ISSN 2519-2558
ПРОБЛЕМИ ЛІНГВІСТИКИ ТЕКСТУ ТА ДИСКУРСУ

Отримано: 14 січня 2020 року
Прорецензовано: 23 січня 2020 року
Прийнято до друку: 29 січня 2020 року
e-mail: lesya.kotsyuk@oa.edu.ua
yuriy.kotsyuk@oa.edu.ua
DOI: 10.25264/2519-2558-2020-9(77)-106-110

Коцюк Л. М., Коцюк Ю. А. Класифікаційна парадигма корпусу текстів за особливостями його дизайну, структури та способами використання, а також способом фіксації та індексації текстових даних. *Наукові записки Національного університету «Острозька академія»: серія «Філологія»*. Острог: Вид-во НаУОА, 2020. Вип. 9(77). С. 106–110.

УДК: 81'33

Коцюк Леся Миколаївна,
кандидат філологічних наук, доцент
Національний університет «Острозька академія»
Коцюк Юрій Анатолійович,
кандидат психологічних наук, ст. викладач
Національний університет «Острозька академія»

КЛАСИФІКАЦІЙНА ПАРАДИГМА КОРПУСУ ТЕКСТІВ ЗА ОСОБЛИВОСТЯМИ ЙОГО ДИЗАЙНУ, СТРУКТУРИ ТА СПОСОБАМИ ВИКОРИСТАННЯ, А ТАКОЖ СПОСОБОМ ФІКСАЦІЇ ТА ІНДЕКСАЦІЇ ТЕКСТОВИХ ДАНИХ

У статті робиться спроба доповнити класифікацію корпусів текстів. Представлено класифікаційну парадигму текстових корпусів з огляду на те, яка його структура та дизайн, зокрема за цим параметром виділено збалансований / репрезентативний корпус, корпус з гнучкою структурою, завершений, повнотекстовий корпус, фрагментарний, паралельний та порівнянний корпуси, а також статичний та динамічний / моніторинговий корпуси. Виявлено, що парадигму за параметром «спосіб фіксації та індексації текстових даних у корпусі» складають друкований корпус, корпус електронних текстів, корпус транскрибованого мовлення, аудіо/відео корпус, мультимедійний корпус, а також простий / нерозмічений / неіндексований / нетегований корпуси та анотований / розмічений / індексований / тегований корпус. Корпуси, в залежності від того, як ними користуються, поділено на категорії "за метою" (як, наприклад, дослідницький та ілюстративний корпуси) та "за доступністю" (корпуси у вільному доступі, закриті корпуси, а також, комерційні корпуси). Також представлено приклади згаданих типів корпусів текстів. У статті представлено термінологічні еквіваленти назв корпусів за типом мовних даних в українській та англійській мовах.
Ключові слова: *корпус, текстові дані, тип корпусу, типологічні характеристики.*

Lesia M. Kotsiuk,
Cand.Sc. (Philology), Associate Professor
Ostroh Academy National University
Yurii A. Kotsiuk,
Cand.Sc. (Psychology), senior lecturer
Ostroh Academy National University

CLASSIFICATIONAL PARADIGM OF A TEXT CORPUS BY ITS DESIGN, STRUCTURE AND USE, AS WELL AS BY THE FIXATION AND INDEXATION METHODS OF ITS TEXT DATA

The article attempts to analyze the typological characteristics of text corpora. The author proposes to classify corpora with consideration of different aspects of this modern linguistic notion, namely the design and structural features of the corpus (balanced / representative corpus, opportunistic corpus, complete corpus, full-text corpus, fragmentary corpus, parallel corpus and comparable corpus, static / sample corpus, dynamic / monitor corpus), the method of fixing and indexing text data in the corpus (printed corpus, electronic text corpus, transcribed speech corpus, audio/video corpus, multimodal corpus, plain corpus, annotated corpus), as well as the way of how the corpus can be used. According to the aim of the corpus use one can distinguish between a linguistic and illustrative corpus. Due to the access possibilities, there can be identified an open-access corpus, closed-access corpus and the commercial one. Examples of these types of text corpora are also presented. The article presents terminological equivalents of corpus names by the type of text data in Ukrainian and English.
Key words: *corpus, text data, corpus type, typological characteristics.*

Класифікаційне різноманіття типів корпусів текстів зумовлено великою кількістю класифікуючих ознак, використовуваних науковцями у дослідження. Вагомий внесок в опис цих корпусів зробили такі науковці як О. Демська-Кульчицька, В. В. Жуковська, В. В. Риков, А. Н. Баранов, Н. Є. Карпіловська, Ґ. Кеннеді та Дж. Синклер, Н. Деш [1; 2; 3; 4; 5; 6; 7; 8]. Стаття є продовженням статті «Парадигма типологічних характеристик корпусу за типом текстових даних» [9], і, на відміну від попередньої в якості класифікуючих ознак розглядає особливості дизайну, структури та способи використання корпусів, а також способи фіксації та індексації текстових даних у них (див. табл. 1).

Завданням цієї статті ставимо дослідити різноманіття корпусів, зважаючи на те, які особливості їх дизайну та структури, спосіб фіксації та методи їх використання бралися до уваги при їх укладанні. Окремою метою вважаємо за потрібне дати визначення англо-українських відповідників назв кожного типу корпусу та представити відомі приклади таких корпусів.

За особливостями дизайну та структури корпусу розрізняють *збалансовані корпуси (balanced / representative corpus), корпуси з гнучкою структурою (opportunistic corpus), завершений (comlpete), повнотекстові (full-text corpus) та фрагментарні (fragmentary corpus), паралельні (parallel corpus) та порівнянні (comparable corpus), статичні корпуси (static/sample corpus)*, *динамічні/моніторингові корпуси (dynamic/monitor corpus)*

У ***збалансований / репрезентативний корпус*** *(balanced / representative corpus)* тексти відбираються у попередньо визначених пропорціях для того, щоб якомога краще відобразити певну мову чи її різноманітність. Наприклад, *корпуси Браунської сім'ї* – корпуси писемного мовлення на 1 млн. слів, які містять 15 текстових категорій, 500 текстів кожен по 2000 слів.

Таблиця 1

Парадигма типологічних характеристик корпусу за особливостями дизайну та структури корпусу, способом фіксації його текстових даних та особливостей його використання

ОСОБЛИВОСТІ ДИЗАЙНУ ТА СТРУКТУРИ КОРПУСУ
збалансований/репрезентативний корпус (balanced/representative corpus) корпус з гнучкою структурою (opportunistic corpus) завершений (complete corpus) повнотекстовий корпус (full-text corpus), фрагментарний корпус (fragmentary corpus) паралельний корпус (parallel corpus), порівнянний корпус (comparable corpus) статичний корпус (static/sample corpus), динамічний/моніторинговий корпус (dynamic/monitor corpus)
СПОСІБ ФІКСАЦІЇ ТА ІНДЕКСАЦІЇ ТЕКСТОВИХ ДАНИХ У КОРПУСІ
друкований корпус (printed corpus) корпус електронних текстів (electronic text corpus) корпус транскрибованого мовлення (transcribed speech corpus) аудіо/відео корпус (audio/video corpus) мультимедійний корпус (multimodal corpus) простий/нерозмічений/неіндексований/нетегований (plain corpus) анотований/розмічений/індексований/тегований корпус (annotated corpus)
ОСОБЛИВОСТІ ВИКОРИСТАННЯ КОРПУСУ
за метою використання корпусу
дослідницький корпус (linguistic corpus), ілюстративний корпус (illustrative corpus)
за доступністю
вільно доступний корпус (open-access corpus), комерційний корпус (commercial corpus), закритий корпус

Структура корпусів, які належать до цієї сім'ї однакова та наперед визначена, хоча вони репрезентують різні варіанти англійського мовлення: наприклад, американської англійської – *Браунський корпус* (*Brown corpus)* 1961 року [10] та *Фрайбург-Браунський корпус* (*Frown corpus*) 1992 року [11]; британської англійської – *Ланкастер-Браунський корпус* (*LOB*) 1961 року [12], *Фрайбург-Ланкастер-Браунський корпус* (*FLob*) 1991 року [13]; індійської англійської (*Kolhapur*) [14], новозеландської англійської (*The Wellington Corpus of Spoken New Zealand English*) [15]; австралійської англійської (*ACE*) [16].

Ще один приклад збалансованого корпусу – відомий *Британський національний корпус* (*British National Corpus*), з чітко організованою структурою, до якого увійшло 100 млн. слів, з них 10% усного мовлення. *Корпус сучасної американської англійської мови* (*The Corpus of Contemporary American English (COCA)*) – найбільший збалансований корпус американської англійської мови.

Корпус з гнучкою структурою (*opportunistic corpus*) – сукупність таких електронних текстів, які можна отримати, модифікувати та опрацьовувати безкоштовно або за символічну плату. Зазвичай такий корпус незавершений та недовершений за своєю формою і користувачі можуть його наповнити чи модифікувати відповідно до своїх потреб. Цінним такий корпус видається для досліджень, у яких його розмір, склад, наявність доступу та практичність не відіграють важливої ролі у продукуванні тверджень про мову чи її різноманітність. По своїй суті, корпуси з гнучкою структурою є «віртуальними» у тому сенсі, що з них є можливість вибирати та користуватися лише тією частиною, яка цікава дослідникові. Вважається, що моніторні корпуси зазвичай мають гнучку структуру.

Паралельні корпуси *(parallel corpus)* **–** корпуси, які містять оригінальні тексти та їх переклади на одну чи більше інших мов.

Паралельні корпуси за кількістю залучених мов можуть бути *двомовними* (*bilingual*) чи *багатомовними* (*multilingual*).

Якщо брати до уваги напрямок перекладу, то виділяють *однонаправлені* (*uni-directional*) *паралельні корпуси*, як наприклад, з української на англійську чи з англійської на українську мови; *двонаправлені* (*bi-directional*) *паралельні корпуси*, які наприклад, включають як оригінальні тексти українською мовою та їх переклади англійською, так і оригінальні тексти англійською та їх переклади українською; *різнонаправлені* (*multi-directional*) *паралельні корпуси* – корпуси, до яких, наприклад, увійшли оригінальні тексти українською мовою та їх переклади англійською, німецькою та французькою. До останньої категорії також можна віднести тексти, які продукуються одночасно декількома мовами [17].

Паралельні корпуси не завжди легкі для опрацювання користувачами. Адже, задля того, щоб корпус був корисним, необхідно ідентифікувати, які речення у підкорпусі є перекладами інших, і які слова є перекладами яких. Корпус, який вказує на ці особливості відомий як *aligned corpus (протиставний корпус)*, оскільки він показує зв'язки між елементами, які є взаємними перекладами один одного. Наприклад, у корпусі речення «Das Buch ist auf dem Tisch» та «The book is on the table» можуть протиставлятися один одному. На подальшому етапі, певні слова також можна протиставити, наприклад. «Das» з «The». Це не завжди простий процес, оскільки часто одне слово у певній мові може протиставлятися декільком словам у іншій мові, наприклад, німецьке слово «raucht» можна протиставити англійському «is smoking».

Паралельні корпуси виникли ще у Середньовіччі, коли були популярні «багатомовні Біблії», які включали в себе тексти Біблії поряд з перекладами на давньоєврейську, латинську та грецьку мови. Зараз паралельні корпуси створюються відділами з комунікацій у багатомовних організаціях, як, наприклад, ООН, НАТО, ЄС та офіційно двомовних країнах, до прикладу – у Канаді. На даний час у наявності є декілька розмічених паралельних корпусів, а ті, що існують, зазвичай двомовні, а не багатомовні. Проекти, фінансовані ЄС (*Multilingual Aligned Annotated Corpus* (*CRATER*) та *Multilingual Text Tools and Corpora* (*MULTEXT*)) мають за свою мету створити унікальний багатомовний паралельний корпус. *Canadian Hansard corpus* складається з паралельних текстів французькою та англійською мовами, але включає в себе обмежену кількість типів текстів (протоколи засідань канадського парламенту). *Корпус паралельних текстів Європарл* (*Europarl parallel corpus*) включає в себе тексти засідань Європейського парламенту 21 європейською мовами: романськими (французькою, італійською, іспанською, португальською, румунською), германськими (англійською, голландською, німецькою, датською, шведською),

слов'янськими (болгарською, чеською, польською, словацькою, словенською), фінсько-угорськими (фінською, угорською, естонською), балтійськими (латвійською, литовською), та грецькою.

Паралельні корпуси набувають особливої актуальності у наш час, оскільки вони є чудовою можливістю поєднати оригінальні тексти з їх перекладами та глибше пізнати саму природу перекладу. Вони можуть стати джерелом для створення інструменту, який спростить процес перекладу [18].

На противагу *паралельному корпусу* ***порівнянний корпус*** (*comparable corpus*) можна визначити як корпус, до якого компоненти відібрані за однаковою схемою, однаково репрезентативні та збалансовані: мають такі ж пропорції текстів одного формату у такій же мовній різноманітності з одного часового періоду, але різними мовами або мовними різноманітностями. Підкорпуси порівнянного корпусу не є перекладами один одного.

В *одномовних порівнянних корпусах* протиставляються діалекти, варіанти мови, наприклад, такі різновиди англійської мови, як англійська мова як іноземна для різних національностей, або англійська мова як офіційна мова у різних країнах. Прикладом паралельного корпусу може слугувати *Корпус міжнародної англійської мови* (*International Corpus of English – ICE*) [19], проект, у якому зібрано одномільйонні підкорпуси національно-особливих англомовних текстів, наприклад, канадійської англійської мови, австралійської англійської мови і т.п. Прикладом багатомовного порівнянного корпусу може слугувати *Корпус контрактного права* (*Aarhus corpus in contract law),* який складається з набору трьох одномовних корпусів текстів з юриспрунденції: датського, французького та англійського контрактного права, які не є перекладами тих же ж самих текстів.

У порівнянному корпусі усі тексти близькі за змістом, але вони різняться мовою чи мовною різноманітністю. Такий корпус може складатися з газетних статей, опублікованих одного дня і на одну тему, але у різних газетах чи журналах. Основна функція порівнянного корпусу – надати можливість порівнювати різні мови чи їх мовні різноманітності, текстові дані яких представляють схожі сфери чи умови спілкування, без залучення будь-яких втручань, яким може бути переклад.

За об'ємом тестових даних, які увійдуть до корпусу розрізняють ***повнотекстові корпуси*** (*full-text corpus*) та ***фрагментарні/фрагментнотекстові корпуси*** (*fragmentary corpus*) [1]. Більшість сучасних корпусів – повнотекстові. До таких відносимо й *авторські корпуси,* й *корпуси спеціальних коротких текстів*, наприклад, *Берлінський корпус змін* (*Berliner Wendekorpus*), сформований з метою створення колекції особистого досвіду участі у соціальному переломі, відомому під назвою «Руйнування стіни 1989 року», або корпус мерфізмів (так званих «законів підлості») [20, 19 с.].

Як відомо, *Браунський корпус (Brown Corpus)* і *Ланкастер-Осло-Бергенський корпуси* повинні були строго відповідати певним критеріям, одним з яких була довжина тексту, яка прирівнювалася до 2000 слів (слововживань). Очевидно, що текстів, які б чітко відповідали таким критеріям, практично немає. Відповідно, ці корпуси є *фрагментнотекстовими* [20, 19 с.].

За часовим параметром (хронологічною ознакою): *статичні корпуси (static/sample corpus)*, *динамічні/моніторингові корпуси (dynamic/monitor corpus)*.

Статичні корпуси *(static/sample corpus)* засвідчують стан мови на певному синхронному зрізі [1]. Першопочатково корпуси текстів створювалися як статичні утворення, які відображали певний часовий стан мовної системи. Статичні корпуси складаються з текстів певного часового проміжку [20, 19 с.]. Типовими представниками цього типу корпусів є корпуси *Браунської сім'ї (Brown corpora), авторські корпуси* – колекції текстів одного письменника.

Динамічні/моніторингові корпуси *(dynamic/monitor corpus)* забезпечують можливість відстежувати зміни у мові, враховуючи аспект діахронії [1]. Значна частина чисто лінгвістичних і не тільки лінгвістичних завдань потребує виявлення функціонування мовних явищ на часовій шкалі – наприклад, зміна значення слова, частоти використання тих чи інших синтаксичних конструкцій і т.д. Для відображення процесуального аспекту проблемної області розробляються технології побудови та експлуатації динамічних корпусів текстів [4]. Динамічні корпуси називають також ***моніторними*** чи ***моніторинговими корпусами***. Їх основна мета – постійно нарощувати свій об'єм. На протязі завчасно запланованого та зафіксованого проміжку часу відбувається оновлення та/чи доповнення множини текстів корпусу [20, 18 с.]. ***Моніторні корпуси*** важливі для лексикографів, які досліджують потік нових текстів на предмет появи нових слів, або зміни значень старими словами. Їх основними перевагами є:

– вони не статичні – завжди можна додати нові тексти, на відміну від синхронних «знімків», представлених у якомусь завершеному корпусі.

– їх всеосяжність – вони представляють широкий та великий спектр мови.

Проте моніторні корпуси не є надійним джерелом для кількісних даних (на противагу якісним характеристикам), оскільки вони постійно змінюються у розмірі і процес відбору не такий суворий, як це відбувається у завершеному корпусі.

Прикладами моніторних корпусів є: *Банк англійської мови* (*Bank of English*), *Корпус сучасної американської англійської мови* (*Corpus of Contemporary American English COCA*) [21], *Гельсінський корпус англійських текстів (Helsinki Corpus of English Texts), Корпус текстів журналу Time* (*Time magazine corpus*) [22].

<u>За способом фіксації текстових категорій у корпусі</u> виділяють *друковані корпуси (printed), корпуси електронних текстів (electronic text corpus), корпуси транскрибованого мовлення (transcribed speech corpus), аудіо та відео корпуси (audio/video corpus), мультимедійні корпуси* (*multimedia corpus*)*.*

Корпуси транскрибованого мовлення *(transcribed speech corpus)*. До такого типу корпусів відноситься, у першу чергу, *усний корпус* (*spoken corpus*) (див. *корпус усного мовлення*), такий корпус, у якому зібрані зразки усного мовлення теговані з допомогою різних типів транскипції (наприклад, орфографічної, фонетичної). Наприклад, *Corpus of Professional Spoken American English (CPSA); Ланкастерський корпус усного мовлення* (*Lancaster/IBM Spoken English Corpus (SEC)); Веллінгтонгський корпус усного мовлення новозеландської англійської (Wellington Corpus of Spoken New Zealand English), Корпус усного мовлення дітей* (*CHILDES*) [23] – усні корпуси, де мовлення представлено у письмовій формі з додатковими символами, які зазвичай використовують у транскрипції.

Аудіо та відео корпуси *(audio/video corpus)* **–** *The Michigan Corpus of Academic Spoken English (MICASE)* містить біля 1,7 млн. слововживань (близько 200 годин записів) сучасного усного університетського мовлення, що було записано в Мі-

чиганському університеті. Проект *Один речевой день* (*ОРД*) Санкт-Перербургського державного університету – звуковий корпус сучасної російської мови щоденного спілкування.

Мультимедійні/мультимодальні корпуси *(multimodal corpus)* включають в себе мультмедійний матеріал (наприклад, відео записи та їх транскрибовані версії). Прикладом такого корпусу є система SACODEYL [24] у якій розміщено відео, аудіо та орфографічну транскрипцію інтерв'ю з підлітками сімома мовами: англійською, французькою, німецькою, італійською, литовською, румунською, та іспанською.

Іншим прикладом мультимодального корпусу є *Російськомовний емоційний корпус* (*REC*), розмічений з врахуванням даних про міміку, рухи рук, брів и т. п., дозволяє вивчити стратегії емоційної взаємодії та конфлікту, неперервану комунікативну поведінку, хезитації [http://www.harpia.ru/rec]. В Іркутському державному лінгвістичному університеті (Росія) ініціювали створення *Навчального Мультимодального Корпусу* (*Учебный Мультимодальный Корпус* (УМКО)) відеозаписів непідготовлених навчальних діалогів носіїв та «не носіїв» російської та китайської мов за певними темами, розмічених в програмі ELAN та представлених також у вигляді паралельних корпусів, вирівняних за смисловими блоками всередині діалогів.

За наявністю індексації виділяють *прості / нерозмічені / неіндексовані / нетеговані* (*plain*) та *анотовані / розмічені / індексовані / теговані* (*annotated*) корпуси.

У розміченому корпусі словам чи реченням присвоюються мітки/теги у відповідності до *характеру розмітки*: морфологічні, синтаксичні, семантичні, просодичні і т.д.

Анотовані транскрипцією – *усні корпуси*.

ОСОБЛИВОСТІ ВИКОРИСТАННЯ КОРПУСУ

За метою використання корпусу: *дослідницький, ілюстративний.*

За А. Барановим, ***дослідницькі корпуси*** (*linguistic corpus*) призначені переважно для вивчення різних аспектів функціонування мовної системи. Цей тип корпусів орієнтований на широкий клас лінгвістичних завдань, а тому вони зазвичай великі за обсягом: від десятків мільйонів до сотень мільйонів слововживань. Згідно Демської О. дослідницькі корпуси застосовують у лінгвістичних дослідженнях із метою формулювання нових теорій, концепцій тощо [1].

Мета ***ілюстративних корпусів*** не стільки виявити нові факти, скільки підтвердити і обґрунтувати вже отримані результати. Такі корпуси зазвичай створюються після проведення наукового дослідження з метою виділення з них лінгвістичних прикладів, які підтверджують ті чи інші мовні (мовленнєві, текстові) факти, які були виявлені раніше із допомогою інших лінгвістичних прийомів [4, 116 с.]. Одним з підвидів такого корпусу можна назвати *перекладний корпус (translational corpus)* (за визначенням Фернандеса Л. [25, p. 92.]), мета якого дослідити процес перекладу та його результати.

За доступністю виділяють *вільно доступні корпуси*, *комерційні корпуси*, *закриті корпуси* [20, 18 с.].

Вільно доступні корпуси дозволяють у будь-який час в режимі он-лайн здійснювати пошук по усіх текстах корпусу у повному об'ємі. У деяких випадках вільний доступ може надаватися до частини корпусних даних та не зі всіма функціональними можливостями. Наприклад, Корпус сучасної американської англійської мови (*The Corpus of Contemporary American English (COCA)*) – найбільший корпус американської англійської мови у вільному доступі.

У роботі з ***комерційними корпусами*** право на його використання он-лайн чи копію на компакт-диску потрібно купляти. Попередньо можна ознайомитися з анотацією до корпусу чи, можливо, навіть попрацювати з корпусом у тестовому режимі, але не зі всіма текстами, а лише з окремими підкорусами: *Лонгманський корпус учнівської англійської мови* (*Longman Learners' Corpus)* та *Кеймбріджський корпус учнівської англійської мови (Cambridge Learner Corpus*).

Закриті корпуси створюються зі спеціальною метою і не призначені для публічного використання.

Література:

1. Демська-Кульчицька О. Дещо про класифікацію текстових корпусів. *Наукові записки. Серія: Мовознавство*. 2004. 1 (11). С. 153–157.

2. Жуковська В. В. Ресурси корпусної лінгвістики у дослідженні історичної динаміки мови. *Слово і речення: синтактика, семантика, прагматика: матер. Міжнар. наук. конф.* / М-во осв. і науки України; Київ. ун-т ім. Б. Грінченка. Київ: Київ. ун-т ім. Б. Грінченка, 2013. С. 151–156.

3. Рыков В. В. Прагматически ориентированный корпус текстов. *Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог – 99»*. (Москва – Таруса). Москва, 1999. URL: http://rykov-cl.narod.ru/t.html (дата доступа 20.12.2015). Название с экрана.

4. Баранов А. Н. Введение в прикладную лингвистику. Москва: УРСС Эдиториал, 2001. 358 с.

5. Карпіловська Н.Є. Вступ до комп'ютерної лінгвістики. Донецьк: Юго-Восток, 2003. 183 с.

6. Sinclair J. M. Corpus typology. *A framework for classification. Studies in anglistics*. Stockholm: Almqvist & Wiksell. 1995. P. 17–33.

7. Sinclair J. M. Corpus Typology Draft. 1996. URL : http://www.ilc.cnr.it/EAGLES/typology/typology.html (access date 20.12.2015). Title from the screen.

8. Dash, Niladri Sekhar. Corpus Linguistics and Language Technology : With Reference to Indian Languages. New Delhi : Mittal Publications, 2005.

9. Коцюк Л. М., Коцюк Ю. А. Парадигма типологічних характеристик корпусу за типом текстових даних. *Наукові записки Національного університету «Острозька акаде-мія»: серія «Філологія»*. Острог : Вид-во НаУОА, 2018. Вип. 4(72), грудень. С. 18–22.

10. Francis W. N., Kucera H. The Brown Corpus / Brown University, Providence. 1998. URL : http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html (access date 12.01.2020).

11. The Freiburg-Brown corpus of American English (Frown). Project leader: Christian Mair. *VARIENG.* Albert-Ludwigs-Universität. Freiburg. 2007. URL : http://www.helsinki.fi/varieng/CoRD/corpora/FROWN (access date 12.01.2020).

12. Lob Corpus. Was compiled by researchers in Lancaster, Oslo and Bergen Lancaster-Oslo/Bergen. URL : http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html (access date 12.01.2020).

13. The Freiburg–LOB Corpus of British English (F-LOB). Project leader: Christian Mair. *VARIENG.* Albert-Ludwigs-Universität. Freiburg. URL : http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/ (access date 12.01.2020).

14. The Kolhapur Corpus. By S. V. Shastri in collaboration with C. T. Patilkulkarni, Geeta S. Shastri / Department of English Shivaji University. Kolhapur. 1986. URL : http://clu.uni.no/icame/manuals/KOLHAPUR/INDEX.HTM (access date 12.01.2020).

15. The Wellington Corpus of Spoken New Zealand English Project director: Prof Janet Holmes *School of Linguistics and Applied Language Studies / Victoria University of Wellington.* 1998. URL : http://www.victoria.ac.nz/lals/resources/corpora-default/corpora-wsc (access date 12.01.2020)

16. Peters Pam, Australian Corpus of English (ACE) / Macquarie University *Australian National Corpus* 1986. URL : https://www.ausnc.org.au/corpora/ace.

17. McEnery A. M. and Xiao R. Z. Parallel and comparable corpora: What are they up to? In: *Incorporating Corpora: Translation and the Linguist*. Translating Europe. Multilingual Matters. Clevedon. 2007. URL : http://eprints.lancs.ac.uk/59/1/corpora_and_translation.pdf (access date 12.01.2020). ISBN 978-1-85359-986-6.

18. Sinclair J. EAGLES Preliminary recommendations on Corpus Typology EAG--TCWG--CTYP/P / School of English, University of Birmingham. Version of May, 1996. URL : http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html (access date 12.01.2020).

19. International Corpus of English (ICE) / The ICE Project. [2000]. URL : http://ice-corpora.net/ice/ (access date 12.01.2020).

20. Захаров В. П., Богданова С. Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика». 2-е изд., перераб. и дополн. СПб. : СПбГУ. РИО. Филологический факультет, 2013.

21. Corpus of Contemporary American English. *English-Corpora.org* Creator Mark Davies. [2019]. URL : https://www.english-corpora.org/coca/ (access date 12.01.2020).

22. TIME Magazine Corpus. *English-Corpora.org* Creator Mark Davies. [2019]. URL : https://www.english-corpora.org/time/ (access date 12.01.2020).

23. Maxine Eskenazi, Jack Mostow, and David Graff. The CMU Kids Corpus LDC97S63. (Computer file : CD for computer : Sound recording). Philadelphia: Linguistic Data Consortium, 1997. URL : https://www.worldcat.org/title/cmu-kids-speech-corpus/oclc/39510571 (access date 12.01.2020).

24. SACODEYL *SACODEYL European Youth Language*. Project Coordinator: Pascual Pérez-Paredes / EU Socrates-Minerva project "System Aided Compilation and Open Distribution of European Youth Language". Corpora Universidad de Murcia [2008] URL : http://www.um.es/sacodeyl/ (access date 12.01.2020).

25. Fernandes Lincoln Corpora in Translation Studies: revisiting Baker's typology. Florianópolis, 2006 Fragmentos, número 30, p. 087-095 URL : http://www.translationindustry.ir/Uploads/Pdf/8217-24982-1-PB.pdf (access date 12.01.2020).