МОВОЗНАВСТВО. ЛІТЕРАТУРОЗНАВСТВО

УДК 811.581.11
DOI https://doi.org/10.24919/2308-4863/48-2-13

Олександр КОЗОРІЗ,
orcid.org/0000-0002-4502-583X
кандидат філологічних наук,
асистент кафедри мов і літератур
Далекого Сходу та Південно-Східної Азії
Інституту філології Київського національного університету
імені Тараса Шевченка
(Київ, Україна) davinci@3g.ua

КРИТЕРІЇ ВИЗНАЧЕННЯ ЯКОСТІ ЛІНГВІСТИЧНИХ КОРПУСІВ НА ПРИКЛАДІ КИТАЙСЬКОЇ, ЯПОНСЬКОЇ ТА ІНДОНЕЗІЙСЬКОЇ МОВ

У результаті дослідження автором було створено вісім лінгвістичних корпусів паралельних текстів на основі різних джерел з порівняльною метою (три основних і п'ять похідних). Дієвими виявилися методи завантаження матеріалу для корпусу на основі частотного списку, електронних словників, а також частотних списків слів попередньо самостійно створених корпусів. Проведено теоретичні розвідки та практичні дослідження задля нормалізації корпусу. Результативними для дослідження корпусу виявилися співвідношення type / token ratio, показники середньої довжини речення ASL тощо. TTR, ASL, type, tokens та інші показники корпусу можуть різнитися відповідно до мови, лівої чи правої частини корпусу. Побудова графіків розподілу лексики по частотності та довжині речень у корпусі яскраво унаочнює результати досліджень, ефективно репрезентує матеріал. Для укладених корпусів було обраховано загальну кількість знаків, слів та речень у корпусі з відповідною узагальнюючою таблицею; встановлено середню довжину речень ASL, співвідношення type / token ratio TTR; написано програму на мові Python для обрахунку зазначених величин; складено частотні списки лексики, пораховано загальну кількість унікальної лексики та побудовано відповідні логарифмічні графіки; запропонована методика аналізу розподілу лексики частотного словника тексту на основі графіків шляхом поділу їх на три частини: початкову, середню та хвостову – вважається нами перспективною. Запропоновано комплексну методику визначення якості корпусу на основі показників TTR, type, показників графіку розподілу лексики корпусу по частотності, а також розподілу речень корпусу по кількості слів. Для визначення продуктивних моделей речень запропоновано методику статистичного зрізу у вигляді окремих корпусів із відповідними довжинами речень. Розроблено методологію роботи з корпусами східних мов – китайської, японської, що не містять проміжків між словами, а також порівняльну методику для індонезійської мови.

Ключові слова: *якість лінгвістичного корпусу, електронний корпус текстів, паралельний корпус, нейронні мережі, системи автоматичного перекладу, китайська мова, японська мова, індонезійська мова.*

Oleksandr KOZORIZ,
orcid.org/0000-0002-4502-583X
Ph. D. in Philology,
Assistant at the Department of Languages and Literatures
of the Far East and Southeast Asia
Taras Shevchenko National University of Kyiv
(Kyiv, Ukraine) davinci@3g.ua

CRITERIA FOR DETERMINING THE QUALITY OF LINGUISTIC CORPORA BASED ON THE EXAMPLE OF CHINESE, JAPANESE AND INDONESIAN LANGUAGES

The result of author's research is creation of eight linguistic corpora of parallel texts based on different sources in comparative purposes (three as the main and five as the derivatives). The methods for downloading material for the corpus, which are based on the frequency list, electronic dictionaries, as well as frequency word lists of previously self-created corpora, proved to be effective. The theoretical investigations and practical researches for normalization of the corpus were carried out. The type / token ratio, indicators of the average sentence length ASL, etc. proved to be effective for the study of the corpus. TTR, ASL, types, tokens and other indexes may vary according to the language, left or right part of the corpus. The construction of the graphs of the vocabulary distribution by frequency and length of sentences in

the body clearly illustrates the results of the research and represents the material effectively. For the enclosed corpora, the total number of signs, words and sentences in the corpus with the corresponding summary table were calculated; the average sentences length (ASL), the type / token ratio (TTR) were determined; the program was written in Python to calculate all these values; frequency lists of vocabulary were compiled, the total number of unique vocabulary was calculated and the corresponding logarithmic graphs were constructed. The proposed method of analysis of the frequency dictionary vocabulary distribution, which is based on the graphs by dividing them into three parts (initial, middle and tail), is considered promising. A complex method for determining the quality of the corpus based on the TTR, type indicators, indicators of the graph of the corpus vocabulary distribution by frequency, as well as the distribution of corpus sentences by the quantity of words is offered. To determine the productive models of sentences, a method of the statistical section in the form of separate corpora with the appropriate lengths of sentences is shown. The methodology of working with the corpora of Oriental languages (Chinese, Japanese), which do not contain spaces between words, has been developed. The comparative methodology for the Indonesian language was conducted.

Key words: *linguistic corpus quality, electronic text corpus, parallel corpus, neural networks, automatic translation systems, Chinese language, Japanese language and Indonesian language.*

Нині спостерігається значний інтерес до використання корпусів в освітній та професійній сферах. Значна кількість словників укладається на основі корпусів. Серйозною проблемою виявляється визначення якості корпусу, відбір необхідного матеріалу, в тому числі серед величезної кількості наявних корпусів. Адже добір текстів для корпусу, а також текстів перекладу – права частина корпусу, визначають якість термінології, що добувається на їхній основі.

Актуальність і новизна теми полягає у тому, що запропоновано аналіз на основі поєднання різноманітних статистичних характеристик корпусу для визначення його якості, у тому числі в автоматичному режимі. Швидке визначення якості лінгвістичних корпусів великих обсягів із навчальною, дослідницькою та прикладною метою на основі статистичних даних раніше не здійснювалося подібним чином. Створення якісних паралельних корпусів східних мов із українською парою наразі є необхідною умовою для покращення роботи систем машинного перекладу. Окремо була розроблена методологія роботи з корпусами східних мов, що не містять проміжків між словами: китайська та японська мови.

Задля досягнення мети дослідження слід виконати такі завдання: визначити джерела лінгвістичного матеріалу; створити чи отримати кілька корпусів із різних джерел, користуючись відмінними методами; розглянути основні можливі характеристики створених корпусів; запропонувати критерії оцінювання якості корпусів. Під якістю корпусу ми розуміємо високий ступінь різноманітності ілюстративного матеріалу для окремої лексичної одиниці. Якість паралельного корпусу, як оцінювання якості перекладених текстів правої частини паралельного корпусу, у межах цієї статті розглядатися не буде.

В Україні у сфері корпусної лінгвістики працювали такі дослідники: О. О. Балабан, Н. М. Бобер, М. М. Брик, Н. П. Дарчук, О. А. Дюндик, А. М. Желєзко, В. В. Жуковська, В. П. Захаров, П. В. Зернецький, О. М. Зубань, Л. С. Івашкевич, Я. В. Капранов, Є. А. Карпіловська, В. І. Качанов, В. В. Комаренко, А. В. Корольова, Ю. В. Кравцова, Н. Є. Леміш, Л. Л. Макарук, Т. Б. Маслова, С. А. Матвєєва, Б. О. Назаров, В. О. Папіжук, В. М. Підвойний, Ю. І. Позніхіренко, В. Ф. Старко, А. А. Таран, О. М. Тищенко, О. В. Ткачик, Т. С. Толчеєва, М. О. Шведова, С. М. Щербина. Основні напрями їхніх досліджень можна побачити у матеріалах першої Міжнародної науково-прикладної конференції «Прикладна і корпусна лінгвістика: розроблення технологій нового покоління» за 25 квітня 2018 року. Закордоном відомі такі прізвища: S. Hoffmann, S. Evert, G. Kennedy, T. MacEnery, C. Meyer, T. Otlogetswe, J. Sinclair, J. Svartvik, E. Tognini-Bonelli, A. Hardie, G. Leech, Ch. Fillmore, P. Baker, Douglas Biber, W. Francis, R. Xiao, Lu. Xiaofei та багато інших. Усі зазначені дослідники мають власний підхід, завдання і мету дослідження корпусів, що лише опосередковано по'вязані з результатами наших досліджень, запропонованими тут методиками створення та аналізу корпусів.

Найбільш вдалим прикладом іноземних досліджень, що корелюють із нашими, буде стаття G. Corpas P. та M. Seghiri «Розмір має значення» (Gloria Corpas Pastor), де описується робота програми ReCor, яка призначена для визначення ідеального розміру репрезентативного корпусу; показує кількість типів, кількість лексем, співвідношення TTR, кількість слів, що з'являються лише один раз (V1) та кількість слів, які з'являються лише двічі (V2); зазначена програма окремо створює файли списки слів із їхньою частотністю поданими в алфавітному порядку та ранжованими. Попри вагомі результати дослідниці доходять висновку, що наразі питання репрезентативності корпусу та його ідеальний розмір лишається невизначеним.

Під лінгвістичним корпусом ми розуміємо великий (не менше одного мільйона речень) за

обсягом набір мовних даних, відібраних за певним організуючим принципом, які зберігаються в електронному вигляді. Відповідно паралельний корпус – це електронний корпус унікальних речень, який окрім оригінальних текстів має переклади іншою мовою, що вирівняні відповідно до оригіналу за реченнями. Головним атрибутом якісного корпусу виступає його репрезентативність, обумовлена домінантою, що об'єднує всі тексти єдиним логічним задумом. Усі корпусно-орієнтовані дослідження покладаються на якість, кількість та репрезентативність корпусу як основу для отримання достовірних результатів.

Оскільки жоден корпус текстів не містить інформацію про всі сфери лексики і граматики, яким би великим цей корпус не був, а будь-який малий корпус може надати факти, які ніколи не можна було б з'ясувати жодним іншим способом, тому основна задача при створені корпусу – це збалансовано відібрати матеріал для дослідження.

На думку Т. МакЕнері репрезентативність більшості корпусів великою мірою обумовлюється набором жанрів, включених до корпусу (збалансованість), та критеріями відбору текстів кожного жанру (відібраність) (McEnery 2006, 11).

На нашу думку, основними напрямками використання паралельних корпусів на сьогодні є: 1) навчальні та дослідницькі цілі, аби перевірити, чи використовується певна лінгвістична форма в мові в певному контексті; 2) для створення систем машинного перекладу, які теж «навчаються» на паралельних корпусах; тому репрезентативним буде відбір речень за двома критеріями: лексичним та граматичним. Лексичний полягає у доборі речень, що містять слова за частотним списком мови або списками-слів створених на базі термінологічних словників, при цьому кожне таке слово має бути представлене мінімум 10 ілюстративними реченнями у корпусі. Визначити граматичний критерій дещо складніше, оскільки відсутній частотний список граматичних моделей речень, а варіативність мовлення наближається до безкінечності.

Однак критерієм, що корелює з частотністю граматичних моделей речень, може бути довжина речення у поєднанні з частотністю кожного слова цього речення. Ми знаємо, що речення короткої і середньої довжини трапляються частіше в усному мовленні, натомість художню прозу і популярні журнали репрезентують середні та довгі речення; самі лише довгі речення характерні для публіцистичних та наукових текстів. Тому, ймовірно, варіативність вживання певного слова для певної довжини речення буде обмеженою.

До нашої уваги потрапили такі статистичні характеристики корпусів як: загальна кількість знаків (*characters*), загальна кількість слів (*words* або *tokens*), загальна кількість речень (*sentences*), загальна кількість різних слів (*type*), співвідношення TTR, ASL тощо. Співвідношення *type / token ratio* (TTR) – це співвідношення між типами та лексемами корпусу. «Типи» – це сукупність різних слів корпусу. Як показали попередні наші дослідження, чим довший текст, тим менший відсоток TTR. Значення TTR рахувалося за англійською частиною паралельного корпусу. ASL – це середня довжина речення (англ. *Average sentence length*) = *words / sentences*). Представлені нижче статистичні характеристики рахувалися за англійською частиною корпусів, оскільки тут легше здійснити поділ на слова, а згадані вище індекси були розроблені саме під англійську мову.

Окремо автором було розроблено методологію дослідження корпусу шляхом поділу логарифмічного графіка розподілу лексики корпусу по частотності на три частини: початкову, середню та кінцеву. Кінцева ділянка графіка представляє лексику з частотою вживання, що дорівнює 1, середня частина графіка репрезентує частотність від 10 до 2 слововживань, початок графіка – лексика з частотністю вище за 11 включно. На нашу думку частотність 10 є оптимальною для логарифмічного графіка.

Скориставшись своїм попереднім досвідом, було створено власний паралельний корпус китайсько-англійських перекладів на основі сайту-словника Quword (Quword). Першим кроком було складання списку слів для пошуку та завантаження. З цією метою, спочатку за основу було взято частотний список англійської мови 5000 слів, який є у вільному доступі; оскільки інформація щодо частотного списку на 60 і 220 тисяч слів дається фрагментарно – лише кожне п'яте слово (Word frequency data).

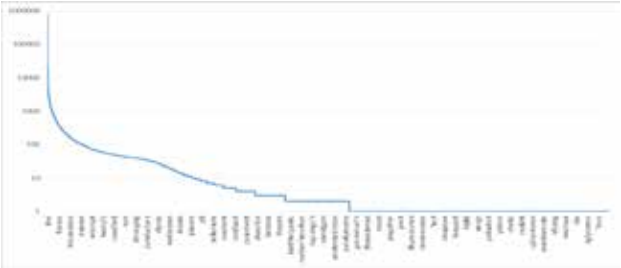
Завантаження сторінок зазначеного словника виконувалося за допомогою програми WinHTTrack Website Copier (HTTrack Website Copier). Пошукові запити: *https://www.quword.com/liju/w/only* або *https://www.quword.com/search?word=*汉语. Для подальшого редагування отриманого html коду за допомогою регулярних виразів використовувався текстовий редактор EmEditor (EmEditor). Об'єднуємо усі отримані сторінки в одну командою: «copy *.* file_name». Фільтруємо всі строчки, що починаються з тегів: «<dl><dt>». Вилучаємо зайве. Не всі словники можна скачувати таким чином, наприклад, сайт https://www.youdao.com/ має захист від скрейпінгу.

Для подальшого збільшення величини корпусу ми застосували інший метод. Шляхом першого завантаження було отримано корпус на 106 000 паралельних пар речень китайської і англійської мов; після складення частотного списку англійської частини цього корпусу, отримано словник-список вже на 42 000 слів, на основі якого була повторена процедура завантаження і отримано корпус до 920 000 пар речень або 12 900 000 лексем, який має словник на 166 000 слів, 46% яких (76 000) вживаються лише один раз – правий «хвіст» графіка; середня частина графіка – частоти від 10 до 2-х – займає близько 28% лексики (47 000); найчастотнішими є перші 26% слів (43 000), див. **Граф. 1.** TTR корпусу – 1,3 % (166 000 / 12 900 000). Середня довжина речення становить 14 слів.

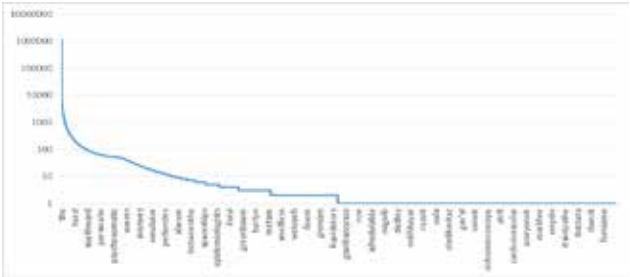
Дещо інакший підхід було застосовано для отримання інформації з іншого сайту Xxenglish (Xxenglish). На основі словника Collins (Collins English Dictionary) було підготовлено список слів (близько 80 000) та проведено завантаження. У результаті отримано корпус на 1 330 000 пар речень або 18 691 000 лексем, який має словник на 322 000 слів, 50% яких (162 000) вживаються лише один раз – правий «хвіст» графіка; середня частина графіка – частоти від 10 до 2-х – займає близько 30% лексики (96 000); найчастотнішими є перші 20% слів (64 000). TTR корпусу – 1,7 % (322 000 / 18 691 000). Середня довжина речення становить 13,68 слів.

У порівнянні помітно аномально високий TTR другого корпусу, хоча із зростанням обсягу корпусу ця величина має зменшуватися, тобто тут він має бути меншим за 1,3% (TTR першого меншого корпусу). Також на **Граф. 2** помітно, що «сходинка» лексики, з якої починається спад, змістилася до початку графіка, тобто кількість гарно проілюстрованої лексики (більше 10 слововживань на корпус) зменшилася щодо загального обсягу лексики. Можна говорити про недостатню якість корпусу, а отже і сайту-першоджерела для потреб створення корпусів.

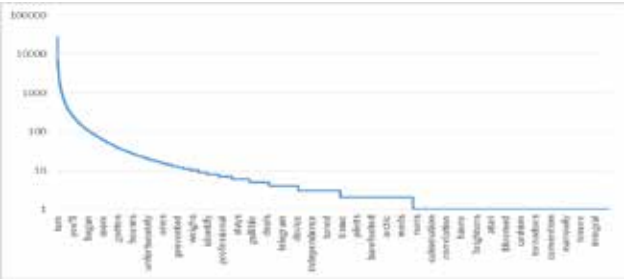
Для порівняння було також досліджено третій корпус отриманий з сайту OPUS (The open parallel corpus) обсягом 106 000 пар речень або 668 000 лексем, який має словник всього на 11 700 слів; 35% яких (4150) вживаються лише один раз – правий «хвіст» графіка; середня частина графіка – частоти від 10 до 2-х – займає близько 41% лексики (4750); найчастотнішими є перші 24% слів (2 800), див. **Граф. 3**. TTR корпусу – 1,75 % (11 700 / 668 000). Середня довжина речення становить 6,3 слів. Для порівняння та узагальнення результатів аналізу корпусів див. **Табл. 1**.



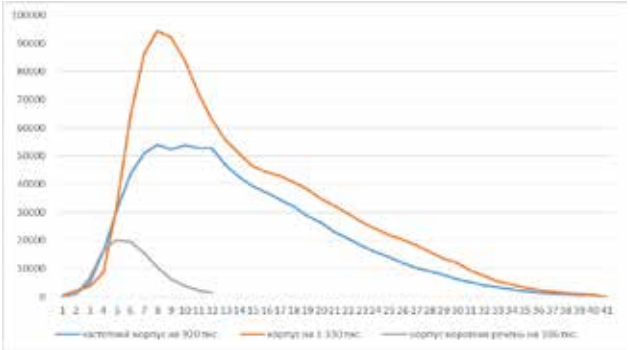
Граф. 1. Логарифмічний графік розподілу лексики корпусу по частотності на 920 000



Граф. 2. Логарифмічний графік розподілу лексики корпусу по частотності на 1 366 000



Граф. 3. Логарифмічний графік розподілу лексики корпусу по частотності на 106 000



Граф. 4. Графіки розподілу кількості слів у реченні для корпусів на 920 000, 1 330 000 та на 106 000 речень

Таблиця 1

Порівняльна таблиця статистичних даних корпусів

graph	characters	words (token)	sentences	type	TTR	ASL	початок	середина	кінець
1	61 372 000	12 900 000	920 000	166 000	1,3 %	14	26%	28%	46%
2	91 182 000	18 691 000	1 366 000	322 000	1,7%	13,68	20%	30%	50%
3	2 476 000	668 000	106 000	11 700	1,75%	6,3	24%	41%	35%

Усі зазначені у таблиці характеристики можна автоматично підрахувати за допомогою невеличкої програми, написаної автором на мові Python, що використовує модуль Natural Language Toolkit (NLTK), спеціально розроблений для аналізу «природних мов»:

```
import nltk as nlp
import re
lines = 0
with open('test.txt', 'r', encoding='utf-8') as f:
        for line in f:
                lines += 1
with open('test.txt', 'r', encoding='utf-8') as f:
        doc=f.read()
        doc=re.sub(r'[^\w]', ' ', doc)
        doc2=re.sub(r'[^\w]', '', doc)
        doc=doc.lower()
```

```
tokens=nlp.word_tokenize(doc)
types=nlp.Counter(tokens)
TTR="%.2f" % ((len(types)/len(tokens))*100)
ASL="%.2f" % (len(tokens)/lines)
print('characters: ' + str(len(doc2)))
print('tokens: ' + str(len(tokens)))
print('sentences: ' + str(lines))
print('types: ' + str(len(types)))
print('TTR: ' + str(TTR))
print('ASL: ' + str(ASL))
```

Середня довжина речення ASL дає слабке уявлення про реальні довжини речень корпусу. Тому було досліджено довжину кожного речення у словах для зазначених корпусів та побудовано графіки, тобто пораховано кількість речень у корпусі відповідної довжини. Кількість слів у реченні визначалася регулярним виразом: ^(\w+\W+){X}$ – де X, кількість слів у реченні; словосполучення тут не враховані, оскільки в кінці виразу обов'язково має стояти розділовий знак; «\W+» – означає один і більше розділовий знак (пробіл, кома, крапка, тире, апостроф тощо); регулярний вираз «\w+» використовувався для пошуку слів, відповідно «\w» – для знаків (символів). Підрахунок загальної кількості знаків із пробілами та іншими розділовими знаками у тексті дещо відрізняється від кількості знаків у самих лише словах, це треба враховувати при визначенні середньої довжини слова.

На **Граф. 4** розподілу кількості слів у реченні добре видно, що другий, найбільший корпус, має пік в районі 8 слів – тобто це досить характерна довжина речення цього корпусу, таких речень тут найбільше. Загалом середня довжина речень двох найбільших корпусів від 6 до 12 слів, що характерно для розмовної мови; менший корпус, створений штучно, спеціально для коротких речень. У другому (найбільшому за обсягом) корпусі помітна аномалія: кількість речень з довжиною у 4 слова менша вдвічі (8600 речень проти 16000 речень), ніж у корпусах суттєво менших за обсягом. Цікаво, що усі корпуси мають відносно незначну кількість речень – на 2, 3, 4 і 5 слів.

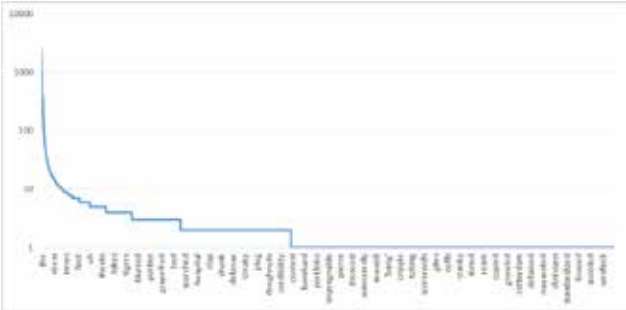
Тому було проведено інший експеримент для речень із довжиною у 4 слова. З цією метою на основі попередніх корпусів зроблено статистичний зріз у вигляді трьох окремих корпусів з довжинами речень у 4 слова. Результати статистичних досліджень представлено у **Табл. 2**, корпуси 1.a, 2.a, 3.a.

Звернімо увагу, що найвагомішим показником якості корпусу є TTR, чим він менший – то краще. На великих корпусах він може становити показники менші за 1, на малих має наближатися до 1.

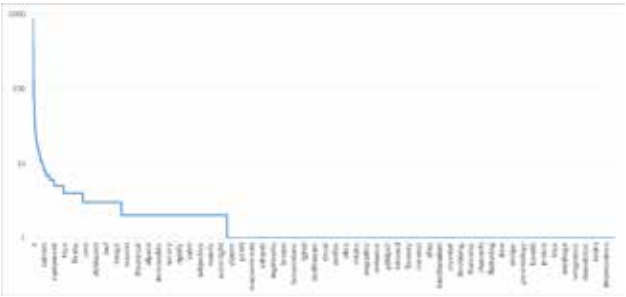
Таблиця 2

Порівняльна таблиця статистичних даних корпусів на 4 слова у реченні

graph	characters	words (token)	sentences	type	TTR	ASL	початок	середина	кінець
1.a	331320	64396	16101	16361	25%	4	3,5%	40,5%	56%
2.a	205445	34492	8623	14853	43%	4	1,7%	32,3%	66%
3.a	261567	66840	16711	4839	7,2%	4	15%	42%	43%
3.b	220480	59116	14780	2765	4,7%	4	25%	70%	5%
3.c	117719	32636	8160	703	2,1%	4	81%	19%	0,01%



Граф. 1.a Логарифмічний графік розподілу лексики корпусу на 920 000 по частотності для 4-слівних речень

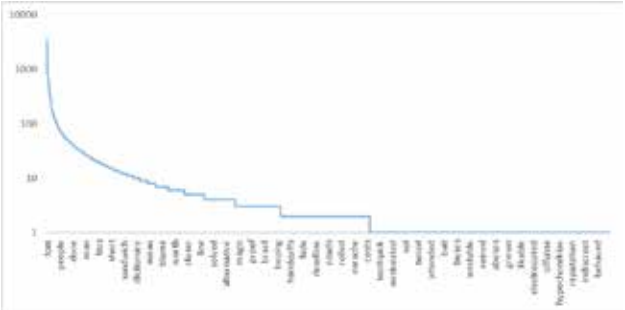


Граф. 2.a Логарифмічний графік розподілу лексики корпусу по частотності на 1 366 000 для 4-слівних речень

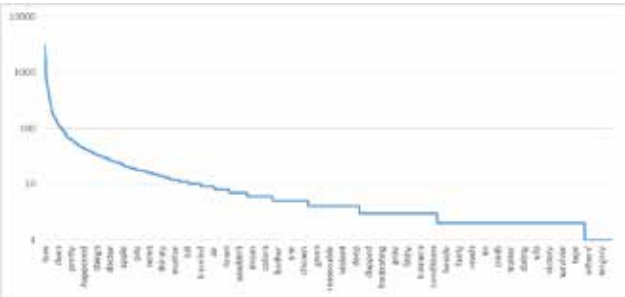
Показники у кілька десятків відсотків означають, що перед нами корпус, який складається зі словосполучень. Зазначені спостереження не стосуються оцінки якості невеличких творів, де TTR має бути навпаки більшим, що буде говорити про елегантність і різноманіття мови автора.

Не менш важливими є кількісні показники частотності лексики, що добре видно на графіках розподілу лексики корпусів по частотності для речень довжиною в чотири слова, див **Граф. 1.a, 2.a, 3.a**.

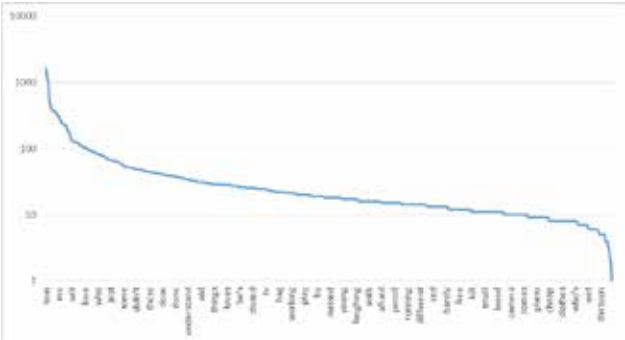
Найгірші результати репрезентації лексики продемонстрував корпус 2.a, який має довжелезний «хвіст» – 66% (слова, що вживаються лише один раз), невеличку середину – 32,3%, і куций початок 1,7% – означає, що корпус складається переважно зі словосполучень, які не повторюються. У корпусі 1.a дещо коротший «хвіст» 56%, дещо довша середина 40,5% і початок 3,5% – це кращий результат ніж у попередньому випадку, але все ще не ідеально як для корпусу, що репрезентує саме текстовий матеріал. Найкраще репрезентує лексику корпус із графіком 3.a, у нього найкоротший «хвіст» – 43% і найбільший початок – 15%, хоча загальний словник корпусу (показник type – 4839) менший за попередні два корпуси, але кожне слово має широкий ілюстративний вжиток.



Граф. 3.a Логарифмічний графік розподілу лексики корпусу по частотності на 106 000 для 4-слівних речень



Граф. 3.b Логарифмічний графік розподілу лексики корпусу по частотності на 106 000 для 4-слівних речень, скоригований по частотності 1



Граф. 3.c Логарифмічний графік розподілу лексики корпусу по частотності на 106 000 для 4-слівних речень, скоригований по частотності від 1 до 10

Проведемо ще один експеримент, який полягає у вилучені з корпусу речень, що мають рідковживану лексику: обріжемо правий «хвіст» корпусу, див. **Граф. 3.b** (тут обрізано речення, що містять слова на одне входження) та **Граф. 3.c** (тут обрізано речення, що містять слова від 1 до 10 входжень). Як видно з цих графіків та статистичної інформації у **Табл. 2.** Нам вдалося покращити TTR корпусу до 2,1%, а початкову частотну лексику довести до 81%. Хоча загальний показник кількості лексики (показник type) зменшився до 703 одиниць, а корпус представляє лише речення з довжиною у чотири слова, але суттєво зросла

якість ілюстративного матеріалу, майже на кожне слово із частотного списку тепер можна знайти більше десяти прикладів вживання у корпусі. Тобто, якщо застосовувати такі тексти у навчальному процесі, студенту достатньо мати лише 703 слова словникового запасу для їхнього розуміння. Такі скориговані корпуси можуть бути використані як дидактичний матеріал для студентів, так і для створення термінологічних словників.

На підтвердження наших слів проілюструємо якість корпусу **3.а** вживанням англійської словоформи «*saw*», див. **Табл. 3**. Хоча наведено лише деякі приклади вживання у невеличкому корпусі речень, втім, очевидно, що моделі речень, представлені тут для довжин у 4 слова, є достатньо репрезентативними. Основна модель речення буде «S – P – O» у різноманітних варіаціях додавання однослівного означення чи обставини або складного додатку (*Complex Object*).

Для роботи зі східними мовами, що не містять проміжків між словами, як наприклад китайська чи японська, можна додатково використовувати програму поділу речень на слова SegmentAnt (SegmentAnt). Програма звісно дає певний відсоток помилок, але в ній дозволено використовувати також власний словник для збільшення точності правильного сегментування речення на слова.

Аби перевірити якість роботи програми та методологію роботи із зазначеними східними мовами загалом було використано відносно невеликий японсько-український та китайсько-український корпус субтитрів із сайту OPUS (The open parallel corpus). Після поділу на слова східна частина текстів виглядає наступним чином, див. **Табл. 4.**

Таблиця 3

Ілюстративна таблиця вживання словоформи «*saw*» для корпусу 3.а на 4 слова у реченні

He also saw it.	他也看到了。	I saw what happened.	我看到发生了什么事。
I saw a dog.	我看到了一条狗。	I saw you yesterday.	我昨天看到你了。
I saw five men.	我看到五个男人。	I saw your father.	我看见了你父亲。
I saw her yesterday.	我昨天见过她。	No one saw anything.	没有人看到任何东西。
I saw him jump.	我看到他跳了。	No one saw us.	没有人看到我们。
I saw him recently.	我最近见过他。	Nobody saw it coming.	没有人看到它的到来。
I saw him running.	我看到他在跑步。	She saw him smile.	她看到他微笑。
I saw the fight.	我看到了打架。	Tom never saw Mary.	汤姆从未见过玛丽。
I saw the sunrise.	我看到了日出。	Tom saw the cat.	汤姆看见了那只猫。
I saw them kissing.	我看到他们接吻。	We saw her dance.	我们看到了她的舞蹈。
I saw Tom blush.	我看到汤姆脸红了。	We saw it all.	我们看到了一切。
They saw something odd.	他们看到了奇怪的东西。	Tom saw what happened.	汤姆知道发生了什么事。

Таблиця 4

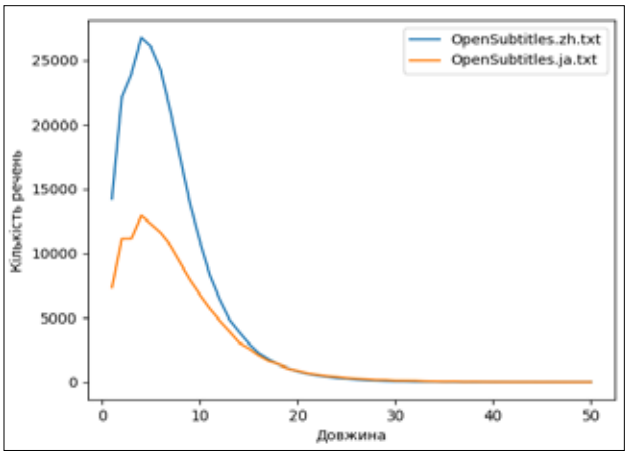
Ілюстративна таблиця роботи програми SegmentAnt для японської та китайської мов

японська мова	**українська мова**
過去 を 支配 する 者 は 未来 まで 支配 する 現在 を 支配 する 者 は 過去 まで 支配 する これ が 我々 の 国 平和 と 豊穣 の 国 調和 と 希望 の 国 ここ が 我々 の 国	Хто контролює минуле – той контролює майбутнє, Хто контролює сьогодення – той контролює минуле Це наша земля. Земля миру та достатку. Земля гармонії та надії. Це наша земля.
китайська мова	**українська мова**
不 ， 我要 生活 ！ 我们 要 玩转 这个 城镇 ！ 小心 ， 你 怎样 开车 的 我 在 开车 吗 ？ 我 喜欢 你 的 车 你 开 它 ， 它 就是 你 的 我 的 朋友 在 哪里 ？ 让 他 进来 ， 他 是 我 的 客人 ！ 我们 去 买 一些 花	Ні, я буду жити! Будемо святкувати на всю. Уважніше за кермом. Я за кермом? Мені подобається твоя машина. Тоді забирай її, вона твоя. Де мій друг? Впусти його, він мій гість! Давай купимо квітів.

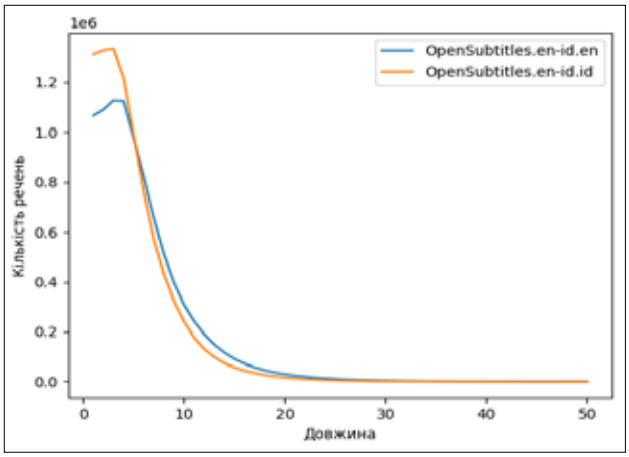
Таблиця 5

Таблиця статистичних даних корпусів субтитрів японської та китайської мов

Language	characters	words (tokens)	sentences	type	TTR	ASL
японська	1 551 151	879 784	132 404	41 372	4.70	6.64
китайська	2 215 564	1 420 580	237 741	70 921	4.99	5.98



Граф. 5. Графіки розподілу кількості слів у реченні для корпусів японської та китайської мов



Граф. 6. Графіки розподілу кількості слів у реченні для англо-індонезійського корпусу субтитрів на 9 млн. речень

Таблиця 6

Таблиця статистичних даних англо-індонезійського корпусу субтитрів

Language	characters	words (tokens)	sentences	type	TTR	ASL
англійська	215 019 441	57 316 396	9268181	210 298	0.37	6.18
індонезійська	244 991 807	46 285 450	9268181	310 553	0.67	4.99

Аналіз статистичних характеристик проводився за східномовними частинами корпусу з використанням зазначеної вище програми на Python. Як бачимо з отриманих даних, див. **Табл. 5** та **Граф. 5**, наша методологія цілком дозволяє проводити статистичний аналіз східних мов – японської та китайської, що не мають поділу тексту на окремі слова. Показники TTR говорять про те, що перед нами корпуси середньої якості, а показники ASL та графіки розподілу кількості слів у реченні підказують розмовний характер текстів, що не суперечить зазначеній тематиці субтитрів. Інші показники дають нам уявлення про розмір корпусу: загальну кількість речень, слів, унікальних слів та символів.

Аби урізноманітнити методологію роботи з корпусами інший тип експерименту ми провели на базі лівої і правої частин англо-індонезійського корпусу субтитрів із сайту OPUS (The open parallel corpus), порівнявши розподіл довжин речень для цих двох мов, див. **Табл. 6** та **Граф. 6**.

Корпус має досить великий обсяг – 9 мільйонів речень, тому результати дослідження, а отже і висновки, є абсолютно достовірними. Як бачимо на **Граф. 6** – кількість речень зазначена у млн., у індонезійській мові суттєво більше (на 200 тис.) кожної довжини коротких речень від 2 до 5 слів, що корелює з показниками tokens та ASL у **Табл. 6** і говорить про більшу довжину англійського речення у порівнянні до індонезійського. Цікавим також є факт, що кількість унікальних слів (type) в англійській мові є набагато меншою ніж в індонезійській.

Отже, найкраще оцінку якості корпусу дає комплексний підхід, визначальними є поєднання таких факторів як: побудова графіків розподілу лексики по частотності у поєднанні з графіками розподілу кількості слів у реченні (довжини речень), співвідношення *type / token ratio* TTR, показник *type* (загальна кількість унікальної лексики), а також запропонована методика аналізу розподілу лексики частотного словника на основі графіків шляхом поділу їх на три частини: початкову, середню та хвостову.

Будь-який корпус допасовується до цілей і мети, що мотивували його створення. Відповідно

до критеріїв зазначених нами вище, графіки розподілу кількості слів у реченні мають бути збалансовані по всій довжині, бажано аби речення кожної довжини у корпусі були представлені рівномірно по кількості. TTR, ASL, type, tokens та інші показники корпусу можуть різнитися відповідно до мови, лівої чи правої частини корпусу. Рекомендований показник TTR якісного корпусу має становити менше 1%, ідеально 0,5% на великих корпусах (більше двох мільйонів речень); показник *type* – що більше, то краще, але він корелює з попереднім показником; графік розподілу лексики по частотності має мати рівномірну висоту по всій довжині осі координат «X», що буде означати широкий ілюстративний вжиток кожної лексичної одиниці у межах корпусу (більше 10 одиниць). Ці показники можна корегувати експериментально, вилучаючи з корпусу речення, що мають рідковживану лексику (частотність = 1 або позиції від 1 до 10), але оскільки вони тісно пов'язані з довжинами речень, то буде відповідне розбалансування довжин речень, прогнозоване їх зменшення. Тому, можна запропонувати іншу методику: додавання до корпусу речень, що містять лексику з малими частотами від 1 до 10, на жаль, як показала практика, не завжди можна знайти необхідний матеріал.

Для визначення продуктивних моделей речень виправданою є методика статистичного зрізу у вигляді окремих корпусів із відповідними довжинами речень, на основі яких можна робити ґрунтовні всебічні узагальнення щодо лексичного вжитку і граматичних моделей. Тексти таких корпусів можуть бути ефективно використані як дидактичний матеріал для студентів на початковому етапі навчання іноземним мовам, для створення термінологічних словників, навчання нейронних мереж, систем автоматичного перекладу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Collins English Dictionary (2006), 8th Ed., Harper Collins Publishers, London.
2. EmEditor. URL: https://www.emeditor.com/ (accessed 12 March 2021).
3. Gloria Corpas Pastor & Miriam Seghiri (2010), "Size matters: A quantitative approach to corpus representativeness", pp. 111–146. URL: https://wlv.openrepository.com/handle/2436/622560 (accessed 12 March 2021).
4. HTTrack Website Copier. URL: http://www.httrack.com (accessed 12 March 2021).
5. McEnery T., Xiao R. and Tono Y. (2006), Corpus-based Language Studies: an Advanced Resource Book / T. McEnery, R. Xiao, Y. Tono. London, Routledge, 386 p.
6. NLTK. URL: http://www.nltk.org/ (accessed 12 March 2021).
7. QuWord. URL: www.quword.com/ (accessed 12 March 2021).
8. SegmentAnt. URL: https://www.laurenceanthony.net/software/segmentant/ (accessed 12 March 2021).
9. The open parallel corpus. URL: https://opus.nlpl.eu/ (accessed 12 March 2021).
10. Word frequency data. URL: www.wordfrequency.info/samples.asp (accessed 12 March 2021).
11. Xxenglish. URL: www.xxenglish.com/ (accessed 12 March 2021).

REFERENCES

1. Collins English Dictionary (2006), 8th Ed., Harper Collins Publishers, London.
2. EmEditor. URL: https://www.emeditor.com/ (accessed 12 March 2021).
3. Gloria Corpas Pastor & Miriam Seghiri (2010), "Size matters: A quantitative approach to corpus representativeness", pp. 111–146. URL: https://wlv.openrepository.com/handle/2436/622560 (accessed 12 March 2021).
4. HTTrack Website Copier. URL: http://www.httrack.com (accessed 12 March 2021).
5. McEnery T., Xiao R. and Tono Y. (2006), Corpus-based Language Studies: an Advanced Resource Book / T. McEnery, R. Xiao, Y. Tono. London, Routledge, 386 p.
6. NLTK. URL: http://www.nltk.org/ (accessed 12 March 2021).
7. QuWord. URL: www.quword.com/ (accessed 12 March 2021).
8. SegmentAnt. URL: https://www.laurenceanthony.net/software/segmentant/ (accessed 12 March 2021).
9. The open parallel corpus. URL: https://opus.nlpl.eu/ (accessed 12 March 2021).
10. Word frequency data. URL: www.wordfrequency.info/samples.asp (accessed 12 March 2021).
11. Xxenglish. URL: www.xxenglish.com/ (accessed 12 March 2021).