Дослідження особливостей проектування корпусу на основі текстової моделі

Уляна Шандрук

асистент кафедри прикладної лінгвістики, Національний університет «Львівська політехніка», Україна,
E-mail: Shandruk.uliana@gmail.com

Dictionaries and grammar books presuppose a description of linguistic facts that can be obtained only from the text. A large amount of texts of a specific author or group of authors is called text corpus. In modern linguistics there is an urgent need for such a tool, as any dictionary can cover all morphological, syntactic, semantic, pragmatic and stylistic properties of the language and speech. Idiolect is an individual style of the language of a particular author. It best reflects the state of language in a particular time and place which is an inexhaustible source for the study of the language and speech in general and the compilation of new corpora in particular. The article is devoted to the different types of corpora and the specific features of their creation, namely the creation of corpus annotation as well as to the investigation of the issues of text models.

Ключові слова — текстовий корпус, корпусна лінгвістика, ідіолект, анотація, модель тексту.

Словники та граматики передбачають опис мовних фактів, який можна вивести лише з текстів. Великий обсяг текстів конкретного автора чи групи авторів називають лінгвістичним корпусом. У сучасному мовознавстві існує гостра потреба у такому інструментарії, адже жодний словник не покаже усі морфологічні, синтаксичні, семантичні, прагматичні, та стилістичні властивості мови, тому інтерес до корпусної лінгвістки нестримно зростає.

Ідіолект – особистий стиль автора – якнайкраще відображає стан мови у певну історичну епоху, що є невичерпним джерелом для дослідження мови та мовлення. Саме тому створення корпусів конкретних авторів дає змогу глибинно вивчати мову.

Характерною особливістю лінгвістичного корпусу є наявність текстової розмітки, або анотування. Для того, щоб якнайкраще зрозуміти процес створення текстової розмітки, потрібно дослідити моделі тексту. Мета статті й полягає у дослідженні текстових моделей і типів та особливостей створення корпусів.

Будь-який текст є відображенням індивідуального стилю автора, який його створює. Ця індивідуальність проявляється вибірковістю тих чи тих лексичних, морфологічних, синтаксичних і фонетичних засобів. Автор, будучи частиною соціуму, у своїх текстах фіксує історичні, національні та територіальні особливості мовлення. Саме тому ідіолект використовують як базу для створення текстових корпусів.

Явище корпусу є продовженням традиційних картотек, які перейшли на новий рівень – комп'ютерний. Із розвитком Інтернету з'явилася можливість зберігати текст у електронному вигляді та мати швидкий доступ до нього. Проте для того, щоб створити такий корпус, необхідно підготувати текст для подальшого опрацювання.

Процес підготовки текстів є доволі складним і тривалим. Такі дослідники, як В. Захаров, С. Богданова, В. Жуковська, виокремлюють декілька етапів створення корпусу текстів: 1) визначення джерел мовного матеріалу та предметної ділянки; 2) введення даних; 3) попереднє опрацювання тексту; 4) конвертування й графемний аналіз; 5) розмітка тексту; 6) коректування результатів автоматичної розмітки; 7) конвертування розмічених текстів у структуру спеціалізованої лінгвістичної інформаційно-пошукової системи; 8) забезпечення доступу до корпусу. Також є прихильники іншої класифікації, наприклад, О. Демська зводить всі етапи до трьох основних: «1) проектування; 2) збір текстів; 3) кодування текстових даних» [2, с. 133]. Проте всі погоджуються, що найважливішим етапом є проектування корпусу (визначення джерел мовного матеріалу та визначення предметної ділянки). Якщо на цьому рівні все зроблено правильно, ймовірність подальших труднощів чи помилок зменшується.

Першим кроком на етапі проектування є визначення предметної ділянки корпусу й постановка чітких завдань, які цей корпус має вирішувати. Тобто слід обрати його об'єкт [4, с. 34]. Наприклад, якщо це корпус західноукраїнських письменників XX століття, дослідник має скласти перелік усіх письменників, які публікували свої твори у цей час, а також, що є набагато складніше, – отримати доступ до творів усіх письменників у списку та визначити, яку інформацію про текст міститиме корпус.

Важливим нюансом на цьому етапі є джерела отримання текстів, адже вони можуть бути як публічно доступні, так і приватно доступні. Перші можна отримати у бібліотеках, архівах, чи в Інтернеті. На них не розповсюджується закон про авторське право, тому їх можна використовувати безкоштовно. Приватно доступні дані є зазвичай власністю конкретної людини, організації тощо. Тому для користування та оприлюднення цих текстів потрібно отримати дозвіл [3, с. 85].

Існує два підходи відбору даних: контрольований корпус (monitor corpus) та збалансований корпус (balanced corpus). У першому випадку, після створення корпусу його можна постійно поповнювати новими текстами, у другому – корпус буде відображати стан мови у той період, коли його було створено [12, с. 6]. Щодо структури корпусів, то бувають моно- та полікорпуси. Монокорпус має простішу структуру та складається лише з

генерального корпусу, а полікорпус є складнішим і містить у собі як генеральний корпус, так і підкорпуси (декілька різних корпусів) [2, с. 141].

Наступний етап – це збирання текстів, результатом якого є готові тексти в електронному вигляді. Оцифровуючи надруковані тексти, слід надавати їм такої форми, щоб відповідне програмне забезпечення могло їх опрацювати. Є три способи, як цього можна досягнути: ручне введення тексту, сканування й розпізнавання та отримання електронних текстів [9, с. 157].

У процесі створення корпусу перш за все треба взяти до уваги такі критерії, як репрезентативність корпусу, збалансованість та розмір. Як зазначає О. Демська, «Репрезентативність – здатність корпусу відображати всі властивості предметної галузі» [2, с. 105]. На репрезентативність впливають два фактори: набір жанрів корпусу та критерії відбору тексту [11, с. 11]. Якщо текст у корпусі не репрезентативний, то такий корпус є дуже обмеженим та втрачає свою придатність. Корпус, у якому текстів одного типу, чи жанру є більше, ніж іншого, називають незбалансованими, а збалансованість є ключовим елементом у розробці корпусу. Питання розміру корпусу є достатньо контроверсійним, адже видається, що чим більшим корпус є, тим краще, але тут з'являється програмне обмеження. Тож обсяг корпусу має відповідати завданням поставленим до нього [9, с. 165].

Наступним етапом створення корпусу є перевірка й виправлення текстів, оскільки у процесі оцифрування ймовірність виникнення технічних помилок досить висока. Деякі тексти проходять також один чи декілька етапів попереднього машинного опрацювання, під час якого відбувається перекодування та видалення всіх нетекстових елементів. Також на цьому етапі зазвичай відбувається сегментація тексту на його структурні компоненти [3, с. 87].

Існують розмічені та нерозмічені корпуси. Розмітка передбачає наявність лінгвістичної інформації про слово, наприклад, морфологічної чи синтаксичної. Розмічені корпуси мають більший лінгвістичний потенціал, аніж нерозмічені, оскільки у других робота з корпусом зводиться лише до пошуку слова, його словоформи, побудови конкордансу або визначення деяких частотних показників [2, с. 205]. Згідно з О. Демською, «розрізняють такі типи розмітки: фонетичну, просодичну, морфолого-синтаксичну (морфосинтаксичну), семантичну, анафоричну, дискурсну, прагматичну тощо» [2, с. 205].

Отже, розмітка – це приписання кожному слову набору певних тегів . Теги можуть бути лінгвістичні, які описують лексичні, граматичні, синтаксичні та семантичні властивості, та екстралінгвістичні, які подають дані про автора, текст, рік написання тощо [4, с. 34].

Із погляду створення розмітка буває трьох типів: автоматична, автоматизована та ручна, жодна з яких не виключає помилок. Однак автоматична є найбільш точною. У статті «CLAWS part-of-speech tagger for English» Р. Гарсид та Н. Сміт доводять, що точність автоматичної розмітки тексту сягає 97 %. Вважають, що автоматична розмітка є найшвидшою, найточнішою та найдешевшою, проте не до всіх текстів її можна використовувати. Для створення малого набору даних і перевірки якості роботи тегерів часто використовують автоматизований тип розмітки [12, с. 31].

Оскільки корпуси призначені для багаторазового використання різними користувачами, то текстова розмітка має бути загальноприйнята й уніфікована. Стандартом розмітки текстів є TEI (Text Encoding Initiative), розроблений на основі CES (Corpus Encoding Standard), який вимагає опису текстів у форматі DTD (Document Type Definition). Для опису структури лінгвістичного корпусу використовують теги [5, с. 239]. Останнім кроком до функціонування корпусу є створення корпусного менеджера, тобто спеціальної системи, за допомогою якої можна здійснювати пошук у корпусі.

Перед тим як текст розмічувати, потрібно побудувати його модель. Існують декілька моделей тексту, серед яких матеріалістична, семантична, структурна. Для створення корпусу найважливішою є структурна модель тексту, яка розглядає його з механічної погляду, розділяючи на структурні частини.

Матеріалістична модель тексту відображає матеріальне середовище, у якому живе текст. Це може бути каміння, дерево, папір тощо. Хоча здебільшого ми звикли думати, що текст існує в документі [7]. Семантична модель тексту підтверджує, що «текст існує тільки тому, що читач надає йому сенс» [8]. Згідно з цією моделлю, текст складається зі створених читачем символів та структур, яких обов'язково наявна неоднозначність і можливість різної інтерпретації. Найкраще семантичну модель проявляється в художньому тексті, де кожний читач може трактувати текст по-своєму. На відміну від матеріалістичної та структурної моделей тексту, семантична модель припускає, що текст є завжди нестійким.

Згідно з І. Кочан, структурна модель тексту складається з послідовності конструктивних блоків. До цих блоків належать: вступний (інтродуктивний), основний і завершальний. Між блоками існують відношення. Слід зазначити, що структура наукового тексту та художнього дещо відрізняються. Остання є складнішою. Відтак принципи побудови моделей цілком залежать від типу тексту [6, с. 57].

Інтродуктивний блок охоплює вступ, передмову, пролог, преамбулу тощо. Цей блок є факультативним, інколи його називають передтекстовим, адже він не існує окремо від тексту. У наступному блоці подано основний текст. Це головна інформація, яку треба донести до читача. Завершальний блок, залежно від стилю тексту, може виступати висновком, післямовою, епілогом [6, с. 69]. Також у тексті є певні знаки або ж зовнішні індикатори, без яких він існувати не може, наприклад, заголовок тексту [6, с. 57]. Він має значну роль, оскільки «саме від заголовка чи не найбільшою

мірою залежить поширюваність опублікованої інформації у просторі та часі» [1, с. 24].

Отже, дослідження авторського ідіолекту має практичне застосування – створення текстового корпусу задля подальшого дослідження мови. Проте процес створення такого корпусу є складним і багатогранним. Основну увагу слід приділити розробленню розмітки, оскільки корпус тим кращий, чим повніша й досконаліша його анотація. Для створення розмітки необхідно чітко розуміти всі елементи текстової моделі та зв'язки, які існують між ними.

Література

1. Базан О. М. Функції риторичних заголовків у публіцистичному тексті (на матеріалі сучасних україномовних ЗМІ) / О. М. Базан // Наукові записки. Серія "Філологічна". – Київ, 2012.
2. Демська О. Текстовий корпус: Ідея іншої форми / О. Демська. – Київ: Видавничо-поліграфічний центр НаУКМА. – 2011.
3. Жуковська В. В. Вступ до корпусної лінгвістики: навчальний посібник / В. В. Жуковська – Житомир: Вид-во ЖДУ ім. І. Франка, 2013. – 142 с.
4. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник для студентов гуманитарных вузов. – Иркутск: ИГЛУ, 2011. – 161 с.
5. Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна та ін. – К.: Довіра, 2005. – 471 с.
6. Кочан І. М. Лінгвістичний аналіз тексту: Навч. пос. – 2-ге вид. / І. М. Кочан. – Київ: Знання, 2008 – 423 с.
7. Brown J., Duguid P. The social life of information. / J. Brown, P. Duguid. – Boston, MA: Harvard Business School Press. – 2000.
8. Cavallo G., Chartier R. A history of reading in the west / G. Cavallo, R. Chartier. – Amherst, MA: University of Massachusetts Press. – 1999.
9. Corpus Linguistics / A. Lüdeling, M. Kytö. – Berlin: Walter de Gruyter, 2008.
10. Garside R., Smith N. A hybrid grammatical tagger: CLAWS4 / R. Garside, N. Smith // Corpus Annotation: Linguistic Information from Computer Text Corpora. – London: Longman, 1997.
11. McEnery A., Xiao Z. Swearing in modern British English: the case of fuck in the BNC / A. McEnery, Z. Xiao // Language and Literature. – 2004. – №13 (3). – P. 237-270.
12. McEnery T., Hardie A. Corpus linguistics : method, theory and practice / Tony McEnery, Andrew Hardie. – Cambridge University Press, 2012. – 294 p.
13. Shillingsburg P. Scholarly editing in the computer age: Theory and practice / P. Shillingsburg. – MI: University of Michigan Press. – 1996.
14. Yin Liu. Ways of Reading, Models for Text, and the Usefulness of Dead People / Liu Yin // Scholarly and Research Communication. – Canada: University of Saskatchewan. – Vol 5. – 2014.

III. КЛАСИЧНА Й КОМП'ЮТЕРНА ЛЕКСИКОГРАФІЯ

Нова джерельна база української лексикографії в системі інтегральних лінгвістичних досліджень

Наталія Сніжко

к. філол. н., старший науковий співробітник відділу лексикології, лексикографії та структурно-математичної лінгвістики, Інститут української мови НАН України, Україна, E-mail: natasnow@ukr.net

This article is devoted to actual problems of modern integrated lexicography, which combines scientific research and dictionary creation activities. The author described the directions of updating of register and textual-illustration foundations of modern lexicography, revealed the structure and functions of the integral lexicographical environment. On base of text illustrations of contemporary writers, the main rubrics of new source base were identified, and the ways of integral research of active, passive, inert and refreshed vocabulary, structure of neo-lexis of writers, system of traditional and modern figurative means of the language were determined.

Ключові слова – інтегральна, тлумачна, авторська лексикографія, джерельна база лексикографії, цитата, образне вживання слова, порівняння, неолексикон Олеся Гончара, активна і пасивна лексика, лексикографічна синергетика, зведений словник.

Початок XXI ст. – період активного утвердження методології інтегральних лінгвістичних досліджень, систематизування знань про світ і мови в інтегральних лексикографічних середовищах.