МОВОЗНАВСТВО

АКТУАЛЬНІ ПРОБЛЕМИ ЛЕКСИКОЛОГІЇ ТА ЛЕКСИКОГРАФІЇ

УДК 811.112'2: 81'33

КОРПУСНА ЛІНГВІСТИКА ТА ГЕРМАНІСТИКА: ТЕОРЕТИЧНІ ЗАСАДИ І ПЕРСПЕКТИВИ

Ковбасюк Л. А.

Стаття присвячена корпусній лінгвістиці та її місцю у сучасній германістиці. Визначено поняття "корпусна лінгвістика", "корпус". Розглянуто теоретичні засади корпусної лінгвістики. Висвітлено найвагоміші корпуси текстів німецької мови та їхні характерні ознаки. Встановлено перспективи корпусних лінгвістичних досліджень у германістиці.
Ключові слова: німецька мова, корпусна лінгвістика, корпуси текстів, перспектива

Статья посвящена корпусной лингвистике и её месту в современной германистике. Определены понятия "корпусная лингвистика", "корпус". Показаны теоретические основы корпусной лингвистики. Представлены самые важные корпусы текстов немецкого языка и их характерные черты. Установлены перспективы корпусных исследований в германистике.
Ключевые слова: немецкий язык, корпусная лингвистика, корпусы текстов, перспектива

The article deals with the corpus linguistics and her place in modern German Studies. The terms "corpus linguistics" and "corpus" are defined. The theoretical background of the corpus linguistics is determined. The most important text corpora of German are described and their features are analyzed. Perspectives on corpus linguistics in German Studies are systematized.
Key words: German, corpus linguistics, text corpora, perspectives

Розвиток сучасної лінгвістичної думки характеризується тісним поєднанням з новітніми досягненнями соціології, психології, когнітивістики тощо та утворенням нових напрямів, що є прикметою нової наукової парадигми сьогодення. Так, синтез мовознавства та кібернетики сприяв появі комп'ютерної лінгвістики, а згодом – і корпусної лінгвістики, підґрунтям якої є певний корпус, великий за обсягом мовний матеріал, відібраний із різних рівнів та функціональних стилів будь-якої мови і зведений у велику комп'ютеризовану систему завдяки відповідному програмному забезпеченню.

Корпусна лінгвістика, як нова лінгвістична галузь, привертає до себе увагу багатьох вітчизняних і зарубіжних мовознавців [1; 2; 4; 5; 11; 15], оскільки залучення великого за обсягом матеріалу для багатоаспектного аналізу сприяє створенню реальної картини функціонування усіх одиниць мови у мовленні. Створення та обробка корпусів німецькомовних текстів у електронному вигляді, упровадження корпусного підходу для лінгвістичного аналізу різного типа текстів, певна низка недостатньо визначених проблем, напр., багатомовність корпусу, загальна доступність, структурне анотування й метарозмітка та ін. визначають **актуальність** та **новизну** обраної нами теми.

Мета даної статті полягає у висвітленні теоретичних засад корпусної лінгвістики, її місця та її досягнень у сучасній германістиці та у визначенні перспективних напрямів розвитку корпусної лінгвістики.

Досягнення поставленої мети передбачає виконання наступних **завдань**: 1) дати визначення термінам "корпусна лінгвістика" та "корпус", 2) окреслити

історичний ракурс формування корпусної лінгвістики, 3) висвітлити теоретичні засади корпусної лінгвістики, 4) проаналізувати найвідоміші корпуси текстів німецької мови, 5) встановити перспективні напрями досліджень у царині корпусної лінгвістики.

Корпусна лінгвістика виокремлюється у сучасній германістиці як прикладна наукова дисципліна, що описує реальні мовні явища, їхні елементи та структуру на основі аналізу автентичних письмових або усних текстів, зібраних у лінгвістичні текстові корпуси. Корпус визначається як масив, зібрання правильно анотованих та структурованих письмових або усних висловлювань у електронному вигляді, що створюються для вирішення певних лінгвістичних завдань [12, с. 13–14]. Корпусна лінгвістика вивчає мову як соціальне явище, що описується заснованими на досвіді даними, тобто в певному мовленнєвому акті. Саме в текстах мова проявляє себе як соціальний феномен, оскільки саме в них її можна записати, описати й проаналізувати. Більшість текстів зустрічаються у вигляді мовленнєвих актів, тобто як соціальна взаємодія між членами певного мовного суспільства [15, с. 112].

Корпусний аналіз визначається низкою ознак: 1) емпіричним підходом до аналізу мовних даних; 2) використанням великих за обсягом, структурованих колекцій природних текстів (корпусів) як основи для аналізу; 3) широким залученням комп'ютерних технологій для дослідження мовного матеріалу; 4) застосуванням кваліативних і квантитативних аналітичних методик, з суттєвою перевагою останніх (вивчення частоти вживання лінгвістичних одиниць, статистичні дослідження сполучуваності і та ін.) [1, с. 9].

Корпусній лінгвістиці властива певна двовекторність, оскільки вона вивчає як теорію та практику створення корпусів, так і самі мовні корпуси. Двовекторність корпусної лінгвістики обумовлена подвійною природою об'єкта її дослідження – текстового корпусу, який, з одного боку, є вихідним мовленнєвим матеріалом для корпусної лінгвістики, а з іншого, є результатом діяльності цього мовознавчого напряму. Предметом корпусної лінгвістики виступають теоретичні основи і практичні механізми створення та експлуатації мовних корпусів [2, с. 10].

Процес корпусного аналізу включає три кроки: 1) ідентифікацію мовних даних за допомогою категоріального аналізу, 2) співвідношення мовних даних з урахуванням статистичних методів та 3) інтелектуальну інтерпретацію результатів дослідником. Перші два кроки є найбільш автоматизованими, останній крок вимагає людської розумової діяльності, оскільки будь-яка інтерпретація є актом залучення розумових здібностей науковця, а тому не перетворюється в алгоритмічну процедуру. Саме у цьому проявляється головна відмінність між корпусною і комп'ютерною лінгвістикою, що зводить мову до набору процедур [15, с. 113].

Типологія текстових корпусів, згідно з Л. Лемницером та Г. Цінзмейстер [12, с. 137–142], ґрунтується на наступних критеріях: 1) функціональність, тобто йдеться про мету створення певний корпус (для проведення досліджень або для надання ілюстративного матеріалу), 2) вибір мови (одномовні, двомовні або багатомовні текстові корпуси, 3) тип даних (Medium), тобто аналізуються тексти писемні, усні чи мультимодальні (напр., відеокорпус), 4) анотація (розмітка), її відсутність чи наявність, а також її тип (більшість анотацій належить до корпусів морфологічного або синтаксичного типу), 5) розмір (великі, маленькі, національні, спеціалізовані корпуси тощо), 6) спосіб існування (Persistenz): динамічний або статичний корпус та 7) доступність (вільний доступ, частковий доступ, необхідна реєстрація, комерційні корпуси).

Найважливішим критерієм, що вирізняє лінгвістичний корпус з-поміж інших інформаційних систем є анотація (розмітка), що полягає в приписуванні тексту певних екстралінгвістичних, структурних та власне лінгвістичних міток. Лінгвістична анотація поділяється в свою чергу на: 1) морфологічну, 2) синтаксичну, 3) семантичну, 4) анафоричну та 5) просодичну [1, с. 76–82].

Одним із основних підходів до аналізу мовних даних у корпусній лінгвістиці є конкорданс, тобто спеціалізована лінгвістична прикладна програма, за допомогою якої здійснюється автоматична вибірка заданих мовних одиниць з електронних текстів, проводиться дослідження корпусу за обраним словом, словосполученням чи фразою. Конкорданс надає інформацію про частотність вживання і сполучення тієї або іншої мовної одиниці, а також звертається до певного тексту, в якому був знайдений приклад, а також демонструє слова, словосполучення або фрази в центрі комп'ютерного екрану, разом зі словами, що знаходяться поруч [12, с. 171,

197]. Крім того, отримати необхідну інформацію з текстового корпусу можна завдяки корпусному менеджеру, тобто спеціальної пошукової системи, що "включає програмні засоби для пошуку даних у корпусі, отримання статистичної інформації й надання результатів користувачеві в зручній формі. Результати цієї процедури подаються у вигляді горизонтальних рядків із пошуковим словом посередині. Ця процедура має назву KWIC (Key Word In Context)" [1, с. 93].

Корпусна лінгвістика почала своє становлення як нова галузь сучасного мовознавства у зв'язку із швидким розвитком комп'ютерних технологій та широким застосуванням комп'ютерів у житті людини, тобто на початку 1960 рр. Перші вагомі дослідження у площині корпусної лінгвістики було проведено на матеріалі англійської мови у 1963 р. у Браунівському університеті (США) У. Френсисом і Г. Кучерою. Цей корпус, що містив 500 текстових уривків загальним обсягом 1 мільйон слів, було створено для дослідження лінгвістичних особливостей американського варіанта англійської мови. Ясність, чіткість і наочність Браунівського корпусу сприяли його швидкій популярності та використанню як певного еталону для створення інших корпусів текстів [3; 12, с. 40]. Пізніше з'явився британський аналог Браунівського корпусу – Ланкастерсько-Осло-Бергенський корпус (Lancaster-Oslo-Bergen), укладений на матеріалі британської масової друкованої продукції 1961 року видання. Анотована версія корпусу з'явилася у 1985 році. Створення зазначених корпусів уможливило різноаспектні лінгвістичні порівняння двох варіантів англійської мови (американського й британського) із залученням текстів різних жанрів, доступних комп'ютерній обробці. У 80-ті роки XX ст. у зв'язку із удосконаленням комп'ютерних технологій, здатних обробляти великі масиви текстів, було створено Британський Національний Корпус (British National Corpus) та Банк Англійської мови (Bank of English) [1, с. 40–41].

У 1992 році була створена організація "Європейська корпусна ініціатива" [10], яка опрацювала близько 50 корпусів текстів об'ємом від 12 тисяч до 5 мільйонів слів кожен на європейських мовах загальним обсягом 98 мільйонів слів. Ці корпуси містять так звані "паралельні" тексти або бітексти (текст однією мовою разом із перекладом іншою мовою), що можуть використовуватись як довідник для пошуку потрібних словосполучень.

Зауважимо, що термін "корпусна лінгвістика" ввійшов до наукового вжитку наприкінці XX ст. завдяки публікації у 1984 році збірника наукових праць "Corpus Linguisitcs: Recent Developments in the Use of Computer Corpora in English Language Research" за матеріалами конференції ICAME "Conference on the Use of Computer Corpora in English Language Research" у м. Неймеген (Голландія) у 1983 р. [6].

Розвиток корпусної лінгвістики у Німеччини розпочався наприкінці XX ст. Підґрунтям сучасної німецької корпусної лінгвістики вважаються текстові зібрання Інституту німецької мови в м. Маннгейм (IDS). Цей академічний корпус має назву "Das deutsche Referenzkorpus" (**DeReKO**) [8], перша частина якого була зібрана завдяки плідній співпраці Інституту німецької мови, Інституту машинної обробки мови університету м. Штуттгарт (IMS) та Мовознавчого семінару університету м. Тюбінген протягом 1999-2002 рр. Метою укладання цього корпусу була найповніша репрезентація сучасної німецької мови з 1956 по 2001 рр. Починаючи з 200 року DeReKO постійно поповнюється новою інформацією. На даний час (станом на 21.03.2016 р.) DeReKO є найбільшим національним лінгвістичним електронним корпусом німецьких писемних текстів та містить у собі 29 мільярдів слів (відповідно 70 мільйонів сторінок словника), поданих у відповідному лінгвальному контексті, що закодований у певному текстовому форматі та повністю відповідає оригіналу. До складу цього корпусу входять, напр., 4 корпуси писемних текстів (W1-W4), корпуси статей Вікіпедії німецькою та англійською мовою, Маннгеймські корпуси історичних газет та журналів та ін. Проаналізовані тексти належать до художньої літератури (твори вибраних письменників XX-XXI ст.), преси та публіцистики сучасної Німеччини, Австрії та Швейцарії (напр., Burgenländische Volkszeitung, die Zeit, Mannheimer Morgen та ін.), значну частину становлять також усні тексти (Archiv für gesprochens Deutsch). Усі приклади у текстових корпусах поділяються за відповідною тематикою (Fiktion, Natur-Umwelt, Wissenschaft та ін.). Корпус містить загальну інформацію про текст (джерело походження, тип тексту та ін.), крім того, усі тексти є морфологічно та синтаксично анотованими, тобто можна шукати певну форму слова, або всі форми

однієї лексеми за початковими літерами та словосполучення. Щоб мати змогу вільно користуватися даними корпусами, необхідно зареєструватися, підписати ліцензійні умови використання (напр., для корпусів текстів Mannheimer Korpus I, II; Bonner Zeitungskorpus і т.д.) та встановити спеціальну програму COSMAS II.

Наступним за вагомістю, важливістю для мовознавчих пошуків та обсягом матеріалу вважається базовий корпус текстів Берлінсько-Бранденбурзької Академія наук "Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart" (**DWDS**) [7], активна розробка якого проходила з 2000 по 2003 рр. Увесь національний академічний корпус містить приблизно 9,2 мільярдів слів, серед яких 1,5 мільярдів знаходяться у вільному доступі, та складається з текстів художньої літератури (28,42%), преси (27,36%), наукових текстів (23,15%) та так званих текстів для ужитку (Gebrauchsliteratur), тобто реклами, підручників, текстів пісень тощо (21,05%). DWDS поділяється на базові корпуси XX ст. й XXI ст. (**DWDS-Korpus, DWDS-Korpus 21**), текстовий корпус (**Deutsches Textarchiv**), що охоплює період з 1600 по 1900 рр. та 3 корпуси преси та публіцистики (Berliner Zeitung, Der Tagesspiegel, Die Zeit). Зазначимо, що корпус усних текстів (**Korpus Gesprochene Sprache**) є окремим складовим компонентом корпусу DWDS, що містить транскрипти промов, доповідей, бесід та інтерв'ю за XX ст. Крім того, структурним компонентом DWDS є так звані спеціальні корпуси, а саме: 1) корпус блогів, 2) корпус титрів до фільмів, 3) корпус текстів Політехнічного журналу (Polytechnisches Journal), 4) корпус текстів Німецької демократичної республіки (DDR). У кожному корпусі подається назва тексту, його автор, дата виходу у світ, видавництво та інформація про тип тексту. Тексти є морфологічно анотованими, тобто можна шукати як певну форму слова, так і всі форми однієї лексеми, а також словосполучення. Для кожного слова надається інформація щодо частотності вживання у кожному структурному підрозділі DWDS. Для вільного безкоштовного користування корпусом необхідно зареєструватися.

Національний академічний текстовий корпус **"Deutscher Wortschatz"** [17] університету м. Лейпциг збирається та аналізується починаючи з 1998 р. та містить більше 500 мільйонів слів, що представлені у 35 мільйонах речень. Крім того, цей текстовий корпус містить 230 корпусів на 200 мовах світу, зміст яких постійно поповнюється. Йдеться про так звані корпуси Лейпцигської колекції (Leipzig Corpora Collection (**LCC**)). Слід зауважити, що всі тексти збираються виключно з онлайн-газет/онлайн-журналів, вебсторінок різного ґатунку та Вікіпедії. Тексти розподілено за певними тематичними групами згідно з класифікацією Ф. Дорнзейфа та є морфологічно та анафорично анотованими, оскільки надають інформацію про референтні зв'язки слів у словосполученнях та реченнях. Зауважимо, що крім інформації, що присутня у більшості текстових корпусах, "Deutscher Wortschatz" надає інформацію щодо класу частотності слова в німецькій мові, що обчислюється за законом Ціпфа (напр., найвищий клас частотності має артикль der) та графічне зображення сполучуваності слова. Для вільного користування корпусом необхідно зареєструватися та підписати ліцензійні умови використання.

Оскільки корпусна лінгвістика є достатньо молодою галуззю мовознавства, їй властиві певні перспективні напрями досліджень, що потребують особливої уваги. Проаналізувавши наукові доробки мовознавців Німеччини, ми дійшли висновку, що до найперспективніших напрямів корпусної лінгвістики, на нашу думку, належать наступні:

1) укладання текстових корпусів інтернетопосередкованої комунікації, оскільки сучасна людина проводить мінімум 3 години на день у Всесвітній мережі Інтернет. Йдеться, напр., про проект "Deutsches Referenzkorpus zur internetbasierten Kommunikation" (**DeRiK**), розробкою якого займаються лінгвісти Інституту німецької мови та літератури Технічного університету м. Дортмунд, університету м. Маннгейм та Берлінсько-Бранденбурзької Академії наук. Завдяки цьому проекту створені текстовий корпус форуму БМВ (**das BMW-Forum-Korpus**), що поєднує один мільйон текстів із різних онлайн-форумів, корпус Дортмундського чату (**das Dortmunder Chat-Korpus**), який пропонує 478 чат-спілкувань обсягом 1,06 мільйона слів та ін. [9]. DeRiK містить тексти лише із різного роду чатів, блогів, онлайн-форумів, твіттера, скайпа тощо та є інтегрованою частиною DWDS [11, с. 151];

2) зібрання та опрацювання корпусів текстів для досліджень у царині історичного мовознавства. Виокремимо, напр., корпус "Deutsch.Diachron.Digital" (**DDD**),

над яким працюють в університетах Берліна, Франкфурта та Єни, анотуючи тексти з 750 по 1050 рр. (загальна кількість опрацьованих слів – 650 тис.). Йдеться про староверхньонімецькі, старосаксонські тексти та тексти, написані латинською мовою. Написання, рубрикація, виділення курсивом та ін. зазначених текстів залишаються без змін та коментуються під час лінгвістичної розмітки [13];

3) створення німецькомовних текстових корпусів мережі Інтернет (**Webkorpora**). Йдеться, напр., про **deWaC-Webkorpus** Інституту когнітивістики університету м. Оснабрюк**,** в якому налічуються 1,5 мільярди слів із німецьких Інтернет-сторінок, заархівованих у 2005 році [9];

4) формування текстових корпусів смс-комунікації, оскільки як дорослі, так і діти достатньо багато часу спілкуються у повсякденному житті за допомогою смс-повідомлень. Як приклад можна навести корпус смс-повідомлень учнів та студентів віком від 12 до 30 років із Оснабрюка та Ганновера [14], що є у вільному доступі та містить 1500 смс-повідомлень;

5) використання корпусів у процесі навчання німецької мови як іноземної. Будь-який текстовий корпус ґрунтується на автентичних текстах різного ґатунку, тому вживання корпусів на заняттях з німецької мови має великий потенціал, оскільки сприятиме опануванню мовних структур німецької мови у певному комунікативному контексті (напр., за допомогою DWDS можна поповнювати словниковий запас, вдосконалювати граматичні навички тощо), проведенню лінгвістичних досліджень при написанні різного виду наукових робіт з лексикології, стилістики тексту та ін. За словами Ф. Валлнер [16], робота з текстовими корпусами вимагає високого рівня володіння мовою, найбільш вдалим є використання конкордансів, за допомогою яких вивчаються словотворчі моделі, полісемія слова, його сполучуваність та ін.

Отже, проведене дослідження дало нам змогу встановити шлях становлення корпусної лінгвістики у германістиці, визначити її основні теоретичні засади, висвітлити найвагоміші німецькомовні корпуси текстів та окреслити перспективні напрями наукового пошуку. Зауважимо, наша стаття не претендує на повноту та вичерпність опису досягнень німецької корпусної лінгвістики. Перспективними для подальших досліджень вбачаємо контрастивний аспект корпусів текстів німецької та української мов, застосування корпусів у процесі дискурсного аналізу, під час викладання німецької мови у вищих навчальних закладах тощо.

Література

1. Жуковська В. В. Вступ до корпусної лінгвістики. Навчальний посібник. / В. В. Жуковська– Житомир: Вид-во ЖДУ ім. І. Франка, 2013. – 142 с.

2. Захаров В. П. Корпусная лингвистика: учебник для студентов гуманитарных вузов/ В. П. Захаров, С. Ю. Богданова. – Иркутск, ИГЛУ, 2011. – 161 с.

3. Френсис У. Н. Проблемы формирования и машинного представления большого корпуса текстов / У. Н. Френсис // Новое в лингвистике. – 1983. – Вып. XIV. – С. 334–352.

4. Beißwenger M. Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation als Zusatzkomponente für die Korpora im Projekt "Digitales Wörterbuch der deutschen Sprache" (DWDS)/M. Beißwenger, L. Lemnitzer //Journal for Language Technology and Computational Linguistics. – 2013. – 26 (2). – S. 1–22.

5. Bubenhofer N. Sprachgebrauchsmuster: Korpuslinguistik als Methode der Diskurs- und Kulturanalyse/ N. Bubenhofer. – Berlin: de Gruyter, 2009. – 388.

6. Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research / [ed. by J. Aarts and W. Meijs]. – Amsterdam: Rodopi, 1984 – 229 p.

7. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart [Die elektronische Ressource]. – Verfügbar über: http://www.dwds.de.

8. DeReKo. Das Deutsche Referenzkorpus [Die elektronische Ressource]. – Verfügbar über :
http://www1.ids-mannheim.de/kl/projekte/korpora/.

9. Empirische Erforschung internetbasierter Kommunikation [Die elektronische Ressource]. – Verfügbar über:
http://www.empirikom.net/Ressourcen/WebHome

10. European Corpus Initiative [Die elektronische Ressource]. – Verfügbar über:
http://www.elsnet.org/eci.html .

11. Köhler R. Korpuslinguistik – zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven / R. Köhler //Zeitschrift für Computerlinguistik und Sprachtechnologie. – 2005. – Band 20. Heft 2. – S. 2–16.
12. Lemnitzer L. Korpuslinguistik: eine Einführung/ L. Lemnitzer, H. Zinsmeister. – Tübingen: Gunter Narr, 2006. – 220 S.
13. Referenzkorpus Altdeutsch [Die elektronische Ressource]. – Verfügbar über: http://www.deutschdiachrondigital.de/ .
14. SMS-Korpus [Die elektronische Ressource]. – Verfügbar über: http://www.mediensprache.net/archiv/corpora/sms_os_h.pdf .
15. Teubert W. Corpus linguistics and lexicography / W. Teubert // Text Corpora and Multilingual Lexicography. – Amsterdam/ Philadelphia: John Benjamins Publishing Company, 2007 – P. 109–134.
16. Wallner F. Lehren und Lernen mit Korpora im DaF-Unterricht [Die elektronische Ressource]. – Verfügbar über: https://www.goethe.de/de/spr/mag/20454877.html .
17. Wortschatz Universität Leipzig [Die elektronische Ressource]. – Verfügbar über: www.wortschatz.uni-leipzig.de .