ПЕРЕКЛАДОЗНАВСТВО

УДК 81’33
DOI: https://doi.org/10.32589/2311-0821.2.2022.274929

Я. В. Капранов
Київський національний лінгвістичний університет, Україна
e-mail: yan.kapranov@knlu.edu.ua
ORCID ID: https://orcid.org/0000-0003-2915-038X

Т. В. Тронь
Київський національний лінгвістичний університет, Україна
e-mail: tetiana.tron@knlu.edu.ua
ORCID ID: https://orcid.org/0000-0003-0266-8461

Б. О. Івановська
Економіко-гуманітарний університет у Варшаві, Республіка Польща
e-mail: b.iwanowska@vizja.pl
ORCID ID: https://orcid.org/0000-0003-1331-2866

КОРПУСНИЙ ІНСТРУМЕНТАРІЙ OPUS ДЛЯ ЗАБЕЗПЕЧЕННЯ ІНТЕЛЕКТУАЛЬНОГО ПЕРЕКЛАДУ
(на прикладі текстів L1 і L2 англійсько-українського кінодискурсу)

Abstract
The article explains the concept of “translation memory” and defines it as a computer database where segments of texts of different L1 discourses are represented, as well as equivalents of these segments in L2. Computer-Aided Translation, Machine Translation and Parallel corpus toolkit are outlined as the main types of translation memory. In particular, Computer-Aided Translation is considered as the process of translating L1 text to obtain L2 by using specialized computer software. In this way, the human factor plays one of the most important missions in the process of performing Computer-Aided Translation, because the L1 text is subjected to three types of processing: pre-, inter- and post-editing. Machine Translation is viewed in a narrow sense as the process of translating a text from L1 to L2, that is performed by a computer in whole and/or in part, and in a broad sense as a branch of scientific research, that is in the focus of Linguistics, Mathematics and Cybernetics, and aims to build a system that implements Machine Translation in the narrow sense of this concept. Parallel corpus toolkit is a database with a set of L1 and L2 texts, that contains a large number of texts of different discourses, issues and topics. In addition, the attention is paid to the OPUS corpus toolkit as one of the translation memory types, which ensures the efficiency of the process of intelligent translation and is currently a free corpus system in the public domain, which contains corpora of texts from L1 and L2 to L3...Ln from numerous Internet resources and is constantly updated. The tested resource capabilities of the OPUS corpus tool have proved their effectiveness in the process of verification of one-, two-, and three-component L2 lexical constructs on the example of L1 and L2 text fragments belonging to film discourse.

Keywords: translation memory, Computer-Aided Translation, Machine Translation, Parallel corpus toolkit, Corpus Linguistics, OPUS.

Анотація

У статті розтлумачено термінопоняття "перекладацька пам'ять" і визначено як комп'ютерні бази даних, де представлено як сегменти текстів різних дискурсів L1, так і еквіваленти цих сегментів L2. Окреслено основні види перекладацької пам'яті: автоматизований переклад, машинний переклад, паралельний корпусний інструментарій. Автоматизований переклад – це процес виконання перекладу тексту L1 для отримання L2 шляхом використання спеціалізованого комп'ютерного забезпечення. У такий спосіб людський фактор відіграє одну з важливих місій у процесі виконання автоматизованого перекладу, адже текст L1 піддається трьом видам обробки – до, інтер- і післяредагуванню. Машинний переклад розглянуто вузько як процес перекладу тексту з L1 на L2, що відбувається за допомогою комп'ютера повністю і / або частково, а також у широкому значенні – як галузь наукових досліджень лінгвістики, математики і кібернетики, яка має на меті побудувати систему, що реалізує машинний переклад у вузькому значенні цього поняття. Паралельний корпусний інструментарій – база даних із набором текстів L1 і L2, де міститься значна кількість текстів різних дискурсів, проблематики, тематики. Окрім цього, увагу звернено на корпусний інструментарій OPUS як один із видів перекладацької пам'яті, що забезпечує ефективність процесу інтелектуального перекладу і на сьогодні є безкоштовною корпусною системою у відкритому доступі, яка містить корпуси текстів від L1 і L2 до L3… Ln із різних інтернет-ресурсів і постійно поповнюється. Апробовані ресурсні можливості корпусного інструмента OPUS засвідчили свою ефективність у процесі верифікації одно-, дво- і трикомпонентних лексичних конструктів L2 на прикладі фрагментів тексту L1 і L2, що належить до кінодискурсу.

Ключові слова: перекладацька пам'ять, автоматизований переклад, машинний переклад, паралельний корпусний інструментарій, корпусна лінгвістика, OPUS.

Вступ. Лінгвістика, здійснивши інтеграцію методології корпусної лінгвістики (Alsop et al., 2020; Stefanowitsch, 2020), перекладознавства (Pylypiuk, 2022), сприяла появі студій із *корпусно-базованого перекладознавства* (англ. *Corpus-Based Translation Studies*) (Попович та ін., 2020; Neumann et al., 2022), основна діяльність якого полягає в дослідженні тексту мови перекладу (далі – L2) на основі аналізу корпусів текстів мови оригіналу (далі – L1) за допомогою комп'ютерного програмного забезпечення (Chitez & Pungă, 2020; Kruger, 2002 та ін.). Сьогодні цей підхід можна розцінювати не просто як додаткову базу даних зі знаннями для прийняття адекватного перекладацького рішення й / або верифікації тексту L2 як різновиду мовної поведінки на основі паралельних (перекладацьких) корпусів текстів (перелік різних корпусів текстів також подано тут: https://www.clarin.eu/resource-families/parallel-corpora), а й підвищення ефективності ***процесу інтелектуального перекладу*** (далі – ПІП), основна сутність якого полягає в залученні різних технологій ***перекладацької пам'яті*** (далі – ПП).

Відкритим і дотепер залишаються два питання: 1) ступінь статистичної презентації бази даних зі знаннями, які містяться в тій чи іншій системі ПП (наприклад, *Déjà Vu*, *memoQ*, *Memsource*, *OmegaT*, *SDLX*, *SmartCAT*, *STAR Transit*, *Trados*, *Wordfast*, *XTM Cloud* та ін.) (від простого слова до цілих текстових фрагментів), які вона запам'ятовує, тобто зберігає, накопичує і т. ін.; 2) вичерпність цих баз даних для забезпечення ефективності ПІП. Спробуємо дати відповідь на ці питання для корпусного ресурсу OPUS як однієї з технологій ПП, яка на сьогодні вже підтвердила свою ефективність у забезпеченні ПІП.

Аналіз останніх досліджень і публікацій.

Аналіз останніх досліджень і публікацій свідчить про застосування ***корпусного підходу*** (Tognini-Bonelli, 2001, p. 85), основна мета якого полягає в залученні різних корпусів текстів, а в межах статті – паралельного корпусу текстів OPUS, зокрема всіх його підкорпусів, який допоможе перевірити збіг / незбіг компонентів у межах одно-, дво- і трикомпонентних конструктів L2 із урахуванням особливостей L1, а

відтак і зафіксувати один із методів їхнього відтворення: або (повної / часткової) еквівалентності, або диференціації значень.

Мета статті – охарактеризувати корпусний інструментарій OPUS й апробувати його ресурсні можливості для забезпечення інтелектуального перекладу тексту L2.

Завдання статті:

– розтлумачити термінопоняття "перекладацька пам'ять" як ключового інструмента процесу інтелектуального перекладу в сучасному цифровому перекладознавстві;

– окреслити види перекладацької пам'яті, які можуть забезпечити процес інтелектуального перекладу в діяльності перекладача-практика;

– надати загальну характеристику корпусного інструментарію OPUS як одного з видів ПП, що забезпечує ефективність процесу інтелектуального перекладу;

– апробувати ресурсні можливості корпусного інструмента OPUS для верифікації одно-, дво- і трикомпонентних лексичних конструктів L2.

Матеріал дослідження – фрагменти тексту кінодискурсу L1 і L2 із фіксацією одно-, дво- і трикомпонентних лексичних конструктів.

Методи та методологія проведення дослідження.

Методологія дослідження фрагментів тексту кінодискурсу L1 і L2 передбачає дотримання й реалізацію трьох послідовних етапів.

На *першому етапі* необхідно виконати перекладацький аналіз фрагментів тексту кінодискурсу L1 і L2, зокрема його лексико-семантичного рівня, для виокремлення одно-, дво- і трикомпонентних лексичних конструктів. На цьому етапі необхідно визначити не просто набір компонентів для певного лексичного конструкта в L1, їхню послідовність (A + B + C), а переконатися в його (лексичному конструкті) кількісній відповідності L2 (наприклад, однокомпонентному лексичному конструкту в L1 може відповідати одно-, а в окремих випадках дво- і навіть багатокомпонентний лексичний конструкт у L2 і т. ін.), а з'ясувати їхнє семантичне навантаження у вигляді фіксації архі- й диференційних сем.

Другий етап спрямовано на безпосереднє залучення корпусного підходу, що відповідає процедурі корпусної верифікації зафіксованого набору компонентів для певного лексичного конструкта в L1 і L2 (зокрема одно-, дво- і трикомпонентних лексичних конструктів L1 і L2) в OPUS з метою фіксації збігів / незбігів у методах перекладу тощо.

Третій етап – це проведення авторської перекладацької експертизи щодо одно-, дво- і трикомпонентних лексичних конструктів L1 і L2 і представленні резолюції щодо забезпечення інтелектуального перекладу.

Виклад та обговорення основного матеріалу дослідження.

Перекладацька пам'ять і її характеристика.

Перекладацька пам'ять (англ. *Translation Memory*, TM, або *Translation Memory Manager*, TMM) – комп'ютерні бази даних, де представлено сегменти текстів різних дискурсів L1, а також еквіваленти цих сегментів L2. У такий спосіб машина зберігає сегменти речень L1 і L2 та повторно використовує їх. Появи цих комп'ютерних баз даних у студіях цифрового перекладознавства (див. працю: Yifan He, 2011) є закономірним. Це пов'язано з тим, що багато фахівців у галузі перекладу почали зауважувати брак можливостей *машинного перекладу* (далі – МП), пов'язаних переважно з проблемою зберігання великого масиву бази даних L2. Окремі аспекти проблеми представлено в статті "The proper place of men and machines in language translation" (Kay, 1980, pp. 1–21), у якій М. Кей намагається описати гіпотетичні дві позиції щодо браку можливостей МП. Перша лінгвістична позиція використовує приклад займенникового посилання (анафора розділення) у

перекладі, щоб проілюструвати труднощі прийняття перекладацьких рішень, що призводить до того, що велика кількість таких проблем ускладнює для машин на той час отримання високоякісного перекладу без втручання людини. Друга позиція – позиція інформатики: коли дослідник зіставляє складність словникового пошуку та перекладу й припускає, що в той час навряд чи буде досить ефективний алгоритм для МП (там само).

На сьогодні ПП має більше переваг, аніж недоліків, серед яких можна виокремити такі:

а) використання застарілих матеріалів. Завдяки ПП перекладачам не потрібно буде працювати над матеріалами, які вже були перекладені раніше. Зі свого боку компаніям і клієнтам не потрібно платити за ці матеріали. Це значно знижує витрати;

б) оцінка вартості. Оцінка нечіткого збігу вимірює схожість вихідної сторони, тому її можна обчислити до фактичного початку перекладу. Це допомагає компаніям ефективно оцінити вартість, перш ніж приступити до роботи;

в) зручне середовище автоматизованого перекладу (CAT). Нечітко узгоджені фрагменти в сегменті можна виділити в середовищі CAT, що допомагає перекладачам знайти місце для подальшого редагування (Yifan He, 2011, pp. 8–9);

г) більшість систем програмного забезпечення, що використовує ПП, оплачується і розробляється лише для Windows (DejaVuX, MetaTexis, MultiTrans, SDLX, Similis, Star Transit, Trados Workbench), але зростає кількість програм Java, які дозволяють добре працювати з Mac OS X, ніж Linux (Heartsome, OmegaT, Open Language тощо). Нарешті, деякі нові системи (Wordbee, XTM-Cloud) можна використовувати з веббраузера без будь-якого завантаження й установки.

Види перекладацької пам'яті.

Коли йдеться про перекладацьку пам'ять (далі – ПП), перекладачі-теоретики і перекладачі-практики зазвичай згадують два види технологій штучного інтелекту, які забезпечують процес інтелектуального перекладу, про який мова нижче.

До технологій ***автоматизованого перекладу*** (англ. *Computer-Aided Translation*) належать *Déjà Vu*, *memoQ*, *Memsource*, *OmegaT*, *SDLX*, *SmartCAT*, *STAR Transit*, *Trados*, *Wordfast*, *XTM Cloud* та ін., основна мета яких зводиться до виконання процесу перекладу на комп'ютері за допомогою людських ресурсів – здебільшого перекладачів, тобто людина забезпечує традиційне використання комп'ютера.

На сьогодні АП – це процес виконання перекладу тексту L1 для отримання L2 шляхом використання спеціалізованого комп'ютерного забезпечення. У такий спосіб людський фактор відіграє одну з важливих місій, адже текст L1 піддається трьом видам обробки – до, інтер- і післяредагуванню.

Найбільш поширеними технологіями машинного перекладу (англ. *Machine Translation*) є Bing Microsoft Translator, Collins Dictionary Translator, DeepL Translate, Google Translate, Internet Slang Translator, M-translate, Translatedict, SYSTRAN Translate.

На сьогодні МП (Ємельянова та ін., 2018) розглядають одночасно з позиції двох значень. У вузькому значенні МП – процес перекладу тексту з L1 на L2, що відбувається за допомогою комп'ютера повністю і / або частково. У процесі МП на вході машини видається текст, словесна частина якого не супроводжується жодними додатковими вказівками, а на виході комп'ютер видає текст L2, що є перекладом тексту L1, причому перетворення тексту L1 у текст L2 виконується без втручання людини. У широкому значенні МП є цілою галуззю наукових досліджень, що перебувають у центрі уваги лінгвістики, математики і кібернетики, і має на меті побудувати систему, що реалізує МП у вузькому значенні цього поняття.

Китайські вчені в галузі комп'ютерної лінгвістики (Cheng et al., 2019) розглядають нейронний МП, заснований на методі глибокого засвоєння інформації (англ. *Deep Learning*), як глибоке структурне навчання або ієрархічне навчання. Воно є частиною більш широкої групи методів машинного навчання, що базуються на інтерпретації результатів навчання, на вибірці алгоритмів конкретних завдань. Навчання може бути як контрольованим, так і без здійснення контролю.

Окремо варто розглянути приклад МП за допомогою нейронних мереж. Слід зазначити, що з початку активного розвитку цього типу моделей (2015–2016 рр.) вони відразу почали показувати найкращі результати на спеціалізованих конференціях. Основна ідея полягає в побудові моделі з використанням кодувальника та декодувальника. Передопрацьовані дані L2 кодуються в числові вектори однакового розміру, а потім кодувальник рекурентно створює один підсумковий вектор і подає його на вхід декодувальника. Декодувальник зі свого боку генерує переклад L1, враховуючи вектор, отриманий від кодувальника, і всі попередні згенеровані слова. Базовою архітектурою для перших моделей були односпрямовані рекурентні нейронні мережі. Серйозний недолік полягав у тому, що контекст враховувався лише для слів, що стоять перед словом, що перекладається. Пізніше ідея була розвинена у вигляді двонаправлених RNN, які враховують контекст після слова.

Ще один вид популярної ПП – ***паралельний корпусний інструментарій***. Це база даних із набором текстів L1 і L2, де міститься велика кількість текстів різних дискурсів, проблематики, тематики. Причому в перекладацькій діяльності здебільшого застосовуються два підходи: якщо перший підхід пов'язаний із використанням уже готових корпусних ресурсів, то другий – із укладанням власного корпусу текстів із проблематики і / або дискурсу, з яким працює перекладач.

Загальна характеристика корпусного інструментарію OPUS і його ресурсних можливостей.

OPUS – безкоштовна корпусна система у відкритому доступі (URL: https://opus.nlpl.eu/), яка містить корпуси текстів від L1 і L2 до L3…Ln з різних Інтернет-ресурсів і яка постійно поповнюється. Усі тексти є конвертованими і вирівняними відповідно до методології корпусної лінгвістики. Назва корпусного ресурсу OPUS (англ. ... *the open parallel corpus*, укр. ... *відкритий паралельний корпус*) була утворена від англійськомовного слова CORPUS шляхом опущення літер C й R.

До ключових характеристик корпусного ресурсу OPUS варто віднести такі: якщо перша характеристика – *мультилінгвальність*, адже OPUS містить більше 90 європейських / неєвропейських мов, то друга – *паралельність*, адже OPUS містить велику кількість паралельних корпусів текстів (наприклад, див. Рис. 1, який демонструє ресурси для англійської та чеської мов із понад один млн паралельних речень (разом L1 і L2)). Мультилінгвальний характер корпусу робить необхідним обробку його документів особливими для кожної мови способами, тому наразі триває робота над створенням спеціальних програм обробки для всіх мов, включених до OPUS.

У статті "Parallel Data, Tools and Interfaces in OPUS" Дж. Тідеманн зазначає, що OPUS містить понад 3800 мовних пар, а це понад 40 млрд лексем у 2,7 млрд паралельних одиницях (Tiedemann, 2012, p. 2216). Окрім цього, варто підкреслити, що OPUS також надає інструменти для паралельної обробки і одномовних даних L1, а також декілька опцій для пошуку даних, що робить його унікальним ресурсом для дослідницької діяльності будь-якого спрямування.

Search & download resources: en (English) | cs (Czech) | >1M | ☐ show all versions

Language resources: click on [tmx | moses | xces | lang-id] to download the data! (raw = untokenized, ud = parsed with universal dependencies, alg = word alignments and phrase tables)

corpus	doc's	sent's	cs tokens	en tokens	XCES/XML	raw	TMX	Moses	mono	raw	ud	alg	dic	freq			other files
CCMatrix v1	1	56.3M	831.2M	941.9M	xces cs en	cs en	tmx	moses	cs en	cs en				cs en		sample	
ParaCrawl v9	1013	50.6M	738.6M	805.4M	xces cs en	cs en	tmx	moses	cs en	cs en				cs en		sample	
WikiMatrix v1	1	2.1M	106.0M	1.0G	xces cs en	cs en	tmx	moses	cs en	cs en				cs en		sample	
wikimedia v20210402	1	0.1M	4.3M	349.2M	xces cs en	cs en	tmx	moses	cs en	cs en				cs en		sample	
CCAligned v1	255	12.7M	163.8M	176.2M	xces cs en	cs en	tmx	moses	cs en	cs en				cs en		sample	
OpenSubtitles v2018	56410	0.3M	44.7M	236.2M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en	query	sample	
DGT v2019	38363	5.2M	95.1M	121.4M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en		sample	
JRC-Acquis v3.0	19818	1.3M	55.6M	62.4M	xces cs en	cs en	tmx	moses	cs en	cs en				cs en		sample	
ELRC-5067-SciPar v1	1	1.1M	20.2M	23.2M	xces cs en	cs en	tmx	moses	cs en	cs en				cs en		sample	
EUbookshop v2	1182	0.5M	16.1M	18.6M	xces cs en	cs en	tmx	moses	cs en	cs en		alg	dic	cs en	query	sample	
Europarl v8	9036	0.6M	15.1M	17.6M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en		sample	
ELRC-EMEA v1	1	0.8M	15.4M	16.1M	xces cs en	cs en	tmx	moses	cs en	cs en				cs en		sample	
ELRC-2713-EMEA v1	1	0.8M	15.4M	16.1M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en		sample	
ELRC_2682 v1	1	0.8M	15.4M	16.1M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en		sample	
EMEA v3	1873	1.1M	11.7M	14.1M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en	query	sample	
XLEnt v1.2	1	3.9M	11.0M	11.6M	xces cs en	cs en	tmx	moses	cs en	cs en				cs en		sample	
ELITR-ECA v1	870	0.3M	8.5M	9.2M	xces cs en	cs en	tmx	moses	cs en	cs en				cs en		sample	
QED v2.0a	4997	0.5M	6.4M	8.5M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en		sample	
News-Commentary v16	6503	0.2M	6.1M	6.3M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en		sample	
Tanzil v1	30	0.2M	4.8M	5.6M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en	query	sample	
GNOME v1	2059	0.7M	4.2M	4.4M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt		cs en		sample	
TED2020 v1	1543	0.2M	2.8M	3.4M	xces cs en	cs en	tmx	moses	cs en	cs en				cs en		sample	
bible-uedin v1	2	62.2k	1.5M	1.8M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en		sample	
ECB v1	1	63.7k	1.4M	1.6M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt		cs en	query	sample	
WMT-News v2019	17	44.9k	0.9M	1.0M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en		sample	
KDE4 v2	1348	0.1M	0.7M	1.2M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en	query	sample	
ELRC-EUR_LEX v1	1	22.6k	0.6M	0.7M	xces cs en	cs en	tmx	moses	cs en	cs en				cs en		sample	
ELRC-3564-EUR_LEX_covid v1	1	22.6k	0.6M	0.7M	xces cs en	cs en	tmx	moses	cs en	cs en		alg smt	dic	cs en		sample	
Mozilla-I10n v1	1	0.1M	0.4M	0.7M	xces cs en	cs en			cs en	cs en				cs en		sample	
total	**145331**	**140.8M**	**2.2G**	**3.9G**	**140.8M**		**168.7M**	**180.3M**									

Рис. 1. **Знімок екрана ресурсів для англійської та чеської мов із понад 1 млн паралельних речень (разом L1 і L2)**

Дж. Тідеманн запропонував модель (Рис. 2), що ілюструє обсяг 100 найчисельніших мовних пар, які було включено до добірки корпусів текстів OPUS. Модель показує, що ці підкорпуси значно перевищують позначку у 100 мільйонів слів, що є високим показником навіть для обробки природної мови (англ. *Natural language processing*, NLP) з інтенсивним використанням даних (Tiedemann, 2012, p. 2216).

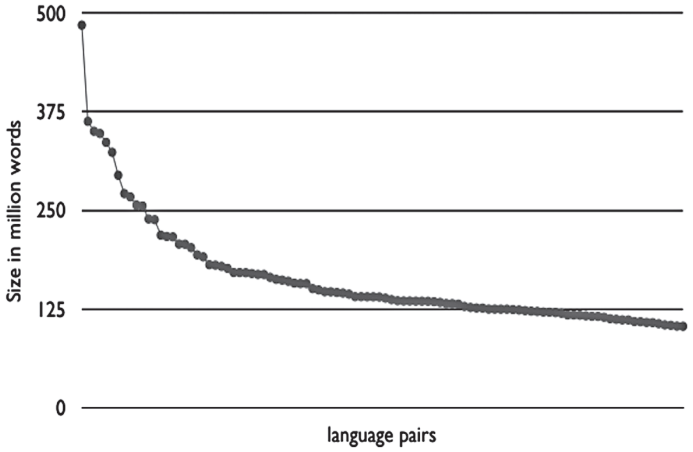

Рис. 2. **Розмір топ-100 мовних пар в OPUS (за версією Дж. Тідеманна)**

На сьогодні, за версією Дж. Тідеманна, *іспанськомовно-англійськомовна* пара з 36 млн паралельних речень, що містять приблизно 500 мільйонів лексем залишається мовною парою з найбільшим обсягом паралельних даних. Попри те, що велика кількість цих популярних мовних пар – це переважно традиційні мови, серед топ-

100 є також різні мовні пари, які, навпаки, мають нижчий ресурсний потенціал. Це паралельні тексти з такими парами, як: *болгарськомовно-угорськомовна* й *румунськомовно-турецькомовна*. Причому вони містять понад 100 млн слів, що є найрідше вживаними (Tiedemann, 2012, p. 2216).

Search & download resources: ar (Arabic) | en (English) | all | ☐ show all versions

Language resources: click on [tmx | moses | xces | lang-id] to download the data! (raw = untokenized, ud = parsed with universal dependencies, alg = word alignments and phrase tables)

corpus	doc's	sent's	ar tokens	en tokens	XCES/XML	raw	TMX	Moses	mono	raw	ud	alg	dic	freq		other files
CCMatrix v1	1	49.7M	908.5M	1.0G	xces ar en	ar en	tmx	moses	ar en	ar en				ar en		sample
WikiMatrix v1	1	2.0M	79.6M	1.0G	xces ar en	ar en	tmx	moses	ar en	ar en				ar en		sample
UNPC v1.0	114067	23.8M	552.6M	552.9M	xces ar en	ar en	tmx	moses	ar en	ar en		alg		ar en		sample
CCAligned v1	507	25.3M	389.8M	412.6M	xces ar en	ar en	tmx	moses	ar en	ar en				ar en		sample
MultiUN v1	67617	10.6M	263.1M	289.6M	xces ar en	ar en	tmx	moses	ar en	ar en		alg		ar en	query	sample
wikimedia v20210402	1	0.4M	24.7M	349.2M	xces ar en	ar en	tmx	moses	ar en	ar en				ar en		sample
OpenSubtitles v2018	40979	0.2M	31.9M	180.3M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en	query	sample
XLEnt v1.2	1	5.8M	19.4M	19.8M	xces ar en	ar en	tmx	moses	ar en	ar en				ar en		sample
QED v2.0a	5033	0.7M	6.6M	9.5M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en		sample
TED2020 v1	3879	0.4M	6.4M	8.1M	xces ar en	ar en	tmx	moses	ar en	ar en				ar en		sample
News-Commentary v16	7185	0.2M	7.0M	7.1M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en		sample
Tanzil v1	30	0.2M	5.6M	7.9M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en	query	sample
Wikipedia v1.0	1	0.2M	3.2M	3.5M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en	query	sample
TED2013 v1.1	1	0.2M	2.4M	3.0M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en	query	sample
GNOME v1	1313	0.5M	2.4M	2.6M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt		ar en		sample
GlobalVoices v2018q4	3875	59.2k	1.3M	1.8M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en		sample
bible-uedin v1	2	62.2k	1.0M	1.8M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en		sample
KDE4 v2	784	0.1M	0.7M	0.8M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en	query	sample
infopankki v1	290	54.9k	0.6M	0.7M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en		sample
Mozilla-I10n v1	1	51.7k	0.2M	0.7M	xces ar en	ar en			ar en	ar en				ar en		sample
ELRC-wikipedia_health v1	1	15.1k	0.3M	0.4M	xces ar en	ar en	tmx	moses	ar en	ar en				ar en		sample
ELRC-3083-wikipedia_health v1	1	15.1k	0.3M	0.4M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en		sample
ELRC_2922 v1	1	15.1k	0.3M	0.4M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en		sample
EUbookshop v2	30	1.7k	80.0k	0.4M	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en	query	sample
Tatoeba v2022-03-03	6	29.3k	0.2M	0.2M	xces ar en	ar en	tmx	moses	ar en	ar en				ar en		sample
tico-19 v2020-10-28	1	3.1k	76.9k	79.9k	xces ar en	ar en	tmx	moses	ar en	ar en		alg smt	dic	ar en		sample
Ubuntu v14.10					xces ar en	ar en	tmx	moses	ar en	ar en			dic	ar en		sample
total	**245608**	**120.6M**	**2.3G**	**3.9G**	**120.6M**		**122.2M**	**131.5M**								

Рис. 3. **Запит на арабськомовно-англійськомовні паралельні дані. Ресурси даних можна завантажити з перерахованих результатів запиту, натиснувши на посилання відповідних форматів даних (XCES/XML, Moses, TMX)**

OPUS містить різні підкорпуси текстів, які подано на домашній сторінці вебсайту (Home: https://opus.nlpl.eu/): від колекції біблійних текстів (Bible (uedin) – Collection of Bible translations) до статей з Вікіпедії (Wikipedia – translated sentences from Wikipedia) (URL: https://opus.nlpl.eu/index.php).

Дж. Тідеманн указує на перспективність OPUS, яка полягає в додаванні інформації про залежність до даних. Дослідник переконаний, що для цього варто покладатися на статистичні аналізатори, підготовлені з відкритої бази даних. Це також передбачає тегування за частинами мови, що вже зроблено для деяких мов і частин корпусу. Ця лінія буде базуватися на сучасних інструментальних засобах, таких, як *hunpos* (Halacsy et al., 2007) та *MaltParser* (Nivre et al., 2007), а також на вже вивчених моделях для різних мов.

Отже, мультилінгвальні й паралельні ресурсні можливості OPUS мають здатність забезпечити ефективність виконання ПІП шляхом залучення близько 3800 мовних пар.

Апробація ресурсних можливостей корпусного інструментарію OPUS для забезпечення інтелектуального перекладу тексту L2.

Спробуємо продемонструвати ресурсні можливості корпусного інструментарію OPUS для готового фрагмента тексту кінодискурсу L1 і L2. Зразу зазначимо, що якщо за L1 обрано англійську мову, то за L2 – українську (див. Табл. 1). Для верифікації одно-, дво- і трикомпонентних лексичних конструктів в OPUS обрано підкорпус OpenSubtitles2018.

Таблиця 1

Фрагмент тексту L1 і L2 кінодискурсу і верифікація одно-, дво- і трикомпонентних лексичних конструктів в OPUS

№	Time	L1	L2	№ за OPUS	Верифікація за OPUS
1.	00:11:10,360 --> 00:11:12,120	Problem?	***А що?***	1595664152	***А що ж ми тоді робитимемо ?***
				712758812	Інші відповідники: ***що з тобою?*** та ін.
2.	00:09:46,360 --> 00:09:48,799	Oh, thank you.	*О,* ***дякую***.	387303477	Так, ***дякую***.
				547110015	Інші відповідники: ***подякувати***
3.	00:26:04,600 --> 00:26:07,000	Yes... because you need me.	*Так… бо* ***я вам потрібен***.	1621751957	***Я потрібен тобі***
				1892114117	Інші відповідники: ***знадоблюся***

Як бачимо, у Табл. 1 якщо для № 1 подано однокомпонентний лексичний конструкт, для № 2 – двокомпонентний, то для № 3 – трикомпонентний, відповідно. Усі лексичні конструкти в L1 підкреслено, а в L2 позначено ***напівжирним курсивом***.

Однокомпонентний лексичний конструкт. Якщо в L1 лексичний конструкт представлено одним компонентом: Problem? (00:11:10,360), що є іменником, то в L2 – двома компонентами: *А що?* (00:11:12,120) – сполучником а і займенником що, що відповідає лексичній перекладацькій трансформації ***диференціації значень***, адже еквівалентом для слова *problem* є "проблема", що підтверджується не лише лексикографічними джерелами, але й OPUS. При цьому зауважуємо, що повний збіг у L1 і L2 простежується в OPUS (див. 1595664152).

Двокомпонентний лексичний конструкт. Якщо в L1 лексичний конструкт представлено двома компонентами: Oh, thank you. (00:09:46,360), то в L2 – одним компонентом: *О,* ***дякую***. (00:09:48,799), що відповідає граматичній трансформації опущення, коли займенник *тобі* в L2 опускається для зменшення додаткового навантаження на речення й відповідає нормам української мови.

Трикомпонентний лексичний конструкт. Кількість компонентів L1 повністю збігається з кількістю компонентів у L2: Yes... because you need me. (00:26:04,600) – *Так ... бо* ***я вам потрібен***. (00:26:07,000). Проте, звернувшись до OPUS, зауважуємо, відповідник, що є найбільш наближеним за значенням до *Я потрібен тобі* (OPUS: 1621751957), хоча і демонструє зміну комбінації компонентів. Це не випадково, а закономірно для контексту. Понад те, можна спостерігати ще один еквівалент, представлений одним компонентом, – ***знадоблюся*** (OPUS: 1892114117).

Висновки та перспективи подальшого дослідження.

У підсумку можемо констатувати, що на сьогодні перекладацька пам'ять і всі її технології займають не просто основне місце в перекладацькій діяльності завдяки наявності різних видів (автоматизований переклад, машинний переклад тощо), а дає поштовх для появи нових систем штучного інтелекту, серед яких, скажімо, паралельні корпусні інструменти різних мов (наприклад, OPUS). Це дає змогу всебічно забезпечити процес інтелектуального перекладу і підвищити його ефективність. Так, проведена верифікація одно-, дво- і трикомпонентних лексичних конструктів L1 і L2 в OPUS засвідчила свою ефективність у контексті перевірки еквівалентів

/ диференційованих еквівалентів у текстах L2, що допомагає перекладачеві переконатися в коректності перекладів.

Перспективи подальшого дослідження вбачаємо в укладанні авторського паралельного корпусу текстів різних дискурсів (наприклад, лише технічних, юридичних текстів тощо) для аналізу багатокомпонентних лексичних конструктів і визначенні методів перекладу L1 і L2.

СПИСОК УМОВНИХ СКОРОЧЕНЬ

L1 – мова оригіналу (букв. *Language 1*)
L2 – мова перекладу (букв. *Language 2*)
OPUS – the open parallel corpus
ПІП – процес інтелектуального перекладу
АП – автоматизований переклад (букв. *Computer-Aided Translation*)
МП – машинний переклад (букв. *Machine Translation*)
ПП – перекладацька пам'ять (букв. *Translation Memory*)

ЛІТЕРАТУРА

Ємельянова, О. В., Мовчан, Д. В., & Баранова, С. В. (2018). XXI століття – нова ера можливостей для студентів перекладачів. *Проблеми освіти : збірник наукових праць*, 89, 134–144.

Попович, Н. М., Луцків, А. М., & Тищук, А. Г. (2020). Corpus-Based Concept Translation. *Фаховий та художній переклад: теорія, методологія, практика*: матеріали Міжнародної науково-практичної конференції, 306–314.

Alsop, S., King, V., Giaimo, G., & Xu, X. (2020), Uses of Corpus Linguistics in Higher Education Research: An Adjustable Lens. In Huisman, J. and Tight, M. (Ed.) *Theory and Method in Higher Education Research* (*Theory and Method in Higher Education Research, Vol. 6*), Emerald Publishing Limited, Bingley, pp. 21–40. https://doi.org/10.1108/S2056-375220200000006003

Cheng, Y., Jiang, L., & Macherey, W. (2019). Robust Neural Machine Translation with Doubly Adversarial Inputs. *Proceedings of ACL*, 4324–4333.

Chitez, M., & Pungă, L. (2020). Digital Methods of Translation Studies: Using Corpus Data To Assess Trainee Translations. *British and American Studies*; Timisoara Vol. 26, 241–270.

Halacsy, P., Kornai, A., & Oravecz, C. (2007). Poster paper: Hunpos – an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, (pp. 209–212), Prague, Czech Republic, June. Association for Computational Linguistics.

Kay, M. (1980). The proper place of men and machines in language translation. *Xerox Palo Alto Research Center*, 1–21.

Kruger, A. (2002). "Corpus-based Translation Research: Its Development and Implications for General, Literary and Bible Translation" in *Acta Theologica Supplementum*, 2, 70–106.

Neumann, S., Freiwald, J., & Heilmann, A. (2022). On the Use of Multiple Methods in Empirical Translation Studies: A Combined Corpus and Experimental Analysis of Subject Identifiability in English and German. In S. Granger & M. Lefer (Authors), *Extending the Scope of Corpus-Based Translation Studies* (pp. 98–129). London: Bloomsbury Academic.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kubler, S., Marinov, S., & Marsi, E. (2007). MaltParser: A Language Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering*, *13*(2), 95–135.

Pylypiuk, K. M. (2022). On the Issue of Interaction of Linguistic Regional Studies and Translation Theory and Practice. *Закарпатські філологічні студії, 22*(1), 221–225. https://doi.org/10.32782/tps2663-4880/2022.21.1.41

Stefanowitsch, A. (2020). *Corpus Linguistics: A Guide to the Methodology*. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.3735822

Tiedemann, J. (2009). News from OPUS – a Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov. *Recent Advances in Natural Language Processing*, V, 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *LREC Conferences*, 2214–2218.

Tognini-Bonelli, E. (2001). Corpus Linguistics at Work. *Studies in Corpus Linguistics*, 6. Amsterdam: John Benjamns.

Yifan He (2011). *The Integration of Machine Translation and Translation Memory*: Thesis. Dublin City University School of Computing.

REFERENCES

Yemel′yanova, O. V., Movchan, D. V., & Baranova, S. V. (2018). KHKHI stolittya – nova era mozhlyvostey dlya studentiv perekladachiv. *Problemy osvity* : zbirnyk naukovykh prats′, 89, 134–144.

Popovych, N. M., Lutskiv, A. M., & Tyshchuk, A. H. (2020). Corpus-Based Concept Translation. *Fakhovyy ta khudozhniy pereklad: teoriya, metodolohiya, praktyka*: materialy Mizhnarodnoyi naukovo-praktychnoyi konferentsiyi, 306–314.

Alsop, S., King, V., Giaimo, G., & Xu, X. (2020), Uses of Corpus Linguistics in Higher Education Research: An Adjustable Lens. In Huisman, J. and Tight, M. (Ed.) *Theory and Method in Higher Education Research* (*Theory and Method in Higher Education Research, Vol. 6*), Emerald Publishing Limited, Bingley, pp. 21–40. https://doi.org/10.1108/S2056-375220200000006003

Cheng, Y., Jiang, L., & Macherey, W. (2019). Robust Neural Machine Translation with Doubly Adversarial Inputs. *Proceedings of ACL*, 4324–4333.

Chitez, M., & Pungă, L. (2020). Digital Methods of Translation Studies: Using Corpus Data To Assess Trainee Translations. *British and American Studies*; Timisoara Vol. 26, 241–270.

Halacsy, P., Kornai, A., & Oravecz, C. (2007). Poster paper: Hunpos – an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, (pp. 209–212), Prague, Czech Republic, June. Association for Computational Linguistics.

Kay, M. (1980). The proper place of men and machines in language translation. *Xerox Palo Alto Research Center*, 1–21.

Kruger, A. (2002). "Corpus-based Translation Research: Its Development and Implications for General, Literary and Bible Translation" in *Acta Theologica Supplementum*, 2, 70–106.

Neumann, S., Freiwald, J., & Heilmann, A. (2022). On the Use of Multiple Methods in Empirical Translation Studies: A Combined Corpus and Experimental Analysis of Subject Identifiability in English and German. In S. Granger & M. Lefer (Authors), *Extending the Scope of Corpus-Based Translation Studies* (pp. 98–129). London: Bloomsbury Academic.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kubler, S., Marinov, S., & Marsi, E. (2007). MaltParser: A Language Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering*, *13*(2), 95–135.

Pylypiuk, K. M. (2022). On the Issue of Interaction of Linguistic Regional Studies and Translation Theory and Practice. *Закарпатські філологічні студії, 22*(1), 221–225. https://doi.org/10.32782/tps2663-4880/2022.21.1.41

Stefanowitsch, A. (2020). *Corpus Linguistics: A Guide to the Methodology*. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.3735822

Tiedemann, J. (2009). News from OPUS – a Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov. *Recent Advances in Natural Language Processing*, V, 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *LREC Conferences*, 2214–2218.

Tognini-Bonelli, E. (2001). Corpus Linguistics at Work. *Studies in Corpus Linguistics*, 6. Amsterdam: John Benjamns.

Yifan He (2011). *The Integration of Machine Translation and Translation Memory*: Thesis. Dublin City University School of Computing.

Дата надходження до редакції 02.12.2022
Ухвалено до друку 22.12.2022

Відомості про авторів

Капранов Ян Васильович, доктор філологічних наук, доцент, професор кафедри теорії і практики перекладу з англійської мови Київського національного лінгвістичного університету e-mail: yan.kapranov@knlu.edu.ua		**Сфера наукових інтересів:** комп'ютерна лінгвістика, корпусна лінгвістика, прикладне перекладознавство
Тронь Тетяна Володимирівна, кандидат педагогічних наук, завідувач кафедри германських і романських мов Київського національного лінгвістичного університету e-mail: tetiana.tron@knlu.edu.ua		**Сфера наукових інтересів:** методика, педагогіка, комп'ютерна лінгвістика, прикладне перекладознавство
Івановська Божена Олександрівна доктор філософії, доцент Економіко-гуманітарного університету в Варшаві, Республіка Польща e-mail: b.iwanowska@vizja.pl		**Сфера наукових інтересів:** експериментальна лінгвістика, комп'ютерна лінгвістика, корпусна лінгвістика, психолінгвістика