КОРПУС ТЕКСТІВ: ОСНОВНІ АСПЕКТИ ВИЗНАЧЕННЯ

Бобкова Т. В.
Київський національний лінгвістичний університет

У статті окреслено основні проблеми визначення терміна "корпус текстів". Окреслені основні підходи до визначення корпусу текстів у широкому й вузькому розумінні. Здійснено аспектуалізацію поняття корпусу з урахуванням широкого й вузького розуміння терміна. Поняття корпусу проаналізовано в аспектах предметної галузі лінгвістичного дослідження; складників, що входять до тіла корпусу, галузевої належності визначальних термінів і детермінативних ознак корпусу текстів. З'ясовано еволюцію терміна "корпус текстів" в українській лінгвістичній традиції. Встановлено, що сучасні спроби визначення корпусу текстів, пов'язані з розвитком типології корпусів і спеціалізації корпусної лінгвістики, базуються на апріорному визнанні частини детермінативних ознак корпусу.

Ключові слова: корпус текстів, корпусна лінгвістика, доелектронний корпус текстів, електронний корпус текстів, корпусна лінгвістика першого покоління, корпусна лінгвістика другого покоління, корпусна лінгвістика третього покоління.

The article is to outline the main problems of the defining a term "textcorpus". The basic approaches to determining the corpus of texts in the broad and narrow sense are identified. An aspectualization of the term was performed with a consideration of wide and narrow understanding of the term. The notion of corpus is analyzed in the aspects of an object sphere of the linguistic research, constituents, which belong within the corpus body, sphere belonging of the defining terms and the determinative features of a text corpus. The evolution of the term "text corpus" in the Ukrainian linguistic school is determined. It is established, that nowadays efforts of defining a text corpus, which are related to the development of corpus typology and corpus linguistics specialization, are based on the antecedent acknowledgement of the part of corpus determinative features.

Key words: corpus, corpus linguistics, pre-electronic corpus, electronic corpus, first generation corpus linguistics, second generation corpus linguistics, third generation corpus linguistics.

Актуальність проблеми визначення корпусу текстів пов'язана з активним упровадженням корпусного підходу в сучасній лінгвістиці. Це вимагає розв'язання низки філософських і теоретико-методологічних питань теорії корпусу, і насамперед потребує визначення фундаментальне поняття корпусної лінгвістики – корпус текстів. Дискусії щодо визначення корпусу виявляють співіснування різних концепцій у сучасній корпусній лінгвістиці [36, p. 19], що зумовлюють теоретичні принципи планування й практичні засади використання корпусів. Зокрема наявність вузького й широкого розуміння терміна "корпус текстів" є причиною існування двох основних підходів до встановлення періодизації, передумов формування й визначення статусу корпусної лінгвістики в сучасному мовознавстві.

У широкому значенні під корпусом розуміють будь-яке зібрання письмових або усних текстів, використовуване з метою дослідження мови [8, с. 270; 16, p. V; 20, p. 67; 28, p. 1–13]. Зазначена інтерпретація основного поняття приводить до виділення періоду ранньої корпусної лінгвістики (кінець XIX ст. – 1960 рр.), що дозволяє значно розширити хронологію розвитку дисципліни за рахунок доелектронних корпусів текстів[4, с. 10; 8, с. 270; 14, p. 34; 20, p. 13–14; 24, p. 20–22; 28, p. 1–13; 33, p. 12] й обгрунтувати концептуальні витоки корпусної лінгвістики. У вузькому значенні під корпусом розуміють зібрання текстів в електронній формі, що презентує певну мову [4, с. 20;5, с. 3;18; 22, p. 29–30;26, p. 215;27, p. 5]. Зрозуміло, що за такою інтерпретацією корпусу текстів, витоки корпусної лінгвістики датуються другою половиною

XX ст. [11, с. 12; 22; 31; 35, p. 136]. Незважаючи на співіснування в сучасній корпусній лінгвістиці двох основних підходів до визначення корпусу текстів, в основу більшості спроб покладено вузьке розуміння основного терміна.

Мета статті – окреслити основні аспекти визначення корпусу залежно від предметної галузі дослідження, складників корпусу, визначальних термінів і детермінативних ознак корпусу. Досягнення поставленої мети передбачає виконання таких завдань:

– окреслити основні проблеми визначення корпусу текстів;

– здійснити аспектуалізацію аналізованого поняття;

– дослідити визначення корпусу текстів у різних аспектах з урахуванням широкого та вузького розуміння;

– представити аргументи на користь вузького розуміння корпусу текстів на базі детермінативних ознак;

– охарактеризувати наявні у вільному доступі українські корпуси на предмет їх відповідності вимогам до планування корпусів текстів.

Від самого початку становлення корпусної лінгвістики ідея корпусу була покладена в основу нового підходу до розуміння й дослідження мови. Усвідомлення корпусу текстів як основного об'єкта корпусної лінгвістики сприяло виокремленню корпусної лінгвістики як самостійної галузі сучасного мовознавства. Однак, незважаючи на те, що поняття корпусу належить до вихідного категоріального апарату корпусної лінгвістики, цей термін часто використовується в лінгвістичній літературі взагалі без визначення [1; 34, p. 24]. Серед причин відсутності загальноприйнятої дефініції корпусу слід виділити, з одного боку, особливості розвитку корпусної лінгвістики, а з другого – складність визначуваного об'єкта.

Становлення корпусної лінгвістики насамперед як емпіричної дисципліни передбачало переважний розвиток практичних засад протягом 1960–90 рр. За таких обставин недостатня увага до розроблення теорії корпусу призвела до невизначеності базових понять, відсутності аналізу питань онтології та епістемології в корпусній лінгвістиці, що негативно вплинуло на розвиток усього напрямку [21, p. 18]. Саме тому на початкових етапах корпусним лінгвістам часто закидали звинувачення у "філософській наївності" розроблених фундаментальних питань або взагалі у відсутності власної розробленої теорії. Проте, незважаючи на те, що корпусна лінгвістика спочатку народилася як методика, а вже потім як наука [26, p. 180], вона не є вільною від теорії, оскільки створення й використання корпусу ґрунтується на теоретичних принципах емпіричного дослідження мови. З впровадженням корпусного підходу в лінгвістичні дослідження корпус текстів у свою чергу стає середовищем для розроблення та верифікації корпусної теорії [21, p. 28–29]. У цьому контексті корпус як система цінується завдяки багатофункціональності, інтерактивності, придатності до широкого використання в теоретичних і практичних дослідженнях мови. З іншого боку, саме ці властивості ускладнюють визначення багатоаспектної системи, у якій переплетені суто лінгвальні та екстралінгвальні параметри – найважливіші критерії корпусу [24, p. 5].

Уведення в науковий обіг терміна "корпус текстів" пов'язують із опрацюванням у 1960–80 рр. електронних корпусів першого покоління: Браунівського (У. Френсис, Г. Куцера, 1963–64 рр.), Ланкастерського (Дж. Ліч, 1970 р.) та Лондонсько-Лундського корпусу текстів (Р. Кверк, Я. Свартвік, 1980 р.). Причому вперше з необхідністю визначення терміна зіткнулися У. Френсис і Г. Куцера при створенні Браунівського корпусу (The Brown Corpusof Standard American English), які за основу використали словникову дефініцію лексеми "корпус" – це сукупність текстів, які вважаються представницькими для лінгвістичного аналізу певної мови [4, с. 19].

Пізніше із впровадженням корпусного підходу в лінгвістичні дослідження зазначене вище визначення зазнало багато уточнень і змін, навіть сформувалася думка про неможливість вичерпної дефініції корпусу текстів [1; 25, p. 4–5]. Однак, як показує вивчення матеріалів

дискусій у сучасній корпусній лінгвістиці, подолання проблем визначення корпусу можливе через аспектуалізацію аналізованого поняття: різні інтерпретації корпусу текстів відображають різні аспекти цього ресурсу [32, p. 1]. У цьому розумінні слід виокремити спроби визначення основного терміна залежно від: 1) предметної галузі лінгвістичного дослідження; 2) складників, що входять до тіла корпусу; 3) галузевої належності визначальних термінів; 4) апріорних детермінативних ознак корпусу.

1) Спроби інтерпретувати поняття корпусу текстів залежно від предметної галузі дослідження спираються на загальне енциклопедичне визначення корпусу Д. Кристала та вузьке галузеве розуміння терміна[17, p. 2–3]. Так, усередині домену загального мовознавства корпус визначається як колекція лінгвальних даних письмових текстів або транскрибованих записів мовлення, які можуть бути використані як відправний пункт лінгвістичного опису або засіб верифікації гіпотез про мову. На увазі мається дослідний масив текстів, і зазначений вище термін за значенням є подібним, зокрема, до корпусу словника.

У контексті власне корпусної лінгвістики корпус розуміється як велика колекція зразків письмових та усних текстів, доступних у машиночитаній формі, зібраних науково обґрунтованим способом для презентації певного різноманіття або вживання мови [17, p. 3]. Отже, залежно від предметної галузі Н. С. Даш пропонує розмежовувати широке розуміння корпусу в домені загального мовознавства як об’єкта лінгвістичного дослідження взагалі й вузьке розуміння – в домені корпусної лінгвістики як об’єкта власне корпусної лінгвістики.

2) Спроби визначити корпус залежно від його складників розрізняються за детермінативною ознакою автентичності текстів, які входять у тіло корпусу. Так, у рекомендаціях Консультативної групи експертів з питань мовних технічних стандартів (EAGLES) пропонується широке розуміння корпусу як зібрання, що потенційно може містити не тільки прозу, поезію, а й реєстри слів і словники[27, p. xi]. За такою інтерпретацією до корпусу включаються не тільки автентичні тексти природною мовою, а й списки слів, що є результатами дослідження текстів (як реєстри) або впорядкування й лексикографічного опису (як словники).

Проте, сучасні лінгвісти, які практикують корпусний аналіз дотримуються більш вузького визначення “для службового використання” [27, p. xi]: корпус інтерпретуються як колекція автентичних текстів або їх фрагментів для здійснення лінгвістичних досліджень. При цьому під автентичними даними, як правило, розуміють дані природної мови, а не штучно створені або виявлені в ході лінгвістичного дослідження [17, p. 1]. Відповідно корпуси, побудовані з використанням уривків текстів для вивчення певного лінгвального явища, не є корпусами в повному сенсі цього слова [27, p. 46]. Отже, спроби визначення корпусу залежно від його складників передбачають розмежування широкого розуміння корпусу як колекції будь-яких текстів і вузького – як колекції автентичних текстів.

3) Дискусії останніх років щодо статусу корпусної лінгвістики в сучасному мовознавстві привели до спроб визначити корпус у термінах різних галузей. Відсутність одностайної думки щодо місця корпусної лінгвістики в науковій парадигмі пояснюється складністю визначення статусу “недисциплінованої дисципліни” [30]. Залежно від статусу корпусної лінгвістики інтерпретації корпусу базуються на врахуванні суто лінгвальних, нелінгвальних або комплексу лінгвальних і нелінгвальних параметрів корпусу як складного об’єкта. При цьому під лінгвальними параметрами, або вимірами, розуміють лінгвальну варіативність і специфіку подібних явищ у мовленні [25, p. 161]. Тоді, як точне визначення й аналіз нелінгвальних параметрів корпусу є неможливим без встановлення меж і параметрів мовлення всередині мови, а також оцінки корпусу як певної вибірки[25, p. 127].

Традиційно корпусну лінгвістику залучають до гуманітарних наук, власне до лінгвістичних дисциплін [3, с. 46; 4, с. 6–7; 6, с. 7; 24, p. 2; 30]. При цьому основним об’єктом корпусної лінгвістики визнається текст [30], і здобуття його сенсу є стандартним завданням гуманітарних

наук [35, p. 140–141]. Спроба визначення корпусу в суто лінгвістичних термінах як дистрибуції слів або моделей [25, p. 5] базується на ототожненні корпусу з текстом та абстрагуванні від нелінгвальних ознак. Слід відмітити, що дефініція корпусу як дистрибуції є неоднозначною, оскільки в певному сенсі збігається з визначенням конкордансу. Ототожнення корпусу з текстом не можна вважати правомірним, оскільки до корпусу входить також корпус-менеджер – інструментарій лінгвістичного й статистичного аналізу текстів. Отже, складність системи потребує врахування комплексу лінгвальних і нелінгвальних параметрів корпусу, визначальних для суті аналізованого поняття.

Визначення корпусної лінгвістики як самостійної дисципліни в межах комп'ютерної [6, с. 7; 7, с. 74] або прикладної лінгвістики [4, с. 8] передбачає інтерпретацію корпусу у відповідних термінах інформатики. Зокрема, корпус може інтерпретуватися як різновид інформаційно-пошукової системи [5, с. 18] або електронної бібліотеки, побудованої за зовнішнім критерієм відповідно до певного завдання [29, p. 36]. Наведені вище визначення корпусу, з одного боку, збігаються з дефініцією електронної бібліотеки або архіву, а з другого – не враховують суттєві лінгвістичні критерії корпусу як колекції текстів. Подібно до цього можна класифікувати спроби визначення корпусу як сформованої за певними правилами вибірки даних [1] або текстозорієнтованої, повнотекстової бази даних [7, с. 74].

Однак, спроби інтерпретувати корпус на базі суто лінгвальних або нелінгвальних параметрів навряд чи можна визнати слушними, оскільки сформульовані в такий спосіб визначення є дещо інтуїтивними [4, с. 19] і не відображають комплексу суттєвих ознак аналізованого поняття. Зокрема, запропонована Дж. Аткінсом типологія текстів базується на 29 екстралінгвальних параметрах, які є релевантними для планування збалансованого корпусу, проте визнається, що неможливо досягти збалансованості корпусу лише на базі екстралінгвістичних параметрів [25, p. 16–17]. Саме тому у випадках необхідності уточнення різновиду інформаційно-пошукової системи вживається термін "лінгвістичний корпус текстів".

Граничний характер корпусної лінгвістики пояснює визначення її статусу як міждисциплінарної галузі, що поєднує інформаційні, комп'ютерні технології і власне лінгвістику [30]. Зазначені особливості статусу корпусної лінгвістики в науковій парадигмі пояснюються тим, що за об'єктом дослідження дисципліна належить до гуманітарних наук, а за методологією вивчення мови – до точних [35, p. 140–141]. Наведене визначення статусу дисципліни сприяє можливості інтерпретації корпусу текстів безпосередньо в контексті корпусної лінгвістики. У термінах корпусної лінгвістики інтерпретація корпусу здійснюється на базі детально розробленої схеми характерних ознак [17, p. 3; 27, p. xi–xiii; 4, с. 19]. При цьому корпусом уважається певний масив, точніше закінчена колекція машиночитаних текстів, дібраних для оптимального представлення мовного різноманіття [13; 15; 19; 20; 23; 25].

Набір детермінативних ознак, або параметрів, що визначає структуру й перспективи використання корпусу, формулюється заздалегідь при плануванні корпусу і може варіюватися залежно від дослідницького призначення [12, с. 344]. Однак, незважаючи на існування різних варіантів визначення [18, p. 17; 27, p. 1], більшістю корпусних лінгвістів [25, p. 5; 14, p. 5; 23, p. 120] за визначальні ознаки корпусу приймаються: 1) машиночитаність, 2) автентичність текстів (включаючи усні транскрибовані), 3) добірність, 4) репрезентативність. При цьому під репрезентативністю, за О. Демською-Кульчицькою, розуміється релевантне відображення предметної галузі в корпусних даних, де задано в пропорції, детермінованій реальною частотою досліджуваного явища, всі властивості відтворюваного в корпусі домену[4, с. 58].

Відсутність одностайності у встановленні необхідних і достатніх ознак передбачає окреслення певних обмежень [32, p. 1], що застосовуються до визнання певної колекції текстів як корпусу. Так, у 1990 рр. набувають більшої ваги ознаки аплікативності [14, p. 5; 23, p. 120] і репрезентативності [23, p. 116]. Проте питання щодо ознаки репрезентативності є найбільш

дискусійним у корпусній лінгвістиці через невизначеність техніки відбору текстів до корпусу. Саме тому більшість лінгвістів визнає, що не варто будувати визначення й планування корпусу на базі однієї детермінативної ознаки: корпус доцільно розглядати як певною мірою розпливчастий і всеосяжний термін [25, p. 5]. На доказ цього положення наводиться Ланкастерський корпус (LCA), при побудові якого використано уривки з текстів для вивчення певного лінгвального явища. Аналізований корпус відповідає лише параметру репрезентативності, і не є корпусом в повному сенсі слова [27, p. 46].

Відповідно, лише урахування комплексу лінгвальних і нелінгвальних параметрів дозволяє оптимально відобразити суттєві ознаки корпусу. Окреслені вище спроби визначити корпус текстів у термінах певної галузі приводять до розмежування вузького розуміння, яке базується на лінгвальних або нелінгвальних параметрах корпусу, й широкого – у термінах корпусної лінгвістики з урахуванням комплексу лінгвістичних і нелінгвістичних детермінативних ознак корпусу.

4) Сучасні спроби визначення корпусу текстів базуються на апріорному визнанні частини детермінативних ознак, що пов'язано з розвитком типології корпусів і спеціалізації галузі. Дослідження різних підходів до визначення корпусу текстів на підставі базових детермінативних ознак, що дозволяють класифікувати певну колекцію текстів як корпус, подано в монографії О. Демської-Кульчицької [4, с. 19–23].

У загальному значенні під корпусом розуміють колекцію текстів чи їх фрагментів, зібраних для лінгвістичних досліджень [37, p. 2; 20, p. 67; 27, p. xi; 24, p. 20–22; 33, с. 12; 8, с. 271–287; 4, с. 10–12]. При широкому розумінні зазначене визначення охоплює доелектронні та електронні типи лінгвістичних корпусів [8, с. 270; 9, с. 1–13]. При цьому доелектронними вважають створені в докомп'ютерний час корпуси, що слугували підґрунтям для ручного аналізу мовного матеріалу: апріорною детермінативною ознакою корпусу стає не машиночитана форма, а його дослідне призначення, яке передбачає текстоорієнтований підхід до вивчення мови.

Ідея вивчення мови за допомогою колекції текстів не є новою в історії лінгвістики. Корпусна методологія, що застосовувалася в структурній традиції на початку XX ст., є значно старшою за комп'ютер. У цей час доелектронні корпуси створювали з метою вивчення текстів [8, с. 270; 10, с. 52], укладання конкордансів, граматик і різноманітних словників [28, p. 1–13]. Однак структуралісти (Фр. Боас (Handbook of Native American Indian Languages, 1911 р.), Л. Блумфільд (Language, 1933 р.), З. Харріс (Methods in Structural Linguistics, 1951 р.) використовували корпусно-базовану методологію дослідження мови без корпусу як такого [24, p. 2–4; 28]. Прості колекції текстів, створені на зразок корпусів [27, p. 13] для фонологічних і граматичних досліджень, не відповідають більшості детермінативних ознак корпусу, зокрема, не вважаються репрезентативними [24, p. 2–4].

Отже, питання про визнання доелектронних корпусів текстів у сучасній корпусній лінгвістиці є дискусійним [20, p. 13–19], зокрема про це свідчить неусталеність термінології щодо корпусів зазначеного типу в сучасній лінгвістичній літературі:

– примітивний корпус (primitive corpus) батьківських щоденників для психолінгвістичних досліджень В. Прейера [27, p. 3];

– колекція текстів (collection of texts) [25, p. 3; 32, p. 2];

– доелектронний корпус (pre-electronic corpus) [8, с. 271; 28, p. 1–13];

– докомп'ютерний корпус (pre-computer corpus) [23, p. 106].

Слід зазначити, що терміни "корпус текстів" і "корпусна лінгвістика" з'явилися в українській лінгвістичній традиції на початку XXI ст. [4, 7, с. 74]. Певний час на позначення емпіричних корпусно-базованих досліджень використовувався термін "текстоорієнтований" [10, с. 14], або "текстозорієнтований" [7, с. 75]. В Україні традицію укладання текстоорієнтованих словників було розпочато в 1981 р. виданням , створеного на вибірці текстів у 500 тис. слововживань, як і всі доелектронні корпуси, вручну.

Загалом використання доелектронних корпусів текстів у лінгвістичних дослідженнях зазнало суворої критики Н. Хомського через високу ймовірність спотворення даних у неправильно організованій вибірці. Пізніше частково критика Н. Хомського була визнана слушною: розмір невеликих корпусів ("shoebox-corpora") дозволяв здійснювати лише фонетичні, зрідка граматичні дослідження через використання лише паперу, людських рук і очей [25, p. 4]. Отже, відсутність репрезентативності й належної фіксації емпіричних даних у доелектронних корпусах ставить під сумнів існування відповідної корпусно-базованої методики в докомп'ютерний час.

Опоненти, які будують свою аргументацію на вузькому розумінні корпусу текстів, вважають окреслений вище підхід наївним. У вузькому значенні під корпусом розуміють зібрання текстів в електронній формі, що презентує певну мову [5, с. 3; 22, p. 29–30]. При такому розумінні апріорною детермінативною ознакою корпусу текстів є машиночитаність: доелектронні корпуси текстів з їх "виснажливим аналізом мовного матеріалу вручну" [28, p. 1] не визнаються. Електронні корпуси текстів, починаючи з першого Браунівського корпусу (1962–1963 рр.), створеного У. Френсисом і Г. Куцерою, і до сучасних корпусів, уважаються наслідком комп'ютерної революції: термін "корпусна лінгвістика" стає синонімом терміна "комп'ютерна корпусна лінгвістика". Як звичай, у сучасній лінгвістичній літературі на позначення аналізованого типу корпусів використовуються терміни:

– комп'ютерний [10, с. 23–24; 23, p. 116],
– машиночитаний [4, с. 19; 17, p. 3; 26, p. 215],
– комп'ютеризована колекція природних текстів [37],
– електронний [5, с. 3; 28, p. 12],
– машинний [10, с. 18].

На користь наведеного вузького розуміння корпусу текстів свідчить також виокремлення в 1980-1990 рр. корпусної лінгвістики як самостійної галузі з власним об'єктом – корпусом текстів. Саме завдяки створенню електронних корпусів у другій половині XX ст. постає необхідність уточнити визначення корпусу текстів і виділити його основні ознаки. Багатьма лінгвістами, слідом за У. Френсисом і Г. Куцерою, апріорно визнається електронна форма існування корпусу, стандартність та можливість його багатократного використання [4, с. 20]. Зокрема на базі апріорного визнання електронної форми побудовано більшість визначень корпусу текстів у сучасній корпусній лінгвістиці [5, с. 3; 22; 26, p. 215].

Подальша деталізація визначення корпусу текстів представлена в російській корпусній традиції. Зокрема, на думку В. О. Плунгяна, електронна форма корпусу дозволяє вважати його лише протокорпусом, а можливість багатократного застосування корпусу передбачає його анотованість або кодованість. За такого розуміння корпусом є масив текстів в електронному вигляді, що супроводжується спеціальною розміткою [18; 11] і є філологічно компетентним для розв'язання лінгвістичних завдань [5, с. 3–7]. За такою інтерпретацією, апріорною детермінативною ознакою корпусу текстів стає анотованість. Зокрема, для порівняння анотована версія Браунівського корпусу з лематизацією словоформ і поверховим синтаксичним аналізом з'явилася майже через двадцять років у 1980 р.

Поява моніторингових, або динамічних корпусів приводить до чергової спроби уточнити визначення корпусу текстів. У 1980 р. для укладання Корпусного словника англійської мови (COBUILD) було розпочато створення моніторингового корпусу . Характерною ознакою зазначеного корпусу є постійне поповнення матеріалу для відображення стану сучасної мови. Однак саме через зазначені принципи побудови моніторинговий корпус не задовольняє вимоги щодо закінченої колекції машиночитаних текстів. Відповідно до цього вважається доцільним не використовувати термін корпус на позначення динамічних текстових колекцій, що постійно змінюють свій склад або зростають[27, p. 5]. У такому разі апріорною для визначення корпусу стає детермінативна ознака скінченності обсягу [26, p. 215] або статичності [4; p. 25–29].

Отже, спроба визначення корпусу текстів на базі апріорних детермінативних ознак передбачає широке розуміння терміна з визнанням різних типів корпусів і вузьке, згідно з яким корпусом уважаються лише певні типи корпусів, як-то електронні, анотовані або статичні. Серед аналізованих вище підходів до визначення корпусу текстів останній є найбільш деталізованим і розробленим у сучасній корпусній лінгвістиці. Зазначений підхід на базі апріорного визнання детермінативних ознак охоплює низку питань щодо корпусної теорії і типології, методології і верифікації результатів корпусного аналізу.

Щодо окреслених вище ознак, а значить і вимог до планування корпусів, українська мова залишається однією з небагатьох європейських мов, що не мають репрезентованого національного корпусу [3, с. 46], його створення лише усвідомлюється як нагальне завдання й перспектива розвитку української лінгвістики. На сьогодні українська корпусна лінгвістика представлена різними типами корпусів:трьома дослідницькими корпусами текстів української мови [39; 40; 41], Навчальним корпусом англійських текстів (Ukrainian Corpusof Learner English – UCLE) [42] і Багатомовним паралельним корпусом усного мовлення [38].

Серед представлених у вільному доступі проектів новітній корпус текстів української мови укладено колективом кафедри української мови та прикладної лінгвістики Донецького національного університету з метою вивчення проблеми граматичної службовості [2, с. 224–225]. У межах проекту реалізовано технічні й програмні аспекти реалізації корпусу, розроблено морфорозмітку й метарозмітку, а також систему тегів для службових частин мови. На сьогодні новітній корпус текстів української мови загальним обсягом близько 5 млн. слововживань функціонує в тестовому режимі[2, с. 224–225].

Багатомовні корпуси представлені паралельним корпусом усного мовлення [38] загальним обсягом біля 8 млн. Корпус розроблено в лабораторії комп'ютерної лінгвістики Київського національного лінгвістичного університету на базі субтитрів серіалів комедійного, драматичного й науково-популярного жанру. Аналізований корпус включає підкорпуси оригінальних текстів англійською мовою загальним обсягом біля 2 млн. та відповідних перекладів німецькою – 0,65 млн., французькою – 0,8 млн., українською – 0,2 млн., російською –1,1 млн., іспанською – 1,2 млн. і грецькою – 1,2 млн. Особливістю розроблення такого паралельного корпусу текстів є розв'язання проблеми автоматичного вирівнювання речень через використання параметру синхронізації часу появи субтитрів на екрані. Програмне забезпечення корпусу дозволяє здійснювати пошук перекладних еквівалентів слів і словосполучень у контексті речення, однак морфологічне анотування й модуль лематизації відсутні.

Навчальні корпуси представлені в українській лінгвістиці тестовою версією UCLE, створеною в лабораторії комп'ютерної лінгвістики Київського національного лінгвістичного університету. Хоча у вільному доступі наявний незначний фрагмент корпусу, загальний обсяг текстів студентських есе становить понад 180 тис. слововживань. Програмне забезпечення навчального корпусу дозволяє будувати KWIC і повні конкордансні списки, здійснювати пошук окремих слів і словосполучень, сортувати списки слів, відображати знайдені словоформи в необмеженому контексті, отримувати статистичну інформацію про окремі елементи корпусу.

Дослідницький корпус сучасної української мови [41] загальним обсягом у 13 млн. словоформ побудовано як інформаційно-довідкову систему, призначену для з'ясування різних питань вивчення української мови. Корпус анотовано за якісними й кількісними ознаками різних мовних одиниць на рівні морфеміки, морфології й синтаксису, а також забезпечено пакетами програм для укладання електронних картотек і параметризованої бази даних, яка включає алфавітно-частотні словники слів і слововживань спільної лексики, неолексем та синтаксичних моделей керування, словники синонімів, антонімів, фразеологізмів, тезаурусів, словник-конкорданс, серію морфемних і словотвірних словників [3, с. 46–47]. Отже, серед

представлених у вільному доступі проектів усім сучасним вимогам до планування корпусів відповідає лише Корпус сучасної української мови, розроблений у лабораторії комп'ютерної лінгвістики Київського національного університету імені Тараса Шевченка.

Здійснене дослідження основних аспектів визначення корпусу текстів дозволяє дійти таких висновків:

1. На сучасному етапі в корпусній лінгвістиці відсутнє загальноприйняте визначення лінгвістичного корпусу текстів.

2. Аспектуалізація аналізованого поняття передбачає визначення корпусу текстів залежно від предметної галузі лінгвістичного дослідження, складників корпусу, галузевої належності визначальних термінів та апріорно встановлених детермінативних ознак корпусу.

3. Залежно від предметної галузі розрізняють широке значення – корпус як об'єкт лінгвістичного дослідження й вузьке – корпус як об'єкт корпусної лінгвістики.

4. Залежно від типу текстів, що становлять тіло корпусу розрізняють широке значення, яке базується на визнанні корпусом колекції будь-яких текстів (прозових, поетичних словників, реєстрів слів), і вузьке: корпусом уважається колекція автентичних текстів.

5. Залежно від визначувальних термінів розмежовують вузьке розуміння корпусу, яке базується або на лінгвістичних, або на нелінгвістичних параметрах корпусу, і широке, власне, в термінах корпусної лінгвістики з урахуванням комплексу всіх параметрів.

6. Найбільш поширеним у сучасній корпусній лінгвістиці є визначення корпусу текстів на базі детермінативних ознак.

7. Залежно від набору апріорно встановлених детермінативних ознак розрізняють широке розуміння терміна з визнанням різних типів корпусів і вузьке – з визнанням певного типу корпусів на базі апріорно встановленої детермінативної ознаки.

8. Серед наявних у вільному доступі українських корпусів усім необхідним параметрам і сучасним вимогам відповідає лише Корпус сучасної української мови, створений у лабораторії комп'ютерної лінгвістики Київського національного університету імені Тараса Шевченка.

Література

1. Баранов А. Н. Введение в прикладную лингвистику: [учебное пособие] / Анатолий Николаевич Баранов. – М. : Эдиториал УРСС, 2001. – 360 с.

2. Данилюк І. Корпус текстів для вивчення граматичної службовості / І. Данилюк // Лінгвістичні студії. Лінгвістичні студії : зб. наук. праць / Гол. ред. А. П. Загнітко. – Донецьк : ДонНУ, 2013. – Вип. 26. –С. 224–229.

3. Дарчук Н. П. Дослідницький корпус української мови: основні засади і перспективи / Н. П. Дарчук // Вісник Київського національного університету імені Тараса Шевченка. Серія: Літературознавство. Мовознавство. Фольклористика. – К. : Видавничо-поліграфічний центр "Київський університет", 2010. – № 21. – С. 45–49.

4. Демська-Кульчицька О. Основи національного корпусу української мови : [монографія] / Орися Демська-Кульчицька. – К. : Інститут української мови НАНУ, 2005. – 219 с.

5. Захаров В. П. Корпусная лингвистика: [учебно-методическое пособие] / Виктор Петрович Захаров. – СПб. : РОПИ СПб. университета, 2005. – 48 с.

6. Корпусная лингвистика: [учебник] / В. П. С. Ю. – Иркутск : Изд. Иркутского гос. линг. университета, 2011. – 161 с.

7. Карпіловська Є. А. Вступ до комп'ютерної лінгвістики: комп'ютерна лінгвістика : [підручник] / Євгенія Анатоліївна Карпіловська. – Донецьк : ТОВ "Юго-Восток, ЛТД", 2006. – 188 с.

8. Лендау С. І. Словники: мистецтво та ремесло лексикографії / Сидні І. Лендау; [пер. з англ.]. – К. : К. І. С., 2012. – 480 с.

9. Михайлов М. Параллельные корпуса художественных текстов: принципы составления и возможности применения в лингвистических и переводоведческих исследованиях (на примере русско-финского параллельного корпуса художественных текстов) : дисс. … докт. философ. наук / Михаил Михайлов. – Тампере : Тамперский университет, 2003. – 255 с.
10. Перебийніс В. І. Традиційна та комп'ютерна лексикографія : [навчальний посібник] / В. І. Перебийніс, В. М. Сорокін. – К. : Видавничий центр КНЛУ, 2009. – 218 с.
11. Плунгян В. А. Корпус как инструмент и как идеология : о некоторых уроках современной корпусной лингвистики / В. А. Плунгян // Русский язык в научном освещении. – М. : Языки славянской культуры, 2008. – № 2 (16). – С. 7–20.
12. Фрэнсис У. Н. Проблемы формирования и машинного представления большого корпуса текстов // Новое в зарубежной лингвистике : Проблемы и методы лексикографии. – М. : Прогресс, 1983. – Вып. XIV. – С. 334–353.
13. Aarts J. Intuition-based and observation-based grammars / J. Aarts // English corpus linguistics / [eds K. Aijmer, B. Altenberg]. – London : Longman, 1991. – P. 44–62.
14. Aston G., Burnard L. The BNC handbook : exploring the British National Corpus with SARA / G. Aston, L. Burnard. – Edinburgh : Edinburgh University Press, 1998. – 256 p.
15. Biber D. Corpus linguistics: investigating language structure and use / D. Biber. – Cambridge : Cambridge University Press, 1998. – 310 p.
16. Corpus linguistics : [an international handbook] / [ed. A. Lüdeling, M. Kytö]. – Vol. 1. – Berlin : Walter de Gruyter, 2008.– 776 ð.
17. Dash N. S. Corpus linguistics and language technology: with reference to Indian Languages / Niladri Sekhar Dash. – New Dehli : Mittal Publications, 2005. – 445 p.
18. Francis W. N. Language Corpora B. C. / W. N. Francis // Directions in Corpus Linguistics / [ed J. Svartvik]. – Berlin and New York : Moutin, 1992. – P. 17–34.
19. Johansson S. Times change and so do corpora / S. Johansson // English corpus linguistics : studies in honour of J. Svartvik / [ed A. Altenburg]. – London : Longman, 1991. – P. 305–314.
20. Kennedy G. Introduction to corpus linguistics / Graeme Kennedy. – London : Longman, 1998. – 315 p.
21. Lager T. A. logical approach to computational corpus linguistics : [Doctoral Dissertation for the degree of Doctor of Philosophy] / Torbjörn Lager. – Göteborg : Göteborg University, Department of Linguistics, 1995. – 326 p.
22. Leech G., Fallon R. Computer corpora – what do they tell us about culture? / G. Leech, R. Fallon // International Computer Archive of Modern English Journal. – 1992. – No 16. – Ð. 29–50.
23. Leech G., Fligelston S. Computers and corpus analysis / G. Leech, S. Fligelston // Computers and written texts / [ed C. S. Butler]. – Oxford : Blackwell Oxford, 1992. – P. 115–140.
24. McEnery T., Wilson A. Corpus linguistics / Tony McEnery, Andrew Wilson. – Edinburgh: Edinburgh University Press, 2001. – 235 p.
25. McEnery T., Xiao R., Tono Y. Corpus-based language studies : an advanced resource book. – Taylor & Francis, 2006. – 386 ð.
26. McEnery T., Hardie A. Corpus linguistics: method, theory and practice / Tony McEnery, Andrew Hardie. – Cambridge : Cambridge University Press, 2012. – 294 p.
27. Meyer Ch. F. English corpus linguistics. An introduction / Charles F. Meyer. – Cambridge : Cambridge University Press, 2002. – 168 p.
28. Meyer Ch. F. Pre-electronic corpora / Ch. F. Meyer // Corpus linguistics: [an international handbook] / [ed A. Lüdeling, M. Kytö]. – Vol. 1. – Berlin : Walter de Gruyter, 2008. – P. 1–14.
29. Ooi V. B. Y. Computer corpus lexicography / Vincent B. Y. Ooi. – Edinburgh : Edinburgh University Press, 1998. – 243 p.
30.Renouf An. Corpus linguistics: past and present / An. Renouf // Corpora in use: in honour of professor Y. Huizhong / [eds W. Naixing, Wenzhong, Li, Pu Jianzhong]. – 2005. – Access Mode :

31. Renouf An. Corpus development 25 years on : from super-corpus to cybercorpus / An. Renouf // Corpus Linguistics 25 Years on / [ed R. Faccinetti]. – Amsterdam – New York : NY, 2007. – P. 27–46.
32. Saldanha G. Principles of corpus linguistics and their application to translation studies research / G. Saldanha // Revista Tradumatica. – No 7. – Barselona : Universitat de Automo di Barselona, 2009. – P. 1–7.
33. Svartvik J. Corpus linguistics 25 + years on / J. Svartvik // Corpus linguistics 25 years on. – Amsterdam – New York : NY, 2007. – P. 11–27.
34. Teubert W., Cermakova A. Corpus linguistics : a short introduction / Wolfgang Teubert, Anna Cermakova. – London : Bloomsbury Academic, 2007. – 153p.
35. Teubert W. Linguistique de corpus: un alternative / W. Teubert // Semen. Les notions de contexte et d'acteurs sociaux /[ed Ad. Petitclerc, Ph. Schepens]. – Vol. 27. – Presses Universitaires de Franche Comté. – Collection Analles Littéraires, 2009. – P. 130–152.
36. Teubert W. Rethinking corpus linguistics / W. Teubert // A mosaic of corpus linguistics : selected approaches / Ed. A. Sánchez, M. Almela. – Frankfurt on Main : Peter Lang, 2010. – P. 19–42.
37. Tognini-Bonelli E. Corpus linguistics at work / Elena Tognini-Bonelli. – Amsterdam : Benjamins, 2001. – 244 p.

Джерела ілюстративного матеріалу

38. Багатомовний паралельний корпус усного мовлення. – К. : КНЛУ, 2010. – Режим доступу : http://www.complinguide.com.ua/Corpus.aspx
39. Корпус текстів Івана Франка. – Львів. – Режим доступу : http://www.ktf.franko.lviv.ua/ ~andrij/science/Franko/concordance.html
40. Корпус текстів української мови кафедри української мови та прикладної лінгвістики Донецького національного університету. – Донецьк. – Режим доступу : http://corpora.pp.ua/bonito/
41. Корпус текстів української мови. – К. : КНУ імені Тараса Шевченка, 2010.– Режим доступу : http://www.mova.info/corpus.aspx?l1=209
42. Навчальний корпус текстів UCLE. – К. : КНЛУ, 2010. – Режим доступу : http://www.complinguide.com.ua/Ucle_index.aspx