ЛЕКСИКОГРАФІЯ ТА МІЖМОВНІ ЗВ'ЯЗКИ

УДК 811.161.2'1'324'38 (038)

Соломія Бук
Львівський національний університет імені Івана Франка

СТАТИСТИЧНІ ХАРАКТЕРИСТИКИ РОМАНУ ІВАНА ФРАНКА «ОСНОВИ СУСПІЛЬНОСТІ»
(НА ОСНОВІ ЧАСТОТНОГО СЛОВНИКА ТВОРУ)

У статті подано основні статистичні характеристики тексту роману І. Франка «Основи суспільності», отримані на матеріалі частотного словника твору. А саме, індекси різноманітності (багатство словника), винятковості, концентрації, співвідношення між рангом слова та величиною покриття тексту тощо. Словник укладено за методикою та принципами, розробленими у проекті створення комплексного корпусу текстів І. Франка.

Ключові слова: частотність слова, частотний словник, ранг, обсяг тексту, індекс різноманітності, індекс винятковості, індекс концентрації, індекс покриття, Іван Франко.

In the article, the main statistical parameters of *Osnovy suspil'nosty (Pillars of Society)*, a novel by Ivan Franko, are calculated: variety, exclusiveness, concentration indexes, correlation between word rank and text coverage, etc. They are obtained from the frequency dictionary for this text, which was compiled due to the methodology and the principles developed in the project of Ivan Franko complex text corpus.

Keywords: word frequency, frequency dictionary, rank, number of the word occurrences, variety, exclusiveness, concentration indexes; text coverage, Ivan Franko.

1. Вступ. Різноаспектне дослідження мови національних письменників — важливе питання для кожного народу. Комп'ютерні технології і, зокрема, базована на них корпусна лінгвістика відкривають нові можливості для якісного та кількісного аналізу ідіолекту письменника. Так, створено корпуси текстів Арістотеля, В. Гюго, В. Шекспіра, Дж. Джойса, Ф. Достоєвського, Г. Сковороди, В. Шевчука, поезії Т. Шевченка та інших зарубіжних та українських письменників. З метою комплексного вивчення мови І. Франка також розпочато проект створення електронного корпусу його текстів [1], що на першому етапі охоплює велику прозу письменника. Морфологічно анотований корпус текстів дає, зокрема, можливість автоматично укладати частотні словники як до окремих творів, так і до їх множин. Дослідженню частотних словників та статистичних характеристик тексту присвячені такі праці: [2–5]. Зокрема, проект автоматичної кількісної параметризації текстів І. Франка викладено в [6]. Так, у межах цього проекту було укладено ЧС до романів «Перехресні стежки» [7] та «Для домашнього огнища» [8], а також конкорданс до «Перехресних стежок», доступний в Інтернеті [9]. У запропонованій статті подано статистичні характеристики роману «Основи суспільності», отримані на основі ЧС роману, який ще не був об'єктом дослідження у квантитативному аспекті.

2. Джерельна база електронного корпусу роману. «Основи суспільності» І. Франко друкував окремими розділами в журналі «Житє і слово» (1894–1895), пишучи його від номера до номера. У листі до Драгоманова (від 01.02.1895) він зазначає, що перервав цей процес, бо хотів видати твір окремою книгою. Проте за життя автора роман окремим виданням не вийшов і

залишився незавершеним. Аж 1918 р. твір побачив світ окремою книгою в Києві з передмовою Василя Верниволі [10].

В основу сюжету покладено судовий процес 1889 р. над поміщицею та її сином, які намагались вбити ксьондза з метою грабунку, щоби покращити своє фінансове становище. І. Франко, як співробітник газети «Kurjer Lwowski», був присутній на процесі, а примірник книги стенограм процесу зберігається в його особистій бібліотеці. За цим сюжетом знято серіал Олега Бійми «Злочин з багатьма невідомими» (1993).

За основу для електронного корпусу роману взято текст твору з 50-томника (1979) з урахуванням оригіналу (1894–1895). У результаті зіставлення цих текстів виявлено значну кількість правописних відмінностей: написання зворотної дієслівної частки -ся окремо, а -сь — разом (*нахиляючи ся, присягай ся, наближу ся, сердила ся*), використання на місці сучасної «ї» після голосних літери «і» або буквосполуки «йі» (*стоіш, ясноі, єі, неі, своєі; йісти, подойів, стойіш, пойіхав, йій, йіх), написання «и» в закінченні Р. в. однини іменників ж.р (пивници, смерти).* Загалом, правопис першодруку «Основ суспільності» збігається із першодруком роману «Для домашнього огнища» (1897), а правопис окремого видання 1918 р. — із правописом «Перехресних стежок» (1900) і «Великого шуму» (1908).

3. Аналіз статистичних характеристик тексту твору. ЧС роману «Основи суспільності» укладено за методикою та принципами, розробленими в згаданих проектах створення корпусу текстів [1] і комплексного кількісного опису творів І. Франка [6], та спільними для двох попередніх ЧС до творів письменника [7; 8]. Для кількісного дослідження лексем у текстах, написаних флективними мовами (до яких належить українська), дуже важливою є лематизація, тобто зведення словоформ до початкової, словникової форми. Так, у тексті «Основи суспільності» здійснено лематизацію усіх частин мови відповідно до типу відмінювання. Роботу виконано напівавтоматичним способом з використанням комп'ютерного лематизатора, що містить словник словоформ двох попередніх романів з відповідними до них лемами.

У кожній словниковій статті подано абсолютну та відносну частоту лексичної одиниці, а також міру покриття тексту.

Таблиця 1.

Частотний словник роману І. Франка «Основи суспільності» (зразок)

Ранґ	СЛОВО	Абс. част.	Відн. част., %	Покрит., %
...	...	...		...
78	БОГ	125	0,186	45,56
79	ПАН	125	0,186	45,74
80	ГОЛОВА	123	0,183	45,93
81	ЧУТИ	123	0,183	46,11
82	СЛОВО	121	0,180	46,29
83	ДЕМЕНЮК(прізв.)	120	0,179	46,47
84	ХОТІТИ	118	0,176	46,64
85	ПЕРЕД(прийм.)—113;ПЕРЕДО—4	117	0,174	46,82
...	...	...		...

В одну словникову статтю зведено фонетичні варіанти слів: дієслова з постфіксами *-ся / -сь*; сполучники *щоб / щоби, і / й;* частки *ж / же, б / би, ще/іще*; прийменники *у / в, з / із / зі / зо* та деякі інші; слова з відповідними префіксами (*вложити/уложити, весь / увесь / ввесь, всякий / усякий* тощо).

Розрізнено омонімію: *як* (прислівник/сполучник/частка), *та* (займенник/ сполучник/ частка), *саме* (займенник/прислівник), *що* (займенник/сполучник/частка), *а* (сполучник / частка / вигук), *брязкало* (іменник/дієслово), *виводи* (іменник/дієслово), *випали* (мин. час від *випасти* і наказ. спосіб від *випалити)* тощо.

Окремо розглянуто скорочення (*гр. (граф), р. (рік), т.є. (то єсть), т.д. (так далі)*) та фрагменти слів (тобто обірвані слова: *про...* (уривок від *прошу), ясне...* (уривок від *ясневельможна*), дві... (уривок від двісті), *...десят* (уривок від *п'ятдесят*).

За обсягом ЧС роману І. Франка «Основи суспільності» — повний, оскільки охоплює власні назви, цифри, слова, написані некириличним алфавітом тощо. Він складається із трьох частин:

1) частотний список слів (лем), 2) частотний список словоформ та 3) алфавітно-частотний список слів (лем). У результаті здійсненої роботи отримано основні статистичні характеристики тексту.

1. Обсяг роману (*N*) — 67 174 слововживань, серед них 149 польських, німецьких, латинських і французьких.
2. Кількість різних словоформ ($V_ф$) — 15 487.
3. Кількість різних слів (*V*) — 8 400.
4. Багатство словника (індекс різноманітності), тобто відношення обсягу словника лексем до обсягу тексту (*V/N*), становить 0,125.
5. Середня повторюваність слова в тексті, тобто відношення обсягу тексту до обсягу словника лексем (*N/V*) — 8,0. Іншими словами, кожне слово в середньому вжито в досліджуваному тексті приблизно 8 разів.
6. Кількість hapax legomena — слів із частотою 1 — (V_1 = 4348) складає 6,47% тексту і 51,76% (більше за половину!) словника.
7. Індекс винятковості у тексті, тобто відношення V_1 до обсягу тексту (V_1/N) становить 0,065.
8. Індекс винятковості у словнику, тобто відношення V_1 до обсягу словника (V_1/V), становить 0,52. Два останні числа — показники варіативності лексики.
9. Протилежним до індексу винятковості є індекс концентрації — відношення кількості слів з високою частотою (10 і вище, їх у тексті роману 51 021 і в словнику — 796) до обсягу тексту (V_{10}/N =75,95) або словника (V_{10}/V = 9,48).

Відносно невелика кількість високочастотної лексики в словнику лем (і, відповідно, низький індекс концентрації) та порівняно велика кількість слів із частотою 1 (і, відповідно, високий індекс винятковості) свідчать про неабияке різноманіття лексики роману.

У ЧС слова розміщено в порядку спадання частот: слово з найбільшою частотою має ранг 1, наступне — 2 і т. д. Таке подання інформації дає змогу обчислити, яку частку тексту (покриття) становлять слова з найбільшою частотою. Величину покриття тексту для певного ранґу обчислюють як відношення суми абсолютних частот усіх слів з меншими ранґами до загальної кількості слів у тексті. Співвідношення між ранґом слова та покриттям тексту подано в Таблиці 2.

Таблиця 2.

Співвідношення між ранґом слова та величиною покриття тексту

Ранґ	Покриття	Ранґ	Покриття	Ранґ	Покриття
1	3,39%	200	59,20%	1500	83,11%
5	11,85%	300	64,54%	2000	86,30%
10	18,80%	400	68,03%	3000	90,40%
25	30,45%	500	70,66%	5000	94,94%
50	39,36%	600	72,76%	7000	97,92%
75	44,99%	750	75,27%	8000	99,40%
100	49,18%	1000	78,53%	8400	100,00%

З Таблиці 2 видно, що перші за частотністю 25 слів покривають майже 31% тексту, перші 100 слів — майже 50%, 1000 слів — майже 79% тексту.

Біля кожного слова в базі даних вказано його частиномовну належність, тому за допомогою комп'ютерної програми можна обчислити кількість слів кожної частини мови. Кількісне співвідношення частин мови в романі зручно подати в таблиці:

Таблиця 3.

Розподіл лексем за частинами мови в романі І. Франка «Основи суспільності»

частини мови	**слів у тексті**		**слів у словнику**	
службові	*17382*	25,88%	*262*	3,12%
іменники	*15758*	23,46%	*2849*	33,92%
дієслова	*12164*	18,11%	*2859*	34,04%
прикметники	*4803*	7,15%	*1567*	18,65%
займенники	*10476*	15,60%	*81*	0,96%
прислівники	*6006*	8,94%	*716*	8,52%
числівники	*558*	0,83%	*51*	0,61%
фраґменти	*27*	0,04%	*15*	0,18%
	67174	**100,00%**	**8400**	**100,00%**

Як видно з Таблиці 3, найчастотнішими словами в тексті очікувано є службові частини мови (17 382). Цей замкнутий і відносно небагатий клас слів виявляє найбільшу активність: займаючи всього 3,12 % у словнику, вони становлять майже 26 % тексту. Подібний замкнутий клас слів становлять займенники. Їх поведінка в тексті ще активніша: загальна кількість займенників у різних формах 10 476, що становить 15,6% від загального обсягу роману. Усього ж різних займенників (лем) тільки 81, що становить 0.96% від загальної кількості слів у словнику. Іншими словами, 81 займенник займає менше ніж 1 % у словнику, проте аж майже 16 % у тексті.

Прислівники та числівники займають приблизно однакову частку в словнику (відповідно 8,94 % і 0,83 %) й тексті (відповідно 8,52 % і 0,61 %).

Відносна кількість усіх інших частин мови в словнику навпаки перевищує їх відносну кількість у тексті. Іменники становлять 33,92 % словника і 23,46 % тексту; частка дієслів у словнику (34,04 %) майже вдвічі більша за їх частку в тексті (18,11 %); а частка прикметників у словнику (18,65 %) перевищує їх частку в тексті (7,15 %) більше ніж удвічі. Ці факти свідчать, що найбагатшим шаром у лексиці «Основ суспільності» є дієслова та прикметники.

Якщо ж з іменників забрати власні імена героїв (а вони, як відомо, належать до найчастотнішої групи слів у художньому творі: Нестор (374), Олімпія (242), Адась (136), Гапка (131), Деменюк (120), Гадина (114), Параска (84), Маланка (66) тощо, усього їх у словнику роману 83 та в тексті 1744), то ця частина мови все одно буде мати лише трохи більшу різноманітність: 32,93 % у словнику та 20,86 % у тексті.

5. Висновки. Роман «Основи суспільності» — важливий твір у творчості Івана Франка. Лінгвостатистичний опис його лексики в цій статті здійснено вперше. ЧС роману дав змогу отримати важливі статистичні характеристики тексту, такі як обсяг роману, кількість різних словоформ та слів, індекс різноманітності (багатство словника), середня повторюваність слова в тексті, кількість hapax legomena, індекс винятковості у тексті та у словнику, індекс концентрації, співвідношення між рангом слова та величиною покриття тексту. На основі даних про розподіл частин мови в романі, отриманих завдяки морфологічній анотації корпусу твору, виявлено особливості поведінки цих класів слів у тексті та словнику, зокрема те, що найбільшу різноманітність показують прикметники та дієслова.

ЧС аналізованого твору укладено за методикою, яка розроблена в проекті комплексного квантитативного опису творів І. Франка, що в перспективі стане запорукою можливості коректного зіставлення статистичних характеристик тексту «Основ суспільності» з аналогічними даними інших романів І. Франка, виявити динаміку росту чи спаду багатства словника, міру появи нових слів в ідіолекті письменника тощо.

1. Бук С. Корпус текстів Івана Франка: спроба визначення основних параметрів / С. Н. Бук // Прикладна лінгвістика та лінгвістичні технології: MegaLing-2006: Зб. наук. пр. — К.: Довіра, 2007. — С. 72–82. 2. Перебийніс В. С. Статистичні методи для лінгвістів: Навчальний посібник / В. С. Перебийніс. — Вінниця: Нова книга, 2002. — 168 с.; 3. Перебийніс В. С. Частотні словники та їх використання / В. С. Перебийніс, М. П. Муравицька, Н. П. Дарчук. — К.: Наук. думка, 1985. — 204 с. 4. Статистичні параметри стилів [за ред. В. С. Перебийніс]. — К.: Наук. думка, 1967. — 260 с.; 5. Алексеев П. М. Частотные словари: Учебное пособие / П. М. Алексеев. — СПб.: Изд-во С.-Петерб. ун-та, 2001. — 156 с. 6. Бук С. Квантитативна параметризація текстів Івана Франка: спроба проекту / С. Н. Бук // Іван Франко: Студії та матеріали: Зб. наук. статей. — Л.: ЛНУ ім. І. Франка, 2010 (у друці), див. препринт arXiv:1005.5466v1 [cs.CL] за адресою http://arxiv.org/abs/1005.5466. 7. Бук С. Частотний словник роману Івана Франка «Перехресні стежки» / С. Бук, А. Ровенчак // Стежками Франкового тексту (комунікативні, стилістичні та лексичні виміру роману «Перехресні стежки»). — Л.: Видавничий центр ЛНУ ім. І. Франка, 2007. — С. 138–369. 8. Бук С. Роман Івана Франка «Для домашнього огнища» крізь призму частотного словника [Електронний ресурс] / С. Н. Бук // Препринт arXiv:1006.0153v1 [cs.CL]. — Режим доступу: http://arxiv.org/abs/1006.0153. 9. Бук С. Он-лайн конкорданс роману Івана Франка «Перехресні стежки» [Електронний ресурс] / С. Бук, А. Ровенчак. — Режим доступу: http://www.ktf.franko.lviv.ua/~andrij/science/Franko/concordance.html. 10. Франко І. Основи суспільности: Повість / Іван Франко. — К.: Українська накладня, 1918. — 304 с.