УДК 81'1

Пасічник Руслана
(м. Острог, Україна)
Національний університет «Острозька академія»

КОРПУС ЯК ОБ'ЄКТ ПРИКЛАДНОГО МОВОЗНАВСТВА

Тези присвячено одному із пріоритетних напрямів сучасних прикладних лінгвістичних досліджень – корпусній лінгвістиці. Сформульовано визначення поняття «корпусна лінгвістика», а також встановлено основні ознаки текстового корпусу.

Ключові слова: *корпусна лінгвістика, текстовий корпус, корпусні ознаки*

Тезисы посвящены одному из приоритетных направлений современных прикладных лингвистических исследований – корпусной лингвистике. Сформулировано определение понятия «корпусная лингвистика», а также установлены основные признаки текстового корпуса.

Ключевые слова: *корпусная лингвистика, текстовый корпус, корпусные признаки*

This research is devoted to one of the most priority directions of the modern applied linguistics studies – corpus linguistics. The basis of corpus linguistics is the development of theoretical statements and practical techniques of construction, machine processing, maintenance and analysis of linguistic data, structured as a corpus of texts.

Key words: *corpus linguistics, text corpus, corpus features.*

У наш час, коли комп'ютерні технології розвиваються доволі стрімко, все більшої популярності набуває корпусна лінгвістика як галузь конкретного мовознавства. Корпусна лінгвістика поступово і впевнено закріплюється як провідний напрям науки про мову, і сьогодні все більше науковців долучаються до її вивчення і сприяють її розвитку. На сьогоднішній день ця галузь науки є перспективним напрямом сучасної прикладної лінгвістики,

яка широко використовує статистичні методи обробки мовного матеріалу.

Вперше термін «корпусна лінгвістика» почали вживати ще в 90-х роках XX століття, у зв'язку з активним розвитком створення лінгвістичних корпусів текстів, поштовхом для якого став розвиток обчислювальної техніки. Першим комп'ютерним корпусом вважається Браунівський корпус, що був створений XX століття. Автори корпусу: У. Френсіс та Г. Кучер вперше вжили слово «корпус» як «сукупність текстів, яка є визначальною для цієї мови і створена для лінгвістичного аналізу».

Виклад основного матеріалу. Корпусна лінгвістика – це розділ мовознавства, що займається створенням, обробкою та використанням корпусів. Відповідно до цього визначення, очевидним є той факт, що саме корпус є предметом дослідження прикладної галузі лінгвістики. Відповідно визначень деяких вчених, корпус розглядають як структуроване зібрання машино читаних текстів найрепрезентативнішого мовного матеріалу тієї або іншої природної мови, є передовсім загальним методом фіксації та дослідження мови, узагальненою моделлю організації та подавання фактичного матеріалу, базованим на машинних технологіях" [1]. Під «корпусом», крім машиночитаного структурованого зібрання текстів природної мови, інколи розуміють будь-який текстовий матеріал або довільний машиночитаний текст, але таке розуміння є некоректним, тому що для того аби зібрання текстів вважалося корпусом, йому повинні бути притаманні певні ознаки [2].

Проте на сьогодні корпусна лінгвістика ще не завершила до кінця процес вироблення єдиного погляду на корпусні ознаки, тобто ті характерні особливості, які необхідні для перетворення будь-якого електронного тексту на корпус. Вперше про ці ознаки, які повинен мати кожен корпус текстів, заговорив Джордж Синклер: "Корпус повинен мати характеристики, значення яких є «значенням за промовчуванням». До таких детермінантних параметрів текстового корпусу, відповідно до думки Дж. Синклера належать обсяг, автентичність, машинне подання і документованість [6]. Однак пізніше цей перелік було дещо змінено іншим науковцем Г. Кенеді, який обов'язковими корпусними ознаками

вважає статичність (динамічність), репрезентативність, збалансованість і обсяг [5]. Щодо такої ознаки як репрезентативність, то її визнають і багато інших вчених, зокрема А. Баранов і В. Риков. Згідно з В. Риковим саме ця характерна властивість необхідна для перетворення "набору текстів на машинному носії на унікальну словесну єдність – корпус текстів" [4]. А Баранов також пропонує ще інші додаткові параметри, необхідні для кожного корпусу текстів: економність, повнота, структурованість матеріалу та комп'ютерна підтримка. Підсумувавши думки вчених, їх релевантні доповнення або схиляння до певних характеристик, було сформовано загальний перелік ознак корпусу текстів: 1) обсяг; 2) автентичність; 3) репрезентативність; 4) збалансованість; 5) електронна форма або комп'ютерна підтримка; 6) документованість; 7) простота подання; 8) повнота; 9) економність; 10) структурованість; 11) статичність або динамічність.

Однак, цей перелік має свої нюанси, адже слід брати до уваги той факт, що не всі вище перераховані ознаки мають однакову важливість і необхідність для сучасного електронного корпусу. Наприклад, корпусні ознаки статичності або динамічності формулює лише Г. Кеннеді [5]. Але він швидше за все має на увазі методики відображення предметного домену в корпусі. Якщо у корпусі представлено якийсь чітко детермінований проміжок часу і стан предметного домену, то корпус буде статичним, якщо ж предметний домен не обмежений часово , то він буде динамічним. Тому статичність або динамічність не варто розглядати як релевантну ознаку корпусу, це, швидше за все, типологічна характеристика текстового корпусу. Подібно можна спростувати ще декілька не зовсім відповідних електронному корпусу текстів ознак [2].

Згідно із деякими вченими, найбільш важливими корпусними ознаками, без яких текстове зібрання практично неможливо кваліфікувати як корпус, є репрезентативність, що надає корпусу можливість відображати всі властивості предметної галузі, тобто рівня реалізації мовної системи, яка включає в себе феномени, які підлягають лінгвістичному описові. Варто зауважити, що предметна галузь для корпусу може бути як завгодно великою

або малою. Інша не менш важлива ознака – це автентичність, що має на меті відбір реально створеного носіями мови писемного або усного тексту, уривка чи тексту у процесі мовної комунікації. Відібраність, що ставить вимогу обмеження фактичного матеріалу шляхом відбору певних фрагментів мови з усього мовного континууму. Беззаперечним є той факт, що навіть найбільший за обсягом корпус порівняно з усіма створеними усними і писемними текстами є мінімальним взірцем, який не в змозі передати весь той мовний матеріал. Для цього, власне, і створюється вибірка, яка передбачає застосування чітких правил екстрагування даних. Збалансованість є також доволі важливою ознакою, яка полягає у введенні до корпусу пропорційної кількості текстових ресурсів. І остання, одна з найважливіших ознак, є машиночитаність, що є визначальною ознакою сучасного електронного текстового корпусу природної мови. Ця вимога передбачає наявність кодування первинних корпусних даних та лінгвістичну анотацію [2].

Отже, вченні вважають, що «корпусна лінгвістика» – це окрема галузь мовознавства, предметом дослідження якої є принципи та методи формування корпусів текстів, а також розроблення комп'ютерних систем для їхнього опрацювання. А корпус, своєю чергою, розглядають як модель мовної системи, що застосовується в текстах різних функціональних стилів, тематики, призначення і структури.

Література:

1. Демська-Кульчицька О. Базові поняття корпусної лінгвістики / О. Демська-Кульчицька // Українська мова. – 2003. – № 1(6). – С. 40–45.

2. Демська-Кульчицька О. Репрезентативіність як ознака текстового корпусу / О. Демська-Кульчицька // Українська мова. – 2005. – № 3. – С. 100–107

3. Рыков В. В. Прагматически ориентированный корпус текстов / В. В. Рыков // Тверской лингвистический меридиан. – 1999. – Вып. 3. – С. 89–96.

4. Kennedy G. Introduction to Corpus Linguistics / G. Kennedy. – London – New-York, 1998. – 309 p.

5. Sinclair J. Corpus, Concordance, Collocation / J. Sinclair. – Oxford, 1991. – 137 p.