так і мережових даних;
- багатомовність – інтерфейс та функції підтримують різні мови користувачів;
- інтеграційна сумісність – наявність відкритих API для взаємодії з іншими цифровими системами;

– адаптивність в оформленні – можливість налаштування зовнішнього вигляду та функціоналу згідно з уподобаннями користувача.

Корпусний менеджер Voyant Tools надає широкі можливості для формування індивідуального корпусу текстів. Користувачі можуть додавати текстовий матеріал у головне поле введення, обираючи один із двох форматів подачі:

- суцільний текст у вигляді одного документа (підтримуються формати TXT, HTML, XML);
- перелік веб-адрес (URL), розміщених по одному в кожному рядку.

Щоб створити корпус на основі локальних даних, потрібно завантажити один або кілька файлів із комп'ютера. Для виділення кількох файлів одночасно зручно використовувати комбінації клавіш Shift або Ctrl. Також замість окремих файлів можна створити zip-архів, який легко завантажується в систему як єдиний об'єкт.

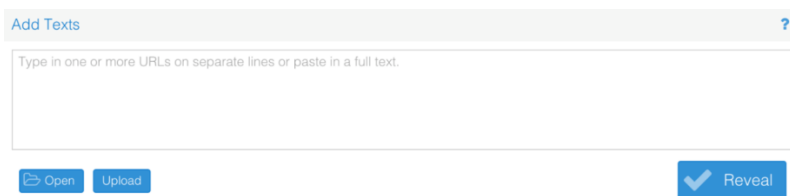


Рис. 1. Інтерфейс завантаження корпусу в Voyant Tools із зазначенням параметрів перед запуском обробки

Усі необхідні параметри для обробки слід зазначити до натискання кнопок Upload або Reveal, що відповідають за початок завантаження або відкриття корпусу (Рис. 1).

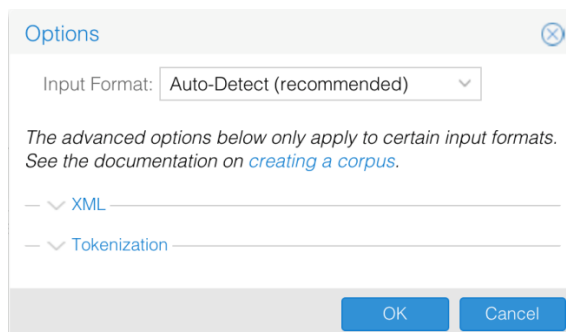


Рис. 2. Підтримувані формати файлів у Voyant Tools та можливості завантаження архівів

Voyant Tools підтримує широкий спектр текстових форматів і зазвичай самостійно розпізнає типи документів. Програма ефективно обробляє тексти, отримані з HTML-сторінок, документів Microsoft Word, Excel, OpenDocument (ODT), Apple Pages, PDF-файлів, звичайного тексту (TXT), RTF, XML та інших. Крім того, можливо завантажувати архіви у форматах .zip, .tar, .tgz, що містять відповідні документи.

У випадку, коли автоматичне визначення формату не працює належним чином, є можливість вказати формат вручну, обравши один із таких варіантів:

- **Atom** – формат на базі XML, що широко використовується у стрічках новин та блогах;
- **DToC (Dynamic Table of Contexts)** – спеціалізований XML-формат, застосовуваний у рамках проєкту Dynamic Table of Contexts;
- **RSS** – XML-формат новинних та блогових стрічок, підтримується починаючи з версії 2.0;
- **TEI (Text Encoding Initiative)** – формат, орієнтований на наукову розмітку текстів, із чіткою структурою заголовків і авторів;
- **TEI Corpus** – варіант TEI, який дозволяє опрацьовувати декілька текстів одночасно, використовуючи теги //TEI.

Працюючи з текстами BBC у Voyant Tools, слід зважати на низку важливих особливостей:

- Voyant Tools продовжує традицію гуманітаристики у сфері створення інструментів для читання, аналізу, пошуку та візуалізації цифрових текстів.
- Сервіс орієнтований на доступність: він не вимагає інсталяції або реєстрації, а робота з ним можлива для файлів у різних форматах, включаючи PDF, DOCX, XML, TXT тощо.
- Voyant – це проєкт з відкритим кодом, що перебуває в активній розробці, тому іноді можуть траплятися технічні неточності. Всі результати варто інтерпретувати критично.
- Програма підтримує тексти більшістю мов, представлених у стандарті Unicode, однак не завжди має глибоку мовну специфікацію, що обмежує, наприклад, можливості семантичного аналізу.
- Інструменти призначені для підтримки інтерпретаційного аналізу, а не для автоматизованого виведення результатів чи постановки запитань.
- Voyant зручно інтегрується у колективні дослідницькі проєкти, дозволяє обмінюватися корпусами між користувачами та вбудовувати інтерфейси у вебсторінки, подібно до вбудовування відео.

Зібраний корпус (див. рис. 3) містить різносторонню інформацію зі статей BBC і є мультимодальним за своєю природою, оскільки містить контент, представлений через кілька функціональних модулів.

2. Особливості використання корпусу статей BBC для встановлення лінгвістичних особливостей науково-популярного тексту

Особливістю корпусу створеного шляхом застосування менеджера Voyant Tools є те, що він доступний іншим користувачам за своєю адресою.

Обсяг анотованого за допомогою Voyant Tools корпусу становить 34,073 слів, 5,420 особливих словоформ.

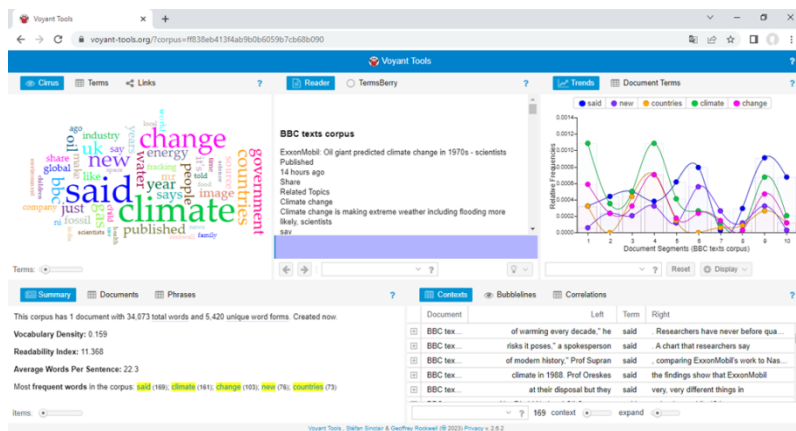


Рис. 3. Основні текстові метрики та частотний аналіз корпусу

Рисунок 3 надає загальні характеристики корпусу:

- щільність словника роману (Vocabulary Density): 0.159
- Індекс зручності читання (Readability Index: 11.368
- Середній відсоток слів у реченні (Average Words Per Sentence): 22.3
- Найчастотнішими словами в корпусі : said (169); climate (161); change (103); new (76); countries (73).

[illegible]

The screenshot shows the Voyant Tools web interface. The top navigation bar includes the Voyant Tools logo and a search bar. The main content area displays a histogram and a line graph. The histogram shows the frequency of words across document segments, with the x-axis labeled 'Document Segments (BBC tests corpus)' and the y-axis labeled 'Relative Frequency'. The line graph, titled 'change', shows the relative frequency of words across the same segments. The data points are as follows:

Document Segment	Relative Frequency (Histogram)	Relative Frequency (Line Graph)
1	0.00095	0.00095
2	0.00025	0.00025
3	0.00035	0.00035
4	0.00105	0.00105
5	0.00025	0.00025
6	0.00035	0.00035
7	0.00025	0.00025
8	0.00005	0.00005
9	0.00045	0.00045
10	0.00025	0.00025

На рисунку 8 представлено відносну частоту унікальних слів.
На рисунку 9 представлено сполучуваність *climate*, та його використання в контекстах.

The screenshot shows the Voyant Tools web interface. At the top, there's a search bar with the URL 'voyant-tools.org/Corpus=H332eb41344b763d6507967c668b090?query=climate&view=Contents'. Below the search bar, the 'Contexts' tab is selected. The main area displays a concordance table for the word 'climate'. The table has four columns: 'Document', 'Left', 'Term', and 'Right'. The 'Document' column lists various sources, mostly 'BBC lex...'. The 'Left' column shows the text immediately preceding the word 'climate'. The 'Term' column contains the word 'climate'. The 'Right' column shows the text immediately following the word 'climate'. The table lists 161 contexts, with the first 11 visible in the screenshot.

Document	Left	Term	Right
BBC lex...	ExxonMobil: Oil giant predicted	climate	change in 1970s - scientists Published
BBC lex...	hours ago Share Related Topics	climate	change Climate change is making
BBC lex...	Share Related Topics Climate change	climate	change is making extreme weather
BBC lex...	flooded more likely, scientists say	climate	change is making extreme weather
BBC lex...	scientists say By Georgina Rannard	climate	and science reporter One of
BBC lex...	oil companies accurately forecast how	climate	change would cause global temperature
BBC lex...	the rest of us that	climate	models were bunk," Naomi Oreskes
BBC lex...	oil industry made us doubt	climate	change A simple guide to
BBC lex...	change A simple guide to	climate	change "ExxonMobil is committed to
BBC lex...	part of the solution to	climate	change and the risks it
BBC lex...	actual temperature increase "Their excellent	climate	modelling was at least comparable
BBC lex...	most influential and well-regarded	climate	scientists of modern history," Prof
BBC lex...	who sounded the alarm on	climate	in 1988. Prof Oreskes said
BBC lex...	avoid the worst effects of	climate	change in a world warmed
BBC lex...	evidence suggesting ExxonMobil's knew about	climate	change, but were accused by
BBC lex...	the company was internally using	climate	science when publicly it called
BBC lex...	over what it knew about	climate	change. Campaigners allege it spread
BBC lex...	over accusations it led about	climate	change. Related Topics Climate change
BBC lex...	about climate change. Related Topics	climate	change Environment Exxon Mobil Top
BBC lex...	approach to external linking. ADVERTISEMENT	climate	change. UAE names oil chief

Рис. 9. Сполучуваність лексеми «climate» та приклади його використання в контекстах

Також, було здійснено кількісний аналіз сполучуваності лексеми *climate* з іншими лексемами.

Як бачимо лексема *climate* сполучається найчастіше із лексемою *change*.

Наведемо приклади сполучуваності лексеми *climate* з лексемою *change* у таблиці 1:

Таблиця 1

Приклади сполучуваності лексеми *climate* з лексемою *change*

№	Лівий контекст	Лексема	Правий контекст
1	2	3	4
1.	ExxonMobil: Oil giant predicted	climate	change in 1970s – scientists
2.	Share Related Topics Climate change	climate	change is making extreme weather
3.	scientists say By Georgina Rannard	climate	and science reporter One of
4.	oil companies accurately forecast how	climate	change would cause global temperature
5.	the rest of us that	climate	models were bunk," Naomi Oreskes
6.	oil industry made us doubt	climate	change A simple guide to
7.	change A simple guide to	climate	change "ExxonMobil is committed to
8.	part of the solution to	climate	change and the risks it
9.	actual temperature increase "Their excellent	climate	modelling was at least comparable
10.	most influential and well-regarded	climate	scientists of modern history," Prof
11.	who sounded the alarm on	climate	in 1988. Prof Oreskes said

Продовження таблиці 1

1	2	3	4
12.	avoid the worst effects of	climate	change in a world warmed
13.	evidence suggesting ExxonMobil's knew about	climate	change, but were accused by
14.	the company was internally using	climate	science when publicly it called
15.	over what it knew about	climate	change. Campaigners allege it spread
16.	over accusations it lied about	climate	change. Related Topics Climate change
17.	about climate change. Related Topics	climate	change Environment Exxon Mobil Top
18.	approach to external linking. ADVERTISEMENT	climate	change: UAE names oil chief
19.	day ago Share Related Topics	climate	change Sultan Ahmed Al Jaber

За тим же принципом можна проаналізувати сполучуваність конкретних термінів чи лексем.

Наприклад лексема *global* міститься в колокаціях *global ocean* та *global target*, лексема *energy* зустрічається у колокаціях *energy bills*, *energy jobs*, *energy systems*, *energy transition*.

ВИСНОВКИ

Проведене дослідження підтверджує ефективність використання корпусного підходу для аналізу науково-популярних текстів, зокрема матеріалів з Інтернет-ресурсу BBC. Створений корпус обсягом понад 34 тисячі слів та понад 5 тисяч словоформ дав змогу виявити структурні й лексичні особливості текстів, а також дослідити типові колокації й тенденції вживання термінів. Застосування Voyant Tools, зокрема інструмента TermsBerry, продемонструвало цінність цифрових методів для візуалізації та глибинного аналізу мови.

Вибір корпусного методу для дослідження науково-популярних статей є доцільним, оскільки науково-популярний стиль поєднує особливості різних мовленнєвих стилів, та дозволяє виокремити закономірності, які реалізуються у всіх функціональних стилях незалежно від тематики наукових текстів, що, у свою чергу, дозволяє залучити для аналізу весь матеріал корпусів, а не обмежитися будь-який тематикою чи функціональним стилем мовлення. Результати дослідження можуть застосовуватися в практиці перекладу, особливо на другому етапі перекладацького семіозису, тобто. при перефразуванням мови оригіналу засобами мови перекладу, на етапі вибору мовних засобів для вираження інваріанту.

Отримані результати засвідчують, що корпусний метод є релевантним інструментом для лінгвістичного аналізу стилістично змішаних текстів і може бути корисним у практиці перекладу, особливо на етапі семантичного переосмислення тексту. Надалі подібні підходи можуть слугувати основою для міжмовних і міжстильових досліджень у сфері прикладної лінгвістики та перекладознавства.

АНОТАЦІЯ

У сучасному мовознавстві спостерігається зростання інтересу до інтеграції цифрових інструментів у процес аналізу мовного матеріалу, зокрема – до застосування корпусних технологій. Лінгвістичний аналіз науково-популярних текстів, як різновиду масової комунікації, вимагає точного, системного та репрезентативного підходу, що забезпечує виявлення мовних закономірностей на великому обсязі даних. У цьому контексті використання корпусних платформ, таких як Voyant Tools, відкриває нові можливості для дослідників – від статистичного опрацювання лексики до виявлення дискурсивних стратегій та тематичних домінант.

Особливий інтерес становить вивчення науково-популярного дискурсу, представленого у публікаціях авторитетних міжнародних ЗМІ, зокрема BBC. Цей медіаресурс поєднує високу інформативність з доступною для широкої аудиторії мовою, що робить його ідеальним матеріалом для лінгвістичного аналізу. Однак системні дослідження цього типу текстів у межах корпусної лінгвістики залишаються недостатньо опрацьованими, що й зумовлює актуальність даного дослідження.

Література

1. Дарчук Н. П., Алексієнко Л. А., Сорокін В. М. Термін у лінгвістичній інформатиці. *Інформаційні технології в освіті*. 2009. № 4. С. 191–199.
2. Король Т. Г. Англomовна науково-популярна стаття в галузі управління як об'єкт перекладу. *Мова і міжкультурна комунікація*. 2019. № 2. С. 230–237. URL: http://nbuv.gov.ua/UJRN/movmk_2019_2_25
3. Bowker L. A Corpus-Based Approach to Evaluating Student Translations. *Translator*. 2000. Vol. 6. Iss. 2.
4. Buts J., Jones H. From Text to Data: Mediality in Corpus-Based Translation Studies. *MonTI. Monografías De Traducción E Interpretación*. 2021. URL: <https://genealogiesofknowledge.net/wp-content/uploads/2021/06/From-text-to-data-Mediality-in-corpus-based-translation-studies.pdf>

5. Corpus-Based Research in Legal and Institutional Translation: Special Issue / ed. by F. P. Ramos. *Translation Spaces*. 2019. Vol. 8. Iss. 1. URL: <https://doi.org/10.1075/ts.8.1>

6. Edina R., Götz A., Pataky E. Translation Studies and Corpus Linguistics: Introducing the Pannonia Corpus. *Acta Universitatis Sapientiae. Philologica*. 2018. Vol. 9. Iss. 3. URL: https://www.researchgate.net/publication/322754677_Translation_Studies_and_Corpus_Linguistics_Introducing_the_Pannonia_Corpus

7. Garcia A. F. Can a Corpus-Driven Lexical Analysis of Human and Machine Translation Unveil Discourse Features that Set Them Apart? 2021. URL: <https://benjamins.com/catalog/target.20065.fra>

Джерела ілюстративного матеріалу

1. The BBC www.bbc.com
2. The BBC www.bbc.com/ukrainian
3. Voyant-tools. URL: <https://voyant-tools.org/>

Information about the author:

Salamin Nadiia Andriivna,

Lecturer at the Department of Ukrainian studies and intercultural communication,

Lviv State University of Life Safety
35, Kleparivska str., Lviv, 79007, Ukraine