СТАТИСТИЧНИЙ АНАЛІЗ ДЕЯКОЇ УЧБОВОЇ ЛІТЕРАТУРИ ХАРЧОВОЇ ПРОМИСЛОВОСТІ

Н.Л. Кузьмінська, Р.М. Семенишин
Національний університет харчових технологій

У статті описані результати статистичного дослідження деяких текстів галузі харчова промисловість. Дослідження проводилось по частотним словникам, складеним за допомогою програми TextAnalyzer, яка була створена спільно з Міжнародним науково-навчальним центром інформаційних технологій та систем НАН України та МОН України.

Вступ.

Генеральна Асамблея Організації Об'єднаних націй проголосила 2008 рік Міжнародним роком мов. В [1] наголошувалось, що позиція ЮНЕСКО полягає у тому, що в багатомовному світі не можуть домінувати всього декілька глобальних мов, а повинні використовуватися і менш розповсюджені мови. Технологічно питання забезпечення мовного різноманіття в кіберпросторі на даний час розв'язане. Багатомовний матеріал Інтернету являє собою безцінний навчальний ресурс, який потрібно уміло використовувати. Вивчаючи багатомовні ресурси, користувачі не тільки поглиблюють знання іноземних мов, але і краще розуміють особливості рідної мови.

Хотілось би відзначити, що досить багато саме різногалузевої учбової літератури потрапляє в кіберпростір, і вона подана не відредагованою. І, в такому випадку, говорити про поглиблення знання або розуміння мови недоцільно. На жаль, з кожним роком ця проблема стає все актуальнішою для української мови. Об'єм учбової літератури збільшується, а якість його подання знижується. Тому проблема аналізу текстової інформації є досить актуальною.

Також підкреслимо, що автори відносяться до тієї категорії людей, які вірять, що мова і мовлення підкоряються статистичним законам. Наявність у мові кількісних характеристик визнається в явному чи неявному виді всіма мовознавцями [2].

Статистичне дослідження великого обсягу матеріалу має своєю метою встановлення деяких характеристик мови і мовлення. Статистичні методи не тільки додають більшої ваги і авторитетності, доказовості мовознавчим висновкам, вони здатні розкрити такі закономірності будови мови та мовлення, які без них

розкрити неможливо, перевірити або і відкинути такі загальноприйняті твердження, існування яких можливе лише внаслідок недостатнього проникнення в глибину структури мови. У плані прикладному статистичні методи мають велике значення для відбору учбового матеріалу [2].

Автори статті вирішили долучитися до приведених вище проблем та провести статистичне дослідження деякої учбової літератури у галузі «харчова промисловість».

Статистичний аналіз текстів галузі «харчова промисловість».

Для аналізу були взяті підручники одного функціонального стилю (предмет «Товарознавство м'яса») на російській [3] та українській [4] мовах. Для складання частотних словників використали програму TextAnalyzer, яка була створена спільно з Міжнародним науково-навчальним центром інформаційних технологій та систем НАН України та МОН України та пройшла апробацію в вищезгаданому центрі. Більш докладно програма описана у [1].

Статистичні характеристики текстів, які вибрані для дослідження наведемо у таблиці 1.

Таблиця 1

Статистична характеристика	Текст [3]	Текст [4]
Всього слів	13045	50655
Всього унікальних слів	4065	10310
Коефіцієнт лексичного багатства	0,31	0,20
Всього абзаців	178	1337
Всього речень	640	3288
Всього літер у словнику	34806	89628
Всього літер у тексті	84549	333550
Середня кількість літер у слові (у тексті)	6,48	6,58
Середня кількість літер у слові (у словнику)	8,56	8,69
Середня кількість слів у реченні	20,38	15,41

На основі даних, отриманих програмою TextAnalyzer, були створені дискретні статистичні розподіли вибірок: (x_i,n_{xi}), (x_i,m_{xi}) i=1,..,21 (x_i – довжина слова, n_{xi} (m_{xi}) – його абсолютна частота у частотному словнику (тексті)) для тексту [3] на російській мові (таблиця 2) та (y_j,n_{yj}), (y_j,m_{yj}) j=1,..,23 (y_i – довжина слова, n_{yj} (m_{yj}) – його абсолютна частота у частотному словнику (тексті)) для тексту

[4] на українській мові (таблиця 3). Також у таблицях 2,3 наведені значення відносних частот w_{nxi} (w_{mxi}), w_{nyj} (w_{myj}).

Таблиця 2

x_i	n_{xi}	w_{nxi}	m_{xi}	w_{mxi}
1	21	0,005166052	1626	0,124702815
2	42	0,010332103	745	0,057136283
3	69	0,016974170	751	0,057596441
4	147	0,036162362	946	0,072551576
5	308	0,075768758	1235	0,094715852
6	479	0,117835178	1201	0,092108291
7	547	0,134563346	1418	0,108750671
8	525	0,129151292	1337	0,102538538
9	479	0,117835178	1008	0,077306542
10	428	0,105289053	903	0,069253777
11	334	0,082164822	679	0,052074546
12	263	0,064698647	482	0,036966025
13	188	0,046248462	338	0,025922233
14	96	0,023616236	168	0,012884424
15	62	0,015252153	96	0,007362528
16	31	0,007626076	44	0,003374492
17	21	0,005166052	32	0,002454176
18	12	0,002952030	13	0,000997009
19	6	0,001476015	7	0,000536851
20	6	0,001476015	9	0,000690237
21	1	0,000246002	1	0,000076693

Таблиця 3

y_i	n_{yi}	w_{nyj}	m_{xi}	w_{myj}
1	33	0,003200776	5350	0,105616425
2	103	0,009990301	3229	0,063744941
3	168	0,016294859	3249	0,064139769
4	373	0,036178468	2610	0,051525022
5	709	0,068768186	4586	0,090534005
6	1001	0,097090204	4902	0,096772283
7	1313	0,127352085	6328	0,124923502

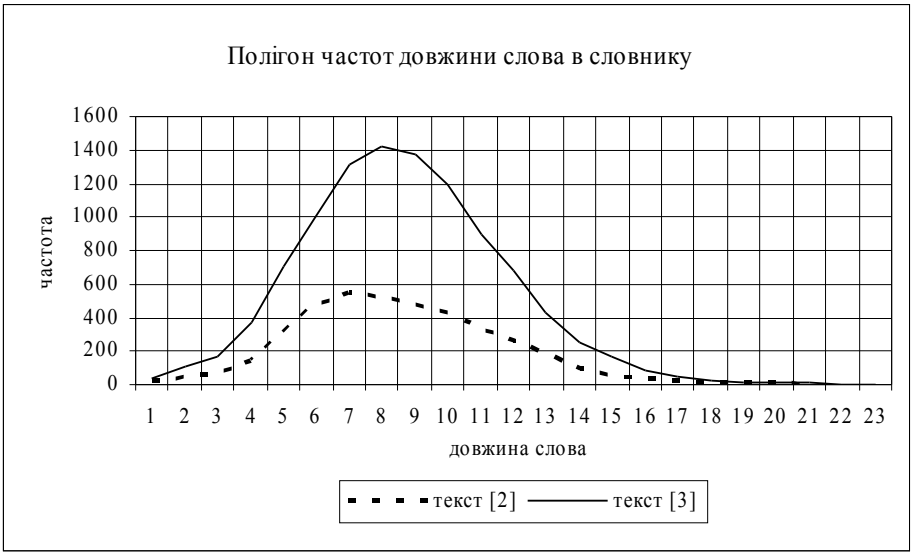
Продолжение табл. 3

8	1420	0,137730359	5244	0,103523838
9	1371	0,132977692	4725	0,093278057
10	1195	0,115906887	3475	0,068601323
11	898	0,087099903	2746	0,054209851
12	681	0,066052376	1973	0,038949758
13	433	0,041998060	907	0,017905439
14	255	0,024733269	593	0,011706643
15	173	0,016779825	376	0,007422762
16	89	0,008632396	223	0,004402329
17	44	0,004267701	74	0,001460863
18	18	0,001745878	25	0,000493535
19	11	0,001066925	17	0,000335604
20	13	0,001260912	13	0,000256638
21	6	0,000581959	6	0,000118448
22	2	0,000193986	2	0,000039483
23	1	0,000096993	2	0,000039483

Використовуючи данні про дискретні статистичні розподіли (таблиці 2, 3) побудуємо полігони частот для вибірок, отриманих по тексту та по частотному словнику (рис.1, 2):

Рисунок 1

Рисунок 2



Оскільки обсяг тексту [4] більший за обсяг тексту [3], то графіки мають різну висоту.

Приведемо значення числових характеристик вибірок, отриманих по частотним словникам (таблиця 4).

Таблиця 4

Числові характеристики	Текст [2]	Текст [3]
Вибіркова середня величина	$\bar{x} = 8,56236$	$\bar{y} = 8,69331$
Дисперсія вибірки	$D_x = 9,30146$	$D_y = 8,89818$
Середнє квадратичне відхилення вибірки	$\sigma_x = 3,04983$	$\sigma_y = 2,98298$
Коефіцієнт асиметрії	$A_x = 0,41167$	$A_y = 0,36888$
Ексцес	$E_x = 0,29236$	$E_y = 0,41503$

На основі отриманих числових характеристик можна сказати, що середні довжини слів, дисперсії та середні квадратичні відхилення у російському та українському текстах близькі за своїми значеннями. Обидва розподіли мають додатну асиметрію (невелику), тобто кількість варіант $x_j > \bar{x}$ ($y_j > \bar{y}$) переважає кількість варіант $x_i < \bar{x}$

($y_i < \bar{y}$). Оскільки $E_x = 0,29236 > 0$ та $E_y = 0,41503 > 0$, то вершини таких законів розподілів будуть гострими, тобто це будуть гостровершинні розподіли. На основі приведених даних (таблиця 4), можна зробити висновок, що отримані розподіли (вибірки по частотним словникам) близькі до нормального закону розподілу.

Висновки.

Найуживанішим словом, якщо не рахувати прийменники, є, як в українській, так і у російській мовах, слово «м'ясо» («мясо»).

Раніше отримані результати досліджень [1], проведених над великою кількістю різнотематичної літератури, показали, що середня довжина слова в українському тексті менша за середню довжину слова в російському. Але наше дослідження цього не виявило, можливо, це пов’язано з вживанням великої кількості складних слів, які в літературній мові частіше замінюються двома або трьома словами. Такими «важкими» для сприйняття є слова, наприклад, вологоутворюючого, смакоаромоутворюючого і т.і. Це, звичайно, може бути специфіка саме цієї галузі, тому робити якісь висновки зарано, потрібно провести ще ряд додаткових досліджень з підручниками інших авторів.

Список літератури.

1. Н.А.Власенко, Н.Л.Кузьминская, А.А.Максименко. ЮНЕСКО о многоязычии и один из путей его эффективного использования.

2. Перебийніс В.І. Статистичні методи для лінгвістів: Навчальний посібник/ Вінниця: «Нова книга», 2001 – 168 с.

3. Товароведение пищевых продуктов. Учебник для технол.фак.торг.вузов. Тылкин В.Б., Кононенко И.Е., Дмитриева А.Б. – 2-е издание, переработанное и дополненное – М.: Экономика, 1980. – 432 с.

4. І.В. Сирохман, Т.М. Раситюк. Товарознавство м’яса і м’ясних товарів. Підручник. – К.: Центр навчальної літератури, 2004. – 384 с.