УДК 81-13:811. 111+81'374

Дубравська Д. М.,
Львівська комерційна академія

ВНЕСОК КОРПУСНОЇ ЛІНГВІСТИКИ У СУЧАСНУ ЛЕКСИКОГРАФІЮ

"Працюючи з корпусом ми відчували потік творчої енергії, яка надихала нас своєю придатністю до застосування, витонченістю та гнучкістю. "

Джон Сінклер[1]

У статті розглянуто поняття корпусу та роль репрезентативного корпусу в кволітативному та квантитативному аналізі лінгвістичних феноменів; розкрито актуальність корпусної лінгвістики та подано результати лінгвістичного дослідження частотності конвертованих іменників в дієслово та композитів із лексемою key (на матеріалах Британського національного корпусу (BNC).

Ключові слова: *корпус, корпусна лінгвістика, корпусні дані, композити, конверсиви, лексикографія.*

В статье рассмотрено понятие корпуса и роль репрезентативного корпуса в кволитативном и квантитативном анализе лингвистических феноменов; раскрыто актуальность корпусной лингвистики и представлены результаты лингвистического исследования частотности конвертированных имён существительных в глагол и композитов с лексемой key (на материалах Британского национального корпуса (BNC).

Ключові слова: *корпус, корпусная лингвистика, корпусные данные, композиты, конверсивы, лексикография.*

The article examines the notion of the corpus as the new methodology for linguists. The role of a representative corpus in qualitative and quantitative analysis of linguistic phenomena is discussed. The importance of the corpus linguistics is revealed. The importance of copora to language study is aligned to the importance of empirical data. Empirical data enable the linguist to make objective statements, rather than those which are subjective, or based upon the individual's own internalised cognitive perception of language. The results of the linguistic research of the parent nouns denoting instrument which are converted into verbs and the list of frequency of the composits with lexeme key are given (based on the British National Corpus (BNC).

Key words: *corpus, corpus linguistics, corpus data, cjmposits, conversion, lexicography.*

Корпусна лінгвістика впродовж останніх десятиріч сформувалась в самостійний науковий напрямок, досягнення якого ознаменували новий етап в розвитку наукової думки. На сьогодні корпусна лінгвістика вважається відносно новим підходом, що має справу з емпіричним дослідженням "реального життя" мови, яке здійснюється за допомогою комп'ютера та електронних корпусів. Корпусна лінгвістика – займається створенням, обробкою та використанням корпусів. Праці провідних представників комп'ютерної та корпусної лінгвістики висвітлюють кардинальні проблеми організації корпусів текстів та застосування результатів їх аналізу в лінгвістичних дослідженнях [1; 6; 9; 12;13].

На основі корпусів текстів граматичні моделі досліджували С. Ханстон та В. Френсіс [7; 8], фразеологію Дж. Сінклер, М. Стабс, М. Холідей, В. Тойберт [5;12; 14; 15; 16], лексичні 'кластери' та їхні функції в різних жанрах Д. Байбер, Р. Реппен, Р. Сімпсон, Дж. Свейлс [2; 11], та семантичну просодію Д. Сінклер, Д. Байбер [2; 12]тощо.

Ця галузь лінгвістики зараз є дуже актуальною, хоча ще 15-20 років тому була захопленням невеликої кількості людей. Термін корпусна лінгвістика виник на початку 1980-х [9, с. 105] і вживається для опису збірки мовних моделей чи цілих текстів: ними можуть слугувати слова, речення, розмови, журнальні та газетні тексти, чи цілі тексти книг. Теоретично корпус здатний представляти потенційно необмежену систематизовану вибірку текстів.

У науковому лінгвістичному словнику з'явилися дуже близькі поняття: "електронні бібліотеки", "масив текстів", "колекція текстів", "електронний архів", "повна текстова база даних" тощо. Серед них можна виділити лінгвістичні корпуси, або мовні корпуси. Варто зауважити, що праця з електронними корпусами давно вже стала одним з основних методів лінгвістичних досліджень.

Аналіз наукових джерел свідчить про те, що лексикографи використовували емпіричні дані ще задовго до появи дисципліни корпусної лінгвістики. Однак, цей процес був тривалий і громіздкий, дані досліджень не були репрезентативними (ґрунтувалися на даних спостереження).

Першим сучасним корпусом текстів був Браунівський корпус текстів на машинних носіях (Brown corpus 1960 – for American English). Його автори У. Френсіс та Г. Кучера спроектували його як набір прозаїчних друкованих текстів американського варіанту англійської мови (1 млн. слів). Його автори під словом "корпус" розуміли "сукупність текстів, яка вважається репрезентативною для даної мови, діалекту … призначена для лінгвістичного аналізу" [4].

Браунівський корпус швидко перетворився в популярний об'єкт дослідження та навіть в певний стандарт для створення інших аналогічних корпусів. Аналогами Браунівського корпусу були LOBC (Ланкастер-Осло/Берген) (1978 рік – 1 млн. слів), London-Lund Corpus (LLC – 1987 рік, 1 млн. слів), Frown Corpus (Freiburg-Brown Corpus of American English) (1992 рік, 1 млн. слів).

У 1990-і роки створений British National Corpus (BNC, 100 млн. слів). В той же час був створений The Bank of English, Birmingham (600 млн. слів). На початку XXI ст. були створені такі корпуси, як American National Corpus (100 млн. слів) та Gigaword corpora (англійський, арабський, китайський – 1 млрд. слів) тощо.

Британський національний стандартизований корпус текстів вважається як таким, що репрезентує сучасну британську англійську мову в цілому, а тому відноситься до загальних корпусів. Містить різні жанри, насичений, збалан-

[1] John Sinclair, Introduction to How to Use Corpora in Language Teaching, Amsterdam: Benjamins, 2004, p. 1

© Дубравська Д. М., 2012

сований (містить в собі писемне і усне мовлення). Писемні тексти підібрані за трьома критеріями: змістовим(галузь), часовим (період появи) та носієм інформації (типи публікації текстів – книги, газети, журнали тощо).

Хоча існує багато визначень корпусу [4; 12; 13; 17] – однак, існують спільні погляди на те, що корпус – це комп’ютеризовані автентичні тексти, підібрані і упорядковані згідно з експліцитними критеріями, визначеними користувачами, вони є репрезентативними зразками певної мови чи мовних варіантів.

Отже, корпусом може називатися репрезентативна вибірка текстів в електронній формі, доступ до якої забезпечується ретельно розробленими дослідницькими комп’ютерними програмами пошуку та аналізу.

Організація корпусу може бути різна – в залежності від прагматичних цілей його творця чи користувача. Хоча створений корпус може використовуватися і для цілей, які не були передбачені автором. Тексти, складові елементи корпусу, можуть бути цілісними оригінальними словесними творами чи їх частинами.

Корпуси діляться на одномовні та багатомовні, паралельні та порівняльні (перекладацькі корпуси), загальні і спеціалізовані, діахронні та синхронні.

Можна навіть стверджувати, що корпуси здійснили революцію в лексикографії. Послідовно застосовуючи принцип комп’ютерної обробки реального мовного матеріалу, що використовується у відповідних сферах комунікації, був сформований принципово новий тип словників [12]. Усі сучасні словники англійської мови створюються на базі корпусів, серед них Oxford, Collins, Longman, Cambridge, Macmillan тощо.

Лінгвістика зазнала значних змін за останні десятиріччя завдяки наявності електронних корпусів, які полегшили аналіз мови як феномену. Для лексикології проведення досліджень з використанням електронних корпусів стало загальноприйнятим стандартом. На цьому фоні перспективними є лексикографічні дослідження, з опорою на дані корпусного аналізу.

На думку багатьох вчених при підборі корпусу для дослідження варто зважати на характерні риси корпусу: кількість (велика вибірка), якість (автентичність), репрезентативність (корпус повинен містити різноманітні зразки з широкого ряду текстів) збалансованість (повинен з максимальною об’єктивністю представляти існування певного феномену в мовній практиці носіїв даної мови), простота формату, можливість верифікації даних тощо.

Отже, лексикографічні дослідження, проведені на основі корпусів будуть достовірними, якщо корпус текстів достатньо великий і варіативний – таким корпусом є *британський національний корпус* (BNC), на основі якого ґрунтуються наші дослідження.

Комп’ютерний аналіз великої кількості текстів дає нам можливість осягнути те, що невидимо неозброєним оком. Програма конкордансів швидко складе список частотності в алфавітному порядку, тому легше здійснити квантитативний та квалітативний аналіз різних лінгвістичних феноменів:

Text model 1

Instrument verbs. The parent noun denotes instrument which is incorporated into a verb.				
Composite	**Total**	**Written**	**Transcribed speech**	**Sample**
to vote	1154	978	76	The next day, of course, they got up and went out to vote.
to hand	1112	1043	69	He pulled out his billfold and tried to hand it to me.
to trace	518	503	15	It took police four days to trace the couple to Bridgnorth in Ontario.
to block	446	422	24	Syria is unlikely to block an attack on Iraq, even if it does not join it.
to lock	233	214	19	They try to lock them out.
to mop	60	53	7	Luke began to mop clumsily at the pool of milk with a teacloth.
to claw	43	39	4	He warned that conservatives were trying to claw back power.
to saw	40	29	11	Did you have to saw this stick?
to shower	37	35	2	She wanted to shower, to change, to relax for a few minutes.
to pipe	22	22	0	It would have been too expensive in electricity to pipe hot water from the house.
to glue	18	18	0	It is a difficult wood to glue, but it takes stain and polishes to an excellent finish when filled.
to button	17	17	0	Miss Ellis stood up and began to button her coat.
to chalk	14	12	2	He had given us the computer codes to chalk on the outside of the boxes…
to strap	14	13	1	He took one of the two pilot seats and began to strap himself in.
to sponge	5	4	1	I did that, and then I went to sponge his jacket.

Корпусні дані – систематизовані і чіткі. З появою корпусів текстів з’явилася можливість використовувати корпусні дані для верифікації та перевірки результатів досліджень отриманих традиційним шляхом. Комп’ютерна лінгвістика змінила спосіб в який лексикографи можуть вивчати мову як феномен, використовуючи готові дані різних періодів. Корпуси містять дані для дескриптивних та теоретичних досліджень і дають можливість вносити радикальні зміни в лінгвістичну теорію на основі очевидних фактів.

Емпіричні дані корпусів дають можливість лінгвістам робити об'єктивні висновки, ґрунтуючись на великій кількості зібраного матеріалу, а не на індивідуальних суб'єктивному сприйманню мови. Матеріали, зібрані в корпусах, це здебільшого зразки мови, яка використовується у реальному житті (зібрано регіональні та соціальні акценти), вони охоплюють різні періоди, жанри тощо.

Д. Літч розглядає корпусну лінгвістику як 'новий філософський підхід' [9, с. 106]. Багато інших вчених і ми в тому ж числі вважаємо її методологією, яка дає нам змогу здійснити всебічний аналіз феномену, який ми вивчаємо. Перш за все це стосується частотності, тексти корпусу – дистрибутивна (або квантитативна ⁄ статистична) інформація: наявність лінгвістичних елементів, наприклад, чи засвідчені в корпусі конверсиви, композити, деривати, фразеологізми і як часто вони вживаються в текстах корпусу; ця інформація представлена у так званому реєстрі частотності в конкордансах у відповідному контексті. Великий розмір корпусу BNC, репрезентативність, наскрізний комплексний аналіз – уможливив нам дослідити таке явище словотворення, як композити.

Так серед 100 мільйонів слів композит *weak tea* зустрічається 16 разів, що підтверджує стійкий зв'язок між компонентами композита. Іншим композитам з лексемою *weak (link. (32), form (30), position (30), spot (24), smile (23) etc.)* притаманна навіть вища частотність, що вказує на статистичну значущість феномену.

Композити – ефективний шлях створення і вираження нових значень. Часто можна вивести значення композита з його складових. Однак, існує доволі велика кількість композитів, зміст яких зовсім не можна взнати з перекладу його складових. При перекладі таких композитів ми орієнтуємося на оточення даного композита в реченні. Для прикладу розглянемо композити із лексемою key:

Text model 2

Composite	Total	Written	Transcribed speech	Sample
keyboard	949	880	69	The keyboard player obviously cares more about advancing his or her career than the future of the band.
keyword	404	403	1	These users are listed at the USER ACCESS keyword.
keynote	167	159	8	The keynote lecture will be given by Prof Violetta N. Fomenko (Ministry of Health, Moscow).
keyhole	125	110	15	'The key was in the keyhole on the inside. '
keystone	83	80	3	That approach would seek to knock out a keystone of Civil Service tradition.
keypad	43	42	1	Use the arrow keys on the numeric keypad to move around
keyboarding	23	21	2	If keyboarding of the two texts could not be run in parallel…
keystroke	18	18	0	A cross-indexing system locates related comments with a single keystroke.
keyboardist	11	11	0	BASSIST singer, keyboardist required to form band with two guitars and drummer.
keyring	9	7	2	That's a good keyring really.
passkey	3	3	0	The night porter used a passkey for those rooms that were empty or where no-one answered.
keyboarder	2	2	0	This procedure is entirely straightforward if one can assume in the keyboarder competence in the handling of the English alphabet...
keypunch	1	1	0	Programs and data were, at that time, punched onto cards using a keypunch.

Отже, виникає необхідність глибокого аналізу композитів в розрізі морфології, семантики та граматики.

Хоча місце корпусів в лінгвістиці є контраверсійним, нині як ніколи виникає широкий інтерес щодо їх використання у лінгвістичних дослідженнях. Це зумовлено тим, що лінгвісти більш зацікавлені у практичному використанні мови, ані ж у мові як абстрактній системі. Безсумнівно, що такий глибокий інтерес появився внаслідок появи і розвитку корпусних методик, які дозволяють швидко і точно аналізувати такі корпуси різні виміри.

Корпусна лінгвістика не лише запропонувала нові методи для лексикографії, але й розширила межі її дослідження. Ініціатором в цьому був Джон Сінклер, який уклав словник англійської мови (COBUILD 1987), що цілком базувався на корпусі.

Корпусна лінгвістика розширила наші лінгвістичні знання завдяки вдалому поєднанню трьох різних підходів: ідентифікація мовних фактів за допомогою категоріального аналізу (процедурний підхід); кореляція фактів за допомогою статистичних методів (математичний підхід); і інтерпретація результатів (когнітивний підхід). Якщо перших два підходи автоматизовані, останній потребує людського інтелекту, оскільки інтерпретація даних потребує людської свідомості. Звичайно інтерпретація даних – це завдання людини, а не корпусу текстів. "… корпуси безсумнівно цінні джерела для лінгвістичного аналізу …" [17, с. 144].

Корпусна лінгвістика враховує, що мова – соціальний феномен. Більшість текстів корпусів – комунікативна діяльність в процесі взаємодії членів певної мовної спільноти. Корпус віртуально поєднує в собі комунікативну

діяльність будь-якої відібраної мовної спільноти (у нашому випадку англомовної). Однак, ми не знаємо як саме мовець чи слухач розуміє слова, речення і тексти, які він промовляє чи чує, нам відомо лише те, що є в тексті корпусу тобто в контексті). Саме в контексті визначається конкретне значення слова. Корпусна лінгвістика дає нам змогу не лише виявити зв’язки між елементами тексту та контекстом, а також сегментами тексту.

Корпусна лінгвістика – доповнення до традиційної лінгвістики, але, на відміну від традиційної, корпусна лінгвістика використовує корпуси не лише для вивчення окремих абстрагованих прикладів, але й для систематичного дослідження феномена в контексті, оскільки вона більш зацікавлена в семантичному зв’язку між елементами тексту і контексту, що можна виявити за допомогою квантитативного аналізу дискурсу чи корпусу.

Дослідження словникового запасу найчастіше зустрічається в корпусній лінгвістиці. Корпуси дають змогу отримати дані по лексемі в цілому (пошук за лемою) та по конкретній словоформі, виявити типові/нетипові вживання та характерні сполучення слів. Ці дані можуть бути різними: контексти, частоти (абсолютні та відносні), частоти за колокаціями, статистика по жанрах/стилях/авторах, тощо. Саме корпусна лінгвістика дає змогу справитися з таким динамічним аспектом мови як словотворення. Корпус дає можливість продемонструвати різні способи вживання лексики в певному контексті.

Джон Ферс наголошував на важливості аналізу оточення, в якому перебуває слово [3]. Саме тому класифікація слів лінгвістами здійснюється не лише за їхнім значенням, але й на основі співіснування з іншими словами.

Корпус, як правило, забезпечує нас емпіричним матеріалом для підтвердження гіпотези; інформацією про частотність вживання слів, фраз чи конструкцій, які можна використати з метою квонтитативних досліджень; та метаінформацією (екстралінгвістичною інформацією) – про такі фактори як вік, рід мовця чи автора, жанр тексту, часові та просторові параметри походження тексту тощо. Така екстралінгвістична інформація дає змогу порівняти різні типи текстів чи різні групи мовців.

Для проведення ефективного корпусного аналізу необхідна чітка методика та інструментарій, оскільки аналіз корпусів текстів включає в себе більше, ніж простий підрахунок лінгвістичних характеристик. Необхідно переконатися, що вибраний вами корпус містить потрібну кількість текстів відповідної довжини для аналізу, саме тоді результати будуть валідними, а висновки матимуть надійне підґрунтя.

Сучасна лексикографія аналізує і використовує методи та можливості корпусної лінгвістики, яка в свою чергу освоює методи та можливості лексикографії. Оскільки це відносно новий підхід до вивчення лінгвістичних явищ, корпус та його інструментарій постійно оновлюється і зазнає якісних змін.

Література:

1. Atkins S., Clear J. and Ostler N. ‘Corpus design criteria’. / S. Atkins, J. Clear, N. Ostler. // Literary and LingComputing 7/1, 2005. – P. 1-16.

Biber D., Conrad S., Reppen R. Corpus linguistics: investigating language structure and use. / Douglas Biber, Susan Conrad, Randi Reppen. Cambridge University Press, 1998. – 300 p.

2. Firth, J. R. A synopsis of linguistic theory. Studies in linguistic analysis. / J. R. Firth, Oxford : Oxford University Press, 1957. – P. 179.

3. Frequency analysis of English usage: lexicon and grammar / by Henry Kuçera and W. Nelson Francis. – Boston : Houghton Mifflin, 1982. – pp. 3-15.

4. Halliday, M. A. K. Lexis as a linguistic level. // In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), In Memory of J. R. Firth. London : Longman, 1966. – pp. 148-162.

5. Hunston S. Corpora in applied linguistics. / Susan Hunston. Cambridge : Cambridge University Press, 2002. – 241p.

6. Hunston S., Francis G. and Manning E. “Grammar and Vocabulary : Showing the Connections”. / S. Hunston, G. Francis and E. Manning. // ELT Journal, 51(3), 1997. – pp. 8-216.

7. Hunston S. and Francis G. Pattern Grammar : A Corpus-driven Approach to the Lexical Grammar of English. / S. Hunston and G. Francis Amsterdam : Benjamins, 2000. – 288 p.

8. Leech Geoffrey N. Corpora and theories of linguistic performance. Directions in corpus linguistics. Proceedings of Nobel Symposium 82, ed. by Jan Svartvik. / Geoffrey N Leech. Berlin, New York : Mouton de Gruyter, 1992. – P. 105-122.

9. Leech Geoffrey N. “Grammars of Spoken English : New Outcomes of Corpus-Oriented Research”. / Leech Geoffrey N. // Language Learning 50(4), 2000. – P. 675-724.

Simpson R. C., Swales J. Corpus linguistics in North America: selections from the 1999 symposium. / Rita C. Simpson, John Swales. – University of Michigan Press, 2001, – 241 p.

10. Sinclair J. M. Corpora for lexicography. In P. van Sterkenberg (Ed.), A Practical Guide to Lexicography. / John Sinclair – Amsterdam : Benjamins, 2003. – pp. 167-178.

11. Sinclair J. M. Introduction to How to Use Corpora in Language Teaching // John Sinclair. – Amsterdam : Benjamins, 2004. – 307 p.

12. Stubbs M. Text and Corpus Analysis. / Stubbs Michael. – Oxford, Blackwell, 1996. – 267 p.

13. Stubbs M. Language corpora. // In A Davies & C Elder eds Handbook of Applied Linguistics. / Stubbs Michael. – Oxford : Blackwell, 2004. – P. 106-132.

14. Teubert, W. (2004) Units of meaning, parallel corpora, and their implications for language teaching. In Connor, U. and Upton, T. A. (2004) Applied Linguistics : A Multidimensional Perspective. Amsterdam : Rodopi, 171-189.

15. McEnery T., Xiao R and Tono Y. Corpus-based language studies : an advanced resource book. Routledge Applied Linguistics. / Tony McEnery, Richard Xiao and Yukio Tono. – New York, 2006. – 389 p.

16. BNC : British National Corpus, URL (22. 02. 2007), http://www.natcorp.ox.ac.uk/.