РОЗДІЛ VI
Теоретичні засади лінгвістичних досліджень

УДК 81’22:81’42

Галина Войтко

Корпусна лінгвістика: історія формування і перспективи розвитку

У статті проаналізовано основні етапи розвитку та становлення корпусної лінгвістики як складової частини прикладної лінгвістики. Розглянуто основні корпуси текстів і їх характеристики, простежено вплив електронних корпусів на подальший розвиток лінгвістичних досліджень. На сучасному етапі розвитку лінгвістики застосування електронного корпусу текстів стало невід’ємним складником багатьох галузей мовознавства та суміжних дисциплін.

Ключові слова: корпусна лінгвістика, корпус, породжувальна граматика, розмітка, конкорданс.

Постановка наукової проблеми та її значення. З розвитком комп’ютерних технологій розвивається така наука, як корпусна лінгвістика. Корпуси – інструмент дослідження мовного матеріалу й моніторингу мовних явищ.

Перш ніж корпусна лінгвістика стала самостійною дисципліною, минуло багато часу. Існує багато різних думок про статус корпусної лінгвістики та доцільності використання корпусів текстів для дослідження мовних явищ.

Сьогодні корпусна лінгвістика й корпуси мовних даних відіграють важливу роль у навчанні та вивченні мов. Створення і аналіз корпусів текстів відкриває нові перспективи для проведення лінгвістичних досліджень, допомагає виявити зміни, які відбуваються в мові під впливом різноманітних зовнішніх факторів. Створення електронного корпусу відкриває нові можливості для аналізу мовного матеріалу та використання корпусів для навчальних цілей.

Аналіз досліджень цієї проблеми. Дослідження у сфері корпусної лінгвістики торкаються питання створення перших електронних корпусів і поступове їх поширення як ефективного засобу дослідження мови, а також становлення корпусної лінгвістики як самостійної дисципліни. Серед дослідників, які зробили вагомий внесок у розвиток корпусної лінгвістики, варто назвати імена Х. Кучери (H. Kucera 1979), Д. Байбера (Biber 1990, 1992), Дж. Синклера (Sinclair 1994), Ґ. Кеннеді (Kennedy 1998), Н. Іде (2000), М. Банька (Backo 1996; 2003), Т. Ерявця (Erjavec 2001), Т. Макенері й А. Вілсон (2001), А. Баранова (2001), С. Шарова (2002), В. Рикова (2001, 2001), Л. Ричкової (2003) та ін.

Мета статті – описати основні етапи в історії розвитку корпусної лінгвістики, поява перших корпусів та їх розвиток.

У статті поставлено такі **завдання**: визначити терміни «корпусна лінгвістика», «корпус»; схарактеризувати основні етапи розвитку корпусної лінгвістики, створення корпусів текстів; визначити перспективні дослідження в галузі корпусної лінгвістики.

Виклад основного матеріалу й обґрунтування отриманих результатів дослідження. Корпусна лінгвістика – напрям, завдання якого – розробити теоретичні засади і практичні прийоми побудови, машинного опрацювання та експлуатації лінгвальних даних, оформлених як корпус текстів. Об’єктом корпусної лінгвістики є корпус, а предметом – текст [1, с. 45].

Корпусна лінгвістика як окремий розділ науки про мову виникла для розв’язання таких проблем:

− спосіб представлення та збереження мовленнєвих репрезентацій;

− вимоги до корпусу текстів із боку укладачів і користувачів;

− специфіка програмного забезпечення корпусів;
− принципи відбору параметрів проблемної галузі;
− способи структуризації корпусу;
− транскрипція текстів усного мовлення;
− мультимедійна підтримка корпусів усного мовлення;
− розробка пошукових систем у корпусі;
− кодування дескрипцій одиниць збереження тощо [6, с. 669].

Тобто корпусна лінгвістика зосереджена на розв'язанні питань відбору та способів представлення інформації в корпусі, а також використання корпусу текстів для розв'язання лінгвістичних завдань.

З появою комп'ютерів та доступу до Інтернету активізуються дискусії щодо визначення і статусу корпусної лінгвістики.

Російський лінгвіст, професор Володимир Олександрович Плунгян переконаний, що корпусна лінгвістика – не просто наука про те, як створювати корпуси і як ними користуватися, а певна ідеологія, основні тенденції якої зародилися ще в класичній філології XIX ст., але значно інтенсифікувалися останніми десятиліттями. Корпусна лінгвістика пропонує новий погляд на мову, яка, на думку вченого, і сама є корпусом [5].

Щодо визначення поняття корпус, то В. П. Захаров наводить таке визначення: «Лінгвістичний корпус текстів розуміють як великий, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, представлений в електронному вигляді й призначений для розв'язання різних лінгвістичних завдань» [3, с. 3].

Тобто корпус – це зібрання текстів, відібраних за певним критерієм, яке певним чином впорядковне, але основною особливістю корпусу є його розмітка.

Розмітку (анотування, маркування) вважають тим критерієм, який відмежовує корпуси текстів від простого зібрання електронних текстів [3].

О. О. Селіванова вважає головним поняттям корпусної лінгвістики корпус мовленнєвої реалізації мови, що кваліфікується як сформована за певними вимогами вибірка мовленнєвого матеріалу, яку можна використовувати для опису й дослідження мови як системи [6, с. 668].

Розмітка лінгвістичної інформації в корпусі найчастіше проводиться з використанням мови SGML/XML.

Однак SGML/XML задає лише синтаксис представлення фрагментів і атрибутів, а не конкретний набір, який використовують при розмітці корпусу. На основі XML останнім часом розроблено кілька рекомендацій: EAGLES (European Advisory Group on Language Engineering Standards), TEI (Text Encoding for Interchange), і XCES (XML Corpus Encoding Standard) [7, с. 14].

Послідовність при розмітці – найбільш важливий чинник у визначенні якості анотованого ресурсу, тобто одні й ті самі мовні явища анотуються таким же чином, і аналогічні або пов'язані явища повинні отримати розмітку, яка представляє їх подібність або спорідненість, якщо це можливо [9, с. 233].

Саме відповідно розмічений корпус дає змогу будувати конкорданс – список усіх уживань даного слова в контексті з посиланнями на джерело. Корпуси можна використовувати для отримання різноманітних довідок і статистичних даних про мовні й мовленнєві одиниці. На основі корпусів можна отримати дані про частоту словоформ, лексем, простежити зміну частот і контекстів у різні періоди часу [3, с. 4].

Щодо зародження корпусної лінгвістики, то тут сформувалося кілька підходів.

Зародження корпусної лінгвістики пов'язують із появою Браунівського корпусу 1963 р., який став першим зібранням текстів, котрі можна було розглядати як корпус текстів, а не просте їх зібрання. Авторами першого корпусу на машинному носії були У. Френсис і Г. Кучера, які створили корпус для дослідження американського варіанта англійської мови. Загальний обсяг корпусу – 1 млн слів. Згодом це число й загалом структура побудови корпусу стала стандартом для створення аналогічних корпусів текстів.

Крім того, деякі вчені датують зародження корпусної лінгвістики початком минулого століття, а не серединою чи початком другої половини XX ст. Відповідно, виділяють два періоди в розвитку корпусної лінгвістики: протокорпусна лінгвістика й корпусна лінгвістика. Протокорпусна лінгвістика – це період фіксування мовних даних і текстових зібрань на паперових носіях. Основні доробки цього періоду – конкорданси Біблії, а згодом і конкорданси літературних творів. Це, наприклад, конкорданс робіт У. Шекспіра (A Concordance of Shakespeare), який було укладено 1787 р.

Найважливішим і найвпливовішим доелектронним корпусом вважають The Survey of English Usage, який уклав Рендольф Квірк 1959 р. в Університеті-коледжі Лондона (University College London). Корпус – велика база даних на картонних картках [2].

Звичайно, створення текстових зібрань та укладання конкордансів на паперових носіях – трудомісткий процес, що вимагає значних затрат часу і праці, тому розвиток комп'ютерних технологій сприяв переходу від протокорпусної лінгвістики до корпусної лігнвістики, безпосередньо пов'язаної з машинними носіями.

Термін «корпусна лінгвістика» поширився після публікації 1984 р. збірника «Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research».

Відповідно до поділу корпусної лінгвістики на протокорпусну і корпусну, 1961 р. був проміжним періодом, коли в лінгвістиці панували ідеї раціоналізму та породжувальної граматики Наома Хомського [1].

Сповільнення розвитку корпусної лінгвістики пов'язане з поширенням ідей раціоналізму й породжувальної граматики Н. Хомського. Н. Хомський критикував корпуси як базу для дослідження мови й загалом корпусний підхід до вивчення мовних явищ.

Щодо текстового корпусу Н. Хомський висловив таку думку: «у будь-якому корпусі природної мови існують спотворення. Деяких речень у них не буде, бо вони очевидні, інших – тому, що вони хибні, ще інших – тому, що вони неввічливі. Отже, природномовний корпус дасть настільки сильно спотворену картину, що базований на ньому опис виявиться звичайним списком мовних одиниць» [8, с. 67].

Наом Хомський заперечував і критикував корпус текстів як базу для дослідження мови, оскільки вважав, що дослідження емпіричних даних – абсолютно беззмістовне заняття, оскільки суть лінгвістики полягає у вивченні мовної компетенції, а не її відображення – мовної діяльності [10].

Критика корпусів зумовила тимчасовий занепад корпусної лінгвістики, але досліджень у цій сфері учені не припиняли.

Після створення Браунівського корпусу розпочався період створення аналогічних корпусів, які відповідали стандартам йому. Перша версія Браунівського корпусу була представлена простим текстовим форматом (із невеликою кількістю структурної розмітки для виділення абзаців, заголовків, цитованих фрагментів і т. п.). Пізніше корпус був доповнений розміткою частин мови й морфологічних ознак слів.

Услід за Браунівським корпусом з'явився ще один корпус, який вплинув на хід лінгвістичних досліджень – Ланкастерсько-Осло-Бергенський корпус (Lancaster-Oslo-Bergen), структура якого була унаслідувана від Браунівського корпусу. Анотована версія корпусу з'явилася 1985 р.

Створення Браунівського й Ланкастерського корпусів відкрило нові можливості для дослідження мови, зокрема порівняння двох варіантів англійської мови (американського та британського), на текстах різних жанрів, доступних комп'ютерній обробці.

Крім згаданих, складалися корпуси для лексикографічних досліджень (American Heritage Intermediate), для вивчення розмовної англійської (Lancaster/IBM Spoken English Corpus, Corpus of Spoken American English), діахронічні корпуси (Helsinki Corpus of English Texts: Diachronic Part), корпуси для лінгводидактичних досліджень (International Corpus of Learner's English) [4].

Створення аналогічних корпусів лише підтвердило той факт, що корпус обсягом 1 млн слововживань не достатньо інформативний і не може дати достовірні результати. І вже у 90-х рр. XX ст. розвиток технологій зберігання та обробки текстів дав можливість створювати корпуси обсягом 100 млн слововживань і більше – це були корпуси другого покоління.

Одним із таких корпусів був *The Bank of English* – так званий моніторинговий корпус, створення якого розпочато 1990 р. за ініціативи видавництва Collins і факультету англійської мови університету Бірмінгема. 1997 р. корпус нараховував 300 млн слововживань, а 2005-го – уже 525 млн. 1989 р. корпус Банк англійської мови мав обсяг у 20 млн слів, а на сьогодні його розмір сягнув 650 млн слів.

Цей корпус став основою для словника Collins COBUILD English Dictionary й низки англійських граматик [2].

Ще одним корпусом, який став зразком для представницьких корпусів, був Британський національний корпус (British National Corpus). Особливість його полягала в тому, що, на відміну від попередніх корпусів, Британський національний корпус був зібранням повних тестів із частиномовною

розміткою та підкорпусом усного мовлення. Обсяг корпусу – 100 млн слів і можливість доступу через Інтернет.

За стандартом, заданим Британським національним корпусом, було укладено національні корпуси іспанської, італійської, хорватської, чеської мов.

За спільним проектом десятки університетів було створено Інтернаціональний корпус англійської мови (International Corpus of English), який дає змогу порівняти та вивчати специфіку слововживання в різних діалектах англійської мови, не лише в британському й американському, а й у кенійському, новозеландському, сінгапурському [4].

Висновки та перспективи подальшого дослідження. Масове поширення електронних корпусів відбувається із середини 90-х рр. XX ст. У цей період корпусна лінгвістика остаточно оформилася як окремий розділ науки про мову.

Отже, пройшовши складний шлях розвитку та становлення, сьогодні корпусна лінгвістика – це пріоритетний напрям сучасних лінгвістичних досліджень, який значно розширив можливості для дослідження мовного матеріалу. Адже корпус будь-якої мови потрібен насамперед для виявлення закономірностей функціонування мовних одиниць, для використання корпусу з навчальною метою. Тому сьогодні актуальні наукові дослідження, присвячені розробленню принципів і методології укладання корпусу, оскільки на сучасному етапі застосування електронного корпусу стало невід'ємним складником багатьох галузей мовознавства та суміжних дисциплін.

Джерела та література

1. Демська-Кульчицька О. Деякі аспекти корпусної лінгвістики / О. Демська-Кульчицька // Українська мова : наук.-теорет. журн. – 2005. – № 1. – С. 44–51.
2. Жуковська В. В. Корпусна лінгвістика: історична перспектива та сучасний стан / В. В. Жуковська // Ключови въпроси в съвременната наука : материали XVIII Междунар. науч. практ. конф. – Т. 18. Филологични науки. – София : ООД «Бял ГРАД-БГ», 2012. – 72 с.
3. Захаров В. П. Корпусная лингвистика : учеб.-метод. пособие / В. П. Захаров. – СПб. : [б. и.], 2005. – 48 с.
4. Кутузов А. Б. Корпусная лингвистика [Электронный ресурс] / А. Б. Кутузов // Курс «Корпусная лингвистика» [сайт к-ры перевода и переводоведения Тюмен. гос. ун-та]. – 2003. – Режим доступа : http://tc.utmn.ru/files/corpus_1.pdf
5. Плунгян В. А. Корпус как инструмент и как идеология: о некотрых уроках современной корпусной лингвистики / В. А. Плунгян // Русский язык в научном освещении. – 2008. – № 2 (16). – С. 7–20.
6. Селіванова О. О. Корпусна лінгвістика / О. О. Селіванова // Сучасна лінгвістика: напрями та проблеми : підручник. – Полтава : Довкілля–К, 2008. – 712 с.
7. Шаров С. А. Представительный корпус русского языка в контексте мирового опыта / С. А. Шаров // НТИ. – Сер. 2. – 2003. – № 6. – С. 9–17.
8. Chomsky N. The Acquisition of Language / N. Chomsky. – New York, 1964. – 187 p.
9. Lьdeling A. Corpus Linguistics: An International Handbook / A. Lьdeling, M. Kytц. – Walter de Gruyter & Co. – 578 p.
10. McEnery T. Corpus Linguistics / T. McEnery, A. Wilson. – Edinburgh : Edinburgh University Press, 2001. – 235 p.

Войтко Галина. Корпусная лингвистика: история формирования и перспективы развития. В статье проанализированы основные этапы развития и становления корпусной лингвистики как составляющей прикладной лингвистики, поскольку создание и использование корпусов открывает новые возможности и перспективы для анализа языкового материала и использования корпусов для учебных целей. Рассмотрены определение понятия «корпусная лингвистика», «корпус». Охарактеризованы два этапа развития корпусной лингвистики: протокорпусний и корпусной. Рассмотрены основные корпуса текстов и их характеристики, влияние электронных корпусов на дальнейшее развитие лингвистических исследований. Рассмотрев развитие корпусной лингвистики и существующие корпуса текстов, можно сказать, что сегодня корпус предстает неотъемлемой частью лингвистических исследований, поскольку позволяет исследовать большие массивы информации для получения достоверных данных и использовать корпус в учебных целях. На современном этапе применения электронного корпуса текстов стало неотъемлемой составляющей многих отраслей языкознания и смежных дисциплин.

Ключевые слова: корпусная лингвистика, корпус, порождающая грамматика, разметка, конкорданс.

Voitko Halyna. Corpus Linguistics: the History of Formation and Development Prospects. The paper analyzes the main stages of development and establishment of corpus linguistics as the integral part of applied linguistics, since the creation and usage of corpus opens up new possibilities and perspectives for the analysis of linguistic material and

the use of corpus for educational purposes. Definitions of such terms as corpus linguistics, corpus are considered. The article outlines main problems and goals of corpus linguistics. The characteristics of corpora and the impact of the emergence of electronic corpus for further development of linguistic research are described. Special attention is given to the description of corpora that were created beginning with corpus linguistics establishment. Today electronic corpora are an integral part of linguistic research, as allow exploring large amounts of information to obtain accurate data and use corpora for educational purposes. At the present stage of linguistic research corpora have become an integral part of many branches of linguistics and related disciplines.

Key words: corpus linguistics, corpus, generative grammar, annotation, concordance.

Стаття надійшла до редколегії
04.04.2014 р.

УДК 811.111’374.822

Євгенія Гороть,
Леся Малімон

Лексикографічна концепція «Англо-українського словника для англомовних студентів-іноземців»

У статті здійснено аналіз принципів укладання навчального «Англо-українського словника для англомовних студентів-іноземців». Проаналізовано семантичні парадигми слів і принципи їх лексикографічної репрезентації. Обгрунтовано необхідність урахування традиційних лінгвістичних та лексикографічних підходів національного мовознавства до граматичного опису слова й одночасно акцентовано на особливостях навчальної лексикографії, яка ґрунтується на неспоріднених мовах. У таких випадках лінгвістична компетентність лексикографа поєднується з дидактичною доцільністю й ефективністю навчання, які мотивуються рівнем знання мови та ментальністю користувачів словника. Вищезазначені принципи визначили критерії відбору реєстрових слів. Основний принцип відбору – частотність цих слів у мовленні. Проаналізовано особливості лексикографічного опису іменника, дієслова й прикметника. Такий підхід лексикографічного опису одиниць мови уможливить укладення навчального «Англо-українського словника для англомовних студентів-іноземців», які починають вивчати українську мову як іноземну.

Ключові слова: словник, функції словника, призначення словника, навчальний словник, авторська лексикографічна модель словника.

Постановка наукової проблеми та її значення. Сучасна міжкультурна комунікація вимагає адекватних часові джерел інформації та способів її передачі. Традиційно інформаційні потоки реалізуються в мові, оскільки саме слово як онтологічна константа цивілізації дає змогу пізнати людину й вербалізовий простір різних народів. Цьому сприяє, передусім, створення словників, які акумулюють духовний досвід і знання.

Аналіз досліджуваної проблеми. Процес укладання словників – один із найстаріших видів філологічної діяльності людства, проте вдосконалення змісту, структури, механізмів укладання лексикографічних праць триває, оскільки кожне покоління користувачів висуває свої вимоги до укладання словників, щоб пристосувати їх до своїх умов життя та спілкування. Мова, як відомо – дієвий засіб міжособистісної, міждержавної, політичної й ділової комунікації, а тому проблема поєднання теоретичних висновків науковців із практикою укладання лексикографічних довідників із метою вдосконалення їх внутрішньої будови не втрачає своєї важливості.

Слово, відображаючи одиничне явище дійсності, є багатогранним феноменом, його можна розглядати з різних боків. Опис слова в словниковій статті індивідуалізований, специфічний, залежний від певних факторів, головні з яких – предметна, адресна й дескриптивна орієнтація словника, його призначення та функція. Вибираючи аспекти слова, що будуть представлені в словнику, а також стиль їх репрезентації, передусім мають бути визначені *тип*, *функції* й *призначення* словника.

На сьогодні в Україні видано велику кількість двомовних словників, що засвідчує активність вітчизняних мовознавців-лексикографів, зорієнтованих на активізацію міжнародних відносин. Біль-