УДК 811.112.2'373.7

О. М. Гордій,
Прикарпатський національний університет, м. Івано-Франківськ

КОРПУСНІ ЛІНГВІСТИЧНІ ДОСЛІДЖЕННЯ І ФРАЗЕОЛОГІЯ

У статті висвітлюються можливості корпусної лінгвістики як методу емпіризації сучасних лінгвістичних досліджень, визначаються його переваги та окреслюються межі застосування для фразеологічних студій і розглядається потенціал Інтернету як гігантського «живого» лінгвістичного корпусу.

Ключові слова: *корпусна лінгвістика, мовний корпус, компілювати, емпірична верифікація, фразеологічні одиниці, Інтернет, пошукова машина.*

В статье освещаются возможности корпусной лингвистики как метода эмпиризации современных лингвистических исследований, определяются его преимущества и границы применения для исследований фразеологии и рассматривается потенциал Интернета как гигантского «живого» лингвистического корпуса.

Ключевые слова: *корпусная лингвистика, языковый корпус, компилировать, эмпирическая верификация, фразеологические единицы, Интернет, поисковая машина.*

The article deals with the capability of Corpus linguistics as a method of the empirization of the contemporary linguistic research, determines the advantages and outlines the limits of use of the corpora method in the phraseological studies and considers the potential of the Internet as a gigantic «live» linguistic corpus.

Keywords: *Corpus linguistics, linguistic corpus, to compile, empirical verification, idioms / set phrases, Internet, search engine.*

В останні роки започатковано багато досліджень, які, з одного боку, переносять випробувану методику інших дисциплін на вивчення фразеології та водночас долають межі традиційної фразеології і намагаються дати їй нове визначення. Прикладом можуть слугувати корпусні лінгвістичні дослідження. Корпусна лінгвістика, яка взяла свій початок у 60-их роках минулого століття на англомовному просторі і з того часу стрімко розвивається, є для мовознавців як викликом, так і великим шансом. Вона ґрунтується на мовних корпусах, тобто великих зібраннях текстів, які слугують базою даних для опису і пояснення мовних феноменів. Корпусна лінгвістика є емпіричною, адже дослідження опираються на реальні мовні дані: її предметом є справжні, а не створені власне з лінгвістичною метою (часто з опорою лише на інтуїцію лінгвіста), тексти, які є частиною природної мови та комунікації, що формуються у корпус за чітко розробленими принципами і критеріями. За результатами обробки цих даних корпусна лінгвістика висуває загальні та специфічні твердження про дискурс.

Хоча сучасна корпусна лінгвістика виникла, як зазначено вище, в середині минулого століття, історія збору мовних даних є набагато старшою. Проте до появи електронних текстів цей процес був надзвичайно складним та потребував значних затрат часу, і при цьому часто не забезпечувалась статистична релевантність. О.Я. Остапович зазначає, що за добором емпіричних відомостей «вручну» закріпилась стереотипна репутація суто технічної, примітивної, часто навіть теоретично безглуздої роботи і нерідко щодо таких «емпірично невтомних» дослідників можна було зустріти зневажливо-поблажливе ставлення як до «фактологів», «збирачів цитат», а не до «вчених у високому розумінні». Проте наведені автором відомості про проведену емпіричну добірку такими відомими мовознавцями як В.В. Виноградов та В. Мідер дійсно гідні подиву [2]. Цікаві дані наводить також С. Дімер у статті «Das Internet als Korpus? : Aktuelle Fragen und Methoden der Korpuslinguistik» [5]. Так близько 1900 р. група німецьких лінгвістів, серед яких були F. Grimm і E. Gasner, збирали і вручну квантифікували мовні дані щодо історії розвитку англійських часток – опрацьовували здебільшого одну частку на одне дисертаційне дослідження. Трудомісткою була також підбірка цитат для Old English Dictionary – протягом десятиліть кореспонденти виписували вручну і передавали у редакційну колегію уривки текстів. Проте з часом зросла потреба у емпіричних мовних даних, які підлягають квантифікації, і в 60-их роках Г. Кучера і Н. Френсіс компілювали Brown-корпус американського варіанту англійської мови, який охоплював один мільйон слів і довший час залишався стандартним джерелом корпусних досліджень. У 70-их роках робоча група на чолі з С. Йохансоном опублікувала відповідний корпус також і британського варіанту англійської – London-Oslo-Bergen (LOB). Наступні десятиліття характеризувалися бурхливим розвитком інформаційних технологій та комп'ютерної техніки, що позначилось також на поширенні та вдосконаленні методів корпусного аналізу. Однак описана вище стереотипна репутація «ненаукового підходу», яка існувала раніше щодо збирачів цитат, певний час поширювалась і на дослідників власне корпусної лінгвістики. О.Я. Остапович, посилаючись на Дж. Семпсона, влучно зауважує, що у мовознавстві, на відміну від природничих наук, ще потрібно було комусь доводити важливість емпіричної верифікації абстрактних гіпотез і теорій [2]. Перші корпуси часто визначались теоретичною лінгвістикою як непотрібні, корпусна лінгвістика довший час вважалась виключно допоміжною наукою, постачальником доказів для лінгвістичних теорій.

В останні два десятиліття з неймовірним технічним прогресом змінилась і корпусна робота. Сучасні лінгвістичні корпуси являють собою велетенські електронні текстові масиви баз даних, обладнані пошуковими машинами з контекстуальними, деривaційними та колокаційними, коокурентними критеріями, що суттєво полегшують дослідникам збір і аналіз емпіричного мовного матеріалу; сьогодні корпусні лінгвісти активно працюють в сфері теоретичних досліджень і до зібрання емпіричного матеріалу подають також інтерпретацію отриманих результатів. Ці корпуси є більш репрезентативними, оскільки включають друковані тексти із максимально широким жанрово-стильовим розмаїттям та завдяки зростанню Всесвітньої мережі доступні тепер в режимі онлайн зареєстрованим користувачам. Крім того, до сучасних корпусів включаються також

аудіофайли із записами усного мовлення (радіо- і телепередачі, політичні промови тощо). Загальновизнаним зразком, на який орієнтуються багато інших сучасних корпусів, є Британський національний корпус (British National Corpus – BNC) із 100 мільйонами слів різножанрових текстів, включаючи усне мовлення. Створений М. Девісом Corpus of Contemporary American English (COCA) охоплює 450 мільйонів слововживань та є єдиним публічно доступним корпусом американського варіанту англійської мови. Для студій з історії мови укладений діахронний Гельсінський корпус (понад 1,5 млн. слововживань). Також були компільовані лінгвістичні корпуси і інших національних мов. Започаткований ще у 80-их роках минулого століття проект «Машинный фонд русского языка» пізніше використовувався для створення Національного корпусу російської мови. Корпус української мови Лабораторії комп'ютерної лінгвістики Київського національного університету ім. Т. Шевченка на сьогодні, на жаль, значно поступається йому за обсягом. Серед найвагоміших корпусів німецької мови слід назвати передусім Cosmas-корпус Інституту німецької мови у Маннгеймі, який налічує майже 2 мільярди слів.

Загалом перевагою емпіричних корпусних даних є те, що вони роблять можливим такий підхід до мовного матеріалу, який «не передбачає – наскільки це можливо – попередні припущення щодо їх характеристик» [13, 64].

Оскільки у корпусних дослідженнях на передньому плані знаходиться вживання мовних одиниць, то вони неодмінно розглядаються як контекстуалізовані феномени, тобто зафіксовані у їхньому контексті, який, так би мовити, «постачається разом з ними», і абстрагування від контексту не завжди легке провести. Згідно з М. Стаббсом (M. Stubbs), корпусна лінгвістика, таким чином, є іманентно соціолінгвістичною, оскільки тексти, з яких складається той чи інший корпус, є справжніми комунікативними актами, які у дискурсивній спільноті служать чи слугували цілям комунікації [14, с. 221].

Корпусна лінгвістика – іманентно квантитативна. Одним з її основних положень є важливість частотності і повторюваних явищ. Це відмежовує корпусний лінгвістичний аналіз від попередніх теоретичних підходів, особливо генеративних. Першочерговою метою корпусної лінгвістики повинно бути дослідження «нормальних», тобто узуальних явищ [14, с. 221]. Одноразові і ідіосинкратичні явища слід реєструвати, але вони можуть розглядатися і описуватися лише на фоні того, що є звичним і очікуваним.

Отже, корпусно-лінгвістичні дослідження є емпіричними і індуктивними, ґрунтуються на автентичних мовних даних природної комунікації і відображають використання мови великої кількості членів мовної спільноти.

Емпіричний комп'ютерний аналіз текстових масивів застосовується сьогодні у багатьох галузях сучасного мовознавства, проте найбільше поширення методи корпусної лінгвістики отримали у працях з лексикографії та лексичної семантики (J. Sinclair (1991) [11]; P. Hanks (1996) [8]; M. Stubbs (2001) [14]). З лексикологічними дослідженнями тісно пов'язане використання мовних корпусів у фразеології. Фразеологічні студії найбільш залежні від підбору автентичних прикладів, адже фразеологічний фонд змінюється досить швидко, і поширені сьогодні фразеологічні одиниці можуть на завтра вийти з ужитку. Традиційний збір емпіричних фразеологічних даних досить трудомісткий, і сучасна корпусна лінгвістика, яка робить можливим пошук в об'ємних цифрових корпусах, стала також і для фразеології незамінним методом. Застосування комп'ютерних методів для отримання емпіричного фразеологічного матеріалу стало логічним продовженням існуючої на той час тенденції у західній науковій школі. Як наслідок прагматичного повороту на зміну традиційним дослідженням, основною проблематикою яких були класифікаційні структурно-типологічні моделі фразеологізмів, концепції широкого і вузького розуміння фразеології, постали питання передусім різножанрового текстового функціонування фразеологічних одиниць (див. огляд п. 1.3.4.) і використання емпіричних мовних даних як основи досліджень фразеології стало тим часом звичним явищем. Також і перед вітчизняною школою, яка, за слушним твердженням О.Я. Остаповича, «надто довго не могла звільнитися від пут відверто схоластичної структурно-семантичної парадигми досліджень, запрограмованої ще в середині минулого століття» [2], відкриваються нові перспективи. Корпуси можуть слугувати об'ємними картотеками, в яких здійснюється пошук та досліджуються приклади певних мовних феноменів, а з іншого боку, статистичний аналіз даних дозволяє у великих корпусах простежити і піддати категоризації ідіоматичне карбування мови.

Вивченню семантики ідіом за допомогою корпусного методу присвячена стаття К. Статі «Korpusbasierte Analyse der Semantik von Idiomen» [12]. На прикладі ідіоми *ins Gras beißen* автор демонструє, що корпусний аналіз розглядає значення лексичних і фразеологічних одиниць як «значення у вжитку», тобто «значення як вживання». Фразеологічні одиниці відзначаються семантичною «гнучкістю» (Flexibilität), яка полегшує чи взагалі робить можливим їх включення у нові контексти. Структура значення ідіоми може розглядатися подібно до категорій в теорії прототипів: можна визначити «центральне значення» (Kernbeutung), яке є статистично значущим, а інші значення випливають з нього завдяки профілюванню чи «приглушенню» семантичних ознак. Звідси випливає, що корпусний метод виявляється затребуваним також у фразеографічній практиці: як для коректної семантизації фразеологічних одиниць, так і для статистичного визначення загальновживаних ФО, зокрема, окреслення т.зв. «фразеологічного / пареміологічного мінімуму» [6; 7]. Корпусний аналіз повинен використовуватися і у студіях з лінгвокультурології. О.Я. Остапович слушно стверджує, що зроблені на основі великих текстових корпусів висновки про переважання у мовному світогляді народу певних тематичних сфер, концептуальних метафор, аксіологічних установок тощо будуть значно вірогіднішими від тих, що почерпнуті із музейних експонатів словників чи фольклорних збірок [2]. Створені за спеціальними критеріями корпуси дозволяють також виявити гендерну специфіку мовного вжитку. Корпусний метод є незамінним і у дослідженнях з лінгвістичної варіантології, адже і тут некритична рецепція застарілих лексикографічних відомостей може призвести до спотворень їх реального функціонування. Відзначимо у цій сфері працю Г. Бікеля [3], який ще в кінці 90-их років, використовуючи як текстовий корпус Всесвітню мережу, розпочав

роботу над словником варіантів німецької мови (Ця робота завершена у 2004 р. [16]). Нижче ми власне розглянемо питання про переваги та обмеження застосування Інтернету як лінгвістичного корпусу.

Описані вище можливості корпусного методу очевидно відкривають перед лінгвістами великі перспективи. Проте не слід переоцінювати роль комп'ютеризованого машинного аналізу загалом для лінгвістики та для фразеології зокрема. Корпус здебільшого залишається лише вихідним джерелом, а комп'ютер – допоміжним інструментарієм, адже багато завдань (пошук стилістичних фігур чи відтінків значень тощо) ні сьогодні, ні, на наше переконання, в осяжному майбутньому не стануть підвладними штучному інтелекту. Також і машинна ідентифікація фразеологічних одиниць як і раніше залишається «міцним горішком» для комп'ютерної і корпусної лінгвістики. Особливо гостро постає так і не вирішене традиційною фразеологією питання про межі фразеології. Для того, щоб комп'ютер міг автоматизовано виявити ФО, слід так операціоналізувати феномен «фразеологізм», щоб його можна було побачити на мовній поверхні за допомогою чітких правил, а розроблені сьогодні різноманітні алгоритми охоплюють інколи більше, а інколи й менше мовних явищ, які традиційно визначаються як фразеологізми. Серед корпусних лінгвістів на даний час не існує єдино прийнятої думки, чи теоретичні концепції фразеології взагалі корисні для квантитативного аналізу. У багатьох сучасних роботах за основу береться широкий підхід до фразеології, такі дослідження «клішованого мовного вжитку» («musterhafter Sprachgebrauch») є цікавими у вирішенні лексикографічних чи дидактичних задач. Особливою складністю відзначається пошук модифікованих одиниць; загалом велику кількість роботи доводиться виконувати «вручну», комп'ютер може лише прискорити процес накопичення первинного матеріалу із різноманітних джерел.

З іншого боку, на даний час не існує ще надійних досконалих корпусів, величина яких достовірно відображала б реальне функціонування тієї чи іншої національної мови. До того ж сучасна мова постійно змінюється. Тому увага дослідників все частіше спрямовується на Інтернет, який дозволяє вивчати мову у великих масштабах і, так би мовити, «вживу».

Поряд із описаними вище систематично побудованими корпусами завдяки Інтернету існує велетенський архів з багатьма мільярдами сторінок, що органічно розростається, і який із застосуванням належних пошукових машин може розглядатися як корпус. Інтернет є корпусом, який щодня зростає і змінюється, мільйони Інтернет-користувачів формують його відповідно до своїх різноманітних потреб і інтересів.

Інтернет як засіб інформації і комунікації зазнав з середини 90-их років небаченого поширення. Розроблений в США у 1969 р. спочатку для наукових і військових цілей (проект «ARPANET» – Advanced Research Projekt Agency), Інтернет завдяки появі Глобальної мережі (WorldWideWeb) зміг стати доступним широкій громадськості і робить можливим електронний обмін інформацією і новинами. Використання Інтернету поступово призвело до утворення комунікативного простору, в якому з'являються нові форми глобального порозуміння, що характеризуються інтерактивним спрямуванням. Виникнення і використання новітніх форм комп'ютерно-опосередкованої комунікації відкриває перед лінгвістикою широке поле для дослідницької діяльності. Д. Крістал писав в 2004 році: «Лінгвістична оригінальність і новизна Інтернету повинні примусити наші серця битися швидше. Перед нами постає майбутнє, в якому наша комунікація буде радикально відрізнятися від комунікації минулого. (…) Захоплююче бути присутнім при цьому з самого початку» [4, с. 35]. Сьогодні вже очевидно, що мова потрапляє під активний вплив мережі Інтернет. За твердженням А. Є. Войскунського, Інтернет являє собою унікальний полігон, на якому розгортається випробування природної мови [1].

Оскільки побудова систематичного мовного корпусу, який давав би відомості щодо частотності лексичних чи фразеологічних одиниць, є надзвичайно трудомісткою і, відповідно, дуже дорогою справою, постає питання, чи може Інтернет як зібрання текстів, що за розміром перевищує усі інші корпуси, використовуватися як корпус для лінгвістичних досліджень?

Для багатьох людей – принаймні у віці до 40 років – увійшло вже у звичку звертатися на порталах Гугл, Вікіпедія, Яндекс та ін. із запитаннями на зразок «карта Мюнхена», «погода в Берліні», «розклад руху поїздів» тощо та керуватися отриманими відповідями у житті. Ця тенденція не пройшла повз лінгвістику. Веб є «скарбничкою» доступних у електронному вигляді різножанрових текстів (блоги, портали новин, дискусійні форуми, газети, журнали і т.д.) і корпусна лінгвістика вже довший час як побачила в Інтернеті вигідне і практично безмежне джерело даних, хоча можливості їх обробки не можуть розглядатися однозначно. Однією з найбільших проблем використання даних з Інтернету є їх негомогенність. Поряд з продуктом мовлення носіїв мови знаходимо зафіксоване мовлення авторів, для яких дана мова не є рідною. Крім того, здебільшого відсутні демографічні дані щодо авторів. Якість розміщених в мережі документів коливається від найбільш авторитетних і інформативних, добре структурованих, до абсолютно «анархічно» організованих. Однією з основних проблем в цій сфері А. Юкер [9] вважає мінливість, «летючість» отриманих з Інтернету мовних фактів, адже не всі Інтернет-документи надійно архівуються.

Дійсно, в Інтернеті наявний такий масивний корпус мовних даних, який у своїй сукупності ніким не може бути переглянутий, і такий дифузний і неозорий корпус не може аналізуватися без будь-яких застережних міркувань, проте неймовірна величина та відносно незначні зусилля, які потрібні для аналізу отриманих у такий спосіб даних, є занадто привабливими, щоб відмовитися від дослідження. Використання Інтернету як бази для лексикографічних і мовностатистичних досліджень можливе, якщо буде доведено, що в такий спосіб можна отримати певною мірою надійні і відтворювані результати, системно пов'язані з мовною реальністю. В останні роки в цьому напрямку проведена вже значна робота, вченими розробляються різноманітні методи і прийоми дослідження мультилінійних, мультимедійних, динамічно-летких текстів з Інтернету [3; 5; 10].

Як ми вже згадували, Г. Бікель ще в кінці 90-их років розпочав роботу над словником варіантів німецької мови, використовуючи Інтернет як текстовий корпус [3]. Пошукові машини (на той час це була Alta Vista) до-

зволяють здійснювати пошук не лише заданою мовою, але і в певному домені; для роботи над вище згаданим словником релевантними були домени *.de*, *.at* та *.ch*. На початку дослідження кількість німецькомовних сторінок складала 20 млн. (1996 рік, Alta Vista), проте з кожним роком це число стрімко зростало. На основі спеціально розроблених лексикографічних тестів (проводився аналіз частотності вибірки слів у різний період часу) автор прийшов до висновку, що Інтернет як динамічний корпус постійно зростає і змінюється, проте запити у певний період часу є консистентними, частота вживання того чи іншого слова вимірюється у практично однаковому процентному відношенні. Г. Бікель вказує, що дослідження проводилось за допомогою пошукової системи Alta Vista, оскільки Google на той час не існувало, проте проведені ним пізніші тести показали, що Google дає порівнювані результати.

С. Дімер, хоча й висловлює застереження щодо відносності оцінки квантитативних показників з Інтернету через відсутність визначеної вихідної кількості даних, проте зазначає, що отримані результати часто порівнювані із результатами з Британського національного корпусу. Ключовим пріоритетом вчений вважає розробку інструментів статистичного аналізу даних для відкритої вихідної кількості [5, с. 54-55].

Отже, Інтернет, який являє собою велетенський інформаційно-комунікаційний простір, за розробки відповідних алгоритмів досліджень може розглядатися також як гігантський лінгвістичний корпус. Очевидно, що можливості, які відкриває Інтернет для лінгвістики та лексикографії зокрема, у майбутньому використовуватимуться у багатьох проектах зі складання словників і Інтернет стане надзвичайно корисним джерелом лінгвістичних досліджень. Перспективним бачимо дослідження функціонування в Інтернет-дискурсі комунікативно-експресивної фразеології, адже велика кількість розміщених матеріалів є текстами здійснюваної тут в асинхронному та квазі-синхронному режимах комунікації.

Література:

1. Войскунский А. Е. Метафоры Интернета / А. Е. Войскунский // Вопросы философии. – 2001. – № 11. – С. 64-79.

2. Остапович О. Я. Корпусний аналіз текстових електронних масивів у вивченні фразеології іноземних мов. Методологічні принципи, практична аплікація, межі й застереження / О. Я. Остапович // Науковий вісник Волинського національного університету імені Лесі Українки. – № 5. – Філологічні науки. – 2008. – С. 391-399.

3. Bickel H. Das Internet als linguistisches Korpus / H. Bickel // Linguistik-online 28. – 2006. – № 3. URL: http://www.linguistik-online.de/28_06/bickel.html

4. Crystal D. Oh what a tangled web we weave / David Crystal // Science & Spirit. – Nov. – Dec. – 2004. – P. 34-35.

5. Diemer S. Das Internet als Korpus? : Aktuelle Fragen und Methoden der Korpuslinguistik / S. Diemer // Saarland Working Papers in Linguistics 2. – 2008. – P. 58-72.

6. Grzybek P. Sinkendes Kulturgut? Eine empirische Pilotstudie zur Bekanntheit deutscher Sprichwörter / P. Grzybek // Wirkendes Wort. – 1991. – № 11. – S. 239-264.

7. Hallsteinsdottir E. Phraseologisches Optimum für Deutsch als Fremdsprache. Ein Vorschlag auf der Basis von Frequenz- und Geläufigkeitsuntersuchungen / Hallsteinsdottir E., Sajankova M., Quasthoff U. // Linguistik-online 27. – 2006. – № 2. – S. 119-138. – URL : http://www.linguistik-online.de/27_06/hallsteinsdottir_et_al.pdf

8. Hanks P. Contextual dependency and lexical sets / P. Hanks // International Journal of Corpus Linguistics. – 1996. – № 1 (1). – P. 75-88.

9. Jucker A. Gutenberg und das Internet. Der Einfluss von Informationsmedien auf Sprache und Sprachwissenschaft / A. Jucker // Net.works. – Nr. 40 – 2004. – URL : http://www.mediensprache.net/de/networx/docs/networx-40.aspx

10. Korpora, Web und Datenbanken : computergestützte Methoden in der modernen Phraseologie und Lexikographie / [Ptashnyk S. (Hrsg.)]. – Baltmannsweiler : Schneider Verlag Hohengehren, 2010. – 267 s.

11. Sinclair J. Corpus, Concordance, Collocation / John Sinclair. – Oxford : Oxford University Press, 1991. – 197 p.

12. Stathi K. Korpusbasierte Analyse der Semantik von Idiomen / K. Stathi // Linguistik-online 27. – 2006. – № 2. – URL : http://www.linguistik-online.de/27_06/stathi.html

13. Storjohann P. Mit digitalem Textmaterial die innere Ordnung des Wortschatzes entdecken / P. Storjohann // Der Deutschunterricht (Wortschatz). – 2006. – № 1. – S. 56-68.

14. Stubbs M. Words and Phrases. Corpus Studies of Lexical Semantics / M. Stubbs. – Oxford : Blackwell Publishing, 2001. – 288 p.

15. Variantenwörterbuch des Deutschen : die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol / [Ulrich Ammon (Hrsg.)]. – Berlin : de Gruyter, 2004. – 954 s.