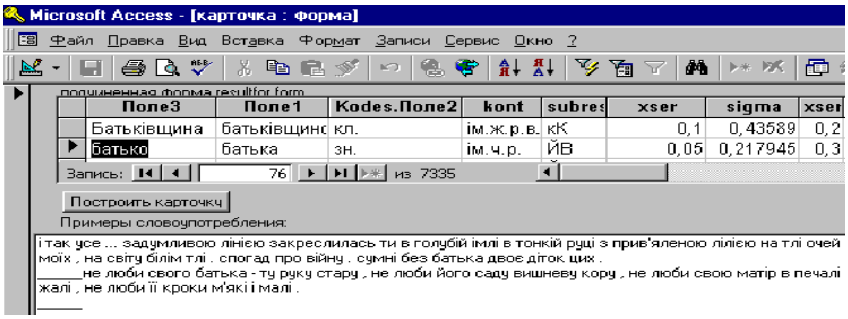
Словник лексики кожного поета через словник-конкорданс модуля-аналізатора має постійний зв'язок із корпусом текстів, що дозволяє автоматично ілюструвати текстове вживання кожної словоформи і будувати електронну картку (рис. 3).

Така картка дозволяє швидко та ефективно створювати різноманітні лексичні словники: *Словники синонімів, омонімів, неологізмів, діалектизмів, архаїзмів, фразеологізмів; Словники тропів: епітетів, метафор, метонімії, синекдох, порівнянь, оксюморонів, гіпербол.*

Ці словники укладаються в автоматизованому режимі, оскільки така інформація не може бути повністю формалізованою і вимагає участі лінгвіста. Наприклад, укладання словника неологізмів, словника синонімічних гнізд, ідеографічного словника супроводжується автоматичним морфемним та словотвірним аналізом, але вимагає тлумачення значення слова, введення його до відповідної лексико-семантичної групи синоптичної схеми або передбачає контроль за значенням слів при зіставленні їх із тією чи іншою тематичною групою електронного словника синонімічних гнізд.

Методика формалізованого лінгвістичного опису мовних одиниць у створенні системи автоматизованого лінгвістичного аналізу є узагальненням теоретичних і прикладних ідей сучасного мовознавства, що робить цю систему надзвичайно ефективним і раціональним інструментом лінгвістичних досліджень.

Рисунок 3. Фрагмент Частотного словника поетичного мовлення М. Вінграновського з текстовою ілюстрацією



Ця методика може бути використана як еталонна при укладанні різноманітних електронних словників, картотек, автоматизованій класифікації лінгвістичного матеріалу,побудові навчальних комп'ютерних тренажерів, тестів тощо. З іншого боку, отримана параметризована інформація про організацію українських текстів на різних рівнях мовної системи дає можливість вивчити закономірності функціонування системи мови в різних стилях, комплексно досліджувати мовні особливості ідіостилів українських поетів та письменників.

В. Старко, к. філол. н.[*]
Волинський державний університет імені Лесі Українки (Луцьк)
УДК 161.2.81'373.374.322

ФРАЗЕОЛОГІЯ, ЛЕКСИКОГРАФІЯ І КОРПУС

The article deals with the problems of multiaspect investigation and lexicographic description of phraseology based on corpus data. The author discusses the frequency of idioms and proverbs and the demand it places on the size and content of corpora and compares three different corpora with respect to their suitability for phraseology research.

Корпусні дослідження фразеології є відносно новою галуззю лінгвістики. Чи не найзначніший внесок в її розвиток внесла Розамунд Мун, чия робота [3] була першою

монографією на цю тему. І до сьогодні, за нашими даними, ця робота залишається найгрунтовнішим корпусним дослідженням фразеології. У своїй книзі авторка заторкнула низку важливих проблем, пов'язаних з дослідженням фразеології, що спирається на корпусні дані. Серед цих проблем варто виокремити такі: типологія фразеологічних одиниць (ФО), їхня частотність та розподіл частотності за типами ФО, вимоги до обсягу, наповнення й репрезентативності корпусу з огляду на специфіку ФО, варіативність ФО та її типи, дискурсивні функції ФО, прагматика й стилістика фразеології, роль ФО в забезпеченні зв'язності тексту тощо. Дальший розвиток ідеї Р. Мун набули у статті [4], де було розглянуто деякі проблеми корпусної фразеографії.

Предметом розгляду цієї статті є питання частотності ФО, зокрема прислів'їв, та обсягу корпусу, потрібного для повноцінного лексикографічного опису фразеології. Як матеріал для аналізу ми використаємо англійські фразеологізми. Наше дослідження є частиною проекту укладення «Англійсько-українського фразеологічного словника».[1]

Почнімо з деяких спостережень, що їх зробила Р. Мун у своїх роботах. Вона зауважує, що до 1990-го року обсяг наявних корпусів (до 20 млн. слововживань) був замалим для надійного дослідження фразеології [4]. Навіть її робота 1998 року [3] спирається лише на 18-мільйонний корпус англійської мови. Тому багато важливих висновків, за визнанням самої авторки, потребують підтвердження на ширшому матеріалі. Поява Британського національного корпусу (БНК, 100 млн. слововживань) та Банку англійської мови (БАМ, 323 млн. слововживань у 1997р., 450 млн. у 2005р., на 2006р. заплановано 530 млн.) означає, за словами Р. Мун, що «ідіоми й подібні до них вислови мають більшу статистичну ймовірність появи [в корпусі], однак цих даних все одно може забракнути, аби адекватно їх описати» [4: 265]. Однією з причин цього є те, що переважна більшість ФО мають набагато нижчу частотність, ніж окремі слова мови. Інша причина полягає в тому, що для опису ФО потрібно мати в наявності суттєво більше контекстів, ніж цього вимагає опис однослівної одиниці. Це зрозуміло, якщо взяти до уваги наявність різних типів фразеологізмів та, що важливіше, комплексну природу ФО, яка виявляється в їхній несистематичній й часто непередбачуваній варіативності, трансформаційному потенціалі, різноманітних дискурсивних й прагматичних функціях тощо. Тому 10 слововживань може бути достатньо, аби дати лексикографічний опис, наприклад, певному терміну, але замало для ФО, що має таку саму частотність вживання. Аби дати приблизну ідею про частотність ФО, використаймо влучний приклад з [4: 266]: англійська ідіома *looking for a needle in a haystack* має (з усіма варіантами) частотність 25 на кожні 100 млн. слововживань. Таку саму частотність мають слова *etiology, disequilibrium, granddaddy, hollowness, igneous, methylated, strychnine, timorous* та *ungentlemanly.*

Оскільки на теперішньому етапі проекту ми працюємо з англійськими прислів'ями, нас цікавить співвідношення частоти ФО взагалі і прислів'їв в англійській мові. Аби продемонструвати це співвідношення проаналізуємо дані, наведені в [3] – їх наведено в таблиці 1.

Таблиця 1. Співвідношення частотності ФО й прислів'їв в англійській мові.

Вживань на 100млн.	**ФО**	**Прислів'я**
< 6	8%	5%
6–22	32%	11%
22–100	32%	3%
100–200	12%	<0, 1%
200–500	9%	<0, 1%
500–1000	4%	0
1000–5000	3%	0
5000–10000	<1%	0
> 10000	<1%	0

[1] Проект був підтриманий грантом від Американської ради наукових товариств (American Council of Learned Societies) та стипендією ім. В. Фулбрайта від уряду США.

У першому стовпчику наведено кількість вживань на 100 млн. одиниць корпусу. Ці дані було отримано шляхом заокруглення даних, що їх дослідниця наводить для використовуваного нею 18-мільйонного корпусу під назвою «Оксфорський експериментальний корпус Гектор» (Oxford Hector Pilot Corpus). У другому стовпчику подано відсоток ФО (від загальної кількості ФО, проаналізованих в роботі Р. Мун, що складає 6776 одиниць), які належать до кожної смуги частотності. В третьому стовпчику в такий самий спосіб наведено відсоток тих ФО, що є прислів'ями (знову ж таки, від загальної кількості ФО). Як бачимо, через невеликий обсяг корпусу частотність вживання прислів'їв фактично перебуває на межі випадку. Тим не менше навіть за таких умов можна простежити тенденцію вживання ФО й прислів'їв.

Слід завважити, що корпус «Гектор» не є збалансованим. Майже на 60% він складається з газетних текстів, приблизно на 11% з художньої літератури і на 18% зі спеціальної (нехудожньої) літератури. Його незбалансованість має і негативні, і певні позитивні наслідки. До негативних слід зарахувати неможливість дійти остаточних висновків щодо відносної частотності вживання мовних одиниць в різних жанрах й їхніх відмінних характеристик у кожному окремому жанрі. Одним з позитивних наслідків є те, що переважання публіцистики в корпусі, як це не парадоксально, сприяє дослідженню ФО, адже, як показують наукові розвідки, зокрема й [3], [4], з усіх жанрів прислів'я і чимало інших типів ФО найчастіше вживають саме в публіцистичному жанрі. У збалансованому корпусі частотність ФО буде нижчою. Ці міркування приводять нас до логічного висновку, що для ґрунтовного обстеження й лексикографічного опису прислів'їв дослідникові необхідно мати корпус максимального розміру, щонайменше на порядок більший, ніж для дослідження інших ФО. Якщо не стоїть завдання порівняльного міжжанрового аналізу ФО, то бажаною є щонайвища частка публіцистичних текстів.

Працюючи над лексикографічним описом англійських прислів'їв, ми маємо змогу користуватися саме таким корпусом. Йдеться про корпус газетних та журнальних публікацій англійською мовою під назвою «ЛексисНексис Академік» (LexisNexis Academic), що до нього надає доступ компанія «Рід Елсевір Інк.»[1] [2]. Нас цікавитиме кілька питань: 1) обсяг цього корпусу; 2) його співвідношення з обсягом БАМ-у; 3) розподіл частоти вживання ФО в корпусі за часовими періодами; 4) достатність/недостатність корпусу для повноцінного дослідження англійських ФО, зокрема прислів'їв.

Обсяг корпусу «ЛексисНексис Академік» точно встановити неможливо з тієї причини, що він щодня поповнюється новими текстами. Однак приблизне уявлення сформувати все ж таки можна, і в цьому нам допоможе порівняння його з БАМ-ом. Ми скористалися даними Р. Мун, наведеними в [4], щодо частоти вживання семи ідіом в БАМ-і 1997р. (323 млн. слововживань) і порівняли їх з вживаністю цих самих ідіом в «ЛексисНексис» станом на 27.02.2006р. Результати порівняння подано в таблиці 2.

У першому стовпчику наведено ідіоми в тій формі, що її було задано в пошуковій команді в роботі з БАМ-ом, а в другому – кількість їх вживань (абсолютні дані). Запис «needle w/5 haystack» означає, що було задано пошук лемми «haystack» в межах п'яти позицій від лемми «needle» (це ключові елементи ідіоми *looking for a needle in a haystack* та її варіантів), дужки позначають факультативні елементи, скісна риска – альтернативні елементи. В наступних чотирьох стовпчиках наведено кількість вживань в абсолютному вимірі в корпусі «ЛексисНексис» за часовими періодами. П'ятий стовпчик подає сумарну кількість, а шостий – відношення даних в другому стовпчику до даних в п'ятому.

Таблиця 2. Порівняння БАМ-у й корпусу «ЛексисНексис».

Ідіома	**БАМ**	**«ЛексисНексис» станом на 27.02.2006р.**					**Пропорція**
		1976–89	**1990–94**	**1995–99**	**2000–06**	**Всього**	
give smb.	64	141	288	443	792	1664	1:26

[1] Ми висловлюємо вдячність керівництву Нью-Йоркського державного університету в Олбані (США) за люб'язний дозвіл користуватися доступом до згаданого корпусу через університетську бібліотеку.

the cold shoulder							
(let) the cat out of the bag	91	302	525	1003	1503	3333	1:36
needle w/5 haystack	85	250	538	1144	1874	3806	1:45
beat about/around the bush	109	271	554	865	1407	3097	1:28
bury the hatchet	110	506	977	1620	2115	5218	1:47
bite the dust	145	636	1537	2573	3220	7088	1:49
(skate) on thin ice	241	400	890	1500	2320	5110	1:21

Ми обмежили пошук в «ЛексисНексис» лише газетними публікаціями, залишивши поза увагою журнальні статті. Підрахунок кількості вживань перших двох ідіом було проведено вручну, причому ми з'ясували, що через різні чинники (головно через повторення тих самих статей та джерела, що їх ми ігнорували, наприклад, англійськомовні газети Японії чи Китаю) нам довелося відкинули 12% вживань ідіом, переважно в посиланнях від 1990-х років і пізніше. Для решти ідіом ми використали формулу: загальна кількість вживань (що її подає пошукова система) х 0, 87. У такий спосіб ми відкинули 13% вживань за всі роки й отримані приблизні дані подали в таблиці.

Як бачимо, співвідношення між частотою вживання вибраних ідіом у БАМ-і й корпусі «ЛексисНексис» коливається в межах від 1:21 до 1:49. Якщо припустити, що «ЛексисНексис» лише втридцятеро більший за БАМ, то його обсяг має становити понад 9 млрд. слововживань. Такий обсяг більш ніж достатній для дослідження й лексикографічного опрацювання ідіом. БАМ має зрости в 1, 6 рази проти 1997 року, але й в такому обсязі він значно поступатиметься корпусу «ЛексисНексис», а через деякий час, цілком ймовірно, ще більше відстане від нього. Тисячі прикладів вживання ФО, що їх дослідник дістає в розпорядження, дають змогу, як ми переконалися на власному досвіді, відтворити їхній повний «лексикографічний портрет» і є унікальним джерелом багатоаспектного дослідження англійської фразеології. З іншого боку, менше 100 контекстів часто не вистачає, аби простежити тенденції вживання, і в цьому полягає обмеженість БАМ-у. «ЛексисНексис» містить тексти не тільки американських і британських газет, а й канадських, австралійських і новозеландських, що уможливлює докладну перевірку географічного поширення тієї чи іншої одиниці англійської мови. З таблиці видно, що кількість вживань ідіом зростає майже в геометричній прогресії від 1976 до 2006 року, а в проміжку 2000–2006рр. складає замалим 50% від загальної. Це створює сприятливі умови для дослідження найновітнішого зрізу мови.

Заради порівняння наведімо дані з роботи [1], де проаналізовано вживання 103 «основних» англійських ідіом у БНК й показано, що найчастотніші з них трапилися в корпусі не більше, ніж 500 разів, переважна більшість – менше, ніж 100 разів, з них чимало – менше, ніж 20 разів.

Слід зауважити, що «ЛексисНексис» є корпусом в тому сенсі, що це збірка задокументованих автентичних електронних текстів із пошуковими засобами. Він цілком складається із публіцистичних текстів (тобто є незбалансованим), і його вміст не було проаналізовано морфологічно чи синтаксично автоматичними чи іншими засобами. В цих аспектах він відрізняється від корпусів на кшалт БНК і БАМ-у і, безумовно, поступається їм. Однак у плані знаходження контекстів вживань ФО його гнучка пошукова система, зручний інтерфейс і величезний обсяг дають йому незрівнянну перевагу над іншими корпусами.

Що стосується придатності корпусу «ЛексисНексис» для дослідження англійських прислів'їв, то на його користь говорить, звичайно, його величезний обсяг. Нижче, в таблиці 3, наведено порівняльні дані частоти вживання 10 англійських прислів'їв у БНК та корпусі «ЛексисНексис»[1].

[1] Оскільки ми не мали змоги провести пошук у більшому за обсягом БАМ-і, ми вдалися до БНК, використавши пошуковий інтерфейс за адресою http://view.byu.edu/. Важливою особливістю цього інтерфейсу є те, що він не має мінімальних обмежень щодо кількості видобутих з корпусу рядків.

У першому стовпчику наведено прислів'я (в основній формі), у другому і третьому – кількість знайдених вживань у БНК і «ЛексисНексис» відповідно.

Таблиця 3. Частота вживання прислів'їв у БНК та корпусі «ЛексисНексис».

Прислів'я	**БНК**	**«ЛексисНексис»**
Necessity is the mother of invention	7	1353
(The love of) money is the root of all evil	7	910
Absence makes the heart grow fonder	7	498
An apple a day keeps the doctor away	2	348
tation is the sincerest form of flattery	1	1493
ke while the iron is hot	1	762
k of all trades and master of none	1	482
apple doesn't fall <never falls> far from the tree	0	576
A good name is better <rather be chosen> than great riches	0	23
Appetite comes with eating	0	10

Для перших восьми прислів'їв ми використали згадану вище формулу обчислення кількості вживань ФО в «ЛексисНексис», а для решти підрахунок було здійснено вручну. Як засвідчує таблиця, БНК значно поступається «ЛексисНексис» і не дає змоги сформувати реальної картини функціонування прислів'їв. Наприклад, маючи в розпорядженні лише БНК, можна було би зробити хибний висновок, що прислів'я *Absence makes the heart grow fonder* є досить вживаним, *Jack of all trades and master of none* – вживаним рідко, а *An apple doesn't fall <never falls> far from the tree* – взагалі вийшло з ужитку. «ЛексисНексис» показує, що ці три прислів'я мають приблизно ту саму – досить високу – частотність і жодне з них не виходить з ужитку. Щодо останніх двох прислів'їв у таблиці, то корпус засвідчує, що вони є досить рідко вживаними, але все ж наявні в колективній свідомості носіїв мови, хай навіть у пасивному запасі.

Отже, для повноцінного, багатоаспектного дослідження фразеології та її лексикографічного опису необхідно мати в розпорядженні збалансований корпус обсягом на порядок більшим, ніж БАМ чи БНК. Таких корпусів англійської мови, за нашими даними, наразі немає, однак корпус «ЛексисНексис», попри свою незбалансованість, має більш ніж достатній обсяг і дає хороші можливості дослідження англійської фразеології, в тому числі й прислів'їв, що мають значно нижчу частотність вживання проти інших фразеологізмів.

Література

1. Grant L. E. Frequency of the 'core idioms' in the British National Corpus (BNC) // International Journal of Corpus Linguistics. – 2005. – С. 429–451.
2. LexisNexis Academic, http://www.lexisnexis.com/academic/universe/academic/, 27.02.2006р.
3. Moon R. Fixed Expressions and Idioms in English. A Corpus-Based Approach. – Oxford: Clarendon Press, 1998. – 338 с.
4. Moon R. Needles and Haystacks, idioms and corpora: Gaining insights into idioms, using corpus analysis // The Perfect Learners' Dictionary. – Tübingen: Max Niemeyer Verlag, 1999. – С. 265–281.

О. Тищенко, к. філол. н.[*]
Інститут української мови НАН України (Київ)
УДК 161.2.81'322.221.24

ЗАСАДИ СТВОРЕННЯ КОРПУСУ УКРАЇНСЬКОЇ ЖЕСТОВОЇ МОВИ ГЛУХИХ

У статті обґрунтовано актуальність і мету створення корпусу української жестової мови глухих як автономної комунікативної системи. Означено змістові, структурні, інструментальні й технічні принципи такої праці.

Актуальність. Уходження в лінгвоукраїністику такого напряму, як корпусна лінгвістика [3; 11], спонукає до розвідок зі створення корпусу не лише природної словесної мови, а й української жестової мови глухих (УЖМГ). Це зумовлено активізацією досліджень, зокрема й лінгвістичних, національних жестових мов як за кордоном [1; 2; 3; 7; 8; 19; 20], так і в Україні [10; 12; 13; 17] у зв'язку з гуманізацією та соціалізацією всіх галузей науки [9].