УДК 811.581.11
DOI https://doi.org/10.24919/2308-4863/35-3-16

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

ПОРІВНЯЛЬНИЙ АНАЛІЗ РІЗНОТЕМАТИЧНИХ ЛІНГВІСТИЧНИХ КОРПУСІВ

У статті досліджено проблему аналізу різнотематичних корпусів великих обсягів. Запропоновано порівняльну методику та критерії розгляду лінгвістичних корпусів.

Загалом було укладено сім корпусів, створено вузькі спеціалізовані термінологічні корпуси на противагу термінологічним словникам для досліджень функціональних особливостей, моделей речень тієї чи іншої терміносистеми. Отримано корпуси медичного, біологічного, політехнічного, нафтогазового спрямування, корпус частотної лексики та корпус сучасної розмовної лексики, а також об'єднаний корпус на 4 млн пар речень. Нами було створено об'єднаний паралельний корпус на 4 млн пар речень або на 68 млн слів англійської частини, що за обсягом становить 10% від відомого корпусу COCA або корпусу GRAC.

Для усіх корпусів пораховано загальну кількість знаків, слів і речень із відповідною узагальнюючою таблицею; встановлено середню довжину речень ASL, визначено автоматичний індекс читабельності ARI, складено частотні списки лексики, пораховано загальну кількість унікальної лексики, визначено співвідношення type/token ratio TTR.

Опрацьована та підтверджена методика скачування корпусу на основі термінологічних списків. Розроблена методика порівняння рангів морфем різних корпусів із морфемами частотного списку СКМ, що може бути застосована для визначення тематики корпусу чи належності тексту до певної галузі у майбутньому. Запропонована методика визначення продуктивних моделей речень корпусу за допомогою регулярних виразів.

Загалом побудовано чотири графічні діаграми (в т. ч. діаграма розподілу за довжиною речень для семи різнотематичних корпусів) і шість таблиць, які виразно унаочнюють результати досліджень, чітко репрезентуючи матеріал. Підрахунки підкріплено формулами та ілюстративним матеріалом, що дозволяє повторити дослідження для будь-яких інших дотичних систем.

Ключові слова: *лінгвістичний корпус, паралельний корпус, частотний список, type/token ratio, середня довжина речення, ступінь термінологічності, продуктивні моделі речень, регулярні вирази.*

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

COMPARATIVE ANALYSIS OF VARIOUS THEMATIC LINGUISTIC CORPORA

The problem of the analysis of various thematic corpora of large volumes is investigated in the article. The comparative technique and criteria of consideration of linguistic corpora are offered.

In total, seven corpora were compiled, and narrow specialized terminological corpora were created in contrast to terminological dictionaries for the study of functional features, sentence models of one or another terminological system. The corpora of medical, biological, polytechnic, oil and gas directions, the corpus of frequency vocabulary and the corpus of modern colloquial vocabulary were obtained, as well as the combined corpus of 4 million pairs of sentences. As a result of our research, we created a combined parallel corpus of 4 million pairs of sentences or 68 million words of the English part, which is 10% of the known COCA corpus or GRAC corpus.

For all corpora, the total number of signs, words and sentences in the corpus with the corresponding summary table is calculated; the average length of ASL sentences is calculated, the automatic readability index ARI is determined, frequency vocabulary lists are compiled, the total number of unique vocabulary is calculated, the type / token ratio TTR is determined.

The method of downloading the corpus on the basis of terminological lists is developed and confirmed. The method of comparing the ranks of morphemes of different corpora with morphemes of the Modern Chinese Character frequency list

has been developed, which can be used to determine the subject of the corpus or the affiliation of the text to a certain field in the future. The technique of definition of models of sentences of the corpus by means of regular expressions is offered.

In total, four graphical diagrams have been constructed (including a sentence-length distribution diagram for seven thematic corpora) and six tables that clearly illustrate the results of the research, undoubtedly representing the material. The calculations are supported by formulas and illustrative material, which allows you to repeat the study for any other similar systems.

Key words: *linguistic corpus, parallel corpus, frequency list, type / token ratio, average sentence length, terminological degree, productive sentence models, regular expressions.*

Постановка проблеми. У багатьох галузях лінгвістики стали популярними дослідження на основі корпусів. Сьогодні спостерігається значний інтерес до використання корпусів в освітній і професійній сферах. Проблема їх створення й опрацювання набуває істотного значення. Велика кількість словників укладається на основі подібних лінгвістичних корпусів.

Виокремлення невирішених раніше частин загальної проблеми. Створення паралельних різнотематичних лінгвістичних корпусів.

Основні напрямки використання корпусів паралельних текстів різної тематики: 1) з навчальною та дослідницькою метою; 2) для створення систем машинного перекладу.

Мета статті – визначити джерела лінгвістичного матеріалу, запропонувати методики створення власних різнотематичних лінгвістичних корпусів; розглянути основні характеристики створених корпусів; запропонувати методики досліджень корпусів.

Аналіз досліджень. 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. За кордоном відомі такі прізвища: S. Hoffmann, S. Evert, G. Kennedy, T. MacEnery, T. Otlogetswe, J. Sinclair, J. Svartvik, E. Tognini-Bonelli та ін. Усі зазначені науковці мають власний підхід, завдання і мету дослідження корпусів, що безпосередньо не пов'язані з результатами наших досліджень, запропонованими тут методиками створення й аналізу корпусів.

Виклад основного матеріалу. Скориставшись власним досвідом, на основі сайту-словника QuWord (QuWord) ми створили оригінальні паралельні корпуси китайсько-англійських перекладів різних тематик. Під паралельним корпусом ми розуміємо електронний корпус, який, окрім оригінальних текстів, має відповідні переклади іншою мовою, що вирівняні до оригіналу за реченнями з видаленням повторів.

Першим кроком була підготовка списку слів для пошуку та скачування паралельних пар речень. З цією метою спочатку за основу було взято частотний список англійської мови 5 000 слів (Word frequency data). Шляхом скачування було отримано корпус на 106 000 паралельних пар речень китайської та англійської мов; або 1 462 000 лексем англійської частини корпусу (загальна кількість слововживань). Після складання частотного списку цього корпусу отримано словник-список на 42 000 слів, на основі якого була повторена процедура скачування й отримано корпус вже на 920 000 пар речень або 12 900 000 лексем (token), котрий має словник на 166 000 слів (type). Таким чином було отримано перший частотний паралельний корпус, оскільки скачування відбувалося на основі частотних списків.

Окремо було виконано скачування на основі списків-слів, створених на базі термінологічних словників медичного (Ривкин, 2004), біологічного (Чибисова та ін., 2003), політехнічного (Столяров та ін., 2003), нафтогазового (OilAndGas, 1998) спрямування й англо-російського словника сучасної розмовної лексики (Глазунов, 2003). Так було отримано ще п'ять окремих корпусів, а також створено загальний корпус на основі всіх шести корпусів. Після об'єднання усіх шести корпусів і видалення повторів загальний обсяг отриманого паралельного корпусу сягає 4 000 000 пар речень, або 67 800 000 лексем англійської частини (token).

Було підраховано основні статистичні характеристики корпусів, що наведено у табл. 1. Представлені характеристики рахувалися за англійською частиною корпусу, оскільки тут легше здійснити поділ на слова, а індекси були розроблені саме під англійську мову.

Автоматичний індекс читабельності (ARI) (Automated readability index) – міра визначення складності сприйняття тексту читачем, що апроксимує складність тексту до номера класу в американській системі освіти. ARI вираховується за формулою: 4.71 * (characters/ words) + 0.5 *

Таблиця 1

Таблиця статистичних даних створених корпусів

№	Corpus	characters	words	sentences	ARI	ASL
1	частотний	61 372 000	12 904 000	920 000	7,98	14,03
2	медичний	35 520 000	6 843 000	410 000	11,36	16,69
3	біологічний	38 073 000	7 564 000	452 000	10,64	16,73
4	політехнічний	113 921 000	22 543 000	1 260 000	11,32	17,89
5	нафтогазовий	39 878 000	7 917 000	437 000	11,35	18,12
6	розмовний	17 149 000	4 026 000	250 000	6,68	16,10
7	об'єднаний	331 330 000	67 800 000	3 919 000	10,24	17,30

* (words / sentences) – 21.43, де Characters – кількість букв і цифр у тексті; Words – кількість слів у тексті; Sentences – кількість речень у тексті.

Середня довжина речення (англ. Average sentence length (ASL) = words / sentences) – величина, тісно пов'язана з метриками індексу легкості читання Флеша (Flesch Reading Ease). FRE = = 206,835 – 1,015 × ASL – 84,6 × ASW, де ASW – середня довжина слова у складах (англ. average number of syllables per word) = syllables / words.

Спостереження виявили, що корпус частотної лексики має найменшу середню довжину речень – 14, а корпус розмовної лексики має найменший автоматичний індекс читабельності – 6,68, що відповідає віку 12-ти років; хоча середня довжина його речень – 16, співмірна з корпусами медичного та біологічного спрямування. Найдовшу середню довжину речень має корпус політехнічного спрямування – 17,89, але це середні показники, які не зовсім чітко відображають реальну картину.

Тому додатково було досліджено довжину кожного речення у словах для всіх корпусів і побудовано відповідні графіки, тобто пораховано кількість речень у корпусі відповідної довжини. Пошук кількості слів у реченні здійснювався 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, кому, крапку, тире, апостроф тощо); регулярний вираз «\w+» використовувався для пошуку слів, відповідно «\w» – для знаків. Корпуси містять незначну кількість словосполучень термінів, які тут не було враховано. Для дослідження словосполучень можна застосовувати «\W*», що означає нуль або більше розділових знаків.

Як бачимо з рис. 1, об'єднаний корпус якнайкраще репрезентує усі типи речень, тут широко представлені речення з довжиною від 4-х до 33-х слів, більше 40 тис. вживань на кожну довжину речення. Графік корпусу частотної лексики на 920 тис. речень має довжини переважно від 6-и до 13-и слів, що корелює із графіком корпусу розмовної лексики на 250 тис. речень меншого обсягу. Графік політехнічного корпусу, найбільший із галузевих, якнайкраще представляє спеціалізовані корпуси – корелює з медичним, біологічним корпусом і корпусом нафтогазогазової лексики та має довжину від 12-и до 30-и слів у реченні. Тобто частотні та розмовні корпуси кардинально відрізняються від галузевих за довжиною речень.

Далі було складено частотні списки морфем китайської частини корпусів, оскільки в англійській мові багато службових слів і важче знайти ключові слова для корпусу (див. табл. 2). Для економії місця тут ілюстровано лише перші 50 позицій. Шляхом порівняння з морфемами частотного списку сучасної китайської мови (СКМ), створеного професором Цзюнь Да (Middle Tennessee State University) (Jun Da), було визначено найбільш значимі морфеми для певної лінгвістичної тематики. Як показує практика, найбільш значимими є перші 125 морфем частотного списку (Козоріз, 2014). Перші 10 частотних морфем є спільними для майже усіх корпусів і текстів китайської мови з невеличкими розбіжностями за рангами.

Звісно, бажано порівнювати корпуси схожі за об'ємом, втім вважаємо, що порівняння рангів може застосовуватися до корпусів, різних за обсягом. Іншою перешкодою для отримання достовірних даних може бути неспеціалізованість словника, на основі якого створювалися корпуси. Втім, методологія дослідження має правомірність.

Спробуємо визначити термінологічність найбільш значимих морфем політехнічного корпусу шляхом порівняння різниці рангів морфем цього корпусу та морфем частотного списку СКМ (Козоріз, 2014). На рис. 2. представлені найбільш характерні перші 33 морфеми політехнічного корпусу із зазначенням ступеня їхньої термінологічності, підрахованого шляхом віднімання від рангу частоти морфеми СКМ рангу морфеми політехнічного корпусу.

Ступінь термінологічності політехнічного корпусу було пораховано для перших 129 частотних

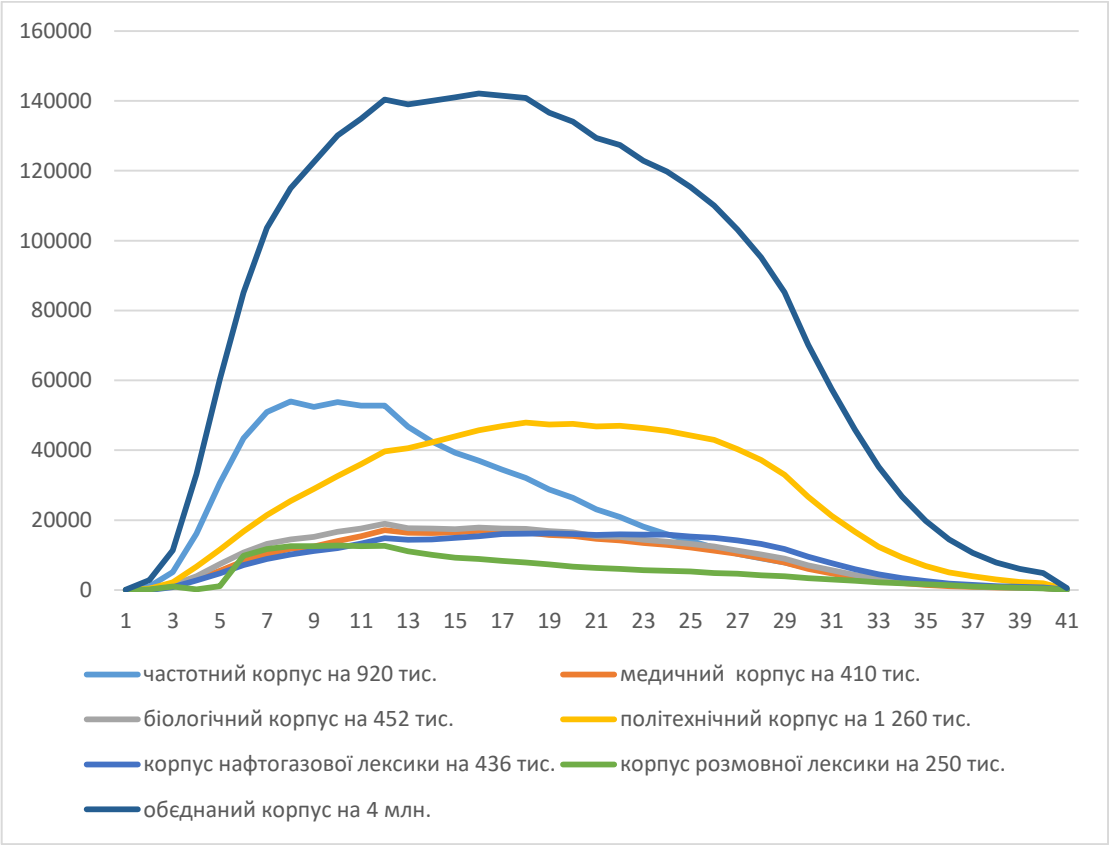

Рис. 1. Графіки розподілу кількості слів у реченні для корпусів

морфем: 析 939, 测 782, 控 651, 压 628, 模 600, 型 473, 料 420, 效 405, 器 392, 构 386, 线 337, 研 335, 究 303, 流 291, 算 261, 程 258, 质 256, 具 244, 设 234, 式 223, 电 213, 统 202, 数 201, 计 198, 量 197, 据 193, 术 184, 基 183, 系 181, 件 174, 结 173, 光 171, 度 150, 形 136, 水 135, 化 133, 接 129, 提 127, 及 126, 制 125, 管 125, 变 122, 通 119, 合 111, 气 106, 性 103, 应 96, 表 90, 机 89, 体 84, 工 79, 果 73, 加 72, 产 71, 高 70, 分 67, 文 66, 或 64, 物 57, 特 56, 并 51, 相 50, 理 49, 间 49, 动 46, 内 44, 使 42, 重 41, 力 40, 方 39, 法 39, 进 38, 将 32, 本 31, 行 30, 实 27, 与 27, 成 23, 定 22, 能 21, 等 20, 点 19, 最 17, 可 15, 面 15, 关 13, 于 12, 作 12, 以 11, 其 7, 种 6, 过 5, 要 2, 为 -2, 时 -4, 这 -5, 现 -5, 发 -7, 上 -9, 它 -9, 明 -11, 生 -16, 开 -21, 部 -23, 到 -24, 地 -26, 因 -27, 们 -29, 多 -30, 下 -32, 子 -33, 学 -35, 同 -35, 如 -39, 来 -42, 些 -44, 人 -45, 后 -47, 他 -48, 会 -55, 所 -56, 主 -58, 而 -62, 自 -65, 经 -74, 你 -81. Можна говорити також про від'ємну термінологічність, тобто непритаманність певних морфем певній терміносистемі, див. рис. 3. Загалом дані про ступінь термінологічності морфем можна застосовувати для автоматичного визначення тематики корпусу чи належності тексту до певної галузі.

Тепер наведімо кілька прикладів вживання морфем «析» та «压» у політехнічному корпусі, аби наочно підтвердити, що корпус дійсно репрезентує саме політехнічну галузь, а морфема справді має високий ступінь термінологічності (див. табл. 3).

Інше цікаве співвідношення для досліджень type/token ratio (TTR) – співвідношення між типами та лексемами корпусу дуже сильно варіюється відповідно до довжини тексту, де «type» – це різні слова в корпусі, word token – це усі лексеми корпусу. Чим довший текст, тим менший буде відсоток.

Спробуємо проаналізувати TTR на найбільшому нашому корпусі на 4 млн пар речень. Після складення частотного списку англійської частини об'єднаного корпусу отримано словник і побудовано відповідний логарифмічний графік лексики корпусу за частотністю (див. рис. 4). Пораховано індекс TTR і деякі інші статистичні характеристики.

Аналіз корпусу показав таке: загальний обсяг різної лексики – 442 000 слів; близько 50% із них (217 000) вживаються лише один раз – правий «хвіст» графіка; середня частина графіка – частоти від 10 до 2-х – займає близько 26% лексики (117 000); найчастотнішими є перші 24% слів (108 000). TTR корпусу – 0,65% (442 000 / 67 800 000). Середня довжина речення становить

Таблиця 2

Таблиця частотних морфем китайської частини корпусів

СКМ	ранг	частотний	медичний	біологічний	політехнічний	нафтогазовий	розмовний	об'єднаний
的	1	的	的	的	的	的	的	的
一	2	一	一	一	一	一	一	一
是	3	是	是	在	在	在	我	是
不	4	在	在	是	了	了	是	在
了	5	了	和	和	用	用	在	了
在	6	我	有	有	是	是	了	有
人	7	他	用	了	和	和	他	和
有	8	有	性	用	有	有	不	不
我	9	这	了	中	中	中	们	我
他	10	们	中	个	个	个	这	个
这	11	个	不	这	对	性	有	这
个	12	不	生	不	分	可	个	用
们	13	人	对	生	以	对	你	人
中	14	和	个	性	能	能	人	他
来	15	为	这	为	可	这	上	中
上	16	中	为	对	这	以	到	们
大	17	用	能	我	电	不	来	为
为	18	上	人	分	不	分	会	以
和	19	以	可	们	性	为	要	对
国	20	到	分	能	为	要	和	上
地	21	可	以	物	方	方	可	可
到	22	来	我	人	机	行	为	能
以	23	你	发	以	行	上	时	大
说	24	大	病	可	要	地	以	要
时	25	能	法	他	上	我	能	时
要	26	对	他	种	法	动	大	到
就	27	地	们	上	动	时	用	生
出	28	会	方	大	于	工	那	来
会	29	时	体	化	时	度	中	性
可	30	要	化	要	数	于	就	分
也	31	生	上	成	我	法	地	于
你	32	些	种	于	出	机	子	地
对	33	于	要	方	大	们	出	行
生	34	出	结	发	度	水	对	会
能	35	过	于	法	系	大	得	方
而	36	子	大	地	成	作	说	出
子	37	那	物	体	作	成	些	你
那	38	行	成	到	制	出	如	作
得	39	作	作	子	工	生	下	法
于	40	成	行	时	理	过	过	发
着	41	国	时	动	过	量	里	成
下	42	发	目	来	们	进	生	过
自	43	就	动	过	进	系	她	国
之	44	多	过	作	量	理	好	种
年	45	种	应	行	化	化	么	动
过	46	而	出	出	到	应	着	子
发	47	她	及	结	地	到	它	多
后	48	说	理	水	应	力	去	理
作	49	得	到	量	器	他	多	机
里	50	方	疗	应	生	人	而	进

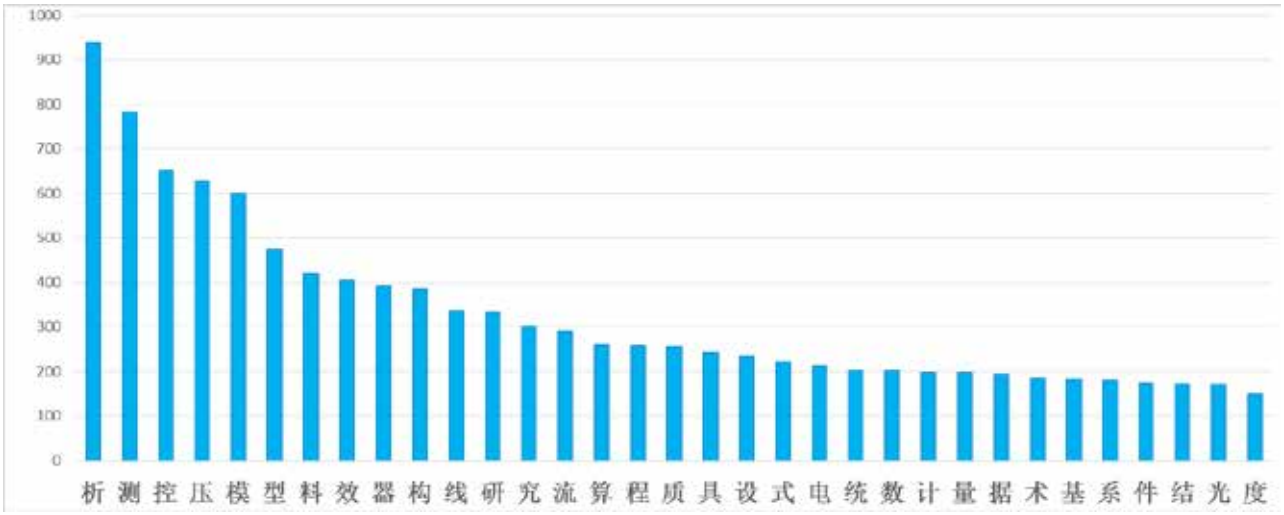

Рис. 2. Найбільш термінологічні морфеми політехнічного корпусу

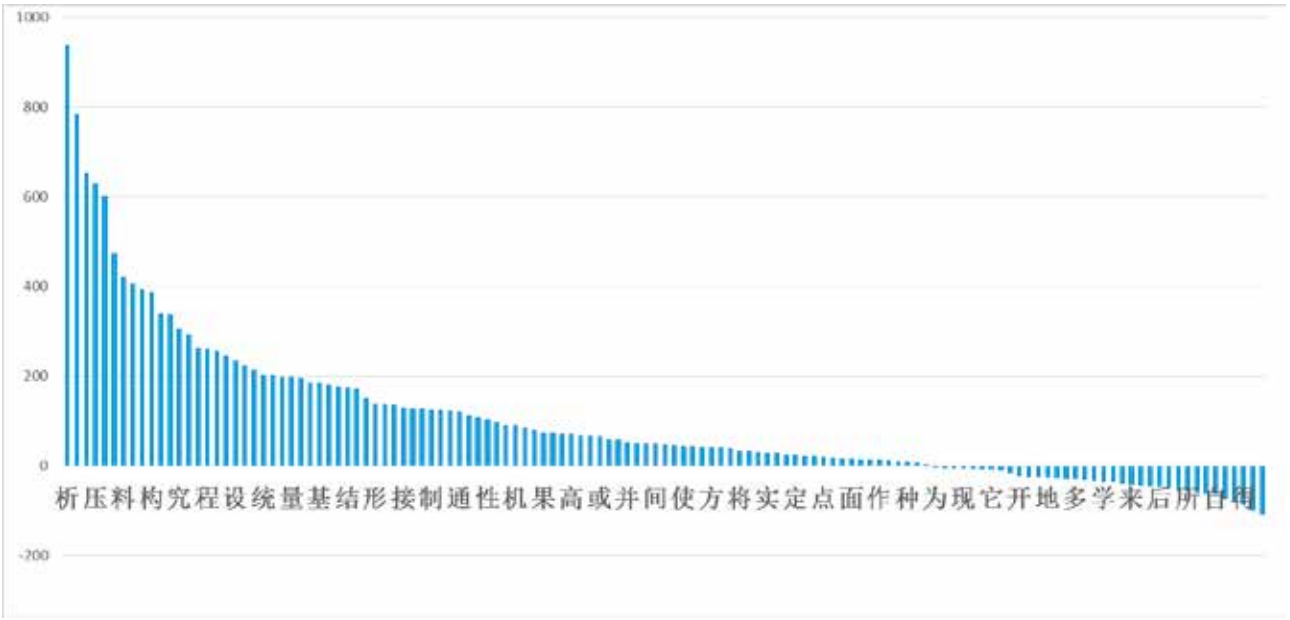

Рис. 3. Загальна термінологічність морфем політехнічного корпусу

Таблиця 3

Приклади вживання морфеми «析» у політехнічному корпусі

介绍了2004欧洲X射线光谱分析会议的简况和会议所讨论的热点问题。	The paper gives a brief introduction to the European Conference on X-ray Spectrometry 2004.
讨论了析出物的析出机制和脱氧产物对析出物的影响。	The precipitation mechanism and the effect of deoxidation production on precipitates were discussed.
利用传输矩阵分析了迭代时反处理的收敛性和时反算子分解方法的原理。	The principle of the time reversal processing and its spatial and time focusing are studied.
反射光线立刻被机器人分析并说出物品的化学成分。	The reflected light is then analyzed in real time (10) to determine the object's chemical composition.
此外，对正弦波激励下的响应进行了减震效果分析。	The shock absorbing effect of the damper isolated system excited by sinusoidal wave is also analysed.
从矿热炉对铜瓦的性能要求方面分析了铸造铜瓦的缺陷。	The shortage of contact shoe in ore smelting electric arc furnace is analysed from its performance.
用有限元方法分析了其反射相位特性及表面波传输特性。	The reflection phase property and surface wave dispersion are analyzed using finite element method.

17,3 слів. Автоматичний індекс читабельності ARI корпусу – 10,24, що відповідає розвитку дитини у 16 років. Виявилося, що правий «хвіст» графіка – лексика, яка вживається лише один раз, на великих корпусах становить 50% корпусу.

Ще один метод дослідження корпусу, що хотілося би втілити у життя: визначення найпродуктивніших моделей речень. З цією метою візьмемо корпус розмовної мови на 250 тис. пар речень, пригадаємо його найчастотніші морфеми: 的, 一, 我, 是, 在, 了, 他, 不, 们, 这, 有. За допомогою регулярних виразів спробуємо знайти в корпусі частотні моделі речень, припускаючи, що тут будуть задіяні поєднання частотних морфем (див. табл. 5), тут не враховується послідовність поєднання та повторення морфем.

Фактично отримано такі моделі речень (див. табл. 6). Здобути вичерпний або абсолютний обсяг моделей вручну без застосування спеціально розробленого програмного забезпечення неможливо – занадто багато варіацій живої мови.

Висновки. Таким чином, ми запропонували та перевірили на практиці методику пошуку найпродуктивніших моделей речень у корпусі, які також можуть корелювати з довжиною речень. Показати та дослідити усі можливі моделі речень у межах цієї статті, на жаль, неможливо.

У підсумку можна сказати, що розроблена методика порівняння рангів морфем різних корпусів із морфемами частотного списку СКМ, яка може бути застосована для визначення тематики корпусу чи належності тексту до певної галузі. Також запропоновано методику визначення продуктивних моделей речень корпусу за допомогою регулярних виразів.

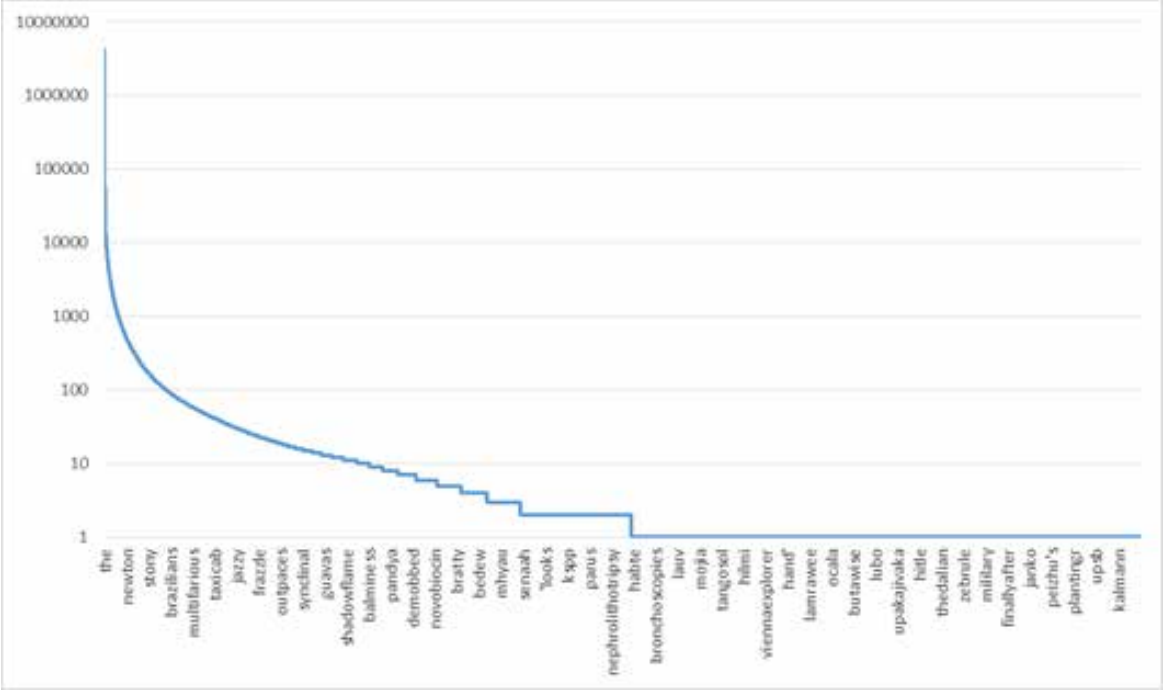

Рис. 4. Логарифмічний графік розподілу лексики об'єднаного корпусу за частотністю

Таблиця 4

Приклади вживання морфеми «压» у політехнічному корпусі

进行了高压共轨柴油机轨道压力控制和匹配研究。	The controlling of rail pressure of high pressure common rail diesel engine is studied.
简要介绍了评价挤压膨化食品营养价值的方法。	The article introduces methods of nutritional appraisal to extruded and expanding foods.
本文就压差换算提出一种方法并举出一个实例。	The article presents a method of pressure difference conversion and provides an example.
有效应力路径随固结压力增大产生偏转。	The deflection of effective stress path occurs with the increase of consolidation pressure.
在电流断开后的一段时间里，可以测量到逐渐下降的电压。	The decaying voltages can be measured for a time after the current is switched off.

Таблиця 5

Теоретично можливі поєднання частотних морфем у реченні

	的	一	我	是	在	了	他	不	们	这
的		+	+	+	+	+	+	+	+	+
一			+	+	+	+	+	+	+	+
我				+	+	+	+	+	+	+
是					+	+	+	+	+	+
在						+	+	+	+	+
了							+	+	+	+
他								+	+	+
不									+	+
们										+
这										

Таблиця 6

Деякі найпродуктивніші моделі речень корпусу розмовної мови

регулярний вираз	**приклади речень**
是.+的.+$	死亡是一种巨大的平等。她实际上真是他的妻子。
了.+的.+$	他脱掉了他的棕色无带鞋。他重新体验了战争的恐怖。他打掉了这艘船上的桅杆。
一.+的	你要到一个奇怪的地方去。这也将是一个艰难的过程。那个水晶花瓶是要做什么的? 我们暂住一个旧的小旅馆。一只肥猫在一个男人的礼帽里。
的.+的.+	黑的像我似的。这里的河流中有大量的鱼。最新的恐慌是有关鸡蛋的。 这件上衣的领子是狐皮的。这么漂亮的一匹小马是我的!
的.+是.+的.+	她的态度是轻率的。他的基因是用于传承的。他的工作是很保密的。
的.+了.+的.+	她的话温暖了他的心。我的鞋子磨痛了我的脚后跟。他的头碰了低矮的天花板。
是.+的.$	他是愿意跟随他到底的。你是如何成为一名摄影师的？许多激光器是用光来激发的。
在.+的.+$	他们住在很偏远的郊区。在那时是一桩很棘手的事。老鹰在它的猎物周围盘旋。
不.+的.+$	不要放弃你的有利条件。你不能改变你的宠物狗的年龄。不会出现更好的时机了。
们.+的.+	他们的棒球投手是一个左撇子。我们有好多的书和玩具。把他们说的话写在横线上。你们没结婚的人太不老实。
的.+了.+	她的眼睛里充满了恐惧。我根据他走路的样子认出了他。我把我的心事告诉了他。
在.+了	你们在黑海中游泳了吗？有人把外衣丢在这儿了。她把海报贴在墙上了。把它随便放在哪儿好了。你在这儿多少年了？他让我在旅店外下了车。
这是.*	这是真的。这是谁的钢笔? 是你的吗? 这是什么？这是一块黑板。
这是.*的	这是事情的一个新的方面。这是一种含有薄荷的药膏。这是一个极其大的区别。 这是万物的天性。这是一个千载难逢的好机会。
地.*了	突然地出现了一只老虎。我们狠狠地惩罚了他们。
的.+不.+	这该死的笔写不出字来。统计的数字加不起来。你的姐姐跟你完全不像。
们.*在.*	你们在黑海中游泳了吗？你知道我们在哪里吃饭吗？
们.*一.*	他们是一群很棒的小伙子。让我们将所有看成一个整体。小孩们在沙堆里挖了一个地道。

Загалом було створено сім різнотематичних корпусів, до яких зведено чотири графічні діаграми, створено зведену таблицю статистичних даних і таблицю частотних морфем корпусів. На основі прикладів вживання морфем «析» та «压» у політехнічному корпусі наочно підтверджено, що корпус дійсно репрезентує саме політехнічну галузь, а методика визначення тематики корпусу чи належності тексту до певної галузі, як і методика скачування корпусу на основі термінологічних списків, є дієвою.

Опрацьовано теоретично можливі поєднання частотних морфем у реченні та виявлено деякі найпродуктивніші моделі речень корпусу розмовної мови.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Глазунов С. А. Новый англо-русский словарь современной разговорной лексики. «Русский язык – Медиа», 2003. 778 с.

2. Козоріз О. П. Статистичні характеристики мовних одиниць юридичної термінології китайської мови. *Вісник Київського національного університету імені Тараса Шевченка. Східні мови та література*. 2014. Вип. 1. С. 15–20. URL: http://nbuv.gov.ua/UJRN/VKNU_Sm_2014_1_6.

3. Ривкин В. Новый англо-русский медицинский словарь. 2004.
4. Столяров Д. Е., Кузьмин Ю.А., Баринов, С. М. Большой англо-русский политехнический словарь : в 2 т. Москва : Руссо, 2003. 1424 с.
5. Чибисова О. И., Смирнов Н. Н., Васецкий С. Г. Новый англо-русский биологический словарь. Москва : Руссо, 2003. 920 с.
6. OilAndGas (En-Ru): Большой англо-русский словарь по нефти и газу. ВНИИГАЗ, РАО «ГАЗПРОМ», 1998. К версии ABBYY Lingvo x3.
7. Jun Da: Modern Chinese Character Frequency List. URL: http://lingua.mtsu.edu/chinese-computing/statistics/char/CharFreq-Modern.xls (дата звернення: 12.02.2021).
8. QuWord. URL: https://www.quword.com/ (дата звернення: 12.02.2021).
9. Word frequency data. URL: https://www.wordfrequency.info/samples.asp (дата звернення: 12.02.2021).

REFERENCES

1. Hlazunov S. A. Noviy anhlo-russkyi slovar sovremennoi razghovornoi leksyky. [New English-Russian dictionary of modern colloquial vocabulary]. “Russian language – Media”, 778 p., 2003 [in Russian].
2. Kozoriz O. P. Statystychni kharakterystyky movnykh odynyts yurydychnoi terminolohii kytaiskoi movy. [Statistical characteristics of the mobile units of the legal terminology of the Chinese language]. Bulletin of the Kiev National University of the Name of Taras Shevchenko. Skhidni movi and literature. 2014. 1. pp. 15–20. URL: http://nbuv.gov.ua/UJRN/VKNU_Sm_2014_1_6 [in Ukrainian].
3. Ryvkyn V. Noviy anhlo-russkyi medytsynskyi slovar. [New English-Russian Medical Dictionary], 2004 [in Russian].
4. Stoliarov, D. E.; Kuzmyn, Yu.A.; Barynov, S.M. Bolshoi anhlo-russkyi polytekhnycheskyi slovar. [The Big English-Russian Polytechnic Dictionary]. 2 vol.: Moskva: Russo; 1424 p; 2003 [in Russian].
5. Chybysova, O. Y; Smyrnov, N.N.; Vasetskyi, S.H. Noviy anhlo-russkyi byolohycheskyi slovar. [New English-Russian Biological Dictionary]. Moskva: Russo; 920 p.; 2003 [in Russian].
6. OilAndGas (En-Ru): Bolshoi anhlo-russkyi slovar po nefty y hazu. [Comprehensive English-Russian Dictionary of Oil and Gas]. VNIIGAZ, RAO “GAZPROM”, 1998. ABBYY Lingvo x3 version [in Russian].
7. Jun Da: Modern Chinese Character Frequency List. URL: http://lingua.mtsu.edu/chinese-computing/statistics/char/CharFreq-Modern.xls (Accessed 12 February 2021).
8. QuWord. URL: https://www.quword.com/ (Accessed 12 February 2021).
9. Word frequency data. URL: https://www.wordfrequency.info/samples.asp (Accessed 11 February 2021).