

допомогла уточнити за формальними показниками межу між центром і периферією ЛСП, що безпосередньо відображається на розміщенні ЛСВ дієслів у синопітичній схемі КТ (місце у межах одного чи декількох концептів; рівень ієрархізації; кількість заповнених зон у мікростатті), а також у вигляді зони додаткової семи у мікростатті КТ дієслів, де фіксуються ЛСВ з ядра інших лексико-семантичних полів, що містять у своєму складі диференційну сему мовлення. Як результат інтеграції КТ дієслів у КТ української мови спроектовано **загальну структуру комп'ютерного тезауруса**, що складається з **трьох** базових компонентів: синопітичної схеми, власне ідеографічної частини (мікро- і макростатті) та пошукової системи з функціями алфавітного та пермутаційного показників.

Розроблення основних теоретико-методичних засад створення тезауруса дає можливість зробити внесок у дослідження таких аспектів мовознавства, як парадигматика й синтагматика, теорія поля, синонімія, антонімія і полісемія у лексичній семантиці, показує органічний взаємозв'язок синтаксичних і семантичних характеристик дієслова. Аналіз і систематизація наявних, а також конструювання нових електронних словників тезаурусного типу сприяють становленню поняттєвого апарату, методології та структури такого якісно нового підрозділу української лексикографії, як комп'ютерна.

Отримані результати, відповідно, можуть знайти відображення в лекціях, спецкурсах і спецсеминарах з проблем комп'ютерної ідеографії. Вони важливі для розвитку прикладної лінгвістики в цілому як логічне продовження її теоретико-методологічних засад, заснованих на глибокому симбіозі класичних лінгвістичних теорій і новітніх комп'ютерних технологій.

#### Література

1. Дарчук Н., Денисенко І., Сірук О., Сорокін В. Теоретичні питання моделювання ідеографічного тезауруса української мови // Українське мовознавство. Міжвідомчий науковий збірник. Вип. 24. — К.: Видавничо-поліграфічний центр "Київський університет", 2002. — С.107–118.
2. Дубичинский В. В. Искусство создания словарей конспекты по лексикографии / Харьковский политехнический ун-т. — Х., 1994. — 102 с.
3. Середницька А. Я. Мовна категоризація світу і її відображення в ідеографічному словнику // Семантика, синтактика і прагматика мовленнєвої діяльності. Матеріали Всеукраїнської наукової конференції. — Львів: Літопис, 1999. — С. 87–90.
4. Синиченко О. До проблеми створення словника понять української мови // Мовознавство: Доп. та повідомл. IV Міжнародного конгресу українців / Відп. ред. В. Німчук. — К.: Пульсари, 2002. — С. 31–35.
5. Сірук О., Сорокін В. Использование компьютерного тезауруса в статистическом анализе лексики художественных текстов (на материале украинского языка) // Texts and Contexts: the Movement of Language. Selected Papers. Vilniaus universiteto Kauno humanitarinis fakultetas. — Kaunas VU Press, 2005. — P. 601–608.
6. Сірук О. Два підходи до побудови комп'ютерного тезауруса дієслів української мови // Українське мовознавство. Міжвідомчий науковий збірник. Вип. 31. — К.: Видавничо-поліграфічний центр "Київський університет", 2004. — С.84–87.
7. Сірук О. Систематизація світових комп'ютерних тезаурусів як підґрунтя для укладання тезауруса української мови // Мовні і концептуальні картини Світу: Збірник наукових праць. — Вип. 16. Кн. 2. — К.: Видавничий Дім Дмитра Бураго, 2005. — С.189–193.
8. Сніжко Н. В., Сніжко М. Д. "Ідеографічний тезаурус" як інформаційно-довідкова система при вивченні закономірностей структурно-функціональної організації лексики // Мовознавство. — Київ, 1996. — № 4–5. — С. 23–28.
9. Соколовська Ж. П. Картина світу та ієрархія сем // Мовознавство. — Київ, 2002. — №6. — С.87–91.

**І. Васильєва\***

Інститут філології КНУ імені Т. Шевченка (Київ)  
УДК 81'33

#### ВИКОРИСТАННЯ КОМП'ЮТЕРНОГО ТЕЗАУРУСА В ДОСЛІДЖЕННІ МОВИ ПОЕТІВ

*In our research we propose the ways for using computer-based Ukrainian thesaurus and the frequency dictionary of Ukrainian poetry for creating poetry language picture. We consider lexico-semantic group "Human" in Ukrainian poetry discourse of 90-th.*

\* © І.Васильєва, 2006

Сьогодні, на щастя, уже не стоїть питання про доцільність застосування ЕОМ у лінгвістичних дослідженнях [2]. Використання комп'ютерних технологій у сучасній науці не лише полегшує роботу дослідників, оптимізує дослідницькі процеси, а й дозволяє ефективно використовувати отримані результати, головним чином як інструмент у дослідженнях суміжних галузей. Прикладами такої плідної "співпраці" у лінгвістиці можуть бути розвідки з галузі термінології, мови поетичних і художніх текстів, національних мовних картин світу тощо, проведені на базі комп'ютерних лексикографічних систем, різноманітних корпусів і комп'ютерних (частотних та ін.) словників [4; 1]. Запозичення інструментарію та даних лексикографічних систем має безліч переваг:

- дослідник отримує у своє розпорядження достовірну інформацію і готовий інструмент дослідження (для пошуку нового наукового результату йому, як правило, вже не потрібно винаходити велосипед);

- це, в свою чергу, позначається на економії ресурсів і часу й дозволяє досліднику оптимізувати роботу над об'єктом, зосередившись на головному;

- залежно від структури й логічної впорядкованості комп'ютерної лексикографічної бази даних або словника можна досліджувати декілька параметрів одночасно (наприклад, частотні, лексико-семантичні, граматичні й стилістичні характеристики тощо).

У цій доповіді ми розглянемо можливість використання комп'ютерного тезауруса української мови, а власне, іменників лексико-семантичної групи "Людина", для дослідження мови поетів. На основі отриманих результатів спробуємо побудувати фрагмент мовної картини світу українських поетів ХХ століття. Є чимало спроб створення тих чи інших картин світу, серед яких і чітко обґрунтовані методи із застосуванням семного аналізу [8]. Але нас зараз більше цікавить художня картина світу (частиною якої є поетична), яка відіграє суттєву роль в сприйнятті дійсності. З. Попова та І. Стернін представляють її як вторинну, що виникає в свідомості читача при сприйнятті художнього твору. [5]. Картина світу в художньому тексті відбиває індивідуальну картину світу в свідомості автора і створюється тими ж мовними засобами, але своєрідно. Ця своєрідність полягає у використанні окремих тематичних груп мовних одиниць, різній (підвищеній чи зниженій) частотності окремих одиниць та їхніх груп, в індивідуальному використанні образних засобів (системи тропів) тощо.

Створення тезауруса, як відомо, може відбуватися декількома шляхами: згори донизу (накладанням заздалегідь розробленої синоптичної (класифікаційної) схеми, яка покликана відображати мовну картину світу, на мовний матеріал) або знизу вгору (виведенням синоптичної схеми шляхом аналізу лексичних значень). Найдоцільнішим, на нашу думку, є поєднання цих двох підходів – самостійна розробка синоптичної схеми шляхом узагальнення мовних понять і порівняння її з відомими класифікаціями. Стосовно ж побудови тезауруса поетичної мови і створення поетичної мовної картини світу, то видається прийнятним варіант використання загальномовного тезауруса шляхом перетину поетичної і загальномовної баз даних. Проте, не маючи повного загальномовного тезауруса української мови, зараз можна говорити лише про дослідження окремих аспектів художніх поетичних явищ (напр., ідіостилю, тропіки, лексика різних семантичних груп) і побудову фрагментів поетичної мовної картини тощо [6].

Інструментом дослідження послужив створений нами на базі тлумачного словника комп'ютерний тезаурус, що представляє собою ієрархічно впорядковану структуру, одиниці якої об'єднані в семантичні поля.

Стаття тезауруса має ономаціологічну спрямованість – складається з реєстрової одиниці (концепту) та лексичних одиниць, що передають це поняття у мові. До складу останніх входять також синоніми, антоніми й деривати. У ролі лексичних відповідників концепту виступають лексеми, у яких категоріальна та диференційна ознаки змісту концепту відповідають категоріально-лексичній та диференційній семам значення слова. Таким чином узгоджуються поняття й значення, логічна та мовна частини словника.

Зауважимо, що лексичними одиницями тезауруса є окремі лексико-семантичні варіанти (ЛСВ), а не лексеми, оскільки лексичній семантиці іменників притаманна багатозначність. Різні ЛСВ одного іменника на позначення людини/особи можуть називати різні характеристики і властивості людини. Наприклад, різні ЛСВ лексеми *безштанько* співвідносяться з такими поняттями: *безштанько 1* – маленький хлопчик,

безитанько 2 – бідна людина. Саме тому при ідеографічній класифікації лексики враховувалися всі ЛСВ іменників, які позначають поняття людини/особи.

Джерелом відбору лексичного матеріалу до тезауруса був словник української мови в одинадцяти томах [7]. Масив лексики, що підлягає ідеографічному опису в тезаурусі, становить близько 15 тис. ЛСВ похідних (переважна більшість) і непохідних іменників, значення яких містять категоріально-лексичну сему “людина/особа”. Аналізуючи визначення іменників у тлумачному словнику, методом суцільної вибірки було відібрано іменники, субстантивовані прикметники та дієприкметники, що вживаються в значенні іменників, на позначення людей.

Оскільки мова служить не тільки засобом об’єктивізації, репрезентації світу предметів та ідей, але й забезпечує також сфери мовленнєво-мисленнєвої та емоційної діяльності людини, то крім нейтральної ми вважали за доцільне включити до вибірки емоційно та стилістично забарвлені, рідковживані, діалектні та застарілі (архаїзми та історизми) ЛСВ слів на позначення людей/осіб. Хоча тезаурус і представляє синхронний зріз лексики, такий вибір зумовлений як суто дослідницьким інтересом, так і тим, що однією з функцій тезауруса є необхідність задоволення пошуку “поняття – слово” і навпаки. До того ж, за запитом користувача до результатів пошуку в електронному тезаурусі можна не включати стилістично марковану лексику.

Далі методом компонентного аналізу виділялися родова (класифікаційна) і додаткові семи значення. За певними ознаками семи порівнювалися, і за результатами аналізу відповідні ЛСВ об’єднувалися в групи і семантичні поля.

Тезаурус представлений у вигляді індексованої електронної бази даних Access, де кожна одиниця має набір міток-параметрів: найближче родове поняття, тип семантичного зв’язку з іншою одиницею, стилістичну інформацію тощо.

Матеріалом для побудови поетичного тезауруса було обрано електронний частотний словник поетичної мови зі статистичними характеристиками, розроблений і створений на творах 15 авторів (періоду 1975–1995) колективом лабораторії комп’ютерної лінгвістики Інституту філології Київського національного університету імені Т. Шевченка (Н. П. Дарчук, Л. А. Алексієнко, О. М. Зубань та ін.) [3]. Словник створено на корпусі текстів обсягом 300 000 слововживань із творів поетів-шістдесятників (Л. Костенко, І. Драч, Б. Олійник, В. Стус, В. Коломієць, Т. Коломієць, Д. Павличко, М. Вінграновський, М. Воробйов), поетів сімдесятих років (І. Жиленко, С. Іовенко, Л. Скірда) і поетів дев’яностих (Ю. Андрухович, О. Забужко, І. Малкович). Довжина вибірки з текстів кожного поета – 20 000 слововживань. Частотний словник надає інформацію про закономірно повторювані лексеми, які є ключовими для індивідуального авторського словника та поетичного в цілому.

Побудова фрагмента тезауруса поетичної мови відбувалася в напівавтоматичному режимі: кожне слово із бази даних тезауруса іменників на позначення людини порівнювалося зі словами з п’ятнадцяти підвбірок поетів. Якщо знаходилася відповідна словоформа, їй приписувався номер лексеми тезауруса і номер значення, у якому вона зустрілася в тексті. Із частотного словника бралася відповідна статистична інформація про частоту лексеми і частоту словоформи. Результати цього етапу наведені у табл. 1. В індивідуальних тезаурусах не враховані поки що авторські неологізми на позначення людей, які не засвідчені в тлумачному словнику.

Таблиця 1. Частотні характеристики словника іменників окремих авторів

№	Автор	Кількість різних:				Кіл-ть ЛСВ на позначення людини	Відношення кількості лексем-назв людей загальної кількості іменників до	Відношення кількості словоформ на позначення людини до загальної кількості іменникових словоформ
		лексем	лексем-назв людей	словоформ	словоформ позначення людини			
1	Андрухович Ю.	3122	421	4933	670	445	13,5%	13,6%

2	Вінграновський М.	1751	118	3252	339	132	6,7%	10,4%
3	Воробйов М.	1608	139	3193	382	155	8,6%	12,0%
4	Драч І.	2404	251	3818	597	269	10,4%	15,6%
5	Жиленко І.	2219	196	3802	576	221	8,8%	15,1%
6	Забужко О.	2712	182	4181	375	192	6,7%	9,0%
7	Иовенко С.	1888	206	3186	513	221	10,9%	16,1%
8	Коломисць В.	2549	263	4188	666	298	10,3%	15,9%
9	Коломисць Г.	2034	186	3536	434	204	9,1%	12,3%
10	Костенко Л.	2606	333	4040	706	357	12,8%	17,5%
11	Малкович І.	1982	175	3159	279	165	8,8%	8,8%
12	Олійник Б.	2097	281	3463	840	316	13,4%	24,3%
13	Павличко Д.	2201	236	3830	496	253	10,7%	13,0%
14	Скирда Л.	1993	145	3338	431	161	7,3%	12,9%
15	Стус В.	2268	253	3664	512	270	11,2%	14,0%
Усього		13233	1548	29150	7816	3659	11,7%	26,8%

Процентні відношення частоти лексем на позначення людини до загальної кількості іменникових лексем (11, 7%) та частоти словоформ на позначення людини до загальної кількості іменникових словоформ (26, 8%) у частотному поетичному словнику дозволяють зробити припущення, що українська поезія другої половини ХХ ст. була певною мірою антропоцентричною. Однак це припущення вимагає додаткових досліджень і розрахунків: для більш точного аналізу необхідно знати кількість та наповнення інших ЛСГ ("Тварини", "Рослини", "Іжа", "Напої", "Засоби пересування", "Одяг" тощо), а залежно від способу поняттєвої класифікації різні дослідники виділяють неоднакову їх кількість від 3 до 100 і більше. Ще точнішими дані могли б бути, якби в розпорядженні дослідника був ще й корпус поетичних текстів.

Найбільший тезаурус назв людей і поезія у Ю. Андруховича – 421 різна лексема і 670 їхніх словоформ на позначення людини. Цікава ситуація у підвиріччі Б. Олійника: його тезаурус не надто великий – 281 лексем, проте вживає він їх найчастіше серед усіх поетів – 840.

Отже, можна зробити висновок: чим більша кількість лексем у поезії автора і висока частота вживання їхніх словоформ, тим багатший тезаурус поета. А невисока частота лексем поряд з високою частотою вживання їхніх словоформ може свідчити про усталеність світогляду автора, мислення усталеними категоріями, концептами.

Проте не всі лексико-семантичні поля і підгрупи однаково представлені: із близько 15 тис. різних ЛСВ іменників на позначення людини загальнономовного тезауруса лише 1714 наявні в поетичному словнику.

Найчастотнішими лексемами в поетичному словнику (див. табл. 2) виявилися слова на позначення загальної назви сукупності – **люди** (270).

Цікаво, що на другому місці за частотою лексема **жінка** (229), а далі – **мати** (203).

Підсумовуючи, можна припустити, що антропоцентризм української поезії має виражену ознаку статі: роль **жінки-матері** в ній займає провідні позиції. Це підтверджує і побудова синонімічної схеми поетичного тезауруса, яка засвідчила, що найчисельнішим є семантичне поле *Сімейних і Родинних стосунків* (чи не в повному обсязі представлено лексемами **мати, мама, діти, син, чоловік, брат, батько, дід, баба, дядько**). Другою за чисельністю і значущістю є група *Соціальних стосунків* (**друг, гість**) і теми **людини і народу**, а також тема мистецтва (частота лексеми **поет** може свідчити про пильну увагу та вимоги авторів до себе та інших митців).

Для більш точних висновків до аналізу можна залучити статистичні дані щодо частоти значень, у яких вжито найчастотніші слова, та їхні контексти. У цьому дослідженні проілюстровано приклад застосування тезауруса і наведено результати апробації методики укладання тезауруса на поетичних текстах.

На основі отриманих результатів проведено порівняння фрагментів мовних картин світу: загальнономовної та індивідуальної (поета). Отримані результати дозволяють зробити

припущення щодо антропоцентричності української поезії другої половини двадцятого століття (проте це припущення вимагає додаткових досліджень) із важливими темами родинних і суспільних стосунків, з провідною роллю жінки в поетичному дискурсі (про що свідчать кількісні характеристики ЛСГ "Людина" поетичного тезауруса).

Таблиця 2. Найчастотніші лексеми на позначення людини в частотному поетичному словнику

	Лексема	Автори															Усього в словнику поетів
		Андрухович Ю.	Драч І.	Йовенко С.	Коломієць Т.	Коломієць В.	Костенко Л.	Малкович І.	Олійник Б.	Павличко Д.	Скирда Л.	Стус В.	Вінграновські й М.	Воробйов М.	Забужко О.	Жигленко І.	
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1.	люди	3	17	7	24	44	50	5	22	16	14	18	10	17	16	7	<b>270</b>
2.	жінка	10	12	40	18	6	15	4	5	5	13	5	13	27	24	32	<b>229</b>
3.	мати	3	22	10	22	14	8		32	24	4	18	14	27	3	2	<b>203</b>
4.	діти	10	9	12	9	22	9	4	14	9	24	3	5	23	11	8	<b>172</b>
5.	син	7	24	4	17	20	8	4	25		10	28	4	4	1	16	<b>172</b>
6.	людина	2	9	14	15	19	11	1	11	14	33	9	8	9	1	15	<b>171</b>
7.	народ	7	8	2	7	34	17	1	25	11	15	3	21	1	5	1	<b>158</b>
8.	мама	1	17	21	11	16	2	10	16	2	21	6	9	3	9	12	<b>156</b>
9.	поет	10	9	13	4	9	20	7	29	4	10	7	6	2	11	10	<b>151</b>
10.	друг	1	7	8	6	8	7		7	10	33	18	6	5	3	21	<b>140</b>
11.	чоловік	4	12	12	1	5	7	3	22	3	11	6	4	25	14	8	<b>137</b>
12.	брат	5	5	9	2	7	3	5	16	13	9	5	4	18	3	7	<b>111</b>
13.	дитина		13	19	2	7		8	8	12	7		7	2		6	<b>91</b>
14.	хлопець	5	4		16	10	7	1	25	2	1	3	1	4	4	1	<b>84</b>
15.	дитя	2	4	19	2	4	5		2	6	18		6			14	<b>82</b>
16.	батько	1	8	6	5	14	3		11	9	5	1	5	5	3		<b>76</b>
17.	дід	1	18		4	11	3	2	16	1	2		7	5		4	<b>74</b>
18.	баба	1	20		2	1	6		4			1	7	4		9	<b>55</b>
19.	дівчина	3	9		3	8	1		4	12	1	3	4	5	1	1	<b>55</b>
20.	гість	4	3		7	2	4	2			2	1	1	8	1	16	<b>51</b>
21.	дядько		15		3	6			5	3			13	2		4	<b>51</b>

## Література

1. Алексеенко Л. А., Дарчук Н. П., Зубань О. Н., Сорокин В. В. Параметризованная база данных поэтической речи как источник и инструмент филологических студий // Материалы международной конференции "Компьютерная лингвистика без границ", Санкт-Петербург, 25–26 марта 2004 г. – СПб., 2004.
2. Баранов А. Н. Автоматизация лингвистических исследований: корпус текстов как лингвистическая проблема // Русистика 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 ХХ–ХХІ 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 Інститутом 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 ( $V_{\phi}$ ), обсяг словника лексем ( $V$ ), багатство словника ( $B$ ), індекс винятковості ( $E$ ), середня повторюваність слова у тексті ( $A$ ) тощо.

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