ISSN 2078-5534. Вісник Львівського університету. Серія філологічна. 2013. Випуск 58. С. 290–307
Visnyk of the Lviv University. Series Philology. 2013. Issue 58. P. 290–307

МОВОЗНАВСТВО

УДК 811.161.2'1'324'38 (038)

КВАНТИТАТИВНА ПАРАМЕТРИЗАЦІЯ ТЕКСТІВ ІВАНА ФРАНКА: ПРОЕКТ ТА ЙОГО РЕАЛІЗАЦІЯ

Соломія БУК

Львівський національний університет імені Івана Франка,
кафедра загального мовознавства,
вул. Університетська, 1, Львів, Україна, 79000,
тел.: (032) 239 47 56, e-mail: solomija@gmail.com

Запропоновано проект квантитативної параметризації усіх текстів І. Франка, що можливо реалізувати, створивши частотний словник усіх творів письменника і лише зі застосуванням сучасних комп'ютерних розробок. Вказано сфери застосування, етапи, методику, принципи і специфіку укладання частотного словника мови другої половини XIX – початку XX століття, якою писав І. Франко. Окреслено співвідношення частотного словника І. Франка зі словником мови письменника та корпусом текстів.

Ключові слова: частотний словник (ЧС), словник мови письменника, ідіостиль І. Франка, лінгвостатистичний аналіз тексту, корпус текстів, квантитативні параметри тексту.

Якщо зіставити кількість літературознавчих та мовознавчих досліджень творчості І. Франка, то шалька терезів схилиться до перших. За лінгвістичними напрямами аналізу доробку письменника розглянуто багато питань: ролі І. Франка у творенні літературної мови [25; 61], внеску в ономастику [54; 85], його мовотворчості й естетичної концепції [44; 53], деяких проблем загального мовознавства у його науковій спадщині [23; 30; 39; 40] та термінології [36]. Чи не найширше висвітлена стилістика творів [6; 51; 64], зокрема наголос у поезії [76], вживання діалектизмів [28; 37; 52], арготизмів [24], синонімів [43], дослідження синтаксису творів [49; 60]. Зазначимо, що творчість І. Франка привертає увагу не лише українських науковців. Наприклад, результати дослідження довжини слів у листах і поетичних творах І. Франка наведено в німецькомовних працях [69].

Проте, незважаючи на згадане, широке коло опрацьованих питань у творчості Генія залишаються майже не вирішені. Серед них зокрема – створення словника мови І. Франка (на цьому наголошують І. Ковалик [31], З. Терлак [59, с. 5–7], І. Ціхоцький [65]), анотованого корпусу його текстів, а також квантитативне дослідження творів

письменника, оскільки у лінгвістиці зламу XX–XXI століть особливо актуальними стали міждисциплінарні дослідження мови, до яких належить і застосування статистичних методів до мовних об'єктів.

Із цього погляду особливо варто зосередитися на реалізації задуму створення «Словника мови Івана Франка» науковцями Львівського університету під керівництвом професора І. Ковалика. Зважаючи на величезний обсяг творчості І. Франка в різних родах і жанрах, було вирішено видати спочатку словник поезії. Першою фазою створення лексикону слушно визначили укомплектування реєстру слів. Професор І. Ковалик розробив принципи та наукові основи укладання словника [31; 32], ґрунтуючись на яких було видано пробний зошит мови художніх творів І. Франка [33]. Карткування віршованої спадщини відбувалося вручну, то ж можна тільки уявити, скільки роботи виконано для виявлення лише списку слів, кількістю понад 35 000, що побачив світ у книзі «Лексика поетичних творів...» [34]. «Мета цього реєстру на основі підрахунків, наявних у картотеці слів, подати всю українську лексику поетичних творів Івана Франка та вказати на частотність використання кожного слова» [34, с. 4]. У словнику свідомо збережено діалектні, фонетичні, морфологічні і граматичні написання, матеріал подано єдиним списком за алфавітом. На основі словопокажчика поезій І. Франка Л. Полюга зробив короткий статистичний аналіз лексики його поезій [50], а І. Ощипко порівняла дані зі словниками мови Т. Шевченка та Г. Квітки-Основ'яненка [42].

Новим кроком, що водночас демонструє й тяглість лексикографічної традиції у франкознавстві, став «Словник мови поетичної збірки Івана Франка "Зів'яле листя"« [59], присвячений 350-річчю Львівського університету. Це праця тлумачного типу, в якій подано детальний опис значень кожного слова вказаної збірки, кількість його вживань, а також усі приклади його контекстуального оточення з урахуванням семантичних та стилістичних особливостей та відповідною паспортизацією. Цьому передувала величезна текстологічна робота укладача З. Терлака, який взяв за основу останнє прижиттєве видання «Зів'ялого листя» 1911 року.

Вказані видання становлять дуже важливу ланку опрацювання ідіостилю І. Франка, проте мають вибіркову природу, що дає підстави вважати проблему комплексного аналізу лексикону письменника, зокрема, створення словника його мови, відкритою.

У мовознавстві останнього часу можна констатувати намагання виявити певні чіткі й конкретні параметри, якими можна було б повно і ненадлишково описати певний лінгвістичний об'єкт, у тому числі ідіолект, мову письменників [1; 3; 29; 55; 67; 81]. Як зазначив Ю. Караулов, «Під **лексикографічною параметризацією** розуміють прагнення сучасної науки про мову подати у формі словника різні, а в ідеалі – всі результати лінгвістичних студій, тобто ословничування лінгвістичних описів» [29, с. 8]. Детально класифікацію словників, об'єктом дослідження у яких є текст письменника, описано у статті [71]. Вони, відповідно до мети і можливостей укладачів, представлені різними типами: словник мови письменника, словопокажчик, конкорданс, частотний словник та порівняно новий тип – лексична будова ідіолекту [55].

Дослідження спадщини видатних особистостей часто завершують укладанням частотних словників їх мови. Такими лексикографічними працями увінчалися, на-

приклад, різноаспектний аналіз творчості українських поетів І. Величковського [4], В. Самійленка [38] тощо.

Частотний словник (далі – ЧС) – тип словника, де наведено кількість вживань, тобто *частоту* певної одиниці мови в обстежених текстах. Звичайно він складається з декількох списків: списків слів, зведених до початкової форми, розташованих за спадом частот і за алфавітом, а також з таких самих списків для словоформ. Мовознавці зазначають, що «... ЧС є важливим інструментом дослідження закономірностей функціонування лексичної системи мови в різних текстах...» [46, с. 78]. В українській лексикографії існують ЧС шести функціональних стилів, проте аналога ЧС письменника чи його окремого твору як, наприклад, ЧС «Войны и мира» Л. Толстого [21], ще не створено.

Усі письменницькі типи словників тісно пов'язані між собою, скажімо, укладаючи словник мови письменника, для розрізнення значень слів та їхнього ілюстрування послуговуються контекстними даними конкордансу. Словникову статтю глосарію будують за таким самим зразком, що й у словнику мови письменника, тільки реєстр слів добирають за іншими критеріями. Той факт, що словопокажчик і словник мови подають, зазвичай, кількість уживань кожного слова, зближає їх із ЧС. Проте є декілька суттєвих рис, на перший погляд, формальних, але насправді принципових, що не дають прирівнювати ЧС із вказаними працями – це послідовність розміщення словникових статей. Щоб словник можна було вважати ЧС, він обов'язково повинен містити ранговий список одиниць, тобто список, у якому першим стоїть слово з найбільшою частотою, на другому – друге за частотністю і т. д. Номер слова за порядком називається рангом. Така форма подання інформації дає можливість визначити обсяг словника слів і словоформ, величину покриття тексту, багатство словника, індексів винятковості й концентрації та інше. Словник мови письменника і словопокажчик подають реєстр слів за алфавітом, що ускладнює обчислення важливих статистичних характеристик тексту, визначення найчастотніших його одиниць, заради чого, зокрема, і укладають ці лексикографічні праці. Ця проблема поглиблюється ще й тим фактом, що вказані словники існують лише в паперовій формі, і виявити найчастотнішу лексику можливо тільки після громіздкої ручної роботи. Для цього доведеться опрацювати все видання і вручну шукати й виписувати слова за спадом частот.

Іншою тенденцією світової письменницької лексикографії вважається поява **письменницьких словників інтегрального типу**, що поєнують у своїй макроструктурі тлумачення, конкорданс, ЧС, а також окремі елементи типу епітетів, порівнянь, метафор тощо. Застосування засобів наочності у вигляді малюнків, таблиць, схем, графіків, діаграм. А це можливо лише за умови використання комп'ютерних корпусів текстів, зокрема тому, що вони уможливлюють автоматичне укладання реєстру усіх лексичних одиниць, ЧС і т. п. Такими інтегральними працями можна вважати, наприклад, словники Ф. Достоєвського, К. Чапека тощо. Для опису мови чеського письменника [81] укладачі словника викристалізували такі параметри: обсяг слововживань, словоформ та лексем кожного тексту й усієї опрацьованої творчості, фонологічні та морфологічні характеристики прямих та переносних, термінологічних та прагматичних номінацій автора, власних та загальних назв, а також фразем та ідіом, авторських метафор, колокацій;

ISSN 2078-5534. Вісник Львівського університету. Серія філологічна. 2013. Випуск 58

кожна лексема у словнику має ранг і кількісну характеристику у шести стилях: проза, драма, публіцистика, поезія, наукова література, кореспонденція. Окремий розділ присвячено статистичним аспектам мови К. Чапека, особливостям його лексикону, тут подано низку статистичних параметрів лексикону письменника, обчислених на основі ЧС, які неможливо було б отримати традиційним ручним способом опрацювання тексту, а саме: обсяг тексту у словоформах, обсяг словника у лексемах, багатство словника, кількість слів за частинами мови, розподіл слів за жанрами та роками тощо. Зокрема, обчислено, скільки слів К. Чапек написав кожного року, а також кількість різних лексем на 1 000 різних слововживань для кожного року. Остання характеристика уможливлює спостерегти, що найбагатшою є лексика поезії, інші ж жанри суттєво їй. Також розкрито багатство словника К. Чапека, порівняно з іншими чеськими письменниками (Я. Гашеком, Б. Грабалом, М. Кундерою). Автори застосували найновішу методику квантитативної лінгвістики для визначення межі між самостійними (семантичними) та службовими (синсемантичними) частинами мови: знайшли для текстів точки *h* (*h*-point) і *k* (*k*-point).

Дослідники творчості Ф. Достоєвського виокремили такі параметри [67]: абсолютна й відносна частотність лексеми в основних жанрах, мікрожанрах (авторська мова, ремарки, пряма мова і т. п.), у мовленні окремих персонажів; квантитативна інформація про граматичні класи лем та афікси. Тут введено поняття «статистичної значущості слова» (що обчислюється за спеціальною формулою), а також «лексичних маркерів підкорпусів». У цьому контексті підкорпусами можуть бути макрожанри, мікрожанри, періоди творчості, окремі тексти, окремі персонажі тощо. Оригінальні дані про позитивні та негативні лексичні маркери Ф. Достоєвського отримано внаслідок порівняння Словника з КТ його сучасників (М. Гоголя, А. Герцена, І. Тургенєва та ін.).

В україністиці параметризовану базу даних створено для творчості Лесі Українки. Це багатоаспектна і багатофункціональна система даних в електронному вигляді. На думку Н. Дарчук і Л. Алексієнко, вона повинна охоплюючи: а) текстовий масив, до якого мають увійти твори з нового академічного видання Лесі Українки, що відповідають рукописним варіантам, або прижиттєві видання, максимально наближені до авторського оригіналу; б) алфавітно-частотний словник поетичного мовлення; в) словник-конкорданс, який подає лексико-семантичну та стилістичну характеристики реєстрового слова; г) словники синонімів, антонімів, паронімів; ґ) словник тропів (епітетів, метонімій, метафор, синекдох, порівнянь тощо); д) інтегрований словник, який містить граматичну, лексико-семантичну, синтаксичну та стилістичну інформацію [1, с. 345]. У лабораторії народознавчої лексики Волинського державного університету розпочато створення ЧС на основі збірок письменниці. Робота проводиться з використанням пакету алгоритмів і програм автоматичного та статистичного аналізу [56].

Іншим проектом, який реалізує Лабораторія комп'ютерної лінгвістики Київського національного університету імені Т. Шевченка під керівництвом Н. Дарчук є «Параметризація українського поетичного мовлення ІІ половини ХХ ст.» (електронна база даних). Це лексикографічна, граматична і статистична електронні бази даних на основі параметризованих текстів п'ятнадцяти видатних поетів 60–90-х років. Фунда-

ментом бази слугує Генеральний ЧС, укладений автоматично за вибірками з поетичних текстів, обсягом 300 000 слововживань. Генеральний ЧС містить 31 000 слів і 69 000 словоформ з такими статистичними параметрами: абсолютна частота, середня частота, середнє квадратичне відхилення, коефіцієнт стабільності [1]. Запропонована методологія обробки лінгвістичних даних в електронній базі є узагальненням комплексу теоретичних і прикладних ідей сучасного мовознавства. Технологія конструювання бази робить її надзвичайно ефективним та раціональним інструментом (вона зберігає багато часу та людських ресурсів) для спеціалістів-філологів різного профілю. Систематизована у серіях словників і списках інформація є важливою для стилістичних, літературознавчих і семантичних досліджень лексичного фонду української поетичної мови в його статиці та динаміці. Електронна база даних зі своєю методологією і технологією допомагає ефективно та оперативно провадити масштабні комплексні філологічні дослідження на рівні сучасної наукометрії.

Отже, проаналізовані проекти лінгвостатистичного опису текстів мають багато спільного: з метою різностороннього опису мови письменників укладено ЧС творів з подібною макро- та мікростуктурою; для цього використано КТ, що, своєю чергою, складаються з підкорпусів основних жанрів, в яких писали автори; КТ мають внутрішнє морфологічне маркування, що уможливлює автоматичне отримання інформації про функціонування частин мови та граматичних категорій.

Подібним за задумом є проект квантитативної параметризації текстів І. Франка, в основі якого лежить електронний КТ [10] зі зовнішнім та внутрішнім маркуванням [18], на першому етапі якого укладено ЧС дев'яти україномовних творів великої прози письменника [9; 11; 14–17; 20; 70]. Коротко окреслимо теоретичні та практичні сфери застосування ЧС певного твору чи письменника.

СФЕРИ ЗАСТОСУВАННЯ ЧАСТОТНОГО СЛОВНИКА ТВОРІВ ПИСЬМЕННИКА

1. Насамперед, природна **мова, зокрема текст, має власні квантитативні закономірності**, тому лінгвістичні дослідження із врахуванням лише якісної її оцінки будуть неповним. «Шлях дисципліни вглиб рано чи пізно наштовхується неминуче на обмеженість якісних методів, на безпорадність неточного способу вираження, на відсутність гіпотез, а також на відсутність теорії», – зауважив знаний німецький лінгвіст Ґ. Альтман [2, с. 6].

Те, що словесне наповнення будь-якого достатньо довгого тексту має власну статистичну структуру, доведено ще на початку XX століття. Вона виявляється у тому, що розподіл частоти одиниць мови в тексті має певну регулярність, він може бути описаний за допомогою певних моделей і теоретичних формул, наприклад, мова і мовлення надають перевагу невеликій кількості одиниць, які часто використовуються і становлять ядро будь-якої мовної чи мовленнєвої підсистеми, тоді як переважна кількість одиниць є низькочастотними (закон переваги) [48, с. 7]. Екстраполюючи це правило на словник письменника, можна стверджувати, що в кожного автора існує строге співвідношення більш і менш частотних лексем. Різниця між статистичною структурою текстів є критерієм унаочнення відмінностей між ними, тобто відмінностей між стилями

ISSN 2078-5534. Вісник Львівського університету. Серія філологічна. 2013. Випуск 58

порівнюваних письменників [58]. «Статистичний аналіз сучасних українських текстів, стильових різновидів української мови засвідчив, що вони відрізняються одиницями всіх рівнів мови, але стилерозрізнювальна потужність цих одиниць різна: найнижчою вона є на фонемному рівні, найвищою – на лексичному та синтаксичному» [47, с. 644]. Власне статистичну структуру лексичного рівня певного твору чи мови письменника і виявляє ЧС.

2. Зазначене **співвідношення більш і менш частотних лексем** у певних письменників читач інтуїтивно сприймає як різноманітний чи одноманітний словниковий запас творів. У науковій літературі (як літературознавчій, так і мовознавчій) часто у дослідженнях оперують такими фразами, як «лексика цього письменника багатша за лексику іншого», «цей письменник використовує більше епітетів, ніж той» і т. д. Ці твердження залишаються непереконливими (або відносними, суб'єктивними, недоведеними) до того часу, поки не зроблено конкретних досліджень, зокрема порівнянь співвідношення частин мови у творах певного письменника з аналогічним співвідношенням у творах іншого, а також таких статистичних характеристик лексикону, як багатство словника чи індекс різноманітності (відношення обсягу словника лексем до обсягу тексту), індекс винятковості (відношення кількості слів із частотою 1 до загального обсягу тексту), індекс концентрації (відношення кількості слів у словнику з абсолютною частотою 10 і більше до загального обсягу тексту). Саме ЧС дає можливість об'єктивно визначити ці величини. «Важливість статистичного дослідження лексики у стилістичному аналізі тексту полягає в його об'єктивності та максимальній точності під час виявлення ваги текстових елементів» [74].

3. Квантитативне співвідношення частин мови. Дослідження частин мови у тексті й реєстрі вважають важливим етапом встановлення індивідуально-авторських особливостей [46, с. 186–187; 79, с. 50]. Деякі закономірності у функціонуванні частин мови встановлено для твору О. Довженка «Поема про море»: «... на іменниковій основі творяться тексти розповідного та описового типу: пейзажі, портретні характеристики. Завдяки іменникам досягається статичність опису, лаконізм, стислість. Прикметники ж увиразнюють ознаки предметів, явищ. На дієслівній основі організовується текст, пов'язаний з описом діяльності людини, різними процесами» [26, с. 102]. Виконані морфологічні дослідження творів І. Франка показують, що його романи дещо відрізняються кількістю вживання частин мови, наприклад, дію романів «Для домашнього вогнища» та «Перехресні стежки» за відносно більшою кількістю дієслів можна вважати динамічнішою, а дію в «Захарі Беркуті», за активністю іменників та прикметників, – більш статичною та описовою. У реєстрі слів І. Франка прикметники та прислівники виявляють подібну тенденцію: кількісно переважають у раніших романах «Борислав сміється» та «Захар Беркут» і зменшуються у пізніших творах «Перехресні стежки», «Великий шум». Стиль І. Франка у частинах мови подібний до усередненого стилю художньої прози за винятком дещо більшої кількості займенників і дещо меншої кількості іменників. Загалом, співвідношення частин мови у різних романах І. Франка приблизно однакове, що свідчить про стабільність авторської манери письма [19].

4. Кількісні реляції між обсягом прямої та авторської мови можуть вважатися

ISSN 2078-5534. Вісник Львівського університету. Серія філологічна. 2013. Випуск 58

статистичним параметром ідіостилю письменника. Різноаспектному дослідженню цих пластів лексики присвячено багато праць, зокрема, наскільки істотно тут можуть різнитися частотні характеристики слів художнього твору представив ЧС сучасної української прози за редакцією В. Перебийніс [66]. Таке розмежування виконано для повістей та романів І. Франка, і виявлено високий рівень діалогічності цих творів – середній обсяг прямої мови тут становить 40,6 %. Це суттєво вищий показник, ніж для художної прози загалом (28,2 %) [12].

5. ЧС письменника (або конкретного його твору) дає інформацію про інші **стилістичні особливості письменника** на рівні лексики, наприклад, кількість слів із територіальних чи соціальних діалектів, їхня частота вживання тощо. Скажімо, А. Бєлий уклав частотні списки іменників, прикметників та дієслів на позначення сонця, місяця, неба, повітря, води в поезіях А. Пушкіна, Є. Баратинського і Ф. Тютчева. Після упущення слововживань, характерних усім трьом авторам, оперуючи рештою з них, він показав особливості сприйняття природи кожним із поетів [5].

6. ЧС є основою для виявлення **ключових слів**. Це «слова, частотність яких в межах визначеної множини текстів (насамперед у творчості одного письменника) виразно переважає середню частотність їх функціонування в літературній мові. Таке відчутне статистичне відхилення від норми появи певних слів вказує на їх особливу значеннєву вартість для цього автора. Визначення відповідного списку ключових слів, а також аналіз стилістичних полів, що творяться довкола них, дають істотну інформацію про механізм поетичної семантики, властивої письменнику або літературній школі» [73, s. 471]. Наприклад, списки таких слів встановлено для польських письменників А. Міцкевича, Б. Пруса, К. Норвіда, М. Конопницької, С. Жеромського, Ю. Словацького та, для напрямів літератури і навіть для цілої національної літератури незалежно від стилю інших [78; 80; 82–84; 86]. Подібну процедуру виконано для згадуваних Ф. Достоєвського та К. Чапека в результаті зіставлення їхнього лексикону із відповідними загальномовними КТ та творами сучасників.

7. ЧС письменників допомагають встановити (ідентифікувати) **авторство творів чи їх фрагментів**, оскільки кожен автор має свої так звані «улюблені» слова чи конструкції, які в його творчості мають найвищу частотність. І, навпаки, можна визначити ті слова, які не функціонували в суспільстві у період його діяльності, тому не могли трапитися в його творчості [41].

8. На основі порівняння ЧС різних функціональних стилів, ЧС письменників і т. ін. можна з високою ймовірністю визначити **лексичну основу мови**, тобто об'єктивно найбільш уживані слова, які треба насамперед засвоїти іноземцеві, що вивчає цю мову. Статистичний підхід до виокремлення словників-мінімумів має довгу традицію в Європі. На жаль, у сучасній лексикографії і дидактичній літературі автори, послуговуючись назвою «найуживаніша лексика», часто не вказують принципів відбору слів до словника, добираючи лексикон інтуїтивно.

9. На підставі зіставлення ЧС письменників, які були сучасниками, можна реконструювати **особливості мовлення певного періоду**, тобто визначити реєстр слів словника мови того часу.

ISSN 2078-5534. Вісник Львівського університету. Серія філологічна. 2013. Випуск 58

Отже, ЧС подає різносторонню статистичну інформацію про текст і має суттєві переваги у виявленні особливостей функціонування усіх одиниць лексичного рівня твору порівняно з іншими лексикографічними працями, об'єктом опису яких є текст письменника. Тому з метою комплексного квантитативного опису ідіостилю письменника насамперед було вирішено укласти повний ЧС текстів І. Франка (далі – ЧСФ).

Методика укладання частотного словника текстів І. Франка. Обсяг спадщини І. Франка за попередніми обчисленнями становить орієнтовно 8 млн. слововживань, то ж у зв'язку зі стрімким розвитком технічних засобів опрацювання мови стає зрозуміло, що вирішити вказане завдання без використання комп'ютера неможливо. Також це зумовлює визначення основних етапів здійснення задуму: опрацювання художньої прози, драми та поезії, літературно-критичних праць, наукових розвідок, епістолярію. У подальшій перспективі можливе й електронне опрацювання рукописів, що, очевидно, є особливо складним для реалізації. ЧС укладають для кожного окремого твору у межах кожного з цих родів й видів літератури із розрізненням прямої та авторської мови.

Зважаючи на складність правописного питання текстів І. Франка, (сам автор у різні періоди творчості писав різними правописами, а сучасні редактори, з метою наближення до сучасного мовлення, вносили ще й свої правки), **джерелами** ЧСФ є академічне «Зібрання творів» у 50-ти томах [63], а також видання творів, що до нього не ввійшли [62]. Усвідомлюючи неповноту цих джерел, потрібно врахувати ширше охоплення матеріалу, зокрема передбачено опрацювання першодруків та прижиттєвих видань І. Франка (їх розглядаємо у кожному конкретному випадку), зіставлення їх текстів. Наприклад, у процесі роботи над ЧС «Перехресних стежок» було зіставлено прижиттєвий першодрук роману 1900 року та академічне його перевидання 1979 року. У результаті виявлено основні відмінності тексту вказаних видань, зокрема відновлено написання літери «ґ» у власних назвах (*Ваґман, Реґіна, Рессельберґ, Ґенцьо, Ґоттесман*), словах, запозичених з польської, німецької, латинської мов (*ґратулювати, ґуст, ведлуґ, абнеґація, резиґнація, морґ* (міра площі), *ґешефт, ґазета*) і давніших, добре засвоєних у мові, запозиченнях (*ґанок, ґрунт, ґрасувати, ґатунок, ґречно*).

З'ясовано правописні відмінності, які не впливають на статистичну структуру тексту: написання назв національностей з великої літери (*Русини, Жиди, Поляки*); використання «ї» на позначення пом'якшення попереднього приголосного на місці сучасного «і» (*лїкар, усї, неділї, дївчата, молодїж*); використання закінчення Н. в. іменників середнього роду «-нє», «-тє» на місці сучасного -ння, -ття (*напруженє, житє*), використання початкової літери «и» (*иньший*) та інші (*фіртка, ґімназияльний, історія, репутацию, усьміх*), а також ті, які мають такий вплив. Серед них – написання зворотної дієслівної частки *-ся* з дієсловом окремо (а частки *-сь* – разом), написання часток *б* і *ж* через дефіс з попереднім словом (*коли-б, чого-ж, повинна-б, се-ж*), написання сучасних прислівників, що починаються колишніми прийменниками, окремо (*з далека, до дому, в низу, у двоє, в десятеро, від разу, до схочу, як найшвидше*). Цікаво, що частка *-ся* у повісті трапляється 2 496 разів у 1 485 дієслівних словоформах – це друге (!) за частотою значення після І / Й.

Аби результати статистичного опису творів можна було коректно порівнювати між собою, важливо, щоби ЧС кожного твору було укладено за єдиною методикою:

«відмінності, інколи вельми відчутні, у методиці побудови словників сильно ускладнюють їх коректне порівняння» [68, с. 300]. Принципи створення ЧСФ розроблено із врахуванням практики укладання трьох ЧС функціональних стилів української мови [7; 8; 13], а також із врахуванням графічних та граматичних відмінностей текстів І. Франка від сучасної літературної мови. Оскільки ці принципи дещо модифікувалися і відшліфовувались у процесі укладання ЧС роману «Перехресні стежки», то приклади, наведені для наочності, взято з цього твору.

Принципи укладання ЧС творів І. Франка. У ЧСФ окремим словом вважаємо послідовність літер (тут апостроф і дефіс розглядаємо як літери) між двома пропусками чи розділовими знаками, тому складні числівники виступають як різні слова. Це стосується займенників типу *абихто*, які в непрямих відмінках з прийменником втрачають єдність написання (*аби з ким*). Написання через дефіс розглядаємо як одне слово (*з-поміж, байдужно-спокійний, адвокат-русин* тощо). ЧСФ подає інформацію про словникові одиниці (тобто леми або слова, зведені до початкової форми) і про словоформи: парадигматичні форми і фонетичні варіанти слів.

Формування ЧСФ зроблено за графічним збігом лем, і кожна частина мови має свою **схему об'єднання словоформ під лемою** (аналогічну, як і в ЧС художньої прози [66], розмовно-побутового [8], наукового [7], офіційно-ділового [13] стилів).

Іменник – до називного відмінка однини зводимо форми всіх відмінків однини та множини. Частоту множинних іменників зводимо до форми називного відмінка множини. Оскільки такі форми іменників як СЕЛЯНИ, РУСИНИ, ЖИДИ, МІЩАНИ і їм подібні позначають осіб і чоловічої, і жіночої статі, їх не об'єднуємо із одниною ані чоловічого, ані жіночого роду.

Прикметник – до називного відмінка однини чоловічого роду зводимо відмінкові форми всіх родів в однині та множині, вищий і найвищий ступені порівняння, за винятком суплетивних форм, які зводимо окремо до називного відмінка однини чоловічого роду вищого ступеня, наприклад, БІЛЬШИЙ, НАЙБІЛЬШИЙ зведено до БІЛЬШИЙ. Суплетивні форми прикметників та прислівників не зводились до звичайного ступеня і в більшості вищезгаданих словопокажчиків і словників.

Займенник – відмінкові форми зводимо відповідно до типу відмінювання.

Числівник – відмінкові форми зводимо відповідно до типу відмінювання.

Дієслово – зводимо до інфінітива синтетичні форми часу (теперішній, минулий і майбутній), форми наказового способу і дієприслівник, а також неособові форми на -но, -то. Аналітичні форми часу вважаємо синтаксичними утвореннями, кожну складову яких зареєстровано як окреме слово.

Дієприкметник – до називного відмінка однини чоловічого роду зводимо відмінкові форми всіх родів в однині та множині, оскільки, за І. Вихованцем, розглядаємо його як різновид віддієслівного прикметника із властивими йому основними категоріями (рід, число, відмінок) та типовою синтаксичною роллю означення [22, с. 54–60.].

Прислівник – зводимо вищий і найвищий ступені порівняння, за винятком суплетивних форм.

Лематизацію слів з частками *-бо*, *-но*, *-таки*, *-то* реалізуємо так: самостійні ча-

ISSN 2078-5534. Вісник Львівського університету. Серія філологічна. 2013. Випуск 58

стини мови зводимо до початкових форм (*говоріть-бо* до ГОВОРИТИ, *брешіть-бо* до БРЕХАТИ, *ходи-но* до ХОДИТИ, *колупнули-таки* до КОЛУПНУТИ, …); у службових частинах мови ці частки зберігаємо (ДУЖЕ-ТО, КОЛИСЬ-ТО, АЛЕ-БО…). Як окремі слова залишаємо також випадки вживання синтетичних особових форм дієслова: ЯКБИ-СТЕ [купували], [потрактував] БИ-С. В окремих випадках, коли початкову форму слова однозначно відтворити складно, зберігаємо словоформу у тому вигляді, в якому вона функціонує в тексті: ПАНЦЮ (ПАНЦЬО?), пор. також словник З. Великодворської [21, с. 8].

У ЧСФ розмежовуємо лексичну та лексико-граматичну (іменник МАТИ і дієслово МАТИ) омонімію, зокрема омографи (*наймúти* (дієслово зі значенням «найняти») і *нáймити* (іменник у множині)); *гóрод* і *горóд*, *мукá* і *мýка*). У цих випадках для розрізнення в дужках подаємо або вказівку на значення (МІЛЯ (*ім'я*)), або на частиномовну належність (МАТИ (*ім.*) і МАТИ (*дієсл.*)). Біля абревіатур та скорочень у дужках вказуємо їх розшифрування: Т. Д. (так далі), О. (отець), Д-Р (доктор) і т. д.

Ті скорочення, які у 50-томнику розшифровано у квадратних дужках (КС[ЬОНДЗ], Т[АК] ЗВ[АНИЙ], ГУЛЬД[ЕНІВ], З[ОЛОТИХ] Р[ИНСЬКИХ]), враховуємо як повну форму (КСЬОНДЗ, ГУЛЬДЕНІВ).

Наголос у словнику подаємо лише в тому випадку, коли він відіграє смислорозрізнювальну роль (СА́МИЙ і САМИ́Й, ВИ́КЛИКАТИ і ВИКЛИКА́ТИ див. також попередній абзац) або поданий у тексті ([річ] НАБУТНА́), оскільки загальна акцентуація мовлення Галичини зламу XIX–XX століть потребує окремого дослідження.

До одної початкової форми (найчастотнішої) зводимо фонетичні варіанти слів, де чергування початкових чи кінцевих літер пов'язане з милозвучністю мови, а саме: дієслова з постфіксами -СЯ / -СЬ; ІТИ / ЙТИ; сполучники ЩОБ / ЩОБИ, І / Й; частки Ж / ЖЕ, Б / БИ; прийменники У / В, З / ІЗ / ЗІ / ЗО, ПІД / ПІДО та деякі інші, а також слова з відповідними префіксами (ВЕСЬ / УВЕСЬ / ВВЕСЬ, ВСЯКИЙ / УСЯКИЙ).

Лексеми ЛЕДВЕ / ЛЕДВО, ТРОХИ / ТРОХА, ТІЛЬКИ / ТІЛЬКО і подібні подаємо в одній словниковій статті, оскільки у прижиттєвому виданні 1900 року послідовно вжито форми ТІЛЬКО, СКІЛЬКО та інші. У 50-томнику розрізнення вказаних форм є штучним [63, т. 1, с. 14–15.]: авторську форму (ТІЛЬКО, СКІЛЬКО, ЛЕДВО) залишено у прямій мові, а сучасну літературну норму подано в інших випадках.

Натомість форми, які відображають особливості мовлення персонажів, зокрема просторічні, подаємо окремо (АДУКА(Н)Т і АДВОКАТ, ПЕРЕГРАФ і ПАРАГРАФ, КАЗЕТА і ҐАЗЕТА).

Лексеми, написані некириличною графікою лематизуємо відповідно до граматики тієї мови, до якої вони належать. Числа, написані цифрами, вважаємо окремим словом.

Цікаво, що в одному лише романі «Перехресні стежки» зафіксовано 45 слів, записаних цифрами, та 208 слів, написаних латинською графікою: німецькою (87), латинською (55), польською (38), французькою (14), чеською (9), їдиш (4), а також один раз – у контексті: «Та ось поперек його дороги простягається чорна стрічка, закривлена по обох краях обрію, мов велике, плазом покладене S». Серед них також трапляються омоніми: *in* (лат.) й *in* (нім.), *a* (лат.) й *a* (польс.), німецькі означені артиклі *die* (жіночого

ISSN 2078-5534. Вісник Львівського університету. Серія філологічна. 2013. Випуск 58

роду і множини), латинське *maxima* функціонує один раз як прикметник (жіночого роду від *maximus*), а другий раз – як іменник (множина від *maximum*). Цікавим є також міжмовний омонім *на* – прийменник в українській мові та частина єврейського словосполучення *на хайрем* (слово честі).

Етапи укладання частотного словника. ЧС кожного твору І. Франка укладаємо напіватоматичним способом у декілька етапів:

1. Створення електронної форми тексту шляхом сканування з подальшим детальним його вичитуванням (графічна та граматична специфіка текстів І. Франка потребує надзвичайно ретельного підходу до цього завдання, оскільки звичайні програми перевірки орфографії не розраховані на тексти західноукраїнського варіанту української мови кінця XIX століття) та вилученням підсторінкових редакційних приміток. Аналізу підлягають усі слова текстів, включно з написаннями латинською графікою та цифрами.

2. Усунення омонімії шляхом додавання до одного із омонімічної пари умовної позначки. Таким способом ці слова стають графічно різними, і програма рахує частоту вживання кожного з них окремо.

3. Автоматичний підрахунок абсолютної частоти кожної словоформи за допомогою спеціально написаної комп'ютерної програми (Автор-розробник – А. Ровенчак). Результатами цієї операції стають *частотний список словоформ за спадом частот*.

4. Лематизацію, тобто зведення словоформ до словникової форми (початкової форми, леми); наприклад, словоформи АДВОКАТА, АДВОКАТАМ, АДВОКАТАМИ, АДВОКАТИ, АДВОКАТОВІ, АДВОКАТОМ – до леми «АДВОКАТ» робимо напівавтоматично. Повністю автоматичну систему розмітки української мови [35, розділ 5, 6.], зорієнтовану на сучасну загальнолітературну норму, у чистому вигляді застосувати до творів І. Франка неможливо через те, що, по-перше, письменник писав західним варіантом літературної мови другої половини XIX – початку XX століття; по-друге, навмисне використовував неправильні форми слів у мовленні персонажів; по-третє, його правопис має графічні особливості (наприклад, лїс, усї); по-четверте, немає підсумованих морфологічно неоднозначних словоформ (омонімів) західного варіанту української літературної мови зламу XIX–XX століть, що допускає неправильну їх автоматичну лематизацію. Скажімо, словоформу ПАНЯ за правилами літературної мови програма мала би вважати іменником IV відміни середнього роду однини в Н. в. (аналогічно до курча, слоненя), тоді як І. Франко вживав його як іменник І відміни жіночого роду однини (зі значенням «пані, жінка»); словоформу *мойого* як слово із прикметниковим закінченням могла звести до *мойий* і т. д.

5. Автоматичний підрахунок абсолютної частоти кожної леми за допомогою спеціально написаної комп'ютерної програми (див. етап 3). Результатом цієї операції стають *ЧС слів за спадом частот*.

6. Розташування всіх зведених лем шляхом сортування в алфавітному порядку. Результатом цієї операції стають допоміжні списки *ЧС слів за алфавітом*.

Отже, ЧС кожного окремого твору І. Франка повинен мати три списки: 1) ЧС слів за спадом частот; 2) ЧС словоформ за спадом частот; 3) ЧС слів за алфавітом. Останній виконує допоміжну роль для знаходження слова. Очевидно, що Проект повної

квантитативної параметризації текстів І. Франка триватиме не один рік. До нього залучено студентів філологічного факультету Львівського національного університету імені І. Франка, зокрема зі спеціальності «прикладна лінгвістика». На сьогодні створено ЧС романів та повістей «Петрії й Довбущуки», «Boa constrictor», «Борислав сміється», «Захар Беркут», «Не спитавши броду», «Для домашнього огнища», «Для домашнього огнища», «Основи суспільності», «Перехресні стежки, «Великий шум» [9; 11; 14–17; 20; 70], деяких казок й оповідань.

Частотний словник творів І. Франка і корпус текстів І. Франка. Під корпусом текстів розуміють зібрання електронних текстів, що відповідає вимогам репрезентативності, збалансованості, розміченості (анотованості), стандартності [27; 75; 77].

Корпус текстів І. Франка буде репрезентативним і збалансованим, якщо міститиме усі тексти письменника в електронній формі. Анотація (маркування, теґування, розмітка) – це позначення в тексті морфологічної, синтаксичної, семантичної інформації (теґів). Морфологічне маркування (вказівка на рід, число, відмінок для іменників, час, вид, спосіб ... для дієслів і т. д.) співзвучне з процедурою лематизації у ЧС. Виникає питання, навіщо спочатку лематизовувати словоформи для ЧС (плодити ще один зайвий(?) продукт), якщо пізніше фактично цю ж процедуру доведеться виконувати для корпусу текстів? Чому відразу не протеґувати текст? Відповідь на це питання лежить у площині співвідношення кількості словоформ (обчислених комп'ютерною програмою автоматично) і кількості слововживань (окремих слів у тексті), тобто у площині обсягів необхідної обробки матеріалу. Зокрема, ЧС роману «Перехресні стежки» зафіксував його обсяг 93 885 слововживань, 19 390 різних словоформ і 9 964 різних слів. Відповідно можна порахувати, що кількість словоформ майже у п'ять разів менша від кількості слововживань: 93 885 / 19 390 = 4,8.

Опрацювавши ці словоформи, вводимо їх разом з початковою формою у словник програми теґування, і вона, натрапивши на цю словоформу в тексті, безпомилково зведе її до правильної леми (про розрізнення омонімів див. вище). Ця процедура також виконує функцію контролю, оскільки дає можливість переконатися, чи результати «ручної» роботи і машини збігаються. Суттєво зазначити, що інтерфейс програми, яка опрацьовує такі складні тексти, як Франківські, обов'язково повинен бути людиноконтрольований, тобто щоби у спірних випадках (наприклад, натрапивши на морфологічно неоднозначні форми) машина ставила питання, а людина, враховуючи контекстний аналіз, сама приймала рішення. Інакше машина помилково присвоїть слову неправильний теґ, а після завершення роботи програми виявити його буде якщо не неможливо, то дуже складно. Доцільно також зазначити, що «ручний» спосіб розмітки не зовсім вийшов з ужитку, оскільки хоча й вимагає непорівнянно більше часу, проте вважається більш якісним. Наприклад, Національний корпус української мови Інституту української мови НАНУ на цьому етапі також теґують вручну [27, с. 111–140.].

На основі ЧС вказаних романів вже зроблено низку лінгвостатистичних досліджень, зокрема проаналізовано частотні розподіли словоформ залежно від кількості складів і фонем, на підставі чого підтверджено закон Менцерата (–Альтманна), за даними

розподілу «ранґ–частота» розраховано параметри законів Ціпфа і Ціпфа–Мандельброта [72], виявлено кількісне співвідношення частин мови у словнику й тексті, зокрема індекси епітетизації, дієслівних означень, ступінь номінальності [19], прямої й авторської мови [12], проаналізовано статистичну структуру творів великої прози І. Франка [9; 11; 14–17; 20; 70].

«Потреба в точних функціональних характеристиках мовних одиниць в різних стилях та жанрах гостро відчувалася в останні десятиліття фахівцями ряду галузей, що мають справу з комп'ютерним аналізом текстів, з укладанням навчальних словників, з теорією та практикою функціонування мови. Пропонована база даних є одним із перших кроків у заповненні цієї лакуни» [45].

Отже, КТ та ЧС творчості І. Франка дають змогу характеризувати ідіолект письменника й окремі його обрані фрагменти (якими можуть бути конкретні твори, авторська та пряма мова, мовлення конкретного персонажа, твори певного періоду тощо) за такими параметрами: обсяг тексту, обсяг словника словоформ і обсяг словника лексем; абсолютна і відносна частота лексем, покриття тексту; багатство словника (індекс різноманітності), середня повторюваність слова, Нарах legomena (слова із частотністю 1), індекс винятковості словника і тексту, індекс концентрації словника і тексту, кількість власних назв; співвідношення частин мови у словнику й тексті, індекси епітетизації, дієслівних означень, ступінь номінальності; тематичні та ключові слова. Зіставлення отриманих характеристик частотності лексем письменника з його сучасниками, а також із їх частотністю в загальномовному КТ, дасть підстави об'єктивно виявляти особливості лексикону автора.

Така квантитативна параметризація текстів І. Франка, що поетапно реалізовується у створенні ЧС творів письменника, дає якісно новий матеріал для різноаспектного дослідження його стилю. Заплановано, що праця виявить цінні дані для укладання словника української мови зламу XIX–XX століть на зразок Словника української мови XVI – першої половини XVII ст. [57], оскільки письменник послуговувався багатьма функціональними стилями у різних царинах людського духу, зокрема художнім, публіцистичним, науковим, епістолярним.

Реалізація цього проекту, окрім самостійної теоретичної й практичної ваги, може слугувати одним з етапів роботи над словником мови письменника (етап визначення реєстру слів), а також над створенням повного корпусу текстів І. Франка.

Список використаної літератури

1. *Алексієнко Л. А.* Принципи створення параметризованої бази даних за поетичними текстами Лесі Українки / Л. А. Алексієнко, Н. П. Дарчук // Леся Українка і сучасність (До 130-річчя від дня народження Лесі Українки) : зб. наук. праць . – Луцьк, 2004 . – С. 344–352.
2. *Альтман Ґ.* Мода та істина в лінгвістиці / Ґ. Альтман // Проблеми квантитативної лінгвістики. – Чернівці, 2005. – С. 3–11.
3. *Апресян Ю. Д.* Избранные труды. Т. 1: Интегральное описание языка и системная лексикография / Ю. Д. Апресян. – М., 1995.

4. *Бадрак Б. М.* Творчість Івана Величковського. Тематика, барокова образність / Бадрак Б. М. : автореф. дис... канд. філол. наук. – К., 2005.
5. *Баевский В. С.* Справочнные труды по поэзии Пушкина и его современников / В. С. Баевский. – [Цит. 03 січня 2013]. – Режим доступу : <http://feb-web.ru/feb/pushkin/serial/v91/ v91-065-.html?cmd=p.htm>
6. *Бацевич Ф.* Імпліцитні текстові засоби в повістях Івана Франка / Ф. Бацевич // Іван Франко – письменник, мислитель, громадянин. – Львів, 1998. – С. 662–666.
7. *Бук С.* 3 000 найчастотніших слів наукового стилю сучасної української мови / С. Бук, Ф. С. Бацевич (наук. ред). – Львів, 2006.
8. *Бук С.* 3 000 найчастотніших слів розмовно-побутового стилю сучасної української мови / С. Бук, Ф. Бацевич (наук. ред). – Львів, 2006.
9. *Бук С.* Інтерпретація лексики роману Івана Франка «Великий шум» у кількісно-статистичному аспекті / С. Бук // Studia Methodologica. – Вип. 30. Тернопіль, 2010. – С. 145–153.
10. *Бук С.* Корпус текстів Івана Франка: спроба визначення основних параметрів / С. Бук // Прикладна лінгвістика та лінгвістичні технології: MegaLing-2006 : зб. наук. праць за ред. В. А. Широкова. – К., 2007. – С. 72–82.
11. *Бук С.* Повість Івана Франка «Захар Беркут» у світлі статистичної лінгвістики / С. Бук // Науковий часопис Національного педагогічного університету імені М. П. Драгоманова : зб. наук. праць. – К., 2010. – С. 172–176. – (Серія 10. Проблеми граматики і лексикології української мови, вип. 6).
12. *Бук С.* Пряма й авторська мова великої прози Івана Франка: лінгвостатистичне дослідження у контексті корпусної лінгвістики / С. Бук // Вісник Львівського університету. 2011. – С. 199–209. – (Серія філологічна, вип. 52 : Мовознавство).
13. *Бук С.* Частотний словник офіційно-ділового стилю: принципи укладання та статистичні характеристики / С. Бук // Лінгвістичні студії : зб. наук. праць. – 2006. – Вип. 14. – С. 184–188.
14. *Бук С.* Частотний словник повісті Івана Франка «Boa constrictor» (редакція 1884 р.) / С. Бук // Стежками Франкового тексту: комунікативні, лінгвосеміотичні, когнітивні, лінгвостатистичні та лексичні виміри прози. – Львів, 2012 (у друці).
15. *Бук С.* Частотний словник роману Івана Франка «Основи суспільності» : Інтерпретація твору крізь призму статистичної лексикографії / С. Бук ; Ф. Бацевич (наук. ред). – Львів, 2012.
16. *Бук С.* Роман Івана Франка «Для домашнього огнища» крізь призму частотного словника / С. Бук // Мовознавство. – 2011. – № 4. – С. 56–66.
17. *Бук С.* Статистична структура роману І. Франка «Борислав сміється» / С. Бук // Уч. зап. Таврич. нац. ун-та им. В. И. Вернадского.– 2010. – С. 114–118. –(Серия «Филология»).
18. *Бук С.* Структурне анотування у корпусі текстів (на прикладі прози Івана Франка) / С. Бук // Українська мова. – 2009. – № 3. – С. 59–71.
19. *Бук С.* Частини мови у словнику та тексті І. Франка (на матеріалі великої прози) / С. Бук // Лінгвістичні студії : зб. наук. праць / Донецький нац. ун-т; наук. ред. А. П. Загнітко. – Донецьк : ДонНУ, 2011. – Вип. 22. – С. 62–65.
20. *Бук С.* Частотний словник повісті І. Франка «Перехресні стежки» / С. Бук, А. Ровенчак // Стежками Франкового тексту (комунікативні, стилістичні та лексичні виміру роману «Перехресні стежки»). Львів, 2007. – С. 138–369.
21. *Великодворская З. Н.* (ред.) Частотный словарь романа Л. Н. Толстого «Война и мир» / З. Н. Великодворская. – Тула, 1978.
22. *Вихованець І.* Частиномовний статус дієприкметників / І. Вихованець // Українська мова. – 2003. – № 2. – С. 54–60.

23. *Галенко І.* Проблеми загального мовознавства в науковій спадщині Івана Франка / І. Галенко // Іван Франко – письменник, мислитель, громадянин. – Львів, 1998. – С. 691–700.
24. *Горбач О.* Вулично-тюремні арготизми у Франковій прозі / О. Горбач // Записки НТШ. – 1963. – Т. 177 (117). – С. 197–206.
25. *Гузар О.* Іван Франко і становлення єдиного українського правопису / О. Гузар // Іван Франко – письменник, мислитель, громадянин. – Львів, 1998. – С. 721–722.
26. *Дарчук Н. П.* Статистичні характеристики лексики як відображення структури тексту Н. П. Дарчук // Мовознавчі студії. – К, 1976. – С. 97–102.
27. *Демська-Кульчицька О.* Основи національного корпусу української мови / О. Демська-Кульчицька. – К., 2005.
28. *Закревська М.* Внесок Івана Франка у розвиток науки про українські діалекти / М. Закревська // Іван Франко – письменник, мислитель, громадянин. – Львів, 1998. – С. 652–656.
29. *Караулов Ю.Н.* Современное состояние и тенденции развития русской лексикографии / Ю. Н. Караулов // Советская лексикография. – М., 1988. – С. 5–18.
30. *Ковалик І. І.* Наукова лінгвістична проблематика в працях Івана Франка / І. І. Кавалик // Іван Франко. Статті і матеріали. – Зб. 12. – К., 1965. – С. 113–121.
31. *Ковалик І. І.* Наукові філологічні основи укладання і побудови Словника мови художніх творів Івана Франка / І. І. Кавалик // Українське літературознавство. Іван Франко. Статті і матеріали. – Львів, 1972. – Вип. 17. – С. 3–10.
32. *Ковалик І. І.* Принципи укладання Словника мови творів Івана Франка / І. І. Ковалик // Українське літературознавство. Іван Франко. Статті і матеріали. – Львів, 1968. – Вип. 5. – С. 174–183.
33. *Ковалик І. І.* Словник мови художніх творів Івана Франка. Пробний зошит / І. І. Ковалик // Українське літературознавство. Іван Франко. Статті і матеріали. – Львів, 1976. – Вип. 26. – С. 63–99.
34. *Ковалик І. І.* Лексика поетичних творів Івана Франка. Методичні вказівки з розвитку лексики / І. І. Ковалик, І. Й. Ощипко, Л. М. Полюга. – Львів, 1990.
35. Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна та ін. – К., 2005.
36. *Кочан І.* Термінологічні проблеми у працях Івана Франка / І. Кочан // Іван Франко – письменник, мислитель, громадянин. – Львів, 1998. – С. 701–703.
37. *Лев В.* Західно-українські елементи мови в ранній творчості Івана Франка / В. Лев // Іван Франко – мислитель і митець : збірка доповідей для відзначення 125-річчя народин і 85-річчя смерті Івана Франка / Редакція Є. Федоренка. – Нью-Йорк ; Париж ; Сідней ; Торонто, 1981. – С. XII.
38. *Малахова О. А.* Мовна модель поетичного світу В. Самійленка : автореф. дис... канд. філол. наук / О. А. Малахова. – Х., 2007.
39. *Мацюк Г. П.* І. Франко про вирішення правописних питань у Галичині / Г. П. Мацюк // Іван Франко і національне відродження. – Львів, 1991. – С. 187–188.
40. *Медвідь А.* Іван Франко: поняття рідної мови / А. Медвідь // Іван Франко – письменник, мислитель, громадянин. – Львів, 1998. – С. 682–686.
41. От Нестора до Фонвизина. Новые методы определения авторства / под ред. Л. В. Милова. – М., 1994.
42. *Ощипко І. Й.* Про укладання словника мови поетичних творів Івана Франка / І. Й. Ощипко // Іван Франко і світова культура. Матеріали Міжнар. симпозіуму ЮНЕСКО (Львів, 11–15 вересня 1986). – Кн. 1. – К., 1990. – С. 82.
43. *Ощипко І.* До вивчення прикметникової та прислівникової синоніміки в художніх творах І. Я. Франка / І. Ощипко // Питання українського мовознавства. – Кн. 4. – Львів, 1960. – С. 97–103.

44. *Панько Т. І.* Мова і нація в естетичній концепції Івана Франка / Т. Панько. – Львів, 1992.
45. Параметризація українського поетичного мовлення ІІ пол. ХХ ст. (електронна база даних) <http://www.mova.info/article.aspx?l1=89&DID=77>. [Цит. 16.01.2013].
46. *Перебийніс В. С.* Частотні словники та їх використання / В. С. Перебийніс, М. П. Муравицька, Н. П. Дарчук. – К., 1985.
47. *Перебийніс В. С.* Статистична стилістика / В. С. Перебийніс // Українська мова : енциклопедія / редкол.: В. М. Русанівський та інші. – 2-ге вид., випр. і доп. – К., 2004. – С. 644.
48. *Перебийніс В. С.* Статистичні методи для лінгвістів: Навчальний посібник / В. С. Перебийніс. – Вінниця: «Нова книга», 2002.
49. *Петличный И. З.* Синтаксис языка призведений Ивана Франко. На материале художественной прозы / И. З. Петличный : автореферат дисс. … докт. филолог. наук. – Львов, 1965.
50. *Полюга Л.* Статистичний аналіз лексики поетичних творів І. Франка / Л. Полюга // Іван Франко і національне відродження. – Львів, 1991. – С. 164–166.
51. *Полюга Л. М.* Слово у поетичному тексті Івана Франка / Л. М. Полюга. – К., 1977. – 167 с.
52. *Рабій-Карпинська С.* Говори Дрогобиччини з узглядненням говірки села Нагуєвичі, Івано-Франківська область / С. Рабій-Карпинська // Іван Франко – мислитель і митець. – С. 128–138.
53. *Сербенська О. А.* Основи мовотворчості журналіста в інтерпретації Івана Франка : текст лекцій / О. А. Сербенська. – Львів, 1992. – 112 с.
54. *Сколоздра О. Р.* Онімна та апелятивна номінація особи в малій прозі Івана Франка: автореф. дис... канд. філол. наук / О. Р. Сколоздра. – Львів, 2009.
55. Словарь языка Достоевского. Лексический строй идиолекта / гл. ред. Ю. Н. Караулов. – М., 2001. – Вып. 1.
56. Словник мови поезії Лесі Українки на комп'ютерній основі / І. Олещук, Н. Тандрик, І. Фролова // Леся Українка і сучасність : зб. наук. праць. – Луцьк, 2008. – Т. 4, кн. 2. – С. 293–298.
57. Словник української мови XVI – першої пол. XVII ст. / НАН України, Ін-т українознавства ім. І. Крип'якевича. – Львів, 1996. – Вип. 3.
58. Статистичні параметри стилів / за ред. В. С. Перебийніс. – К., 1967.
59. *Терлак З.* Словник мови поетичної збірки Івана Франка «Зів'яле листя» / З. Терлак. – Львів, 2010.
60. *Труш О. М.* Синтаксис наукового мовлення Івана Франка (складне речення) : автореф. дис. ... канд. філол. наук / О. М. Труш. – Львів, 2008.
61. *Франко З. Т.* Мова творів Івана Франка / З. Т. Франко // Курс історії української літературної мови. – Т. 1. – К., 1958. – С. 476–519.
62. *Франко І. Я.* Мозаїка : Із творів, що не ввійшли до Зібрання творів у 50-ти томах / І. Франко – Львів, 2002.
63. *Франко І.* Зібрання творів : у 50 т. / І. Франко. – К., 1979–1989.
64. *Ціхоцький І. Л.* Мовна характеристика персонажів у прозі Івана Франка : автореф. дис. … канд. філол. наук / І. Ціхоцький. – Львів, 2004.
65. *Ціхоцький І.* Франкознавчі дослідження кафедри української мови Львівського університету (наукові традиції і перспективи) / І. Ціхоцький // Вісник Львівського університету. – 2010.– С. 91– 98. – (Серія філологічна ; вип. 72).
66. Частотний словник сучасної української художньої прози : у 2 т. / за ред. Перебийніс В. С. – К., 1981.

67. *Шайкевич А. Я.* Статистический словарь языка Достоевского / А. Я. Шайкевич, В. М. Андрющенко, Н. А. Ребецкая. – М., 2003.
68. *Якубайтис Т. А.* О статистических пластах лексики / Т. А. Якубайтис // Вопросы статистической стилистики. – К., 1974. – С. 300.
69. *Best K.-H.* Wortkomplexität im Ukrainischen und ihre linguistische Bedeutung / K.-H. Best, S. Zinenko // Zeitschrift für Slavische Philologie. – 1998. – Bd. 58. – S. 107–123.
70. *Buk S.* The novel Ne Spytavšy Brodu (Without Asking a Wade) by Ivan Franko in the Light of Statistical and Quantitative Linguistics / S. Buk // Glottotheory. – Vol. 4. – 2011.
71. *Buk S.* Writer Vocabularies of Ukrainian Authors / S. Buk // VARIA XVI: Zborník materiálov zo XVI. Kolokvia mladých jazykovedcov (Častá-Papiernička 8.–10.11.2006). – Bratislava, 2009. – S. 57–63.
72. *Buk S.* Statistical Parameters of the Novel Perekhresni stežky (The Cross-Pathes) by Ivan Franko / S. Buk, A. Rovenchak // Quantitative Linguistics. – V. 62: Exact Methods in the Study of Language and Text. – Berlin ; New York, 2006. – P. 39–48.
73. *Głowińska M. and other* Słownik terminów literackich / M. Głowińska, T. Kostkiewiczowa, A. Okopień-Sławińska, J. Sławiński. – Warszawa, 1999. – S. 471.
74. *Lizisowa M.-T. «O* Słowach-kluczach polskiej epiki romantycznej» Edwarda Stachurskiego / M.-T. Lizisowa // Konspekt. Pismo akademii pedagogicznej w Krakowie. – 2002. – No. 2. <www.wsp.krakow.pl/konspekt/konspekt2/lizis.html> [Цит. 28.12.2012].
75. *Meyer C. F.* English Corpus Linguistics: An Introduction / C. F. Meyer. – Cambridge, 2002.
76. *Ovcharenko M.* Stress in Ivan Franko's poetry / M. Ovcharenko // Annals of the Ukrainian Academy of Arts and Science in U. S. – 1960. – Vol. 8. – № 1–2 (25–26). – P. 121–140.
77. Podstawy językoznawstwa korpusowego / red. B. Lewandowska-Tomaszczyk. – Lódź, 2005.
78. *Rudnicka-Fira E.* Słownictwo „Dziadów» A. Mickiewicza w świetle analizy statystycznej (wybór problematyki) / E. Rudnicka-Fira. – Kraków, 1986. – S. 72–98.
79. *Ruszkowski M.* Wskaźnik epitetyzacji w badaniach stylistycznych / M. Ruszkowski // Respectus Philologicus. – 2004. – № 5(10). – S. 48–53.
80. *Skubalanka T.* Podstawy analizy stylistycznej. Rozważania o metodzie / T. Skubalanka. – Lublin, 2001. – S. 263.
81. Slovník Karla Čapka / F. Čermák. (Hl. ed.). – Praha, 2007.
82. *Stachurski E.* Próba ustalenia listy wyrazów charakterystycznych dla tekstu literackiego (niezależnie od epoki) / E. Stachurski // Studia historyczno-językowe III, red. K. Rymut. – Kraków, 1998. – S. 265–271.
*83. Stachurski E. S*łownictwo polskich naturalistów. Badania statystyczne / E. Stachurski. – Kraków, 1989.
84. *Stachurski E.* Słowa-klucze polskiej epiki romantycznej / E. Stachurski. –Kraków, 1998.
85. Vlasenko-Bojchun A. Franko's contribution to onomastics / A. Vlasenko-Bojchun // Іван Франко – мислитель і митець : зб. доп. для відзначення 125-річчя народин і 85-річчя смерті Івана Франка. – Нью-Йорк ; Париж ; Сідней ; Торонто, 1981. – С. 120–127.
86. *Zelech W.* Statystyczna struktura poezji Marii Konopnickiej / W. Zelech // Rocznik Naukowo-dydaktyczny WSP w Krakowie. z. 192, Prace językoznawcze IX. – Kraków, 1997. – S. 285–293.

Стаття надійшла до редакції 19.10.2012
Прийнята до друку 26.10.2012

ISSN 2078-5534. Вісник Львівського університету. Серія філологічна. 2013. Випуск 58

QUANTITATIVE PARAMETRIZATION OF TEXTS WRITTEN BY IVAN FRANKO: THE PROJECT AND ITS REALISATION

Solomiya BUK

Ivan Franko National University of Lviv,
Department for General Linguistics,
1, Universytetska St., UA-79000 Lviv, Ukraine,
tel.: (032) 239 47 56, e-mail: solomija@gmail.com

In the article, the project of quantitative parameterization of all texts by I. Franko is manifested. It can be made only by using modern computer techniques after the frequency dictionaries for all Franko's works are compiled. The paper describes the application spheres, methodology, stages, principles and peculiarities in the compilation of the frequency dictionary of the second half of the 19th century – the beginning of the 20th century.

The relation between the I. Franko frequency dictionary, explanatory dictionary of writer's language and text corpus is discussed.

Key words: frequency dictionary, writer's dictionary, idiostyle of I. Franko, linguostatistic analysis of text, text corpus, quantitative parameters of text.

КВАНТИТАТИВНАЯ ПАРАМЕТРИЗАЦИЯ ТЕКСТОВ ИВАНА ФРАНКО: ПРОЕКТ И ЕГО РЕАЛИЗАЦИЯ

Соломия БУК

Львовский национальный университет имени Ивана Франко,
кафедра общего языкознания,
ул. Университетская, 1, Львов, Украина, 79000,
тел.: (032) 239 47 56, e-mail: solomija@gmail.com

В статье предложен проект квантитативной параметризации текстов И. Франко, который возможно реализовать, создав частотный словарь всех произведений писателя и только при применении современных компьютерных программ. Указаны сферы использования, этапы, методика, принципы и специфика составления частотного словаря языка второй половины XIX – начала XX веков, на котором писал И. Франко. Описаны соотношения частотного словаря И. Франко со словарем языка писателя и корпусом текстов.

Ключевые слова: частотный словарь, словарь языка писателя, идиостиль И. Франко, лингвостатистический анализ текста, корпус текстов, квантитативные параметры текста.