DOI 10.33190/0027-2833-320-2021-5-005

Бук С. Н.

ВЕЛИКА ПРОЗА ІВАНА ФРАНКА: ЕЛЕКТРОННИЙ КОРПУС, ЧАСТОТНІ СЛОВНИКИ ТА ІНШІ МІЖДИСЦИПЛІНАРНІ КОНТЕКСТИ

Львів : ЛНУ імені Івана Франка, 2021. 424 с.

У монографії Соломії Бук описано досвід побудови корпусу великої прози Івана Франка, укладання частотних словників і здійснення статистичних досліджень на основі цих текстів.

Представлення і дослідження української літературної спадщини в електронних корпусах текстів, створення й удоступнення в інтернеті українських лінгвістичних ресурсів є дуже актуальним завданням і для української філології, і для славістики в цілому. Завдяки сучасним комп'ютерним технологіям створені унікальні можливості для масштабних порівняльних славістичних корпусних досліджень, у яких українська мова довго була «білою плямою». Причини цього прикрого запізнення не лише організаційні та фінансові. Існують і суто філологічні, текстологічні чинники, що ускладнюють процес створення українських корпусів, зокрема історичних, до яких належить Корпус великої прози І. Франка. Стандартні електронні корпуси текстів ґрунтуються на паперових виданнях, а сучасні видання українських класичних творів часто не точно відповідають оригіналу, містять редакторський варіант тексту, наближений до сучасної літературної норми, а отже, не можуть бути надійним джерелом для дослідження історії мови. Зокрема, в академічному 50-томному виданні творів І. Франка (Київ : Наук. думка, 1976–1986) є не тільки численні купюри, а й значні текстові заміни порівняно з прижиттєвими виданнями[1] . З іншого боку, текстологічно надійніші старі видання XIX — першої половини XX ст. мають граматичні та правописні особливості, які потребують додаткових інструментів для комп'ютерного аналізу або навіть ручного опрацювання текстів.

Соломія Бук уклала корпус, до якого увійшли дев'ять великих художніх творів І. Франка: «Борислав сміється» (1880–1881), «Захар Беркут» (1883), «Не спитавши броду» (1885–1886), «Для домашнього огнища» (1892), «Основи суспільності» (1894–1895), «Перехресні стежки» (1900), «Boa constrictor» (1905–1907), «Великий шум» (1907), «Петрії і Довбущуки» (1909–1912). Загальний обсяг корпусу — 500 тис. слововживань.

Корпус текстів І. Франка містить різні редакції кожного твору (тексти за останніми прижиттєвими та сучасними виданнями) і побудований за принципом паралельного корпусу, що, з одного боку, зручно для детального порівняння варіантів, з другого боку, — усуває проблему вибору однієї редакції (такий вибір доводиться робити укладачам великих референтних корпусів, куди не можна додати декілька варіантів одного твору). Сподіваємося на оприлюднення паралельного корпусу варіантів творів І. Франка в

[1] Друль О. Поправлюваний Франко. *Збруч*. 2015. URL: https://zbruc.eu/node/35977 (дата звернення: 07.06.2021).

інтернеті, поповнення і створення на його базі чи за його зразком корпусів варіантів інших класичних творів української літератури XIX ст., що було б дуже корисно, враховуючи історію перевидання і редагування багатьох старих українських текстів.

У праці С. Бук представлено також класичні паралельні українсько-польський та польсько-український корпуси автоперекладів І. Франка. Хоча для пари української і польської мов існують паралельні корпуси, автопереклади І. Франка в таких корпусах досі не були представлені.

Паралельний аналіз різних видань та редакцій творів І. Франка дозволив С. Бук укласти перелік правописних особливостей прижиттєвих видань І. Франка і рекомендувати зберегти деякі з них у сучасних перевиданнях (с. 68, 110–111). Крім цих правописних відмінностей, наведені у 50-томнику великі прозові тексти І. Франка і тексти останніх прижиттєвих видань С. Бук визнає загалом ідентичними (с. 107–117), саме тексти з 50-томного видання стали основним матеріалом рецензованої монографії.

Корпус великої прози І. Франка лематизовано напівавтоматично. С. Бук опрацювала тексти прижиттєвих видань І. Франка і вручну поповнила словник програми для коректної лематизації цих текстів, зокрема західноукраїнських мовних рис і нестандартних написань, що передають фонетично мову персонажів. Такий словник є цінним ресурсом, який потенційно може бути застосований для автоматичної лематизації більшого корпусу текстів, що належать до сучасного Франкові західного варіанта літературної мови. У корпусі вручну знято граматичну омонімію. Текст, написаний іншими мовами, також лематизований відповідно до граматики цих мов. Цікавим є аналіз референції займенника *я*, кожне вживання якого розмічено в корпусі за анафоричним референтом (с. 97–102). Щоправда, *моє «я»* (с. 99: «Регіно, Регіно! <…> Віддай мені найкращу частину мойого "я"»!) є вже не займенником, а субстантиватом, тому в нього, на наш погляд, не може бути анафоричного референта.

З усіх описаних у монографії корпусів С. Бук оприлюднила для пошуку в інтернеті корпус повісті «Перехресні стежки». Текст подано за сучасним правописом, є список лем (у тому числі лематизовані польські, німецькі, чеські, французькі, латинські фрагменти тексту), конкорданси для всіх лем, пошук здійснюється за словоформою і за частиною словоформи[2].

Значне місце в книжці відведено обчисленню різних статистичних показників для кожного твору і для великої прози І. Франка в цілому, зокрема кількість слововживань і слів, індекс різноманітності словника, кількість високочастотної і низькочастотної лексики, кількість слів за частинами мови, співвідношення прямого та авторського мовлення тощо. Здійснено кількісне зіставлення двох редакцій «Boa constrictor» (1884 і 1907 рр.), причому знайдено цікаві свідчення роботи І. Франка над стилем. Наприклад, виявлено, що в другій редакції він послідовно замінив *к* на *до*, *огромний* на *великий*/*здоровенний*, *сьогодня* на *сьогодні*, *хороший* на *гарний*, *способности* на *здібності* тощо.

У праці здійснюється розгорнутий порівняльний аналіз словників різних романів, що містить деякі цікаві спостереження над хронологічними тенденціями і специфікою окремих текстів (с. 232 — частотність частин мови, с. 267 і далі — рівень діалогічності різних творів). На наш погляд, добре було б

2 Бук С., Ровенчак А. Онлайн-конкорданс роману Івана Франка «Перехресні стежки». 2010. URL: http://ktf.franko.lviv.ua/users/andrij/science/Franko/concordance.html (дата звернення: 07.06.2021).

прокоментувати це докладніше. Чи можна пов'язати такі статистичні відмінності з загальною еволюцією творчості І. Франка? С. Бук наводить часткову зведену таблицю статистичного профілю всіх романів (с. 188–189), де можна побачити, як багатство словника Франка хронологічно змінюється (спочатку зростає, потім, у повістях «Основи суспільності» і «Перехресні стежки», зменшується, потім знову зростає), що може бути пов'язано як з жанровою специфікою текстів, так і зі зміною творчої манери автора. Зокрема, цікавими були б докладніший аналіз еволюції частки епітетів у творах І. Франка (с. 182, 238) і, можливо, якісь ілюстративні приклади з текстів.

Кількісні характеристики великої прози І. Франка зіставлено з характеристиками української прози XX ст. за «Частотним словником сучасної української художньої прози» (Київ : Наук. думка, 1981. Т. 1–2) як зразком української прози в цілому (с. 272–276). Цей словник укладено за повоєнними творами 22 письменників (1945–1970-х рр.). Здається, що для виявлення індивідуальних характеристик мови І. Франка доречніше було б порівняти його твори з корпусом сучасної йому прози чи навіть сучасної йому західноукраїнської прози. Для цього є достатній матеріал у референтному корпусі української мови[3]. Продовження зіставних досліджень мови І. Франка з референтним корпусом було б, на нашу думку, перспективним напрямком.

У дослідженні С. Бук застосовано статистичні методи. На текстах І. Франка показано, як виконується закон Ципфа. Закон Менцерата–Альтмана обчислено на двох рівнях: склад–слово і клауза–речення. Введено новаторський статистичний критерій — температура. Цей параметр стосується передусім низькочастотної лексики і показує динамічні зміни характеристик тексту при збільшенні обсягу вибірки. На прикладі корпусу творів І. Франка описано розподіли частиномовних характеристик ланцюжків словоформ (колокацій частин мови). Досліджено ступінь зв'язності різних лексем і побудовано мережі їх зв'язності.

Можна запропонувати ще декілька статистичних критеріїв, які в перспективі було б цікаво підрахувати для характеристики лексичної своєрідності текстів І. Франка. Це показник TF-IDF (який виокремлює лексику і колокації, що характерні для досліджуваного корпусу і рідко трапляються в інших) та показник Delta (який використовують у стилеметрії для характеристики авторського сигналу і атрибуції тексту). Для оцінки статистичної надійності тих чи інших різниць і порядків частот можна застосовувати статистичні критерії значущості (χ^2, Фішер).

Праця С. Бук є значним кроком уперед для української корпусної і квантитативної лінгвістики. Вона є узагальненням багаторічних досліджень і об'єднує різні підходи до аналізу обраного матеріалу. Деякі статистичні критерії і методи застосовано до української мови вперше. Ці результати вагомі не тільки для мови конкретного автора, а й для вивчення властивостей мови в цілому.

Розроблені С. Бук корпуси і частотні словники можуть стати універсальною основою для поглибленого вивчення мови І. Франка, імпульсом для розвитку української авторської лексикографії та дослідження ідіолектів.

[3] ГРАК: Генеральний регіонально анотований корпус української мови / М. Шведова, Р. фон Вальденфельс, С. Яригін, А. Рисін, В. Старко та ін. Київ; Львів; Єна, 2017–2021. URL: http://uacorpus.org/ (дата звернення: 07.06.2021).

М. Шведова
Київський національний лінгвістичний університет
м. Київ, Україна
Електронна пошта: corpus.textiv@gmail.com
http://orcid.org/0000-0002-0759-1689

M. Shvedova
Kyiv National Linguistic University
Kyiv, Ukraine
E-mail: corpus.textiv@gmail.com
http://orcid.org/0000-0002-0759-1689

Buk S. N.
LONG PROSE FICTION BY IVAN FRANKO: ELECTRONIC CORPUS, FREQUENCY DICTIONARIES, AND OTHER INTERDISCIPLINARY CONTEXTS
Lviv: Ivan Franko Lviv National University, 2021. 424 p.

Дата надходження до редакції — 05.06.2021
Дата затвердження редакцією — 12.06.2021