Література

1. Алексеенко Л. А., Дарчук Н. П., Зубань О. Н., Сорокин В. В. Параметризированная база данных поэтической речи как источник и инструмент филологических студий // Материалы международной конференции "Компьюерная лингвистика без границ", Санкт-Петербург, 25–26 марта 2004 г. – СПб., 2004.
2. Баранов А. Н. Автоматизация лингвистических исследований: корпус текстов как лингвистическая проблема // Руссистика сегодня. – 1998. – №1–2. – С. 179–191.
3. Дарчук Н. П. Структурно-статистическая база данных современного украинского языка на основе частотных словарей // Слово и словарь = Vocabulum et vocabularium: Сб. науч. тр. по лексикографии. – Гродно: ГрГУ, 2005. – С. 194–197.
4. Ляшевская О. Н., Плунгян В. А., Сичинава Д. В. Национальный корпус русского языка как инструмент лексикографа // Слово и словарь = Vocabulum et vocabularium: Сб. науч. тр. по лексикографии. – Гродно: ГрГУ, 2005. – С. 197–202.
5. Попова З. Д., Стернин И. А. Язык и национальная картина мира. – Воронеж, 2002. – 59 с.
6. Сирук Е. Б., Сорокин В. М. Использование Компьютерного тезауруса в статистическом анализе лексики художественного текста (на материале украинского языка) // Материалы международной конференции "Тексты и контексты: движение языка", Каунас, 6–7 мая 2004 г. – Каунас, 2004.
7. Словник української мови: В 11 т. – Київ, 1970–1980.
8. Соколовська Ж. П. Картина світу та ієрархія сем // Мовознавство. – Київ, 2002. – №6. – 87–91.

С. Бук, к. філол. н.[*]
Львівський національний університет імені Івана Франка (Львів)
УДК 161.2.81’373.374.322

СТАТИСТИЧНІ ХАРАКТЕРИСТИКИ ЛЕКСИКИ ОСНОВНИХ ФУНКЦІОНАЛЬНИХ СТИЛІВ УКРАЇНСЬКОЇ МОВИ: СПРОБА ПОРІВНЯННЯ

Résumé: In the article, for the first time, the main statistical characteristics, such as the variety, exclusiveness, concentration indices are compared for the vocabulary of six modern Ukrainian language functional styles: belles-lettres, poetry, journalistic, colloquial, scientific and official. The research has been done at the material of the 6 appropriate frequency dictionaries and the number of words and wordforms is given as well. The result is the complete correlation of all mentioned statistical characteristics.

Key words: frequency dictionary, (absolute and relative) word frequency, word occurrence, sampling, statistical description of vocabulary.

У лінгвістиці зламу XX–XXI століть особливо актуальними стали міждисциплінарні дослідження мови. Таким напрямком є й статистична лінгвістика. Як зауважив німецький лінгвіст Г. Альтманн, “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” [1: 5].

Мова як складна система дискретних одиниць має окрім якісних (які є об'єктом вивчення фонетики, дериватології, лексикології, синтаксису, теорії тексту, комунікативної лінгвістики тощо), також й кількісні характеристики. Останні властиві усім рівням мовної системи, та особливо виразні для лексичного.

Перші статистичні дослідження української мови було здійснено Інституті мовознавства ім. О. Потебні [11; 10]. Пізніше з метою виявлення лексичних особливостей різних стилів укладено їх частотні словники (далі — ЧС). Так, зараз в українській лексикографії існують ЧС для п’яти функціональних стилів, які більшість підручників зі стилістики визначають основними: для художньої прози [12], публіцистики [7], розмовно-побутового [4], наукового [3], офіційно-ділового [5]; а також для поетичного мовлення [6]. Аналіз залежності ранг-частота для текстів, що становлять джерельну базу цих словників, здійснено в [16]. Цікаво порівняти й базові статистичні характеристики їх словникового складу: обсяг словника словоформ ($V_ф$), обсяг словника лексем (V), багатство словника (Б), індекс винятковості (Е), середня повторюваність слова у тексті (А) тощо.

Існують різні підходи до способів здійснення цієї процедури. Так, наприклад, В. Перебийніс з цією метою запропонувала поняття “нульовий стиль” [13], М. Арапов зі

співавторами висунув теорію визначення "кількісної відстані" між словниками [2] тощо. Проте незаперечним є факт, що питання, пов'язані із методами порівняння ЧС, тісно переплетені з принципами їхнього укладання, "відмінності, інколи вельми відчутні, у методиці побудови словників сильно ускладнюють їх коректне порівняння" [15: 300], іншими словами, для отримання достовірних результатів важливо, щоби порівнювані словники були створені на одинакових засадах. Здійснимо короткий опис згаданих ЧС.

1. ЧС сучасної української художньої прози [12] укладено вручну на вибірці текстів 25 письменників загальним обсягом 500 000 слововживань.
2. В основу ЧС публіцистики [7] лягли тексти всеукраїнських газет за 1994 р., розрахованих на читачів різних регіонів, різних вікових та соціальних груп. Загальний обсяг вибірки — 300 000 слововживань.
3. ЧС розмовно-побутового стилю [4] укладено на вибірці 300 000 слововживань. Оскільки в українській мовознавчій практиці Національного корпусу текстів усного мовлення ще не існує, то матеріал дослідження склали 45 текстових уривків з прозових драм (хронологічні межі 1982–2002 рр.) приблизно по 6 000 слововживань кожен. Аналогічні джерела для відповідних словників використовували, наприклад, автори публікацій [17], [9].
4. ЧС наукового стилю [3] побудовано на 104 уривках приблизно по 3 000 слововживань кожен. Загальний обсяг — понад 300 000 тис. Тут представлено майже у рівних пропорціях такі галузі наук: біологія та хімія, психологія і педагогіка, фізика та математика, техніка, географія та геологія, історія та мовознавство. Зауважимо, що в ЧС наукового стилю [8], віднедавна доступному в Інтернеті, опрацьовано тільки гуманітарні науки: філософію, літературознавство та мовознавство.
5. Джерелами ЧС офіційно-ділового стилю [5] стали тексти підручників ділової мови та офіційних документів загальним обсягом 300 000 слововживань: Конституція України, кодекси, українські та міжнародні закони, міжнародні договори, конвенції, меморандуми, економічні й адміністративні документи та ін.
6. На текстах 1975–1995 рр. п'ятнадцяти поетів загальним обсягом 300 000 слововживань побудовано й ЧС поетичної мови [6].

Усі названі вище ЧС було укладено за майже ідентичними принципами (в тому числі застосовано практично однакові схеми лематизації слів) з використанням комп'ютера (окрім першого, який укладався вручну). Також видно, що величина вибірки усіх словників — 300 тис. слововживань, окрім ЧС художньої прози, який створено на текстах довжиною 500 тис. слововживань. Тому, для правомірності порівнянь словників, частоти усіх слів в ньому було відповідно прошкальовано за формулою:

$$n_2 = n_1 / 500 \times 300, \qquad (1)$$

де n_2 — частотність одиниці у вибірці 300 тис. слововживань, n_1 — частотність одиниці у вибірці 500 тис. слововживань.

Обчислення параметрів для словника художньої прози. При укладанні пробного зошита ЧС художньої прози було помічено, що "... подовження тексту неоднаково впливає на збільшення кількості слів і словоформ: кількість словоформ зростає дещо швидше, ніж кількість слів" [12: 14], тому показник багатства словника, обчисленого для текстів різної довжини, зіставляти не можна — простої формули для шкалювання частот (1) недостатньо. Проте для ЧС художньої прози відома кількість слів W та словоформ F для двох точок: 100 тис. і 500 тис. [12: 14], тому можна обчислити таку кількість для деякої третьої точки — 300 тис. слововживань. Згідно з Ю. Тулдавою [14: 62], частотні характеристики підлягають так званій алометричній залежності на зразок

$$y = Ax^b,$$

де A, b — константи (їх треба рахувати у нашому випадку), а x, y — певні величини.

Якщо W — кількість слів (для 100 тис. — 13 258, для 500 тис.— 33 391), F — кількість словоформ (для 100 тис. — 26 275, для 500 тис.— 86 284, а N — обсяг тексту (відповідно 100 і 500 тис.), то відношення F / W для деякого третього значення (N = 300 000) на підставі даних для двох інших (при N = 100 000 і N = 500 000):

$$\frac{F}{W} = AN^b \qquad (2)$$

[14: 64]. Виконавши обчислення, матимемо $A \approx 0.297$, $b \approx 0.165$, звідки для $N = 300\,000$ за цією ж формулою маємо відношення кількості словоформ до кількості слів 2, 38.

Враховуючи, що величини F і W також пов'язані подібною залежністю [14: 63]:

$$W = CF^d \qquad (3)$$

з параметрами $C \approx 4.89$, $d \approx 0.777$ (які також розраховано на ПК на підставі даних для масивів 100 тис. і 500 тис. слововживань), можна обчислити F і W для $N = 300\,000$. Справді, підставляючи (3) в (2), матимемо:

$$\frac{F}{C\,F^d} = AN^b; \quad F^{1-d} = AC\,N^b; \quad F = \left(AC\,N^b\right)^{1/(1-d)}.$$

Вийде значення кількості слів 24 906, словоформ — 59 161 (дані заокруглено до цілих).

Кількість hapax legomena (слів, що зустрілися в досліджуваному масиві текстів лише один раз), яку позначимо H, пов'язана з обсягом тексту подібним законом:

$$H = KN^q,$$

де K і q — деякі константи (див., наприклад [18: 81]). Для частотного словника художньої прози значення $K \approx 26,\ 7$, $q \approx 0,\ 48$, що дає при $N = 300\,000$ кількість hapax legomena 11 342. Далі отримані вказаним способом значення параметрів словника художньої прози позначено зірочкою (*).

Порівняння основних статистичних характеристик лексики функціональних стилів української мови найзручніше показати в таблиці:

Таблиця 1. Зведена порівняльна таблиця основних частотних характеристик п'яти ЧС української мови

	V	$\boldsymbol{V_ф}$	$\boldsymbol{V / N}$	$\boldsymbol{V_1 / N}$	$\boldsymbol{V_1 / V}$	$\boldsymbol{V_{10т} / N}$	$\boldsymbol{V_{10} / V}$
ЧС поетичної мови	31194	69012	0, 103	0, 052	0, 495	0, 789	0, 098
ЧС худ. прози	24906* (33391)	59161* (86284)	0, 083* (0, 067)	0, 038* (0, 029)	0, 455* (0, 430)	— (0, 821)	— (0, 149)
ЧС публіцистики	20824	43363	0, 070	0, 031	0, 450	0, 789	0, 161
ЧС розм.-поб. стилю	20541	42106	0, 073	0, 034	0, 465	0, 804	0, 121
ЧС наукового стилю	19367	42802	0, 059	0, 025	0, 427	0, 890	0, 189
ЧС офіційно-ділового стилю	9045	24263	0, 030	0, 0085	0, 280	0, 935	0, 303

Тут $\boldsymbol{V}$ — кількість різних слів (обсяг словника лексем), тобто слів у словнику, $\boldsymbol{V_ф}$ — кількість словоформ (обсяг словника словоформ), $\boldsymbol{V / N}$ — індекс різноманітності, $\boldsymbol{V_1 / N}$ — індекс винятковості тексту, $\boldsymbol{V_1 / V}$ — індекс винятковості словника, $\boldsymbol{V_{10т} / N}$ — індекс концентрації тексту, $\boldsymbol{V_{10} / V}$ — індекс концентрації словника.

Обсяг словника словоформ ($V_ф$) та **обсяг словника лексем** (V) — це величини, що показують, скільки окремих лексем (лематизованих, тобто зведених до початкової форми, слів) чи словоформ вжито у відповідних проаналізованих корпусах текстів конкретних функціональних стилів. Як видно із порівняльної таблиці, найбільше як слів, так і словоформ є в поетичному мовленні, далі за зменшенням ідуть художній, публіцистичний, розмовно-побутовий та науковий стилі. Найменше слів і словоформ (у два рази (!) менше, ніж в усіх інших) зафіксовано в офіційно-діловому. Це експериментальне підтвердження апріорно зрозумілого факту, що при написанні офіційно-ділового тексту користуються усталеними зворотами, обмеженою лексикою, уживають менше лексичних засобів, ніж в інших функціональних стилях.

Багатство словника (*Б*), тобто індекс різноманітності — це відношення обсягу словника лексем (*V*) до обсягу тексту (*N*), що обчислюється за формулою:

$$Б = V / N.$$

Підставивши відповідні дані, ми отримали, що в поезії ця величина становить 0, 103, в художній прозі — 0, 083* (0, 067), в публіцистиці — 0, 070, в розмовно-побутовому стилі — 0, 069, в науковому — 0, 065, в офіційно-діловому — 0, 030.

Індекс різноманітності абсолютно точно корелює із даними обсягу словника: найвищий він в поетичному мовленні, найнижчий — в офіційно-діловому, посередині за спадом — художній, публіцистичний, розмовно-побутовий та науковий.

Середня повторюваність слова у тексті (*А*), тобто відношення обсягу тексту (*N*) до обсягу словника лексем (*V*) — величина, обернена до індексу різноманітності, — обчислюється за формулою:

$$A = N / V.$$

Цей показник для поезії становить 9, 6, для художнього стилю — 12, 1* (14, 9), для публіцистики — 14, 3, для розмовно-побутового — 14, 5, для наукового — 15, 5, для офіційно-ділового — 33, 17. Іншими словами, кожне слово в середньому зустрілося в конкретному досліджуваному корпусі текстів 10–16 разів (виняток — офіційно-діловий стиль).

Індекс винятковості *для тексту* (у нашому випадку для тексту 300 000 слововживань) ($E_т$), тобто відношення кількості слів із частотою 1 — hapax legomena — (V_1) до обсягу тексту (*N*), обчислюється за формулою:

$$E_т = V_1 / N,$$

Індекс винятковості *для словника* (тобто для загальної кількості окремих слів, зведених до початкової форми *V*) ($E_с$) — за формулою:

$$E_с = V_1 / V,$$

Результати обчислень показують, що найбільший індекс винятковості і для словника, і для тексту, як і слід було очікувати, — в поезії, найменший — в офіційно-діловому. Далі за зменшенням цього показника ідуть художній, розмовно-побутовий, публіцистичний та науковий стилі. Цей параметр є показником варіативності лексики.

Індекс концентрації — величина, протилежна до індексу винятковості, що вказує, яку частку *тексту* (*N*) або *словника* (*V*) займає високочастотна лексика (з абсолютною частотою 10 і більше). Позначається відповідно $V_{10т}$ і V_{10}. Обчислюється за формулами $V_{10т} / N$ — індекс концентрації тексту та V_{10} / V — індекс концентрації словника. Показники цієї величини є найнижчими для поезії і, логічно, — найвищими для офіційно-ділового стилю.

Отже, порівняння статистичних даних лексики п'яти основних функціональних стилів і поезії вперше в українській лінгвістиці дає конкретні кількісні дані до тези, що найбагатша лексика — в художній поезії та прозі, далі — в публіцистиці, розмовно-побутовому, науковому та офіційно-діловому стилях[1]. Про це свідчить кореляція відповідних величин: кількості слів в ЧС, індекси різноманітності та винятковості, середня повторюваність слова. Суттєво, що наведені факти виявлено на однаковій за розміром вибірці текстів згідно з вимогами до порівнянь ЧС. Це також може слугувати підтвердженням, що вже така невелика з погляду сучасної корпусної лінгвістики вибірка — 300 000 слововживань — репрезентує основні відмінності функціонування слів у різних стилях української мови.

Література

1. Альтман Ґ. Мода та істина в лінгвістиці // Проблеми квантитативної лінгвістики.— Чернівці: Рута, 2005.— С. 3–11.
2. Арапов М. В., Тер-Гаспрян Л. И., Херц М. М. Сравнение частотных словарей // Научно-техническая информация. Серия 2. Информационные процессы и системы.— 1978.— №4.— С. 20–29.
3. Бук С. З 000 найчастотніших слів наукового стилю української мови.— Львів: Львівський національний університет імені Івана Франка, 2006.—192 с.
4. Бук С. З 000 найчастотніших слів розмовно-побутового стилю української мови.— Львів: Львівський національний університет імені Івана Франка, 2006.— 180 с.

[1] У попередньо виданій монографії [11], аналізуючи статистичні параметри лексичного рівня, автори зосередили свою увагу на характеристиці особливостей вживання дієслівних форм у різних стилях.

5. Бук С. Частотний словник офіційно-ділового стилю сучасної української мови // Лінгвістичні студії: Зб. наук. праць.— 2006.— Випуск 14.— С. 184–188.
6. Дарчук Н. П. (ред.) Частотний словник сучасної поетичної української мови // Лінгвістичний портал MOVA.info.— 2003–2006.— [Цит. 11 березня 2006].— Доступно з <http://www.mova.info/Page2.aspx?l1=89>; Дарчук Н. П. (ред.) Частотний словник сучасної поетичної української мови.— [Цит. 01 грудня 2001].— Доступно з <http://www.philolog.univ.kiev.ua/WINS/chast/chast.htm>.
7. Дарчук Н. П. (ред.) Частотний словник сучасної української публіцистики // Лінгвістичний портал MOVA.info.— 2003–2006.— [Цит. 11 березня 2006].— Доступно з <http://www.mova.info/Page2.aspx?l1=91>.
8. Критська В. І., Колєснов І. В., Недозим Т. І. та ін. Частотний словник наукового стилю // Лінгвістичний портал MOVA.info.— 2003–2006.— [Цит. 11 березня 2006].— Доступний з: <http://www.mova.info/article.aspx?11=176&DID=1700>.
9. Морковкин В. В. Сравнительный список наиболее употребительных русских слов (на материале шести словарей) // Лексические минимумы русского языка / Под ред. П. Н. Денисова.— М.: МГУ, 1972.— С. 16–74.
10. Муравицька М. П., Олексієнко Л. А. (відп ред.) Структура мови і статистика мовлення.— К.: Наукова думка, 1974.— 176 с.
11. Перебийніс В. С. (ред.) Статистичні параметри стилів.— К.: Наук. думка, 1967.— 260 с.
12. Перебийніс В. С. (ред.) Частотний словник сучасної української художньої прози.— К.: Наук. думка, 1981.— Т. 1.— 863 с.; Т. 2.— 855 с.
13. Перебейнос В. С. Методы и уровни моделирования нулевого стиля // Вопросы статистической стилистики.— К.: Наука, 1974.—331 с.— С. 16–35.
14. Тулдава Ю. П. Проблемы и методы квантитативно-системного исследования лексики.— Таллин: Валгус, 1987.— 204 с.
15. Якубайтис Т. А. О статистических пластах лексики // Вопросы статистической стилистики.— К.: Наукова думка, 1974.— С. 299–314.
16. Buk, S. N., Rovenchak, A. A. Rank–Frequency Analysis for Functional Style Corpora of Ukrainian // Journal of Quantitative Linguistics.— 2004.— V. 11, No. 3.— P. 161–171.
17. Juilland A., Brodin D., Davidovitch C. Frequency Dictionary of French Words.— The Hague; Paris, 1970.
18. Kornai A. How many words are there? // Glottometrics.— 2002.— V. 4.— P. 61–86.

Н. Андрейчук, І. Волошиновська*
Національний університет „Львівська політехніка” (Львів)
УДК 81’373.322.324

ДІЄСЛОВА, ЩО ВИЗНАЧАЮТЬ ГЛИБИНУ ОПРАЦЮВАННЯ ПРЕДМЕТА ДОСЛІДЖЕННЯ, ТА ЇХ СТАТИСТИЧНІ ХАРАКТЕРИСТИКИ У НАУКОВО-ТЕХНІЧНИХ ТЕКСТАХ

The list of key verbs denoting the stage of scientific research is suggested and their frequency analysis when used in scientific texts is viewed as means of determining of the depth of the subject of research cognition.

Вступ

У процесі пізнання людиною довкілля одним з основних моментів є збереження набутої інформації та її передача наступним поколінням. Для забезпечення поступу у певній науковій галузі корисним, а в більшості випадків навіть необхідним, є аналіз робіт попередників у напрямку запланованих та споріднених досліджень. Систематизація напрацьованого матеріалу, встановлення критеріїв та розробка методів пошуку і відбору базової інформації сприяють ефективній передачі знань, що є особливо важливим в умовах високого темпу розвитку сучасних технологій. Передові розробки передбачають діяльність науковців у міжждисциплінарних сферах, що, разом з неналежним опрацюванням базової інформації, доволі часто зумовлює невідповідність хронології робіт їх науковій новизні. Такі обставини ускладнюють пошук джерел з інформацією про новітні досягнення у заданому напрямку. З іншого боку, новизна не завжди виступає критерієм відбору інформації, необхідної для успішного вирішення завдань, тому потрібно прослідкувати також початкові та проміжні етапи розвитку досліджень. Таким чином, вартою уваги є проблема визначення стадії досліджень, описаних в роботі, що піддається аналізу.