Ученые записки Таврического национального университета им. В. И. Вернадского
Серия «Филология. Социальные коммуникации». Том 23 (62). № 3. 2010 г. С. 114-118.

НАЦИОНАЛЬНЫЕ ЯЗЫКИ И ИХ ВЗАИМОДЕЙСТВИЕ

УДК 811.161.2’1’324’38 (038)

СТАТИСТИЧНА СТРУКТУРА РОМАНУ ІВАНА ФРАНКА “БОРИСЛАВ СМІЄТЬСЯ”

Бук С.

Львовский национальный университет имени Ивана Франко, г. Львов, Украина
E-mail: solomija@gmail.com

У статті з'ясовано поняття статистичної структури тексту, яку на лексичному рівні, як правило, визначають за даними частотного словника (ЧС). На підставі ЧС роману І. Франка “Борислав сміється” отримано кількісні характеристики тексту та зіставлено їх з відповідними даними для інших романів письменника.

Ключові слова: статистична структура тексту, частотний словник, кількісні характеристики тексту, багатство лексики.

Постановка проблеми. Статистична структура тексту (ССТ) — розподіл частоти одиниць мови в тексті, що має певну регулярність. Він різний для різних мовних елементів. Наприклад, статистичні параметри стилів, що встановлюються на різних рівнях, мають неоднакову стилерозрізнювальну потужність для різних пар стилів: більш споріднені стилі найвиразніше розмежовуються на синтаксичному рівні, менш споріднені — на лексичному ” [7, с. 239]. ССТ розуміється як його кількісна організація, як його модель [6, с. 130].

ССТ описують певні закони й теоретичні формули (Закон переваги, Закон Ціпфа, Закон Мандельброта тощо). СТТ на рівні лексем, як правило, визначають за даними частотного словника (ЧС), що подає до кожної реєстрової одиниці її частотність, тобто кількість вживання у тексті. Різниця між ССТ є одним із критеріїв унаочнення відмінностей між різними текстами, стилями, авторами. Визначенню й уточненню структурно-кількісних закономірностей будови тексту присвячені роботи В. Перебийніс, Н.Дарчук, М. Муравицької, М. Арапова, Ю. Тулдави, Р. Фрумкіної, В. Левицького, Б. Головіна та інших авторів, які розглядають ЧС як лінгвістичну модель, вивчення якої сприяє виявленню законів функціонування мови та мовлення. Так, ЧС укладено до творів багатьох письменників (В. Шекспіра, В. Гюго, К. Чапека, М. Павича, Ф. Достоевського), у т. ч. й до поетичних [5] та прозових [2-4] творів І. Франка.

Роман І. Франка “Борислав сміється” (1881) привертав увагу багатьох мовознавців (О. Горбача, З. Франко, О. Сербенської, І. Ощипко, С. Жилко, І. Ціхоцького, І. Петличного), довгий час твір входив у шкільну програму як такий, у якому вперше відображено початкові форми революційної боротьби робітництва та стихійне пробудження його класової свідомості, проте у лінгвостатистичному

ракурсі він аналізується вперше. Такий підхід до твору є логічною частиною проекту квантитативної параметризації великої прози І. Франка [1].

Важливою проблемою укладення ЧС є добір джерел. У нашому випадку ними стали першодрук 1881-1882 рр. (див. рис. 1) та видання твору 1979 р. [8], які поза правописними відмінностями є ідентичними. Цікавою орфографічною деталлю тексту є використання у власних назвах на місці сучасного "г" літери латинської графіки "g": *Gотліб, Гаммершляг.* Твір друкувався у журналі "Сьвіт", видання якого припинилось, і роман залишився незакінченим. До "Борислава..." І. Франко, як відомо, не хотів ні повертатися, не хотів його ні дописати, ні видати окремою книгою, мабуть тому, що його захоплення соціалістичним вченням з часом зменшилося, а зі смертю М. Драгоманова в 1895 р. втратилось.

Ві Львові, дня 25. лютого 1882.

СЬВІТ

Рочник II. ІЛЮСТРОВАНА ЛІТЕРАТУРНО-ПОЛИТИЧНО-НАУКОВА ЧАСОПИСЬ. Число 14 (2 за 1882).

БОРИСЛАВ СЬМІЄСЯ.

VIII.

Звільна, важкою ходою повзли одностайні робучі дні в Бориславі. Бенедьо працював цілими днями при своїй фабриці, нитчував пляни будинків, заправляв робітниками, наглядав за вчасним довозом цегли, каміня, вапна і всього потрібного, — а при всім тім обходився з робітниками так по братерськи, так щиро та приязно, немов хотів на кождім кроці показати їм, що він їм рівний, їх брат і такий же бідний робітник, як усі вони, немов хотів перепрашати їх за те, що отсе, не по своїй волі став над ними наставником. А вечорами, по роботі, він верзав до пізної ночи ходив в важкій задумі по болотистих улицях Борислава, заглядав до брудних шинків, до тісних хат та комірок, де жили робітники, за-

не згинути! — се був загальний оклик, загальний стогін, що хмарою носився понад головами тих тисячів висохлих, посинілих, виголоджених людей. Небо і землю запер бог на зелізні ключі, — вся надія мужиків-хліборобів вигоріла разом з їх житом та вівсом на порудівих від спраги загонах. Худоба гибла за недостачею паші. Не осталося нічого, як іти на заробітки, а заробітків то як раз і не було тоді ніяких в нашім Підгірю — крім Борислава. От і поперлися туди бідні люде з усіх сторін, хапаючися за послідне, так як той потапаючий хапався за стебелинку. Небо і землю запер бог на зелізні ключі, — а бідні люде думали, що бориславські богачі будуть милостивійші від бога і отворять їм брами своїх богацтв!...

А бориславські богачі тілько того й бажали! Вони від давна потішалися надією, що аж порядний голод причиниться до великого зросту їх „гешефтів". І ось вони не помили-

Рис. 1 — Фрагмент першої сторінки VIII розділу роману Івана Франка "Борислав сміється", надрукованого у львівському часописі "Сьвіт" 25 лютого 1882 р. (центральну частину сторінки займає портрет Михайла Максимовича).

Проектуючи ЧС "Борислава..." на прижиттєве видання, цікаво відзначити написання частки -ся разом, на відміну від інших Франкових романів, в оригінальних виданнях яких саме вона займає стало друге за частотністю місце.

У ЧС розрізнено омонімію, зведено фонетичні варіанти слів, здійснено структурну, морфологічну та ономастичну анотації. Так, у романі виявлено 80 власних назв (у 2204 слововживаннях), серед яких кількісно домінують власні назви головних персонажів: *Бенедьо* (306), *Герман* (261), *Леон* (252), *Готліб* (128), *Рифка* (127), *Матій* (119) та назва міста *Борислав* (183).

У результаті укладення ЧС роману “Борислав...” було отримано основні його кількісні характеристики. Обсяг тексту (кількість слововживань): 77 456, тобто серед творів І. Франка, до яких укладено ЧС, це другий за величиною після “Перехресних стежок” (93 888), що перебільшує “Основи суспільності” (67 174), “Для домашнього огнища” (44 840), “Великий шум” (37 005). Відповідно, і обсяг словника словоформ (16 064), і обсяг словника лексем (8 576) у цьому творі також більші.

Зведений список лем п'яти згаданих творах великої прози містить 20980 різних слів, що значно перевищує Словник мови Т. Шевченка (6 116) та Г. Квітки-Основ'яненка (11 772).

Багатство словника, яке обчислюється як відношення обсягу словника лексем до обсягу тексту, обернено пропорційне довжині тексту, тобто, чим довший текст, тим потенційно менше з'являється у ньому нових слів [6, с. 143]. Тому цей показник у “Бориславі...” (8 572/77 456 = 0,111) більший, ніж у “Перехресних...” (0,106) і менший за інші романи: “Основи...” (0,125), “Для домашнього...” (0,145), “Великий шум” (0,175),

Середня повторюваність слова у тексті — величина, обернена до попередньої, і становить 9,04, тобто в середньому кожне слово трапляється у тексті 9 разів. Проте ця величина дуже узагальнена, адже кількість слів, що трапилися в романі один раз (hapax legomena) — 4 370, тобто вони займають більше, ніж половину словника (50,98%). Приблизно такі ж результати і для інших творів: у “Перехресних...” вони займають 49,18% словника, “Основах...” — 51,76%, “Для домашнього...” — 51,85%, “Великому шумі”— 56,7%. Саме в них криється основне багатство лексикону письменника.

За допомогою величини hapax legomena обчислюють індекс винятковості словника (0,51) та тексту (0,056). Ці величини логічно корелюють із відповідними даними ЧС інших романів: “Перехресні...” (0,49/0,052), “Основи...” (0,52/0,065), “Для домашнього...” (0,52/0,075), “Великий шум” (0,57/0,099).

Протилежним до індексу винятковості є індекс концентрації словника/тексту, що вказує частку словника/тексту, яку займають слова із великою частотою (умовно із частотою 10 і більше). У тексті “Борислава...” таких слів 61 328, що становить 79,18% обсягу його тексту, а в словнику — 916, що становить 10,69% його словника. У “Перехресних...” відповідні показники логічно більші: 74 651 (79,5%) та 1 123 (11,3%), в решти творах меншого обсягу ці показники логічно менші: в “Основах...” 51 021 (75,95%) та 796 (9,48%), “Для домашнього...” 32 516 (72,5%) та 598 (9,2%), “Великому шумі” 25 456 (68,8%) та 479 (7,4%). В принципі, чим менше у тексті високочастотних слів, тим різноманітніша лексика тексту і навпаки. Отже, справджуються попередньо отримані результати.

Унаочнити розподіл кількості слів із певною частотою залежно від цієї частоти можна на графіку (рис. 2).

Треба зауважити, що цей графік відрізняється від рангово-частотної кривої, хоча й отриманий за її даними і візуально дуже до неї подібний.

На підставі отриманих рангово-частотної залежності та частотного спектру тексту було знайдено значення т. зв. *h*- і *k*-точки. Координати цих точок визначаються з умови рівності значень функції і аргумента [10, p. 17, 35]. Отримане значення *h*-точки 109 означає, що слово з частотою 109 має ранг 109, а значення *k*-точки 20

означає, що у тексті є 20 слів із частотою 20. Існують гіпотези, що ці точки можуть слугувати межею зон словника, де переважають семантичні (самостійні) або синсемантичні (службові) частини мови [10, p. 18, 37]. Суттєва відмінність між цими значеннями, а також аналіз частотного списку, дає підстави стверджувати, що така границя знаходиться посередині.

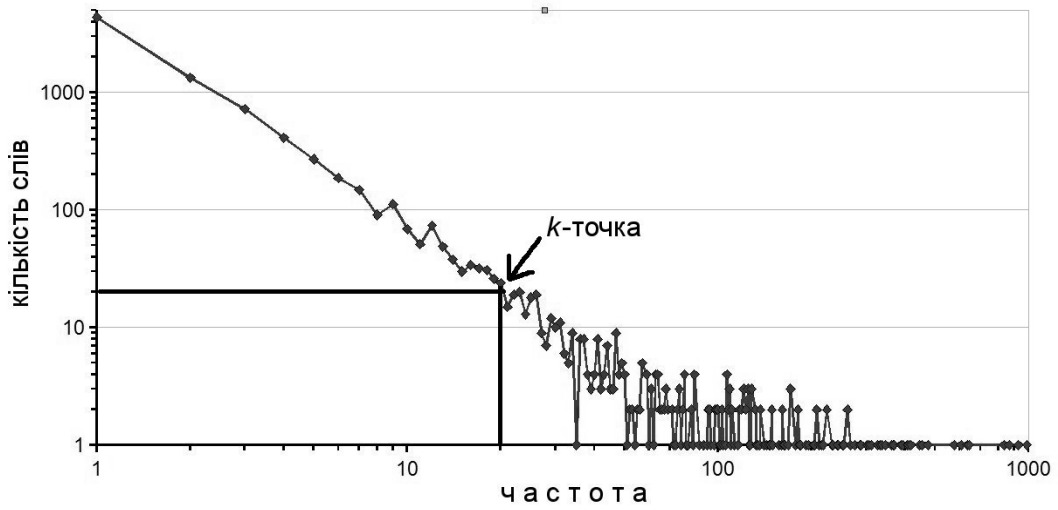

Рис. 2 — Частотний спектр роману Івана Франка “Борислав сміється”.

На думку польського науковця В. Маньчака, “Проблеми, які розглядають мовознавці, розпадаються насамперед на дві категорії: 1) ті, які можна розв'язати за допомогою статистики і 2) ті, яких за допомогою статистики розв'язати не можна. Мовознавство, яке розуміють як точну науку, займається тільки проблемами першої категорії. Іншими словами, йдеться про те, щоби так формулювати проблеми, щоби їх можна було розв'язати за допомогою статистики. Якщо це не можливо, ними не варто займатися, як не варто займатися жодними дослідженнями, про які наперед відомо, що вони не можуть привести до висновків, які можна перевірити” [9, с. 8]. Характеристика роману у світлі статистичної лінгвістики, а саме визначення ССТ твору, безперечно, належить до проблем першого типу.

Висновки і перспектива. Таким чином, в результаті аналізу ЧС роману “Борислав...” отримано важливі кількісні характеристики роману, які становлять статистичну структуру його тексту. Квантитативні параметри твору корелюють з аналогічними величинами інших романів І. Франка і дають змогу визначити його місце серед них. ССТ роману у перспективі також може увиразнити лексичні особливості твору, що становить окремий науковий інтерес і стане темою подальшого дослідження твору, оскільки кількісна та якісна сторони мови та мовлення корелюють і взаємопов'язані.

Список литературы

1. Бук С. Квантитативна параметризація текстів Івана Франка: спроба проекту // Іван Франко: Студії та матеріали. — Львів, 2010 (у друці); див. препринт arXiv:1005.5466v1 [cs.CL]. — Електронний ресурс <http://arxiv.org/abs/1005.5466>.
2. Бук С. Роман Івана Франка "Для домашнього огнища" крізь призму частотного словника / С. Н. Бук // Препринт arXiv:1006.0153v1 [cs.CL]. — Електронний ресурс <http://arxiv.org/abs/1006.0153>.
3. Бук С. Статистичні характеристики роману Івана Франка "Основи суспільності" (на основі частотного словника твору) // Вісник: Проблеми української термінології. — Львів: Національний університет "Львівська політехніка". — 2010 (у друці).
4. Бук С., Ровенчак А. Частотний словник роману Івана Франка "Перехресні стежки" // Стежками Франкового тексту. — Львів: Видавничий центр ЛНУ імені Івана Франка, 2007. — С. 138-369.
5. Лексика поетичних творів Івана Франка: Методичні вказівки з розвитку лексики / уклад. І. І. Ковалик, І. Й. Ощипко, Л. М. Полюга. — Львів: ЛДУ, 1990. — 264 с.
6. Перебийніс В. С., Муравицька М. П., Дарчук Н. П. Частотні словники та їх використання.— К.: Наук. думка, 1985. — 204 с.
7. Статистичні параметри стилів / за ред. В. С. Перебийніс.— К.: Наук. думка, 1967.— 260 с.
8. Франко І. Борислав сміється // Зібрання творів у 50-ти томах. — Т. 15: Повісті та оповідання. — К.: Наук. думка, 1979. — С. 256–480.
9. Mańczak W. Problemy językoznawstwa ogólnego. — Wrocław; Warszawa; Kraków: Ossolineum, 1996. — 257 s.
10. Popescu I.-I. et al. Word frequency studies. — Berlin; New York: Mouton de Gruyter, 2009. — xii, 278 p.

Бук С. Н. Статистическая структура романа Ивана Франко "Борислав смеется" / С. Н. Бук // Ученые записки Таврического национального университета им. В. И. Вернадского. Серия «Филология. Социальные коммуникации». – 2010. – Т. 23 (62), № 3. – С. 114-118.

В статье выясняется понятие статистической структуры текста, определяемое на лексическом уровне, как правило, по данным частотного словаря (ЧС). На основании ЧС романа И. Франко "Борислав смеется" получены количественные характеристики текста, которые сравниваются с соответствующими данными других романов писателя.

Ключевые слова: статистическая структура текста, частотный словарь, количественные характеристики текста, богатство лексики.

Buk S. N. Statistical structure of Boryslav Laughs, a novel by Ivan Franko / S. N. Buk // Scientific Notes of Taurida V. I. Vernadsky National University. – Series: Philology. Social communications. – 2010. – Vol. 23 (62), No 3. – P. 114-118.

In the article, the notion of the statistical structure of text is explored. On the lexical level it is defined as a rule from the frequency dictionary data. On the base of *Boryslav Smijet'sja [Boryslav Laughs]*, a novel by Ivan Franko, the quantitative parameters of text are obtained. They are compared with the respective data from other novels of the writer.

Key words: statistical structure of text, frequency dictionary, quantitative parameters of text, vocabulary richness.

Поступила в редакцию 01.09.2010 г

.