Відновлена українська кирилиця: проєкт автоматизації додавання кириличних полів до українських записів в OCLC WorldCat

Дженні Товес
OCLC

Роман Ташліцький
Торонтський університет

Лана Согласнова
Торонтський університет

Переклад з англійської Дениса Ткачівського[1]

Анотація: Цей робочий звіт стосується спільного проєкту, результатом якого стало успішне додавання полів кирилиці до близько 30,000 українських записів у бібліографічній базі даних WorldCat, найбільшому в світі онлайн-каталозі. Історично склалося так, що українські записи в англомовних бібліотеках були транслітеровані тільки відповідно до таблиці латинізації Бібліотеки Конгресу. Однак чинні стандарти також вимагають оригінального написання, як-от українською кирилицею. Хоч автоматизація кирилізації українських історичних записів теоретично проста, на практиці з'явилося кілька проблем – від низької якості транслітерації до історичних змін в українській орфографії. У звіті представлено проєкт OCLC Ukrainian Cyrillicization та обговорено кроки щодо його реалізації як приклад успішної співпраці в галузях бібліографічної автоматизації, української філології та культури, слов'янської каталогізації та лінгвістики.

Keywords: Українська мова, кирилиця, автоматична де-транслітерація, транслітерація Бібліотеки Конгресу, каталогізація, OCLC WorldCat.

[1] Український переклад виконано в рамках навчально-виробничої практики студентів-магістрів з перекладу та редагування, яка є частиною освітньої програми «Художній переклад з англійської мови, літературне редагування та менеджмент перекладацьких проєктів» кафедри теорії та практики перекладу з англійської мови Інституту філології Київського національного університету імені Тараса Шевченка. Керівники навчально-виробничої практики: д.філол.н., професор Лада Коломієць, к.філол.н., асистент Олена Підгрушна.

Volume VIII, No. 2 (2021) DOI: https://doi.org/10.21226/ewjus626

0. Основи і вступ

У західній науці загальноприйнятою практикою є транслітерація українських (а також всіх нелатинських) назв і термінів, зокрема бібліографічної інформації. Англомовні країни в основному дотримуються правил Американської бібліотечної асоціації – таблиці латинізації Бібліотеки Конгресу (ALA-LC RT) для української мови. В англомовній каталогізації рекомендується, щоб бібліографічні записи включали як транслітерацію, так і оригінальне написання (Група зацікавлених матеріалами нелатинським шрифтом (Non-Latin Script Materials Affinity Group); «Рекомендації PCC (Програми спільної каталогізації)» ("PCC Guidelines")). Однак в історичних записах бібліотечних каталогів часто нема оригінального написання, що не відповідає сучасним стандартам. На Рисунку 1 нижче показано приклад бібліографічного поля лише в транслітерації (запис OCLC № 1114559628; Поле MARC 245 "Title and Statement of Responsibility").

Рисунок 1. Бібліографічне поле "Title and Statement of Responsibility" у транслітерації відповідно до таблиці української латинізації Бібліотеки Конгресу США ("Ukrainian [2011]").

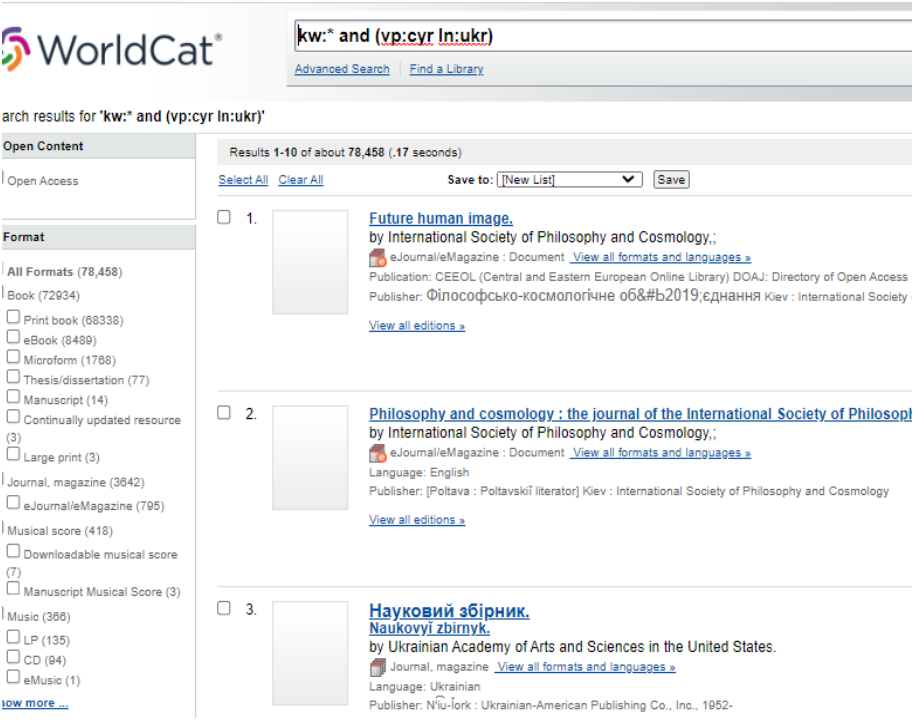
245 10 Kruta arkhitektura: Superovi fakty dliā͡ diteĭ - malykh i velykykh / Saĭmon Armstrong, pereklad z anhliĭs'koï Hanny Leliv.

У грудні 2020 року, на момент написання цього звіту, OCLC WorldCat, «найбільший у світі каталог», містив близько 2 мільярдів записів, згідно з веб-сайтом (*WorldCat*). Для української мови WorldCat містив 757,789 записів для різних матеріалів: книг, періодичних видань, карт, музичних партитур, візуальних матеріалів, цифрового контенту і т.д., на багатьох різних мовах каталогізації. З цього всього до реалізації нашого проєкту тільки близько 30,000 українських записів з англійською мовою каталогізації містили поля кирилиці. Точний витяг такої статистики, особливо стосовно наявності полів кирилиці, здійснено на вихідних даних OCLC першим автором.[2] Для того, щоб оцінити кількість матеріалів українською мовою у WorldCat за допомогою його загальнодоступного інтерфейсу, може знадобитися певна бібліографічна майстерність, а саме – використання індексних міток OCLC Expert Search («Мітки індексу» ("Index Labels")). Щоб отримати всі записи кирилицею українською мовою, можна

[2] Дж. Товес (J. Toves). [Вираховано з копії дослідження з даних worldcat.org] [Неопубліковані сирі дані]. OCLC.

використати такий рядок пошуку у пошуковому вікні : kw:* та (vp:cyr ln:ukr). Із знімків екрана загальнодоступного інтерфейсу OCLC WorldCat, який представлений на Рисунку 2, можна побачити, що ця комбінація пошукових термінів дала 78,457 записів. Зокрема, пошуковий термін kw:* видає записи з принаймні одним символом (*) в індексі ключових слів (kw), які також (‘and’) індексуються в покажчику мови (ln) як українською мовою (ukr.), та в індексі набору символів (vp) як присутні кириличні символи (cyr) («Набори символів» (“Character Sets”)). Хоча ці цифри можуть бути не такими точними, як результати, отримані за допомогою маніпуляцій необробленими даними, все ж можна отримати загальне уявлення про обсяг українських матеріалів представлений у WorldCat, а також відстежити зміни.

Рисунок 2. Параметри «Розширений пошук» OCLC WorldCat для отримання записів українською мовою.[3]



3 https://www.worldcat.org/search?q=kw%3A*+and+(vp%3Acyr+ln%3Aukr). Дата звернення: 25 червня 2021 р.

Поля кирилиці можна додавати до застарілих записів постфактум, щоб покращити та трансформувати якість бібліотечних каталогів і доступ до бібліотечних матеріалів. Таке покращення, також відоме як кирилізація (або детранслітерація), може виконуватися автоматично за допомогою програмного зіставлення латинських символів з їхньою відповідністю в кирилиці відповідно до таблиці латинізації Бібліотеки Конгресу. Джейкобз та інші (Jacobs et al.) повідомили про покращення 13,099 записів для російськомовних матеріалів у Публічній бібліотеці району Квінз за допомогою програми детранслітерації під назвою «Cyril», яка була спеціально написана мовою програмування Perl. У 2011–13 роках завдяки оновленій версії Cyril, безкоштовного програмного забезпечення MarcDeTrans, доступного в інтернеті, слов'янська секція бібліотек Торонтського університету також здійснила пілотну трансформацію кількох тисяч російських та українських записів у каталозі бібліотеки (Саммерз (Summers)). В той час як для російськомовних записів цей проєкт приніс бажані результати з вибірковим ручним контролем якості, якість записів українською мовою постраждала через велику кількість помилок у транслітерації (питання транслітерації в українській мові обговорюється детальніше нижче).

Нещодавно OCLC Research запустили глобальний проєкт "Kirillitsa v WorldCat" (Товес та ін. (Toves et al.)). Першим кроком у проєкті у співпраці з UCLA Metadata and Cataloging стала кирилізація російськомовних записів, у результаті якої зявилося «близько 958,000 записів російською мовою у WorldCat, що представляють 3,7 млн. бібліотечних фондів», доповнених полями кирилиці постфактум (Товес та ін. (Toves et al.)). У березні 2020 року у двох розсилках для слов'янських бібліотекарів (SlavLibs) та каталогізаторів слов'янських бібліотек (SlavicCats) було опубліковано заклик «допомогти OCLC з їхньою українською автообробкою кирилиці» (Флетчер (Fletcher)). Понад десяток слов'янських бібліотекарів зголосилися долучитися до проєкту у ролі рецензентів. Крім того, автори цієї статті разом працювали над конкретними стратегіями для створення хорошого набору записів для кирилиці. Протягом літа 2020 року кілька зразків розширених записів було надіслано рецензентам для оцінки якості кирилиці та виявлення помилок. У підсумку, протягом вихідних (4-7 вересня 2020 р. – День праці) у процесі пакетної обробки даних, поля кирилиці додали до 28,333 українських записів у OCLC WorldCat, що майже вдвічі збільшило кількість українських записів із кириличними полями, до 58,928. У цьому звіті описано, як команда це зробила та що ще потрібно зробити.

1. Критерії для кирилізації українських записів у WorldCat

Записи, які слід було кирилізувати, повинні були відповідати критеріям якості. Їх було кілька.

Один набір критеріїв базувався на бібліографічних стандартах. Для початку були включені лише записи повного рівня для фізичних матеріалів, створених відповідно до двох поточних стандартів: опису та доступу до ресурсів (Resource Description and Access (RDA)) та англо-американських правил каталогізації, версії 2a (AACR2a). Також англійська була взята за єдину мову каталогізації. У числах це виключило 66,717 записів, не пов'язаних з AACR2 та RDA, 370,269 записів для цифрових активів і 131,436 записів на інших мовах каталогізації, а також більшість записів від постачальника (які, зазвичай, не мали повного запису для каталогізації).

Таблиця латинізації Бібліотеки Конгресу для української мови базується на українській орфографії згідно з орфографічною реформою, прийнятою в 1933 році урядом УСРР (Мазніченко та ін.; «Українська мова (2011)» ("Ukrainian (2011)")). З цієї причини хронологічний діапазон для першого етапу проєкту обмежувався публікаціями після 1933 року, виключивши близько 18,000 записів. OCLC WorldCat містить значну кількість записів для публікацій, виданих за межами України, особливо українською діаспорою в Канаді, США, Німеччині та інших країнах. Однак більшість емігрантських громад та діаспор прийняли орфографію 1933 року лише у посткомуністичній версії, що додатково ускладнювало проєкт. Нашим робочим рішенням на першому етапі процесу було обмеження географічного обсягу проєкту публікаціями з України. Дата та місце публікації українських бібліографічних записів є прикладами бібліографічних критеріїв, визначених соціолінгвістичними та культурними міркуваннями, що вимагають знання української культури й історії, зокрема історії кодифікації української мови. Важливо, що Україна як країна видання охоплює місця видання, як-от Львів, зі складною політичною історією. Наше робоче рішення (зокрема щодо літери «ґ») полягало в тому, щоб урахувати правопис таких публікацій. Воно обговорюється нижче в Розділі 3.1.

Значною частиною цього проєкту було визначення критеріїв на основі якості транслітерації. Сучасні каталогізатори знають про деякі поширені помилки в транслітерації українських записів, а побоювання двох каталогізаторів з-поміж авторів були підтверджені та доповнені після перегляду першого набору зразків, а також підтримані вхідними даними від колег-рецензентів. У Таблиці 1 наведено найпоширеніші помилки в українській транслітерації та отримані помилки в

кирилізації. Ці помилки не тільки поширені, але й трапляються дуже часто.

Низька якість транслітерації була однією з основних проблем проєкту. Кирилізація залежить від точності транслітерації, щоб згенерувати правильний рядок кириличних символів. Стратегія полягає в застосуванні автоматичної кирилиізації тільки до «хороших» записів, які не містять помилок, особливо в плані транслітерації. Як демонструють приклади в таблиці, якщо неправильний латинський символ використовується помилково, він також буде кирилізований як неправильний кириличний символ, або навіть як нісенітниця. Наприклад, існують дві поширені помилки при транслітерації українського «ї»: замість збереження символу як «ї» використовуються різні символи: а саме «ĭ» (з діакритичним «бревіс»), або «ǐ» (з діакритичним «гачеком»). При виникненні цих помилок, програма кирилізації видасть або неправильний символ «й» замість «ĭ» (згідно з українською LC RT), або беззмістовний набір символів и&#x0220E замість «ǐ» (оскільки «ǐ» нема в українській LC RT). Багато помилок, перерахованих у Таблиці 1, ймовірно, пов'язані з мовним втручанням у таблицю російської латинізації (наприклад, в російській мові немає символу «і», тільки «й»). Загалом висока кількість помилок транслітерації в українських записах призвела до високої кількості помилок у кирилізації в початковому наборі.

Таблиця 1. Поширені помилки в українській транслітерації (таблиця латинізації ALA-LC).

Український символ кирилицею	**Латинізація (транслітерація) Бібліотеки Конгресу**	**Помилки транслітерації**	**Примітки**
г	h	g	Неправильний символ
		g кирилізується як ***г***	
є	i͡e (з нероздільною лігатурою)	e, ie	Неправильний символ, відсутній діакритик
		e кирилізується як **е** **ie** кирилізується як **іе**	
ж	ž̑h (з нероздільною лігатурою)	zh	Відсутня лігатура
		zh кирилізується як **зг**	

и	у	і	Неправильний символ
		і кирилізується як **і**	
ї	ї (і-дієрезис)	і, ĭ, ǐ (i-breve, i-caron)	Неправильні символи
		ĭ кирилізується як **й** **ǐ** кирилізується як набір символів и&#x0220E	
й	ĭ (і-бревіс)	ǐ (i-caron)	Wrong diacritic
		ǐ кирилізується як набір символів и&#x0220E	
ц	t͡s (з нероздільною лігатурою)	ts	Відсутня лігатура
		ts кирилізується як **тс**	
ь	' (штрих)	‘ (apostrophe)	Неправильний символ
ю	i͡u (з нероздільною лігатурою)	iu	Відсутня лігатура
		iu кирилізується як іу	
я	i͡a (з нероздільною лігатурою)	ia	Відсутня лігатура
		ia кирилізується як іа	

Було складено робочий список помилок, наприклад таких, як показано в Таблиці 1. Наша мета полягала в складанні стратегій для виявлення помилок, щоб створити прийнятну базу записів для кирилізації, зазвичай шляхом виключення тих записів, які містять помилки, або іноді при можливості виправляючи помилки у записах. Важливо зазначити, що ці емпіричні стратегії на основі даних не є ані алгоритмами виправлення написання, ані модулями обробки природної мови. Ми шукали незвичні закономірності, які могли б вказувати на помилку. Основні стратегії обговорюються нижче в Розділі 2. У Розділі 3 представлено обговорення окремих помилок транслітерації, а саме літери «ґ» і «г», літери «ж», а також помилкове вживання апострофа «‘» замість штриха при транслітерації м’якого знаку «ь». Загалом було виключено близько 19,000 записів, де закономірності, які ми вважали підозрілими, неможливо було виправити.

2. Стратегії для визначення помилок та виправлення

2.1 Записи, які містили символи, яких нема в таблиці латинізації Бібліотеки Конгресу, зазвичай пропускалися. Ця стратегія виключала такі помилки транслітерації, як «ĭ» замість «ї» або «ĭ».

2.2 Таблиця сполучення голосних

Для визначення найбільш поширених поєднань голосних було складено таблицю з переліком поєднань голосних у наборі даних. Це дозволило ідентифікувати менш поширені шаблони, які ймовірно вказують на помилки. Наприклад, слова, що закінчуються на сполучення голосних «ої» та «оі», були визначені як помилки та виключені з кирилізації, оскільки вони зазвичай були поширеною помилкою транслітерації українського родового відмінка однини прикметника жіночого роду із закінченням «-ої». Крім того, особливо поширеною помилкою є відсутність лігатур. У каталогізації таблиця латинізації вимагає лігатур над диграфами, що представляють «м'які» голосні «є», «ю», «я», щоб відрізнити їх від комбінацій простих літер. Наприклад, «є» транслітерується як i͡e, а «я» транслітерується як i͡a. Коли транслітерація помилково пропускає лігатуру, це призведе до помилок кирилізації, які показані в Таблиці 1 вище: наприклад, "ie" буде кирилізовано як «ie» (замість «є»), а «ia» як «ia», а не «я». Сполучення голосних без лігатур були переглянуті й виключені з кирилізації, якщо вони були визначені як неправильні (насправді багато з них були словами з російських паралельних полів, наприклад, слово «воссоединение»). Для деяких слів були внесені виправлення, наприклад, «komediia» було змінено на «komediiā», або «derzhȁvotvorennia» на «derzhȁvotvorenniā», на основі аналізу переліку двох тисяч найуживаніших слів, які описані в наступному розділі.

2.3 Список 2,000 поширених слів

Іншою стратегією було складання списку з 2,000 найпоширеніших слів, що зустрічаються в українських бібліографічних описах. Ми сподівалися, що список дозволить певне обмежене виправлення транслітерації для збільшення кількості записів, у яких є текст кирилицею. Ми переглянули поля, що включали заголовки, відомості про відповідальність, видання, інформацію про публікацію (включаючи місце та видавця), серію, загальні примітки та примітки до змісту (коди полів MARC '245', '246', '250', '260', '264', '362', '490',

'500', '505'). Більшість полів, які були транслітеровані, стосувалися «назви та відомостей про відповідальність» (MARC код 245) та інформації про публікацію (MARC код 260 та 264). У списках наведено численні приклади помилок транслітерації.

Всі слова в списках були по одному перевірені другим автором статті за певної допомоги третього автора, і були визначені правильні форми. Для багатьох слів не було варіантів, наприклад, для найпоширенішого слова та= ta. У багатьох випадках, як і очікувалося, крім правильної транслітерації були знайдені варіантні транслітерації з помилками. Зазвичай існувало більше одного способу зробити помилку. Абсолютним анти-чемпіоном у цій категорії виявилося слово «української = ukraïns'koï» (родовий відмінок однини, жіночий рід). У кирилиці слово «української» містить м'який знак «ь», зазвичай неправильно транслітерований як апостроф або навіть пропущений, і два вживання символу «ї», зазвичай неправильно транслітерованого принаймні трьома різними способами. Внаслідок різних комбінацій цих поширених помилок було виявлено 25 (!) різних помилкових варіантів транслітерації слова «української».

Для випадків, коли всі варіанти стосувалися єдиної правильної форми слова (наприклад, «української»), усі помилки транслітерації були виправлені перед застосуванням кирилиці. Нарешті, в деяких випадках були можливі дві різні форми, зазвичай відповідні різним граматичним формам слова, наприклад, «традиції» в називному відмінку множини та «традиції» в родовому відмінку однини. У деяких випадках, коли менш поширений варіант траплявся лише кілька разів, коректний правопис можна було підтвердити за допомогою ручної експертизи, наприклад, "shkil" («школа», у родовому відмінку множини), "Shkil'" (може бути відносно поширеним прізвищем); або "im." (абревіатура «імені») та "їm" (може бути дієсловом «їсти» у першій особі однини теперішнього часу або займенником «вони» у давальному відмінку). Однак в інших випадках обидва варіанти траплялися з високою частотою, наприклад, «традиції» в називному відмінку множини та «традиції» в множині родового відмінку. За відсутності надійних обчислювальних лінгвістичних процедур для автоматичного підтвердження правильного правопису такі записи були пропущені.

2.4 Імена та власні назви

Імена та власні назви в українській мові становлять окрему категорію іменників, складних для автоматичного аналізу через деякі граматичні та стилістичні особливості. Наприклад, «Валерій» – чоловіче ім'я в

називному відмінку однини або жіноче ім'я в родовому відмінку множини, і «Валерії» – жіноче ім’я в родовому відмінку однини або називному відмінку множини (як жіночого, так і чоловічого роду); «Михайла», «Михаїла» (родовий або знахідний відмінок однини; друге ім’я може бути більш давнім варіантом, що використовується в релігійній літературі, або графічним відображенням російського імені «Михаил»). Інші закономірності включають чоловічі імена, що закінчуються на «-ій» в називному відмінку однини і на «-ії» в називному відмінку множини, що робить обидві версії правильними. Урешті-решт навіть власна назва «України» (родовий відмінок однини), яка зустрічається в українських записах з високою частотою, іноді може використовуватися як “Ukraïny” (родовий відмінок однини) в поетичному стилі. Щоб уникнути помилкових автоматичних замін, такі випадки повинні бути виключені або розглянуті окремо.

3. Помилки української транслітерації: три приклади кирилізації

3.1 Літери «ґ» та «г»

Відповідно до ALA-LC RT, українська літера «г» транслітерується як “h”, а «ґ» як “g”. Однак ми помітили, що літера “g” («ґ») траплялася набагато частіше, ніж можна було очікувати в українських записах. Насправді літера була відкинута з української абетки в 1933 році і повернена лише в 1989 році (Мазніченко та ін. 7). (Історично український правопис, прийнятий у 1933 році, використовувався до 1989 року з кількома незначними правками в 1930-х, 1946 та 1960 роках [Мазніченко та ін.]). Тому записи українською мовою радянського періоду навряд містили цю літеру. Вона могла траплятися лише у назвах російських колофонів, які містили необхідний російський переклад всіх неросійських назв, опублікованих у радянські часи. Інтуїція виявилася обґрунтованою, оскільки більшість “g” в записах між 1933 і 1989 роками виявилися або з російських колофонних назв, або просто помилками транслітерації, які помилково вживали “g” замість “h” для українського «г».

Як зазначено в Розділі 1 вище, нашим робочим рішенням було обмежити хронологічний обсяг записів, включивши лише матеріали, опубліковані після 1933 року, і обмежити географічний обсяг Україною як країною публікації. Однак така географічна сфера охоплювала також назви, опубліковані після 1933 року за межами території Радянського Союзу до Другої світової війни, де все ще широко використовувалася орфографія до 1933 року. Для вирішення цієї проблеми було введено таке правило: якщо дата <= 1939 та поле 26x$a

(місце публікації) відповідає регулярному виразу, що включає найпоширеніші місця публікації в нерадянській Україні (такі як Львів, Коломия, Чернівці тощо), то допускалася наявність “g”. Це дозволило записувати книги кирилицею, видані до початку Другої світової війни в цих важливих центрах українського книговидавництва. Проте, навіть з усіма згаданими обмеженнями, ми все ще не могли бути повністю впевнені, що літера “g” («ґ») була на своєму місці, оскільки всередині України протягом вищезазначеного часу ситуація часом була неспокійною, і використання орфографічних правил після 1933 року не могло бути гарантовано. Примітно, що під час Другої світової війни значна частина України деякий час була окупована Німеччиною, а опубліковані там матеріали не контролювалися радянським урядом. Тим не менш, ми все ще повинні були бути дуже обережними під час застосування транслітерації до цих записів і вручну переглянути будь-які незвичайні випадки в записах (наприклад, коли літера “g” («ґ») займає позицію в слові, що суперечить закономірностям, які зустрічаються в більшості інших записів).

Ще одна поява “g” у транслітерації виявилася шаблоном українського закінчення прикметника «-ого» (чоловічий рід, родовий відмінок однини або знахідний відмінок однини [з одухотвореними іменниками]), помилково представленого як “-ogo”, яке виявилося досить поширеним. Оскільки літера «ґ» зазвичай не зустрічається в цьому закінченні українською мовою, було вирішено, що “-ogo” завжди означає російське слово, і запис, що містить його, повинен бути пропущений.

3.2 Літера «ж»

Окрім складностей соціолінгвістичного та культурного характеру, ми зіткнулися з деякими питаннями, викликаними взаємодією окремих транслітераційних конвенцій Бібліотеки Конгресу та особливостями української граматики. Наприклад, згідно таблиці транслітерації Бібліотеки Конгресу, літера «ж» відображається у вигляді комбінації літер z+h з лігатурою (zh͡h). При цьому, якщо лігатура пропущена, комбінація z+h передає просто з+г. Помилкове пропущення лігатури було звичайним явищем у багатьох записах, і важливо було визначити, яка з комбінацій z+h потребує лігатури, особливо враховуючи той факт, що численні українські префікси, що закінчуються на “z”, з наступною основою, що починається на “h”, становили значну частину української лексики (як у словах «розгадати», «розганяти» тощо). Рішення можна було знайти в числовому відслідковуванні закономірностей, де велика кількість однакових закономірностей

зазвичай означало їхню правильність, а рідкісні закономірності проглядалися по одній, щоб підтвердити або виправити їх.

3.3 М'ЯКИЙ ЗНАК «Ь» ТА АПОСТРОФ «‘»

Використання апострофа та м’якого знаку в транслітерації було ще одним серйозним викликом. У транслітерації ALA-LC м’який знак «ь» передається як спеціальний символ ' («штрих»), дуже схожий на символ апострофа ‘. Історично, апостроф ‘ дійсно часто помилково використовувався в транслітерації замість простого символу ' для м’якого знаку в багатьох записах, особливо у старих. Ця поширена помилка транслітерації призведе до великої кількості помилок у кирилиці. Для того, щоб очистити дані, ми розглянули набір правил для автоматичного алгоритму, щоб чітко розрізняти фактичний апостроф і апостроф, що використовується помилково замість символу м’якого знака. Орфографічне правило для апострофа в українській мові досить просте: апостроф вживається після літер б, п, в, м, ф і перед я, ю, є, ї; також за апострофом ніколи не може слідувати приголосна, він не може стояти в кінці слова. Ці правила допомогли виправити помилково використані апострофи на позначення м’яких знаків у багатьох випадках, наприклад, “na͡tsional'na=національна” (від неправильно транслітерованого “na͡tsional’na”, що кирилицею помилково передається як «національна»). Однак існувало також багато винятків, з якими було важко впоратися автоматично. Зокрема всі власні назви треба було перевіряти вручну (у тому числі імена осіб, назви місць тощо), що було зроблено шляхом відкидання всіх слів з апострофами, які починаються з великої літери, якщо вони не зустрічалися як перше слово в підполі заголовка.

ЗАКЛЮЧНІ ЗАМІТКИ

На Рисунку 3 нижче показано приклад автоматично кириличного бібліографічного поля, створеного з транслітерованого поля в україномовній публікації (OCLC №796946868). На додачу до поля 245 «Назва та відомості про відповідальність» у транслітерації було створено нове, парне поле, що містить назву та відомості про відповідальність на кирилиці (MARC код 880). Всі записи з кирилицею в цьому проєкті також автоматично отримують примітку (MARC код поле 588), яка звучить так: “Non-Latin script generated programmatically” («Нелатинські символи згенеровані автоматично»).

Рисунок 3. Запис WorldCat з автоматично дертанслітерованими кириличними полями українською мовою.

245 10 6880-01 Vcheni Ukraïny --laureaty mizh͡narodnykh premiĭ i nahorod / Vitaliĭ Ablit͡sov.
588 Non-Latin script generated programmatically.

880 10 6245-01 Вчені України --лауреати міжнародних премій і нагород / Віталій Абліцов.

Результатом цього проєкту є не тільки можливість відображати бібліографічні дані, відповідні до їхнього оригінального представлення. Можливість отримувати метадані в оригінальному написанні також має важливе значення для надання даних спільнотам, які їх потребують. Хоча сучасні практики каталогізації вимагають паралельних даних в оригінальному написанні та транслітерації латинською мовою, все ще існує велика кількість старих записів, в яких відсутні дані оригінального написання. Як показує наша дискусія, хоча здається, що перетранслітерувати латинську мову назад на кирилицю просто, насправді виникає багато проблем. Починаючи з технічної сторони, найпоширенішими проблемами є відсутні або неправильні діакритичні позначки. Цьому є багато причин. Старіші системи, можливо, унеможливили або ускладнили введення позначок. Також іноді позначки втрачаються, коли записи передаються між системами. Оскільки текстові поля зазвичай мають діакритичні знаки, нормалізовані для пошуку, помилки непомітні, якщо ви покладаєтеся на латинський текст. Однак при спробі кирилізації вони одразу впадають в очі. Виявлення проблем з позначками та проблем з використанням неправильної схеми транслітерації (тобто використання російської таблиці латинізації замість української таблиці латинізації) є вимогою для формування точного кириличного тексту. Ми наразі обмежили наші дані та виключили такі записи, але сподіваємося, що зможемо переглянути їх пізніше, коли наші методи вдосконаляться.

З соціально-культурної точки зору, проєкт спеціально та значною мірою успішно зосередився на публікаціях з України як першому кроці. Проте багато центрів українського життя в діаспорі були досить плідними у публікаціях за роки радянської влади в Україні. Зростаючий інтерес до українських публікацій про еміграцію серед вчених робить вирішення питань кирилиці в записах для матеріалів за межами України (особливо з таких міст, як Мюнхен, Балтимор, Торонто або Вінніпеґ – це добре відомі центри української культурної діяльності в еміграції) потенційним напрямком для майбутніх

проєктів. Розуміння культурного контексту для метаданих та описаної роботи є важливим для розуміння транслітерації, присутньої в записах. Де публікуються роботи, коли і де створюються метадані – усі ці фактори повинні бути вивчені, щоб отримати найкращі дані кирилиці.

Future Work

З червня 2019 року по листопад 2020 року кількість записів з кириличним текстом у WorldCat зросла з 1,6 мільйонів до 3 мільйонів – майже вдвічі![4] Це результат двох переглядів записів російською мовою та одного перегляду записів українською мовою. Підрозділ OCLC Research продовжує працювати зі слов'янськими мовними експертами, щоб додати кириличний текст до WorldCat. Інші мови, як-от болгарська, наразі переглядаються. Окрім таблиць латинізації ALA-LC для англійської мови як мови каталогізації, вивчаються інші схеми транслітерації для інших мов каталогізації. На даному етапі основна увага приділяється виявленню сумнівних латинських даних та пропуску цих записів, щоб ми могли зосередитись на швидких здобутках із найперспективнішими даними, як це успішно продемонстрував проєкт української кирилиці. Подальші етапи проєкту будуть зосереджені на способах опрацювання погано транслітерованого тексту.

[4] Дж. Товес (J. Toves). [Вираховано з копії дослідження з даних worldcat.org] [Неопубліковані сирі дані]. OCLC.

Список використаної літератури

Мазніченко, Є. І., та ін., ред. *Український правопис.* Нац. акад. наук України, Наукова думка, 2019.

"Character Sets Present." *OCLC,* 13 May 2021, https://help.oclc.org/Librarian_Toolbox/Searching_WorldCat_Indexes/Bibliographic_records/Bibliographic_record_indexes/Indexes_A_to_C/Character_Sets_Present. Accessed 25 June 2021.

Fletcher, Peter. "Ukrainian Help Needed by OCLC." *SlavLibs / SlavCats*, 11 June 2020.

"Index Labels and Examples of an Expert Search in WorldCat." *OCLC,* 29 Jul. 2020, https://help.oclc.org/Discovery_and_Reference/FirstSearch/Search/Expert_search_in_WorldCat_indexes/Index_labels_and_examples_of_an_expert_search_in_WorldCat?sl=en. Accessed 25 June 2021.

Jacobs, Jane W., et al. "Cyril: Expanding the Horizons of MARC21." *Library Hi Tech*, vol. 22, no. 1, Jan. 2004, pp. 8-17. DOI: 10.1108/07378830410524459.

Non-Latin Script Materials Affinity Group, ALA ALCTS CaMMS Committee on Cataloging: Asian & African Materials. "Linked Data for Production: Pathways to Implementation (LD4P2) Survey on Romanization: Report." *ALA Connect,* 2 March 2020, http://connect.ala.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=4199970e-0b71-4182-a5fc-1d53c1731458. Accessed 25 June 2021.

"PCC Guidelines for Creating Bibliographic Records in Multiple Character Sets." *Library of Congress,* 28 June 2016, rev. 7 Sept. 2017, https://www.loc.gov/aba/pcc/bibco/documents/PCCNonLatinGuidelines.pdf. Accessed 25 June 2021.

Summers, Ed. "MARC-Detrans-1.41 - De-Transliterate Text and MARC Records." *Metacpan.org*, 16 Nov. 2009. https://metacpan.org/dist/MARC-Detrans. Accessed 15 Dec. 2020.

Toves, Jenny, et al. " Кириллица в WorldCat." *Hanging Together: The OCLC Research Blog,* 15 April 2020, https://hangingtogether.org/?p=7868. Accessed 15 Dec. 2020.

"Ukrainian (2011)." *Library of Congress ALA-LC Romanization Tables.* https://www.loc.gov/catdir/cpso/romanization/ukrainia.pdf. Accessed 15 Dec. 2020.

WorldCat. https://www.worldcat.org. Accessed 15 Dec. 2020.