ДЖЕРЕЛА ІЛЮСТРАТИВНОГО МАТЕРІАЛУ

10. Barnes HWC — Barnes J. A History of the World in 10½ Chapters / J. Barnes. — London : Jonathan Cape, 1989. — 314 p.

В статье выделены различные виды репрезентации рассказчика в постмодернистском романе Дж. Барнса «История мира в 10 ½ главах». Выявлена специфика функционирования лингвостилистических приемов актуализации нарративной маски как своеобразного способа повествования и воплощения постмодернистского принципа текстообразования. Определены различные виды нарративных масок и их языковой код в произведении.

Ключевые слова: *нарратор, нарративная маска, дискурс, коммуникация, нарративная дистанция.*

This article discusses various kinds of narrators in postmodern novel "A History of the World in 10½ Chapters" by J. Barnes and suggests the notion of "narrative mask" as a specific way of narration. It is assumed that different kinds of narrative masks provide the implementation of postmodern principles of text composition. The article aims at explicating expressive means and stylistic devices of narrative mask representation.

Key words: *narrator, narrative mask, discourse, communication, narrative distance.*

УДК 81'1:811.111

OXFORD TEXT ARCHIVE: ДОСВІД ОБРОБКИ КОРПУСНИХ ДАНИХ

Юзькова І.В.,
Київський університет імені Бориса Грінченка

У статті розглянуто новітні шляхи та інструменти для обробки корпусних даних. Особлива увага надається визначенню понять «корпус» та «лінгвістичний корпус». Описані основні характеристики та етапи роботи з електронним Oxford Text Archive та комп'ютерною програмою WordSmith.

Ключові слова: *корпус, лінгвістичний корпус, ОТА, конкорданс.*

Новітніми інструментами для дослідження лінгвістичних даних і прикладних задач стали корпуси текстів. Сьогодні корпусна лінгвістика, як і будь-який новий науковий напрям «не тільки відкрила невідомі раніше перспективи досліджень, а й започаткувала створення певних правил і закономірностей роботи з матеріалом, а також, що теж неминуче для революційного напряму, спричинила необхідність розв'язання цілої низки нових проблем, досі невирішених лінгвістам» [2]. Завдяки стрімкому розвитку в науковому лінгвістичному словнику з'явилися дуже близькі поняття: «електронні бібліотеки», «масив текстів», «колекція текстів», «електронний архів», «повна текстова база даних», які стали предметами наукових пошуків як вітчизняних, так і зарубіжних дослідників. Саме тому **актуальним** вбачається розгляд найбільшого та широковживаного електронного архіву *Oxford Text Archive* (далі — ОТА).

Метою статті є висвітлення характерних ознак Оксфордського текстового архіву і шляхів обробки наявних в ньому даних. **Матеріал дослідження** взятий з джерела фактичного матеріалу ОТА.

Для того щоб зрозуміти, з чого складається окремий текстовий архів у нашому випадку ОТА, необхідно розглянути тексти, безпосередньо з яких і складаються корпуси текстів, а з них — архіви. Під *корпусом текстів* В.Н. Шевчук розуміє величезний масив текстів (як письмових, так і усних) природної мови, представлених в комп'ютерному вигляді, тобто на машинному носії, і належним чином упорядкованих з метою їх використання в наукових і практичних дослідженнях [5]. *Лінгвістичний корпус* — це масив текстів, зібраних в єдину систему, сформовану за певними ознаками (мовою, жанром, часом створенням, автором тощо) і забезпечених пошуковою системою. Він може містити як письмові тексти (газет, журналів, літературних творів), так і транскрипти радіо- і телепередач. Організація корпусу може бути найрізноманітнішою. Залежно від цілей його створення в корпус можуть входити тексти конкретною мовою, одного або кількох авторів і літературних жанрів, написані в певний історичний період і т.д. Весь масив текстів систематизований. Це означає, що в корпусі зафіксоване розташування кожного слова в реченні щодо інших слів, а також враховується частота його використання у цьому корпусі [4].

Першим досить великим корпусом, тексти якого зберігалися на машинному носії, був Браунівський (Brown corpus 1960 — for American English). Його розробники У. Френсіс та Г. Кучера розглядали поняття корпусу як сукупність текстів, яка вважається репрезентативною для певної мови чи діалекту, що призначена для лінгвістичного аналізу [6]. Браунівський корпус швидко перетворився в популярний об'єкт дослідження та навіть у певний стандарт для створення інших корпусів. Аналогічними були структури побудови наступних корпусів: Lancaster-Oslo/Bergen Corpus (LOB — 1978 р.), London-Lund Corpus (LLC — 1987 р.), The Freiburg-Brown corpus of American English (Frown — 1992 р.). Найвідоміший на сьогодні British National Corpus (BNC) було створено в 1990-ті рр. [1, 3]. Проте мало хто знає, що Oxford Text Archive (OTA) був започаткований набагато раніше і став основою для створення у 2006 р. The Oxford English Corpus (OEC).

OTA — це архів електронних текстів та інших літературних і мовних ресурсів, які були створені, зібрані й розподілені з метою дослідження літературних та лінгвістичних питань в університеті Оксфорд.

OTA був створений Луї Бюрнардом у 1976 р. спочатку як Оксфордський архів електронної літератури. Він вважається одним із перших архівів цифрових навчальних текстових ресурсів для збору і поширення матеріалів з усіх доступних наукових центрів. OTA продовжує співпрацю з Оксфордським університетом обслуговування з використанням електронно-обчислювальних машин OUCS, який володіє відповідними науково-дослідними проектами, що здійснюються в Оксфордському електронному науково-дослідному центрі на факультеті лінгвістики, філології та фонетики Оксфордського університету [7]. OTA також керує розподілом британського національного корпусу (BNC).

Структура архіву містить 8 розділів: 1) власне архів (OEC); 2) рекомендації щодо завантаження ресурсів; 3) новини; 4) Оксфорд (для обмеженого кола користувачів); 5) електронні мовні ресурси; 6) проекти; 7) поради та 8) проблемні питання.

Особливо значущим надбанням в OTA є власне архів (OEC), в якому зосереджений найбільший у своєму роді текстовий корпус англійської мови, що містить понад два мільярди слів. У свою чергу, власне архів має 3 підрозділи: TEI texts, Corpora та Legacy formats

TEI texts (The Text Encoding Initiative texts). Ініціатива кодування тексту — консорціум, що згалом розвиває і підтримує стандарт для представлення текстів у цифрову форму. Його головним результатом є набір керівних принципів, які визначають методи кодування для машинописних текстів в основному в гуманітарних, соціальних науках і лінгвістиці [9]. Тексти в у цьому підрозділі доступні в різних форматах для читання, завантаження або посилання.

Corpora. Колекції мовних даних, що містять тексти з різних джерел, як правило, складені для цілей лінгвістичного дослідження. Підрозділ являє собою своєрідний каталог власних накопичених корпусних даних та корпуси інших університетів із відкритим (безкоштовним) чи обмеженим доступом до інформації.

Legacy formats (застарілі файли). З'явився на світ у 1976 р. Деякі з ресурсів обмеженого доступу, багато файлів мають формат, який досить важко розшифрувати та використати, хоча більшість з них звичайний текст. Оскільки OTA не в змозі запропонувати підтримку роботи з таким типом файлів, тому й був створений такий підрозділ.

Через OTA проходить безліч наукових документів, розмічених відповідно до останніх вимог кодування матеріалів та текстів, а це обов'язкове зазначення таких даних, як:

- назва документа;
- автор документа (якщо відомо);
- стать автора (якщо відомо);
- тип мови (наприклад, британська чи американська англійська);
- жанр джерела;
- рік (дата, якщо відомо);
- дата збору матеріалу архівом;
- доступ до матеріалів;
- статистика документа (кількість символів, слів тощо) [7].

Цифрова версія Оксфордського англійського корпусу форматується в XML і зазвичай аналізується за допомогою спеціального програмного забезпечення.

Для досліджень у сфері корпусної лінгвістики, де фігурують великі за обсягом вибірки текстів, необхідне використання декількох типів програмного забезпечення: комерційні комп'ютерні програми (*LEXA, MonoConc, MicroConcord, TACT, WordSmith, WordCruncher, Manatee (Bonito), IMS Corpus Workbench (CQP), XAIRA, Visual Corpus Manager (VCM), EXMARaLDA, Corpus-Manager(Co-Ma)*), а також програми, розроблені для специфічних процедур аналізу, наприклад для граматичних моделей [3, 92].

Основні процедури, які доступні дослідникові при здійсненні аналізу корпусу текстів, містять:

- пошук заданого слова, словосполучення в корпусі;
- висновок результатів пошуку з урахуванням оточення в окремому полі;
- підрахунок кількості прикладів вживання слова в корпусі;
- сортування результатів пошуку за необхідними параметрами.

Всі дані процедури швидко і точно виконуються за допомогою комп'ютерної програми складання *конкордансу* (пошуку відповідностей).

У нашому випадку для обробки корпусних даних використовувалась програма *WordSmith 6.0*. Перед початком роботи з програмою на головній сторінці контролера інструментів користувачу пропонуються три кнопки основних інструментів і декілька клавіш для уточнення налаштувань. Клавіша "Concord" укладає конкорданси, "KeyWords" знаходить у текстах ключові слова, а "WordList" створює списки слів у тексті або колекції текстів. Для вдалого використання програмного забезпечення спочатку необхідно завантажити чи створити власний корпус текстів *(рис. 1)*. Особливої уваги потребує оформлення документа для подальшого завантаження файлу в програму.

<header>
<title> THE DUKE OF YORK TO PRINCE HENRY </title>
<year> 1610 </year>
<addresser> CHARLES </addresser>
<addressee> PRINCE HENRY </addressee>
</header>
Good brother, I hope you are in good health and merry, as I am, God be thanked.
In your absence I visit sometimes your stable, and ride your great horses, that at your return I may wait on you in that noble exercise. So committing you to God, I rest
Your loving and dutiful brother York.
To my brother the Prince. [8]

Рис. 1. Приклад правильного оформлення тексту для роботи з *WordSmith Tools*

Після впорядкування та оформлення матеріалу за допомогою *WordSmith* можна створювати список частотності вживання слів у одному тексті чи в корпусі текстів. За допомогою порівняння власного корпусу даних із BNC отримуємо список ключових слів. Поряд із кожним ключовим словом розміщені різні цифри, які містять інформацію про те, як часто вживається кожне слово у вихідному тексті (текстах) і наскільки ця частотність відрізняється від частотності його вживання у референтному корпусі *(рис. 2)*.

KeyWords

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	Keyness	P	L	S
1	YOUR	17	4,52	5	134 393	0,14	87,10	0,000		
2	BROTHER	9	2,39	4	7 568		85,79	0,000		
3	LOVING	5	1,33	4	1 409		58,50	0,000		
4	I	21	5,59	5	732 523	0,74	49,54	0,000		
5	ESTEEMING	2	0,53	1	2		44,41	0,000		
6	KINDE	2	0,53	2	3		43,22	0,000		
7	PRINCE	4	1,06	3	5 344		34,39	0,000		
8	YOU	14	3,72	5	588 503	0,59	28,33	0,000		

KWs plot links clusters filenames source text notes

8 entries Row 1 YOUR

Рис. 2. Список ключових слів, укладених на основі PCEEC (17 ст.) порівняно з BNC

Найбільш репрезентативний та швидкий спосіб обробки інформації за допомогою *WordSmith* — це укладання конкордансу. Конкорданс — це список усіх уживань слова, перед і після якого є слово-розділювач, такий як знак пунктуації, пробіл тощо.

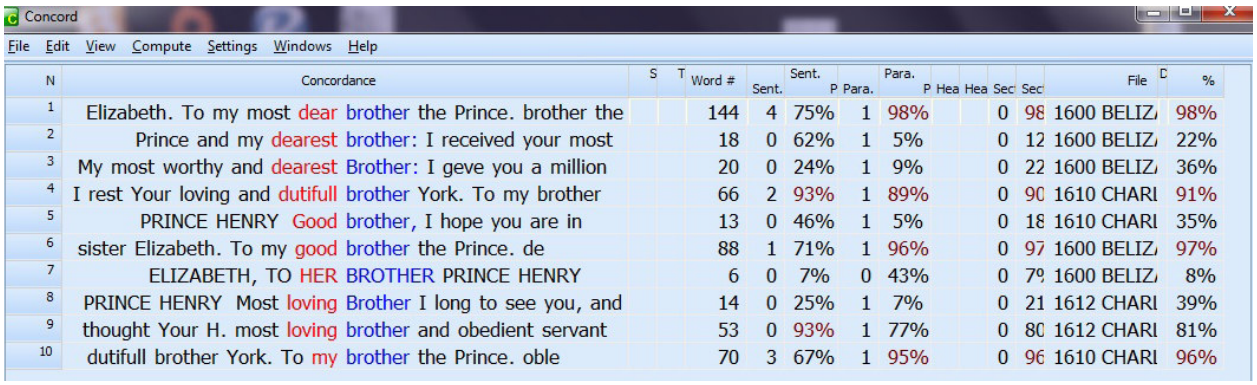

Рис. 3. Список усіх уживань слова “brother” з його лівим і правим оточенням

Для наочності зазначено в програмі пошукове слово *brother* для вибраних текстів з *Parsed Corpus of Early English Correspondence* (PCEEC) *(рис. 3)*. Отримані результати засвідчують, що вибране слово має найбільшу сполучуваність зі словами *dearest*, *good*, *loving* ліворуч від центрового слова; із займенником *I* та сполучником *and* — праворуч.

Отже, можна дійти висновку про те, що слово *brother* є не тільки ключовим для створеного нами корпусу на основі OTA, але й найбільш сполучуваним з його лівим і правим оточенням.

ЛІТЕРАТУРА

1. Ванівська О.І. Основні підходи до аналізу мовних даних у корпусній лінгвістиці / О.І. Ванівська // Наукові записки. — Острог : Вид-во Національного університету «Острозька академія», 2012. — Вип. 27. — 368 с. — (Серія «Філологічна»).
2. Голубкова Е.Е. Вестник Московского государственного лингвистического университета. Языкознание / Е.Е. Голубкова. — М. : МГЛУ, 2009. — Вып. 572. — С. 30.
3. Жуковська В.В. Вступ до корпусної лінгвістики : навч. посіб. / В.В. Жуковська. — Житомир : Вид-во ЖДУ ім. І. Франка, 2013. — 140 с.
4. Сысоев П.В. Иностранные языки в школе / П.В. Сысоев. — М. : ООО «Методическая мозаика», 2010.— Вып. 4. — С. 12.
5. Шевчук В.Н. Электронные ресурсы переводчика : справочные материалы для начинающего переводчика. — М. : Либрайт, 2010. — С. 44.
6. Halliday M.A.K. Lexis as a linguistic level / C.E. Bazell, J.C. Catford, M.A.K. Halliday & R.H. Robins (Eds.) // In Memory of J.R. Firth. — London : Longman, 1966. — P. 148–162.
7. [Електронний ресурс]. — Режим доступу : http://en.wikipedia.org/wiki/Oxford_English_Corpus-cite_ref-oec_2-0MacEnery T. and Wilson A. Corpus Linguistics. — Edinburgh : University Press, 1996. — P. 23.

ДЖЕРЕЛА ІЛЮСТРАТИВНОГО МАТЕРІАЛУ

8. Parsed Corpus of Early English Correspondence (PCEEC) [Електронний ресурс]. — Режим доступу : http://ota.ahds.ac.uk/desc/2510
9. Text Encoding Initiative [Електронний ресурс]. — Режим доступу : http://www.tei-c.org/index.xml

В статье рассмотрены новейшие пути и инструменты для обработки корпусных данных. Особое внимание уделяется определению понятий «корпус» и «лингвистический корпус». Описаны основные характеристики и этапы работы с электронным Oxford Text Archive и компьютерной программой WordSmith.

Ключевые слова: *корпус, лингвистический корпус, OTA, конкорданс.*

The article deals with the latest methods and instruments for the corpus data analysis. Particular attention is paid to definition of the notions “corpus” and “linguistic corpora”. The basic characteristics and stages of work with electronic Oxford Text Archive and computer program WordSmith are described.

Key words: *corpus, linguistic corpora, OTA, concordance.*