ISSN 2311-259X

https://doi.org/10.28925/2311-259x.2021.3.9
УДК 811’33.161.2’373’374

Оксана Тищенко
Інститут української мови НАН України
вул. Михайла Грушевського, 4, Київ, 01001, Україна
https://orcid.org/0000-0002-5709-1252
tom-73@ukr.net

АРХІВНА КАРТОТЕКА VS. TRANSKRIBUS: МАШИННЕ РОЗПІЗНАВАННЯ РУКОПИСНИХ МАТЕРІАЛІВ

Предметом дослідження є машинне розпізнавання рукописних матеріалів Архівної картотеки (АК) — лексико-фразеологічних матеріалів словникової комісії Всеукраїнської Академії наук, зокрема картотеки «Російсько-українського словника» 1924–1933 рр. за редакцією А. Кримського та С. Єфремова. Вивчення АК має наукове значення в контексті культурно-національного відродження в Україні на початку ХХ ст., а також у розвитку української мови, теорії та практики україністики ХХ — початку ХХІ ст. Актуальність і цінність АК стали передумовою переведення її матеріалів у цифровий формат: 2018 р. в Інституті української мови НАН України створено комп’ютерну систему «Архівна картотека», що в онлайні удоступнює матеріали насамперед у вигляді сканованих зображень. Проблема, яка потребує нагального розв’язання, — це переведення рукописних текстів у машинописний формат. Складність ручного розпізнавання спонукає до вивчення й застосування можливостей ресурсу Transkribus, що передбачає застосування методу машинного навчання. Метою розвідки є з’ясування шляхом аналізу, систематизації, класифікування та опису матеріалу особливостей підготовки карток АК для машинного опрацювання текстів. Новизна дослідження полягає в тому, що вперше розглянуто питання забезпечення двигуна HTR навчальними даними АК (завантаження на платформу, сегментування зображень на рядки й текстові ділянки, транскрибування вмісту кожної сторінки).

Головним результатом є з’ясування змісту підготовчого етапу, завданнями якого було усунення огріхів автоматичного сегментування: нетекстових елементів, непосутніх текстових елементів, некоректного автоматичного визначення текстового регіону чи рядка. Окреслено перспективи лексикографічної толоки в процесі розпізнавання карток, для чого передбачено використати колективний доступ до колекції транскрибованих документів у Transkribus. До розпізнавання ж карток вручну можна долучитися в межах нового проєкту «Усеукраїнська толока: Архівна картотека» — онлайнової платформи на сайті «АК».

Ключові слова: лексикографія; Архівна картотека; електронна система «Архівна картотека»; машинне розпізнавання; Transkribus; лексикографічна толока (краудсорсинг).

Сьогодні шляхи розв’язання багатьох мовних проблем — аналізу новотворів, іншомовних запозичень, правописних змін, зросійщених форм тощо — ми шукаємо, зокрема, у «золотій добі» української лексикографії, у словниках кінця ХІХ — початку ХХ ст. Особливо цінним є «Російсько-український словник» за ред. А. Кримського та С. Єфремова (далі — РУС). Це перший український академічний нормативний, сучасний (відповідно до епохи), створений на народній основі словник, багатий на синоніми, фразеологію, щедро ілюстрований цитатним матеріалом із різних джерел (Поздрань, 2018). Як відомо, у зв’язку зі згортанням українізації наприкінці 20-х — на початку 30-х рр. ХХ ст. репресували як науковців, так і їхні праці — словники та навіть матеріали до них, що призвело до втрати або свідомого забуття вагомого шару лінгвістичних напрацювань, зокрема IV тому РУС. Сьогодні фахівці та зацікавлені поціновувачі Слова зацифровують збережені здобутки й удоступнюють їх в інтернеті (див. електронний ресурс «Російсько-українські словники», 2020). Також ці словники останніми роками перевидають і в традиційній паперовій формі — радше для відновлення історичної справедливості, унаочнення символічного поля українознавства, коли книги після понад півстолітньої павзи можна поставити на полиці поряд з іншими доробками національної науки та культури. Безумовно, ці праці треба сприймати критично, марно шукати там *селфі*, *булінгу* та *локдауну*, однак саме там ми можемо бачити способи розв’язання багатьох проблем відновлення самобутнього обличчя нашої мови. Отже, матеріали про мовну й мовознавчу практику того періоду — актуальна та цінна інформація. Не вся лексикографічна продукція 20–30-х рр. ХХ ст. збереглася, та укладачів і користувачів нових українських словників цікавлять як завершені праці тієї доби, так і будь-які вцілілі матеріали до них.

Такими є картотечні записи, **словесний знадібок**, що його у 20-х рр. ХХ ст. зібрали для роботи Комісії для складання Словника живої української мови (КСЖУМ) Всеукраїнської Академії

ISSN 2311-259X

наук, — *Архівна картотека*[1] (далі — АК), яку після «чисток» у 30-х рр. минулого століття від «націоналістичного мотлоху» було рознесено між мільйонами карток Лексичної картотеки (далі — ЛК) Інституту мовознавства імені О. О. Потебні АН УРСР, а згодом цю картотеку успадкував Інститут української мови НАН України. Оскільки це — матеріали саме РУС, зокрема його знищеного тому (Тищенко, 2016), брак якого відчувають сучасні лексикографи й користувачі, ми переконані, що зарано називати паперову словозбірню пам'ятником минулого. Перед нами — матеріали із загубленою, забутою історією, відновити яку треба в сучасному полі питань для пошуку відповідей.

Академічний РУС базувався на виваженому колі джерел, залучених для розписування АК: це *художня* література (твори класиків: П. Куліша, Т. Шевченка, І. Франка, О. Кониського, Б. Грінченка, М. Коцюбинського, В. Стефаника, В. Самійленка, Лесі Українки; сучасників: Миколи Хвильового, М. Рильського, Григорія Косинки; переклади творів В. Шекспіра, О. Пушкіна, Біблії та ін.); *етнографічні* та *фольклорні* джерела, матеріали з *живих уст* (наприклад, села Бандурова на Чигиринщині, із Черкаського повіту, із села Франтівки на Липовеччині, з Києва, Звенигородщини, Бессарабії, Житомирщини тощо); оскільки РУС було зорієнтовано на літературний стандарт, тобто нормативну мову, до джерел залучили *наукову, навчальну, публіцистичну* літературу (твори С. Єфремова, А. Кримського, переклади з О. Потебні, а також Леніна, Сталіна, видання «Азбука комунізму» та іншої тогочасної літератури); *лексикографічні* джерела (насамперед — словники за редакцією Б. Грінченка, М. Уманця і А. Спілки, В. Дубровського, а також словники І. Верхратського, М. Левченка та ін.; дані тогочасних академічних термінологічних словників). АК містить одномовні картки (українські) та двомовні (українсько-російські, російсько-українські); заголовні слова зазвичай супроводжено цитатою та паспортом джерела.

Ці картки протягом XX ст. були в полі зору лексикографів, але пам'ять про їхню «націоналістичну» долю поступово стерлася, їх сприймали як застарілий уламок історії. Утім подивімося на них із погляду сьогодення. По-перше, це не просто виписані цитати й слова — перед нами робочі матеріали лексикографа, де видно його думку, пошуки, роздуми (рис. 1). По-друге, цінним є те, що можна побачити тільки з Архівної картотеки, наприклад, лексеми, яких не було зафіксовано в словниках: *росієць, висада* («фасад»), *самочутний* тощо. Окрім заголовних слів зважаємо й на одиниці в цитатах, де натрапляємо на авторські перлини, як-от у перекладах Б. Грінченка: *А спрагу я гасив з потоків тих, що з* ***крижників*** *біжать (Шіллер, Вільг. Телль)* — слово *крижники* («льодовики») більше ніж варте уваги й заслуговує потрапити до сучасних текстів і словників (рис. 1).

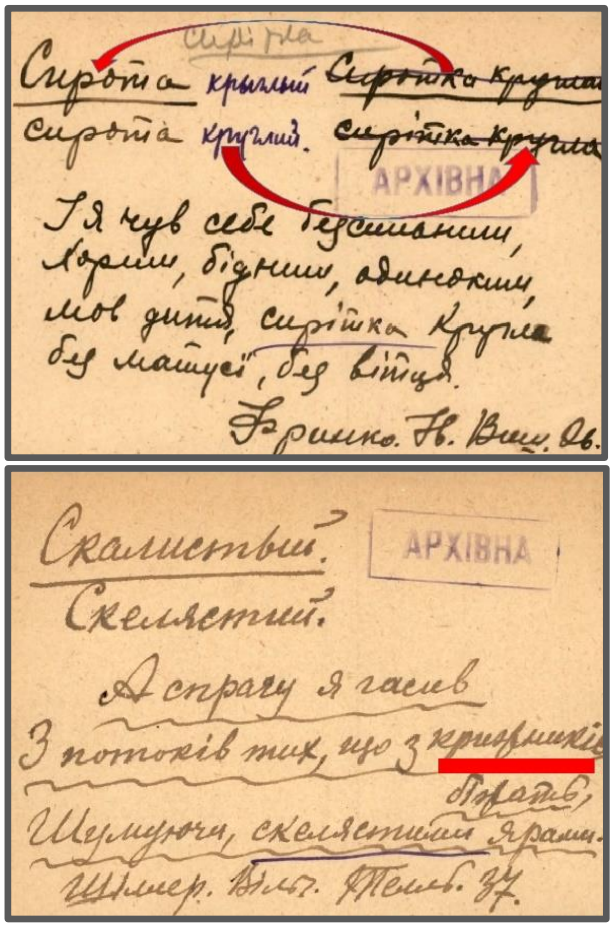

Рис. 1. Картки Архівної лексичної картотеки

Важливо досліджувати і внесок певного автора та роботу укладача картки з його текстами: які лексеми чи їхні варіанти, словосполучення, неосемантизми ще не потрапили до словників, які нові переклади чи коментарі нам пропонують автори карток. *Кусає собі губи, дивлячись* ***політично*** (*П. Куліш)* — автор картки перекладає лексему *політично* як *хитро.* РУС фіксує слово *політично* як переклад до рос. *политически*. Можливо, у IV т. ним би переклали й російське *хитро* поряд з іншими відповідниками. Такого варіанта перекладу ми не знайшли в інших словниках. *Самочуття* («самосвідомість») фіксують ще словник за редакцією Б. Грінченка та словники «золотої доби». А от сьогодні *самочуття* в цьому значенні не вживають і загалом не фіксують, хіба що як застаріле. П. Куліш (і тільки він) вживає прикметник *самочутний,* що його в картці АК перекладено як *самосознательный.* Струнко й логічно письменник витворює парадигму лексеми *самочуття*, отже, слова життєздатні. І в IV т. це була б окраса української частини словника, шанс новому слову ввійти до широкого обігу.

Це матеріал для лінгвістичних досліджень у багатьох напрямах, рідкісні матеріали, які потребують збереження й долучення до сьогоденних мовних і мовознавчих процесів, тож логічним було рішення про зацифрування АК й удоступнення її в інтернеті. Чотири місяці 2018 р. тривало лаштування АК: сотні волонтерів узяли участь

[1] Штамп «АРХІВНА», яким позначили картки в 50-х рр. XX ст., дав умовну назву й самій картотеці.

ISSN 2311-259X

у всеукраїнській акції «Збереження Архівної лексичної картотеки» та опрацювали близько 6 млн одиниць ЛК, аби толокою вибрати з-поміж них картки, позначені штампом «АРХІВНА».

2018 р. в Інституті української мови НАН України зреалізовано проєкт[2] зі створення цифрового формату лексико-фразеологічних матеріалів КСЖУМ Всеукраїнської Академії наук. Силами невеликого колективу[3] засканована вибрані картки і створено **електронну систему «Архівна картотека»** (Архівна картотека, 2020) (рис. 2). Окрім загальної інформації про історію цього знадобу та його збереження, на сайті міститься основне — власне зацифровані картки з можливим пошуком картки за потрібним словом, копіюванням тексту й зображення картки (Тищенко, 2020).

Без сумніву, ця електронна система має стати неодмінним складником Нової електронної картотеки — лексикографічного знадобу[4], щоб уже на цій основі разом з іншими джерелами (словниками, картотеками, корпусами тощо) уможливити укладання нових словників (рис. 3).

Рис. 3. Лексикографічний електронний знадіб

Рис. 2. Електронна система «Архівна картотека»: вебінтерфейс (2018–2021)

Фрагмент АК розпізнано, тобто набрано вручну тексти карток відповідно до полів картки (близько 3 000 карток). Якщо картку ще не розпізнано, її можна переглянути тільки як зображення, тому натепер пошук нерозпізнаних карток за заданим словом неможливий. Розпізнавання карток — це набирання тексту у відповідні поля, що відбивають мікроструктуру картки, наприклад заголовні слова або описові конструкції: *смеркнути* —

[2] За підтримки Українського культурного фонду, грант № 1109, вересень — листопад 2018 р.

[3] Кандидатка філологічних наук О. Тищенко (керівниця проєкту) й докторка філологічних наук Л. Кислюк (наукові співробітниці Інституту української мови НАН України); кандидатка філологічних наук Ю. Поздрань (старша викладачка Вінницького національного технічного університету) та Ю. Вознюк (аспірантка Інституту української мови НАН України). Електронну систему зреалізовано за участі програмістів на чолі з М. Ткаченком.

[4] Див. Тищенко, О. (2019). Електронна лексична картотека: шлях створення інструментарію сучасного словникаря. *Українська мова, 2,* 37–52.

ISSN 2311-259X

наступить сумеркам (рис. 4), також ми послідовно фіксували всі виправлення в картках або пізніші додавання елементів: до *стемнеть* додано *повечереть,* що ми їх фіксуємо як додаткові до заголовного слова одиниці (у цьому випадку — синонім) (Тищенко, 2020). Усе це — для зручного пошуку потрібної інформації, аналізу макроструктури картотеки, дослідження зафіксованого в ній фактажу.

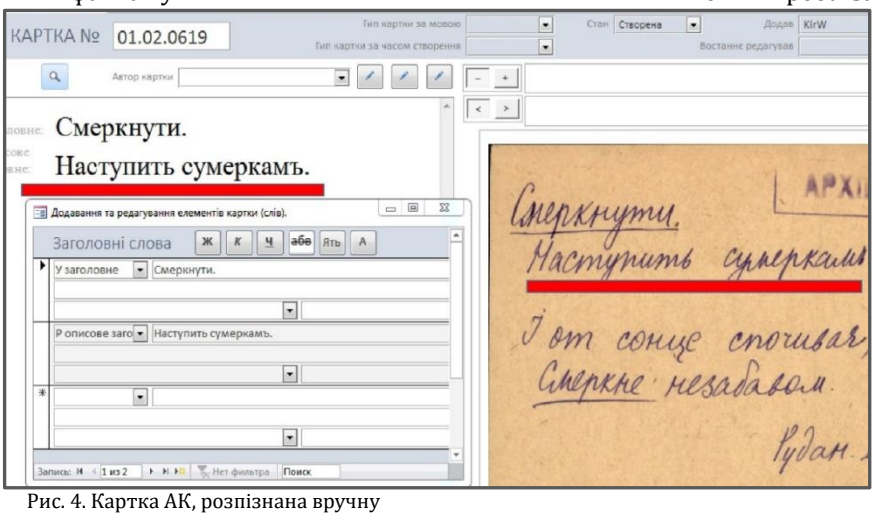

Рис. 4. Картка АК, розпізнана вручну в системі «Архівна картотека»

Однак ручне розпізнавання потребує значних зусиль великого колективу — із 350 тисяч карток АК набрано близько трьох тисяч. Тому ми бачимо сенс застосувати систему автоматичного розпізнавання тексту — програму **Transkribus** (2020). Це комплексна платформа[5] для автоматизованого розпізнавання, транскрипції та пошуку документів, яка складається з експертного інструмента (Transkribus), вебінтерфейсу (http://transkribus.eu) та кількох хмарних сервісів. Основною метою Transkribus є підтримка користувачів, які займаються транскрипцією друкованих або рукописних документів, — філологів, істориків, архівістів, представників громадськості та ін. Технологія, на яку спирається Transkribus, нова, і можливість тестування цього програмного забезпечення для розпізнавання рукописного тексту є важливою як для користувачів, так і для розробників (Данли, 2018).

Проєкт READ надає загальну модель, але наразі програму потрібно навчити розуміти конкретний стиль написання документів. Рекомендовано починати з приблизно 15–20 тис. слів для рукописного тексту — це набір даних, які можна використовувати для навчання механізму HTR[6] (для АК це близько 1 000 карток). Для забезпечення двигуна HTR навчальними даними зображення треба завантажити на платформу, сегментувати кожне зображення на рядки й текстові ділянки за допомогою автоматичних інструментів, а потім транскрибувати вміст кожної сторінки (рис. 5).

Щойно базові лінії з'являться на зображенні, можна вводити інформацію в текстовому редакторі — для цього автоматично формується поле введення транскрипції. Для кожної базової лінії в текстовому редакторі є відповідний рядок. Транскрибувати текст треба порядково, точно відтворюючи його розміщення на зображенні: 1. *Скеля* 2. *Скала* тощо (рис. 6).

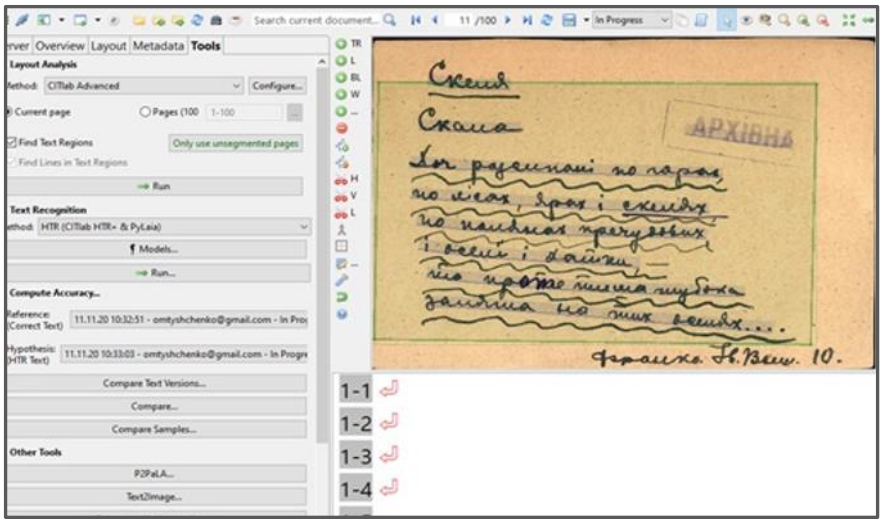

Рис. 5. Підготування картки до розпізнавання в Transkribus: сегментування на рядки

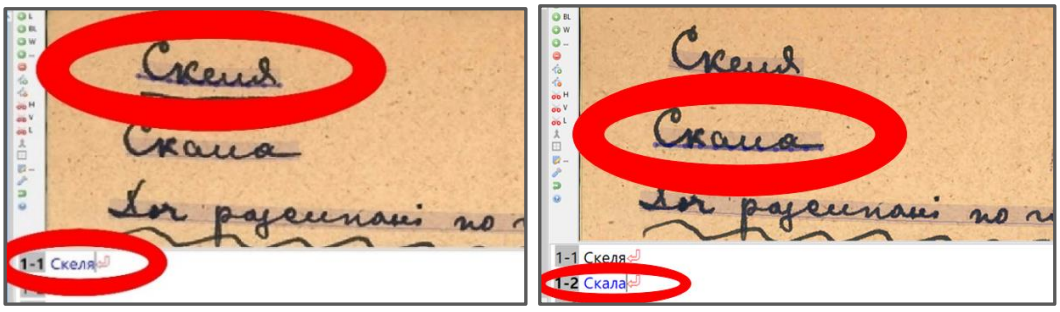

Рис. 6. Транскрибування тексту в Transkribus

[5] Фінансує Європейська Комісія в межах проєкту READ.

[6] HTR (*англ.* handwritten text recognition) — програмне забезпечення для розпізнавання рукописного тексту.

ISSN 2311-259X

Перед набиранням транскрипції треба виправити **огріхи сегментування на рядки**. Насамперед усуваємо *нетекстові елементи* — описки, плями, вади паперу, помилково марковані базовими лініями (рис. 7).

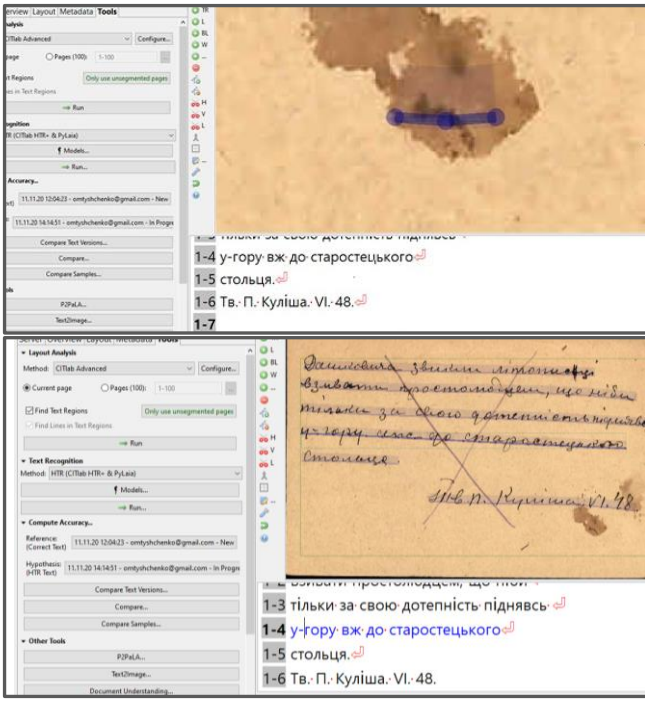

Рис. 7. Усунення огріхів: нетекстові елементи

Вилучаємо з поля транскрипції *непосутні текстові елементи*, які не входять до змісту картки. Передусім це напис штампу «Архівна», який є на всіх картках і не має лексикографічної цінності для жодної з них (рис. 8). Також це можуть бути уваги лексикографа, наприклад знак запитання в кутку картки, напис *ст. / ст / Ст* на вагомій частині карток (можливо, так маркували *старі* за якоюсь ознакою картки), напис *дуб. / дуб / дубл* (означає, що для цієї картки створили *дублікат*, замінивши ним оригінал у великій ЛК). Під час ручного розпізнавання такі уваги вводимо в окреме поле, однак вони не є об'єктом для пошуку філологічної інформації.

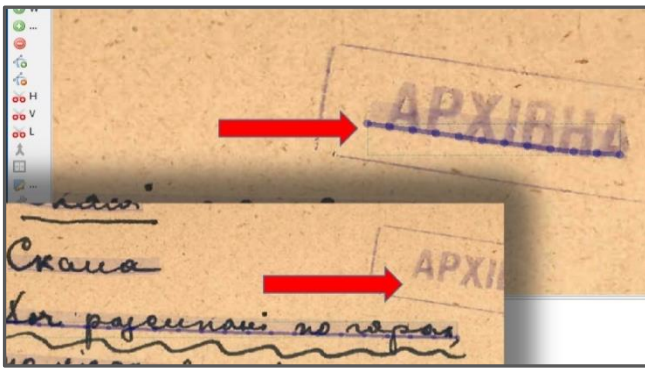

Рис. 8. Усунення огріхів: непосутні текстові елементи

Огріхом розбивання на рядки може бути *некоректне створення базових ліній*, як-от неповне охоплення тексту в рядку, яке ми усуваємо вручну, продовживши лінію (рис. 9). Часто в одному рядку формується кілька базових ліній, що в полі для транскрипції репрезентовано окремими рядками — це потребує ручного об'єднання кількох базових ліній в одну.

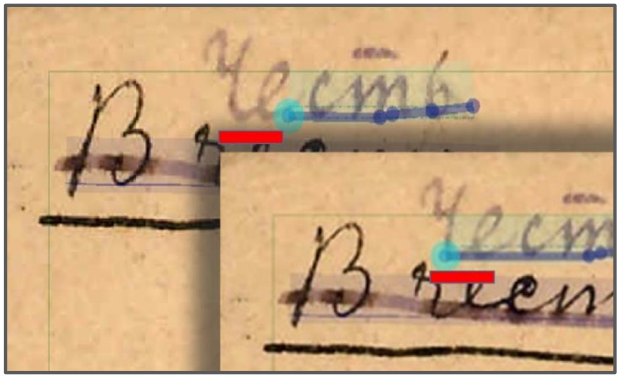

Рис. 9. Усунення огріхів: некоректне охоплення тексту

Ще огріхом автоматичного створення базових ліній є *некоректне визначення текстового регіону*, поля тексту, наприклад помилкове сегментування на стовпці суцільного тексту або об'єднання в один стовпець різнорядних елементів, коли випадкова відстань між його елементами в одному рядку створює хибне враження розподілу на стовпці (рис. 10). Це також усуваємо вручну, формуючи текстовий регіон не як два стовпці, а як текст із суцільною нумерацією рядків.

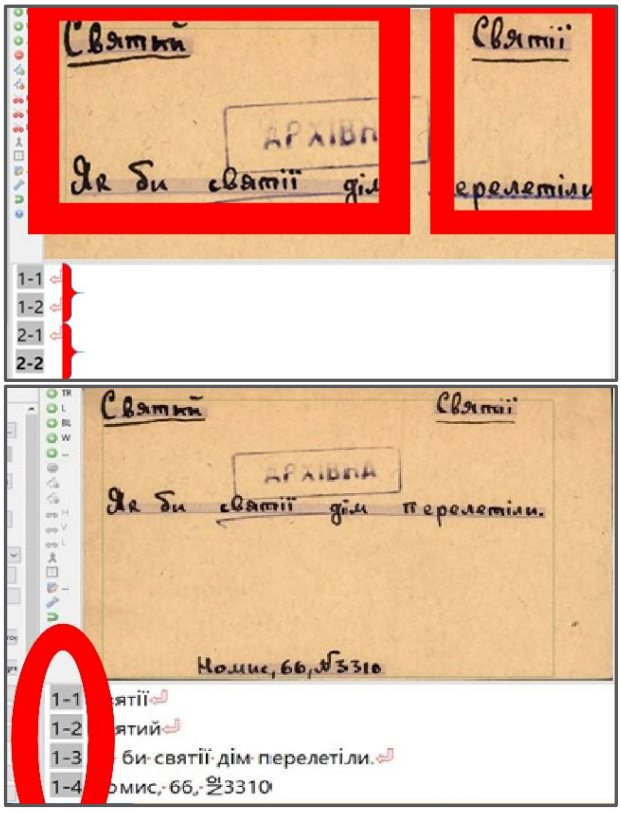

Рис. 10. Усунення огріхів:
помилкове сегментування на стовпці

Отримавши коректне поле для введення транскрипції, **набираємо текст**. Велика кількість карток містять *виправлення (правки) в тексті*,

ISSN 2311-259X

які ми розрізнювали під час ручного розпізнавання: фіксували пізнішу вставку або закреслення і вставку — динаміка оформлення картки демонструє пошук слова, уточнення та ін. Звичайні ж описки, що не мають посутнього значення, ми не фіксували, аби не засмічувати поля набраних карток. Однак вважаємо, що для Transkribus такі правки вводити доцільно для тренування програми (рис. 11).

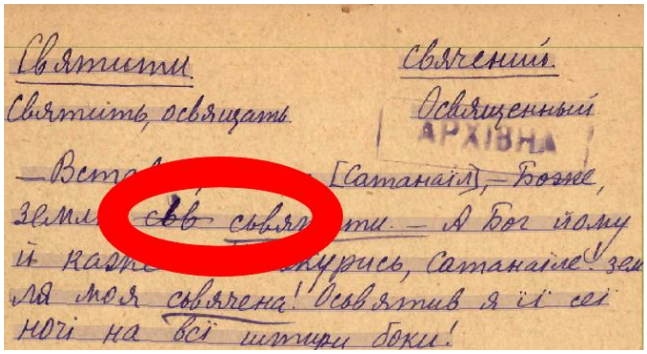

Рис. 11. Правки в тексті: в системі «АК» та в Transkribus

Transkribus пропонує також низку інструментів для визначення *стилю напису*: заголовок, колонтитул, виноска тощо, що теж можна використовувати як параметризацію заголовних чи додаткових слів, цитати, паспорта, а також граматичних, етимологічних, семантичних відомостей та ін. (рис. 12). Також можливим є форматування тексту приступними засобами текстового редактора Word: підкреслення, закреслення, курсив тощо.

Створивши достатній набір даних, можемо зв'язатися з командою Transkribus, яка активує для нас навчання в програмі. Після завершення цього процесу ми оберемо свою модель HTR та застосуємо її до решти сторінок у нашій колекції документів. Для цього маємо підготувати решту зображень: визначити текстові регіони та рядки / базові лінії — це ще один великий фрагмент роботи. Можна буде скористатися текстовим редактором у Transkribus, щоб виправити автоматично розпізнаний текст, але це наступний етап опрацювання АК, який передбачатиме й вимірювання ефективності HTR та OCR[7] (для невеликої кількості друкованих карток) із коефіцієнтом помилок Word та коефіцієнтом помилок символів.

Дослідники багатьох країн покладають великі надії на цю нову модель розпізнавання рукописних текстів, однак для безумовного опертя на штучний інтелект час іще не настав — програма потребує вдосконалення шляхом тривалих тренувань на великих обсягах інформації. Однак уже зараз технологія цього типу пропонує інші потенційні можливості, насамперед це пошук за орієнтирними, ключовими словами, що оптимізує використання архівних даних уже найближчим часом навіть за недостатньо високого рівня точності розпізнавання. Це пов'язано з тим, що транскрипція відбиває тільки один варіант для певного слова, тоді як саме програмне забезпечення генерує безліч можливих варіантів для кожного слова. З урахуванням цих варіантів можна знаходити правильне слово з набагато більшою ймовірністю (Данли, 2018). Для швидшого опрацювання в Transkribus передбачено колективний доступ до колекції транскрибованих документів у Transkribus Web interface — це полегшена версія Transkribus, зручна й проста для вжитку. Тож усі охочі можуть доєднатися до нашої спільної справи.

Пришвидшити опрацювання карток можна й у межах нашого нового проєкту «*Усеукраїнська толока: Архівна картотека*» (2020) — онлайнової платформи на сайті «АК» для *ручного розпізнавання* текстів карток усіма охочими. Поки ми будемо навчати штучний інтелект розпізнавати старі рукописні картки автоматично, зацікавлені люди зможуть зареєструватися та взяти участь в удоступненні текстового масиву АК для широкого кола користувачів у межах *лексикографічної толоки*. Також буде потреба вичитувати розпізнані автоматично тексти та вводити інформацію у відповідні поля системи «АК» — для автоматизації роботи з матеріалами картотеки.

Однак електронна система — це тільки *інструмент* вивчання лексикографічних клейнодів Архівної картотеки. Тому перспективними й важливими є власне наукові студії АК в лінгвістичній, лінгвокультурологічній оптиці, аби не дати карткам укритися зеленою тугою та безвістю заборон, використати їх потенціал для покріплення нашого питомого мовного «я» в сучасних вимірах.

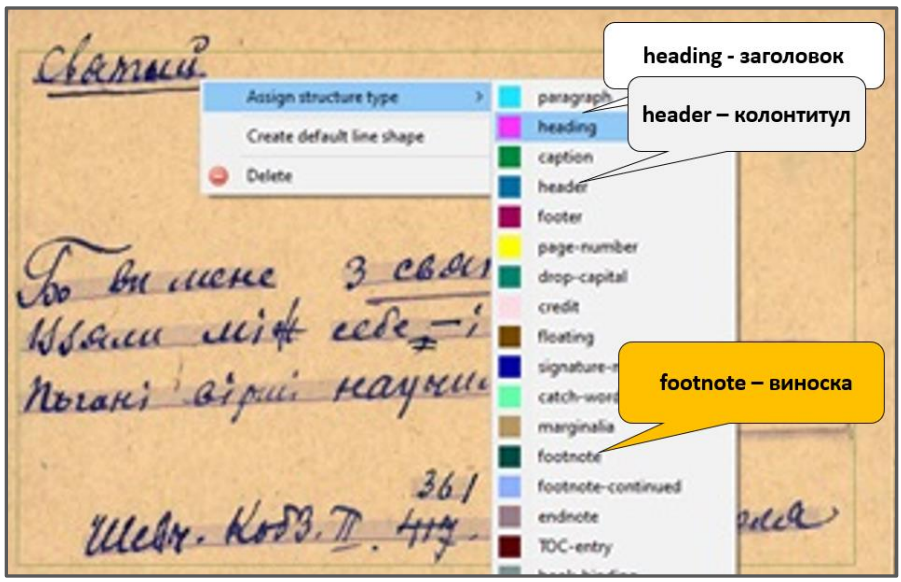

Рис. 12. Стиль тексту в Transkribus

[7] OCR (*англ.* optical character recognition) — технологія оптичного розпізнавання символів, спрямована на друковані тексти.

ISSN 2311-259X

Покликання

Архівна картотека (2020). https://ak.iul-nasu.org.ua

Данли, Р. (2018). *Машины читают архивные документы: программное обеспечение для распознавания рукописного текста.* Блог Национальных Архивов Великобритании. http://blog.nationalarchives.gov.uk/blog/machines-reading-the-archivehandwritten-text-recognition-software/
Цит. за перекладом https://tsdea.archives.gov.ua/wp-content/uploads/2018/03/26032018_st.pdf

Кримський, А., Єфремов, С. (Гол. ред.). (1924–1933). *Російсько-український словник.* Т. I–III. Київ.

Поздрань, Ю. (2018). *«Російсько-український словник» за редакцією А. Ю. Кримського та С. О. Єфремова в історико-лінгвістичному контексті*.

Російсько-українські словники (2021). https://r2u.org.ua

Тищенко, О. (2016). Архівна картотека як лексико-ілюстративна база «Російсько-українського словника» за ред. А. Ю. Кримського та С. О. Єфремова. I. Лексична картотека: історія створення та репресій; II. Мікро- та макроструктура архівної картотеки. *Українська мова, 2,* 44–71; *3,* 57–78.

Тищенко, О. (2020). Архівна картотека української мови в цифровому форматі: від пам'ятки мови до сучасного лексикографічного інструментарію. *Rocznik Slawistyczny, LXIX,* 185–197.

Усеукраїнська толока: Архівна картотека (2020) http://work.iul-nasu.org.ua

Transkribus (2021). https://readcoop.eu/transkribus

References (translated and transliterated)

Arkhivna kartoteka [Archival Card Index] (2018–2021). https://ak.iul-nasu.org.ua

Danli, R. (2018). Mashiny chitayut arkhivnye dokumenty: programmnoe obespechenie dlya raspoznavaniya rukopisnogo teksta [The machines read archival documents: handwriting recognition software]. Blog Natsionalnykh Arkhivov Velikobritanii. http://blog.nationalarchives.gov.uk/blog/machines-reading-the-archivehandwritten-text-recognition-software/
Quoted from translation: https://tsdea.archives.gov.ua/wp-content/uploads/2018/03/26032018_st.pdf

Krymskyi, A., Yefremov, S. (Ed.). (1924–1933). *Rosiisko ukrainskyi slovnyk* [Russian-Ukrainian Dictionary]. Vol. I–III.

Pozdran, Yu. (2018). *"Rosiisko-ukrainskyi slovnyk" za redaktsiieiu A. Yu. Krymskoho ta S. O. Yefremova v istoryko-linhvistychnomu konteksti* ["Russian-Ukrainian Dictionary" edited by A. Yu. Krymsky and S. O. Yefremov in the historical-linguistic context].

Rosiisko-ukrainski slovnyky [Russian-Ukrainian Dictionaries] (2021). https://r2u.org.ua

Transkribus (2021) https://readcoop.eu/transkribus/

Tyshchenko, O. (2016). Arkhivna kartoteka yak leksyko-iliustratyvna baza "Rosiisko-ukrainskoho slovnyka" za red. A. Yu. Krymskoho ta S. O. Yefremova. I. Leksychna kartoteka: istoriia stvorennia ta represii; II. Mikro- ta makrostruktura arkhivnoi kartoteky [The archival card index as the lexical and illustrative base of "Russian-Ukrainian dictionary" ed. A. Krymsky and S. Yefremov. I. Lexical card index: history of creation and repression; II. Micro- and macrostructure of archival lexical card index]. *Ukrainska mova, 2,* 44–71; *3,* 57–78.

Tyshchenko, O. (2020). Arkhivna kartoteka ukrainskoi movy v tsyfrovomu formati: vid pamiatky movy do suchasnoho leksykohrafichnoho instrumentariiu [Archival card index of the Ukrainian language in digital format: from a language monument to modern lexicographic tools]. *Rocznik Slawistyczny, LXIX,* 185–197.

Useukrainska toloka: Arkhivna kartoteka [All-Ukrainian Toloka: Archival Card Index] (2020) http://work.iul-nasu.org.ua

Oksana Tyshchenko
Institute of Ukrainian Language of the NAS of Ukraine, Ukraine

ARCHIVE CARD INDEX VS. TRANSKRIBUS: MACHINE RECOGNITION OF HANDWRITTEN TEXT

The subject of the research is machine recognition of handwritten materials of the Archival Card Index (ACI) — lexical and phraseological materials of the dictionary commission of the All-Ukrainian Academy of Sciences, in particular, card index of the "Russian-Ukrainian dictionary" 1924–1933 ed. A. Krymsky and S. Yefremov. The study of the ACI should be considered in the context of cultural and national revival in Ukraine in the 20th — early 21st centuries. The relevance and value of the ACI became a prerequisite for the transfer of its materials to the digital format. In 2018 the Institute of Ukrainian Language of the NAS of Ukraine created a computer system "Archival Card Index", which accessibles materials primarily in the form of scanned images. The problem that needs urgent resolution is the transfer of handwriting to a typewriter format. The complexity of manual recognition, which requires considerable effort and time, encourages the study and application of *Transkribus* resource capabilities, which involves the use of the machine teaching. The Aim of the study is to clarify by analyzing, systematizing, classifying and describing the material features of the preparation of ACI cards for machine processing of texts. The scientific novelty of the study is that for the first time, the issue of providing the HTR engine with ACI training data (loading to the platform, segmenting images into lines and text areas, transcribing content each page).

The main result is finding out the content of the preparatory stage, the tasks of which are to eliminate the flaws of automatic segmentation: non-text elements, non-substantial text elements, incorrect automatic detection of text region or line. The prospects of lexicographic toloka (crowdsourcing) in the process of card recognition are outlined, for which it is envisaged to use collective access to the collection of transcribed documents in *Transkribus*. To recognize the cards manually and for the future check and adjustment of automatically recognized ones, you can join the new project "All-Ukrainian Toloka: Archival Card Index" — online platform on the website "ACI".

Keywords: lexicography; Archive Card Index; Electronic System "Archive Card Index"; machine recognition; Transkribus; lexicographical toloka (crowdsourcing).

Стаття надійшла до редколегії 03.02.2021