КОРПУСНИЙ ПІДХІД У СУЧАСНІЙ ЛІНГВІСТИЦІ: ПЕРСПЕКТИВИ І МОЖЛИВОСТІ ЗАСТОСУВАННЯ

Мейзерська І. В.
Київський національний лінгвістичний університет

У статті висвітлено основні напрями застосування корпусного підходу в сучасній лінгвістиці (науково-теоретичні дослідження, лексикографія, лінгводидактика). Проаналізовано, яким чином у корпусі актуалізується функціонування та вживання окремої лексичної одиниці. Розмежовано поняття "текст" і "корпус". Особливу увагу звернено на застосування анотованих корпусів.
Ключові слова: корпусна лінгвістика, корпус, анотований корпус, тег, текст.

В статье освещены основные направления корпусных исследований в современной лингвистике (научно-теоретические исследования, лексикография, лингводидактика). Показано, каким образом в корпусе актуализируется функционирование и употребление отдельной лексической единицы. Разведены понятия "текст" и "корпус". Особое внимание уделено применению аннотированных корпусов.
Ключевые слова: корпусная лингвистика, корпус, аннотированный корпус, тег, текст.

The article explores the main directions of corpus studies and ways of applying corpora in modern linguistics (research, lexicography, language teaching (ESL). It clarifies the mechanisms used in corpora to show the functioning and real usage of a given lexical unit. Besides, the author distinguishes between the terms "text" and "corpus", paying special attention to annotated corpora as tools for both research and reference.
Key words: corpus linguistics, corpus, annotated corpus, tag, text.

Корпусна лінгвістика як самостійний та самодостатній науковий напрям склалася наприкінці 80-х – початку 90-х років XX століття. Останнім часом корпусно-орієнтовані дослідження стали невід'ємною частиною діяльності лінгвістів. Терміном "корпусна лінгвістика" прийнято називати галузь вивчення мови на основі текстових або акустичних корпусів з інтенсивним залученням комп'ютера на етапах збирання, збереження та аналізу даних [3, с. 62]. Такий підхід передбачає формування та використання корпусів реальних текстів, що можуть застосовуватися як база даних для розв'язання широкого кола теоретичних та практичних завдань.

На думку одного з чільних представників цього напрямку досліджень Джефрі Ліча, ключовими рисами комп'ютерної корпусної лінгвістики є:

1) зосередженість на вживанні мови, а не на компетентності;

2) зосередженість на лінгвістичному описі, а не на лінгвістичних універсаліях;

3) зосередженість на кількісних, а не лише якісних моделях мови;

4) зосередженість радше на емпіричних, ніж раціоналістичних, поглядах на науковий пошук [11, с. 158].

Корпусну лінгвістику можна розглядати одночасно і як науку, і як методологію [6]. Як наука, а саме галузь мовознавчої науки, вона має конкретний об'єкт дослідження (мова в її безпосередньому природному вживанні в усних та писемних висловленнях / текстах) та послуговується суто науковими статистичними методами. Як методологія, вона забезпечує чіткі критерії та принципи збирання, опрацювання, збереження та аналізу даних. При цьому прикметною є та риса, що для корпусних досліджень не можна визначити єдино прийнятого методу або системи принципів, оскільки вони кожного разу добираються залежно від типу корпусу та його призначення, а також відповідно до конкретних завдань та сфери застосування того чи іншого корпусу.

Основними сферами застосування електронних корпусів є:

1) наукові лінгвістичні дослідження в галузях лексичної семантики, граматичної та лексичної сполучуваності, стилістики, прагматики, діалектології тощо. Перевагами застосування корпусу в мовознавчих студіях, на думку дослідника Я. Свартвіка, є те, що він:

– об'єктивний, оскільки мовці часто не можуть дати точний звіт про те, що вони говорять;

– верифіковний;

– корисний у вивченні мовних варіантів, діалектів, стилів, а також в історичних порівняннях;

– встановлює частотність слововжитку;

– є теоретичним ресурсом;

– корисний для машинного перекладу, розпізнавання та синтезу мовлення, а також програм, пов'язаних зі вживанням мови;

– дає більш репрезентативну картину мови, ніж добірки цитат;

– є єдиним способом вивчати слововживання носіїв інших мов, оскільки жодна інша методика не працює;

– той самий корпус можна використовувати для різних потреб [13, с. 7];

2) лексикографія, зокрема створення конкордансів, словників слів та словосполучень на основі одномовних корпусів, а також багатомовних лексиконів і конкордансів із залученням паралельних текстових масивів із різних мов. Наразі, одним із важливих напрямків корпусної лінгвістики є створення автоматизованих лексикографічних систем для укладання ефективних двомовних перекладних словників. Слід зазначити, що лексикографічні студії потребують особливо великих репрезентативних корпусів. Багато слів та словосполучень, представлених у словниках, не є високочастотними, тому, аби уможливити вивчення їхнього вживання, слід мати багатомільйонний корпус, укладений на основі багатьох текстів, представлених різними типами [7, с. 249];

3) лінгводидактика, добір навчального матеріалу та створення ефективних посібників для осіб, що вивчають ту чи іншу мову як іноземну. Із цією метою створюються спеціальні корпуси, які дають змогу проаналізувати типові помилки носіїв різних мов при вивченні певної іноземної мови, зокрема, англійської. Одним із таких інструментів є, наприклад, Російський навчальний корпус англійської мови (Russian Learner Corpus of English), що становить частину Міжнародного навчального корпусу англійської мови (International Corpus of Learner English, ICLE). Цей корпус місить есе, написані російськими студентами англійською мовою і оформлені як електронна база даних з метою подальшого аналізу типових помилок.

Усі корпуси поділяються на дві великі групи: першу формують такі, які містять "чисті" тексти; другу – такі, які містять тексти анотовані, тобто спеціально структуровані та підготовлені таким чином, аби забезпечити досліднику можливість опрацьовувати потрібний саме йому матеріал [1, с. 525]. Іншими словами, анотовані корпуси містять позатекстову лінгвістичну інформацію, представлену спеціальними *тегами* (маркерами на позначення релевантної інформації про частиномовні, граматичні, стилістичні, прагматичні та інші особливості тієї чи іншої лексичної одиниці). Для тегування, або ж маркування тексту, прийнято використовувати єдину Стандартну узагальнену маркувальну мову (СУММ) (Standardized Generalized Markup Language, SGML) [5, с. 329].

Крім того, за типом та призначенням виділяють: статичні та динамічні корпуси (залежно від можливості їх поповнення новими даними); одно-, дво- та багатомовні; навчальні корпуси (створені спеціально для потреб іноземців, які вивчають мову); паралельні (на матеріалі різних мов); порівняльні (тексти певної тематики та проблематики різними мовами); узгоджені (aligned) – корпуси з підрядковим перекладом, що укладаються на основі порівняльних.

Із дослідницькою метою найзручнішим є використання анотованих корпусів, оскільки вони є потужним інструментом для аналізу лексичних одиниць та різноманітних синтаксичних

конструкцій у їх реальному функціонуванні в мовній практиці. Зокрема, така організація корпусу дозволяє відстежувати варіативні граматичні моделі та одержувати швидкий доступ до ілюстративних прикладів, одержаних із реальних текстів.

Скажімо, в англійському підрядному реченні прийменник може або передувати займеннику (wh-clause), або ж відділятися від нього і переміщуватися в кінець речення (preposition stranding):

a) I want a data source [*on which*] *I can rely;*

b) I want a data source [*which*] *I can rely on.*

В українській мові дещо подібним явищем є випадки "редукції субстантива" [4] після прийменника, коли все семантичне навантаження падає саме на нього: *Ти питимеш чай* ***з*** *цукром чи* ***без****?; – Ліки приймати* ***до*** *їди? – Бажано* ***після****.*

У Міжнародному корпусі англійської мови (International Corpus of English, ICE) ця граматична особливість відображена за допомогою спеціальних тегів:

I very much enjoyed the work that I was involved in <PS; PREP> [10, с. 16]

У цьому прикладі слово *in* марковане тегом PREP, що вказує на його частиномовну належність – прийменник. Крім того, оскільки цей прийменник не має фіксованої позиції і здатен переміщуватися в реченні, вживаючись окремо від займенника, йому присвоєно тег PS (stranded preposition). За допомогою цього тегу можна отримати доступ до всіх інших випадків уживання нефіксованих прийменників, які трапляються в корпусі, що значно розширює можливості аналізу.

Об'єктивність корпусних даних забезпечується широкою електронною базою усних та писемних текстів найрізноманітніших жанрів та тематик, здатних відображати максимальну кількість слововживань в усіх стилях. Наведемо схему текстових категорій, представлених у Міжнародному корпусі англійської мови (International Corpus of English, ICE):

Усні	діалоги	приватні	особисті розмови
			телефонні розмови
		публічні	класні уроки
			обговорення в ефірі
			інтерв'ю в ефірі
			парламентські дебати
			перехресні допити
			бізнес-переговори
	монологи	імпровізовані	спонтанні коментарі
			непідготовлені промови
			демонстрації
			презентації
		підготовлені	промови в ефірі
			новини
Писемні	недруковані	непрофесійне письмо	студентські есе
			студентські письмові екзаменаційні відповіді
		кореспонденція	соціальне листування
			ділове листування
	друковані	наукові праці	гуманітарні науки
			соціальні науки
			природничі науки
			технічні науки

		науково-популярні праці	гуманітарні науки
			соціальні науки
			природничі науки
			технічні науки
		репортажі	огляди новин у пресі
		пояснення	адміністративні / правознавчі
		інструкції	хобі / вміння та навички
		публіцистика	передові статті / колонки
		художня література	романи / оповідання

Виходячи з різноманіття представлених текстових категорій, можемо зробити висновок, що корпусні дані на відміну від традиційних лексикографічних джерел дають можливість проаналізувати безпосереднє функціонування лексичної одиниці в різних стилях, комунікативних ситуаціях тощо. Зокрема, корпусний аналіз найближчого лексичного та граматичного оточення слова дозволяє відстежити його вживання в усіх характерних для нього:

1) колокаціях, або ж сталих словосполученнях, що їх також трактують як "сталі, повторювані комбінації слів" [8, с. 277], "передбачувані комбінації окремих (індивідуальних) слів" [12, с. 93]. Сполучуваність у колокаціях може варіюватися від відносно вільної (наприклад, keep: *keep house / a diary / a shop / a hotel / pets / a boat / to diet* etc.; go: *go grey / pale / sour / bankrupt / mad* etc.) до зв'язаної: *stark naked, gin and tonic, white coffee, wide awake, safe and sound, to and fro* etc.;

2) колігаціях, які, на думку М. Хоуві можна визначити як "граматичного компаньйона слова та позицію, яка є для нього бажаною; іншими словами, колігації слова вказують на те, яким чином воно "поводиться" з погляду граматики" [9, с. 226]. Як ілюстрацію можна навести диференційоване вживання *much* або *many* залежно від граматичного класу, до якого належить іменник (обчислювані / необчислювані іменники), або ж уживання форм герундія чи інфінітива після певних дієслів (*He offered to help / He suggested going for a walk*);

3) синтаксемах, які, за Г. А. Золотовою, є мінімальними, далі неподільними семантико-синтаксичними одиницями мови, що виступають одночасно як носії елементарного смислу і як конструктивної компоненти більш складних синтаксичних утворень [2].

І, нарешті, слід наголосити на тому, що корпус у жодному разі не можна трактувати як текст або як просту сукупність текстів. Це складна цілісна система, що має свою логічну структурну організацію та диференційні ознаки, для якої текстові принципи аналізу недійсні. Корпус укладається на основі великої репрезентативної групи текстів, оброблених таким чином, що мовний матеріал розміщується за принципом конкордансу. Наведемо фрагмент корпусу зі словом *rabbit* на прикладі електронної бази даних Bank of English словників серії Collins:

*of a girl with your qualifications. He **rabbited** on like that for half a minute,*
*Chronicle History of King Leir, who **rabbited** on about Christian piety, was simply*
*with a male friend the other day as he **rabbited** on obsessively about the beautiful new*
*While his colleagues **rabbited** on about guns and conspiracies, he*
*wished to put up with said driver's **rabbiting**, I mean has the world gone mad, this is*
*owner and non-**rabbiting** ferret-owner, I'm deeply familiar with*
*mind what all these silly critics go on **rabbiting** about, it's one of the world's best. He*
*did not have a clue what he was **rabbiting** on about. < p > Racing has paid too big a*

Ліве й праве оточення ключового слова становлять контекст, який можна ввести до словникової статті з метою ілюстрування значення слова [3, с. 61]. Типовий конкорданс дає дослідникові змогу знайти будь-яке слово і висвітлює його, як показано вище, у форматі КСВК (Ключове слово в контексті) (Key Word in Context, KWIC), що складається з окремих рядків із виділеними

ключовим словом посередині кожного рядка. Формат КСВК звичайно можна також перетворити на формат показу речень чи цілих абзаців. Деякі системи забезпечують змогу вивести на екран повний зразок тексту, оскільки деякі абзаци, зокрема в діалогах, можуть бути дуже короткими, тому треба побачити більше, щоб зрозуміти контекст [5, с. 329].

Проаналізувавши структуру вищенаведеного фрагменту корпусу, можемо виділити його диференційні ознаки порівняно з текстом. Для будь-якого тексту характерними є структурна, смислова цілісність, а також наявність певної комунікативної мети. Але якщо цей текст використовується як матеріал корпусу, ці ознаки не беруться до уваги. У процесі корпусного аналізу текст маркується і розбивається на фрагменти, що ілюструють контекст, найближче лексичне та граматичне оточення тієї чи іншої мовної одиниці. Ці фрагменти групуються в конкорданс, тобто вибірку всіх можливих контекстів, засвідчених у репрезентативній групі текстів.

На відміну від лінійної, горизонтальної організації тексту, для корпусу характерним є вертикальний принцип упорядкування мовного матеріалу. Контексти, представлені в конкордансі, засвідчують функціонування мовної одиниці в різних стилях, сферах уживання, даючи широку картину реальної мовної практики. Для тексту ж здебільшого характерною є певна жанрова та стильова специфіка, а також авторська індивідуальність, яка знаходить своє відображення в ідіостилі.

І нарешті, оскільки корпус подає мовний матеріал у фрагментованому вигляді, то не йдеться про притаманні тексту зв'язність та загальну комунікативну мету. Корпус позбавлений зв'язності, а комунікативну мету можна простежити лише на рівні окремо взятого речення, але не на рівні тексту. Наочно проілюструвати основні відмінності між власне текстом та електронним текстовим корпусом можемо за допомогою такої таблиці:

ознаки тексту	**ознаки корпусу**
1. структурна цілісність	1. фрагментованість
2. лінійність, горизонтальна організація	2. нелінійність, вертикальна організація
3. смислова цілісність	3. наведення конкретних контекстів та мовних структур
4. авторська індивідуальність	4. зразки загальної мовної практики
5. зв'язність	5. відсутність зв'язності
6. наявність загальної комунікативної мети на текстовому рівні	6. комунікативна мета простежується лише на рівні окремого речення

Література

1. Гладкова А. П. Модель анотації текстового корпусу як засіб дослідження художньої картини світу // Studia Linguistica. – Вип.4. – К. : КНУ, 2010. – С. 524–528.
2. Золотова Г. А. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса / Г. А. Золотова. – М. : Наука, 1988. – 286 с.
3. Карпова О. М. Английская лексикография : учебное пособие / О. М. Карпова. – М. : Академия, 2010. – 176 с.
4. Конюшкевич М. Предлог как синтаксемообразующий формант и структура синтаксемы / М. Конюшкевич // Лінгвістичні студії. – Донецьк, 2006. – Вип. 14. – С. 73–79.
5. Лендау С. І. Словники: мистецтво та ремесло лексикографії / С. І. Ландау / [пер. з англ.]. – К. : К.І.С., 2012. – 480 с.
6. Baker P. A Glossary of Corpus Linguistics / Paul Baker, Andrew Hardie, Tony McEnery. – Edinburgh: Edinburgh University Press, 2006. – 187 p.
7. Biber. D. Corpus Linguistics: Investigating Language Structure and Use / [Susan Conrad, Randi Reppen]. – Cambridge : CUP, 1998. – 214 p.

8. Clear J. From Firth Principles. Computational Tools for the Study of Collocation / Baker, Mona et al. (eds.), Text and Technology. In Honour of John Sinclair. – Philadelphia, Amsterdam : John Benjamins Pub. Co., 1993. P. 271 – 292.
9. Hoey M. A world beyond collocation: New perspectives on vocabulary teaching. // Lewis M. (Ed.) Teaching collocations: Further developments in the lexical approach. 2000 – P. 224–245.
10. Hoffman T. Preposition Placement in English. T. Hoffman – Cambridge, 2011. – 297 p.
11. Leech G. Principles and Applications of Corpus Linguistics / G. Leech // Perspectives on Corpus Linguistics / [ed. by V. Viana, S. Zyngier and G. Barnbrook.] – Amsterdam : John Benjamins Publishing Company, 2011. – P. 155 – 170.
12. Lewis M. Implementing the lexical approach: Putting theory into practice. – Hove, England : Language Teaching Publications. – 1993. – 181 p.
13. Svartvik J. Corpus Linguistics Comes of Age // Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82. – Berlin : Mouton de Gruyter, 1992. – P. 5–21.