МОВОЗНАВСТВО

УДК 811.581.11

О. Козоріз, канд. філол. наук, асист.
e-mail: davinci@3g.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

СТВОРЕННЯ ВЛАСНИХ ЛІНГВІСТИЧНИХ КОРПУСІВ

Досліджено проблему створення власних корпусів паралельних текстів великих обсягів. Запропоновано методику та критерії конструювання паралельних лінгвістичних корпусів. У результаті дослідження створено об'єднаний корпус на 3 850 000 пар речень або на 65 мільйонів слів англійської частини, що за обсягом становить 10 % відомого корпусу COCA або корпусу GRAC. Дієвими виявилися методи завантаження матеріалу для корпусу на основі частотного списку, термінологічних 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, автоматичний індекс читабельності ARI, показники середньої довжини речення ASL тощо. Побудова графіків розподілу лексики за частотністю та довжиною речень у корпусі яскраво унаочнює результати наших досліджень, ефективно репрезентує матеріал. Також можна говорити про вдалий досвід створення вузьких спеціалізованих термінологічних корпусів на противагу термінологічним словникам для подальших досліджень саме функціональних особливостей, моделей речень тієї чи іншої терміносистеми. Отримано корпуси медичного і біологічного спрямування (приблизно 500 тис. пар речень кожен), а також політехнічного на 1,3 млн. Загалом було укладено вісім корпусів, для п'яти з них пораховано загальну кількість знаків, слів та речень у корпусі з відповідною узагальнювальною таблицею; встановлено середню довжину речень ASL, визначено автоматичний індекс читабельності ARI, співвідношення type / token ratio TTR; для корпусів 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 через поділ їх на три частини: початкову, середню та хвостову – вважається нами перспективною.

Ключові слова: лінгвістичний корпус, електронний корпус текстів, паралельний корпус, частотний список, ключові слова, type / token ratio, середня довжина речення.

В умовах зростання кількості інформації проблема її опрацювання набуває істотного значення. У цій статті ми хочемо зосередити увагу на дослідженнях на основі корпусів, які стали популярними в багатьох галузях лінгвістики. Сьогодні спостерігається значний інтерес до використання корпусів в освітній та професійній сферах. Значна кількість словників укладається на основі подібних лінгвістичних корпусів. Актуальності представленій розвідці додає активізація процесу створення паралельних корпусів великих обсягів із навчальною, дослідницькою та прикладною метою. Основними напрямами використання корпусу паралельних текстів є такі: 1) з навчальною та дослідницькою метою, аби перевірити, чи використовується певна лінгвістична форма в мові у певному контексті; 2) для створення систем машинного перекладу – сучасні нейронні мережі сьогодні використовують як альтернативу всім наявним алгоритмам для машинного перекладу.

Задля досягнення мети, а саме створення власних паралельних корпусів, слід виконати такі завдання: визначити джерела лінгвістичного матеріалу; створити власний корпус; розглянути основні характеристики створеного корпусу.

Розглянемо ключові поняття. Т. МакЕнері та Е. Вілсон подають таке визначення: корпус – це зібрання мовних фрагментів, відібраних відповідно до чітких мовних критеріїв для використання як моделі мови [McEnery, Wilson, 2001].

Лінгвістичний корпус – це достатньо велика колекція мовних даних, відібраних за певним упорядкувальним принципом, які зберігаються в електронному вигляді. Головним атрибутом корпусу є його репрезентативність, зумовлена домінантою, що об'єднує всі тексти єдиним логічним задумом.

Співвідношення type / token ratio (TTR) – співвідношення між типами та лексемами корпусу, дуже сильно варіюється відповідно до довжини тексту або корпусу текстів. Якщо текст завдовжки 1000 слів, кажуть, що він має 1000 "лексем" (word token). Але багато цих слів будуть повторюватися, і в тексті може бути лише 400 різних слів. Отже, "типи" – це різні слова, у цьому прикладі їх буде 40%. Чим довший текст, тим менший буде відсоток.

Автоматичний індекс читабельності (англ. Automated readability index (ARI)) – 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 * (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, що 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 в складах (англ. average number of syllables per word) = syllables / words.

У 2018 р. у своїй статті О. Скобнікова вживає термін "ключові слова" і тлумачить його так: "ключові слова – обчислюються шляхом порівняння частотності слова в досліджуваному корпусі з частотністю слова в корпусі довідковому (референтному, або опорному). У результаті порівняння частоти слововживань система привласнює одиницям, що відрізняються несподівано високою частотністю, статистичну назву ключового слова" [Скобнікова, 2018]. Дещо схоже, "ключові морфеми юридичного дискурсу", у своїй статті 2014 р. на практичному матеріалі досліджували і ми, називаючи їх також "ядерною зоною" [Козоріз, 2014]. Утім, технологія порівняння частотних списків не нова і досліджувалася за кордоном уже давно, відома під назвою "Матричний метод: статистичне порівняння частотних профілів" [Rayson, 2002].

Інший дослідник В. Бісікало, посилаючись на дослідження О. Каніщевої від 2014 р., у своїй статті згадує гіперболічну залежність розподілу Парето, що дає змогу під час аналізу тексту виявити три групи слів: "початковий пік – найбільш частотні та короткі стоп-слова, що мають службове значення; ключові або найзначиміші з огляду на зміст тексту слова (середня частина гіперболи, оптимальний вибір ширини якої дозволяє досягти найвищої якості ключових слів); хвостова частина – слова, що зустрічаються рідко та майже не характеризують зміст тексту" [Бісікало, 2015]. Хоча автор у своїй статті не підкріплює цю думку практичними дослідженнями, але така гіпотеза

розподілу лексики частотного словника тексту на три частини: початкову, середню та хвостову – вважається нами перспективною і буде застосована на практиці.

Деякі дослідники також зазначають обов'язкову наявність у корпусі металінгвістичної інформації [Захаров, 2011, с. 7]. Виділяють зовнішню або ж екстралінгвістичну розмітку (автор, назва, дата, обсяг, жанр, тематика тексту тощо), структурну (розділ, абзац, речення, словоформа), лінгвістичну (лексичні, граматичні та інші характеристики елементів тексту) [Жуковська, 2013, с. 76]. Однак для сучасних наукових потреб, оскільки така розмітка потребує суттєвої затрати робочого часу, може вважатися опціональною, як на нашу думку.

Розглянемо деякі провідні корпуси: The Corpus of Contemporary American English – COCA [Corpus of Contemporary…], який є найбільшим корпусом американського варіанту англійської мови. Корпус складається з понад 600 млн слів, містить тексти, які систематично добираються з 1990 р. до сьогодні, рівномірно репрезентуючи усне мовлення, художню прозу, популярні журнали, газети та наукову літературу – щороку відбирається однаковий обсяг текстів (4 млн слів) із кожної тематики.

Інший відомий корпус англійської мови Oxford English Corpus пропонує такі можливості для досліджень: **word sketch** – сполучення слів, класифіковані за граматичними відношеннями; **thesaurus** – синоніми та подібні слова до кожного слова; **keywords** – термінологічне вилучення односкладових та багатоскладових одиниць; **word lists** – списки англійських іменників, дієслів, прикметників тощо, впорядковані за частотністю; **n-grams** – перелік частот багатоскладових одиниць; **concordance** – приклади в контексті; **trends** – діахронічний аналіз автоматично ідентифікує неологізми та зміни у використанні [Oxford English Corpus].

Для дослідження української мови пропонується Генеральний регіонально анотований **корпус української мови** (ГРАК; англ. General Regionally Annotated Corpus of Ukrainian, GRAC) обсягом понад 650 млн токенів, може бути використаний для здійснення лінгвістичних досліджень, а також під час укладання словників та граматик [Генеральний…].

Відомий Корпус китайської мови CCL 语料库检索系统 пропонує користувачу автоматичну **сегментацію китайського тексту**. Виглядає це так: 一些/m 人/n 喜欢/v 加入/v 桂皮/n 和/c 其它/r 调味品/n ，/w 以/p 增加/v 风味/n 。/w [语料库在线网站]. Також він містить програму підрахунку частотних характеристик корпусу користувача.

Є Корпус британської англійської мови [British National Corpus]; Національний корпус російської мови [Национальный…]; Чеський національний корпус [Czech National Corpus] тощо [Corpus Survey]. На основі корпусів створюють **словники сполучуваності слів**: Оксфордський словник сполучуваності [English Collocations…], безкоштовний онлайн-словник сполучуваності [ProWritingAid] тощо.

Іншим мірилом, що розрізняє типи корпусів, є кількість мов, представлених у корпусі. Окрім корпусів однією мовою (монокорпуси), створюються **багатомовні корпуси**, що побудовані на матеріалі двох або більше мов. Багатомовні корпуси так само поділяються на порівняльні та паралельні корпуси. **Паралельний корпус** – єдність підмножини оригінальних текстів та підмножини їхніх перекладів іншою мовою [Демська, 2011, с. 90]; корпус, який складається з як мінімум двох підкорпусів, один із яких є вихідним, а інший містить тексти-переклади вихідного корпусу [Жуковська, 2013, с. 24, 136], або велике зібрання паралельних текстів. Також існує поняття **паралельний текст** або ж **бітекст** – текст однією мовою та поряд його переклад іншою. Відомі паралельні корпуси: Russian-Chinese Translation Corpus [Russian-Chinese Translation Corpus]; Glosbe [Glosbe]; Linguee [Англо-русский словарь и система…]; MyMemory [MyMemory]; Opus [Open parallel corpus OPUS]; Reverso [Reverso context]; TAUS Data Cloud [TAUS Data Cloud] тощо [Corpus Survey].

Процес створення паралельного корпусу зазвичай відбувається або автоматичними системами сегментування, або ж людиною. Проблеми, притаманні таким корпусам, вмотивовані людським чинником: недосконалість текстів, з яких складаються корпуси, низька лінгвістична якість текстів оригіналу, наявність помилок перекладу, неправильне вирівнювання правої перекладної частини, невідповідність зазначених мов корпусу (як права, так і ліва частини), використання неавтентичних текстів як оригінальних під час створення паралельних корпусів (досить велика кількість текстів у мережі є продуктами машинного перекладу) тощо.

За нашими спостереженнями, останнім часом у мережі з'явилася велика кількість текстів, продукованих машиною, перекладених автоматично. Цілеспрямовано створений **механічний паралельний корпус** – це паралельний корпус, перекладений машиною на основі підготовленого монокорпусу. Такий корпус можна використовувати з подальшим редагуванням людиною для тих самих цілей, що і звичайні паралельні корпуси. Основною його вадою буде наявність саме системних помилок перекладу, оскільки машина перекладає завжди однаково прогнозовано. Натомість не буде вад, притаманних корпусам, що створені людиною: низька лінгвістична якість текстів перекладу, вільний переклад, неправильне вирівнювання правої перекладної частини, невідповідність зазначених мов корпусу тощо.

Основне завдання під час створення корпусу із зазначеною вище метою – це збалансовано відібрати матеріал для дослідження. На нашу думку, репрезентативним буде відбір речень за двома критеріями: лексичним і граматичним. Лексичний полягає в доборі речень, що містять слова за частотним списком мови. З граматичним критерієм дещо складніше, оскільки відсутній частотний список граматичних моделей речень. На жаль, жоден підручник із граматики не подає абсолютно всіх можливих моделей речень, за якими слова можуть поєднуватися, та навіть не вказує їхньої частотності.

Критерієм, що корелює з частотністю граматичних моделей речень, може бути довжина речення у поєднанні з частотністю кожного слова цього речення. Речення короткої і середньої довжини будуть траплятися частіше в усному мовленні, натомість художню прозу і популярні журнали репрезентують середні та довгі речення; самі лише довгі речення – характерні для публіцистичних і наукових текстів.

Теоретично процес та особливості створення власного корпусу описує у своїй фундаментальній праці В. Жуковська [Жуковська, 2013, с. 85–87]. Через значний обсяг вибірки текстів природної мови роботу з корпусом реалізують за допомогою спеціальних програмних засобів – **конкордансерів** або **корпусних менеджерів**, наприклад: Wordsmith, MonoConc, AntConc [AntConc Homepage], Xiara тощо.

Ми створили свій власний паралельний корпус китайсько-англійських перекладів на основі сайту-словника https://www.quword.com/ [QuWord]. Створення паралельних корпусів у площині англійська – китайська мови, а не українська – китайська мови, визначена наявністю у мережі великих обсягів якісного матеріалу двосторонніх перекладів саме цього напряму. Українсько-китайський к

орпус може бути створений уже на базі готових англо-китайських корпусів за допомогою автоматичного перекладу гуглом як механічний паралельний корпус для подальшого редагування перекладів людиною, оскільки наразі його створення, на нашу думку, унеможливлюється необхідністю суттєвих витрат людських ресурсів.

Першим кроком було створення 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; інформація щодо частотного списку на 60 і 220 тисяч слів дається фрагментарно – лише кожне п'яте слово [Word frequency data].

Через завантаження 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 складення частотного списку цього корпусу, отримано словник-список на 42 000 слів, побудовано відповідний логарифмічний графік лексики корпусу за частотністю (див. рис. 1).

Загальний обсяг лексики словника – 42 000 слів; 40 % яких (17 000) вживаються у корпусі лише один раз – правий "хвіст" графіка; середня частина графіка – частоти від 10 до 2-х – займає приблизно 40 % лексики (17 000); найчастотнішими є перші 20 % слів (8 000). TTR корпусу – 2,9 % (42 000 / 1 462 000). Середня довжина речення становить 14,6 слова. Автоматичний індекс читабельності ARI корпусу – 7,3, що відповідає 13-річному віку.

На основі словника-списку попереднього корпусу 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 приблизно 28 % лексики (47 000); найчастотнішими є перші 26 % слів (43 000). TTR корпусу – 1,3 % (166 000 / 12 900 000). Середня довжина речення становить 14 слів. Автоматичний індекс читабельності ARI корпусу – 8.

Як бачимо на рис. 2, така технологія гарно збалансовує корпус лексично, урізноманітнюючи ілюстративний матеріал частотної лексики, зріс початок графіка, звузилася середина, але також дещо подовжився "хвіст", що очевидно характерно для великих корпусів.

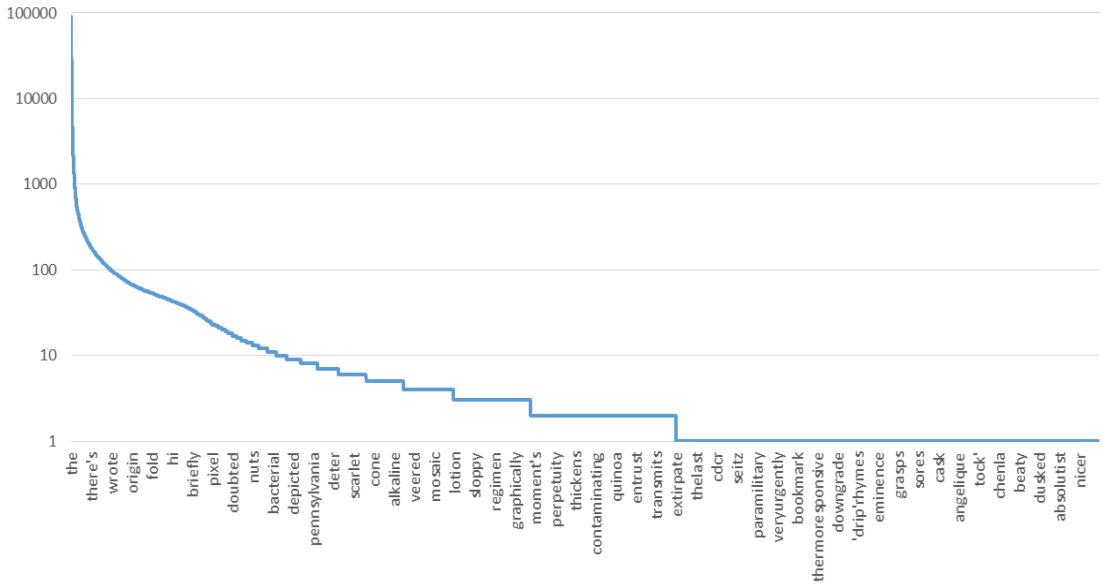

Рис. 1. Логарифмічний графік розподілу лексики корпусу за частотністю на 106 000

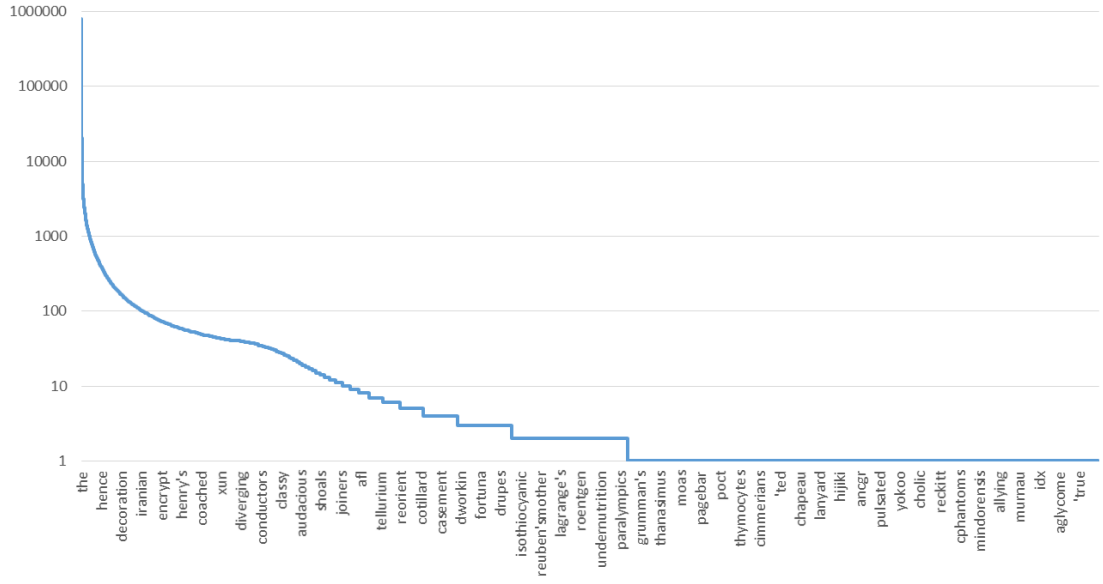

Рис. 2. Логарифмічний графік розподілу лексики корпусу за частотністю на 920 000

Окремо з корпусом на 920 000 було проведено експеримент щодо зменшення "хвоста" корпусу, тобто з корпусу було вилучено речення, що містять нечастотну лексику – частоти від 1 до 10 входжень. Як наслідок отримано корпус до 710 000 пар речень або 9 289 000 лексем англійської частини, який має словник усього на 44 000 слів, тобто вчетверо менше за початковий; 7,9 % яких (3500) вживаються лише один раз – правий "хвіст" графіка; середня частина графіка – частоти від 10 до 2-х – займає приблизно 7,5 % лексики (3 300); найчастотнішими є перші 85 % слів (37 700). TTR корпусу – 0,47 % (44 000 / 9 289 000). Середня довжина речення становить 13 слів. У такий спосіб ми отримали найбільш лексично збалансований корпус. Як видно на рис. 3, корпус має лише невеличкий правий "хвіст" і таку само маленьку середину. Автоматичний індекс читабельності ARI корпусу – 7,1.

Далі було проведено завантаження на основі частотного списку китайської мови на 56 000 слів отримано корпус ще приблизно на 1 000 000 унікальних пар речень. Також було виконано завантаження на основі списків-слів медичного [Новый англо-русский медицинский…, 2004], біологічного [Новый англо-русский биологический…, 2003] та політехнічного [Большой англо-русский политехнический…, 2003] термінологічних словників. У такий спосіб отримано корпуси приблизно 500 000 пар речень кожний для перших двох і 1 300 000 для останнього. Дослідження цих корпусів буде виконано окремою роботою.

Після об'єднання та видалення повторів загальний обсяг отриманих паралельних корпусів сягає 3 500 000 пар речень або 59 700 000 лексем. Було побудовано частотний список лексики англійської частини цього об'єднаного корпусу та відповідний логарифмічний графік. Спостереження виявили таке: загальний обсяг лексики – 410 000 слів; 50 % яких (200 000) вживаються лише один раз – правий хвіст графіка; середня частина графіка – частоти від 10 до 2-х – займає приблизно 30 % лексики (120 000); найчастотнішими є перші 20 % слів (80 000). TTR корпусу – 0,69 % (410 000 / 59 700 000). Середня довжина речення становить 17 слів.

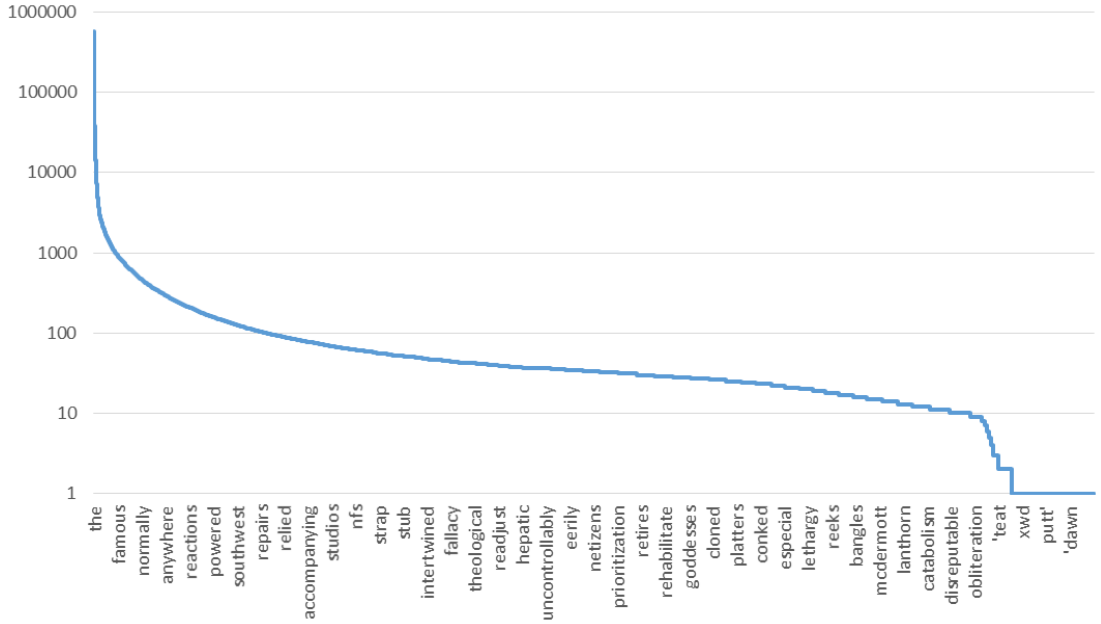

Рис. 3. Логарифмічний графік розподілу лексики корпусу за частотністю на 710 000

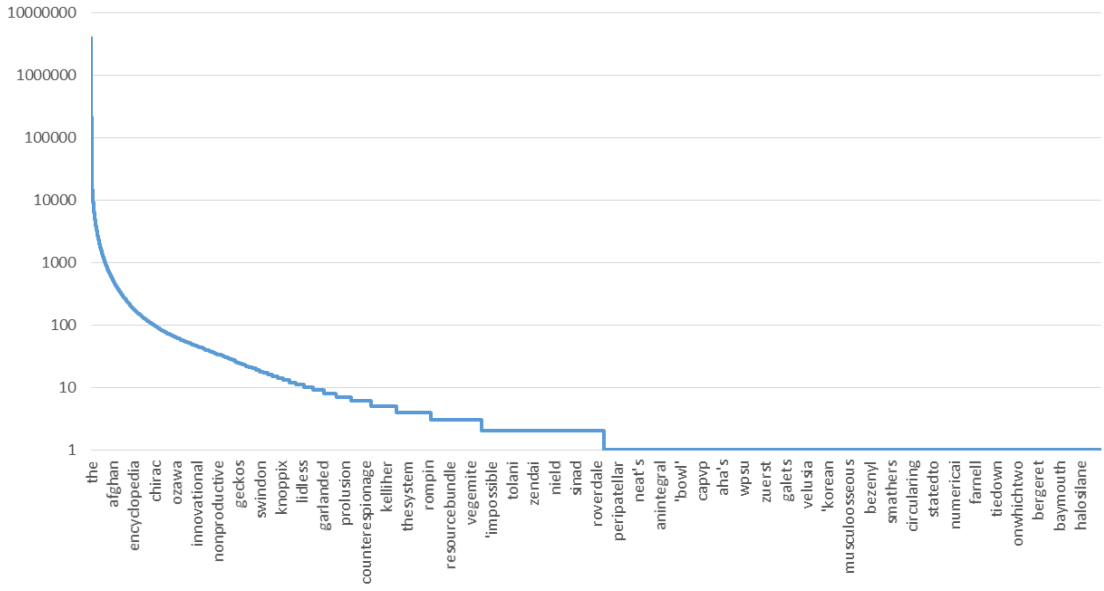

Рис. 4. Логарифмічний графік розподілу лексики об'єднаного корпусу за частотністю на 3 500 000

Спробуємо підняти початок та середню частину корпусу через додаткове завантаження матеріалу на основі списку слів із частотами від 10 до 2-х попереднього корпусу. Після додаткового завантаження і додавання матеріалу до вже наявного корпусу отримано загальний корпус на 3 850 000 пар речень або 65 200 000 лексем. Його аналіз показав таке: загальний обсяг різної лексики – 442 000 слів; приблизно 50 % яких (217 000) вживаються лише один раз – правий "хвіст" графіка; середня частина графіка – частоти від 10 до 2-х – займає приблизно 26 % лексики (117 000); найчастотнішими є перші 24 % слів (108 000). TTR корпусу – 0,67 % (442 000 / 65 200 000). Середня довжина речення становить 17 слів. Автоматичний індекс читабельності ARI корпусу – 10,32, що відповідає розвитку дитини в 16 років.

Як бачимо, технологія вирівнювання ілюстративного матеріалу гарно працює навіть у межах одного ресурсу, ми можемо підтягнути вгору початок та середню частину лексики корпусу, зменшивши кількість лексики з малою кількістю вживань. Натомість виявилося, що правий "хвіст" графіка на великих корпусах лишається майже сталою величиною і складає 50 % лексики корпусу. Лексика з частотою 1 – це або помилки, або дуже рідкісні слова, що не представлені на ресурсі. Ймовірно, що користуючись кількома ресурсами для завантаження одночасно, можна суттєво підняти початок та середню частину корпусу і майже уникнути малих величин вживання слів у корпусі; також можна використати метод відтинання правого "хвоста" графіка, тобто вилучити із корпусу всі речення, що містять слова з малою частотою вжитку.

Загалом було складено вісім корпусів та досліджено п'ять із них. Відповідно виявлено такі закономірності: зі зростанням величини корпусу, знижується TTR корпусу. Цікавим є факт, що третій штучно обрізаний корпус дає показники ARI, TTR та інші показники, які порушують загальну тенденцію (див. табл. 1). Очевидно, що корпус № 3 є збалансованішим із лексичного погляду.

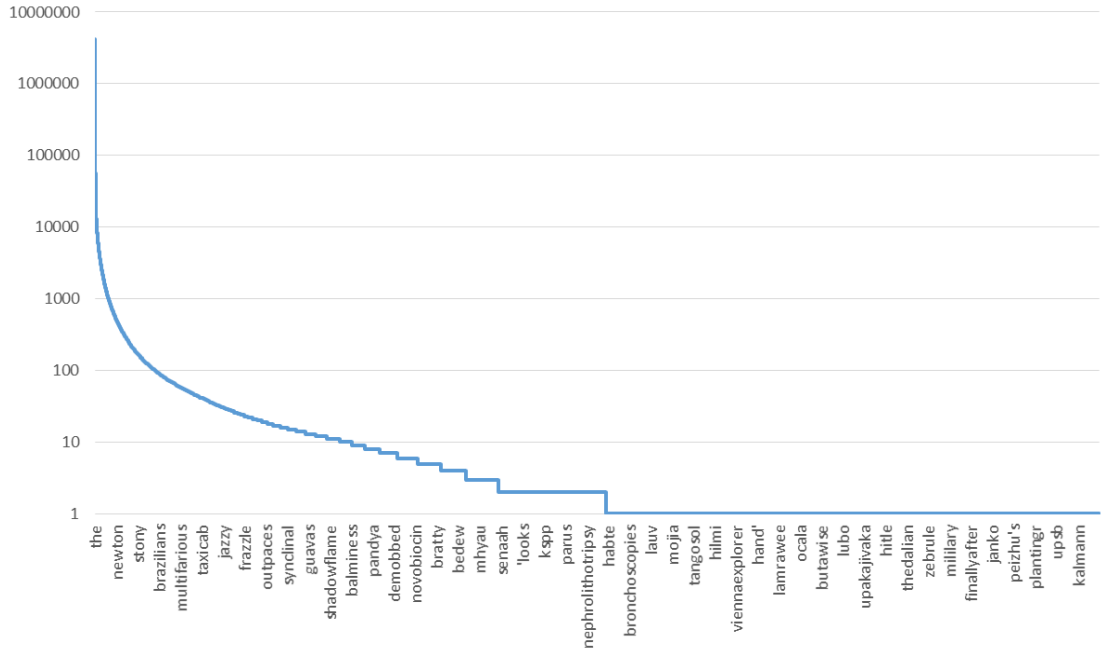

Рис. 5. Логарифмічний графік розподілу лексики об'єднаного корпусу за частотністю на 3 850 000

Таблиця 1

Статистичні дані створених корпусів

graph	characters	words	sentences	ARI	TTR	ASL	початок	середина (від 10 до 2-х)	кінець
1	6 632 000	1 462 000	100 000	7,245746	2,9 %	14,6	20 %	40 %	40 %
2	61 372 000	12 900 000	920 000	7,988786	1,3 %	14	26 %	28 %	46 %
3	43 465 000	9 289 000	710 000	7,150536	0,47 %	13	85 %	7,9 %	7,5 %
4	292 666 000	59 700 000	3 464 000	10,27694	0,69 %	17	20 %	30 %	50 %
5	319 840 000	65 200 000	3 770 000	10,32222	0,67 %	17	24 %	26 %	50 %

Автоматичний індекс читабельності ARI можна критикувати для корпусів змішаного типу, оскільки він абсолютно не враховує семантичного навантаження лексики, а спирається виключно на довжину речень і слів. Так, в останньому зведеному корпусі широко представлені тексти медичної, біологічної та технічної тематики, що аж ніяк не можуть відповідати розвитку дитини в 16 років, але, ймовірно, це середній показник щодо корпусу, який враховує як прості, так і складні його частини.

Тому далі було пораховано довжину речень у словах для кількох корпусів та побудовано відповідні графіки. Пошук кількості слів у реченні здійснювався регулярним виразом: ^(\w+\W+){X}$ – де X, кількість слів у реченні, словосполучення тут не враховані, оскільки наприкінці виразу обов'язково має стояти розділовий знак. Як бачимо з рис. 6, у штучно обрізаному корпусі № 3 – нижчий графік, що був збалансований лексично, відбулося зменшення кількості речень за довжиною – звуження і зниження верхньої частини графіка, що вірогідно зменшить різноманіття ефективних моделей речень, тому така корекція корпусу виглядає досить сумнівно; натомість збільшення корпусу додатковим завантаженням нечастотної лексики дає позитивний загальний результат, середній графік. Найзбалансованішим за довжиною речень виглядає верхній графік корпусу № 5 на 3 850 000, тут широко представлені речення з довжинами від 6-и до 33-х слів, а загальний обсяг різної лексики складає 442 тис. слів.

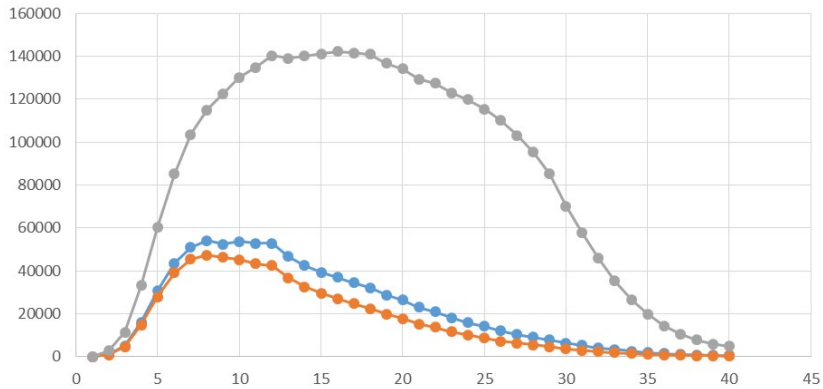

Рис. 6. Графіки розподілу кількості слів у реченні для корпусів на 710 000, 920 000 та 3 850 000 речень

Унаслідок дослідження нами одноосібно було створено об'єднаний корпус на 3 850 000 пар речень або на 65 млн слів (англійська частина), що за обсягом становить 10 % відомого корпусу COCA або корпусу GRAC. Проведено теоретичні розвідки та практичні дослідження задля нормалізації корпусу. Ефективними для дослідження корпусу виявилися показники ARI, TTR, ASL тощо. Найважливішими є графіки розподілу лексики корпусу за частотністю та довжиною речень у корпусі. Також можна говорити про вдалий досвід створення вузьких спеціалізованих термінологічних корпусів на противагу термінологічним словникам для подальших досліджень саме функціональних особливостей (моделі речень) тієї чи іншої терміносистеми.

Список використаних джерел

Бісікало О. В. Статистичний аналіз складних залежностей у тексті. *Вісник Національного університету "Львівська політехніка"*. Серія : інформаційні системи та мережі. 2015. № 814. С. 228–236.

Большой англо-русский политехнический словарь : в 2 т. / Д. Е. Столяров та ін. Москва : РУССО, 2003. 1424 с.

Демська О. Текстовий корпус : ідея іншої форми. Київ : НаУКМА, 2011. 282 с.

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

Захаров В. П., Богданова С. Ю. Корпусная лингвистика : учеб. для студентов гуманитарных вузов. Иркутск : ИГЛУ, 2011. 161 с.

Козоріз О. П. Статистичні характеристики мовних одиниць юридичної термінології китайської мови. *Вісник Київського національного університету імені Тараса Шевченка. Східні мови та літератури*. 2014. Вип. 1(20). С. 15–20.

Новый англо-русский биологический словарь / О. И. Чибисова, Н. Н. Смирнов, С. Г. Васецкий. Москва : РУССО, 2003. 920 с.

Новый англо-русский медицинский словарь / В. Л. Ривкин, М. С. Бенюмович. Москва : РУССО, 2004. 876 с.

Скобнікова О. В. Створення власного корпусу американських кіносценаріїв. *Науковий вісник Дрогобицького державного педагогічного університету імені Івана Франка*. Серія : філологічні науки (мовознавство). 2018. № 9. С. 204–207.

McEnery T., Wilson A. Corpus Linguistics: An Introduction. 2nd ed. Edinburg : Edinburg University press, 2001. 235 p.

Rayson P. E. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison : thesis … Ph. D. / Lancaster University, 2002. 208 p.

Джерела ілюстративного матеріалу

Англо-русский словарь и система контекстуального поиска по переводам Linguee. URL : http://linguee.ru (дата обращения : 11.02.2021).

Генеральний регіонально анотований корпус української мови (ГРАК). URL : http://uacorpus.org/ (дата звернення : 11.02.2021).

Национальный корпус русского языка. URL : http://ruscorpora.ru/new/ (дата обращения : 11.02.2021).

AntConc Homepage. URL : http://www.laurenceanthony.net/software/antconc/ (Last accessed : 11.02.2021).

British National Corpus. URL : http://www.natcorp.ox.ac.uk/ (Last accessed : 11.02.2021).

Corpus of Contemporary American English. URL : https://www.english-corpora.org/coca/ (Last accessed : 11.02.2021).

Corpus software and related tools. URL : http://ucrel.lancs.ac.uk/tools.html (Last accessed : 11.02.2021).

Corpus Survey. URL : https://www.lancaster.ac.uk/fass/projects/corpus/cbls/corpora.asp (Last accessed : 11.02.2021).

Czech National Corpus. URL : http://web.archive.org/web/20131029222327 http://ucnk.ff.cuni.cz/english/stahni.php (Last accessed : 11.02.2021).

English Collocations Dictionary online. URL : http://ozdic.com/collocation-dictionary/ (Last accessed : 11.02.2021).

Glosbe (багатомовний онлайн-словник). URL : http://glosbe.com (Last accessed : 11.02.2021).

MyMemory. URL : http://mymemory.translated.net (Last accessed : 11.02.2021).

Open parallel corpus OPUS. URL : http://opus.lingfil.uu.se (Last accessed : 11.02.2021).

Oxford English Corpus. URL : https://www.sketchengine.co.uk/oxford-english-corpus (Last accessed : 11.02.2021).

ProWritingAid. URL : https://prowritingaid.com/ (Last accessed : 11.02.2021).

QuWord. URL : https://www.quword.com/ (Last accessed : 11.02.2021).

Reverso context. URL : https://context.reverso.net (Last accessed : 11.02.2021).

Russian-Chinese Translation Corpus. URL : http://www.rucorpus.cn/ (Last accessed : 11.02.2021).

TAUS Data Cloud. URL : http://data-app.taus.net (Last accessed : 11.02.2021).

Word frequency data. URL : https://www.wordfrequency.info/samples.asp (Last accessed : 11.02.2021).

语料库在线网站。 URL : http://corpus.zhonghuayuwen.org/CpsWParser.aspx (登录时间 : 11.02.2021).

References

Bisikalo, O. V., 2015. Statystychnyi analiz skladnykh zalezhnostei u teksti. *Visnyk Natsionalnoho universytetu "Lvivska politekhnika". Informatsiini systemy ta merezhi*, 814, pp. 228–236.

Chibisova, O. I., Smirnov, N. N. and Vaseckij, S. G., 2003. *Novyj anglo-russkij biologicheskij slovar'*. Moscow: Russo.

Demska, O., 2011. *Tekstovyi korpus: ideia inshoi formy*. Kyiv: NaUKMA.

Kozoriz, O. P., 2014. Statystychni kharakterystyky movnykh odynyts yurydychnoi terminolohii kytaiskoi movy. *Bulletin of Taras Shevchenko National University of Kyiv. Oriental languages and literatures*, 1(20), pp. 15–20.

McEnery, T. and Wilson, A., 2001. *Corpus Linguistics: An Introduction*. 2nd ed. Edinburg: Edinburg University press.

Rayson, P. E., 2002. *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph. D. Lancaster University.

Rivkin, V. L. and Benjumovich, M. S., 2004. *Novyj anglo-russkij medicinskij slovar'*. Moscow: Russo.

Skobnikova, O. V., 2018. Stvorennia vlasnoho korpusu amerykanskykh kinostsenariiv. *Naukovyi visnyk Drohobytskoho derzhavnoho pedahohichnoho universytetu imeni Ivana Franka. Filolohichni nauky (movoznavstvo)*, 9, pp. 204–207.

Stoljarov, D. E. et al., 2003. *Bol'shoj anglo-russkij politehnicheskij slovar'*. Moscow: Russo.

Zaharov, V. P. and Bogdanova, S. Ju., 2011. *Korpusnaja lingvistika*. Irkutsk: IGLU.

Zhukovska, V. V., 2013. *Vstup do korpusnoi linhvistyky*. Zhytomyr: Zhytomyrskyi derzhavnyi universytet imeni Ivana Franka.

Sources

Anglo-russkij slovar' i sistema kontekstual'nogo poiska po perevodam Linguee, [online]. Available at: <http://linguee.ru> [Accessed 11 February 2021].

AntConc Homepage, [online]. Available at: <http://www.laurenceanthony.net/software/antconc/> [Accessed 11 February 2021].

British National Corpus, [online]. Available at: <http://www.natcorp.ox.ac.uk/> [Accessed 11 February 2021].

Corpus of Contemporary American English, [online]. Available at: <https://www.english-corpora.org/coca/> [Accessed 11 February 2021].

Corpus software and related tools, [online]. Available at: <http://ucrel.lancs.ac.uk/tools.html> [Accessed 11 February 2021].

Corpus Survey, [online]. Available at: <https://www.lancaster.ac.uk/fass/projects/corpus/cbls/corpora.asp> [Accessed 11 February 2021].

Czech National Corpus, [online]. Available at: <http://web.archive.org/web/20131029222327/http://ucnk.ff.cuni.cz/english/stahni.php> [Accessed 11 February 2021].

English Collocations Dictionary online, [online]. Available at: <http://ozdic.com/collocation-dictionary/> [Accessed 11 February 2021].

Glosbe (multilingual online dictionary), [online]. Available at: <http://glosbe.com> [Accessed 11 February 2021].

Heneralnyi rehionalno anotovanyi korpus ukrainskoi movy (HRAK), [online]. Available at: <http://uacorpus.org/> [Accessed 11 February 2021].

MyMemory, [online]. Available at: <http://mymemory.translated.net> [Accessed 11 February 2021].

Nacional'nyj korpus russkogo jazyka, [online]. Available at: <http://ruscorpora.ru/new/> [Accessed 11 February 2021].

Open parallel corpus OPUS, [online]. Available at: <http://opus.lingfil.uu.se> [Accessed 11 February 2021].

Oxford English Corpus, [online]. Available at: <https://www.sketchengine.co.uk/oxford-english-corpus> [Accessed 11 February 2021].

ProWritingAid, [online]. Available at: <https://prowritingaid.com/> [Accessed 11 February 2021].

QuWord, [online]. Available at: <https://www.quword.com/> [Accessed 11 February 2021].

Reverso context, [online]. Available at: < https://context.reverso.net> [Accessed 11 February 2021].

Russian-Chinese Translation Corpus, [online]. Available at: <http://www.rucorpus.cn/> [Accessed 11 February 2021].

TAUS Data Cloud, [online]. Available at: <http://data-app.taus.net> [Accessed 11 February 2021].

Word frequency data, [online]. Available at: <https://www.wordfrequency.info/samples.asp> [Accessed 11 February 2021].

Yuliaokuzaixianwangzhan, [online]. Available at: <http://corpus.zhonghuayuwen.org/CpsWParser.aspx> [Accessed 11 February 2021].

Надійшла до редколегії 11.02.21

O. Kozoriz, Ph. D. in Philology, teaching assist.
e-mail: davinci@3g.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

CREATING OWN LINGUISTIC CORPORA

The problem of creation of own corpora of parallel texts of large volumes is investigated in the article. The technique and criteria of construction of parallel linguistic corpora are offered. As a result of our research, we created a combined corpus of 3,850,000 pairs of sentences or 65 million words of the English part, which is 10 % of the known COCA corpus or GRAC corpus. Methods for downloading material for the corpus based on the frequency list, terminological dictionaries, as well as frequency lists of words of previously self-created corpora proved to be effective. Theoretical investigations and practical researches for normalization of the corpus are carried out. The type / token ratio, the automatic readability index ARI, the average sentence length ASL, etc. were effective for the study of the corpus. The construction of graphs of the distribution of vocabulary by frequency and length of sentences in the corpus clearly illustrates the results of our research, effectively represents the material. We can also talk about the successful experience of creating narrow specialized terminological corpora as opposed to terminological dictionaries for further research of functional features, sentence models of a particular terminological system. Medical and biological corpora (about 500 thousand pairs of sentences each), as well as a polytechnic corpora for 1.3 million were compiled. A total of eight corpora were compiled, for five of them the total number of characters, words and sentences in the corpus with the corresponding summary table was calculated; the average length of sentences ASL was determined, the automatic readability index ARI were determined, the ratio type/token ratio TTR was calculated. For each corpora frequency lists of vocabulary are made, the total amount of unique vocabulary is calculated and the corresponding logarithmic graphs are constructed; proposed method of analysis of the distribution of vocabulary of the frequency dictionary of the text on the basis of graphs by dividing them into three parts: initial, middle and tail – is considered promising for us.

Keywords: linguistic corpus, electronic corpus of texts, parallel corpus, frequency list, keywords, type / token ratio, average sentence length.

А. Козорез, канд. филол. наук, ассист.
e-mail: davinci@3g.ua
Киевский национальный университет имени Тараса Шевченко, Киев, Украина

СОЗДАНИЕ СОБСТВЕННЫХ ЛИНГВИСТИЧЕСКИХ КОРПУСОВ

Исследована проблема создания собственных корпусов параллельных текстов больших объемов. Предложена методика и критерии конструирования параллельных лингвистических корпусов. В результате исследования был создан объединенный корпус на 3850000 пар предложений или на 65000000 слов английской части, по объему составляет 10 % видимого корпуса COCA или корпуса GRAC. Действенными оказались методы загрузки материала для корпуса на основе частотного списка, терминологических словарей, а также частотных списков слов предварительно самостоятельно созданных корпусов. Проведены теоретические разведки и практические исследования для нормализации корпуса. Результативным для исследования корпуса оказались соотношение type / token ratio, автоматический индекс читабельности ARI, показатели средней длины предложения ASL и тому подобное. Построение графиков распределения лексики по частоте и длине предложений в корпусе ярко иллюстрирует результаты исследования, эффективно представляет материал. Также можно говорить об удачном опыте создания узких специализированных терминологических корпусов в противовес терминологическим словарям для дальнейших исследований именно функциональных особенностей, моделей предложений той или иной терминосистемы. Получено корпуса медицинского и биологического направлений (около 500 тыс. пар предложений каждый), а также политехнического на 1,3 млн. Всего было создано восемь корпусов, для пяти из них посчитано общее количество знаков, слов и предложений в корпусе с соответствующей обобщающей таблицей; установлено среднюю длину предложений ASL, определен автоматический индекс читабельности ARI, соотношение type / token ratio TTR; для корпусов составлены частотные списки лексики, посчитано общее количество уникальной лексики и построены соответствующие логарифмические графики; предложена методика анализа распределения лексики частотного словаря текста на основе графиков путем разделения их на три части: начальную, среднюю и хвостовую – считается нами перспективной.

Ключевые слова: лингвистический корпус, электронный корпус текстов, параллельный корпус, частотный список, ключевые слова, type / token ratio, средняя длина предложения.