О. Й. Максимів

СПЕЦИФІКА ЧАСТОТНИХ СЛОВНИКІВ ПУБЛІЦИСТИЧНОГО СТИЛЮ СУЧАСНИХ ПЕРСЬКОЇ ТА УКРАЇНСЬКОЇ МОВ

Дані частотних словників у лінгвістиці є предметом зіставлення головним чином для типологічних [див. напр. Алексеев 2001, *103–109* та ін.] та стилістичних потреб [див. напр. Микерина 1966, *19–20*; Перебийніс 1985, *74–76*, *160–183*; Перебийніс 2002, *68–97*; Бук 2006, *166–172* та ін.]. Зазвичай акцентується на подібності та близькості словників [див. напр. Тисенко 1977, *922–933*; Арапов 1978, *20–29*], зіставляються здебільшого лише словники, укладені за однаковими принципами, пропонуються складні статистичні підрахунки, існують також роботи такого типу і в концептуальній лінгвістиці [див. напр. Мухин 2008, *73–74*]. Проте давно помітили, що частотний словник поділяється на певні частотні зони [див. напр. Засорина 1966, *23*]. Нас цікавить зіставлення таких зон із можливістю виявити не подібні, а специфічні елементи, виокремити етноспецифіку мови та словник, специфічний для галузі знань, на текстах якої укладалися ці словники. Тому єдиною вимогою для зіставлення такого плану є галузева (чи тематична) близькість словників зіставлюваних мов.

Джерелом цієї статті є статистичні дані частотних словників публіцистичного стилю мовлення української [Дарчук 2003–2014] (далі – ЧСУ) та перської (далі – ЧСП) мов. Детальніше про принципи укладання ЧСП див. [Максимів 2009]. Обидва словники укладені за матеріалами лексики публіцистичних текстів (стилістично однорідні [пор. з Арапов 1978, *137*]) із центральних та обласних українських та іранських газет (дотримання просторової єдності) 2003–2004 (відповідно 1381–1382) років (достатньо вузький часовий зріз). Увага до публіцистичного стилю пояснюється тим, що "сфера його використання – політичне, суспільно-культурне життя людей, виробнича діяльність, а основні ознаки стилю – спрямованість на новизну, актуалізація сучасності, політична, суспільна, морально-етична оцінка життя" [Дарчук 2005, *107*].

Мета розвідки полягає у висвітленні статистичних особливостей частотних словників публіцистичного стилю перської та української мов.

Розглянемо процедуру, за якою укладалися ЧСУ та ЧСП. Вона передбачає послідовність певних етапів [див. Засорина 1966; Перебийніс 1985; Нелюбин 1991; Алексеев 2001]. У зіставлюваних словниках ці етапи набули такого вигляду:

1) за одиницю підрахунку в обох словниках обрали словоформу, оскільки розмітка корпусу перської мови [2011–2004 عاصی], на базі якого укладали ЧСП, наразі не є завершеною, і також доступним є частотний словник словоформ української мови [Частотний… 2003–2014];

2) укомплектували вибірковий корпус обсягом 3 044 009 слововживань із шести наявних у корпусі перської мови неспеціальних газетних текстів. Оскільки у частотних списках за матеріалами кожної газети наводили абсолютні частоти, то об'єднання списків звелося до додавання цих частот [див. Перебийніс 1985, *77*]. Вибірковий корпус ЧСУ налічував 300 000 слововживань;

3) автоматично підрахували абсолютну частоту однакових словоформ у новоствореному перському корпусі. Підрахунок здійснили за допомогою спеціально написаного макросу для програми Microsoft Excel (автор – *Б. Рудий*, науковий співробітник Лінгвістичного музею при КНУ імені Т. Шевченка). Корпус налічував 69 300 словоформ. Нелітерні символи (цифри, розділові знаки, слова, написані не перською графікою, тощо) зі списку вилучили, і в результаті ми одержали список обсягом 59 215 словоформ. ЧСУ налічує 60 673 словоформи. Ми оперуємо даними з інтернетресурсу, бо саме звідти взяли матеріал словника для подальшої роботи, у статті [Дарчук 2005, *108*] наведена інформація, що цей словник містить 52 257 різних словоформ.

Зважаючи на незавершеність розмітки перського корпусу та на відсутність відповідного комп'ютерного програмного забезпечення, лематизацію і морфологічну кваліфікацію словоформ списку робили вручну. Оскільки "обсяг словника з урахуванням омографів збільшується несуттєво" [Максимів 2009, *57*], основну масу загальновживаних слів розпізнали надійно, а решту слів, які складають меншу частку словника будь-якого тексту, ми контролювали шляхом перегляду в конкордансах і прийняття рішень у кожному випадку окремо [див. також про це Засорина 1966; Тисенко 1977]. Лексико-граматичну омонімію, як правило, не враховували (напр., не розмежовували вживання слова خوب *khub*

‘добрий, добре’, زیبا *zibā ‘гарний, гарне’,* آزاد *āzād ‘вільний, вільно’* як прикметника і як прислівника), адже у словниках різні частини мови того самого слова реєструють у межах однієї словникової статті. Під час укладання ЧСУ омонімію знімали.

Основні вимоги до того, щоб порівняння частотних словників було коректне, такі [Бук 2008, *34*]:

1) словники повинні бути укладені на матеріалі однакового обсягу. У нашому випадку цієї вимоги не було дотримано, проте нас цікавлять не абсолютні числові результати порівняння словників, а відносні, які б дали можливість апроксимувати числові дані словників і виокремити частотні зони;

2) словники повинні бути укладені на однакових принципах. Що стосується одиниці підрахунку, то більшість наявних частотних словників мають вхідними одиницями слова або словоформи [див. Алексеев 2001, *74*]. Ми спробуємо зіставити дані словників словоформ перської та української мов. Хоча під час укладання ЧСУ омонімію усували, а ЧСП уклали на базі корпусу перської мови без її урахування, це суттєво не вплине на результати порівняння даних словників словоформ. Відомо, що омоніми (чи омографи) становлять незначну частину наповнення будь-якого тексту [див. Использование… 1990, *171*].

Визначимо основні статистичні характеристики ЧСУ та ЧСП і спробуємо зіставити їх між собою. Прошкалювати частоти слів будь-якого з даних словників, як це зробили для визначення статистичних характеристик лексики основних функціональних стилів української мови [див. Бук 2006], ми не можемо, бо маємо справу зі словниками різних мов.

Найзагальнішими статистичними характеристиками частотного словника [див. Тулдава 1987, *57;* Бук 2008, *34–37*] є:

N – обсяг тексту, на основі якого укладений словник (тобто кількість слововживань у тексті). $N_{перс.}$ = 3 044 009, $N_{укр.}$ = 300 000;

V – обсяг словника словоформ. $V_{перс.}$ = 59 215, $V_{укр.}$ = 60 673.

Відносні показники аналізованих словників мають такий вигляд:

V/ N – відношення обсягу словника до обсягу тексту, що показує відносне “багатство”, чи “різноманітність” словника. V/ $N_{перс.}$ = 0,0195, V/ $N_{укр.}$ = 0,2022;

N/ V – середня частота (повторюваність) слова у даному тексті. N/ $V_{перс.}$ = 51,41, N/ $V_{укр.}$ = 4,94.

Наведені відносні статистичні показники залежать від обсягу і від типу тексту [див. Тулдава 1987, *57*]: при збільшенні обсягу тексту відносна різноманітність лексики зменшується, а середня повторюваність слів збільшується. Для обох мов тип тексту однаковий: публіцистика. Врахувавши, що вибірка для перського словника у десять разів більша від вибірки для українського, спостерігаємо дуже незначну відмінність багатства та середньої частоти слова обох словників: 5 % і 8,69 % відповідно на користь ЧСУ. Це можна пояснити типологічною відмінністю перської (аналітичної) та української (синтетичної) мов, хоча підтверджується те, що "сумнівною є залежність від системних характеристик мови показника різноманітності, або багатства лексики досліджуваних творів" [Перебийніс 1981, *14*].

Порівняння основних статистичних характеристик лексики частотних словників публіцистики перської та української мов найзручніше показати в Таблиці 1.

Таблиця 1. Основні статистичні характеристики лексики ЧСП і ЧСУ

	V	V/N	N/ V	V_1/N	V_{10}/N
ЧСП	59 215	0,0195	51,41	0,0068	0,966
ЧСУ	60 673	0,2022	4,94	0,12	0,811

Тут V – обсяг словника словоформ, V/N – багатство словника, N/ V – середня повторюваність слова у тексті, V_1/N – індекс винятковості тексту, V_{10}/N – індекс концентрації словника.

Дані словника лексем (L) частотних словників публіцистики перської та української мов можуть бути предметом окремої статті.

З першого погляду, неможливо порівнювати між собою дані, одержані за результатами підрахунків текстів різного обсягу. Так, для укладання ЧСУ використали вибірку на 300 000 слововживань, ЧСП укладали на матеріалі корпусу обсягом 3 044 009 слововживань, тобто у десять разів більшого, ніж ЧСУ. Більше того, вибірковий корпус, який складається з великої кількості коротких текстів (саме таким способом укладався ЧСУ), завжди дає більший обсяг словника, ніж укладений з невеликої кількості довгих текстів (такою була процедура укладання ЧСП) [див. Алексеев 2001, *82–83*]. Це твердження дещо нівелюється особливостями

композиції газет, де практично не зустрічаються тексти одного автора на одну тему, які б займали більше, ніж одну сторінку.

Як виявилося, словники словоформ обидвох мов мають приблизно однаковий обсяг – близько 60 тисяч словоформ (див. Таблицю 1)! Роблячи попередні висновки, припускаємо, що окрема субмова (у даному випадку публіцистичний стиль мовлення) оперує певною відносно обмеженою кількістю слів – загальних для всієї мови і ключових для даної субмови, тобто виходимо на спеціальний лексикон галузі (публіцистики). Заперечується теза про те, що обсяг вибірки суттєво впливає на склад частотного словника. Підтверджується те, що визначальним є якісний склад вибірки.

Зрозуміло, що індекс винятковості для тексту, який характеризує варіативність лексики, частку тексту, яку займають слова, що трапилися один раз, буде різним. Для ЧСУ він становить 0,12, для ЧСП 0,0068. Таку велику різницю цього показника пояснюємо значно меншим обсягом українського тексту, використаного для укладання частотного словника. Хоча, як бачимо з Таблиці 2, кількість hapax legomena ЧСУ в 1,4 разу перевищує їхню кількість у ЧСП.

Індекс концентрації тексту – частка тексту, яку займають високочастотні слова, що трапилися більше, ніж десять разів, – становить: для перського тексту 0,966, для українського 0,811. Впадає у вічі невелика різниця цього показника, що є ознакою стилістичної однорідності зіставлюваних словників.

Таблиця 2. Розподіл словоформ за частотою

Частота	Кількість словоформ	
	перського	українського
Більше 2000	170	12
1500–1999	75	1
1000–1499	188	7
500–999	463	24
400–499	223	8
300–399	331	21
200–299	552	39
100–199	1359	119
50–59	544	105
40–49	747	200
30–39	1130	323
20–29	1870	684
10–19	4318	2144
9	795	499
8	961	668
7	1132	805
6	1459	1124
5	1833	1598

90–99	268	43
80–89	304	44
70–79	357	70
60–69	385	81

4	2628	2443
3	4110	4190
2	7635	9111
1	25 377	36 310
Всього	59 214	60 673

Оскільки сенс укладання частотного словника полягає передусім у "стратифікації різних за статистичною вагою шарів лексики" [Засорина 1966, *70*], тобто у виявленні основних частотних зон слів, то окремо заслуговують на увагу дані розподілу словоформ цих словників за частотою. Виділяти частотні зони можливо за різними ознаками [див. Тулдава 1987, *65*], залежно від цілей і завдань конкретного дослідження. Найдоречнішим є таке розмежування частотних зон, при якому кожна зона охоплювала б діапазон, який дорівнює одному порядку [див. Малаховский 1980, *101*]. Розділивши таким чином дані Таблиці 2 на чотири зони за інтервалом рангів різницею на порядок, можна помітити цікаві закономірності, відображені у Таблиці 3.

Таблиця 3. Розподіл даних ЧСП та ЧСУ на зони

Частота	Кількість словоформ	
	перського	українського
Більше 1000	433 (0,73 %)	20 (0,03 %)
100–1000	2928 (4,94 %)	187 (0,31 %)
10–100	9923 (16,76 %)	3694 (6,09 %)
1–9	45930 (77,57 %)	56748 (93,53 %)

Бачимо, що український словник насиченіший у нижній своїй частині, тобто більша частина словоформ, які увійшли до нього, мають невисоку частотність (93,53 %). Нижня частина перського словника становить 77,57 %. Це пояснюємо аналітичним типом перської та синтетичним – української мови. Наприклад, у перській мові відсутні відмінки [див. також Бекбаева 1982], тоді як українське слово може вживатися по одному разу в різних відмінках, і всі ці слововживання увійдуть до низькочастотної зони словника. Тому предметом зіставлення на рівні лексичного значення не можуть бути частотні словники словоформ, щонайменше, треба відштовхуватися від словників лексем або навіть груп

лексем. Процедура зіставлення списків частотних словників стане предметом окремого дослідження.

Вже зараз можна припустити, що лексика, яка визначає багатство словника публіцистики кожної мови, кожного автора зокрема, належить саме до цієї зони. Таку лексику не можна вважати етноспецифічною. Сюди також належить, з іншого боку, застаріла чи неусталена лексика [див. також Тисенко 1977].

Третя зона Таблиці 2 (інтервал рангів 10–100) є відносно найоднорідніша для обох словників: слова, розташовані у ній, складають 16,76 % перського і 6,09 % українського списків. Можемо припустити, що саме сюди потрапить найбільша кількість слів, специфічних для публіцистичного стилю мовлення загалом.

До другої зони Таблиці 2 потраплять різноманітні за своєю специфічністю слова: як етноспецифічні для певної мови, так і специфічні для певної галузі (публіцистики у нашому випадку). Саме на її базі і доречно буде дослідити етноспецифіку перської та української мов. Наприклад, за результатами зіставлення перших двох сотень морфемних груп слів перського та українського частотних словників публіцистики можна виокремити специфічно перські слова, специфічно українські та специфічні для публіцистичного стилю мовлення. У дужках подано порядковий номер (ранг) морфемної групи, до якої входить це слово, у частотному словнику морфемних груп кожної із зіставлюваних мов.

До специфічно перських слів належать: اسلامی *eslāmi іслам-ський (101)*, آمریکا *āmrikā Америка (56),* نفت *naft нафта (90),* ایران *irān Іран (27),* جهان *jahān світ (49)*. До специфічно українських *молодий (143), влада (67), український (27), Росія (86), Янукович (98)*. Специфічними для публіцистичного стилю мовлення є такі слова, як رئیس *ra'is (93) президент (48),* دولت *dowlat (60) держава (55),* اطلاع *ettelā' (115) повідомлення (157),* سیاست *siyāsat (67) політика (72),* قانون *qānun (96) закон (74)*.

До першої зони, очевидно, у першу чергу, потраплять службові та малоінформативні загальні повнозначні слова.

Для зручності опрацювання і відповідно до прийнятого у науці за законом Бредфорда [про цей закон див. дет. Чурсин 1982, *91–100*; Хайтун 1983, *71–85*] щодо поділу даних саме на три зони, дані першої і другої зон варто об'єднати. Доцільність такого об'єднання підтверджують ще й ті факти, що кількісний склад

новоутвореної групи зміниться не набагато, і що за попередніми результатами, наведеними вище, вже навіть у цій зоні можна виокремити етноспецифічну лексику та лексику, специфічну для досліджуваної галузі.

Таблиця 4 ілюструє покриття публіцистичного тексту найчастотнішими словоформами перської та української мов. Те, що відсоток покриття українського тексту зі збільшенням рангу словоформ збільшується повільніше, ніж відсоток покриття перського тексту, є результатом різнотипності української та перської мов.

Таблиця 4. Покриття тексту найчастотнішими словоформами

Ранг словоформ	Покриття перського тексту, %	Покриття українського тексту, %
10	22,23	11,01
50	34,68	22,22
100	40,81	27,23
500	60,26	40,09
1000	70,36	47,01
1500	76,26	51,46

Отже, зіставлення статистичних даних лексики частотних словників публіцистики перської та української мов вперше дає конкретні кількісні дані про існування окремих частотних зон, у яких представлена лексика, етноспецифічна для кожної із зіставлюваних мов, а також специфічна для досліджуваного стилю мовлення (публіцистичного). Про це свідчить кореляція відповідних величин: кількості словоформ у частотних словниках, індексів концентрації словника. Важливо, що наведені факти виявлено на різних за розміром вибірках текстів, що дає змогу говорити про лексику, специфічну (ключову) в різних мовах для певних однакових типів текстів. Це також може служити підтвердженням того, що для частотного словника репрезентативність вибірки є визначальною, а її обсяг вирішального значення не має.

У літературі [див. напр. Чернышев 1966, *60–61*; Бектаев 1971, *47–112*; Малаховский 1980, *99–105*; Тулдава 1987, *65*; Перебийніс 2002, *3–12* тощо] відзначали можливість застосування страти-

фікації лексики за ознакою частотності для лексикологічних, лінгвостилістичних, лінгводидактичних, лексикографічних, психологічних досліджень, для вирішення проблем автоматичної переробки тексту. Вказували на те, що кількісні характеристики слів, які належать до певної частотної зони, багато в чому корелюють з якісними властивостями цих слів, наприклад, з нейтральністю (загальновживаністю), тематичністю ("ключові слова"), інформативністю тощо [див. напр. Тищенко 2007, *84–91*]. Також відзначають перспективність пов'язання даних східних мов із контрастивним підходом та етномовною відносністю, вивчення семантичної етноспецифіки окремих східних мов [див. Тищенко 2005–2011]. У подальших студіях ми застосуємо дані частотних словників публіцистики перської та української мов для етнолінгвістичних потреб, а саме, зіставивши порціями лексику аналізованих ЧСП та ЧСУ, виявимо етноспецифіку лексичної системи перської та української мов.

ЛІТЕРАТУРА

Алексеев П. М. **Частотные словари.** Санкт-Петербург, *2001*.

Андрющенко В. М. Новые работы в области статистической лексикографии // **Вопросы языкознания,** *1968*, № 5.

Арапов М. В., Тер-Гаспарян Л. И., Герц М. М. Сравнение частотных словарей // **Научно-техническая информация.** Серия 2. Информационные процессы и системы. *1978*. № 4.

Бекбаева К. А. Персидский язык // **Квантитативная типология языков Азии и Африки.** Ленинград, *1982*.

Бектаев К. Б., Лукьяненков К. Ф. О законах распределения единиц письменной речи // **Статистика речи и автоматический анализ текста.** Ленинград, *1971*.

Бук С. **Основи статистичної лінгвістики.** Львів, *2008*.

Бук С. Статистичні характеристики лексики основних функціональних стилів української мови: спроба порівняння // **Лексикографічний бюлетень.** *2006*. Випуск 13.

Дарчук Н. Параметризована база даних сучасної української мови на основі частотних словників // **Проблеми квантитативної лінгвістики:** зб. наук. праць. Чернівці, *2005*.

Дарчук Н. П. (ред.) Частотний словник українського публіцистичного стилю – 2004 // **Лінгвістичний портал MOVA.info.** –

2003–2014. – [Цит. 24 лютого 2009]. – Доступно з http://www.mova.info/Page2.aspx?11=178.

Засорина Л. Н. **Автоматизация и статистика в лексикографии.** Ленинград, *1966*.

Использование ЭВМ в лингвистических исследованиях / Т. А. Грязнухина, Н. П. Дарчук, Н. Ф. Клименко и др.; отв. ред. *В. И. Перебейнос*. Киев, *1990*.

Максимів О. Й. Принципи укладання частотного словника газетної лексики сучасної перської мови // **Вісник КНУ ім. Т. Шевченка.** Східні мови та літератури, *2009*, № 14.

Малаховский Л. В. Принципы частотной стратификации словарного состава языка // **Статистика речи и автоматический анализ текста.** Ленинград, *1980*.

Микерина Т. А. О сопоставлении частотных списков // **Межвузовская конференция по вопросам частотных словарей и автоматизации лингвостатистических работ.** Тезисы докладов и сообщений. Ленинград, *1966*.

Мухин М. Ю. Частотный спектр как вид представления концептуальной системы автора // **MegaLing'2008. Горизонти прикладної лінгвістики та лінгвістичних технологій** // Доповіді міжнародної конференції. 22–28 вересня *2008*, Україна, Крим, Партеніт. Сімферополь, *2008*.

Нелюбин Л. Л. **Компьютерная лингвистика и машинный перевод.** Москва, *1991*.

Перебийніс В. С. Вступ // **Частотний словник сучасної української художньої прози.** У 2-х. т. Т. 1. Київ, *1981*.

Перебийніс В. І. **Статистичні методи для лінгвістів.** Вінниця, *2002*.

Перебийніс В. С., Муравицька М. П., Дарчук Н. П. **Частотні словники та їх використання.** Київ, *1985*.

Тисенко Э. В. Статистические параметры словаря // **Частотный словарь русского языка** / Под ред. Л. Н. Засориной. Москва, *1977*.

Тищенко К. Східні мови і метатеорія мовознавства [Електронний ресурс] // **Віртуальна Русь,** *2005–2011*. – Режим доступу до статті: http://www.ruthenia.info/txt/tyschenkok/me.html.

Тищенко К. М. **Основи мовознавства.** Київ, *2007*.

Тулдава Ю. **Проблемы и методы квантитативно-системного исследования лексики.** Таллин, *1987*.

Фрэнсис У. Н. Проблема формирования и машинного представления большого корпуса текстов // **Новое в зарубежной лингвистике.** *1983*. № 14.

Хайтун С. Д. **Наукометрия. Состояние и перспективы.** Москва, *1983*.

Частотний словник наукового стилю [Електронний ресурс] / [ред. Дарчук Н. П.] // Лінгвістичний портал MOVA.info. – 2003–2014. – Режим доступу: http://www.mova.info/Page2.aspx?l1=176.

Чернышев В. А. Опыт подразделения лексики специального текста на основе частотных характеристик // **Межвузовская конференция по вопросам частотных словарей и автоматизации лингвостатистических работ.** Тезисы докладов и сообщений. Ленинград, *1966*.

Чурсин Н. Н. Популярная информатика. Київ, *1982*.

29. عاصی م. **پایگاه‌داده‌های زبان فارسی.** – تهران: پایگاه‌داده‌های زبان فارسی، 2004 – Режим доступу : مصطفی عاصی / [Електронний ресурс] .2011 http://pldb.ihcs.ac.ir