РОЗДІЛ 6
СТРУКТУРНА, ПРИКЛАДНА ТА МАТЕМАТИЧНА ЛІНГВІСТИКА

УДК 81’33
DOI https://doi.org/10.32782/tps2663-4880/2022.21.1.47

КОРПУСНА ЛІНГВІСТИКА: СУЧАСНИЙ СТАН ТА ПЕРСПЕКТИВИ ДОСЛІДЖЕНЬ

CORPUS LINGUISTICS: MODERN APPROACH AND RESEARCH PERSPECTIVE

Голощук С.Л.,
orcid.org/0000-0001-9621-9688
кандидат філологічних наук, доцент,
доцент кафедри іноземних мов
Національного університету «Львівська політехніка»

У статті розглянуто аналіз характерних рис корпусної лінгвістики як одного з пріоритетних інструментів розвитку філологічної науки. Корпусна лінгвістика – це широкомасштабне вивчення мовних даних – комп’ютерний аналіз дуже великої колекції транскрибованих висловлювань або письмових текстів. У статті описуються основні методи корпусної лінгвістики, окреслюється вплив генеративної лінгвістики на корпусну лінгвістику та досліджуються основні підходи до використання корпусних даних. Корпусна лінгвістика розглядається як один із засобів, що використовується для тлумачення лінгвістичних даних у гуманітарних та соціальних науках.

Представлено аналітичний підхід до базових понять корпусної лінгвістики. Поняття корпусу тлумачиться як сукупність комп’ютеризованих автентичних текстів, підібраних і упорядкованих відповідно до вимог користувача, що є репрезентативними зразками певної мови чи мовних варіантів. Зроблено спробу висвітлити основні вимоги до створення анотації (розмітки) корпусу, яка обмежується автоматичною анотацією електронного тексту, що є найпоширенішим видом розмітки в контексті сучасної корпусної лінгвістики. Розглядаються наступні три типи: (1) метатекстова розмітка, (2) структурна та (3) лінгвістична розмітка.

Окреслено можливості використання корпусного підходу для лінгвістів та спеціалістів інших галузей науки. У прикладних галузях корпус активно використовується викладачами як рідної, так і іноземної мови, оскільки забезпечує фахівців необхідними інструментами для обробки фактологічного матеріалу. Корпусний підхід слугує невичерпним джерелом інформації для укладання словників і граматик нового покоління, тому що дає можливість використовувати найновіші зразки ілюстративних джерел. За допомогою корпусного підходу легко досліджувати й мову інтернет-комунікації, що дозволяє отримати інформацію про мову, яка, як правило, не піддається інтуїтивному огляду, і простежити її зміни на лексичному, семантичному та граматичному рівнях.

Методологія корпусної лінгвістики використовується і в нових професійних галузях, зокрема в криміналістичній лінгвістиці – науці, що виникла на стику лінгвістики і кримінології. Її завдання полягає в тому, щоб дослідити достовірність юридичних текстів за допомогою корпусних даних та його інструментарію.

Ключові слова: текст, корпус, корпусна лінгвістика, розмітка корпусу, прикладне значення корпусної лінгвістики.

The article deals with the analysis of the characteristic features of corpus linguistics. Corpus linguistics is the study of language data on a large scale – the computer-aided analysis of very extensive collections of transcribed utterances or written texts. The article outlines the basic methods of corpus linguistics, explains the influence of the generative linguists on corpus linguistics, and surveys the major approaches to the use of corpus data. Clear and detailed explanations lay out the key issues of method and theory in contemporary corpus linguistics. Corpus linguistics is viewed as a key methodology to explain large linguistic data in the humanities and social sciences. An overview of the main concepts relating to corpus annotation is provided. It is restricted to automatic annotation of electronic text, which is the most common kind of annotation in the context of contemporary corpus linguistics. We focus on the annotation of texts and consider the following three main types of annotation: (1) metatext annotation, (2) structural annotation, and (3) linguistic annotation. Possibilities of using the corpus approach for linguists and specialists of other branches of science are outlined. Using corpora in language teaching has enabled language research with its natural language approach. Corpus tools help to investigate actual usages and characteristics of certain genres in order to improve syllabus design and develop more effective classroom exercises. E-communication analysis allows extracting information on language that does not tend to be open to intuitive inspection and trace its changes on the lexical, semantic, and grammar level. Besides pure linguistic inquiry, researchers had begun to apply corpus linguistics to other academic and professional fields, such as the emerging sub-discipline of law and corpus linguistics called forensic linguistics, which seeks to understand legal texts using corpus data and tools.

Key words: text, corpus, corpus linguistics, corpus annotation, application of corpus linguistics.

Постановка проблеми. Розвиток комп'ютерних технологій спричинив появу нових технічних можливостей для відбору й аналізу функціональних характеристик мовних одиниць у різних типах текстів. Створення корпусів певної мови є характерною ознакою сучасного розвитку філологічної науки, оскільки надає дослідникові доступ до необхідної лінгвістичної інформації з метою її подальшого вивчення та опрацювання. Корпусна лінгвістика (КЛ) стає одним із пріоритетних інструментів сучасних лінгвістичних досліджень та "зосереджується на аналізі природної мови в умовах реального функціонування з використанням комп'ютерних технологій на основі великих за обсягом, ретельно відібраних та впорядкованих текстових корпусів" [1].

Мета даного дослідження полягає в тому, щоб окреслити сучасну корпусну лінгвістику як інноваційну методику емпіричного мовознавства та продемонструвати переваги залучення корпусних даних у сучасні мовознавчі розвідки та суміжні галузі науки.

Об'єктом дослідження виступає корпусна лінгвістика та її понятєва база. Предмет дослідження становить вивчення основних тенденцій практичного використання інструментарію КЛ та окреслення подальших перспектив її впровадження. Матеріалом вивчення є аналіз напрацювань спеціалізованої літератури в галузі за останні роки.

Теоретичне дослідження, яке розкриває і обґрунтовує більш глибинні і суттєві аспекти корпусної лінгвістики є методологічною основою нашої праці, а тому було використано методи аналізу та синтезу інформації.

Аналіз останніх досліджень і публікацій. Корпусне мовознавство пропонує новий погляд на мову, яка, по суті, і сама є корпусом [2]. Праці провідних представників корпусної лінгвістики висвітлюють фундаментальні проблеми організації корпусів текстів та застосування результатів їх аналізу в лінгвістичних дослідженнях [3, 4, 5].

Незважаючи на значну кількість підходів до визначення поняття корпусу, існують спільні погляди на те, що корпус – це комп'ютеризовані автентичні тексти, підібрані і упорядковані згідно з експліцитними критеріями, визначеними користувачами, вони є репрезентативними зразками певної мови чи мовних варіантів [4, 6]. Отже, корпусом може називатися репрезентативна вибірка текстів в електронній формі, доступ до якої забезпечується ретельно розробленими дослідницькими комп'ютерними програмами пошуку та аналізу [7, с. 80]. Основними рисами корпусу текстів є: мета або логічна ідея, машиночитаний формат, репрезентативність як результат певної відбіркової процедури, а також наявність металінгвістичної інформації [8]. Електронне зібрання текстів на машинному носії дозволяє використовувати стандартні програми для швидкого пошуку слів і конструкцій із заданими граматичними та іншими властивостями, що відповідають науковим гіпотезам лінгвістів.

У США корпусна лінгвістика зазнала критики з боку основоположника генеративізму Н. Хомського. Вчений розглядав корпусний спосіб накопичення мовних даних неадекватним і хибним для опису породжувальної здатності природної мови, пояснюючи це тим, що лише інтуїція мовця може замінити корпус і стати джерелом мовного матеріалу (див.: [9, 6, 10]). На думку дослідника, завдання лінгвістів полягає у моделюванні мовної компетенції, яку трактував як знання мови (граматична правильність) і протиставляв її використанню мови (прийнятність, мовленнєва діяльність). Відповідно до цього положення, науковець трактував корпусну лінгвістику як не ефективне джерело фактологічного матеріалу, оскільки її база даних формується на вибірці зовнішніх висловлювань, які не дозволяють моделювати мовну компетенцію (див.: [6, с. 6]).

У нашому дослідженні ми поділяємо думку представників Ланкастерської школи корпусної лінгвістики, які стверджують, що не можна використовувати корпусну лінгвістику як єдиний засіб пояснення природньої мови, тому що жодний корпус не може адекватно представити мову [6, с. 9]. Дослідник Е. МкЕнері зазначає, що дослідження мови на основі реальних мовних даних є більш надійним джерелом, аніж внутрішні роздуми над нею, а тому для пояснення мовних явищ слід брати до уваги обидва підходи [6, с. 14].

Виклад основного матеріалу. Кількість і різноманітність створюваних корпусів щоразу зростає й на сьогодні їх зафіксовано понад 600 [8]. Відповідно до обсягу текстової вибірки, корпуси поділяються на малі, великі та середні. Корпуси, які налічують менше одного мільйона слововживань вважаються малими, від одного мільйона до десяти мільйонів – середніми, та від десяти і понад сто мільйонів – великими. До розряду середніх корпусів належать: American Heritage Intermediate (AHI) на 5 млн. слововживань; Esti kirjakeele korpus (корпус текстів естонської мови) на 1 млн. слововживань; а великих: FRANTEXT – один із найбільших французьких лінгвістичних проектів, розпочатий у 1963 році, в межах якого

створено корпус обсягом понад 90 млн. слововживань; Bank of English на 320 млн. слововживань; Mannheimer Korpora (корпус німецької мови) обсягом 778 млн. слововживань [3]. З-поміж сучасних корпусів англійської мови (як американського, так і британського варіантів) найбільш відомими є Британський національний корпус (British National Corpus – BNC), Міжнародний корпус англійської мови (International Corpus of English – ICE), Лінгвістичний банк англійської мови (Bank of English), Корпус сучасної американської англійської мови (Corpus of Contemporary American English – COCA) тощо.

Серед різноманіття можливостей, які пропонує корпусний підхід для автоматичного опрацювання лексичних одиниць мови, слід виокремити укладання конкордансів — особливий тип словника, що подає до кожного реєстрового слова (чи словоформи) всі або вибіркові контексти його вживання. Це дає можливість дослідникам упорядковувати й подавати конкорданси як окремих творів, так і збірки творів письменників. Найпоширенішим джерелом конкордансів різними мовами є Біблія; упорядковано конкорданс поетичних творів Тараса Шевченка, Луция Аннея Сенеки, Данте, В.Шекспіра, П.де Ронсара, О.С.Пушкіна, М.Горького, У.Блейка, У.Уїтмена та ін. [11, с. 84].

Анотацію корпусу вперше було здійснено у 1978 році і з того часу зацікавленість у виконанні такої роботи поступово зростає і стає все більш різноманітною. Джофрей Ліч визначає анотування корпусу як практику додавання до електронного писемного чи усного корпусу інтерпративної, лінгвістичної інформації. Анотація також визначається і як кінцевий продукт цього процесу: лінгвістичні символи, які додаються до електронної репрезентації мовного матеріалу. Відзначимо, що саме анотація, або розмітка, — головна характеристика корпусу, яка і відрізняє його від електронних колекцій, бібліотек, енциклопедій, широко представлених в сучасному Інтернеті. В.А. Плунгян зазначає, що корпусом може називатися лише те електронне зібрання текстів, яке супроводжується розміткою, незалежно від його об'єму [2].

Розроблено певні рекомендації для укладачів корпусу щодо створення анотації з метою спрощення правил її укладання та уніфікації використання. Найважливішими аспектами є: можливість використовувати як анотований корпус, так і його первинний текст; анотація повинна існувати поза текстом; користувач корпусу повинен мати доступ до усієї документації, яка містить інформацію про схему анотування, дані розробників, дані апробації анотації; забезпечувати можливість її використання для різноманітних дослідницьких цілей, містити теоретично нейтральний аналіз інформації; та не претендувати на абсолютний стандарт розробки [4, с. 6-7].

Як відомо, чим багатша і різноманітніша розмітка, тим вищою є наукова і навчальна цінність корпусу. Базуючись на корпусне зібрання текстів, дослідник має можливість фактично перевірити не лише теорію мови, але й інші прикладні гіпотези з певної галузі.

Існують різні типи розмітки:

- метатекстова розмітка: автор, назва, дата створення, обсяг, тематика тексту і т. д., яка характеризує текст в цілому;
- структурна розмітка: інформація про структуру тексту, яка дозволяє відокремити одне слово від іншого, виділити межі словосполучення, речення, тексту;
- лінгвістична розмітка: приписування одиницям тексту певної лінгвістичної інформації (заперечне речення або питальне, спонукальне або примикання і т. д.).

Якісна анотація корпусу також дозволяє дослідникові швидко і ефективно знайти ті слова і конструкції, які йому потрібні в межах певного наукового дослідження. Для цього програма пошуку повинна розуміти як мінімум те, які форми у тексті відносяться до одного й того ж слова (мати –ім. жін.р., мати – ім. чол. р. мн., мати – дієсл., інф.), тобто хоча б частково "розуміти" граматичну структуру даної мови.

Корпусний підхід до вивчення мови відкриває нові можливості як для лінгвістів, так і для фахівців інших галузей науки. У прикладних галузях корпус активно використовується викладачами як рідної, так і іноземної мови. Для фахівців, чия діяльність пов'язана зі словом, актуальною є проблема допустимого вживання слова чи словосполучення, його використання в певних типах текстів. Саме на основі корпусної вибірки, можна чітко та зі статистичними даними простежити ту чи іншу слово сполучуваність та підібрати найкращу з них.

Корпус може слугувати джерелом для вивчення етимології певної лексичної одиниці – датованість текстів корпусу дозволяє досить легко простежити ці зміни і зрозуміти причину їх появи чи згасання. За допомогою корпусу також можна спостерігати над динамікою змін у мові: які нові слова чи конструкції з'явилися, а які стають застарілими чи вже вийшли з ужитку.

Корпусний підхід є невичерпним джерелом інформації для укладання словників і граматик

нового покоління, тобто таких, які укладені на основі певного корпусу. Такий підхід дає можливість інтегрального опису мовних фактів. При цьому морфологічні, синтаксичні, семантичні, прагматичні, стилістичні, комунікативні, сполучувальні властивості мовних одиниць аналізуються й пояснюються лише на основі текстів.

За допомогою корпусного підходу легко досліджувати й мову інтернет-комунікації (переписка електронною поштою, чати, форуми, блоги). Електронні тексти поповнюються новими лексемами за рахунок запозичень (драйвовий, меседж), професіоналізмів (запатчити, онлайн) та слів інших груп лексики. Запозичення нерідко використовуються в англійському написанні, проте транслітеруються українською абеткою. Як наслідок — одиниці писемного тексту перебирають на себе функції усної мови. Все це як своєрідний сленг, але він цікавий і важливий для лінгвістів, тому що прогнозує шляхи майбутнього розвитку мови.

Отримав новий виток розвитку і автоматичний переклад тексту - переклад окремого словосполучення чи речення можна шукати в паралельному корпусі [2]. Це досить новітній підхід, який потребує подальшого вивчення й опрацювання.

Напрацювання корпусної лінгвістики стає активним джерелом ресурсу для фахівців у сфері криміналістичної лінгвістики (forensic linguistic analysis), яка досліджує формальні методи встановлення авторства текстів сучасних документів з метою використання отриманих результатів у слідчих діях чи судочинстві. Дослідження в криміналістичній лінгвістиці проводять для того, щоб з'ясувати час та місце створення документа, писаря й автора тексту, наявність силових дій щодо писаря в процесі створення документа, а загалом визначити — підроблений документ чи автентичний. Використання методів корпусної лінгвістики дозволяє здійснити точний та достовірний аналіз таких даних.

Для програмістів у галузі автоматичної обробки текстів (в тому числі і різного роду пошукових систем), аналіз мовних одиниць за допомогою корпусу є також одним із цікавих і актуальних напрямків дослідження. Розробка програмного забезпечення тісно пов’язана з природньою мовою, і розпізнавати структуру текстів, написаних тією ж мовою сприяє як розвитку корпусних досліджень, так і вдосконаленню інформаційних технологій.

Висновки. Можливості залучення корпусних даних у царину лінгвістичних розвідок є невичерпним джерелом інформації для лінгвістів. Тому перспективою наших подальших розвідок буде практичне опрацювання інструментарію корпусної лінгвістики для аналізу фактологічного матеріалу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Жуковська В.В. Корпусна лінгвістика: історична перспектива та сучасний стан. URL: http://www.rusnauka.com/12_KPSN_2012/Philologia/3_108393.doc.htm (дата звернення: 29.06.2020).

2. Плунгян В.А. Почему современная лингвистика должна быть лингвистикой корпусов?: (публичная лекция, прочитанная 01.10.2009). URL: http://www.polit.ru/lectures/2009/10/23/corpus.html (дата звернення: 12.11.2020).

3. Демська-Кульчицька О.М. Репрезентативність як ознака текстового корпусу. Українська мова, 2005. – №3. – С. 100-107.

4. Leech G. Introducing Corpus Annotation. Corpus Annotation. Linguistic Annotation from Computer Text Corpora. – London, New-York: Routledge, 1997. – P. 1-19.

5. McEnery T. Why use a corpus? URL: https://www.futurelearn.com/courses/corpus-linguistics-2014-3/steps/14706/progress

6. McEnery T., Wilson A. Early corpus linguistics and the Chomskyan revolution. Corpus Linguistics. An Introduction. – Edinburgh: Edinburgh Universty Press, 1996. – P. 2-27.

7. Голощук С.Л. Історичні передумови розвитку корпусної лінгвістики. International Academy Journal: Web of Scholar. – 6 (15), September 2017. – с. 80-84.

8. Таценко Н.В. Методи корпусної лінгвістики в підготовці фахівців-філологів. Scientific and pedagogic internship “Organization of educational process in the field of philological sciences in Ukraine and EU countries”: Internship proceedings, August 24 – October 2: тези доповідей. Венеція: Venice : Izdevnieciba “Baltija Publishing”, 2020. С. 177-181. URL: https://essuir.sumdu.edu.ua/bitstream-download/123456789/81070/1/Tatsenko_corpus.pdf;jsessionid=A87ACDDFA0B7C776542F181F209F634C (дата звернення: 03.02.2021).

9. Селіванова О.О. Сучасна лінгвістика: напрями та проблеми: Підручник. – Полтава: Довкілля-К, 2008. – 712 с.

10. The Linguistic Encyclopedia. Second edition. Edited by Kirsten Malmkjaer. URL: https://books.google.com.ua/books?id=uCrXOLvD7fMC&printsec=frontcover&hl=uk&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false (дата звернення: 19.11.2020).