Валентина Балог, к. ф. н.[*]
Орися Демська-Кульчицька, к. ф. н.[*]
Інститут української мови НАН України (Київ)
УДК 161.2.81’374.72’22

НАЦІОНАЛЬНИЙ КОРПУС УКРАЇНСЬКОЇ МОВИ: ПІДКОРПУС ХУДОЖНЬОЇ ЛІТЕРАТУРИ

У статті висвітлено засади формування та формалізації матеріалів підкорпусу художньої літератури в складі генерального корпусу НКУМ: обґрунтовано принципи добору текстів, визначено кількість та обсяг текстових фрагментів, подано адаптовану систему кодування первинних даних.

Створення *Національного корпусу української мови* [1] є першою спробою застосування загальної теорії корпусної лінгвістики до української мови, наслідком чого має стати стандартний загальномовний корпус, зіставний з аналогічними світовими загальномовними текстовими корпусами національного типу. Такий підхід, на наш погляд, повинен розв’язати низку завдань, найважливішими з яких є впровадження корпусно базованих методик дослідження української мови, формування корпусної лінгвоукраїністики як окремого напряму, а також забезпечення технологічного статусу української мови в умовах інформаційного суспільства.

Національний корпус – це зібрання текстів, що репрезентують національну мову на певному(их) етапі(ах) її існування в усьому різноманітті жанрів, стилів, історичних, територіальних і соціальних варіантів. Більшість мов світу й усі європейські мови на сьогодні вже мають власні національні корпуси, які статусом прирівнюють до

академічних граматик та великих тлумачних словників національних мов. Відповідно до цієї дефініції сформулюємо визначення поняття *'національний корпус української мови'*.

Національний корпус української мови (НКУМ) – *це перетворена на електронну форму, стандартно організована й програмно оброблена вибірка текстів сучасної української мови, репрезентативна для всіх функціональних рівнів загальнонародної мови, призначена для лінгвістичного аналізу й технологічного застосування*.

Стосовно лінгвістичного аналізу, то йдеться про академічні дослідження різних рівнів мовної системи (лексики, фразеології, термінології, морфології та синтаксису, словотвору, орфографії тощо), про методику української мови як рідної й як іноземної тощо. Натомість, технологічне застосування передбачає використання корпусу з метою побудови машинної мовної моделі як основи для розробок у галузі інформаційних технологій, створення програм автоматичного розпізнавання й синтезу українського мовлення, забезпечення автоматичних методів перетворення українськомовної текстової інформації, лінгвістичної підтримки автоматичних систем управління.

З погляду типологічно-аплікативних характеристик НКУМ розглядаємо як

- **дослідницький**: орієнтований на широкий клас передусім лінгвістичних завдань;
- **фрагментний**: збудований з текстових фрагментів чи уривків текстів, відібраних за попередньо визначеними засадами відбору текстових даних до корпусу;
- **мішаний**: передбачає введення текстових фрагментів обох варіантів реалізації мовної системи – писемних і усних;
- **динамічний**: передбачає константне поповнення множини корпусних текстів;
- **синхронний**: охоплює рівень сучасної української мови;
- **загальнонародної (національної) мови**: з урахуванням територіальної специфіки як у межах України, так і за її межами;
- **мономовний**: тексти, що увійшли до корпусу, є результатом мовної діяльності носіїв української мови;

– **морфологічно анотований**: усі текстові дані розмічені до рівня слова й кожне слово передбачає маркування частиномовної належності та відповідних морфологічних значень.

Побудова довільного корпусу на етапі проекту вимагає обов'язкової параметризації предметної галузі, репрезентованої корпусом. Оскільки національний корпус є зібранням електронних текстів, завдання яких репрезентувати національну мову на певному етапі(ах) її існування в усьому різноманітті жанрів, стилів, історичних, територіальних і соціальних варіантів, то предметну галузь НКУМ повинна становити українська загальнонародна мова в таких формах її існування, як літературна мова, територіальний діалект і соціолект. Найповніше в НКУМ повинна бути репрезентована літературна мова – основна наддіалектна форма існування природної мови, ознаками якої є опрацьованість, унормованість, поліфункціональність, стилістична диференціація, фіксованість.

Змістова параметризація предметної галузі НКУМ передбачає визначення джерельної текстової бази корпусу з наступною стилістично-жанровою стратифікацією відібраних текстів. Тут стильова диференціація мови можлива, по-перше, за характером мовної експресії, і, по-друге, за характером суспільної функції мови.

За характером мовної експресії традиційно розрізняють високий, середній і низький стиль. В основі такої тристильової диференціації лежить концепція залежності між предметом викладу, тематикою і добором мовних засобів та жанрів. Цей поділ, успадкований європейською традицією доби Відродження й бароко з александрійської філософської школи античного періоду [2, с. 39], детермінував жанровий підхід до параметризації предметного домену низки національних корпусів європейських мов.

В українській традиції тристильова диференціація мови, яку, зокрема, розробляли Ф. Прокопович, М. Довгалевський, Г. Кониський, утратила актуальність у процесі історичного розвитку стилістичної системи української мови, коли три традиційні стилі ('слоги') занепали, а основними одиницями стильової диференціації української мови стали структурно-функціональні стилі. Таким чином, в українську лінгвістичну традицію увійшов функціональний принцип стилістичної

диференціації мови, і ця диференціація покладена в основу змістової параметризації НКУМ.

Приймаємо, що сучасна українська мова на рівні стильової диференціації вкладається в семиелементну систему: 1) художній, 2) науковий, 3) офіційно-діловий, 4) публіцистичний, 5) конфесійний, 6) розмовний і 7) епістолярний стилі. А залежно від реалізації мовної системи виділяємо тексти усного й писемного варіантів національної мови. Художній, науковий, офіційно-діловий і публіцистичний стилі в межах НКУМ головно репрезентовані писемними текстами, теологічний – писемними і усними, а розмовний – усними.

Зазначимо, що корпус текстів лише тоді відповідатиме загальноприйнятим вимогам корпусу національного типу і виконуватиме всі передбачені для нього функції, коли в ньому адекватно репрезентовано обидві форми функціонування мовної системи. Але для багатьох галузей мовної практики писемна форма національної мови є пріоритетною і вона зазвичай реалізує себе в художньому, науковому, офіційно-діловому і публіцистичному стилях. У НКУМ приймаємо, що фактичний матеріал писемного варіанту української мови становитимуть тексти таких функціональних стилів:

1) художнього;
2) наукового;
3) офіційно-ділового;
4) публіцистичного;
5) конфесійного;
6) епістолярного.

Художній стиль у НКУМ репрезентуватимуть прозові, поетичні й драматургічні тексти відповідних хронологічних меж (підкорпуси сучасної української мови й української мови зламу XX / XXI ст.) з такою жанровою диференціацією:

– проза: роман, повість, оповідання, новела;
– поезія: вірш, сонет, поема, балада, сатира;
– драматургія: драма, комедія, трагедія.

Загальний обсяг корпусу детермінує обсяг його складових як засіб досягнення збалансованості корпусу. Незбалансованість дослідницького

корпусу національного типу призводить до неможливості проведення коректних досліджень на його основі, оскільки такий корпус не відтворює загальної мовної картини через відхилення в бік того чи іншого мовного матеріалу. Наприклад, якщо не збалансувати дані усної і писемної форми мови, то за умови перенасичення даних усними текстами матимемо неадекватну картину функціонування норм. Отже, пропонуємо такі загальнокорпусні кількісні параметри НКУМ (див. табл. 1).

Таблиця 1. *Загальні кількісні параметри НКУМ*

характер текстових даних	**відсоток від підкорпусу**	**кількість слововживань**
ПИСЕМНІ ТЕКСТИ (850 тис. слів – 85 %)		
Художній стиль (350 тис. слів – 35%)		
Проза (250 тис. слів – 25%)		
роман	10%	100 тис.
повість	10%	100 тис.
оповідання	3%	30 тис.
новела	2%	20 тис.
Поезія (50 тис. слів – 5%)		
вірш (лірика)	1%	10 тис.
сонет	1%	10 тис.
поема	1%	10 тис.
балада	1%	10 тис.
сатира	1%	10 тис.
Драматургія (50 тис. слів – 5%)		
драма	2%	20 тис.
комедія	1,6%	16 тис.
трагедія	1,4%	14 тис.
Науковий стиль (100 тис. слів – 10%)		
Власне науковий підстиль (20 тис. слів – 2%)		
тексти гуманітарних наук	1%	10 тис.

тексти природничих і точних наук	1%	10 тис.
Науково-популярний підстиль (20 тис. слів – 2%)		
тексти гуманітарних наук	1%	10 тис.
тексти природничих і точних наук	1%	10 тис.
Науково-методичний підстиль (20 тис. слів – 2%)		
тексти гуманітарних наук	1%	10 тис.
тексти природничих і точних наук	1%	10 тис.
Професійно-технічний підстиль (40 тис. слів – 4%)		
культура і мистецтво	1%	10 тис.
міжнародні відносини	0,5%	5 тис.
бізнес	1%	10 тис.
медицина	1%	10 тис.
техніка	0,5%	5 тис.
Офіційно-діловий стиль (50 тис. слів – 5%)		
законодавчо-правові державні документи	2,5%	25 тис.
організаційно-службові документи	1%	10 тис.
суспільна документація	1,5%	15 тис.
Публіцистичний стиль (250 тис. слів – 25%)		
періодика	24%	240 тис.
візуальна реклама	0,5%	5 тис.
учнівські та студентські неопубліковані твори	0,5%	5 тис.
Конфесійний стиль (50 тис. слів – 5%)		
Біблія	3%	30 тис.

твори літургійного призначення	2%	20 тис.
Епістолярний стиль (50 тис. слів – 5%)		
Класична кореспонденція (40 тис. слів – 4%)		
офіційно-ділове внутрішньодержавне листування	1%	10 тис.
офіційне міждержавне листування	1%	10 тис.
приватне листування	2%	20 тис.
Електронна кореспонденція (10 тис. слів – 1%)		
e-mail листування	1%	10 тис.
УСНІ ТЕКСТИ (150 тис. слів – 15 %)		
Офіційно-діловий стиль (70 тис. слів – 7%)		
живі службово-організаційні розмови	2%	20 тис.
телефонні службово-організаційні розмови	1%	10 тис.
виступи, повідомлення, оголошення на зборах	4%	40 тис.
Розмовний стиль (50 тис. слів – 5%)		
побутові розповіді	1.5%	15 тис.
побутове діалогічне мовлення	0,5%	5 тис.
повідомлення в транспорті	0,5%	5 тис.
аудіореклама	1%	10 тис.
телефонна побутова комунікація	0,5%	5 тис.
діалектне мовлення	1%	10 тис.

Конфесійний стиль (30 тис. слів – 3%)		
проповіді	3%	30 тис.
Всього підкорпус:	100%	1 млн.

Загальний обсяг НКУМ становитиме **1 млн. с**лововживань. Очевидно, що згідно із запропонованою стратегією розвитку корпусу, ця величина є початковою чи вихідною.

Наступним кроком проектування статистики НКУМ є визначення кількісних параметрів текстових фрагментів кожного із функціональних стилів / підстилів і кількості слововживань у кожному конкретному текстовому фрагменті (див. табл. 2). Точкою відліку для цих розрахунків є загальна кількість слововживань генерального корпусу, у нашому випадку 1 млн. слововживань.

Таблиця 2. *Індивідуальні кількісні параметри тестових фрагментів НКУМ*

характер текстових даних	**кількість фрагментів**	**кількість слів у фрагменті**
Писемні тексти		
Художній стиль		
Проза (250 тис. слів)		
роман (100 тис. слів)	20	5 тис.
повість (100 тис. слів)	20	5 тис.
оповідання (30 тис. слів)	10	3 тис.
новела (20 тис. слів)	10	2 тис.
Поезія (50 тис. слів)		
вірш (лірика) (10 тис. слів)	20	500
сонет (10 тис. слів)	20	500
поема (10 тис. слів)	10	1 тис.
балада (10 тис. слів)	10	1 тис.
сатира (10 тис. слів)	20	500
Драматургія (50 тис. слів)		
драма (20 тис. слів)	10	2 тис.
комедія (16 тис. слів)	8	2 тис.
трагедія (14 тис. слів)	7	2 тис.

Науковий стиль (100 тис. слів)		
Власне науковий підстиль (20 тис.)		
тексти гуманітарних наук (10 тис. слів)	10	1 тис.
тексти природничих наук (10 тис. слів)	10	1 тис.
Науково-популярний підстиль (20 тис. слів)		
тексти гуманітарних наук (10 тис. слів)	10	1 тис.
тексти природничих наук (10 тис. слів)	10	1 тис.
Науково-методичний підстиль (20 тис. слів)		
тексти гуманітарних наук (10 тис. слів)	10	1 тис.
тексти природничих наук (10 тис. слів)	10	1 тис.
Професійно-технічний підстиль (40 тис. слів)		
культура і мистецтво (10 тис. слів)	10	1 тис.
міжнародні відносини (5 тис. слів)	5	1 тис.
бізнес (10 тис. слів)	10	1 тис.
медицина (10 тис. слів)	10	1 тис.
техніка (5 тис. слів)	5	1 тис.
Офіційно-діловий стиль (50 тис. слів)		
законодавчо-правові державні документи (25 тис. слів)	10	2.5 тис.
організаційно-службові документи (10 тис. слів)	20	500
суспільна документація (15 тис. слів)	10	1.5 тис.
Публіцистичний стиль (250 тис. слів)		
періодика (240 тис. слів)	120	2 тис.

візуальна реклама (5 тис. слів)	100	50
учнівські та студентські неопубліковані твори (5 тис. слів)	10	500
Конфесійний стиль (50 тис. слів)		
Біблія (30 тис. слів)	15	2 тис.
твори літургійного призначення (20 тис. слів)	10	2 тис.
Епістолярний стиль (50 тис. слів)		
Класична кореспонденція (40 тис. слів)		
офіційно-ділове внутрішньодержавне листування (10 тис. слів)	10	1 тис.
офіційне міждержавне листування (10 тис. слів)	10	1 тис.
приватне листування (20 тис. слів)	20	1 тис.
Електронна кореспонденція (10 тис. слів)		
e-mail листування (10 тис. слів)	20	500
Усні тексти		
Офіційно-діловий стиль (70 тис. слів)		
живі службово-організаційні розмови (20 тис. слів)	40	500
телефонні службово-організаційні розмови (10 тис. слів)	40	250
виступи, повідомлення, оголошення на зборах (40 тис. слів)	40	1 тис.

Розмовний стиль (50 тис. слів)		
побутові розповіді (15 тис. слів)	15	1 тис.
побутове діалогічне мовлення (5 тис. слів)	10	500
повідомлення в транспорті (5 тис. слів)	33	150
аудіореклама (10 тис. слів)	66	150
телефонна побутова комунікація (5 тис. слів)	20	250
діалектне мовлення (10 тис. слів)	10	1 тис.
Конфесійний стиль (30 тис. слів)		
проповіді (30 тис. слів)	15	2 тис.

Виходячи із загальних корпусних кількісних параметрів та індивідуальних кількісних параметрів тестових фрагментів, тексти художньої літератури мають становити 35% від усього корпусу, тобто якщо НКУМ планово на 1 млн. слововживань, то обсяг художніх текстів становитиме 350 тис. слововживань, де, відповідно, проза – 250 тис. слововживань чи 35%, поезія – 50 тис. слововживань чи 5% і драматургія – 50 тис. слововживань чи 5%. Далі в межах прози, поезії та драматургії пропонуємо жанрову диференціацію текстів і, відповідно до визначених жанрів, забезпечення корпусу фактичним матеріалом: роман – 100 тис. слововживань чи 10%, повість – 100 тис. слововживань чи 10%, оповідання – 30 тис. слововживань чи 3%, новела – 20 тис. слововживань чи 2%; лірика – 10 тис. слововживань чи 1%, сонет – 10 тис. слововживань чи 1%, поема – 10 тис. слововживань чи 1%, балада – 10 тис. слововживань чи 1%, сатира – 10 тис. слововживань чи 1%; драма – 20 тис. слововживань чи 2%, комедія – 15 тис. слововживань чи 1,5% і трагедія – 15 тис. слововживань чи 1,5%.

Зважаючи на такі загальні корпусні кількісні параметри, індивідуальні кількісні параметри тестових фрагментів повинні становити для:

- прози − по 20 творів для роману і повісті, фрагменти яких повинні охоплювати по 5 тис. слововживань, і по 10 творів для оповідання й новели з фрагментами, відповідно, по 3 і 2 тис. слововживань;
- поезії – по 20 творів для лірики, сонету й сатири, з фрагментами по 500 слововживань, та по 10 творів для поеми і балади, з фрагментами по 1 тис. слововживань;
- драматургії – 10 творів з фрагментами на 2 тис. слововживань для драми, 8 творів на 2 тис. слововживань для комедії й 7 творів на 2 тис. слововживань для трагедії.

За аналогією до *Британського національного корпусу*, який розглядаємо як еталонний у сучасній корпусній лінгвістиці, вважаємо за доцільне, особливо на етапі впровадження напряму корпусної лінгвістики в українську мовознавчу традицію, визначити хронологічні межі НКУМ сучасною українською мовою. Тобто від останніх років XVIII ст. і до сьогодні, де окремо розглядати субперіод зламу XX / XXI ст., залишаючи поза увагою усі попередні періоди існування української мови, а саме: давньоукраїнський період від середини XI ст. до кінця XIV ст., ранньосередньоукраїнський від початку XV ст. до середини XVI ст., середньоукраїнський від середини XVI ст. до перших років XVIII ст., пізньосередньоукраїнський від середини і до кінця XVIII ст.

Виокремлення субперіоду зламу XX / XXI ст. мотивоване передусім екстралінгвальними чинниками, зокрема, утворенням Української держави і, відповідно, розширенням сфери функціонування української мови аж до входження в комп'ютерне середовище.

Отже, підкорпус художньої літератури у складі генерального корпусу НКУМ матиме обсяг 350 тис. слововживань, які репрезентуватимуть 155 текстів, і охоплюватиме художні твори сучасної української літературної мови від останніх років XVIII ст. і до сьогодні, де окремо слід розглядати субперіод зламу XX / XXI ст.

До укладання підкорпусу художньої літератури, як і до НКУМ загалом, пропонуємо застосувати принцип випадкової вибірки і

стосовно авторів, і стосовно творів. У результаті випадкової вибірки ми отримали такий список авторів, тексти яких становитимуть джерельну базу для підкорпусу художньої літератури:

1. Андрій Головко
2. Андрій Малишко
3. Андрій Чайковський
4. Архип Тесленко
5. Богдан Лепкий
6. Богдан-Ігор Антонич
7. Борис Антоненко-Давидович
8. Борис Грінченко
9. Борис Олійник
10. Валер'ян Підмогильний
11. Василь Барка
12. Василь Еллан Блакитний
13. Василь Земляк
14. Василь Пачовський
15. Василь Симоненко
16. Василь Стефаник
17. Василь Стус
18. Віктор Неборак
19. Віктор Петров
20. Володимир Винниченко
21. Володимир Діброва
22. Володимир Дрозд
23. Володимир Малик
24. Володимир Самійленко
25. Володимир Сосюра
26. Всеволод Нестайко
27. В'ячеслав Сахно
28. Гнат Хоткевич
29. Григорій Квітка-Основ'яненко
30. Григорій Чубай
31. Дмитро Білоус
32. Дмитро Павличко
33. Емма Андієвська
34. Євген Гуцало
35. Євген Маланюк
36. Іван Багряний
37. Іван Гнатюк
38. Іван Драч
39. Іван Карпенко-Карий
40. Іван Котляревський
41. Іван Кочерга
42. Іван Нечуй-Левицький
43. Іван Франко
44. Катерина Мотрич
45. Катря Гриневичева
46. Леонід Глібов
47. Леся Українка
48. Ліна Костенко
49. Максим Рильський
50. Марійка Підгірянка
51. Маркіян Шашкевич
52. Марко Вовчок
53. Марко Кропивницький
54. Микола Бажан
55. Микола Вінграновський
56. Микола Вороний
57. Микола Зеров
58. Микола Куліш
59. Микола Олійник
60. Микола Сингаївський
61. Микола Хвильовий
62. Михайло Драй-Хмара
63. Михайло Коцюбинський
64. Михайло Петренко
65. Михайло Старицький
66. Михайло Стельмах
67. Михайль Семенко

68. Наталена Королева
69. Наталка Білоцерківець
70. Оксана Забужко
71. Оксана Іваненко
72. Олег Ольжич
73. Олександр Довженко
74. Олександр Олесь
75. Олександр Підсуха
76. Олексій Коломієць
77. Олена Теліга
78. Олесь Бердник
79. Олесь Гончар
80. Олеся Садова
81. Ольга Кобилянська
82. Осип Назарук
83. Остап Вишня
84. Павло Глазовий
85. Павло Грабовський
86. Павло Загребельний
87. Павло Тичина
88. Панас Мирний
89. Пантелеймон Куліш
90. Петро Гулак-Артемовський
91. Платон Воронько
92. Роман Іваничук
93. Семен Скляренко
94. Сергій Плачинда
95. Сидір Воробкевич
96. Степан Васильченко
97. Степан Руданський
98. Тарас Шевченко
99. Тимофій Бордуляк
100. Тодось Осьмачка
101. Улас Самчук
102. Юрій Андрухович
103. Юрій Винничук
104. Юрій Клен
105. Юрій Липа
106. Юрій Мушкетик
107. Юрій Покальчук
108. Юрій Федькович
109. Яр Славутич

Оброблення текстів у НКУМ доцільно реалізовувати за допомогою засобів SGML [3] у форматі TEI [4]. Де SGML (Standard Generalized Markup Language) – це міжнародний стандарт на визначення незалежних від пристроїв уведення/виведення інформації, незалежних від обчислювального середовища методів подання текстів у електронній формі. А TEI (Text Encoding Initiative) – система кодування текстів, яка є міжнародним і міждисциплінарним стандартом подання усіх типів текстів, функціональних у бібліотечній, музейній, видавничій справах, лінгвістиці.

Схема кодування TEI використовує стандартну мову узагальненої розмітки (SGML) і спрямована на забезпечення обміну інформацією, що зберігається в електронній формі. Використання SGML і TEI – це шлях до універсального оброблення тексту, завдяки чому будь-яке програмне забезпечення загального призначення, яке має змогу працювати із SGML, може опрацьовувати TEI-сумісні тексти.

У процесі програмного оброблення текстів для формування підкорпусу художньої літератури НКУМ використаємо адаптований набір тегів (кодів), співвідносний зі стандартним базовим набором тегів системи кодування TEI Lite для прозових творів.

Так, обов'язковими елементами для всіх текстів, поданих у форматі TEI, є *електронний заголовок* і *власне текст*.

1. **Електронний заголовок.**

Цей елемент тексту TEI містить загальну інформацію про текст (ми обрали й уніфікували систему кодів, яка подає бібліографічні дані та параметри тексту). Заголовок уводиться за допомогою елемента <**headerText**> і має чотири основних компоненти з відповідними атрибутами та значеннями:

<author>		містить повне ім'я автора;
<title>		містить назву твору;
<soursDesk>		групує бібліографічний та жанрово-стильовий опис опрацьовуваного тексту, роблячи це за допомогою таких елементів:
	<edition>	подає особливості цієї редакції тексту;
	<vol>	містить інформацію про номер періодичного видання, серії, тому тощо; подається за допомогою атрибута *type*, який має такі значення: *number* (номер), *series* (серія), *volume* (том);
	<style>	подає інформацію про стиль художнього твору, що вводиться в корпус;
	<genre>	подає інформацію про жанр художнього твору, що вводиться в корпус;
	<extend>	подає інформацію про розмір електронного тексту; для зручності розмір указується в кількості слів (оскільки слово є одиницею виміру самого корпусу), за допомогою атрибута *type*[1], який має значення w – скороченої форми терміна *word* (слово);
<publicationStmt>		групує інформацію про

[1] Глобальний атрибут *type* конкретизує текст через присвоєння йому різних значень і активно використовується у випадках типологічного розмаїття

	публікації / видання / перевидання опрацьовуваного тексту;
<publisher>	містить назву видавництва;
<pubPlace>	подає назву місця, де розташоване видавництво;
<date>	містить дату виходу видання (дата може бути подана в будь-якому форматі);
<address>	містить адресу видавництва, представництва, електронної бібліотеки, сайту тощо, тобто місця, звідки взято текст. Передбачені такі значення для уточнення типу адреси: e-*mail* – для електронної, a-*post* – для поштової.

2. **Основна частина – власне текст.**

Систему теґів для розмітки основної частини умовно можна поділити на дві групи. Першу групу становлять коди для розділення тексту, виокремлення як основної частини тексту, так і її структурних елементів і субелементів, а саме частин (розділів, книг тощо), абзаців, речень, а також пропусків у тексті. Другу групу формує система теґів для виділення окремих елементів тексту.

Текстову частину вводять за допомогою елемента <**text**>. Розділення тексту на першому рівні супроводжується використанням таких елементів:

<front>	містить довільну вступну інформацію, яка розміщена перед основним текстом;
<group>	містить монотекст, множину текстів або груп текстів;
<body>	містить усю основну частину одного монотексту, крім того, що стосується вступної чи завершальної частин;
<back>	містить різні додатки і все, що розташоване після основного тексту.

Основна частина прозового тексту може бути згрупована в глави (розділи, підрозділи, книги тощо), або представлена у вигляді простого набору абзаців. У першому випадку використовуємо елемент <**div**>. Взагалі цей теґ застосовуємо у випадку виділення будь-якої чітко вираженої частини, наприклад, діалогу без слів автора або віршованого тексту, із типологічним визначенням такої частини за допомогою відповідних елементів. У другому випадку абзац відмічається теґом <**p**>. У межах основної частини всі речення виділяються елементом <**s**>. Пропуски тексту позначаються теґом <**gap**>.

Таким чином, для розділення тексту на другому рівні використовуються такі елементи:

** ** містить розділ або чітко виражену частину прозового тексту; обов'язковими для цього елемента є атрибути *type* і *n*; останній означає коротку (але зрозумілу) назву або номер розділу; атрибут *type* подає загальноприйняту назву для відповідної категорії частини або вказує на тип частини, яку необхідно виділити; в НКУМ для атрибута *type* використовуються такі значення:

- <chapter> розділ, глава, частина;
- <book> книга;
- <poem> вірш у прозовому творі; розмітка самого вірша відповідно до стандартів TEI з використанням елементів <lg> (виокремлює групу віршованих рядків, що становлять певну структурну одиницю) і <l> (виокремлює віршований рядок);
- <speach> пряма мова;

<p> відмічає абзаци прозового тексту;

<s> відмічає речення прозового тексту;

<gap> відмічає пропуск у тексті.

Наступний перелік теґів та їхніх значень стосується кодування елементів тексту.

<abbr>		скорочення будь-якого типу;
<address>		будь-яка адреса в тексті;
<cit>		цитата з будь-якого іншого документу;
<date>		дата у будь-якому форматі;
<num>		будь-яке число;
<lg>		група віршованих рядків, що становлять певну структурну одиницю;
<l>		віршований рядок;
<name>		будь-яка власна назва; потребує використання атрибута *type* з такими значеннями:
	<antrop>	антропоніми;
	<toponim>	топоніми;
	<place>	назви країв, країн, територій
	<astrolog>	назви астрономічних об’єктів;
	<teonim>	назви божеств;
	<zoonim>	клички тварин;
	<organisation>	назви організацій;
	<literat>	назви художніх творів;
	<print>	назви друкованих видань
<term>		термін; у разі вживання терміна іноземною мовою передбачено використання атрибута *lang* із зазначенням перших літер назви мови англійською мовою, наприклад: *engl* – англійська, *kirg* – киргизька, *chin* – китайська, *lat* – латина; *slav* – старослов’янська мова тощо;
<sp>		пряма мова;
<speaker>		мовець прямої мови;

<foreign>		запозичення; у разі вживання іноземною мовою також передбачено використання атрибута *lang* із зазначенням відповідної мови;
<hi>		відмічає слово чи фразу, які графічно відрізняються від загального тексту; потребує атрибут *rend*, якому, відповідно, надаються наступні значення:
	<italic>	курсив;
	<bold>	напівжирний шрифт;
	<undo>	підкреслений текст;
	<ll>	велика літера (*large letter*).

Приклад:

```
<header Text>

   <author> В'ячеслав Сахно </author>
   <title> Циркулятор </title>

   <sourceDesc>
         <edition> Сучасність </edition>
         <vol> 3 </vol>
         <style> художній </style>
         <genre> роман </genre>
         <extend type=w> 5000 </extend>
   </sourceDesc>

   <pablStmt>
         <pubplace> Київ </pubplace>
         <date> 2001 </date>
         <address>  </address>
   </pablStmt>

</header Text>

<text>
```

<s>Я зрідка бачу сни.</><s>Або просто забуваю.</><s>Тому я заздрю людям, які їх бачать і пам'ятають.</><s>Адже вони мають змогу жити у двох різних світах.</><s>Одне життя вдень - нудне й однобарвне, геть інше вночі - непередбачуване й різнобарвне.</><s>Щоправда, вранці настає неабияке розчарування – настирливо дзеленчить будильник, мусиш збиратися на осоружну працю.</><s>Гадаю, якби я міг добирати собі сни до вподоби, на власний смак, то, можливо, й змирився б із білим днем.</>

<s>Мені люба нічна пора.</><s>Тим-то я не люблю спати, а вставати вранці й поготів.</><s>Тобто я є, за прийнятими мірками, совою, або, радше, пугачем чи сичем.</><s>Слово чоловічого роду мені імпонує більше.</><s>Цілком імовірно, що людина насправді нічна істота, бо ж саме ніч вона обрала для мрій, любощів і збереження породи.</>

<gap>

<s>Власне, втрачати мені було нічого.</><s>По мені ніхто б і не заплакав!</><s>Але ж це справжнісіньке самогубство!</><s>Можливо, в мене ще будуть шанси, не такі ризиковані!..</>

<sp>Твоєму життю ніщо не загрожує!</sp> немов гіпнотизуючи мене, провадив <name type=antrop>Богдан</>.<sp>На твоєму місці я б не відмовлявся від такого шансу.</ <s>Він підвівся.</> <sp>Думай, старий! Думай! Остаточну відповідь <gap></sp>

</text>

Література

1. Демська-Кульчицька О. М. Основи Національного корпусу української мови. – К., 2005.
2. Передрієнко В. А. Трьох стилів теорія // Енциклопедія. Українська мова. – К.: Видавництво Українська енциклопедія ім. М. П. Бажана, 2000. – С. 640–641.
3. Standard Generalized Markup Language ISO 8879: Information processing – Text and office systems. – Geneva, 1986.
4. Guidelines for Electronic Text Encoding and Interchange / C. M. Sperberg-McQueen, L. Burnard. – 2001. – http://www.hcu.ox.ac.uk/TEI/P4X/index.html.

Тетяна Цимбалюк-Скопненко, к. ф. н.[*]
Інститут української мови НАН України (Київ)
УДК 161.2.81’374

УКЛИНЮВАННЯ ЯК ОДИН ІЗ ВИДІВ ЛЕКСИКО-СЕМАНТИЧНИХ ПЕРЕТВОРЕНЬ ФРАЗЕОЛОГІЗМІВ У МОВІ ПЕРЕКЛАДУ МИКОЛИ ЛУКАША

У статті проаналізовано специфіку вклинювання як лексико-семантичного перетворення у творчій спадщині видатного українського перекладача М. Лукаша, установлено особливості інкорпорації елементів різних частин мови до структури традиційно вживаної фразеологічної одиниці в перекладних поетичних і прозових текстах.

У другій половині ХХ ст. інтенсивність інновацій у фразеологічних системах сучасних слов’янських мов помітно зросла. Ці динамічні процеси мають багаторівневий характер, оскільки культурно-політичні події згаданого періоду справили надзвичайний вплив на всі мови європейського ареалу. Лексико-семантичні перетворення у фразеологічній системі літературної мови передусім полягають у тому, що оказіональні елементи доповнюють або змінюють компонентний склад традиційно вживаної одиниці. Використання вклинювання[1] як способу перетворення фразеологічної одиниці (далі – ФО) свідчить „про

[1] Термін запровадив О. Кунін [2, с. 13].