УДК 811’112.2

Л. В. Щигло,
Сумський державний університет, м. Суми

КОРПУСОЗОРІЄНТОВАНЕ ДОСЛІДЖЕННЯ МОВНОЇ СИСТЕМИ: ОСНОВНІ ВЕКТОРИ

Сучасні інформаційні технології знаходять широке застосування у науковій парадигмі лінгвістики. Методи і підходи корпусної лінгвістики не просто дозволяють прискорити дослідження мови, але й підвищити ефективність, достовірність та перевірку результатів обробки даних. Висвітлюються аргументи стосовно переваг та недоліків корпусних методів до студіювання лінгвістичних явищ.

Ключові слова: *корпусне дослідження, квантитативний аналіз, методи.*

CORPUS-ORIENTED RESEARCH OF LANGUAGE SYSTEM: BASIC VEKTORS

Modern information technologies are getting widely used in the science paradigm of linguistics. The article deals with observing the issue of language computer-corpus research as a new and promising tendency in linguistic science paradigm being up-to-date rather than temporary. Methods and approaches of corpus linguistics not only allow to speed up researching language phenomena but also help to reach both significant effectiveness and authenticity and a higher level of checking processed-data results. They make it possible to settle those issues which the linguistics of the previous century did not determine and research because of voluminousness of their completing. The arguments concerning advantages and disadvantages of corpus methods and approaches to studying linguistic phenomena are covered as well. One also determines aspects that prove practical importance and future advanced character of corpus methodology for settling linguistic problems.

Key words: *corpus research, quantitative analysis, methods.*

КОРПУСНО-ОРИЕНТИРОВАННОЕ ИССЛЕДОВАНИЕ ЯЗЫКОВОЙ СИСТЕМЫ: ОСНОВНЫЕ ВЕКТОРЫ

Современные информационные технологии находят широкое применение в научной парадигме лингвистики. Методы и подходы корпусной лингвистики не просто позволяют ускорить исследования языка, но и повысить эффективность, достоверность и проверку результатов обработки данных. Освещаются аргументы относительно преимуществ и недостатков корпусных методов для изучения лингвистических явлений.

Ключевые слова: *корпусное исследование, квантитативный анализ, методы.*

Постановка наукової проблеми та її значення. Розвиток науки і техніки відкриває нові можливості для дослідників як технічної сфери діяльності, так і гуманітарної. Сучасні інформаційні технології знаходять широке застосування у лінгвістичних дослідженнях у форматі електронних ресурсів різних типів: електронні словники, бази даних, текстові корпуси, що, в свою чергу, сприяє швидкій та об’єктивній обробці лінгвістичних даних. Отже, слід зазначити, що наявність електронних автоматично оброблюваних лінгвістичних баз даних не тільки значно прискорює й оптимізує об’ємний процес збору мовного матеріалу, але й веде до зміни наукової парадигми в лінгвістиці [1]. В рамках сьогодення такі завдання успішно вирішує корпусна лінгвістика – розділ комп’ютерної лінгвістики, що займається розробкою загальних принципів побудови та використання лінгвістичних корпусів (корпусів текстів) із застосуванням комп’ютерних технологій. Це уможливлює отримання більш швидких результатів стосовно обробки значних масивів текстів, що до тепер займало досить багато часу. Корпус не просто дозволяє прискорити дослідження мови і багаторазово підвищити ефективність, достовірність та перевірку результатів обробки даних – він дозволяє вирішувати такі завдання, які лінгвістика попередніх століть практично не ставила через об’ємність їх виконання. До таких завдань відноситься, наприклад, багато видів статистичних та інших квантитативних досліджень мови. До того ж корпусна лінгвістика не тільки квантитативний і статистичний інструмент, але й своєрідна «стратегія, методологія дослідження» [2, с. 19]. Необхідність в об’єктивних кількісних даних, потреба великого масиву прикладів, а також потреба широкої «географії» джерел передбачає і зовсім інший методологічний підхід до вирішення завдань, що й зумовлює актуальність нашого дослідження.

Корпусна лінгвістика виходить з того, що дослідник займає, з одного боку, позицію стороннього спостерігача стосовно мовних явищ, з іншого боку, довільно задає параметри для вибірки та аналізу даних корпусу, тобто корпусна лінгвістика об’єднує в собі теоретичні та емпіричні принципи лінгвістики. **Об’єкт статті** – доробки сучасних вітчизняних та зарубіжних науковців стосовно застосування корпусної методології в сучасній лінгвістиці. **Предмет дослідження** – огляд напрацювань сучасних вітчизняних та зарубіжних науковців, в яких висвітлюються проблеми дослідження лінгвістичних явищ в ракурсі методів і підходів корпусної лінгвістики. **Мета роботи** – обґрунтування актуальності застосування корпусної методології для вирішення лінгвістичних проблем.

Виконання поставленої мети передбачає вирішення таких **завдань:**

– з’ясувати роль та місце корпусної методології в сучасній науковій парадигмі лінгвістичної науки;

– висвітлити аргументи стосовно переваг та недоліків корпусних методів до студіювання мовних явищ;

– окреслити аспекти, що підтверджують значущість та перспективність корпусних методів дослідження для вирішення лінгвістичних проблем.

Виклад основного матеріалу й обґрунтування отриманих результатів дослідження. Визначення терміну «корпус» є неоднозначним і досить різноплановим. Корпус (текст, масив):

– приблизна сукупність висловлювань, відібраних для аналізу і представлених у вигляді письмового тексту, аудіо запису, тощо;

– вся сума (сукупність) творів мовлення, створених колективом носіїв певної мови [3, с. 209]. Під лінгвістичним, або мовним, корпусом текстів розуміється великий, представлений в електронному вигляді, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, призначений для вирішення конкретних лінгвістичних завдань [4, с. 8].

В ареалі корпусної лінгвістики виокремлюють такі типи корпусів: ілюстративний, дослідницький, динамічний, статистичний. Дослідницький корпус призначений переважно для вивчення різноманітних аспектів функціонування мовної системи. Цей тип корпусних даних, зазвичай, зорієнтований на широкий загал лінгвістичних завдань. В якості основних вимог, що пред’являються дослідниками до подібного типу корпусу, виокремлюються такі як: репрезентативність, повнота, економічність, самодостатність, комп’ютерна підтримка, структуризація матеріалу [5, с. 118–119].

Пріоритетною характеристикою корпусу, що відрізняє його від простих збірок текстів, полягає в наявності додаткової інформації про властивості вхідних текстів, що до нього належать (розмітки чи анотації). Кожен текст повинен мати лінгвістичну та екстралінгвістичну розмітку. У інформацію про текст необхідно включити відомості про інформантів, про

час, місце запису, про конкретну ситуацію спілкування, відомості про діалекти, тощо. Метатекстова інформація повинна бути універсальною, типовою для лінгвістичних корпусів різного типу, щоб не обмежувати параметри пошуку, а, навпаки, зробити корпус доступним для багатьох дослідників з їх різною метою, підходами та вихідними гіпотезами.

Об'ємними і повними є корпуси національних мов, наприклад Національний корпус російської мови, Браунський корпус американського варіанту англійської мови, Британський національний корпус та інші. У багатьох країнах ведуться роботи зі створення корпусів за різновидами мови (корпус діалектів, усної, письмової мови, корпус смс-повідомлень, дитячої мови, публіцистичних текстів, тощо [6, С. 113–123]. Обширною інформаційної системою є корпус розмовної німецької мови (Datenbank Gesprochenes Deutsch (DGD des DSav)), що розроблюється Мангеймським інститутом німецької мови.

В європейських корпусах розмовного та діалектного мовлення використовують систему EXMARaLDA (Extensible Markup Language for Discourse Annotation), тобто розширену марковану систему лінгвістичної анотації розмовної мови. Це система програм та інструментів для створення, управління, анотування та обробки корпусу розмовного мовлення. Базовою програмою для первинного створення корпусу текстів і їх анотування є Partitur Editor, назва якої вже сама визначає тип введення інформації: партитурна нотація. На відміну від так званої драматургічної нотації, яка передбачає вертикальне розташування тексту, партитурна нотація, вважається більш вдалою, будується як музична партитура, але замість інструментів виступають учасники комунікації. Це дозволяє точніше відобразити процес спілкування в абсолютному вимірі (на часовій осі) і у відносному вимірі, що характеризує мову учасників комунікації в порівнянні один з одним (одночасне говоріння, паузи, вставки). Суто технічно партитурна нотація вимагає більшої точності і більш складна в написанні. Однак використання спеціальних засобів комп'ютерної підтримки дозволяє спростити створення партитурних транскриптів мовлення [5, с. 124].

Отже, електронні корпуси діалектних текстів є принципово новим джерелом, що сприяє залученню діалектології до сучасної наукової лінгвістичної парадигми, в якій вивчення основних лінгвістичних характеристик діалектних одиниць автоматизується, що, в свою чергу, забезпечить перехресні дослідження в текстах різних діалектів, полегшить пошук і вибірку необхідних даних і дозволить проводити діахронічні дослідження в темпоральному просторі декількох десятиліть.

Робота над більшістю корпусів продовжується. Однак, питання доцільності застосування корпусних методів стосовно опису лінгвістичних явищ супроводжується не лише сумнівами, але й скептичними відгуками стосовно надійності такого джерела інформації, як корпус.

Патріарх генеративної лінгвістики Н. Хомський вважає, що корпусний підхід зводиться до простого спостереження за великим обсягом даних і «не є методом наукового пізнання і не може забезпечити ні успішного вирішення пізнавальних і практичних проблем, ні приросту знання» [7].

Широкий спектр сучасних корпусних досліджень показано опонентами генеративістів з питань пізнавального потенціалу корпусних досліджень, представниками корпусної лінгвістики з Кембріджського університету в навчальному посібнику «Corpus Linguistics: method, theory and practice», в якому дослідники окреслюють межі між двома основними підходами до вивчення лінгвістичних явищ Corpus-driven vs Corpus-based і доводять їх паралельність [8]. Якщо в дослідженні, заснованому на даних корпусу (corpus-based study), вирішується питання перевірки валідності теорії або гіпотези з використанням корпусних методів, то лінгвіст-корпусник за даними корпусу (corpus-driven study), будує свою теорію, повністю і цілком покладаючись на матеріал корпусу, описуючи таким чином узус. Як бачимо, у зарубіжній лінгвістиці не стоїть питання про звернення до корпусів, питання полягає здебільшого в підході до цього лінгвістичного ресурсу, однак і ця відмінність досить умовна. Представники когнітивного напрямку в зарубіжному мовознавстві інкорпорують емпіричні методи в лінгвістичний опис та прирівнюють в правах застосування експерименту і корпусних даних.

Вчені-лінгвісти також зазначають, що «в лінгвістиці сталася корпусна революція». Після появи корпусів ця наука стала зовсім інша»; і навіть якщо «цей пафос трохи прибрати, трохи знизити градус, то ступінь значущості все ж таки залишиться. Корпусне дослідження – це більше, ніж методика аналізу» [9]. Вітчизняні вчені на сьогоднішній день не тільки використовують корпусні методи і дані у своїй роботі, але й створюють і анотують корпусні ресурси, тому слід сказати, що сучасна лінгвістика повинна стати лінгвістикою корпусів, коли відбір матеріалу дослідження буде фундуватися не на даних словників та інших лексикографічних джерел, а буде здійснюватися за допомогою грамотно сформованих пошукових запитів [10]. Є вчені, які налаштовані скептично до використання корпусних технологій в лінгвістиці. Вони висловлюють думку, що звернення до корпусних ресурсів в лінгвістичному описі – це лише данина моді. Як на рубежі століть модно було вивчати концепти та інші запозичені із західної лінгвістики конструкти, так і сьогодні вважається модним застосовувати корпусні дані, оскільки вчені-скептики впевнені, що захоплення корпусами мине, як минула мода на опис концептів.

Прихильники зваженого, обережного поводження з корпусним даними, особливо при зверненні до Wikipedia Corpus або Google Books, вважають, що деякі лінгвісти сліпо слідують «моді на корпус», і це тотальне захоплення часто веде до фальсифікації результатів та зловживанню кількісними даними. На користь ненадійності результатів, що отримуються в рамках корпусного дослідження є узуальна природа корпусу, що вміщує ненормативне і неправильне вживання, зокрема це стосується інтернет-ресурсів. Такий стан речей виправдовує побоювання прибічників традиційних методів дослідження і інтроспекції в обмеженості корпусного підходу. Фактично, результати такого дослідження обмежені описом узусу, що, на думку скептиків, не дозволяє робити висновки про теоретично важливі закономірності мовної системи.

Застосування корпусних технологій передбачає знайомство з основами корпусної та ІТ-термінології, володіння навичками формування оптимального для цілей дослідження пошукового запиту і методами кількісної та статистичної обробки даних. Ускладнює роботу лінгвіста-корпусника і недосконалість пошукового інструментарію, що породжує певний скептецизм. Пошук за запитом може видавати сотні і навіть тисячі результатів (контекстів слововживання), які просто фізично нереально переглянути в обмежених часових рамках. Це провокує вчених-скептиків критично ставитися до заяв вчених-революціонерів і вчених-романтиків про те, що корпусні технології заощаджують час, а пошукова система сприяє вирішенню проблем архітектоніки і розвитку мови. Вдосконалення пошукових систем і методик запиту – одне з важливих завдань, що стоять перед корпусною лінгвістикою.

На разі розглянемо й питання, на чому базується переконаність лінгвістів-корпусників в тому, що сучасні дослідження мови не можуть проводитися поза корпусною лінгвістикою. По-перше, корпусні дослідження мови відрізняються більшою репрезентативністю даних, що передбачає і реально квантитативні, і статистичні дослідження [11; 12]). Однак питання стосовно збалансованості та репрезентативності корпусу є відкритим [13]. По-друге, корпусні дослідження – це все більш і більш вдалі спроби з не наявного (наприклад, усного дискурсу) зробити наявне (дискурс, представлений у розмічених текстах, який можна вивчати [14]). Окрім того, корпусні технології дозволяють приступити до спостереження

за рідкісними мовними явищами і прослідкувати динаміку мовних змін на малому часовому відрізку. У корпусному дослідженні мови знаходять відображення і отримують інтерпретацію як частотні явища, так і оказіональні. Деякі дослідницькі завдання передбачають звернення не до одного, а кількох ресурсів. Порівняння і грамотний аналіз даних, отриманих за допомогою різних корпусів, уможливлює встановлення мовної варіативності і закономірностей мовних змін, прогнозування подальшого розвитку описуваного явища, а також осмислення таких вживань, які суперечать усталеним уявленням про мовну норму. На відміну від інших видів дослідження (інтроспекції, словникової роботи), корпусна методологія дозволяє перевіряти гіпотези щодо мовних змін і закономірностей. Використання корпусів надає можливість об'єктивізувати лінгвістику, знайти більш вагомі аргументи стосовно фактів, дозволяє створити ситуацію повторюваності, що є важливим елементом науки, оскільки перевірка результатів корпусного дослідження забезпечує його ефективність, повторюваність і достовірність.

Корпусні технології дозволяють отримати принципово нові дані стосовно еволюції мови, її функціонування. Можливість отримання репрезентативної вибірки мови на різних етапах історії в машинописній формі дозволяють історикам мови швидше і ефективніше проводити їх діахронічні дослідження. Отримані результати характеризуються більшою системністю та всеохопністю, дозволяють прослідкувати динаміку мовних змін, встановити закономірності розвитку національної мови та виявити специфіку функціонування діалектів [15, С. 131]. Для цього створюються діахронічні (історичні) корпуси, які уможливлюють дослідження мови упродовж якого-небудь (досить довгого) проміжку часу або вивчати мову в її сучасному стані. До числа діахронічних корпусів належать, наприклад, гельсінський корпус як відомий і часто використовуваний корпус текстів різних періодів, корпус Lampeter ранньоновоанглійської мови, що містить добірку матеріалів, виданих між 1640 і 1740 роками, Боннський корпус сформований з ранньонововерхньонімецьких текстів, а Бохумський корпус вміщує рукописи середньоверхньонімецького періоду розвитку німецької мови. Отже, слід зауважити, що корпус автентичних пам'яток є достовірним першоджерелом дослідження мовної системи, що дозволяє визначити основні закономірності її історичного розвитку.

Висновки дослідження. Вчені застерігають про «небезпеку новизни» в лінгвістиці, називаючи це парадоксом її внутрішнього розвитку [9], але з точки зору практичного використання корпусної методології стосовно досліджень лінгвістичних явищ це нерозумно і недоцільно. Аналіз доробок вітчизняних і зарубіжних дослідників щодо застосування методів та підходів корпусної лінгвістики для вирішення лінгвістичних проблем дозволяє констатувати той факт, що методи і підходи корпусної лінгвістики мають як теоретичну, так і практичну значущість, принципову новизну та перспективність використання в сучасній науковій парадигмі.

Література

1. Крючкова О. Ю. Корпус русской диалектной речи: концепция и параметры оценки / О. Ю. Крючкова, В. Е. Гольдин, А. П. Сдобнова. – URL: http: // www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/36.pdf.

2. Perkuhn R. Korpuslinguistik / R. Perkuhn, H. Keibel, M. Kupietz. – Paderborn : Wilhelm Fink Verlag, 2012. –144 S.

3. Ахманова О. С. Словарь лингвистических терминов / О. С. Ахманова. – М. : КомКнига, 2007. – 576 с.

4. Захаров В. П. Корпусная лингвистика: учебник для студентов гуманитарных вузов / В. П. Захаров, С. Ю. Богданова. – Иркутск : ИГЛУ, 2011. – 161 с.

5. Баранов О. Н. Введение в прикладную лингвистику / О. Н. Баранов. – М. : Едиториал УРСС, 2003. – 360 с.

6. Lemnitzer L. Korpuslinguistik. Eine Einführung / L. Lemnitzer, H. Zinsmeister. – Tübingen : Narr Verlag, 2010. – 214 S.

7. Andor J. The master and his performance : An interview with Noam Chomsky / J. Andor // Intercultural Pragmatics. – 1-1. – 2004. – P. 93–111.

8. McEnery T. Corpus Linguistics: Method, theory and practice / T. McEnery, A. Hardi. – Cambridge : Cambridge University Press, 2012. – Support website for Corpus Linguistics: Method, theory and practice. – Mode of access. – Режим доступа : http: // corpora.lancs.ac.uk/clmtp.

9. Плунгян В. А. Почему современная лингвистика должна быть лингвистикой корпусов. Лекция, прочитанная в рамках проекта «Публичные лекции» / В. А. Плунгян. – Режим доступа : http://polit.ru/article/2009/10/23/corpus/.

10. Методы когнитивного анализа семантики слова : компьтерно-корпусный подход / под общ. ред. В. И. Заботкиной. – М. : Языки славянской культуры, 2015. – 344 с.

11. Boriskina O. O. A Corpus-based Study of Noun Cryptotypes in English / O. O. Boriskina // Компьютерная лингвистика и интеллектуальные технологии: материалы ежегодной Международной конференции / А. Е. Кибрик, В. И. Беликов, И. М. Богуславский, Б. В. Добров, Д. О. Добровольский [и др.]. – 2011. – С. 135–145.

12. Донина О. В. Криптоклассные данные для определения меры языковой эквивалентности / О. В. Донина // Вестник Воронежского государственного университета. – Сер.: Лингвистика и межкультурная коммуникация. – 2015. – № 1. – С. 108 110.

13. Шилихина К.М. Роль контекста в интерпретации иронии / К. М. Шилихина // Вестник Воронежского государственного университета. – Сер. Лингвистика и межкультурная коммуникация. – 2008. – № 2. – С. 10–15.

14. Шилихина К. М. Использование корпусов в исследованиях дискурса / К. М. Шилихина // Вестник Воронеж. гос. ун-та. Сер.: Лингвистика и межкультурная коммуникация. – 2014. – № 3. – С. 21–26.

15. Лук'янець Г. Г. Основні напрямки сучасних корпусних досліджень мови та перспективи їх подальшого розвитку / Г. Г. Лук'янець // Наукові праці НУХТ. – 2012. – № 44. – С. 131.