ROCZNIKI HUMANISTYCZNE
Tom LXXIII, zeszyt 2 – 2025
ZESZYT SPECJALNY
DOI: https://doi.org/10.18290/rh25732.5s

MAGDALENA MAJDAK
EWA RODEK
DOROTA ADAMIEC
KATARZYNA KRYŃSKA
JAGODA MARSZAŁEK

BAZA HISTORYCZNYCH LEKSYKONÓW POLSKICH. EDYCJA CYFROWA SŁOWNIKÓW Z XVII I XVIII WIEKU[1]

Abstrakt. Artykuł przedstawia koncepcję oraz efekty wstępnej fazy realizacji grantu „Digitalizacja trzeciego stopnia wielkich słowników XVII i XVIII wieku – stworzenie Bazy Historycznych Leksykonów Polskich" (BazHiLek), realizowanego w Pracowni Historii Języka Polskiego XVII i XVIII wieku IJP PAN (finansowanie: NPRH 2024–2029). Pierwszy etap prac poświęcono edycji cyfrowej trzech słowników: *Thesaurusa* Knapiusza (1643, wyd. II), *Nowego dykcjonarza* Troca (1764, t. III) oraz *Forytarza* Ernestiego (1674), stanowiących zalążek Bazy. Leksykony te odznaczają się oryginalnym warsztatem i są bogatym źródłem informacji lingwistycznych. Przygotowanie materiału wymagało połączenia działań filologicznych i informatycznych: konwersji obrazów do postaci tekstowej, rozpoznania mikrostruktury artykułów hasłowych, przygotowania modeli OCR, wyboru znaczników TEI,

Dr hab. MAGDALENA MAJDAK, prof. IJP PAN – Polska Akademia Nauk, Instytut Języka Polskiego; adres do korespondencji: al. A. Mickiewicza 31, 31-120 Kraków; e-mail: magdalena.majdak@ijppan.pl; ORCID: 0000-0002-5261-5141.

Dr EWA RODEK – Polska Akademia Nauk, Instytut Języka Polskiego; adres do korespondencji: al. A. Mickiewicza 31, 31-120 Kraków; e-mail: ewa.rodek@ijppan.pl; ORCID: 0000-0002-9967-0332.

Dr DOROTA ADAMIEC – Polska Akademia Nauk, Instytut Języka Polskiego; adres do korespondencji: al. A. Mickiewicza 31, 31-120 Kraków; dorota.adamiec@ijppan.pl; ORCID: 0000-0001-8179-038X.

Dr KATARZYNA KRYŃSKA – Polska Akademia Nauk, Instytut Języka Polskiego; adres do korespondencji: al. A. Mickiewicza 31, 31-120 Kraków; e-mail: katarzyna.krynska@ijppan.pl; ORCID: 0000-0003-3446-122X.

Mgr JAGODA MARSZAŁEK – Polska Akademia Nauk, Instytut Języka Polskiego; adres do korespondencji: al. A. Mickiewicza 31, 31-120 Kraków; e-mail: jagoda.marszalek@ijppan.pl; ORCID: 0000-0003-1694-544X.

[1] Artykuł powstał w ramach grantu NPRH „Digitalizacja trzeciego stopnia wielkich słowników XVII i XVIII wieku − stworzenie Bazy Historycznych Leksykonów Polskich" (nr NPRH/DN/SP/0003/2023/12).

strukturalnego oznakowania materiału i opracowania reguł automatycznego tagowania danych pod kątem przeszukiwania i porównywania zasobów leksykograficznych w Bazie. Celem projektu jest opracowanie standardów umożliwiających poszerzanie jej o kolejne słowniki oraz stworzenie zaawansowanej wyszukiwarki. W artykule omówiono poszczególne zadania i trudności w ich realizacji oraz zarysowano planowane etapy prac.

Słowa kluczowe: metaleksykografia; leksykografia historyczna; słownik wielojęzyczny; Knapiusz; Troc; Ernesti; modelowanie OCR; OCR4all; ligatury; abrewiatury

THE DATABASE OF HISTORICAL POLISH LEXICONS.
DIGITAL EDITION OF THE 17TH AND 18TH CENTURY DICTIONARIES

Abstract. The article presents the concept and outcomes of the initial phase of the grant project entitled "Third-level digitisation of large 17th- and 18th-century dictionaries: The development of the Database of Historical Polish Lexicons", carried out at the Department of the History of the 17th and 18th Century Polish, Institute of the Polish Language, PAS (NPRH 2024–2029). The first stage of the project focused on the digital edition of three dictionaries: Knapiusz's *Thesaurus* (1643, 2nd edition), Troc's *Nowy dykcjonarz* (1764, vol. III), and Ernesti's *Forytarz* (1674), which will form the foundation of the Database. These lexicons are notable for their original scholarly methods and constitute a rich source of linguistic data. Their preparation required the integration of philological and computational methods: converting scans into machine-readable text, analyzing the microstructure of lexical entries, selecting appropriate TEI tags, structural markup, training OCR models, and developing rules for automatic TEI-based annotation to enable advanced search and comparative analysis within the Database. The projects aims to establish standards for incorporating further historical dictionaries into it and to develop a robust search engine. The article discusses individual tasks and difficulties in their implementation, and outlines the planned stages of work.

Keywords: metalexicography; historical lexicography; multilingual dictionary; Knapiusz; Troc; Ernesti; OCR modelling; OCR4all; ligatures; abbreviatons

WPROWADZENIE

Projekt „Digitalizacja trzeciego stopnia wielkich słowników XVII i XVIII wieku − stworzenie Bazy Historycznych Leksykonów Polskich" (BazHiLek) powstaje w Pracowni Historii Języka Polskiego XVII i XVIII w. Instytutu Języka Polskiego PAN dzięki wsparciu Narodowego Programu Rozwoju Humanistyki na lata 2024–2029. Głównym celem tego przedsięwzięcia jest stworzenie platformy cyfrowej gromadzącej słowniki dawne i umożliwiającej ich zaawansowane przeszukiwanie. W niniejszym artykule przedstawiono koncepcję grantu oraz efekty realizacji jego pierwszego etapu.

Zalążek Bazy mają stanowić dwa najważniejsze dzieła w historii polskiej leksykografii XVII i XVIII wieku: *Thesaurus Polonolatinograecus seu promptuarium linguae Latinae et Graecae* Grzegorza Knapiusza (1643, wyd. II, rozszerzone, t. 1)

oraz *Nowy dykcjonarz to jest mownik polsko-niemiecko-francuski z przydatkiem przysłów potocznych, przestróg gramatycznych, lekarskich, matematycznych, fortyfikacyjnych, żegla*[*r*]*skich, łowczych i inszym naukom przyzwoitych wyrazów* Michała Abrahama Troca (1764) , a także *Forytarz języka polskiego, osobne, rzeczy niemal wszystkich, słowa*, *mowy różne i rozmowy dwie w sobie zawierający* Jana Ernestiego (1674) ze względu na jego oryginalną budowę. Głęboka digitalizacja tych źródeł ułatwi dostęp do zawartego w nich materiału, którego znaczenie dla badań nad polszczyzną XVII–XVIII wieku jest trudne do przecenienia. Wybór tak zróżnicowanych i skomplikowanych, a jednocześnie sztandarowych dzieł polskiej leksykografii historycznej został dokonany ze względu na potrzebę wypracowania metod umożliwiających poszerzanie Bazy o kolejne słowniki cechujące się dowolną mikrostrukturą.

W pierwszej kolejności uwagę poświęcono elektronicznej edycji trzech wymienionych dzieł. Skany egzemplarzy *Thesaurusa* i *Nowego dykcjonarza* pozyskano z biblioteki cyfrowej CRISPA Uniwersytetu Warszawskiego, a *Forytarz* z Die Sächsische Landesbibliothek – Staats- und Universitätsbibliothek. Przygotowanie materiału do głębokiej digitalizacji (w rozumieniu Piotra Żmigrodzkiego[2]) wymagało równoległych działań filologicznych i informatycznych. Podjęto szczegółowe analizy makro- i mikrostruktury w perspektywie porównawczej oraz pod kątem stworzenia reguł pozwalających na wyszukiwanie danych w obrębie poszczególnych elementów artykułu hasłowego w dowolnym słowniku umieszczonym w Bazie.

CELE I ZAŁOŻENIA

W projekcie zadeklarowano realizację następujących zadań badawczych:

1. Przeprowadzenie analizy makro- i mikrostruktury wybranych słowników.
2. Automatyczne rozpoznanie tekstu i korekta wyników.
3. Automatyczna anotacja struktury słowników za pomocą znaczników TEI XML i korekta wyników.
4. Dodanie transkrypcji, przeprowadzenie automatycznej lematyzacji i anotacji morfosyntaktycznej materiału.
5. Stworzenie multiwyszukiwarki słownikowej na platformie TEI Publisher.

[2] Piotr Żmigrodzki, *Słowo – słownik – rzeczywistość. Z problemów leksykografii i metaleksykografii* (Kraków: Universitas, 2008), 102–3.

6. Opracowanie reguł API umożliwiających automatyczne pobieranie treści z platformy i integrowanie ich z innymi zasobami językowymi, w tym z *Elektronicznym słownikiem języka polskiego XVII i XVIII wieku*.

7. Przeprowadzenie badań leksykograficznych i leksykologicznych na pozyskanym materiale.

Artykuł zawiera omówienie zadań realizowanych w pierwszym roku trwania grantu.

STRUKTURA SŁOWNIKÓW

Sposób opisu materiału leksykalnego w słownikach z XVII i XVIII wieku znacznie różnił się od dzisiejszego, w związku z tym niejednokrotnie konieczne było rozstrzyganie kwestii podstawowych, takich jak postać wyrazu (wyrażenia) hasłowego czy zakres definicji (opisu semantycznego). Główna trudność polegała nie tyle na odkrywaniu koncepcji autorów i rekonstrukcji warsztatu leksykograficznego (choć także), ile na mierzeniu się ze specyfiką opisu. Poczynione ustalenia stały się podstawą katalogu wyabstrahowanych wariantów strukturalnych artykułów hasłowych wyodrębnionych ze względu na różnice w układzie elementów. Z uwagi na skomplikowaną strukturę, wielojęzyczność (polszczyzna, łacina, greka, niemiecki, francuski), układy alfabetyczny i *a tergo* oraz różne rodzaje czcionek (fraktura, szwabacha, antykwa, kursywa) każde z dzieł wymagało przygotowania osobnego modelu odczytania tekstu. Niejednorodność materiału sprawiła, że było to znaczące wyzwanie, którego realizacja – jak już można stwierdzić – zakończyła się powodzeniem. W kolejnym etapie poszczególne składowe artykułów hasłowych zostaną poddane znakowaniu. Prace nad wyborem i przypisaniem im znaczników TEI podjęto z uwzględnieniem najnowszych wytycznych TEI Lex-0 – *A baseline encoding for lexicographic data*[3].

Ze względu na zróżnicowanie struktury słowników konieczne było przeprowadzenie standaryzacji na takim poziomie ogólności, który z jednej strony umożliwiałby wprowadzenie historycznego materiału leksykograficznego do

[3] Toma Tasovac, Laurent Romary, Piotr Banski, Jack Bowers, Jesse de Does, Katrien Depuydt, Tomaž Erjavec, Alexander Geyken, Axel Herold, Vera Hildenbrandt, Mohamed Khemakhem, Boris Lehečka, Snežana Petrović, Ana Salgado, Andreas Witt, *TEI Lex-0: A Baseline Encoding for Lexicographic Data. Version 0.9.4. DARIAH Working Group on Lexical Resources*, 2018, https://dariah-eric.github.io/lexicalresources/ pages/TEILex0/ TEILex0.html [dostęp: 2.04.2025].

bazy w sposób jak najpełniejszy, z drugiej strony zatrzymywał analizę na możliwym, wspólnym dla wszystkich słowników poziomie. Pogłębione interpretacje będą powstawały w ramach indywidualnych prac badawczych.

MAKROSTRUKTURA

Knapiusz i Troc dążyli do możliwie pełnego opisu zasobu leksykalnego języka polskiego, choć różnili się w podejściu do doboru materiału: Knapiusz, dbając o czystość języka, ograniczał obecność wyrazów uznawanych za wulgarne (choć niektóre z nich rejestrował, opatrując stosownymi kwalifikatorami), natomiast Troc włączał do siatki haseł nie tylko terminologię specjalistyczną (z zakresu prawa, medycyny czy teologii), ale również liczne nazwy własne. Ernesti w *Przedmowie* odwołał się do celów glottodydaktycznych swojego słownika: „onego do Podánia do Druku publičznego/ nic inszego/ jáko Niedostátek sposobney jákiey kśiążečzki/ dla Młodźi Języká Polskiego učzyć śię prágnącey/ ná Autorze [...] wymogł”[4]. Zgromadzony w *Forytarzu* zasób leksykalny stanowił więc zamierzony przez autora wybór słownictwa polskiego. Znacząca większość haseł to rzeczowniki, jednak reprezentowane są też inne klasy gramatyczne (por. *czwórnasob*, *trzykroć*, *rad*, *wzajem*, *naukos*, *latoś*, *moy*, *czyż*). Obok rzeczowników pospolitych wśród haseł występują także nazwy własne różnego typu.

Estreicher opisał *Forytarz* jako „Dzieło ułożone w formie słownika polsko-niemieckiego, nie abecadłowo, bez wyjaśnienia systemu w układzie zdań”[5]. Nie było to stwierdzenie całkowicie uprawnione. O ile Knapiusz i Troc stosowali w układzie haseł porządek alfabetyczny, o tyle u Ernestiego najogólniejszą regułę układu artykułów hasłowych stanowił porządek *a tergo*. Został on zastosowany w kilku kolejnych ciągach wydzielonych na podstawie niezbyt przejrzystych kryteriów. Przedstawione wyżej obserwacje pokazują znaczną różnorodność makrostruktury omawianych słowników, która może wynikać m.in. z odmiennych celów i założeń autorów, choć pełne wyjaśnienie tych różnic wymaga dalszych szczegółowych badań.

[4] Jan Ernesti, *Forytarz języka polskiego, osobne, rzeczy niemal wszystkich, słowa, mowy różne i rozmowy dwie w sobie zawierający*, t. 16 (Wrocław: Drukarnia Mikołaja Kornachera, 1674), ijv.

[5] Karol Estreicher, *Bibliografia polska*, t. 16 (Kraków: Akademia Umiejętności, 1898), 97.

MIKROSTRUKTURA

Analizowane tu leksykony mają charakter przekładowy – zasadniczymi elementami mikrostruktury są tu więc, oprócz wyrazu hasłowego, obcojęzyczne odpowiedniki, najczęściej w postaci równoznacznych wyrazów obcych: łacińskich, greckich, niemieckich, francuskich. Niekiedy jednak ekwiwalenty te mogą przybrać postać opisową, zbliżoną do definicji realnoznaczeniowej charakterystycznej dla słowników ogólnych. Niekonsekwencje te powodują konieczność uwzględnienia w modelu reprezentacji ich treści w bazie licznych pól opcjonalnych. Na przykład występująca często w słowniku Troca informacja gramatyczna może mieć dwojaki charakter: oznaczenia metajęzykowego (*adv.*, *conj.*, *f.*, *g.*, *plur.*) albo zakończenia poprzedzonego skrótem odpowiedniej kategorii gramatycznej. U Knapiusza ten element mikrostruktury pojawia się rzadko, u Ernestiego zaś go brak. W strukturze bazy pole takie należy jednak uwzględnić, w dodatku z możliwością wypełnienia go bardzo zróżnicowanym typem danych. Mikrostruktura wszystkich analizowanych słowników charakteryzuje się daleko posuniętym brakiem regularności w zawartości kolejnych elementów artykułów hasłowych (takich jak wyraz hasłowy, informacja gramatyczna, ilustracja materiałowa, związki wyrazowe, podhasła i inne).

AUTOMATYCZNE ROZPOZNANIE TEKSTU SŁOWNIKÓW

Do uzyskania wersji edytowalnej słowników posłużył program OCR4all. Jego twórcy wykorzystali różne narzędzia i silniki obliczeniowe: do rozpoznawania tekstu silnik obliczeniowy Calamari, natomiast do segmentacji, podziału na linie odczytu i korekty – program Larex. Przebieg poszczególnych procesów jest zoptymalizowany i czytelny dla użytkownika. Program OCR4all sam dostosowuje format i rozdzielczość dostarczonych skanów tekstów, a następnie dokonuje redukcji ewentualnych zanieczyszczeń obrazu.

Prace w programie OCR4all odbywały się dwutorowo: stworzono modele do odczytu każdego ze słowników, a następnie partiami sprawdzano wyniki automatycznego rozpoznania tekstu. Przygotowanie modeli stanowiło duże wyzwanie ze względu na wspomnianą wielojęzyczność, a także druk z użyciem rozmaitych czcionek i wykorzystanie alfabetów pochodzących z różnych systemów pisma. Wielość znaków drukarskich utrudniała odpowiednie skomponowanie zestawu tekstów do trenowania modelu. Aby ograniczyć liczbę alografów, a tym samym usprawnić sam proces, podjęto decyzję o wytrenowaniu osobnych

modeli dla każdego słownika. Dzięki temu w niedługim czasie udało się uzyskać satysfakcjonujące wyniki. Wskaźnik ewaluacyjny przyjęty w programie OCR4all, czyli średni znormalizowany współczynnik błędów etykiet (znaków) (*Mean Normalized Label Error Rate*), w automatycznym odczycie *Forytarza* Ernestiego wyniósł około 3%, w odczycie *Nowego Dykcjonarza* Troca około 1,5%, a *Thesaurusa* Knapiusza około 1%, co świadczy o średniej i wysokiej jakości rozpoznania tekstu[6].

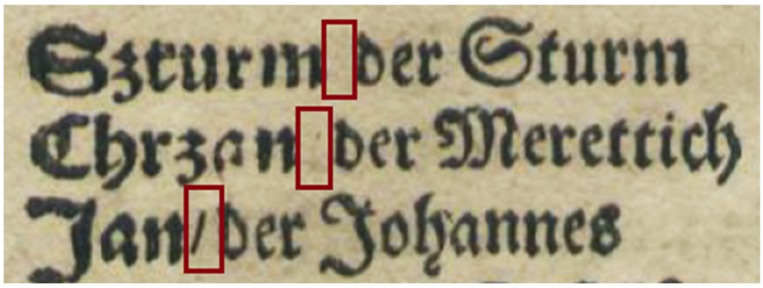
Pewnym wyzwaniem okazał się dwukolumnowy układ każdej z publikacji. Niestety ani program OCR4all, ani Larex (do którego OCR4all się odwołuje) nie dają możliwości wytrenowania modelu struktury tekstu lub oznaczenia typu układu tekstu na rozpoznawanej stronie[7], dlatego wyznaczanie regionów odczytu przeprowadzono ręcznie.

Na etapie trenowania modeli OCR, a następnie korekty odczytu zastosowano różnego typu zabiegi redaktorskie, ujednolicające rozpoznawany tekst haseł słownikowych i ułatwiające przeprowadzenie automatycznego znakowania strukturalnego. Nadrzędną zasadą było oryginalnego wyglądu dzieła. Zapisy niekonsekwentne, które mogłyby utrudnić dalsze przetwarzanie tekstu słowników, ujednolicono. Takie postępowanie pozwoliło skrócić czas kolejnych zadań i uniknąć wielu błędów. Wszelkie szczegółowe rozstrzygnięcia dotyczące transliteracji i transkrypcji opisano w dokumentacji projektu.

Jedną z problemowych kwestii stanowiła interpunkcja. Program wykorzystany do OCR nie zachowuje krojów czcionek odczytywanego tekstu, w związku z tym uznano, że jedyną pomocą w anotacji struktury słowników mogłyby być charakterystyczne znaki przestankowe. Z tego względu np. oprócz szczegółowego sprawdzania oryginalnej interpunkcji w trakcie korekty *Forytarza* Ernestiego uzupełniano ukośniki oddzielające część polskojęzyczną od niemieckich ekwiwalentów (por. ryc. 1), a w *Nowym Dykcjonarzu* Troca – znaki paragrafu § sygnalizujące kolokacyjną część hasła.

[6] Wartość MN-LER = 1% oznacza, że średnio tylko 1% znaków w każdej linii jest błędny. Mimo że OCR4all używa miary CER do ewaluacji trenowanych modeli, użytkownik nie ma do jej wartości bezpośredniego dostępu. Twórcy programu skupili się na praktycznym wykorzystaniu każdego modelu, dlatego ewaluacja jest dokonywana osobno dla każdej partii rozpoznawanego materiału wskaźnikiem Mean Normalized Label Error Rate (MN-LER). Jest to miara stosowana w narzędziu Calamari, w której podstawą ewaluacji jest analiza poszczególnych linii tekstu, a nie pojedynczych znaków (jak we wskaźniku CER) czy słów (jak w wypadku WER) (Bernhard Liebl, Michael Burghardt, „On the Accuracy of CRNNs for Line-Based OCR: A Multi-Parameter Evaluation", w *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, red. Folgert Karsdorp, Barbara McGillivray, Adina Nerghes, Melvin Wevers, CEUR Workshop Proceedings 2723 (Amsterdam: CEUR-WS.org, 2020), https://doi.org/10.48550/arXiv.2008.02777).

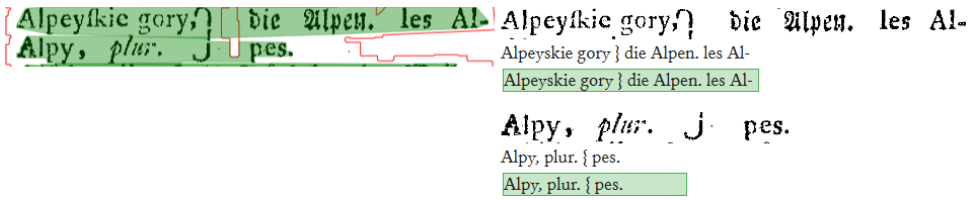
[7] Inne popularne programy do OCR dają taką możliwość: modele strukturalne można tworzyć w programie eScriptorium, natomiast układ kolumn można ustawić w programie Transkribus.



Ryc. 1. Fragment *Forytarza* J. Ernestiego z niewidocznymi ukośnikami, *fol.* Bv v

Innym rozwiązaniem mającym zapewnić późniejsze poprawne oznaczenie mikrostruktury było takie zakodowanie haseł podwójnych i wielokrotnych w słownikach Ernestiego i Troca, które zapewniłoby odpowiednie połączenie ich treści. Trzeba pamiętać, że programy OCR wyznaczają poziome linie odczytu, a następnie łączą tekst linearnie. W wypadku haseł podwójnych doprowadziłoby to do rozbicia treści obcojęzycznej eksplikacji. W tym celu w miejscu oryginalnej dużej klamry zastosowano sekwencję znaków, która nie występuje w żadnym ze słowników: przy pierwszym wyrazie hasłowym zakodowano sekwencję _}_ na oznaczenie otwarcia klamry, a przy ostatnim wyrazie hasłowym _{_ na zamknięcie klamry (por. ryc. 2 i 3).

Ryc. 2. Hasło podwójne w *Forytarzu* J. Ernestiego na różnych etapach opracowania OCR



Ryc. 3. Hasło podwójne w *Nowym Dykcjonarzu* M. A. Troca na różnych etapach opracowania OCR

Największym wyzwaniem w tworzeniu modeli automatycznego odczytu okazały się znaki, abrewiatury i ligatury alfabetu greckiego występujące w słowniku Knapiusza. Po pierwsze, w piśmie greckim oprócz liter występują akcenty i przydechy, które mogą pełnić funkcję dystynktywną, a więc różnicować znaczenie, dlatego tak ważne jest ich prawidłowe odczytanie. Po drugie, w starych drukach powszechnie stosowano w grece ligatury i skrócenia, np.: końcówka -ος (-os) często występuje u Knapiusza w formie znaku , ligatura dwuznaku στ (st) przyjęła formę , dwuznaku ου (ou) – , abrewiatura rodzajnika τῆς (tes) występuje w formie , a przyimka παρά (para) – w formie . Można z pewnością stwierdzić, że w *Thesaurusie* znajduje się przeszło sto ligatur i abrewiatur. Dodatkowym utrudnieniem przy odczycie greki jest równoczesne występowanie liter, ligatur i abrewiatur wraz z zapisem standardowym lub nieskróconym (przy czym na tym etapie trenowania modelu trudno było określić frekwencję tych form w całym słowniku). Można zauważyć, że istnieje np. dwojaki zapis litery *beta* – w postaci β oraz ϐ. Obydwa występują wariantywnie w materiale źródłowym i mogą się pojawiać nie tylko wielokrotnie w obrębie jednego hasła, ale nawet w tym samym słowie (np. ekbibastes: ἐκβιβαστὴς ‘wykonawca wyroku’). Jak dotąd nie zauważono zasady, która miałaby rządzić doborem konkretnej litery dla oznaczenia tej samej głoski. W różnych formach zapisu występują jeszcze m.in. litery τ (tau) – standardowe τ: , wysokie τ: i w ligaturze ττ: czy też litera π (pi): , . Przyjęto zasadę, zgodnie z którą wszelkie postaci zapisu liter będą ujednolicane do współczesnego zapisu alfabetu greckiego. Wariantywnie występują również abrewiatury wraz z formami nieskróconymi. Warto zwrócić uwagę, że w słowniku Knapiusza oprócz standardowego zapisu spójnika και (kai): , słowo to występuje w dwóch różnych skróconych postaciach: oraz . Tak samo końcówka bezokolicznika – σθαι (-sthai) – została zapisana jako jedna abrewiatura: , natomiast w innym miejscu jako dwie ligatury: .

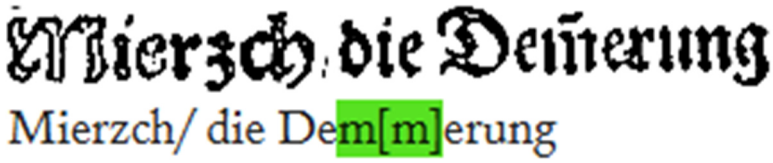
Jak wspomniano, w języku greckim występuje wiele znaków diakrytycznych na oznaczenie przydechów (słabego i mocnego), akcentów (*circumflexus*, *acutus* i *gravis*), dierezy, ι (joty) *subscriptum* czy apostrofu przy elizjach. Stanowiły one spore wyzwanie dla programu OCR4all, głównie ze względu na różnorodność ich zastosowania w obrębie takiej samej litery lub w bliskim sąsiedztwie. Uczący się program w kolejnych próbach długo pomijał diakryty nad literami lub całe litery ze znakami. W ostatecznej wersji również zdarzają się te problemy, jednak w porównaniu z pierwszą są one nieliczne.

Wielość znaków powodowała, że materiału *Ground Truth* (czyli wzorcowych danych) było początkowo zbyt mało, aby model dla *Thesaurusa* Knapiusza generował dobrej jakości wyniki, dlatego trzeba go było trenować znacznie dłużej niż dla pozostałych słowników i na większej liczbie − specjalnie w tym celu przygotowanych − danych wzorcowych.

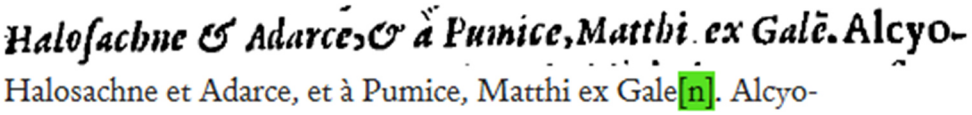
KOREKTA WYNIKÓW OCR

Aby zapewnić możliwie najlepszą jakość materiału mającego stanowić podstawę Bazy Historycznych Leksykonów Polskich, po uzyskaniu automatycznego odczytu tekstu w programie OCR4all przeprowadzono niezbędną korektę wyników, podczas której teksty źródłowe zostały sprawdzone pod kątem zgodności z przyjętymi w projekcie zasadami edycji.

Dla każdego słownika przygotowano instrukcje dotyczące transliteracji i szczegółowych rozstrzygnięć w przypadku miejsc problemowych. Podstawowym założeniem było przeprowadzenie transliteracji. Zachowano pisownię łączną i rozdzielną oryginału oraz wielkie i małe litery, niektóre zapisy wymagały jednak ujednolicenia ze względu na późniejsze etapy automatycznego przetwarzania materiału (z tego względu np. wyrazy hasłowe w słowniku Ernestiego zostały zapisane wielkimi literami). We wszystkich językach zrezygnowano z wariantywnego zapisu niektórych liter (np. alografów czy niemieckich umlautów), gdyż różnica ta miała jedynie charakter graficzny. Rozwinięto wszelkie oryginalne zapisy z tyldami, najczęściej oznaczające opuszczenie liter *m* i *n* w geminatach lub zakończeniach wyrazów (por. ryc. 4 i 5) − rozszerzenia te opatrzono nawiasami kwadratowymi.



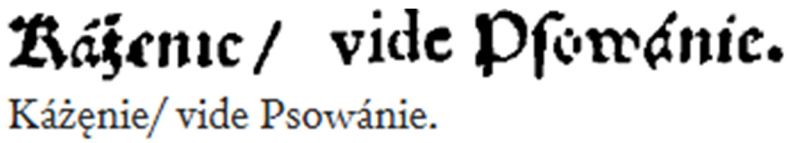
Ryc. 4. Fragment *Forytarza* J. Ernestiego z rekonstrukcją skrótu wskazanego tyldą, k. Av v



Ryc. 5. Fragment *Thesaurusa* G. Knapiusza z korektą zapisu tyldy, s. 681

Istotnym zadaniem korekty było poprawne rozwinięcie wspomnianych wyżej ligatur i abrewiacji łacińskich i greckich w słowniku Knapiusza. Niezwykle pomocne okazało się tu kanoniczne opracowanie Williama Wallace'a, zawierające indeks poszczególnych ligatur i abrewiatur oraz ich różnych wariantów[8].

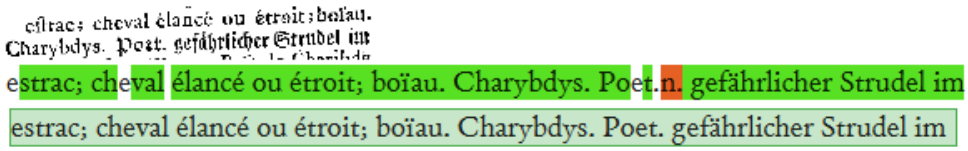
Newralgiczny punkt korekty stanowiło poprawne rozpoznawanie liter *ę* i *ą* w polskich tekstach złożonych czcionkami gotyckimi. W szwabasze polskiej samogłoski nosowe oznaczano w sposób, który tylko nieznacznie różnił się od oznaczania samogłosek ustnych, przez co podczas ich automatycznego odczytu w miejscach ze zniekształconym drukiem niekiedy dochodziło do mylenia tych liter. Zadanie korektorów polegało zwłaszcza na wychwyceniu zapisów z wtórną lub antycypacyjną nosowością (por. ryc. 6).



Ryc. 6. Widok hasła *każenie* w *Thesaurusie* G. Knapiusza, s. 272

[8] William Wallace, „Index of Greek Ligatures and Contractions", *The Journal of Hellenic Studies* 43, nr 2 (1923): 183–93.

Ponadto należało uzupełnić lub skorygować miejsca pominięte albo źle wyodrębnione przez program OCR (np. w sytuacji zaznaczenia dwóch wierszy jako jednego, por. ryc. 7). Dzięki tej weryfikacji materiał poddany optycznemu rozpoznawaniu jest kompletny.



Ryc. 7. Przykład skorygowania błędnego podziału na wiersze w *Nowym Dykcjonarzu* M. A. Troca

Należy skonstatować, że stworzenie licznego zespołu korektorów o odpowiednich kompetencjach nie było zadaniem łatwym. Ostatecznie korekta, oparta na jasno sformułowanych zasadach edycyjnych, zapewniła spójność i integralność materiału, który posłużył jako podstawa do dalszych etapów opracowania cyfrowego.

AUTOMATYCZNA ANOTACJA STRUKTURY SŁOWNIKÓW ZA POMOCĄ ZNACZNIKÓW TEI XML

Materiał tekstowy pozyskany z trzech słowników poddano strukturyzacji i standaryzacji za pomocą znaczników TEI P5, powszechnie stosowanych w naukowych edycjach cyfrowych (*Digital Scholarly Edition* − DSE)[9]. Korpus znaczników TEI stanowi substandard XML – uniwersalnego formatu zapisu danych tekstowych i pozwala w sformalizowany sposób opisać strukturę i treść tekstu. Punktem odniesienia były dla nas wytyczne TEI Lex-0, czyli zestaw dobrych praktyk wypracowany przez zespół leksykografów w ramach europejskiego projektu DARIAH[10]. Zostały one uwzględnione w projekcie na etapie planowania i tworzenia bazy z myślą o ujednoliceniu i właściwej interpretacji zapisów słowników, tak aby zagwarantować spójny model wyszukiwania i wizu-

[9] TEI P5, *TEI: Guidelines for Electronic Text Encoding and Interchange*, 2025, https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html [dostęp: 2.04.2025].

[10] Tasovac i in., *TEI Lex-0*.

alizacji oraz spójne zasady kodowania. Materiał leksykograficzny będzie poddany automatycznej anotacji strukturalnej, podczas której zostaną oznakowane wszystkie części artykułów hasłowych oraz pozostałe elementy struktury dzieł (kustosze, numery składek, numery stron itp.). Etap ten zakończy się korektą.

TRANSKRYPCJA, AUTOMATYCZNA LEMATYZACJA I ANOTACJA

W kolejnych etapach projektu tekst polskojęzyczny zyska warstwę transkrybowaną, a następnie zostanie poddany automatycznej anotacji morfosyntaktycznej (polegającej na wyznaczeniu formy podstawowej każdego wyrazu i szczegółowym oznaczeniu wartości kategorii gramatycznych). Do zadań tych będą wykorzystane narzędzia wytrenowane na tekstach z XVII i XVIII wieku, użyte m.in. do budowy innych zasobów leksykalnych z tego okresu, np. Elektronicznego Korpusu Tekstów XVII i XVIII wieku „KorBa". Narzędzia te stworzyli badacze z Zakładu Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN, którzy od lat zajmują się zagadnieniem przetwarzania języka naturalnego. Do uwspółcześnienia ortografii posłuży transkryber autorstwa Marcina Wolińskiego i Bartłomieja Nitonia, a do znakowania tager Hydra Katarzyny Krasnowskiej-Kieraś i Marcina Wolińskiego. Użytkownicy bazy zyskają możliwość pełnotekstowego przeszukiwania tekstu polskiego zgodnie z uwspółcześnionym zapisem oraz wyszukiwania zaawansowanego według części mowy, kategorii gramatycznych i innych. Dzięki szczegółowej anotacji znacznikami TEI możliwe będzie wyszukiwanie informacji w dowolnej części artykułu oraz wielokierunkowe filtrowanie wyników.

Zgodnie z założeniami grantu w kolejnych latach ma powstać projekt multiwyszukiwarki umożliwiającej przełączanie pomiędzy warstwami transliteracji, transkrypcji oraz widoku oryginału. Baza zostanie umieszczona na platformie TEI Publisher, która daje możliwość zrealizowania wszystkich zaplanowanych funkcjonalności.

*

Słowniki wybrane do opracowania cyfrowego w projekcie to jedne z podstawowych dzieł polskiej leksykografii XVII i XVIII wieku. Po przeprowadzeniu dotychczasowych analiz można po raz pierwszy z dużą dokładnością podać liczbę zawartych w nich znaków, a więc także stron znormalizowanych,

i tym samym poznać realną objętość tych leksykonów, niezależną od liczby stron druku, formatu czy wielkości czcionek. *Thesaurus* G. Knapiusza (wyd. II, 1643, t. I) liczy ponad 5,5 tys. stron znormalizowanych (1645 kart), *Nowy dykcjonarz* Troca (1764, t. III) – około 3,5 tys. stron znormalizowanych (1567 kart), a *Forytarz* – 133 strony znormalizowane (228 kart). Materiał będący podstawą dalszych działań w projekcie to zatem ponad 9 tys. stron znormalizowanych.

Prace służące stworzeniu Bazy Historycznych Leksykonów Polskich mogą się przyczynić do weryfikacji dotychczasowych sądów opartych na cząstkowym materiale, dotyczących m.in. zależności między słownikami. Pozwolą także na poznanie kwestii podstawowych, których dotychczas nie zbadano, takich jak dokładna liczba artykułów hasłowych czy znajdujących się wewnątrz nich wyrazów polskich. Zastosowanie wyszukiwarki umożliwi realizację przewidzianych na kolejne lata trwania projektu analiz porównawczych. Będą one prowadzone w zakresie metaleksykografii (pogłębione badania makro- i mikrostruktury słowników), semantyki historycznej (analiza znaczeń ekwiwalentów w wybranych hasłach rzeczownikowych), pragmatyki (zakres użycia kwalifikatorów). Dzięki głębokiej digitalizacji badacze zyskają możliwość pełniejszego dostępu do tych niezwykle cennych zasobów leksykalnych i leksykograficznych doby średniopolskiej. Mamy nadzieję, że przygotowywane przez nas narzędzie spełni oczekiwania językoznawców i wszystkich zainteresowanych historią i kulturą XVII i XVIII wieku.

BIBLIOGRAFIA

Bibliografia przedmiotowa

Estreicher, Karol. *Bibliografia polska*. T. XVI. Kraków: Akademia Umiejętności, 1898.

Frączek, Agnieszka. *Słowniki polsko-niemieckie i niemiecko-polskie z przełomu XVII i XVIII wieku. Analiza leksykograficzna*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego, 2010.

Iwanowska, Aleksandra. „O metodzie leksykograficznej «Nowego dykcyjonarza to jest mownika polsko-niemiecko-francuskiego» (1764) Michała Abrahama Troca". *Prace Filologiczne* 39 (1994): 291–325.

Liebl, Bernhard, Michael Burghardt. „On the Accuracy of CRNNs for Line-Based OCR: A Multi-Parameter Evaluation". W *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*, red. Folgert Karsdorp, Barbara McGillivray, Adina Nerghes, Melvin Wevers. CEUR Workshop Proceedings 2723. Amsterdam: CEUR-WS.org, 2020. https://doi.org/10.48550/arXiv.2008.02777.

Puzynina, Jadwiga. „O metodzie leksykograficznej w «Thesaurusie» Knapskiego". *Poradnik Językowy* 4 (1956): 121–9; 5 (1956): 168–75, 6 (1956): 216–21.

Puzynina, Jadwiga. *„Thesaurus" Grzegorza Knapiusza. Siedemnastowieczny warsztat pracy nad językiem polskim*. Wrocław–Warszawa–Kraków: Zakład Narodowy im. Ossolińskich, 1961.

Tasovac, Toma, Laurent Romary, Piotr Banski, Jack Bowers, Jesse de Does, Katrien Depuydt, Tomaž Erjavec, Alexander Geyken, Axel Herold, Vera Hildenbrandt, Mohamed Khemakhem, Boris Lehečka, Snežana Petrović, Ana Salgado, Andreas Witt. *TEI Lex-0: A Baseline Encoding for Lexicographic Data. Version 0.9.4. DARIAH Working Group on Lexical Resources*. 2018. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html [dostęp: 2.04.2025].

Wallace, William. „Index of Greek Ligatures and Contractions". *The Journal of Hellenic Studies* 43, nr 2 (1923): 183–93. The Society for the Promotion of Hellenic Studies.

Żmigrodzki, Piotr. *Słowo – słownik – rzeczywistość. Z problemów leksykografii i metaleksykografii*. Kraków: Universitas, 2008.

Elektroniczny słownik języka polskiego XVII i XVIII wieku. Red. Włodzimierz Gruszczyński (2004–2023), Magdalena Majdak (2024–). https://sxvii.pl [dostęp: 2.04.2025].

Elektroniczny Korpus Tekstów XVII i XVIII wieku. „KorBa". https://korba.edu.pl [dostęp: 2.04.2025].

TEI P5. *TEI: Guidelines for Electronic Text Encoding and Interchange*. 2025. https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html [dostęp: 2.04.2025].

BIBLIOGRAFIA PODMIOTOWA

Ernesti, Jan. *Forytarz języka polskiego, osobne, rzeczy niemal wszystkich, słowa, mowy różne i rozmowy dwie w sobie zawierający*. Wrocław: Drukarnia Mikołaja Kornachera, 1674.

Knapiusz, Grzegorz. *Thesaurus Polonolatinograecus seu promptuarium linguae Latinae et Graecae […]*, t. 1. Kraków: Drukarnia Łazarzowa, 1643. Wydanie II.

Troc, Michał Abraham. *Nowy dykcyonarz to jest mownik polsko-niemiecko-francuski z przydatkiem przysłow potocznych, przestrog gramatycznych, lekarskich, matematycznych, fortyfikacyjnych,* żegla[r]skich*, łowczych i inszym naukom przyzwoitych wyrazow*, t. 3. Lipsk: Officina Gleditschiana, 1764.