УДК 81.33
DOI https://doi.org/10.32782/tps2663-4880/2022.25.1.44

ВИКОРИСТАННЯ ТЕКСТОВИХ КОРПУСІВ У ЛІНГВІСТИЧНИХ ДОСЛІДЖЕННЯХ

THE USE OF THE TEXT CORPORA IN LINGUISTIC RESEARCH

Капінус О.Л.,
orcid.org/0000-0003-4535-8850
кандидат історичних наук, доцент,
доцент кафедри іноземних мов
Національного університету «Одеська юридична академія»

Байло І.Я.,
orcid.org/0000-0002-1953-0023
старший викладач кафедри іноземних мов
Національного університету «Одеська юридична академія»

Метою статті є представлення результатів вивчення сучасних літературних джерел з проблеми використання текстових корпусів у дослідницьких цілях. На підставі проведеного аналізу встановлено, що електронний корпус не просто дає можливість прискорити дослідження мови та багатократно підвищити їхню ефективність, вірогідність та перевірюваність – він допомагає вирішувати такі завдання, які лінгвістика попередніх епох практично не ставила через їхню трудомісткість чи нездійсненність. До таких завдань належать, наприклад, численні види статистичних та інших квантитативних досліджень мови, а також завдання, пов'язані з моніторингом мовних змін та описом їхніх механізмів.

З'ясовано, що відправною точкою виникнення корпусної лінгвістики стала з'ява та усвідомлення об'єктивного лінгвістичного поняття – дихотомії «мова – мовлення». Концепція, згідно з якою мова й мовлення вважаються різними реальними об'єктами, що взаємодіють у процесі мовної діяльності, завоювала й продовжує завойовувати все більше прихильників. На їхню думку, мова має лише мовленнєвий потенціал, різні можливості дискурсотворення.

Зазначено, що корпусна лінгвістика перебуває на перетині теоретичних та прикладних завдань. Можна сказати, що теоретичні дослідження готують вирішення практичних завдань і становлять їхню основу.

Увагу акцентовано на постанні нових прикладних аспектів корпусної лінгвістики. Зазначено, що зовсім недавно у цій сфері були лише два важливі та складні напрями: компіляція частотних словників на основі текстових корпусів, або статистична лексикографія, та автоматична обробка текстів, або комп'ютерна лінгвістика. Нині вже створено передумови використання ідей корпусної лінгвістики в дидактичних цілях.

Ключові слова: корпус, корпусна лінгвістика, корпусний підхід, лінгвістичне дослідження, навчання іноземних мов.

The aim of the article is to present the results of the study of modern literary sources on the problem of the use of text corpora for research purposes. On the basis of the conducted analysis, it was established that the electronic corpus not only makes it possible to speed up language research and increase its efficiency, reliability and verifiability many times over – it helps to solve the tasks the linguistics of previous eras practically did not assign because of their workload or impracticability. Such tasks include numerous types of statistical and other quantitative language studies, as well as tasks related to monitoring language changes and describing their mechanisms.

It was found out that the starting point of the emergence of corpus linguistics was the phenomenon and awareness of the objective linguistic concept – the 'language – speech' dichotomy. The concept that language and speech are considered to be different real objects that interact in the process of language activity has won and continues to win more and more adherers. In their opinion, language has only speech potential and various possibilities of discourse creation.

It is noted that corpus linguistics is at the intersection of theoretical and applied tasks. It can be said that theoretical studies prepare solutions to practical problems and form their basis. Attention is focused on the emergence of new applied aspects of corpus linguistics. It is noted that quite recently there were only two important and complex directions in this field: the compilation of frequency dictionaries based on text corpora, or statistical lexicography, and automatic text processing, or computer linguistics. Currently, the prerequisites for using the ideas of corpus linguistics for didactic purposes have already been created.

Key words: corpus, corpus linguistics, corpus approach, linguistic research, foreign language learning.

Постановка проблеми. Розвиток та використання електронних текстових корпусів – це один із найперспективніших напрямів сучасної лінгвістики. У межах саме цього напряму можна прогнозувати проривні досягнення як у галузі теоретичної лінгвістики (одержання нових знань про влаштування мови), так і в галузі прикладної лінгвістики (отримання технологій нового покоління для автоматичної обробки текстів та прискорена модернізація методів лінгвістичних досліджень) [1, с. 131]. Річ у тім, що тільки електронний корпус дає змогу в реальному часі отримувати результати, що вимагають обробки таких масивів текстів, з якими звичайний дослідник упоратися не в змозі. Адже для отримання тих самих даних вручну, шляхом, скажімо, простого

перегляду текстів і виписування прикладів на картки, можуть знадобитися місяці, а то й роки.

Утім, електронний текстовий корпус не просто дає можливість прискорити дослідження мови й багатократно підвищити його ефективність, вірогідність та перевірюваність – він допомагає вирішувати такі завдання, які лінгвістика попередніх епох практично не ставила через їхню трудомісткість чи нездійсненність. До них можна віднести, наприклад, різні види статистичних та інших квантитативних досліджень мови, які частково проводилися й у докорпусну епоху, але бурхливо розвиваються саме останнім часом. Такими ж практично нерозв'язаними є також завдання, пов'язані з моніторингом мовних змін та описом їхніх механізмів, бо, як відомо, кожна мова перебуває в процесі сталих, але повільних змін, результати яких здебільшого стають помітні лише в масштабі кількох століть.

Розуміння механізмів таких змін, на думку фахівців [2; 3], можуть дати принципово нові знання про сутність природної мови в цілому. Проте дослідження в цій галузі будуть найбільш ефективними в разі залучення даних так званих історичних, або діахронних, корпусів, що містять тексти, створені за великий проміжок часу, зазвичай не менше п'яти-семи століть.

Результатом такого діахронного розширення стає створення загальномовних корпусів різних мов. Так, вже укладено національні корпуси англійської мови (британського та американського варіантів), стали реальністю національні корпуси багатьох слов'янських мов (зокрема російської, польської, чеської, хорватської), активно поповнювано ресурси Генерального регіонально анотованого корпусу української мови тощо. Такі корпуси містять зазвичай сотні мільйонів словоживань чи інших мовних одиниць, тобто створювано ті самі «нескінченні генеральні сукупності», про які неодноразово згадували вчені, що працюють у галузі статистичної лінгвістики [4].

На думку багатьох дослідників, корпус може також стати науковою базою для відродження вже втрачених або близьких до зникнення мов у вигляді наукових граматик, академічних словників, документованих мовних матеріалів.

З погляду сучасної теоретичної та описової лінгвістики корпус є не тільки особливим інструментом вивчення мови, а й необхідним компонентом його інтегрального опису. До класичної пари «словник – граматика» сучасна наука додала третій елемент – корпус, розуміючи повний опис мови як такий, що включає корпус та побудовані на його основі словник і граматику. Відтак правомірно говорити про «корпусні словники» й «корпусні граматики» нового покоління, що створені – й верифіковані – стосовно саме конкретного фіксованого корпусу. Корпусний характер словників і граматик підвищує їхню надійність й перевірюваність, дає змогу уникнути суб'єктивності й неповноти, якими часто грішать традиційні описи.

Очевидно, що залучення даних електронних текстових корпусів поступово стає невіддільним складником лінгвістичної теорії та практики. Цим пояснюємо актуальність питань про специфіку й перспективи використання текстових корпусів для вирішення широкого кола завдань у різних галузях мовознавства, а також у суміжних науках.

Аналіз останніх досліджень і публікацій свідчить про постійний дослідницький інтерес до проблем корпусної лінгвістики та її технологій. Відправною точкою для виникнення корпусної лінгвістики науковці вважають з'яву та усвідомлення об'єктивного лінгвістичного поняття – дихотомії «мова – мовлення». Концепція, згідно з якою мова і мовлення вважаються різними реальними об'єктами, що взаємодіють у процесі мовної діяльності, завоювала й продовжує завойовувати все більше прихильників. На їхню думку, мова має лише мовленнєвий потенціал, різні можливості дискурсотворення. Аналогічні міркування висловлювано вже давно [4], при цьому лінгвісти розглядали можливості дискурсу як пристосування людини до навколишньої мовної ситуації.

Однак і на сучасному етапі розвитку корпусної лінгвістики спостережено певне протистояння традиційного (домінантного) і нового (що ще перебуває на стадії становлення) підходів до збору мовного матеріалу. Це відзначають дослідники, які використовують корпусні методи аналізу мовних явищ. Вони вважають необ'єктивним той факт, що лінгвістичні описи ґрунтуються на інтуїтивних судженнях, коли найкращим способом отримання даних виявляється не робота з текстами, а використання інтуїції носія мови, наведення штучних прикладів, а також використання різних словників [5].

Попри те, науковці зазначають, що в лінгвістиці загалом відбувається поступове зрушення в бік дослідження мовної варіативності, від мови до мовлення, від норми до узусу [1, с. 12].

Деякі труднощі викликає також відсутність єдиного розуміння термінів, використовуваних у корпусній лінгвістиці, що є звичайним у процесі розвитку будь-якого складного питання. Зрозуміло, що в корпусній лінгвістиці ключовим

є поняття корпусу, і дотепер по-різному визначуване дослідниками. Так, Е. Фінеган називає корпусом представлене в машиночитаному форматі репрезентативне зібрання текстів, що включає інформацію про ситуацію, у якій текст створено (зокрема інформацію про автора, адресат, аудиторію) [6]. Українська мовознавиця О. Демська-Кульчицька визначає текстовий корпус як перетворену на електронну форму репрезентативну вибірку текстів природної мови, призначену для наукового і практичного її вивчення. При цьому дослідниця наголошує, що тексти мають бути «систематизованими, закодованими й організованими відповідно до вимог Стандарту кодування корпусу» [7, с. 72]. Лінгвісти Т. МакЕнері, Е. Харді вважають, що корпус – це зібрання мовних фрагментів, відібраних відповідно до чітких мовних критеріїв для використання як моделі мови [8].

Наявність, здавалося б, різноманітних за своїм характером визначень самого феномену «корпус», при більш уважному розгляді показує, що вони не суперечать один одному, а просто дають його більш детальний, повний опис. Тож під лінгвістичним корпусом текстів можна розуміти машиночитане, збалансоване, репрезентативне зібрання особливо розмічених (анотованих) текстів, відібраних згідно фіксованих параметрів для досягнення визначеної лінгвістичної мети та досліджуваних нелінійно за принципом гіпертексту [5].

Очевидно, що сьогодні ми вже можемо говорити про досить різноплановий науковий доробок сучасних мовознавців щодо проблем корпусної лінгвістики. Критичне осмислення, систематизація та популяризація цих напрацювань сприятиме продуктивному опануванню можливостями текстових корпусів лінгвістичною спільнотою.

Отже, **метою** цієї розвідки є огляд та опис літературних джерел, які дають змогу оцінити значущість корпусного підходу в лінгвістичних дослідженнях, визначити його місце в сучасній теоретичній та прикладній лінгвістиці.

Виклад основного матеріалу. Базовими концепціями, якими оперували лінгвісти на самому початку досліджень текстових корпусів, були, як зазначено вище, відмінність у лінгвістичних об’єктах «мова-мовлення», і навіть стилістична диференціація текстів. На основі цих фундаментальних понять і почали розвиватися найпродуктивніші напрями корпусної лінгвістики.

Ми почнемо свій опис з тих теоретичних завдань, які можуть бути вирішені у рамках корпусної лінгвістики. Цей напрям не є широко розгалуженим об’єктом із безліччю плідних аспектів, що відійшли від основної ідеї. На перший погляд, теоретичні розвідки мають досить скромні результати у порівнянні з дуже успішними і яскравими продуктами аналізу, одержуваними у прикладних дослідженнях. Але теоретичні проблеми, розроблені на базі текстових корпусів, дають можливість у своїх результатах представляти нові знання та факти про будову мови. Так, учені неодноразово зазначали, що багато граматичних конструкцій і явищ виявляються тільки в роботі з текстовими корпусами, а вивчення типологічних явищ лінгвістики, граматичний (синтаксичний) аналіз поряд з лексичним, є найчастішим типом дослідження, для якого використовують корпуси. Можна стверджувати, що теоретична лінгвістика забезпечує знаннями та навичками всі сфери прикладної лінгвістики. Тобто практична цінність теоретичного дослідження полягає в застосовності отриманих результатів до аналізу мовленнєвих творів, до їх, певною мірою, оптимізації, а отже, може знайти та знаходить широке застосування в навчальному процесі та практиці навчання й використання письмового варіанту мови.

На перший погляд, відмінності в теоретичних та прикладних напрямах корпусної лінгвістики мають бути величезними та протилежними одне одному. Проте в реальності цього немає, бо в основу корпусної лінгвістики покладено певне розуміння того, що мова – це цілком соціальне явище, її можна описати даними, заснованими на досвіді, тобто на мовленнєвих актах. Усі, хто говорить і пише якоюсь мовою, обов’язково пристосовуються до соціальних обставин. Тож відмінність цих напрямів (теоретичний, описовий та прикладний) не означає їхньої ізоляції. Тут можна говорити про триєдину спрямованість усієї лінгвістики [9]. Ба більше, у складі власне прикладного дослідження завжди присутні елементи теоретичного.

Отже, корпусна лінгвістика перебуває на перетині теоретичних та прикладних завдань, бо текстові корпуси є унікальною основою для поєднання та координування явищ дихотомії «мова та мовлення». Можна сказати, що теоретичні дослідження готують вирішення практичних завдань і становлять їхню основу.

Зовсім недавно корпусна лінгвістика була підґрунтям лише для двох дуже важливих та складних напрямів прикладної лінгвістики – компіляції частотних словників на основі текстових корпусів, або статистичної лексикографії, та автоматичної обробки текстів, або комп’ютерної лінгвістики. Нині вже створено передумови використання ідей корпусної лінгвістики в дидактичних цілях.

Але розглянемо, як можна поєднати такі різні, на перший погляд, аспекти в річищі корпусної лінгвістики, як статистична лексикографія, комп'ютерна лінгвістика та лінгводидактика.

Статистичну лексикографію характеризує те, що вона має справу виключно з мовними явищами, а також те, що використовує математичні (кількісні) та статистичні методи дослідження цих явищ. Застосування статистичних методів є суттєвим для формування репрезентативного корпусу, аналіз якого дає досить надійні з погляду статистики результати.

Під час аналізу текстового корпусу та створення ймовірнісно-статистичних моделей (частотних словників) перевагу надають насамперед спеціалізованим текстам будь-якої галузі, наприклад, текстам юридичного дискурсу, що включає основні жанри юридичної та ділової документації.

Чому такі тексти кращі, ніж, скажімо, тексти художньої літератури? Прикладні семантичні моделі характеризують такі специфічні риси (принципи): моделювання вузькопрофесійних аспектів мовної поведінки; суворо обмежене використання даних тих чи тих рівнів мови; більша увага до аналізу, ніж до синтезу; більш високий (ніж у пояснювальних моделях) ступінь формалізації; вибір конкретних інструментів моделювання відповідно до заданої сукупності практичних вимог; прив'язка до обмеженої підмови (або комплексу підмов); повна інтерпретовність усіх елементів моделі, що випливає з обов'язкової попередньої семантичної інвентаризації підмов; вибір різних текстових утворень (не обов'язково речення) як робочого об'єкта аналізу тощо [10]. Усі представлені характеристики дають можливість сформувати особливу властивість комп'ютерної лінгвістики – оптимізацію, коли об'єкт зберігає в результуючому поданні ті істотні властивості, які необхідні лише для цієї практичної задачі [11, с. 8]. Прикладні моделі, орієнтовані на конкретні комунікативні ситуації, конкретні мови (підмови), суттєво округляють модельований об'єкт і припускають широкі можливості вибору інструменту моделювання.

Синтез проблем теоретичного та прикладного характеру сприяв з'яві сучасних напрямів. Як уже згадувано, фахівці в галузі прикладної лінгвістики вже використовують електронні текстові корпуси для вирішення певних професійних завдань, наприклад, під час навчання рідної та іноземних мов, написання підручників та посібників, для впровадження найсучасніших методик навчання мови тощо. Тож прикладна лінгвістика включає, по суті, всяке практичне завдання, що вимагає лінгвістичних знань: від такої традиційної мети лінгвістики, як навчання рідної чи іноземної мови, до автоматизації процесів управління.

Таке розуміння предмета відбито в працях українських та зарубіжних науковців [12; 13; 14; 15; 16; 17]. Називаючи корпусну лінгвістику перспективним напрямом у сфері викладання іноземних мов, вони мають на увазі як використання корпусних об'єктів (сукупності текстів), так і методи створення та аналізу корпусів. При цьому наголошувано, що корпусний підхід як метод лінгвістичного дослідження орієнтований на прикладне вивчення мови, її функціонування в різних типах дискурсу (наприклад, юридичного, що має гуманітарний характер, або науково-технічного дискурсу), що важливо для викладання іноземної мови для спеціальних цілей.

Своєю чергою у вишах України викладачі [18], які мають не лише багаторічний досвід роботи з текстами наукової комунікації, а й результати проведених корпусних досліджень, досвід формування ймовірнісно-статистичних моделей та використання частотних словників, також розпочинають поступово, поряд із використанням теоретичної граматики, вводити у процес навчання елементи структурної лінгвістики. Вони вважають, що для розв'язання проблеми адекватного аналізу текстів письмової наукової комунікації такий підхід цілком можливий. Він логічно обумовлений аналітичною знаковою системою мови, доступною для розуміння здобувачами, і більш економний з погляду часових витрат.

Наведемо лише деякі приклади, які розкривають величезні можливості застосування корпусного підходу в лінгводидактиці. Так, у процесі навчання іноземної мови в будь-якій спеціалізованій науковій галузі (гуманітарній, технічній) цей підхід допомагає суттєво впорядкувати та точно визначити поетапне впровадження мовних елементів: на першому етапі, звичайно, мають вводитися найчастотніші елементи мови, їх поєднання та реалізацію семантичної структури в реальних текстових корпусах; потім менш частотні й т. д. Наявність частотних словників може точно вказати викладачу, з яких саме одиниць мови потрібно починати мовне навчання.

Корпусна лінгвістика, яка займається також вивченням різних мовних явищ із теоретичних позицій, може надати всі дані про певні мовні явища, що функціонують у спеціалізованому тексті практично всіх типів дискурсу (гуманітарному чи науково-технічному), для впровадження у навчальний процес – від найбільш частотних до низькочастотних. Наприклад, частотність

граматичних явищ – синтаксичні словосполучення будь-якого типу (фразеологічні поєднання, синтагматика, багатокомпонентні конструкції, будь-які типи речень, структурні компоненти); структурний синтаксис; словотвірна типологія будь-яких частин мови; форми та функції частин мови; частотність лексичних компонентів, їхня віднесеність до різних стратифікаційних шарів; реалізація семантичних дефініцій слів, врахованих у нормативних словниках; можливість простежити зміни семантики слів у процесі реалізації у спеціалізованих текстах тощо.

Таким чином, можна стверджувати, що необхідність і важливість використання результатів корпусних досліджень для проведення якісного навчального процесу не підлягає сумніву.

Висновки і перспективи подальших розвідок у даному напрямку. Проаналізувавши лінгвістичну літературу, що висвітлює застосування в сучасних дослідженнях теоретичного та прикладного характеру, ми дійшли таких висновків.

У проаналізованих літературних джерелах зазначено, що завдяки корпусній лінгвістиці під час теоретичного та прикладного аналізу різних мовних явищ та явищ дискурсології загалом відбувалося та відбувається поступове зрушення в бік дослідження мовної варіативності, від мови до мовлення, від норми до узусу.

Джерела свідчать, що використання текстових корпусів вивело лінгвістику на більш високий рівень, дозволило використовувати не тільки інтуїтивні та суб'єктивні методи, але й ті, що пов'язані з матеріальними об'єктами – текстами.

Корпусна методологія впливає не лише на теоретичні (дослідні) аспекти, а й на прикладні. Крім того, вона сприяє взаємовигідному поєднанню завдань теоретичної та прикладної лінгвістик, а також збагаченню кожної з них.

Подальші дослідження будуть зосереджені на лінгводидактичних можливостях корпусних технологій.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Лук'янець Г. Г. Основні напрямки сучасних корпусних досліджень мови та перспективи їх подальшого розвитку. *Наукові праці Національного університету харчових технологій.* 2012. № 44. С. 127–133.

2. Шипнівська О. О. Структурно-семантичні та функціональні характеристики міжчастиномовної морфологічної омонімії сучасної української мови : автореф. дис. ... канд. філол наук : 10. 02. 01. Київ, 2007. 19 с.

3. Смашнюк О. І. Маркери емоційності у спонтанній комунікації (на матеріалі британського національного корпусу текстів) : автореф. дис. ... канд. філол. наук : 10.02.04. Київ, 2009. 20 с.

4. Перебийніс В. І. Статистичні методи для лінгвістів : посібник. 2-ге вид., випр. і допов. Вінниця : Нова Книга, 2013. 176 с.

5. Жуковська В. В. Вступ до корпусної лінгвістики : навч. посіб. Житомир : Вид-во ЖДУ ім. І. Франка, 2013. 142 с.

6. Finegan E. Language : its structure and use. Stamford, CT : Cengage Learning, 2015. 575 p.

7. Демська-Кульчицька О. М. Що нового в науці про мову? *Культура слова*. 2002. Вип. 61. С. 70–74.

8. McEnery T., Hardie A. Corpus Linguistics : Method, Theory and Practice. Cambridge University Press, 2011. 312 p.

9. Дарчук Н. П. Традиційна лінгвістика – структурна лінгвістика – комп'ютерна лінгвістика – триєдина сутність. *Вісник Київського національного університету імені Тараса Шевченка. Літературознавство, мовознавство, фольклористика*. 2016. Вип. 1. С. 24–27.

10. Стахмич Ю. С. Теоретичні засади вивчення моделювання природної мови у комп'ютерній лінгвістиці. *Наукові записки. Сер. : «Філологічні науки»*. Кіровоград, 2016. Вип. 144. С. 225–229.

11.Основи інформатики та прикладної лінгвістики зі змістовим модулем: копірайтинг : конспект лекцій / укладачі : А. В. Прокопенко, Л. І. Гарцунова. Суми : Сумський державний університет, 2020. 108 с.

12. Саєнко Н. С. Корпусний підхід у навчанні іноземних мов у технічному університеті. *Педагогічні науки: теорія, історія, інноваційні технології*. Суми, 2016. №1 (55). С. 142–151.

13. Жуковська В. В. Лінгвістичний корпус як новітній інформаційно-дослідницький інструментарій сучасного мовознавства. *Вчені записки ТНУ ім. В. І. Вернадського. Серія : Філологія. Соціальні комунікації*. Київ, 2020. Т. 31 (70), № 3 Ч. 1. С. 113–119.

14. Gavioli L. Exploring Corpora for ESP Learning. Amsterdam ; Philadelphia : John Benjamins, 2006. 176 p.

15. Basanta C. P., Martín, M. E. R. The application of data-driven learning to a small-scale corpus: using film transcripts for teaching conversational skills. *Corpora in the Foreign Language Classroom*. Leiden, 2007. P. 148–158.

16. Boulton A. Separating fact and fiction: The real story of corpus use in language teaching. 20 Years of EUROCALL : Learning from the Past, Looking to the Future : Proceedings of the 2013 EUROCALL Conference. Evora, Portugal, 2013. P. 51–56.

17. Liu D., Lei L. Using corpora for language learning and teaching. TESOL International Association, 2017. 144 p.

18. Шапа Л. Н. Методические указания по работе с лексикой по специальности «Электроснабжение» (частотный словарь-минимум). Одесса : ОГПИ, 1993. 36 с.