УДК 811

Л. М. Коцюк, О. О. Очеретович,
Національний університет «Острозька академія», м. Острог

ПАРАМЕТРАЗИЦІЙНІ ХАРАКТЕРИСТИКИ КОРПУСУ ТЕКСТІВ БІАТЛОНУ СУЧАСНОЇ АНГЛІЙСЬКОЇ МОВИ

У статті йдеться про особливості параметризації електронного корпусу текстів біатлонної тематики та можливості його практичного використання. Автори дають характеристику основним напрацюванням корпусної лінгвістики у створенні спеціалізованих корпусів текстів. Виводиться набір параметрів для створеного корпусу текстів біатлону: широта, глибина, пропорційність, часовий вимір та об'єм. Стаття є апробацією студентського наукового дослідження, проведеного у науково-практичній лабораторії LEXILAB, що у Національному університеті «Острозька академія», і є частиною проекту створення корпусу текстів наукового мовлення.

***Ключові слова**: корпусна лінгвістика, корпус текстів, параметризація, біатлон.*

PARAMETRIZATIONAL CHARACTERISTICS OF THE BIATHLON TEXT CORPUS OF THE ENGLISH LANGUAGE

The article deals with the parameterization characteristics of electronic text corpora on biathlon and the possibilities of its practical use. The lack of specialized dictionaries or glossaries of sports in general, biathlon in particular, caused the interest in creating the terminology survey in this sphere. One of the modern trends in the sphere of term studies is corpus-based. The authors summarize the basic developments of corpus linguistics in the creation of specialized text corpora. Having analyzed some types of specialized corpora, the coprus of biathlon texts was created. A set of parameters for the prepared corpus is as follows: width, depth, proportionality, time dimension, and volume. The article serves the approbation of the student research conducted in the scientific laboratory LEXILAB, which is functioning at the National University of «Ostroh Academy», and is the part of the project aimed at creating the text corpora of scientific language.

***Key words**: corpus linguistics, text corpus, parametrization, biathlon.*

ПАРАМЕТРИЗИЦИОННЫЕ ХАРАКТЕРИСТИКИ КОРПУСА ТЕКСТОВ БИАТЛОНА СОВРЕМЕННОГО АНГЛИЙСКОГО ЯЗЫКА

В статье говорится об особенностях параметризации электронного корпуса текстов биатлонной тематики и возможности его практического использования. Авторы дают характеристику основным наработкам корпусной лингвистики в создании специализированных корпусов текстов. Выводится набор параметров для созданного корпуса текстов биатлона: широта, глубина, пропорциональность, временное измерение и объем. Статья является апробацией студенческого научного исследования, проведенного в научно-практической лаборатории LEXILAB в Национальном университете «Острожская академия» и является частью проекта создания корпуса текстов научной речи.

***Ключевые слова:** корпусная лингвистика, корпус текстов, параметризация, биатлон.*

Сьогодні спостерігається тенденція посилення інтересу до сфери спорту в цілому та до біатлону зокрема, що передусім пов'язано із регулярним проведенням чемпіонатів та кубків світу з цього виду спорту, зимових Олімпійських ігор, інших змагань. Для ефективнішого розв'язання лінгвістичних завдань науковці все частіше звертаються до сучасних технологій, що в свою чергу докорінно змінює технологію дослідницької праці та відкриває перед лінгвістами нові можливості у вивченні функціонування мови, як всеохоплюючої системи людської взаємодії. Оскільки на сучасному етапі розвитку суспільства робота з комп'ютером становить один з різновидів комунікації, то лінгвістам відведено особливе місце в створенні нових засобів опрацювання інформації та її систематизації.

Недостатній рівень дослідження електронних корпусів текстів, упереджене ставлення багатьох мовознавців до можливості їх практичного використання у різних галузях та науках, обмеження сфери застосування корпусів текстів при розв'язанні окремих прикладних завдань, а також відсутність базових навичок користування корпусами у переважної частини лінгвістів зумовлюють **актуальність** та доцільність наукового дослідження на цю тему.

Метою дослідження є ознайомлення з основними напрацюваннями корпусної лінгвістики, вивчення характерних ознак електронних корпусів, а також створення власного спеціалізованого корпусу текстів біатлону. **Об'єкт дослідження**: письмові англомовні тексти біатлонної тематики як особлива складова терміносистеми спорту. **Предмет дослідження**: корпусні технології представлення мовної системи.

Лексика біатлону становить значний пласт спортивної термінології у зв'язку з популяризацією цього виду спорту. Терміносистема спорту є однією з найбільш активних і мобільних у своєму формуванні, розвитку й поширенні у професійних та позапрофесійних сферах комунікації. Кожен вид спорту має власну систему термінологічних найменувань, професійну лексику й жаргон, що й зумовлює необхідність і актуальність їх всебічного вивчення, лінгвістичного аналізу та лексикографічного кодифікування термінології спорту. Дослідженню спортивної термінології присвятили свої наукові розвідки такі українські вчені: І. Янків, В. Осінчук, Л. Бардіна, Н. Назаренко, О. Боровська, Н. Юрко, О. Матвіяс та ін. Серед іноземних науковців термінологію спорту досліджували А. Рилов, З. Буляж, Є. Птушкіна, Ф. Суслов, Д. Тишлер, І. Юрковський, Р. Попов, У. Свінкс, А. Рум та ін. Однак, незначна кількість словників та глосаріїв, що спеціалізуються на термінології біатлону засвідчує недостатній рівень опрацювання цієї терміносистеми.

Наприкінці минулого століття в мовознавчих дослідних методиках сформувався новий підхід до відбору та організації мовного матеріалу для подальшого його вивчення, опрацювання, опису та аналізу. Цей підхід отримав назву корпусний, що походить від назви об'єкту, в який організовано фактичний матеріал – корпус текстів. Поняття «корпус» у лінгвістиці означає «електронний вигляд системно організованої та програмно обробленої вибірки текстів, що представляють всі історичні й географічні варіанти та форми існування довільної мови». У науці використовується для різноманітних мовознавчих досліджень і програмних застосувань [1]. На думку лінгвіста Захарова В., корпус текстів – це значний за обсягом, представлений в електронному вигляді, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, створений для вирішення конкретних лінгвістичних завдань [3, с. 3].

Завдяки комп'ютерним технологіям відкрилися нові технічні можливості для обробки, збереження і відбору лінгвістичних даних. Це значно сприяло виникненню та розвитку нової галузі мовознавчих досліджень: корпусної лінгвістики. Актуальність та важливість цієї науки постійно зростає у зв'язку з величезною кількістю інформаційних потоків і проблемою автоматичного опрацювання лінгвістичних даних. Як стверджує білоруський мовознавець Ричкова Л., матеріал корпусу дозволяє не лише оптимізувати і об'єктивізувати лінгвістичні дослідження, але і по-новому окреслити багато

традиційних лінгвістичних понять [6, с. 185]. Таким чином, корпусні дослідження, ключовим елементом яких є реальний мовний матеріал, а не мовна інтуїція, зводять до мінімуму суб'єктивізм дослідника і допомагають здійснювати максимально об'єктивний аналіз мовного матеріалу.

Корпусна лінгвістика як галузь прикладного мовознавства займається визначенням загальних принципів побудови, обробки та експлуатації даних лінгвістичних корпусів (корпусів текстів) із використанням сучасних комп'ютерних технологій, розробленням методики збору реальних мовних явищ – писемних та усних текстів, а також способів їх збереження та аналізу [3, с. 3]. Варто зауважити, що етапі становлення корпусної лінгвістики корпуси укладалися на папері. Такі перші доелектронні корпуси були конкордансами, тобто алфавітними списками всіх ужитих у певному тексті слів разом із контекстуальним оточенням. Укладання цих паперових корпусів-конкордансів займало багато часу і зусиль та вимагало напруженого аналізу, який здійснювався вручну. Тим не менше, доелектронні корпуси, як початковий етап розвитку сучасних корпусів, відіграли значну роль в таких лінгвістичних проектах, як укладання конкордансів Біблії й літературних творів, а також написання граматик і словників [8, с. 1]. Завдяки активному розвитку комп'ютерних технологій у США, саме тут було створено перший комп'ютеризований корпус, що став прикладом для наслідування для багатьох лінгвістів по всьому світу. У Браунівському університеті Нельсон Френсіс та Генрі Кучера розпочали укладання одномільйонного корпусу, який було названо за місцем його створення Браунівським корпусом (the Brown Corpus) [2, с. 37–40].

Що ж включає в себе поняття «корпус текстів»? «Корпус – це певне зібрання текстів, в основі яких лежить логічний задум, логічна ідея, що об'єднує ці тексти. Логічна ідея втілюється в правилах організації текстів в корпус, алгоритмі і програмі аналізу корпусу текстів та в пов'язаних з цим ідеологією та методологією. Корпус є четвертою фактурою мовлення (тексти на машинному носії)» [5]. Лаконічнішу дефініцію знаходимо у Меєра Ч.: «Корпус – це організована певним чином словесна єдність, елементами якої є цілі тексти чи спеціальним чином відібрані уривки з текстів, що доступні для лінгвістичного аналізу» [9, с. 36]. МакЕнері Т. та Вілсон Е. констатують, що корпус, як правило, складається з вибірок, що «максимально репрезентують досліджувану область/сферу» [7, с. 24]. Отже, можна стверджувати, що корпус – це зменшена модель мови чи підмови, створена з метою вирішення конкретних лінгвістичних завдань.

Зараз існує чимало класифікацій текстових корпусів, а оскільки у нашому дослідженні ми займалися розробкою саме спеціалізованого корпусу, то цей різновид корпусів заслуговує особливої уваги. Спеціалізований корпус – це жанрово чи галузево специфічний корпус, що має за мету відобразити певну підмову. Такі текстові корпуси створюються для вирішення конкретних лінгвістичних задач і протиставляються національним корпусам. Сьогодні спостерігається підвищений інтерес до створення та використання спеціалізованих корпусів в освітній та професійних сферах. Наприклад, the Corpus of Professional Spoken American English (CPSA) складається з транскриптів комунікативних ситуацій з академічної та політичної професійних галузей. The Michigan Corpus of Academic Spoken English (MICASE) містить біля 1,7 млн. словоживань (близько 200 годин записів) сучасного усного університетського мовлення, що було записано в Мічиганському університеті [2, с. 66-69].

Спеціалізовані корпуси мають особливий тип – це так звані корпуси учнівського мовлення або учнівські корпуси (*learner corpora*), які укладаються з усних і/або писемних текстів, спродукованих особами, що вивчають мову як іноземну. Тут англійський термін *learner* перекладається лексемою «учнівський», похідною від іменника учень зі значенням «той, хто вчиться, вивчає щось», тобто під цим прикметником слід розуміти людину, яка навчається, незалежно від віку. Такі корпуси почали створюватися ще в кінці 80-х на початку 90-х років XX століття.

Після з'ясування значення терміна «корпус» виникає питання: чим же характеризується корпусно-базований підхід до вивчення мови? Для корпусно-базованих досліджень властиві високий рівень точності та надійності зберігання великих обсягів мовної інформації. Ричкова Л. вирізняє таку низку властивостей корпусного аналізу:

1) емпіричний підхід до аналізу мовних даних (досліджуються реальні моделі мовної реалізації у природних текстах);
2) використання великих за обсягом, структурованих колекцій природних текстів (корпусів) як основи для аналізу;
3) широке залучення комп'ютерних технологій для дослідження лінгвального матеріалу;
4) застосування кваліативних і квантитативних аналітичних методик [6, с.185].

Як же відбувалося корпусно-базоване дослідження терміносистеми біатлону сучасної англійської мови? Для здійснення загального аналізу лексики біатлону, нами було відібрано тексти загальним обсягом близько 10 тисяч лексичних одиниць. Такий невеликий у порівнянні з іншими корпусами обсяг дає можливість зробити певні висновки на основі наявної інформації, а також передбачає можливість подальшого глибокого аналізу. У ході дослідження виникла необхідність параметризації корпусу. Під цим терміном розуміємо його побудову за визначеним набором параметрів, таких як наприклад: 1) широта (охоплення максимальної кількості типів текстів першого рівня); 2) глибина (виокремлення підкатегорій на нижчих рівнях); 3) пропорційність (заповнення окремих «клітинок» текстами, дібраними в певній пропорції); 4) часовий вимір [4, с. 76].

Підбираючи тексти для корпусу, ми послуговувалися досить поширеним різновидом методу випадкової вибірки – стратифікованою випадковою вибіркою, щоб більш-менш рівномірно охопити різноманітні за стилем та жанром пласти інформації. Більше того, підбір текстів необхідної тематики здійснювався з метою об'єктивної і всебічної оцінки доступних лінгвістичних даних Усі джерела є англомовними, за винятком англо-російського глосарію спортивної термінології зимових видів спорту. До прикладу, до вибірки увійшли тексти, які належать до наукової, довідкової, навчальної літератури та публіцистики. За жанрами текстів можна умовно виділити наукові статті, глосарій, енциклопедію, науково-популярні статті та підручник. Ці дані подані нижче (див. Табл. 1). Перша група під назвою *наукова література* становить 24.2% (2501 слово) від усієї сукупності матеріалу, друга – *довідкова література* вміщує найбільшу кількість термінів, а саме – 31.3% (3231 слово). Третя група (*публіцистика*) становить 30% (3093 слова), і остання група цієї класифікації має назву *навчальна література* і містить 14.4% (1485 слів) інформації. Саме жанрове різноманіття, на противагу однобокій систематизації, забезпечує репрезентативність добору текстів, що є однією з найважливіших ознак корпусів.

Таблиця 1

Класифікація текстів за типом та жанром

Тип тексту	Жанр тексту	Назва тексту	Кількість слів
Наукова література	Наукова стаття	SCIENCE_ARTICLE_1	1490
		SCIENCE_ARTICLE_2	510
		SCIENCE_ARTICLE_3	501

Довідкова література	Глосарій	REFERENCE_GLOSSARY_1	1882
	Енциклопедія	REFERENCE_ENCYCLOPAEDIA_1	1349
Публіцистика	Науково-популярна стаття	PUBLICISTICS_ARTICLE_1	1661
		PUBLICISTICS_ARTICLE_2	1432
Навчальна література	Підручник	EDUCATION_HANDBOOK_1	1485

Аналізуючи вищезазначені дані, варто вказати, що їхнім головним критерієм є репрезентативність і збалансованість текстів. Тобто, для кожної підгрупи підбиралися приблизно однакові за обсягом тексти, а також уривки з рівномірною кількістю слів. Після цього, назви текстів шифрувалися відповідно до їхнього типу. Наприклад, SCIENCE_ARTICLE_1 (наукова стаття) для наукової літератури, REFERENCE_ENCYCLOPAEDIA_1 (енциклопедія) для довідкової літератури, EDUCATION_HANDBOOK_1 (підручник) для навчальної літератури і т.п.

Наступним кроком стала розмітка або анотування самих текстів. Задля цього використовувалися теги, які у системах обробки інформації прийнято розуміти як ознаки даних. Під час розмітки текстів були використані такі теги: *<назва>, <автор>, <підзаголовок>, <текст>, <абзац>, <визначення> (дефініцію), <ключове слово> і <переклад>* (*<title>, <author>, <header>, <text>, <p>, <definition>, <headword>* і *<translation>* відповідно). Анотування текстів здійснювалося у документі Microsoft Word, де з допомогою вищезгаданих тегів можна з легкістю виокремити потрібні дані. Після завершення цієї операції, розмічений текст було перенесено у текстовий документа (формат txt) і завантажено у навчальну систему «Moodle» Національного університету «Острозька академія». Як бачимо, параметризація текстів біатлону є необхідною умовою створення електронного корпусу цих текстів, адже у такий спосіб мовний матеріал набуває систематизованості, що завдяки можливості подальшого аналізу та опрацювання надає йому практичну цінність. Корпуси текстів в електронному вигляді – інноваційна можливість зробити навчальний процес більш наочним, а також суттєво полегшити наукові дослідження мовного матеріалу. Комп'ютерні технології породили відносно нове поняття «корпусного методу», що базується на здобутках однієї з галузей прикладної лінгвістики – корпусної лінгвістики.

Корпус текстів з біатлону може стати в нагоді не лише людям, зацікавленими у цій сфері, а й дослідникам мовних явищ, адже будь-який корпус текстів (параметризований відповідно до певних критеріїв) є прикладом ефективних напрацювань у галузі корпусної та прикладної лінгвістики. Такий корпус може лягти в основу загального корпусу спортивної тематики, як його структурний підрозділ або стати компонентом цілісного спеціалізованого корпусу текстів мови. Оскільки терміносистема біатлону на сучасному етапі є малодослідженою, цей корпус текстів стане надійним джерелом термінолексики цієї сфери. Завдяки цьому значно полегшиться процес написання підручників, довідників, енциклопедій, статей, подальших досліджень термінів біатлонного спорту. До того ж, створення подібних корпусів відкриває перспективу створення багатомовних корпусів, де тексти подаватимуться паралельно різними мовами, або створення поліваріантних корпусів, до яких входить кілька різних перекладів тексту на одну й ту ж саму мову.

Отже, це дослідження дозволяє зробити висновок, що такі невід'ємні складові корпусу текстів як репрезентативність і збалансованість досягаються виключно методом зведення текстів до певних критеріїв, тобто параметризацією корпусу текстів. Тим не менше, можна з упевненістю сказати, що на практиці неможливо досягнути як ідеальної репрезентативності, так і абсолютної збалансованості. Корпусна лінгвістика – наука, за якою стоїть майбутнє мови, тому рівень дослідження електронних корпусів текстів потребує подальшої наукової розробки і детальнішої систематизації задля можливості їх практичного використання у різних галузях та науках.

Література:

1. Демська-Кульчицька О. М. Що таке корпусно-базовані дослідження мови / О. М. Демська-Кульчицька. – Українська мова та література. – № 40, 2004. – С. 21–22.

2. Жуковська В. В. Вступ до корпусної лінгвістики: навчальний посібник / В. В. Жуковська. – Житомир : Вид-во ЖДУ ім. І. Франка, 2013. – 142 с.

3. Захаров В. П. Корпусная лингвистика : Учебно-метод. пособие / В. П. Захаров – СПб., 2005. – 48 с.

4. Карпіловська Є. А. Вступ до комп'ютерної лінгвістики / Є. А. Карпіловська. – Донецьк, 2003.

5. Рыков В. В. Прагматически ориентированный корпус текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог- 99». – М., 1999.

6. Рычкова Л. В. Праблема састауных аб'ектау у корпусах славянскімоу і лінгвістычных базах дадзеных / Л. В. Рычкова // Мовознауства. Літаратура. Культуралогія. Фальклярыстыка. XIII Міжнародньї з'езд славыстау. Даклады беларускай дзлегацьіі. – Мінськ, 2003. – С. 184–195.

7. McEnery T. Wilson A. Corpus Linguistics An introduction / T. McEnery, A. Wilson. – Edinburgh : Edinburgh University Press, 2001. – 235 p.

8. Meyer Ch. F. Pre-electronic corpora / Ch. F. Meyer // Corpus Linguistics. An International Handbook. Edited by A. Lüdeling, M. Kytö. – 2008. – Volume 1. – P. 1–15.

9. Meyer Ch. P. English Corpus Linguistics. An introduction / Ch. P. Meyer. – Cambridge University Press, 2004. – 168 p.