УДК 81’33(=162.1=172)(=162.1=161.2)
DOI: 10.18523/lcmp2522-9281.2020.6.146-170

Павло Левчук
https://orcid.org/0000-0001-7865-6833
Данута Рошко
https://orcid.org/0000-0001-5566-0522
Роман Рошко
https://orcid.org/0000-0002-2291-6939

БАГАТОМОВНІ КОРПУСИ ІНСТИТУТУ СЛАВІСТИКИ ПОЛЬСЬКОЇ АКАДЕМІЇ НАУК – CLARIN-PL. ПОЛЬСЬКО-ЛИТОВСЬКИЙ ПАРАЛЕЛЬНИЙ КОРПУС «2» ТА ПОЛЬСЬКО-УКРАЇНСЬКИЙ ПАРАЛЕЛЬНИЙ КОРПУС

У статті описано групу Clarin-PL, яка є польським представництвом Європейської дослідницької інфраструктури CLARIN ERIC. Представлено завдання та цілі інфраструктури CLARIN ERIC та групи Clarin-PL. Як приклади подано окремі мовні засоби та ресурси, розроблені групою Clarin-PL. Особливу увагу присвячено тим багатомовним ресурсам, головна роль у побудові яких належить команді Інституту славістики Польської академії наук (ІС ПАН), зокрема це два розширені багатомовні корпуси сучасних текстів Polish-Lithuanian Parallel Corpus «2» і Polish-Ukrainian Parallel Corpus. Схарактеризовано провідну роль ІС ПАН у побудові групою Clarin-PL багатомовних корпусів. Окреслено нові, вже розпочаті та заплановані завдання, пов’язані з побудовою багатомовних ресурсів Clarin-PL.

Ключові слова: CLARIN ERIC, Clarin-PL, паралельні корпуси, польсько-український паралельний корпус, польсько-литовський паралельний корпус.

1. Вступ

Протягом майже 40 років в Інституті славістики Польської академії наук (далі – ІС ПАН) досліджують слов'янські та балтійські мови в зіставному аспекті, застосовуючи методологію, що передбачає залучення мови-посередника. Особливістю цих досліджень є теоретичний опис – від значення, викладеного мовою-посередником, до формальних характеристик рівня порівнюваних мов. З-поміж результатів такої роботи можна назати багатотомну *Gramatyka konfrontatywna bułgarsko-polska* (Koseska-Toszewa, & Penčev, 1988–2009) та *Polsko-bułgarska gramatyka konfrontatywna* (Korytkowska, Koseska-Toszewa, & Roszko, 2007). Ці граматики – фундаментальні праці, поява яких уможливлена ретельним та надійним аналізом великих обсягів мовної інформації. Традиційний аналіз друкованих джерел фактичного матеріалу потребує багато часу, тому збирання та різноаспектне оцифрування мовних ресурсів значно полегшують дослідження і, що найважливіше, роблять їх надійнішими, адже аналіз великих оцифрованих ресурсів дає змогу ефективно відокремити сигнал від шуму. Перші спроби створити «пам'ять перекладів» [1] в ІС ПАН здійснила команда, що складалася з Данути Рошко та Романа Рошка. У 1990-х рр. вони розробили перші паралельні корпуси для польської та литовської мов, які щодня використовували як у науковій діяльності (див. Roszko, 2004; Roszko, 2006a, 2006b), так і в перекладацькій роботі. Підбадьорена успіхом застосування паралельних корпусів у зіставних дослідженнях, керівник групи семантики ІС ПАН Віолетта Косеска-Тошева визнала будівництво корпусів пріоритетними завданнями групи. У співпраці з Інститутом славістики Польської академії наук та Інститутом математики та інформатики Болгарської академії наук було створено тримовний *Експериментальний болгарсько-польсько-литовський корпус.* Над укладанням корпусу у 2006–2014 рр. працювала міжнародна команда у складі Людмили Димитров, Віолетти Косески-Тошевої, Данути Рошко та Романа Рошка. Було розроблено два підкорпуси: паралельний і порівняльний. Паралельні ресурси *перевищували обсяги в 3,5 мільйона словоформ, тоді як порівняльні ресурси становили лише 0,2 мільйона словоформ. Спочатку автори корпусу залучали до паралельних корпусів тексти, написані однією з трьох мов, та їх переклади двома іншими мовами, напр.:*

[1] База даних, що містить набір раніше перекладених текстів (*англ.* MT).

1. Stanisław Lem, *Solaris*, Kraków: Wydawnictwo Literackie, 1961 (оригінал написаний польською);
2. Станислав Лем, *Соларис* (перекладачка Андреана Радева), София: Отечество, 1980 (переклад болгарською);
3. Stanislavas Lemas, *Soliaris* (перекладачка Giedrė Juodvalkytė), Vilnius: Vaga, 1978 (переклад литовською).

З часом робота просувалась, і виявилось, що кількість текстів, що відповідають зазначеним критеріям, є невеликою, тому було вирішено залучити переклади з третіх мов, не представлених у корпусі. Паралельний корпус розділений на дві частини: основну, яка охоплює переважно польські тексти, перекладені литовською та болгарською мовами, та вторинну, що містить твори, перекладені з третіх мов. Тексти, які охоплює «пам'ять перекладів», були анотовані на рівні абзаців і речень та позначені тегами. Для позначення ресурсів корпусу було використано такі мовні засоби: TaKIPI (http://nlp.pwr.wroc.pl/narzedzia-i-zasoby/narzedzia/takipi) – для польської мови, MultTex-East (https://www.researchgate.net/publication/266472851_Bulgarian_MULTEXT-East_Corpus_-_Structure_and_Content) – для болгарської мови, MorfoLema (http://donelaitis.vdu.lt/MorfoLema/) – для литовської мови. За кінцеву мету поставили опис усіх засобів згідно з єдиним стандартом MULTEXT-East (див. Roszko, D., & Roszko, R., 2009; Roszko, 2009). На проміжному етапі було складено перелік взаємних формальних відповідностей між польською, болгарською та литовською системою морфосинтаксичних теґів.

Через те, що корпус містив твори з ліцензією, його не можна було публікувати в інтернеті. Пошук ресурсів корпусу здійснювали в комерційній програмі пошуку багатомовних ресурсів – ParaConc (http://www.athel.com/para.html).

Корпусні ресурси використовували у зіставних (польсько-литовських та польсько-болгарських студіях, що охоплювали питання семантичної категорії визначеності-невизначеності, часу та гіпотетичної модальності) та лексикографічних (переважно польсько-болгарських) дослідженнях. З-поміж найважливіших опублікованих праць такі: Koseska-Toszewa, & Mazurkiewicz, 2010; Duškin, 2010; Roszko, 2015; Dimitrova, Koseska-Toszewa, Roszko, D., & Roszko, R., 2009, 2010, 2014; Koseska-Toszewa, & Satoła-Staśkowiak, 2014; Satoła-Staśkowiak, 2010.

Болгарсько-польські ресурси цього корпусу використовували й використовують у лексикографічних та лексикологічних працях дослідники

з ІС ПАН, Гуманітарно-економічної академії в Лодзі у спільних проєктах з болгарськими та українськими колегами.

Європейська інфраструктура CLARIN ERIC [1]. 29 вересня 2006 р. на першій опублікованій Європейській Дорожній Карті Дослідницької Інфраструктури (від 2006 р. ESFRI, European Strategy Forum on Research Infrastructures) з'явилася інфраструктура CLARIN, співзасновниками якої були сім держав, з-поміж них і Польща. Сьогодні європейську інфраструктуру CLARIN разом творять 20 держав і міждержавних організацій. Чотири держави (Франція, Ісландія, Південно-Африканська Республіка та Велика Британія) є членами-спостерігачами, США в цій інфраструктурі не є повним членом, а має статус країни-партнера. Варто зазначити, що CLARIN є новаторською інфраструктурою, що ідеально вписується у русло досить поширених у світі загалом і в Європі зокрема міждисциплінарних досліджень (на межі інформатики й мовознавства). Інфраструктура CLARIN виникла як відповідь на задоволення потреб користувачів, а також у контексті загальносвітового тренду розвитку штучного інтелекту. Завданням штучного інтелекту є оброблення природної мови, що неможливо без тісної співпраці мовознавців та інформатиків.

Польська група в інфраструктурі CLARIN ERIC – Clarin-PL [2]. Від самого початку польську групу дослідницької інфраструктури CLARIN ERIC становила мережа з шести наукових установ: Вроцлавський політехнологічний університет (керівник групи Clarin-PL), Інститут інформаційних технологій Польської академії наук, Інститут славістики Польської академії наук, Польсько-японська академія інформаційних технологій, Лодзький університет та Вроцлавський університет. Основною метою, що визначає побудову польської дослідницької інфраструктури Clarin-PL, є підтримка розвитку гуманітарних та соціальних наук у Польщі в тих сферах, що потребують аналізу всіх (малих та великих) мовних даних (як-от письмовий текст або мовлення). Група Clarin-PL створює та надає вченим цілісну інфраструктуру, забезпечує істотну підтримку, завдяки якій можна проводити дослідження, використовуючи сучасні методи, основані на технологіях оброблення мови (якісних та кількісних). Варто наголосити, що такі дослідження гарантують науковцям

[1] Опис європейської інфраструктури CLARIN ERIC подаємо згідно з (Levchuk, & Roszko, 2020).

[2] Опис польської групи Clarin-PL можна знайти в (Levchuk, & Roszko, 2020).

досягнення результатів, які відчутно впливають на форму сучасної світової науки.

Перший етап будівництва польської інфраструктури Clarin-PL відбувся у 2013–2018 рр. За цей період група Clarin-PL тричі отримувала підтримку Міністерства науки та вищої освіти. Другий етап розвитку польської інфраструктури Clarin-PL триває з другої половини 2018 р. Він полягає у підтримці інфраструктури, її обмеженому розширенні та адаптації ресурсів і мовних засобів до змін світових стандартів. Фазу технічного обслуговування також фінансує Міністерство науки та вищої освіти. На початку 2020 р. група Clarin-PL отримала фінансування престижного проєкту, поданого в рамках Оперативної програми «Розумний розвиток» на 2014–2020 рр., придатна вартість проєкту становить близько 132 мільйонів злотих. Основна мета цього проєкту – значно розширити орієнтовану на Clarin-PL дослідницьку інфраструктуру, яка стане платформою подальших розробок та впроваджень для оброблення природних мов та вивчення великих мовних даних (текстів і мовлення), а також мультимодальних даних.

Роль кожного наукового підрозділу, що входить до групи Clarin-PL, є важливою. До прикладу, славісти й балтисти ІС ПАН не лише будують багатомовні корпуси з польською мовою як вузловою, а й беруть участь у випрацюванні концепцій, потрібних для моделювання мовних засобів, перевіряють ці мовні засоби та ресурси. Усі члени групи рекламують інфраструктуру Clarin-PL, спільно організовують семінари (групові та індивідуальні), де користувачі інфраструктури, теперішні та потенційні, не лише ознайомлюються із сучасним станом та перспективами розвитку цієї інфраструктури, а й насамперед отримують знання про те, як ефективно використовувати всі зібрані в інфраструктурі ресурси та мовні засоби.

Брак ресурсів та мовних засобів для тієї чи тієї мови значно обмежує можливі сфери застосування інженерії природних мов, тому працівники ІС ПАН послідовно беруть участь у створенні багатомовних ресурсів. Будь-хто з користувачів інфраструктури CLARIN може ретельно проаналізувати створені ресурси, використовуючи всі системи оброблення мови, розроблені та опубліковані польською групою Clarin-PL.

2. Теоретичне підґрунтя мовних ресурсів і мовних засобів Clarin-PL

Мовні ресурси – це бази даних, що формалізовано описують природну мову в різних аспектах, наприклад, це можуть бути багатомовні корпуси та «пам'яті перекладів», а також словники, граматики, стохастичні мовні моделі та інші. Ось кілька прикладів мовних ресурсів:

- **Słowosieć** (PLWordNet) – це велика мережа слів (191 тисяча слів) та лексико-семантична база даних (285 тисяч значень і понад 600 тисяч облікових записів) для польської мови з функцією польсько-англійського словника (255 тисяч записів). Це найбільший семантичний словник реляційної моделі даних у світі;
- **Spokes** – пошукова система розмовних даних, побудована на основі 247 588 тверджень загальною кількістю майже 2,5 мільйона слів;
- **KonText – о**дно- і багатомовні корпуси Clarin-PL (докладніше – у пункті 3) та інші.

Мовні засоби – це насамперед програми для автоматичного аналізу тексту та мовлення на різних рівнях опису: формальному (морфологійному, синтаксичному), семантичному та прагматичному. По-друге, це програми, призначені для конкретних завдань з оброблення тексту (наприклад, для розпізнавання потенційних термінів, пошуку власних назв у тексті тощо). Приклади мовних засобів Clarin-PL:

- система розвідки літературного тексту (**LEM**);
- додаток для вилучення з корпусу словників та створення словників лексичних одиниць (**MeWeX**);
- мовні засоби та послуги для оброблення мовлення (**Mowa**);
- мовні засоби для перетворення орфографічного запису на фонетичний (**Transkrypcja fonetyczna**);
- токенізація та морфосинтаксичне позначення (**Tagger WCRFT2**);
- пошук і класифікація власних назв (**NER**);
- синтаксичний аналізатор залежностей для польської мови (**Parser**);
- синтаксичний аналізатор (**Spejd**);
- мовні засоби для узагальнення (скороченої форми) текстів (**Summarize**);
- мовні засоби для визначення ключових слів у тексті (**Słowa kluczowe – ReSpa**);
- мовні засоби для виявлення термінів у тексті (**TermoPL**).

Повний список доступний за посиланням: https://drive.google.com/file/d/1w4znaJgYOH_VAfjgGwT4q19EQRSusVIC/view.

3. Методи та матеріал дослідження

Попит на багатомовні корпуси зростає з кожним роком. Сьогодні користувачі корпусів – це не лише мовознавці, представники гуманітарних та соціальних наук, перекладачі, викладачі університетів, а й ІТ-спеціалісти, які послуговуються багатомовними корпусами для побудови штучного інтелекту, навчання алгоритмів автоматичного перекладу, мовних засобів програмування. Завдання побудови багатомовних корпусів у структурах Clarin-PL взяла на себе команда науковців з ІС ПАН. Учені працюють над дво- та тримовними корпусами слов'янських і балтійських мов. У 2016 р. на веб-сайті сховища Clarin-PL DSpace було опубліковано тримовний польсько-болгарсько-російський корпус (Polish-Bulgarian-Russian Parallel Corpus, https://clarin-pl.eu/dspace/handle/11321/308), а в 2018 р. –двомовні корпуси паралельних текстів обсягом, що перевищує 50 мільйонів контрактних сегментів:

- польсько-литовський (16 543 470 словоформ, Polish-Lithuanian Parallel Corpus «2», https://clarin-pl.eu/dspace/handle/11321/539, див. пункт 4.1);
- польсько-болгарський (27 504 783 словоформи; Polish-Bulgarian Parallel Corpus, https://clarin-pl.eu/dspace/handle/11321/536);
- польсько-російський (5 615 274 словоформи, Polish-Russian Parallel Corpus, https://clarin-pl.eu/dspace/handle/11321/534);
- польсько-український (1 156 579 словоформ, Polish-Ukrainian Parallel Corpus, https://clarin-pl.eu/dspace/handle/11321/535, див. пункт 4.2).

Ці корпуси доступні у сховищі Clarin-PL DSpace як файли пам'яті перекладів у форматі TMX (Memory Memory eXchange). Кожен файл TMX має опис, що містить метадані, збережені у форматі CMDI (Component Metadata infrastructure). Двомовні корпуси також доступні в багатомовному браузері KonText за адресою https://kontext.clarin-pl.eu/run.cgi/first_form [1]. На рисунку наведено фрагмент відповіді на одночасний запит про

[1] Для доступу до ресурсів у браузері KonText потрібна реєстрація користувача Clarin-PL (на сайті https://ctj.clarin-pl.eu/auth/). KonText – це загальновідомий інструмент пошуку в інтернеті одномовних та багатомовних мовних ресурсів. Його використовують не тільки в інфраструктурі Clarin, а й в інших проєктах. Наприклад, засновники Чеського національного корпусу (див. Klimowa (s. d.), без дати) (https://kontext.korpus.cz/first_form) роками

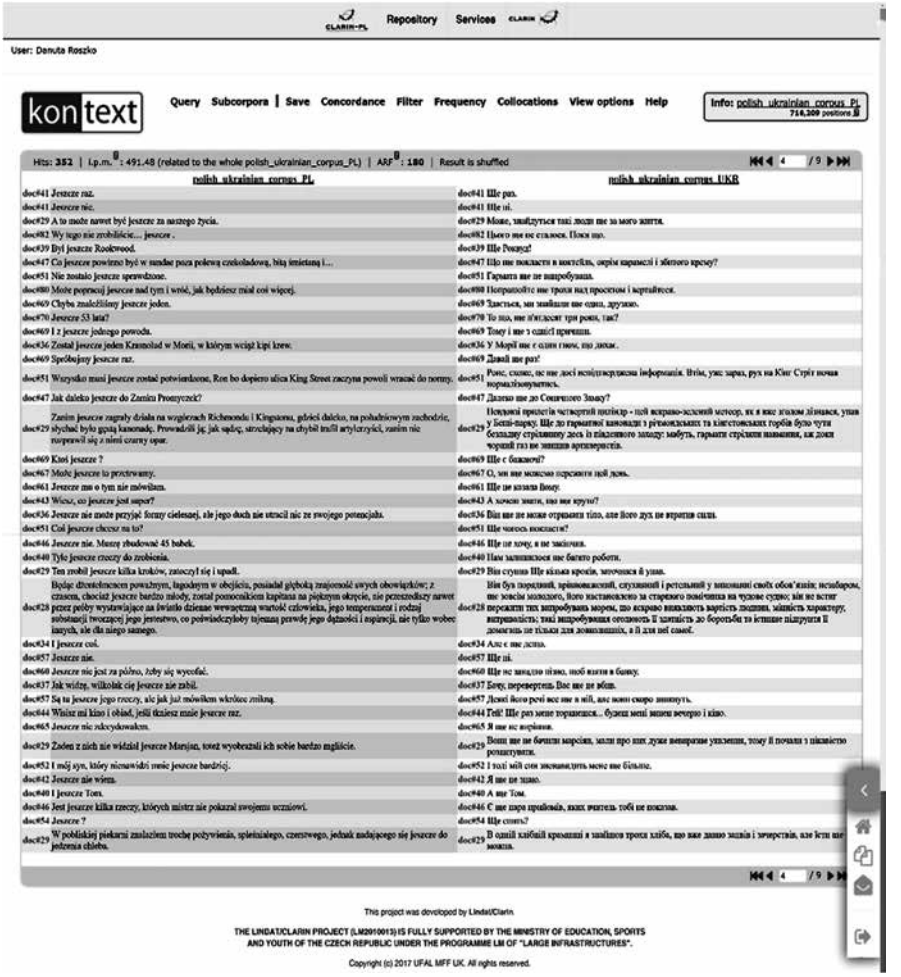

Рисунок. Результат одночасного пошуку вживання українського *ще* та польського *jeszcze*.
Вигляд четвертої із дев'яти сторінок результатів

використовують KonText для представлення лінгвістичних ресурсів, з-поміж них і багатомовних, відомий InterCorp. KonText – це інструмент, який постійно розвивається. Розвиток може стосуватися багатьох напрямів, оскільки KonText – це програма з відкритим кодом, розроблена за ліцензією GPL 2 як графічно модифікована та розширена версія оригінальної програми NoSketchEngine.

паралельний пошук лексем: української *ще* та польської *jeszcze*. Меню, яке видно у верхній частині зображення, має безліч розширених опцій та функцій, які можна вибрати кілька разів на будь-якому етапі аналізу результатів. Користувач може експортувати результати пошуку у визначеному форматі або за стандартом CSV, XML, TXT. Він також може створювати власні підкорпуси, щоб звузити кількість пошукових текстів. Варто наголосити, що користувач має змогу поглиблено аналізувати ресурси лише однієї мови.

Багатомовні корпуси доступні на найрізноманітніших платформах. Доступ до певних частин корпусу платний, див. *Sketch Engine* (https://www.sketchengine.eu/). Вони представлені 143 мовами світу. Якнайдетальніше там описані польські ресурси (вони позначені тегами, лематизовані, адаптовані до мовних засобів Word Sketch Grammar та Terms), а литовські й українські ресурси – менш деталізовані. Альтернативою платним платформам є пропозиції щодо ресурсів доступних за відкритою ліцензією. Можна назвати велику кількість платформ, де зосереджено спільні польсько-литовські ресурси, наприклад, *InterCorp* (39 мов, 1,6 мільярда слів, http://www.korpus.cz/intercorp/) (Rosen, 2016). Що ж до польсько-українських ресурсів – асортимент уже не таких великий, наприклад: *ParaSol* (http://parasolcorpus.org) (Waldenfels, & Meyer, 2006) забезпечує обидві мовні пари, хоча з обмеженим обсягом, *PolUKR* – польсько-український паралельний корпус (Turska, & Kotsyba, 2006; Kotsyba, 2012) та його запланований розвиток *PolUKR-2* (Kotsyba, 2016). Також варто згадати запланований корпус із такими мовами: польська, французька, англійська та у подальшому українська (Grabar, Kanishcheva, & Hamon 2019). *Національний корпус російської мови* в частині *Паралельний корпус* (https://ruscorpora.ru/new/search-para-multi.html) містить тексти польською, українською та литовською мовами, однак їх порівнюють лише з російською мовою. Окрім того, в мережі можна знайти багато умовно паралельних польсько-литовських та польсько-українських текстів. До прикладу, *Opus* – відкритий паралельний корпус, http://opus.nlpl.eu/ (Tiedemann, 2016), до якості якого можна мати багато зауважень, починаючи хоча б із того, що його неможливо переглядати у браузері, а потрібно завантажити. Більшість із таких корпусів створені автоматично, їх коригують у мережі. Це призводить до численних неточностей, зокрема сплутування мов, наприклад, українські ресурси містять численні російськомовні тексти.

У пропонованій розвідці ми обмежились обговоренням лише польського, литовського та українського корпусів, хоча багатомовні корпуси Clarin-PL також охоплюють інші мови (переважно слов'янські, германські та балтійські). Місія групи Clarin-PL – це допомога користувачам. Потенційні користувачі повідомили про попит на литовсько-болгарський, литовсько-російський, литовсько-український, болгарсько-російський, болгарсько-український та російсько-український корпуси. Паралельно зі створенням нових корпусів провадиться робота з розширення корпусів, які вже доступні на веб-сайті Clarin-PL. Особливу увагу спрямовано на польсько-український корпус, який уклали без попереднього планування. Зацікавлення до цих ресурсів перевищило сміливі сподівання розробників, саме тому було вирішено значно розширити польсько-український корпус.

Група науковців Інституту славістики Польської академії наук прийняла загальні правила щодо побудови багатомовних корпусів слов'янських та балтійських мов, а саме: корпуси містять сучасні тексти, що представляють усі функційні стилі; перевагу надано взаємним перекладам. На першому етапі роботи ресурси очищено, перевірено правопис, уніфіковано кодування, наповнення створених файлів метаданими. Наступний етап – робота анотаторів, які вручну вирівнюють тексти. Сегментація ресурсів відбувається на рівні речень із дотриманням вимог щодо змісту. Окремі сегменти можуть містити два речення або й більше, якщо це потрібно для розуміння змісту. Вирівнення перевіряє другий анотатор. Потім ресурси автоматично позначають: кожній словоформі присвоюють лему (основна форма) і морфосинтаксичний опис. Неоднозначні та нерозпізнані форми коментують вручну.

Добору ресурсів для корпусу передують певні обговорення. Зважаючи на той факт, що ідея Clarin-PL – це прагнення забезпечити вільний доступ до ресурсів та мовних засобів, яких потребує користувач, у текстах корпусів має бути розв'язане питання ліцензування. Насамперед до корпусу залучають тексти з відкритою ліцензією, щодо решти – ведуть перемовини з власниками авторських прав. Зазвичай це автори, перекладачі, видавці. Мета перемовин – отримати згоду/ліцензію на безкоштовне залучення певного тексту до ресурсів корпусу та на використання з дотриманням правил корпусної лінгвістики щодо оброблення твору. Варто додати, що отримати ліцензію на залучення твору до корпусу – це дуже

складний і тривалий процес: деякі видавництва хоча й дозволяють залучити твір до корпусу, відмовляються надати цифрову версію. Тоді такий текст (книга) потребує копіткого сканування, розпізнавання (перетворення відсканованих зображень у текст), очищення. Це ще одна причина того, чому кількість ліцензованих творів у різних корпусах не однакова. Корпус, створення якого розпочато раніше, містить більше ліцензованих текстів, наприклад, польсько-литовський. Польсько-український корпус зараз перебуває на початкових стадіях розбудови, тому містить набагато менше текстів. Отримані ліцензії на використання творів у корпусі є частковими.

4. Багатомовні корпуси з центральною мовою польською Clarin-PL

4.1. Польсько-литовський паралельний корпус «2» (The Polish-Lithuanian Parallel Corpus «2»)

Цей корпус розробив авторський колектив – Данута Рошко та Роман Рошко. Обидві версії цього корпусу доступні на веб-сайті Clarin-PL: перша версія – Polish-Lithuanian Parallel Corpus (Польсько-литовський паралельний корпус: https://clarin-pl.eu/dspace/handle/11321/309) і розширена – Polish-Lithuanian Parallel Corpus «2» (https://clarin-pl.eu/dspace/handle/11321/539). Також обидві версії доступні у сховищі DSpace на вебсайтах CLARIN-PL у форматі файлу TMX разом із вихідними метаданими у форматі CMDI. Корпус «2» охоплює 11 439 996 слів, обсяг польської частини становить 6 021 862 словоформи, а литовської – 5 472 134. На період липень 2018 – червень 2021 рр. запланована робота над цим корпусом передбачає збільшення його функційних можливостей, розширення ресурсів та забезпечення узгодженості з мовними засобами, що розробляються (наприклад, KonText). Відбувається ручна корекція тагування. Ресурси є напівавтоматично лематизованими та анотованими. Автоматично нерозпізнані мовні форми позначають вручну. Нова версія корпусу – «3» – з'явиться в 2021 р. Таблиця 1 містить основні відомості про польсько-литовський паралельний корпус «2».

Таблиця 1

Характеристика Polish-Lithuanian Parallel Corpus «2»

	Польська	**Литовська**
Кількість словоформ	6 021 862	5 472 134
Кількість позицій [1]	10 695 720	9 832 984
Кількість флексем [2]	212 512	231 675
Середня довжина слова	5,7 літери	5,9 літери
Середня довжина слова (звужено до спеціальних текстів)	7,9 літери	8,5 літери
Середня тривалість речення	10,93 словоформи	10,15 словоформи

Нижче ми представляємо внутрішній баланс корпусу.

Обговорюваний польсько-литовський корпус охоплює **78 художніх творів**, 25 із них – твори, написані литовською мовою та перекладені польською, решта творів – написані німецькою (5 творів), польською (3 твори), латиською (2 твори), французькою (2 твори) та словацькою (1 твір). **Поетичні тексти** представлені скромно: по одному тексту польською (оригінал) та литовською (переклад) мовами.

Юридична мова репрезентована 32 текстами різного обсягу. 16 з них – це закони й постанови, які є взаємними перекладами. Переважно це тексти, написані литовською мовою, загалом 11, решта 5 текстів написані польською мовою. Ще 16 текстів – це коментарі до законів та записи судових процесів, 4 з них були складені польською мовою та перекладені литовською.

Окрему групу становлять великі тексти ЄС, що не представлені з-поміж вищезазначених юридичних праць. Ця група налічує 18 файлів. Серед них 8 текстів – окремі документи великого обсягу, 10 текстів є компіляцією більшої кількості тематично подібних документів, що зберігаються в одному тематичному файлі (наприклад, законодавство, що стосується транспорту). Мова оригіналу всіх текстів – англійська.

Технічні тексти (загалом 92) охоплюють роботи в галузі медицини, програмування, енергетики, перероблення сирої нафти та інструкції для

[1] До цієї кількості входять усі слова та інші символи й вислови, наприклад, числа, оформлені словами.

[2] Флексема – змінювана форма лексеми.

побутової техніки. Це переважно переклади англійською мовою, невелику групу становлять взаємні переклади з польської та литовської мов.

Наукові тексти – це 10 статей/розділів монографій, більшість з яких опубліковані польською мовою та перекладені литовською мовою. Також сюди залучено деякі тексти, перекладені з російської та німецької мов.

До складу корпусу було включено 40 текстів із **Вікіпедії** (із галузі політології, історії та соціології), здебільшого це взаємні переклади. У двох випадках це переклади з третьої мови – англійської та російської.

Кінодіалоги були додані до корпусу у версії «2» на прохання користувачів. Більшість діалогів – це переклади з англійської мови, невелика частина – це діалоги, перекладені з російської. Їх загальна кількість – 32 файли. Окремі серії були об'єднані в єдине ціле. Якби кожен епізод витягувався в окремий файл, кількість файлів діалогу перевищувала б тисячу.

Презентації на конференціях. До корпусу було вміщено презентації конференцій, присвячених спільним європейським проєктам. Маючи згоду авторів, ми намагалися відібрати тексти так, щоб мова оригіналу була однією з двох мов, представлених у корпусі. Незважаючи на численні пошуки, усі презентації, представлені в корпусі, написані литовською та перекладені польською мовою.

Щодо **застосування** The Polish-Lithuanian Parallel Corpus «2», то його ресурси були використані у польсько-литовському зіставному дослідженні, проведеному в ІС ПАН (див. Roszko, D., & Roszko, R., 2014; Roszko, 2015; Koseska-Toszewa, & Roszko, 2016). Окрім того, цим корпусом послуговуються присяжні перекладачі, редакції та видавці в Польщі, польські компанії, які співпрацюють з Литовською Республікою або мають філії в Литві. Також ресурси цього корпусу використовують під час університетських занять із перекладу, описової граматики та практичного викладання литовської мови.

Окремо варто схарактеризувати використання корпусу The Polish-Lithuanian Parallel Corpus «2» у зіставних студіях. Одним із найяскравіших прикладів цього є дослідження гіпотетичності в польській та литовській мовах. Гіпотетичність потрактовуємо як одну з модальних категорій, яка слугує для вираження суб'єктивного ставлення мовця до переказуваного змісту. У цьому значенні було виділено 6 груп із різними гіпотетичними характеристиками, починаючи з найнижчого ступеня (динаміка вираження тіні сумнівів, напр. пол. *Może i był pijany* – лит. *Gal jis ir buvo girtas* – «Можливо, він був п'яний») і закінчуючи най-

вищим (мовець упевнений, напр. пол. *Niewątpliwie przywieziono z Ziemi.* – лит. *Be abejonės atsivežta iš Žemės* – «Безсумнівно, принесений із Землі»). Представлені показники гіпотетичності виражені лексичними, морфологійними засобами (лише у литовській мові) та синтаксичними конструкціями. Завдяки використанню корпусу кількість визначених показників значно зросла (порівняно з попередніми дослідженнями, проведеними традиційним способом – ручне обстеження), що уможливило зарахувати всі фіксовані показники до однієї з шести груп, виокремлених відповідно до ступеня гіпотетичної виразності. Результати цієї багаторічної роботи були вже опубліковані (Рошко, Д., & Рошко, Р., 2012).

4.2. Польсько-український паралельний корпус (The Polish-Ukrainian Parallel Corpus)

У початковий період (липень 2016 – червень 2018) цей корпус готував колектив авторів, до якого входили Максим Душкін, Роман Рошко, Войцех Сосновський та Роман Тимощук. Перша версія польсько-українського корпусу була опублікована в липні 2018 р. (Polish-Ukrainian Parallel Corpus – Польсько-український паралельний корпус), файли у форматі TMX розміщені у сховищі Clarin-PL DSpace (https://clarin-pl.eu/dspace/handle/11321/535). Обсяг корпусу становив 1 156 579 словоформ. У першій версії корпусу не було лематизації та морфосинтаксичної анотації. Ресурси були підключені до багатомовного браузера KonText. У табл. 2 подано основні дані про польсько-український паралельний корпус.

Таблиця 2

Основні характеристики польсько-українського паралельного корпусу

	Польська	**Українська**
Кількість словоформ	558 188	598 391
Кількість позицій	716 209	783 508
Кількість флексем	72 242	70 792
Середня довжина слова	5,5 літери	5,3 літери
Середня довжина слова (звужено до списків діалогів)	5,2 літери	4,8 літери
Середня тривалість речення	10 словоформ	11 словоформ
Середня тривалість речення (звужено до списків діалогів)	4,7 словоформи	4,9 словоформи

Опублікований польсько-український корпус містить загалом 168 текстів: 4 із них – художні тексти, юридична мова представлена в 50 текстах, спеціалізовані тексти – це 4 твори, решта, 110 текстів, – це кінодіалоги. Усі тексти польською та українською мовами є перекладами з англійської мови. Значна кількість кінодіалогів, що залучені до цього корпусу, була відповіддю на потреби потенційних користувачів. Clarin-PL систематично організовує семінари, тренінги, де автори корпусу представляють ресурси та мовні засоби. Користувачі пишуть свої пропозиції щодо побудови мовних засобів та ресурсів. У випадку польсько-українського корпусу багато потенційних користувачів пропонували охопити тексти, найближчі до розмовної мови. Цьому критерію найкраще відповідають діалоги.

Новий етап у розвитку польсько-українського корпусу почався з липня 2018 р., коли змінився склад команди, яка його розбудовує. Команду залишили Войцех Сосновський та Роман Тимошук, натомість долучилися Данута Рошко та Павло Левчук. На липень 2018 – червень 2021 р. запланована робота над польсько-українським корпусом полягає у тому, щоб виправити різні помічені помилки, перевірити та стандартизувати український правопис, збільшити функційні можливості, істотно розширити ресурси та забезпечити узгодженість з мовними засобами, що перебувають у розробці (наприклад, KonText). Зараз відбувається ручна корекція вирівнювання. Оскільки польські ресурси напівавтоматично лематизовані та анотовані, автоматично нерозпізнані мовні форми доводиться анотувати вручну. Публікацію версії «2» польсько-українського корпусу заплановано на 2021 р.

Варто зазначити, що з публікацією версії «2» польсько-українського корпусу відбудуться відчутні якісні зміни: значно збільшиться кількість текстів, буде залучено взаємні переклади (з польської українською та з української польською) текстів, що репрезентують різні функційні й жанрові стилі – наукові, художні, технічні, масмедійні, юридичні тощо. Наразі ми не можемо заявити, чи будуть українські ресурси лематизовані та морфосинтактично анотовані до випуску версії «2». Ми звернулись до двох українських дослідницьких центрів, які вже певний час заявляють про побудову тагування щодо можливості позначення українських ресурсів, але на момент надсилання статті до друку жодної відповіді не надійшло. Варто додати, що група

науковців ІС ПАН найближчим часом не планує працювати над тагуванням української мови [1].

Очевидно, що запорукою успіху є зразкове зрівноваження корпусних засобів. Користувачі позитивно оцінюють залучення кінодіалогів до вже опублікованої версії польсько-українського корпусу. Цілком можливо, що подальше розширення корпусу відбуватиметься з-поміж іншого й завдяки доповненню новими діалогами й текстами масмедійних жанрів.

Щодо перспектив **використання Польсько-українського паралельного корпусу**, то насамперед треба зазначити, що його ресурси вже кілька років є платформою для студіювання лексики науковцями з Болгарії, Польщі та України, які об'єднані в польсько-болгарсько-українську та польсько-українську дослідницькі групи. До прикладу, доктор П. Ковальський очолює дослідження під назвою «Інноваційні процеси в слов'янських мовах в етнокультурному та етнолінгвістичному контексті. Специфіка лексики та словотворення слов'янських мов на початку XXI століття», доктор В. Сосновський – «Мовне протистояння активної слов'янської фразеології (на матеріалі польської, болгарської, російської та української) – лінгвістичний, культурний та соціальний аспекти» та «Протистояння сучасних процесів сучасною польською та українською мовами» (співпраця з УМІФ НАН). Найважливіші результати цих досліджень віддзеркалені у таких публікаціях: Jaskot, Ganoszenko, Sosnowski, & Tymoshuk, 2017; Jaskot, & Sosnowski, 2017; Sosnowski, Blagoeva, & Jaskot, 2019a–c; Sosnowski, Blagoeva, & Tymoshuk, 2018; Sosnowski, & Tymoshuk, 2017a, 2017b; Сосновський, & Тимощук, 2017a, 2017b та інші. Окрім того, цей корпус використовують польські та українські перекладачі, а також польські шкільні вчителі в тих класах, де є учні з України.

Детальних прикладів використання польсько-українського паралельного корпусу в проведенні лінгвістичного аналізу в цій статті не наводимо, оскільки самі (тобто автори статті) жодне таке дослідження наразі не здійснюємо.

---

[1] Автори статті висловлюють вдячність рецензентам за те, що вони вказали на два потенційні теги української мови. Після перевірки їхньої ефективності буде прийнято рішення щодо їх можливого використання. Ідея побудови морфологійного сегментатора українського тексту (http://www.mova.info/Page2.aspx?l1=101) нам відома, проте важко визначити, коли цей сегментатор буде доступний, адже навіть розробники цього мовного ресурсу заявляють: *«У перспективі планується робота демонстраційної версії сегментатора в режимах онлайн на цьому сервері»*.

Наведемо кілька ілюстративних прикладів, вибраних із польсько-українського корпусу, що ілюструють реалізацію різних ступенів інтенсивності гіпотетичної семантики (про яку йшлося в пункті 4.1.): пол. ***Pewnie*** **palił dziesięć galonów na milę.** – укр. *Він* ***напевне*** *спалював бензину по десять галонів на милю;* пол. *To* ***pewnie*** *wtedy właściciel się wyprowadził.* – укр. *Я* ***подумав****, що саме того року власники й виїхали;* пол. *Albo* ***może*** *trochę cementu, i poproszę o odcisk stopy.* – укр. *Або,* ***можливо****, мішок цементу, і попрошу відбитків Ваших ніг;* пол. ***Moim zdaniem****, to jeszcze jeden nieudolny Amerykanin.* – укр. ***Мені*** *він* ***здався****, ще одним телепнем-американцем.*

4.3. Завдання Clarin-PL, заплановані до 2024 року

Команда ІС ПАН зосереджує свої зусилля на багатомовних корпусах балтійських та слов'янських мов. Зараз триває робота щодо розширення новоствореного корпусу, розбудови запланованих та обмірковування абсолютно нових корпусів.

Двомовні корпуси, опубліковані в репозитарії Clarin-PL (Polish-Lithuanian Parallel Corpus «22», Polish-Bulgarian Parallel Corpus, Polish-Russian Parallel Corpus, Polish-Ukrainian Parallel Corpus [1]), постійно удосконалюють. До них додають нові ресурси, перевіряють орфографію, розпаралелювання, лематизацію та анотацію. Розширено зокрема функції веб-браузера KonText.

Зараз час команда ІС ПАН зосередила зусилля на створенні таких багатомовних корпусів: литовсько-болгарського, литовсько-російського, литовсько-українського, болгарсько-російського, болгарсько-українського та російсько-українського. Їх оприлюднення заплановане на 2021 рік.

У другій половині 2020 р. розпочалась побудова польсько-болгарського, польсько-литовського, польсько-російського та польсько-словенського багатомовних квазідовідкових корпусів. Усі згадані корпуси будуть паралельно вручну лематизовані та анотовані. Метою такої роботи є укладання типових корпусів, які в майбутньому можна використовувати для створення удосконалених лінгвістичних ресурсів та інструментів, а також для сприяння у побудові штучного інтелекту.

[1] Польсько-литовський паралельний корпус «2», Польсько-болгарський паралельний корпус, Польсько-російський паралельний корпус, Польсько-український паралельний корпус.

5. Висновки

Європейська інфраструктура Clarin-ERIC стабільно зростає. Розсіяні ресурси (раніше створені та новостворені) об'єднують в одне ціле. Польська група Clarin-PL розробляє переважно ресурси та мовні засоби для польської мови. Однак підвищений інтерес до польської мови не звужує масштаби діяльності ІС ПАН, про що свідчать описані тут слов'янські та балтійські паралельні корпуси, зокрема польсько-український (Polish-Ukrainian Parallel Corpus) та польсько-литовський (Polish-Lithuanian Parallel Corpus «2»). Українського користувача, безумовно, зацікавить польсько-український корпус, який фактично був створений у рамках експерименту, а зараз відбувається його активний розвиток. Українців також можуть зацікавити інші корпуси, як-от литовсько-український, болгарсько-український та російсько-український, завдяки українській мові, що входить до них. На 2021 рік заплановане оприлюднення великого польсько-українського корпусу (Polish-Ukrainian Parallel Corpus «2») та нових двомовних корпусів: литовсько-українського, болгарсько-українського та російсько-українського.

До 2025 року будуть створені власноруч виготовлені та описані квазідовідкові корпуси: польсько-болгарський, польсько-литовський, польсько-російський та польсько-словенський.

Функції багатомовного браузера KonText постійно розширюються (відповідно до потреб користувачів), також постійно триває робота над поповненням та коригуванням багатомовних ресурсів, розширенням метаданих (у файлах CMDI). Мета цієї праці – надати користувачеві корпусу продукт найвищої якості, який відповідає постійно змінюваним стандартам, а також доступний для повноцінного користування як офлайн (програми, що підтримують роботу перекладача CAT), так і онлайн (у браузері KonText та інших, доступних на Clarin- EN мовні засоби).

Список використаної літератури

Левчук, П., & Рошко, Р. (2020). Багатомовні корпуси слов'янських та балтійських мов Clarin-PL. У Наталія Михальчук, & Світозара Бігунова (Ред.), *Сучасні проблеми германського та романського мовознавства: Матеріали V Міжнародної науково-практичної конференції* (с. 18–27). Retrieved from https://drive.google.com/file/d/1w4znaJgYOH_VAfjgGwT4q19EQR-SusVIC/view.

Рошко, Д., & Рошко, Р. (2012). Значения гипотетичности в литовском, польском языках и в литовском говоре окрестностей Пунска в Польше. *Baltistica*, *47* (1), 73–88. https://doi.org/10.15388/baltistica.47.1.2133

Сосновський, В., & Тимощук, Р. (2017a). Нові підходи до створення сучасних фразеологічних словників (на матеріалі «Лексикона польської та української активної фразеології»), *Мовознавство, 2,* 69–77.

Сосновський, В., & Тимощук, Р. (2017b). О работе над «Лексиконом польской и украинской активной фразеологии». В L. Janovec, R. K. Brabcová, V. Skibina, Z. Wildová (Eds.), *Svět v obrazech a ve frazeologii / World in Pictures and in Phraseology* (pp. 269–276). Univerzita Karlova, Pedagogická fakulta.

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko, R. (2009). Bulgarian-Polish-Lithuanian Corpus – Current Development. In C. Vertan, S. Piperidis, E. Paskaleva, M. Slavcheva (Eds.), *International Workshop. Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages held in conjunction with The International Conference RANLP-2009, Proceedings. Borovets* (pp. 1–8).

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko R. (2010). Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus). *Cognitive Studies/Études cognitives*, *10*, 217–239. https://dx.doi.org/10.11649/cs.2010.009

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko, R. (2014). Trilingual Aligned Corpus – Current State and New Applications. *Cognitive Studies | Études cognitives*, *14*, 13–20. https://dx.doi.org/10.11649/cs.2014.002

Duškin, M. (2010). *Wykładniki przybliżoności adnumeratywnej w języku polskim i rosyjskim*. Warszawa: Instytut Slawistyki PAN.

Grabar, N., Kanishcheva, O., & Hamon, T. (2019). Multilingual aligned corpus with Ukrainian as the target language. In *SlaviCorp. Prague, Czech Republic. ffhalshs-01968343*.

InterCorp. Retrieved from http://www.korpus.cz/intercorp/.

Jaskot, M., Ganoszenko, Ju., Sosnowski, W., & Tymoshuk, R. (2017). *Leksykon aktywnej frazeologii polskiej i ukraińskiej.* Warszawa: KJV Digital.

Jaskot, M., & Sosnowski, W. (2017). O fałszywych przyjaciołach tłumacza na przykładzie Leksykonu aktywnej frazeologii polskiej i ukraińskiej. In Barbara Borkowska-Kępska, Grzegorz Gwóźdź (Eds.), *LSP Perspectives 2. Języki specjalistyczne – nowe perspektywy 2* (pp. 55–62). Dąbrowa Górnicza: Wyższa Szkoła Biznesu w Dąbrowie Górniczej.

Kisiel., A., Koseska-Toszewa, V., Kotsyba, N., Satoła-Staśkowiak, J., and Sosnowski, W. (2016). *Polish-Bulgarian-Russian Parallel Corpus*, CLARIN-PL digital repository, http://hdl.handle.net/11321/308).

Klimova, J. (s. d.), Czech National Corpus (CNC). Retrieved from http://www.sfs.uni- tuebingen.de/~dm/events/EastWest96/cnc.html.

Korytkowska, M., Koseska-Toszewa, V., & Roszko, R. (2007). *Polsko-bułgarska gramatyka konfrontatywna.* Warszawa: Wydawnictwo Akademickie Dialog.

Koseska-Toszewa, V., & Mazurkiewicz, A. (2010). Constructing catalogue of temporal situations. *Cognitive Studies/Études cognitives*, *10*, 71–109. https://doi.org/10.11649/cs.2010.004

Koseska-Toszewa, V., & Roszko, R. (2015). On Semantic Annotation in CLARIN-PL Parallel Corpora. *Cognitive Studies/Études cognitives*, *15*, 211–236. https://doi.org/10.11649/cs.2015.016

Koseska-Toszewa, V., & Penčev, J. (Eds.) (1988–2009). *Gramatyka konfrontatywna bułgarsko-polska* (Vol. I–IX). Sofia; Warszawa.

Koseska-Toszewa, V., & Roszko, R. (2016). Języki słowiańskie i litewski w korpusach równoległych CLARIN-PL. *Studia z Filologii Polskiej i Słowiańskiej*, *51*, 191–217. https://doi.org/10.11649/sfps.2016.011

Koseska-Toszewa, V., & Satoła-Staśkowiak, J. (2014). Wprowadzenie teoretyczno-metodologiczne do „Współczesnego słownika bułgarsko-polskiego". In A. Kisiel (Ed.), *Współczesny słownik bułgarsko-polski* (pp. 1–18). Warszawa: Instytut Slawistyki PAN.

Kotsyba, N. (2012). PolUKR (a Polish-Ukrainian Parallel Corpus) as a Testbed for a Parallel *Corpora Toolbox. Prace Filologiczne, LXIII,* 181–196.

Kotsyba, N. (2016). Polsko-Ukraiński Korpus Równoległy PolUKR i jego następca PolUKR-2. In E. Gruszczyńska, A. Leńko-Szymańska (Eds.), *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora* (pp. 133–142). Warszawa: Instytut Lingwistyki Stosowanej.

MultTex-East. Retrieved from https://www.researchgate.net/publication/266472851_Bulgarian_MULTEXT-East_Corpus_-_Structure_and_Content.

MorfoLema. Retrieved from http://donelaitis.vdu.lt/MorfoLema/.

ParaConc. Retrieved from http://www.athel.com/para.html.

Rosen, A. (2016). *InterCorp – a look behind the façade of a parallel corpus*. Retrieved from https://rownolegle.ils.uw.edu.pl/files/2016/03/02_Rosen.pdf.

Roszko, D. (2006a). *Funkcjonalne odpowiedniki litewskiego perfectum w litewskiej gwarze puńskiej i w języku polskim*. Warszawa: Instytut Slawistyki PAN.

Roszko, D. (2006b). Formy perfectum i ich funkcje w litewskiej gwarze puńskiej, *Acta Baltico-Slavica*, *30, 519–531*.

Roszko, D., & Roszko, R. (2009). Morphosyntactic Specifications for Polish and Lithuanian [Description of Morphosyntactic Markers for Polish and Lithuanian Nouns within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004)]. In V. Koseska-Toszewa, L. Dimitrova, & R. Roszko (Eds.), *Representing Semantics in Digital Lexicography. Innovative Solutions for Lexical Entry Content in Slavic Lexicography. MONDILEX Fourth Open Workshop. Warszawa, Poland, 29 June – 1 July, 2009. Proceedings* (pp. 145–158). Warsaw: Institute of Slavic Studies, Polish Academy of Sciences.

Roszko, D. (2015). O innej anotacji leksykalnej w Eksperymentalnym korpusie gwary puńskiej. In D. Roszko, J. Satoła-Staśkowiak (Eds.), *Semantyka a konfrontacja językowa* (Vol. V, pp. 293–300). Warszawa: Instytut Slawistyki PAN.

Roszko, D., & Roszko, R. (2014). A Net Presentation of Lithuanian Sentences Containing Verbal Forms with the Grammatical Suffix *-dav-*, *Cognitive Studies | Études cognitives*, *14*, 173–182. https://doi.org/10.11649/cs.2014.014

Roszko, D., & Roszko, R. (2016a). Polsko-litewskie korpusy równoległe. Elementy anotacji semantycznej z zakresu modalności możliwościowej i kwantyfikacji zakresowej. In E. Gruszczyńska, A. Leńko-Szymańska (Eds.), *Polskojęzyczne korpusy równoległe. Polish language Parallel Corpora* (pp. 119–132). Warszawa. Retrieved from http://repozytorium.ceon.pl/bitstream/handle/123456789/9717/07_Roszko_Roszko.pdf?sequence=1&isAllowed=y, http://rownolegle.blog.ils.uw.edu.pl/files/2016/03/0000_Korpusy.pdf.

Roszko, D., & Roszko, R. (2016b*). Polish-Lithuanian Parallel Corpus. CLARIN-PL digital repository*. Retrieved from http://hdl.handle.net/11321/309.

Roszko, D., & Roszko, R. (2018a). Polsko-litewskie korpusy IS PAN i CLARIN-PL. In *Prace bałtystyczne*.

Roszko, D., & Roszko, R. (2018b). *Polish-Lithuanian Parallel Corpus "2". CLARIN-PL digital repository*. Retrieved from http://hdl.handle.net/11321/539.

Roszko, D., Roszko, R., & Sosnowski, W. (2018). Polish-Bulgarian Corpora ISS PAS (IS PAN) and CLARIN-PL. *Slavica Lodziensia*, *2*.

Roszko, D., Roszko, R., Sosnowski, W., & Satoła-Staśkowiak, J. (2018). Polish-Bulgarian Parallel Corpus. *CLARIN-PL digital repository*. Retrieved from http://hdl.handle.net/11321/536.

Roszko, R. (2004). *Semantyczna kategoria określoności/nieokreśloności w języku litewskim (w zestawieniu z językiem polskim)*. Warszawa: Instytut Slawistyki PAN.

Roszko, R. (2009). Description of Morphosyntactic Markers for Polish Verbs within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004). In V. Koseska-Toszewa, L. Dimitrova, R. Roszko (Eds.), *Representing Semantics in Digital Lexicography. Innovative Solutions for Lexical Entry Content in Slavic Lexicography. MONDILEX Fourth Open Workshop. Warszawa, Poland, 29 June – 1 July, 2009. Proceedings* (pp. 159–163). Warsaw: Institute of Slavic Studies, Polish Academy of Sciences.

Roszko, R., Sosnowski, W., Duszkin, M., Roszko, D., & Tymoshuk, R. (2018). *Polish-Russian Parallel Corpus*. *CLARIN-PL digital repository*. Retrieved from http://hdl.handle.net/11321/534.

Roszko, R., Tymoshuk, R., Duszkin, M., & Sosnowski, W. (2018). *Polish-Ukrainian Parallel Corpus*. *CLARIN-PL digital repository*. Retrieved from http://hdl.handle.net/11321/535.

Satoła-Staśkowiak, J. (2010). From momentarity to perfective multiplicity. Different aspects of the aorist. *Cognitive Studies/Études cognitives*, *10*, 127–132, https://doi.org/10.11649/cs.2010.007

Sketch Engine. Retrieved from https://www.sketchengine.eu/.

Sosnowski, W., & Tymoshuk, R. (2017a). Konfrontacja językowa polskich i ukraińskich jednostek frazeologicznych na przykładzie materiału z leksykonu aktywnej frazeologii polskiej i ukraińskiej. In D. Blagoeva, & L. Andreĭchin (Eds.), *Bilgarsko-polski studii* (pp. 91–108). Bŭlgarska akademiya na naukite institut za bŭlgarski ezik.

Sosnowski, W., & Tymoshuk, R. (2017b). On "The dictionary of active Polish and Ukrainian phraseology". Contrastive linguistics and culture. *Cognitive Studies/ Études cognitives*, *17*. https://doi.org/10.11649/cs.1317

Sosnowski, W., Blagoeva, D., & Jaskot, M. (2019a). Към въпроса за междуезиковата еквивалентност при фразеологията (лексикографски аспекти). In Vanya Micheva, Diana Blagoeva, Sia Kolkovska, Tatyana Aleksandrova, & Hristina Deykova (Eds.), *International Annual Conference of the Institute for Bulgarian Language* (pp. 76–82). Sofia: Інститут болгарської мови BAN.

Sosnowski, W., Blagoeva, D., & Jaskot, M. (2019b). A lexicographical approach to the contrastive analysis of Bulgarian and Polish phraseology. *Cognitive Studies/ Études cognitives*, *19*, 1–17. https://doi.org/10.11649/cs.1923

Sosnowski, W., Blagoeva, D., & Jaskot, M. (2019c). O koncepcji "Leksykonu aktywnej frazeologii bułgarskiej i polskiej". *Izvestiya na Instituta za bŭlgarski ezik 'Prof. Lyubomir Andreĭchin', 32*, 134–159.

Sosnowski, W., Blagoeva, D., & Tymoshuk, R. (2018). New Bulgarian, Polish, and Ukrainian phraseology and language corpora. *Cognitive Studies/Études cognitives*, *18*, 1–13. https://doi.org/10.11649/cs.1768

Sosnowski. W. (2017). Od słowa do działania, czyli o nauczaniu słownictwa poprzez tekst. *Języki Obce w Szkole*, *3*, 41–46.

TaKIPI. Retrieved from http://nlp.pwr.wroc.pl/narzedzia-i-zasoby/narzedzia/takipi.

Tiedemann, J. (2016). OPUS – parallel corpora for everyone. In *Proceedings of the 19th annual conference of the European Association for Machine Translation (EAMT)*. Riga.

Turska, M., & Kotsyba, N. (2006). Polsko-ukraiński korpus równoległy (PolUKR). *Biuletyn Polskiego Towarzystwa Językoznawczego*, *62*, 83–92.

Waldenfels, R. von, & Meyer, R. (2006). *ParaSol, a Corpus of Slavic and Other Language*s. Retrieved from parasol.unibe.ch.

References

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko R. (2010). Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus). *Cognitive Studies/Études cognitives*, *10*, 217–239. https://dx.doi.org/10.11649/cs.2010.009

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko, R. (2009). Bulgarian-Polish-Lithuanian Corpus – Current Development. In C. Vertan, S. Piperidis, E. Paskaleva, M. Slavcheva (Eds.), *International Workshop. Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages held in conjunction with The International Conference RANLP-2009, Proceedings. Borovets* (pp. 1–8).

Dimitrova, L., Koseska-Toszewa, V., Roszko, D., & Roszko, R. (2014). Trilingual Aligned Corpus – Current State and New Applications. *Cognitive Studies | Études cognitives*, *14*, 13–20. https://dx.doi.org/10.11649/cs.2014.002

Duškin, M. (2010). *Wykładniki przybliżoności adnumeratywnej w języku polskim i rosyjskim*. Warszawa: Instytut Slawistyki PAN.

Grabar, N., Kanishcheva, O., & Hamon, T. (2019). Multilingual aligned corpus with Ukrainian as the target language. In *SlaviCorp. Prague, Czech Republic. ffhalshs-01968343*.

InterCorp. Retrieved from http://www.korpus.cz/intercorp/.

Jaskot, M., Ganoszenko, Ju., Sosnowski, W., & Tymoshuk, R. (2017). *Leksykon aktywnej frazeologii polskiej i ukraińskiej.* Warszawa: KJV Digital.

Jaskot, M., & Sosnowski, W. (2017). O fałszywych przyjaciołach tłumacza na przykładzie Leksykonu aktywnej frazeologii polskiej i ukraińskiej. In Barbara Borkowska-Kępska, Grzegorz Gwóźdź (Eds.), *LSP Perspectives 2. Języki specjalistyczne – nowe perspektywy 2* (pp. 55–62). Dąbrowa Górnicza: Wyższa Szkoła Biznesu w Dąbrowie Górniczej.

Kisiel., A., Koseska-Toszewa, V., Kotsyba, N., Satoła-Staśkowiak, J., and Sosnowski, W. (2016). *Polish-Bulgarian-Russian Parallel Corpus*, CLARIN-PL digital repository, http://hdl.handle.net/11321/308).

Klimova, J. (s. d.), Czech National Corpus (CNC). Retrieved from http://www.sfs.uni- tuebingen.de/~dm/events/EastWest96/cnc.html.

Korytkowska, M., Koseska-Toszewa, V., & Roszko, R. (2007). *Polsko-bułgarska gramatyka konfrontatywna.* Warszawa: Wydawnictwo Akademickie Dialog.

Koseska-Toszewa V., & Mazurkiewicz A. (2010). Constructing catalogue of temporal situations. *Cognitive Studies/Études cognitives*, *10*, 71–109. https://doi.org/10.11649/cs.2010.004

Koseska-Toszewa, V., Roszko, R. (2015). On Semantic Annotation in CLARIN-PL Parallel Corpora. *Cognitive Studies/Études cognitives*, *15*, 211–236. https://doi.org/10.11649/cs.2015.016

Koseska-Toszewa, V., & Penčev, J. (Eds.) (1988–2009). *Gramatyka konfrontatywna bułgarsko-polska* (Vol. I–IX). Sofia; Warszawa.

Koseska-Toszewa, V., & Roszko, R. (2016). Języki słowiańskie i litewski w korpusach równoległych CLARIN-PL. *Studia z Filologii Polskiej i Słowiańskiej*, *51*, 191–217. https://doi.org/10.11649/sfps.2016.011

Koseska-Toszewa, V., & Satoła-Staśkowiak, J. (2014). Wprowadzenie teoretyczno-metodologiczne do „Współczesnego słownika bułgarsko-polskiego". In A. Kisiel (Ed.), *Współczesny słownik bułgarsko-polski* (pp. 1–18). Warszawa: Instytut Slawistyki PAN.

Kotsyba, N. (2012). PolUKR (a Polish-Ukrainian Parallel Corpus) as a Testbed for a Parallel *Corpora Toolbox. Prace Filologiczne, LXIII,* 181–196.

Kotsyba, N. (2016). Polsko-Ukraiński Korpus Równoległy PolUKR i jego następca PolUKR-2. In E. Gruszczyńska, A. Leńko-Szymańska (Eds.), *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora* (pp. 133–142). Warszawa: Instytut Lingwistyki Stosowanej.

Levchuk, P., & Roszko, R. (2020). Bahatomovni korpusy slov'yans'kykh ta baltiys'kykh mov Clarin-PL. In Nataliya Mykhal'chuk, & Svitozara Bihunova (Eds.), *Suchasni problemy hermans'koho ta romans'koho movoznavstva. Materialy V Mizhnarodnoyi naukovo-praktychnoyi konferentsiyi [Modern Issues in Germanic and Romance Linguistics. Materials V International Research Scientific and Practical Conference]* (pp. 18–27). Retrieved from https://drive.google.com/file/d/1w4znaJgYOH_VAfjgGwT4q19EQRSusVIC/view [in Ukrainian].

MultTex-East. Retrieved from https://www.researchgate.net/publication/266472851_Bulgarian_MULTEXT-East_Corpus_-_Structure_and_Content.

MorfoLema. Retrieved from http://donelaitis.vdu.lt/MorfoLema/.

ParaConc. Retrieved from http://www.athel.com/para.html.

Rosen, A. (2016). *InterCorp – a look behind the façade of a parallel corpus*. Retrieved from https://rownolegle.ils.uw.edu.pl/files/2016/03/02_Rosen.pdf.

Roszko, D. (2006a). *Funkcjonalne odpowiedniki litewskiego perfectum w litewskiej gwarze puńskiej i w języku polskim*. Warszawa: Instytut Slawistyki PAN.

Roszko, D. (2006b). Formy perfectum i ich funkcje w litewskiej gwarze puńskiej. *Acta Baltico- Slavica*, *30, 519–531*.

Roszko, D. (2015). O innej anotacji leksykalnej w Eksperymentalnym korpusie gwary puńskiej. In D. Roszko, J. Satoła-Staśkowiak (Eds.), *Semantyka a konfrontacja językowa* (Vol. V, pp. 293–300). Warszawa: Instytut Slawistyki PAN.

Roszko, D., & Roszko, R. (2009). Morphosyntactic Specifications for Polish and Lithuanian [Description of Morphosyntactic Markers for Polish and Lithuanian Nouns within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004)]. In V. Koseska-Toszewa, L. Dimitrova, R. Roszko (Eds.), *Representing Semantics in Digital Lexicography. Innovative Solutions for Lexical Entry Content in Slavic Lexicography. MONDILEX Fourth Open Workshop. Warszawa, Poland, 29 June – 1 July, 2009. Proceedings* (pp. 145–158). Warsaw: Institute of Slavic Studies, Polish Academy of Sciences.

Roszko, D., & Roszko, R. (2012). Znachenija gipotetichnosti v litovskom, pol'skom jazykah i v litovskom govore okrestnostej Punska v Pol'she. *Baltistica*, *47* (1), 73–88 [in Ukrainian]. https://doi.org/10.15388/baltistica.47.1.2133

Roszko, D., & Roszko, R. (2014). A Net Presentation of Lithuanian Sentences Containing Verbal Forms with the Grammatical Suffix *-dav-*, *Cognitive Studies | Études cognitives*, *14*, 173–182. https://doi.org/10.11649/cs.2014.014

Roszko, D., & Roszko, R. (2016a). Polsko-litewskie korpusy równoległe. Elementy anotacji semantycznej z zakresu modalności możliwościowej i kwantyfikacji zakresowej. In E. Gruszczyńska, A. Leńko-Szymańska (Eds.), *Polskojęzyczne korpusy równoległe. Polish language Parallel Corpora* (pp. 119–132). Warszawa. Retrieved from http://repozytorium.ceon.pl/bitstream/handle/123456789/9717/07_Roszko_Roszko.pdf?sequence=1&isAllowed=y, http://rownolegle.blog.ils.uw.edu.pl/files/2016/03/0000_Korpusy.pdf.

Roszko, D., & Roszko, R. (2016b*). Polish-Lithuanian Parallel Corpus. CLARIN-PL digital repository*. Retrieved from http://hdl.handle.net/11321/309.

Roszko, D., & Roszko, R. (2018a). Polsko-litewskie korpusy IS PAN i CLARIN-PL. In *Prace bałtystyczne*.

Roszko, D., & Roszko, R. (2018b). *Polish-Lithuanian Parallel Corpus "2". CLARIN-PL digital repository*. Retrieved from http://hdl.handle.net/11321/539.

Roszko, D., Roszko, R., & Sosnowski, W. (2018). Polish-Bulgarian Corpora ISS PAS (IS PAN) and CLARIN-PL. *Slavica Lodziensia*, *2*.

Roszko, D., Roszko, R., Sosnowski, W., & Satoła-Staśkowiak, J. (2018). Polish-Bulgarian Parallel Corpus. *CLARIN-PL digital repository*. Retrieved from http://hdl.handle.net/11321/536.

Roszko, R. (2004). *Semantyczna kategoria określoności/nieokreśloności w języku litewskim (w zestawieniu z językiem polskim)*. Warszawa: Instytut Slawistyki PAN.

Roszko, R. (2009). Description of Morphosyntactic Markers for Polish Verbs within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004). In V. Koseska-Toszewa, L. Dimitrova, R. Roszko (Eds.), *Representing Semantics in Digital Lexicography. Innovative Solutions for*

Lexical Entry Content in Slavic Lexicography. MONDILEX Fourth Open Workshop. Warszawa, Poland, 29 June – 1 July, 2009. Proceedings (pp. 159–163). Warsaw: Institute of Slavic Studies, Polish Academy of Sciences.

Roszko, R., Sosnowski, W., Duszkin, M., Roszko, D., & Tymoshuk, R. (2018). *Polish-Russian Parallel Corpus*. *CLARIN-PL digital repository*. Retrieved from http://hdl.handle.net/11321/534.

Roszko, R., Tymoshuk, R., Duszkin, M., & Sosnowski, W. (2018). *Polish-Ukrainian Parallel Corpus*. *CLARIN-PL digital repository*. Retrieved from http://hdl.handle.net/11321/535.

Satoła-Staśkowiak, J. (2010). From momentarity to perfective multiplicity. Different aspects of the aorist. *Cognitive Studies/Études cognitives*, *10*, 127–132, https://doi.org/10.11649/cs.2010.007

Sketch Engine. Retrieved from https://www.sketchengine.eu/.

Sosnowski, W., & Tymoshuk, R. (2017a). Konfrontacja językowa polskich i ukraińskich jednostek frazeologicznych na przykładzie materiału z leksykonu aktywnej frazeologii polskiej i ukraińskiej. In D. Blagoeva, & L. Andreĭchin (Eds.), *Bilgarsko-polski studii* (pp. 91–108). Bŭlgarska akademiya na naukite institut za bŭlgarski ezik.

Sosnowski, W., & Tymoshuk, R. (2017b). Novi pidkhody do stvorennya suchasnykh frazeolohichnykh slovnykiv (na materiali "Leksykona pol's'koyi ta ukrayins'koyi aktyvnoyi frazeolohiyi"). *Movoznavstvo*, *2*, 69–77 [in Ukrainian].

Sosnowski, W., & Tymoshuk, R. (2017c). On "The dictionary of active Polish and Ukrainian phraseology". Contrastive linguistics and culture. *Cognitive Studies/ Études cognitives*, *17*. https://doi.org/10.11649/cs.1317

Sosnowski, W., & Tymoshuk, R. (2017d). O rabote nad "Leksikonom pol'skoj i ukrainskoj aktivnoj frazeologii". In L. Janovec, R. K. Brabcová, V. Skibina, Z. Wildová (Eds.), *Svět v obrazech a ve frazeologii / World in Pictures and in Phraseology* (pp. 269–276). Univerzita Karlova, Pedagogická fakulta [in Ukrainian].

Sosnowski, W., Blagoeva, D., & Jaskot, M. (2019a). Kŭm vŭprosa za mezhduezikovata ekvivalentnost pri frazeologiyata (leksikografski aspekti). In Vanya Micheva, Diana Blagoeva, Sia Kolkovska, Tatyana Aleksandrova, Hristina Deykova (Eds.), *International Annual Conference of the Institute for Bulgarian Language* (pp. 76–82). Sofia: Instytut Języka Bułgarskiego BAN.

Sosnowski, W., Blagoeva, D., & Jaskot, M. (2019b). A lexicographical approach to the contrastive analysis of Bulgarian and Polish phraseology. *Cognitive Studies/ Études cognitives*, *19*, 1–17. https://doi.org/10.11649/cs.1923

Sosnowski, W., Blagoeva, D., & Jaskot, M. (2019c). O koncepcji "Leksykonu aktywnej frazeologii bułgarskiej i polskiej". *Izvestiya na Instituta za bŭlgarski ezik 'Prof. Lyubomir Andreĭchin', 32*, 134–159.

Sosnowski, W., Blagoeva, D., & Tymoshuk, R. (2018). New Bulgarian, Polish, and Ukrainian phraseology and language corpora. *Cognitive Studies/Études cognitives*, *18*, 1–13. https://doi.org/10.11649/cs.1768

Sosnowski. W. (2017). Od słowa do działania, czyli o nauczaniu słownictwa poprzez tekst. *Języki Obce w Szkole*, *3*, 41–46.

TaKIPI. Retrieved from http://nlp.pwr.wroc.pl/narzedzia-i-zasoby/narzedzia/takipi.

Tiedemann, J. (2016). OPUS – parallel corpora for everyone. In *Proceedings of the 19th annual conference of the European Association for Machine Translation (EAMT)*. Riga.

Turska, M., & Kotsyba, N. (2006). Polsko-ukraiński korpus równoległy (PolUKR). *Biuletyn Polskiego Towarzystwa Językoznawczego*, *62*, 83–92.

Waldenfels, R. von, & Meyer, R. (2006). *ParaSol, a Corpus of Slavic and Other Language*s. Retrieved from parasol.unibe.ch.

Abstract

Pavlo Levchuk, Danuta Roszko, Roman Roszko

MULTILINGUAL CORPS INSTITUTE OF SLAVIC STUDIES, POLISH ACADEMY OF SCIENCES - CLARIN PL. POLISH-LITHUANIAN PARALLEL CORPUS "2" AND POLISH-UKRAINIAN PARALLEL CORPUS

Background. This article describes the Clarin-PL consortium, which represents the Polish contribution to the CLARIN ERIC European research infrastructure. The aims and tasks of both CLARIN ERIC and Clarin-PL are presented.

Purpose. Presentation of the achievements of researchers from the Institute of Slavic Studies of the Polish Academy of Sciences in the field of creating and developing multilingual corpora, including tagging and parallelizing texts.

Methods. The team of the Institute of Slavic Studies of the Polish Academy of Sciences adopted common assumptions for the construction of multilingual corpora of the Slavic and Baltic languages. Namely, the corpora contains selected modern texts that represent all functional styles to the greatest extent. Mutual translations are preferred.

Results. The article presents a description of selected multilingual resources created by Clarin-PL and made available online via the Clarin-PL website, which a team from the Institute of Slavic Studies of the Polish Academy of Sciences (IS PAN) played a key role in creating. These resources are two expanded multilingual corpora of parallel contemporary texts: the Polish-Lithuanian Parallel Corpus 2 and the Polish-Ukrainian Parallel Corpus. Due to the fact that IS PAN played a leading role in the development of the multilingual corpora in the Clarin-PL consortium, it was decided to present an outline of corpus linguistics development in IS PAN.

Discussion. The European Clarin-ERIC infrastructure is steadily developing. Scattered resources (previously created and newly emerging) are combined into a coherent whole. The Polish Consortium Clarin-PL primarily creates and develops resources and tools for the Polish language. The aim of these works is to provide the recipient with the highest possible quality of corpora compatible with constantly changing standards, allowing for the versatile use of tools.

Keywords: CLARIN ERIC; Clarin-PL; Parallel Corpus; Polish-Lithuanian Parallel Corpus; Polish-Ukrainian Parallel Corpus.

Матеріал надійшов 02.09.2020