ROCZNIKI HUMANISTYCZNE
Tom LXXIII, zeszyt 2 – 2025
ZESZYT SPECJALNY
DOI: https://doi.org/10.18290/rh25732.6s

ALBERT BYRSKI

WIELOJĘZYCZNE EDYCJE ŹRÓDEŁ W PRZETWARZANIU JĘZYKA NATURALNEGO – STUDIUM PRZYPADKU WYBRANYCH AKT SEJMIKOWYCH Z WIELKIEGO KSIĘSTWA LITEWSKIEGO

Abstrakt. Artykuł omawia wyzwania związane z komputerową analizą wielojęzycznych tekstów historycznych na przykładzie akt sejmikowych z terenu Wielkiego Księstwa Litewskiego. Przedstawiono proces przygotowania danych: ekstrakcję tekstu z plików PDF, czyszczenie oraz anotację językową. Szczególną uwagę poświęcono problemom wynikającym z braku ujednoliconych edycji cyfrowych, współczesnych modyfikacji ortograficznych oraz wielojęzyczności tekstów (polski, ruski, łacina). Wykorzystano narzędzia NLP, takie jak Morfeusz (Korbeusz), Concraft oraz Stanza. Podkreślono znaczenie dostosowania narzędzi do specyfiki materiału historycznego i konieczność dalszej standaryzacji anotacji w ramach Universal Dependencies.

Słowa kluczowe: przetwarzanie języka naturalnego; akta sejmikowe; edycje źródeł; wielojęzyczność; Wielkie Księstwo Litewskie

MULTILINGUAL SOURCE EDITIONS IN NATURAL LANGUAGE PROCESSING: A CASE STUDY OF A SELECTED DIETINES ACTS FROM THE GRAND DUCHY OF LITHUANIA

Abstract. The article discusses the challenges of computational analysis of multilingual historical texts, using the example of sejmik records from the Grand Duchy of Lithuania. It presents the data preparation process: text extraction from PDFs, cleaning, and language annotation. Particular attention is given to problems stemming from the lack of standardized digital editions, modernized orthography, and the multilingual character of the texts (Polish, Ruthenian, Latin). NLP tools such as Morfeusz (Korbeusz), Concraft, and Stanza were used. The importance of adapting tools to the specifics of historical material and the need for further standardization of annotations within the Universal Dependencies framework are emphasized.

Keywords: natural language processing; dietines acts; source editions; multilingualism; Grand Duchy of Lithuania

Mgr ALBERT BYRSKI – Uniwersytet Warszawski, Wydział Historii; adres do korespondencji: ul. Krakowskie Przedmieście 26/28, 00-927 Warszawa; e-mail: a.byrski@uw.edu.pl; ORCID: https://orcid.org/0000-0003-0595-2839.

Komputerowa analiza tekstów o charakterze historyczno-prawnym jest obecna w nauce od dawna. Wcześniejsze badania[1], ograniczone możliwościami technicznymi, koncentrowały się zwykle na wybranych słowach i pojęciach analizowanych przez badaczy w sposób ręczny lub półautomatyczny. Dopiero rozwój metod przetwarzania języka naturalnego (NLP – Natural Language Processing), optycznego rozpoznawania znaków (OCR – Optical Character Recognition) oraz wzrost dostępnej mocy obliczeniowej umożliwiły zastosowanie bardziej zaawansowanych metod analizy historycznych tekstów źródłowych. W kontekście badań nad nowożytnymi dziejami dawnej Rzeczypospolitej szczególne nadzieje na rozwój metod komputerowej analizy języka historycznego wiążą się z projektem Elektroniczny korpus tekstów polskich z XVII i XVIII wieku (do 1772 r.)[2] (dalej cyt. Korpus barokowy). Ciekawym i cennym uzupełnieniem tego nurtu badań cyfrowych jest Korpus dokumentów sejmikowych Rzeczypospolitej przedrozbiorowej (dalej cyt. Korpus sejmikowy) autorstwa Szymona Rutkowskiego[3], zawierający edytowane akta sejmikowe z ziem koronnych. W obu przypadkach celem jest nie tylko gromadzenie tekstów źródłowych, lecz także zapewnienie badaczom języka historycznego dostępu do narzędzi umożliwiających ich przeszukiwanie oraz prowadzenie analiz statystycznych i lingwistycznych.

Ze względu na wciąż niezadowalającą jakość wyników automatycznej transkrypcji rękopisów z epoki nowożytnej, w pracach nad korpusami tekstów historycznych często korzysta się z edycji źródłowych, które dostępne są zazwyczaj wyłącznie w formie plików PDF, pozbawionych strukturalnej warstwy tekstowej. Oznacza to konieczność ich wcześniejszego przygotowania do przetwarzania – od konwersji do formatu możliwego do automatycznej analizy. Proces ten obejmuje

[1] Warto tutaj wspomnieć badania Zespołu Kultury Staropolskiej IH UW, które były prowadzone w sposób ręczny; zob. Urszula Augustyniak, „Polska i łacińska terminologia ustrojowa w publicystyce politycznej epoki Wazów", w *Łacina jako język elit*, red. Jerzy Axer (Warszawa: Wydawnictwo DiG, 2004); Urszula Augustyniak, *Koncepcje narodu i społeczeństwa w literaturze plebejskiej od końca XVI do końca XVII w.* (Warszawa: Państwowe Wydawnictwo Naukowe, 1989). Warte uwagi są również badania Karola Mazura, który w jednym z rozdziałów dokonał analizy słownikowej terminologii akt sejmikowych; zob. Karol Mazur, *W stronę integracji z Koroną. Sejmiki Wołynia i Ukrainy w latach 1569–1648* (Warszawa: Wydawnictwo Neriton, 2006), 226–55. Dla kontrastu, badanie przy wykorzystaniu dostępnych zasobów cyfrowych do analizy pojęcia „wolność" zob. Urszula Augustyniak, „Wolność szlachcica w Rzeczypospolitej XVII w. Propozycje badawcze", *Przegląd Historyczny* 114 (2023): 23–49.

[2] Włodzimierz Gruszczyński, Dorota Adamiec, Maciej Ogrodniczuk, „Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.)", *Polonica* 33 (2013): 311–8. Korpus dostępny pod linkiem: https://korba.edu.pl/query_corpus/ [dostęp: 22.05.2025].

[3] Szymon Rutkowski, *Korpus dokumentów sejmikowych Rzeczypospolitej przedrozbiorowej*, https://sejmiki.nlp.ipipan.waw.pl/overview [dostęp: 22.05.2025].

nie tylko optyczne rozpoznawanie znaków, lecz także dalsze etapy: czyszczenie[4] oraz znakowanie tekstu zgodnie z przyjętymi standardami anotacji. Przedmiotem niniejszego artykułu jest projekt stworzenia cyfrowego korpusu źródłowych akt sejmikowych z terenu Wielkiego Księstwa Litewskiego (dalej cyt. WKL)[5] przeznaczonego do analiz kwantytatywnych i badań nad dawnym językiem. Charakterystyczne dla tego materiału jest występowanie tekstów w różnych językach – przede wszystkim polskim, ruskim[6] i łacińskim – często współobecnych w ramach jednego dokumentu. Taka wielojęzyczność wymusza zastosowanie różnych podejść przetwarzania dla poszczególnych segmentów tekstu.

Niniejszy artykuł stanowi jedną z możliwych propozycji opracowania i analizy tego typu źródeł. Przedmiotem refleksji są zarówno wybrane praktyczne aspekty przygotowania danych historycznych, jak i wyzwania oraz możliwości związane z przetwarzaniem wielojęzycznego języka naturalnego w kontekście tekstów dawnych. Podstawowym celem był pragmatyzm rozwiązań, a w miarę możliwości starano się stosować rozwiązania automatyczne lub półautomatyczne.

EDYCJE AKT SEJMIKOWYCH Z TERENU WIELKIEGO KSIĘSTWA LITEWSKIEGO

Edycje akt sejmikowych wprawdzie mają w polskiej historiografii długą tradycję, niemniej dopiero dwie ostatnie dekady przyniosły wyraźne ożywienie: opublikowano nowe tomy akt wielkopolskich oraz materiały z sejmików podolskich, chełmskich, lubelskich, bełskich i rawskich[7], a kolejne edycje są w przygotowaniu. Podkreślano ich znaczenie nie tylko dla badania historii politycznej,

[4] Pod określeniem „czyszczenie" rozumiem usunięcie wszystkich zbędnych elementów, które mogą zaburzyć analizę tekstu właściwego.

[5] Analizie poddano korpus obejmujący instrukcje i uchwały sejmikowe z terenu Wielkiego Księstwa Litewskiego. Kryterium doboru tekstów stanowiła definicja masowości tego typu źródeł zaproponowana przez Jaremę Maciszewskiego (por. Jarema Maciszewski, „Kultura polityczna Polski «złotego wieku»", w *Dzieje kultury politycznej w Polsce*, red. Jerzy A. Gierowski, 11–5 (Warszawa: Państwowe Wydawnictwo Naukowe, 1977). Na tej podstawie wybrano 233 dokumenty z lat 1570–1655, obejmujące łącznie około 328 000 słów. Ponadto do korpusu włączono przygotowywane w IH PAN do wydania akta sejmikowe z województwa nowogródzkiego (50 tekstów). Składam serdeczne podziękowania Profesorowi Andrzejowi Rachubie za udostępnienie tych materiałów.

[6] Na potrzeby artykułu używam określenia „język ruski" jako terminu najbardziej rozpowszechnionego w polskiej historiografii. Inne określenia są często uzależnione od przynależności narodowej historiografii.

[7] Z najnowszych edycji m.in. zob. Jarosław Stolicki, oprac., *Akta sejmiku podolskiego in hostico 1672–1698* (Kraków: Towarzystwo Naukowe „Societas Vistulana", 2002); Wiesław Bondyra, Henryk Gmiterek, Jerzy Ternes, oprac., *Akta sejmikowe ziemi chełmskiej 1572–1668* (Lublin: Archiwum

lecz także dla rekonstrukcji procesów społecznych, gospodarczych oraz wojskowych w epoce nowożytnej. Na tym tle o wiele gorzej prezentuje się stan prac edytorskich nad aktami sejmikowymi z terenów Wielkiego Księstwa Litewskiego[8]. Wprawdzie już w XIX wieku badacze rosyjscy rozpoczęli wybiórcze wydawanie akt sejmikowych w ramach prac Wileńskiej Archeograficznej Komisji, jednakże dobór materiału był podyktowany wymaganiami wyznaniowymi i politycznymi, tzn. kształtowaniem takiego, a nie innego obrazu dawnej Rzeczypospolitej (a zwłaszcza Wielkiego Księstwa Litewskiego). Znaleźć tam możemy przede wszystkim instrukcje i uchwały powiatu brzeskiego, grodzieńskiego, wileńskiego oraz sporadycznie z pozostałych powiatów. Niestety ich wartość obniża niedokładność odczytu oraz liczne błędy[9]. W XXI wieku sytuacja zaczęła się diametralnie zmieniać. Początkowo zaczęto wydawać pojedyncze instrukcje i uchwały, natomiast nowy etap w badaniach zapoczątkowany został przez Henryka Lulewicza, który skupił się na edycji akt zjazdów stanów litewskich[10]. W kolejnych latach rozpoczęto projekty mające na celu wydanie akt sejmikowych poszczególnych sejmików: nowogródzkiego, trockiego, kowieńskiego[11], wołkowyskiego, połockiego oraz województwa wileńskiego[12]. W 2024 roku

Państwowe w Lublinie–Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, 2013); Henryk Gmiterek, oprac., *Akta sejmikowe województwa lubelskiego 1572–1632* (Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, 2016); Mirosław Nagielski, red. nauk., oprac. Michał Bąk i in., *Lauda ziemi rawskiej i województwa rawskiego 1583–1793* (Warszawa: Neriton, 2017); Michał Zwierzykowski, Robert Kołodziej, oprac., *Akta sejmikowe województwa bełskiego. Lata 1572–1655* (Kraków: Księgarnia Akademicka, 2020); Henryk Gmiterek, oprac., *Akta sejmikowe województwa lubelskiego 1633–1668* (Lublin–Radom: Uniwersytet Marii Curie-Skłodowskiej, 2021); Michał Zwierzykowski, Robert Kołodziej, Andrzej Kamieński, oprac., *Akta sejmikowe województwa bełskiego: lata 1656–1695* (Kraków: Księgarnia Akademicka, 2023); Igor Kraszewski, Michał Zwierzykowski, Robert Kołodziej, Andrzej Kamieński, oprac., *Akta sejmikowe województw poznańskiego i kaliskiego. Lata 1656–1668* (Poznań: Grupa Wydawnicza FNCE–Wydawnictwo Naukowe FNCE, 2024); Wiesław Bondyra, Jerzy Ternes, oprac., *Akta sejmikowe województwa lubelskiego: 1697–1763* (Lublin–Radom: Uniwersytet Marii Curie-Skłodowskiej, 2024).

[8] Omówienie aktualnego stanu edycji akt sejmikowych z terenu Wielkiego Księstwa Litewskiego zob. Andrzej Rachuba, „Edycje akt sejmikowych z terenu Wielkiego Księstwa Litewskiego". *Miscellanea Historico-Iuridica*, nr 1 (2022): 347–63.

[9] Rachuba, „Edycje akt sejmikowych", 349. *Акты, издаваемые Археографическою Коммиссиею* (dalej cyt. AVAK), t. 1–4, 7–8, 13 (Вильно: Виленская Археографическая Комиссия, 1865–1886).

[10] Henryk Lulewicz, oprac., *Akta zjazdów stanów Wielkiego Księstwa Litewskiego*, t. 1: *Okresy bezkrólewi (1572–1576, 1586–1587, 1632, 1648, 1696–1697, 1706–1709, 1733–1735, 1763–1764)*, t. 2: *Okresy panowań królów elekcyjnych XVI–XVII wiek* (Warszawa: Neriton, 2006–2009).

[11] Na uwagę zasługuje bardzo dobra edycja akt kowieńskich, zob. Monika Jusupović, wyd., *Akta sejmiku kowieńskiego z lat 1733–1795* (Warszawa: Neriton, 2019).

[12] Tomasz Kempa, Tomasz Ambroziak, oprac., *Akta sejmikowe województwa wileńskiego (1566–1655)* (Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, 2024) (dalej cyt. Akta wileńskie).

wydano także *Źródła do dziejów Żmudzi (1522–1648)*[13], w których znajdują się instrukcje i uchwały sejmiku żmudzkiego z tego okresu.

Dostępne obecnie edycje mają charakter papierowy lub zostały udostępnione w formie prostych plików PDF. Nie spełniają zatem wymogów nowoczesnych edycji cyfrowych, a ich użycie w badaniach ilościowych wymaga wcześniejszej digitalizacji oraz uzupełnienia o strukturę i wielopoziomową anotację lingwistyczną. O ile ich układ typograficzny jest przejrzysty i przyjazny czytelnikowi, to z punktu widzenia dalszego przetwarzania maszynowego stanowi ograniczenie: brak jest zakodowanych nagłówków, segmentacji czy wyróżnionych nazw własnych. Choć każdy dokument ma nagłówek (z datą i miejscem powstania) oraz część metadanych edytorskich (informacje o podstawie źródłowej, zachowanych rękopisach, wcześniejszych edycjach i zasadach redakcji), elementy te nie zawsze są rozpoznawalne w sposób automatyczny przez komputer[14]. Dlatego każda próba wykorzystania ich w systemach przetwarzania języka naturalnego wymaga wcześniejszej konwersji i anotacji, np. w formacie TEI XML (dla struktury dokumentu) lub CoNLL-U (dla anotacji morfosyntaktycznej).

Na potrzeby niniejszego artykułu analizie poddano edycje obejmujące akta sejmikowe do połowy XVII wieku. Wybór ten wynika stąd, że właśnie dla tego okresu mamy obecnie najwięcej wydanych współczesnych XXI-wiecznych edycji, a ich struktura wygląda podobnie. Każdy dokument opatrzony jest wyraźnym nagłówkiem, z datą i miejscem sporządzenia, po którym następuje uporządkowana część metadanych edytorskich. Zawiera ona informacje o zachowanych oryginałach i kopiach, ewentualnych wcześniejszych edycjach oraz uwagi dotyczące redakcji tekstu. Część właściwa dokumentu prezentowana jest w języku oryginału i uzupełniona o przypisy dolne – zarówno tekstowe, jak i rzeczowe – objaśniające kontekst, słownictwo lub niejasności. Obecne są zarówno przypisy cyfrowe, jak i literowe.

Podstawą wydania dla tekstów w języku polskim była we wszystkich przypadkach instrukcja K. Lepszego[15]. Na jej podstawie dokonywano selektywnej modernizacji językowej – rozwijano skróty (w tym formuły urzędowe i grzecznościowe), upraszczano archaiczne formy graficzne (np. „ph" zastępowano

[13] Eugenijus Saviščevas, Jonas Drungilas, oprac., *Źródła do dziejów Żmudzi (1522–1648)*, red. Tomasz Kempa (Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, 2023) (dalej cyt. Akta żmudzkie).

[14] Część tych informacji można wydobyć dzięki wyrażeniom regularnym. Takie podejście wykorzystał Szymon Rutkowski w projekcie *Tribunica*, na potrzeby tworzenia metadanych dla dokumentów; link: https://gitlab.com/szmer/tribunicia/-/tree/master?ref_type=heads [dostęp: 22.05.2025].

[15] Kazimierz Lepszy, oprac., *Instrukcja wydawnicza dla źródeł historycznych od XVI do połowy XIX wieku* (Wrocław: Ossolineum, 1953).

przez „f”[16], „oe” przez „e”[17]), a także ujednolicano pisownię słów pochodzenia łacińskiego występujących w polskiej składni[18]. Modernizacji podlegała także interpunkcja oraz układ akapitów, dostosowywany do logicznego przebiegu treści dokumentów. Jednocześnie pozostawiano wiele cech średniopolskiej fleksji i fonetyki, takich jak formanty *-tski*, końcówki *-szy*, archaiczne warianty (*bydź*, *braciej*) czy fonetyczne oboczności typowe dla gwar WKL. Błędy pisarskie poprawiano najczęściej bez oznaczenia, natomiast istotniejsze zmiany sygnalizowano poprzez konwencjonalne znaki redakcyjne, jak *[s]*, *[ss]*, *[?]* lub *[...]*[19]. Tego rodzaju edytorski kompromis – między transkrypcją a modernizacją – ma istotne konsekwencje dla przetwarzania tekstów w ramach NLP: teksty te nie oddają w pełni oryginalnej warstwy językowej dokumentu, dlatego stanowi to ograniczenie w dalszych badaniach, zawężając możliwości analizy języka[20].

Kolejną kwestią są zasady wydawania tekstów w cyrylicy. Również dla wszystkich edycji zastosowano te same zasady. Dokumenty ruskie publikowane zgodnie z tradycją edytorską Metryki Litewskiej[21] cechują się specyficzną konwencją graficzną, która uwzględnia zarówno elementy transliteracyjne, jak i modernizacyjne[22]. Wyniesienia górne liter zostały przesunięte do linii zasadniczej i oznaczone kursywą, a skróty rozwinięto, uzupełniając brakujące litery w nawiasach okrągłych. Litery opuszczone przez skrybów oznaczano w nawiasach kwadratowych. Unowocześniono także zapis fonetyczny, rezygnując z niektórych liter o jednakowym brzmieniu, które przeniesiono ze wcześniejszej tradycji piśmienniczej. Litery takie jak *ŭ* oddano przez wyniesione *u*, a całe słowa wpisane nad linijką zaznaczano kursywą. Stosowane w tekście źródła litery, które we współczesnych językach wschodniosłowiańskich wyszły z użytku, zostały oddane za pomocą liter używanych współcześnie, tj. litera *є* oraz ѣ jako *e*; *ѫ* i *ѧ* jako *я*; *ѡ* jako *o*; *ѳ* jako *a*; *ѹ* jako *y*. Zrezygnowano z kustod oraz

[16] Zmiana np.: „phara” na „fara”.

[17] Zmiana np.: „economia” (czy „oekonomia”) albo „Oestonia” zamieniano na „ekonomia” i „Estonia”.

[18] Na podstawie Instrukcji K. Lepszego edytorzy ujednolicali zapis wyrazów łacińskiego pochodzenia używanych w polskiej składni. Ponadto, zgodnie z Instrukcją, wyrazy z końcówką *-tia* oddawano w polszczyźnie przez *-yja*, np. „konstytucyja” („constitutia”), „egzorbitancyja” („exorbitantia”).

[19] Akta żmudzkie: XIV–XVI; Akta wileńskie: 46–8, Akta zjazdów, t. 1: 13.

[20] Mowa o ograniczeniach m.in.: ingerencje modernizacyjne w zakresie grafii i interpunkcji; brak pełnego odzwierciedlenia oryginalnej ortografii i niuansów fleksyjnych; brak oznaczenia znaków diakrytycznych i innych cech charakterystycznych rękopisów.

[21] Анна Леонидовна Хорошкевич, Сергей Михайлович Каштанов, ред., *Методические рекомендации по изданию и описанию материалов Литовской метрики* (Москва–Вильнюс, 1985).

[22] Transliteracja – zachowanie cech zapisu rękopiśmiennego w wersji drukowanej. Modernizacja – ujednolicenia dla ułatwienia czytania.

niekonsekwentnie używanych wielkich liter i znaków interpunkcyjnych[23]. Taka konwencja, choć ułatwia czytanie i automatyzację segmentacji, wymaga świadomego traktowania tekstu jako materiału pośredniego.

PRZYGOTOWANIE TEKSTÓW

Podstawową kwestią w przygotowaniu tekstów do skutecznego przetwarzania języka naturalnego jest sprowadzenie ich do postaci tekstowej – ciągu znaków i słów – a nie graficznej, czyli obrazów stron, które same w sobie nie są możliwe do analizy lingwistycznej. Gdy plik PDF zawierał już warstwę tekstową (tzw. OCR-owany PDF), spośród testowanych narzędzi najlepsze wyniki dała biblioteka pypdfium2[24], stabilnie obsługująca dokumenty mieszane (alfabet łaciński + cyrylica) i poprawnie eksportująca do formatu .txt. W sytuacji, gdy źródłem był obraz lub zeskanowana strona bez warstwy tekstowej, warto skorzystać z narzędzia, takiego jak PyTesseract[25] – opartego na silniku Tesseract OCR, rozwijanego przez Google – które umożliwia konwersję grafiki na tekst poprzez rozpoznawanie znaków. Testowano w tym przypadku modele rus i ukr, jednak typowe błędy obejmowały zamianę znaków o podobnym kształcie, brak rozpoznania wszystkich diakrytyków oraz sklejanie liter[26]. Jednakże takie automatyczne rozwiązanie niesie ze sobą poważny minus. Wraz z warstwą tekstową wydobywane są również wszelkie komentarze edytorskie[27], przypisy dolne, paginacja, nagłówki stron oraz inne elementy typograficzne, które nie należą do właściwej treści dokumentu historycznego[28]. Ostatecznie zdecydowano się na podejście półautomatyczne – wykorzystano program ABBYY FineReader,

[23] Tomasz Ambroziak, „Jak wydawać cyrylickie akta sejmikowe? Analiza rosyjskich, ukraińskich i białoruskich współczesnych zasad wydawniczych oraz wybranej praktyki edytorskiej", cz. 1, *Miscellanea Historico-Iuridica* 21, nr 1 (2022): 321–45.

[24] https://pypi.org/project/pypdfium2/ [dostęp: 22.05.2025].

[25] https://pypi.org/project/pytesseract/ [dostęp: 22.05.2025].

[26] Nie zdecydowano się na dotrenowanie modelu z powodu zbyt małej skali projektu. Ponadto liczba tekstów nie była na tyle duża, aby uzasadniała tworzenie dodatkowego, ręcznie anotowanego korpusu treningowego, co wymagałoby znacznych nakładów pracy i czasu, przekraczających możliwości jednej osoby.

[27] W tym: informacje na temat przetrzymywania rękopisów, literatury przedmiotu oraz miejsca powstania dokumentu.

[28] W tym przypadku jest mowa o problemach towarzyszących podstawowemu podejściu opartemu na bezpośrednim rozpoznaniu obrazu i prostym czyszczeniu warstwy tekstowej. W praktyce wiele nowoczesnych narzędzi umożliwia wcześniejsze rozpoznanie regionów, a we współczesnych rozwiązaniach proces ten i rozpoznania traktuje się jako odrębne zadania.

komercyjny silnik OCR, który okazał się bardziej niezawodny w przypadku cyrylicy i niskiej jakości skanów (np. tomów AVAK), który pozwala na ręczne zaznaczanie interesujących nas stref tekstu jeszcze na poziomie skanu. Dzięki funkcji definiowania obszarów odczytu możliwe jest wyodrębnienie jedynie zasadniczej treści dokumentu, z pominięciem przypisów dolnych, paginacji, nagłówków oraz komentarzy edytorskich. Takie podejście znacząco redukuje liczbę artefaktów w danych wyjściowych, usprawnia także dalsze etapy czyszczenia tekstu. Ponadto ten program także poradził sobie bez problemu ze słabej jakości skanami AVAK[29].

Jednym z problemów napotkanych podczas ekstrakcji tekstów w cyrylicy była kwestia zachowania kursywy, która w edycjach źródłowych często służy do oznaczania wyniesień, dopisków lub wyróżnień redakcyjnych[30]. ABBYY pozwala wprawdzie na eksport do formatów, które tę informację zachowują, jednak w projekcie zdecydowano się na najprostszy eksport do .txt – w tym trybie formatowanie zostaje utracone. Ponieważ kursywa należy do warstwy prezentacyjnej, a nie treści właściwej, podczas kopiowania do edytorów nieobsługujących formatowania (takich jak Visual Studio Code, terminale, surowe edytory tekstu lub narzędzia do NLP) zostaje bezpowrotnie utracona. W efekcie dane wejściowe do dalszego przetwarzania tracą tę dodatkową informację strukturalną. Choć z punktu widzenia semantycznego nie wpływa to na analizę treści, ogranicza możliwość wiernej rekonstrukcji układu edytorskiego[31].

Tab. 1. Przykłady zmian w tekstach cyrylickich[32]

Edycja	Ekstraktowany tekst
све*т*ских	светских
вря*д*никовъ	врядниковъ
зе*м*ских	земских
Люде*и*	людеи
Упе*в*ненье	Упевненье
тепере*ш*ны	теперешны

[29] Mam tu na myśli poprawne rozpoznanie tekstu zapisanego cyrylicą oraz jego ekstrakcję graficzną.

[30] Kursywa jest też używana do wyróżnienia łaciny, ale nie zauważono takiego problemu, jak w przypadku ruszczyzny.

[31] Podstawowym celem projektu obejmującego korpus instrukcji uchwał z terenu WKL jest analiza terminologii politycznej; dlatego utrata kursywy – pełniącej głównie funkcję edytorską i typograficzną – nie miała wpływu na rezultaty moich badań.

[32] Przykłady pochodzą z: Akta żmudzkie: 149–57.

Wszystkie analizowane teksty zostały zapisane w kodowaniu UTF-8, które stanowi obecnie standardowy sposób reprezentacji znaków w systemach informatycznych. Jego główną zaletą jest pełna zgodność z Unicode oraz możliwość zapisu znaków z różnych systemów pisma, co miało istotne znaczenie w kontekście wielojęzycznego materiału, obejmującego m.in. alfabet łaciński i cyrylicę. Utrata informacji o kursywie i pogrubieniu wynikała wyłącznie z przyjętego formatu eksportu (.txt), a nie z ograniczeń samego kodowania. Należy również pamiętać, że wszystkie wykorzystane teksty zostały wcześniej zredagowane zgodnie ze współczesną ortografią, co ogranicza obecność historycznych cech pisma, takich jak dawne ligatury czy warianty literowe[33].

Po zakodowaniu i wstępnym zapisaniu plików konieczne było przeprowadzenie etapu czyszczenia tekstu. W celu przygotowania tekstu do dalszej analizy zastosowano zestaw operacji czyszczących, opartych na wyrażeniach regularnych[34]. Usunięto oznaczenia redakcyjne typowe dla edycji źródłowych, takie jak samodzielne [s] i [ss][35], nawiasy z rozwinięciami skrótów, pozostawiając samą treść, a także fragmenty formatowania graficznego, takie jak znaki cudzysłowu typograficznego czy symbole redakcyjne. Zredukowano artefakty wynikające z OCR, takie jak nieprawidłowe przeniesienia wyrazów na koniec wiersza, pojedyncze znaki nowej linii w środku zdań, wielokrotne spacje czy cyfry doklejone do słów. Jednocześnie zadbano o zachowanie struktury logicznej tekstu, ograniczając nadmiarowe podziały linii i oczyszczając fragmenty z niewidocznych znaków specjalnych. W efekcie uzyskano tekst jednorodny, pozbawiony większości zanieczyszczeń, gotowy do dalszej pracy i anotacji językowej. Niestety nie udało się usunąć wszystkich elementów, ponieważ bardziej inwazyjne wyrażenia mogły skutkować usunięciem fragmentów tekstu właściwego. Jednym z takich przykładów są pozostałości po przypisach literowych.

[33] Więcej o Unicode zob.: Michael Piotrowski, *Natural Language Processing for Historical Texts* (San Rafael: Morgan & Claypool Publishers, 2012), 53–7.

[34] M.in. zob.: https://swtch.com/~rsc/regexp/regexp1.html [dostęp: 22.05.2025].

[35] Mogą one pełnić również funkcję korekty tekstu, dlatego usunięciu podlegają tylko samodzielne. Natomiast uzupełnienia tekstu zostały pozostawione dzięki zastosowaniu dodatkowych wyrażeń regularnych.

PRZETWARZANIE WIELOJĘZYCZNEGO HISTORYCZNEGO TEKSTU – WYBRANE ASPEKTY

Pierwszym krokiem w pracy z wielojęzycznym tekstem, takim jak instrukcje czy uchwały sejmikowe z terenu Wielkiego Księstwa Litewskiego, było przeprowadzenie anotacji językowej z uwzględnieniem jego heterogenicznego charakteru. Ze względu na liczne łacińskie wtrącenia w tekstach polskich oraz potrzebę wypracowania ujednoliconego schematu anotacji, zdecydowano się na rozpoznawanie języka na poziomie każdego tokenu[36].

Dla tekstów w języku ruskim, zapisanych cyrylicą, zastosowano prosty mechanizm identyfikacji oparty na wyrażeniu regularnym rozpoznającym alfabet cyrylicki. W przypadku języka polskiego i łacińskiego, których teksty wykorzystują ten sam alfabet, konieczne było zastosowanie dokładniejszej metody. W tym celu wykorzystano bibliotekę Lingua (Python), która okazała się najskuteczniejszym narzędziem do detekcji języka w kontekście krótkich jednostek leksykalnych[37]. Po ręcznym sprawdzeniu próbek rozpoznanego tekstu zrezygnowano z wykorzystania podwójnej walidacji[38], ponieważ skuteczność rozpoznania języka była na bardzo wysokim poziomie[39]. Ponadto podobieństwa morfologiczne wynikające z łacińskich końcówek fleksyjnych w wyrazach polskich nie spowodowały zaburzenia w rozpoznaniu[40].

Na etapie przygotowania tekstu dokonano tokenizacji przy użyciu wyrażeń regularnych, unikając wcześniejszego przypisywania informacji językowych przez narzędzia NLP[41]. Wszystkie znaki interpunkcyjne oraz cyfry potraktowano jako neutralne i nieuwzględnione w detekcji języka. Tak podzielony tekst został poddany detekcji za pomocą biblioteki Lingua.

Szczególnym wyzwaniem były krótkie tokeny (zwłaszcza jedno- lub dwuliterowe wyrazy), które występowały w tekstach we wszystkich trzech językach.

---

[36] Wszystkie anotacje przygotowano zgodnie z zasadami TEI XML, więcej o tym standardzie: Piotrowski, *Natural Language Processing for Historical Texts*, 60–8.

[37] https://pypi.org/project/lingua-language-detector/1.1.0/ [dostęp: 22.05.2025]. Pozostałe warte uwagi biblioteki obsługujące łacinę: *Langid*, *Polyglot*, *FastText*.

[38] Pod terminem „walidacja" rozumiem sprawdzenie i uwiarygodnienie wyniku przetwarzania.

[39] Skuteczność mierzono ręcznie na próbie 1000 tokenów, porównując wyniki z rozpoznaniem człowieka. Testowano również *Langid*, *Polyglot*, *FastText*, jednak dawały one więcej błędów przy krótkich tokenach.

[40] Np. takie formy, jak *konstytucyja* (constitutio) czy *egzorbitancyja* (exorbitantia) nie były błędnie klasyfikowane jako łacina, lecz poprawnie – jako polszczyzna.

[41] Za zastosowaniem takiego rozwiązania przemawiała przede wszystkim także prostota, elastyczność, jak również wydajność.

Ze względu na ich często niejednoznaczny charakter językowy, nie były one bezpośrednio przypisywane żadnemu językowi na podstawie samej formy. W takich przypadkach zastosowano mechanizm kontekstowej identyfikacji języka, polegający na analizie sąsiedztwa danego tokenu: jeśli zarówno lewy, jak i prawy kontekst (tj. poprzedzający i następujący token) zostały rozpoznane jako należące do tego samego języka, krótki token dziedziczył tę klasyfikację. Natomiast w sytuacji, gdy konteksty były niejednoznaczne lub wskazywały na różne języki, token otrzymywał oznaczenie „nieznany" (*null*) i trafiał do pliku przeznaczonego do weryfikacji ręcznej. Taki zabieg pozwolił ograniczyć ryzyko błędnego przypisania języka w przypadku formalnie identycznych słów funkcyjnych, a jednocześnie zachować wysoki poziom precyzji anotacji.

W dalszej części artykułu chciałbym się przyjrzeć kilku wybranym narzędziom i pakietom wykorzystywanym w przetwarzaniu historycznego języka naturalnego, które testowałem dla litewskich akt sejmikowych. Nie jest to pełne zestawienie dostępnych rozwiązań, raczej krótki przegląd najistotniejszych przykładów w kontekście analizowanych materiałów.

Katalog dostępnych rozwiązań otwierają narzędzia opracowane dla języka średniopolskiego, który dominuje w analizowanych aktach sejmikowych z epoki nowożytnej. Kluczowym komponentem w tym zakresie jest Morfeusz – analizator morfologiczny, pierwotnie stworzony dla współczesnej polszczyzny, który doczekał się specjalnej wersji dostosowanej do tekstów z XVII i XVIII wieku. Wersja ta, znana pod nazwą Korbeusz, bazuje na słowniku i gramatyce historycznej opracowanej w ramach projektu Korpus barokowy i umożliwia analizę morfologiczną form występujących w tekstach średniopolskich. Korbeusz generuje pełny zestaw możliwych interpretacji morfosyntaktycznych dla danej formy, jednak samodzielnie nie dokonuje rozstrzygnięcia w przypadku niejednoznaczności. W celu przeprowadzenia dezambiguacji zastosowano narzędzie Concraft 2 – kontekstowy tagger, który wykorzystuje statystyczne modele do wyboru najbardziej prawdopodobnej interpretacji na podstawie kontekstu sąsiednich wyrazów. Podobnie jak Morfeusz, Concraft został dostosowany do potrzeb przetwarzania języka średniopolskiego i korzysta z modeli trenowanych na danych historycznych[42]. W eksperymentach przeprowadzonych na

[42] Marcin Woliński, „Morfeusz Reloaded", w *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, red. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, Stelios Piperidis (Reykjavík: European Language Resources Association (ELRA), 2014): 1106–11; Jakub Waszczuk, Witold Kieraś, Marcin Woliński, „Morphosyntactic Disambiguation and Segmentation for Historical Polish With Graph-Based Conditional Random Fields", w *Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11–14, 2018, Proceedings*, red. Petr

Korpusie Barokowym Concraft uzyskał poprawność na poziomie 94%, radząc sobie z analizą słów nieznanych analizatorowi fleksyjnemu[43]. Choć zestaw narzędzi do analizy języka średniopolskiego powstał pierwotnie z myślą o korpusach literackich i kaznodziejskich, z powodzeniem został zaadaptowany do przetwarzania tekstów politycznych i prawnych, takich jak instrukcje czy lauda sejmikowe. Przykładem praktycznego zastosowania tych rozwiązań jest Korpus sejmikowy opracowany przez Szymona Rutkowskiego[44].

Nieodłącznym elementem akt sejmikowych jest język łaciński. Choć rzadziej występuje jako język główny dokumentu, regularnie pojawia się w formie wtrąceń, zwłaszcza w kontekście formuł urzędowych, cytatów lub elementów tytulatury. Z tego względu skuteczne przetwarzanie tego typu materiału wymaga uwzględnienia narzędzi wspierających analizę języka łacińskiego. Dodatkowym utrudnieniem jest fakt, że mamy tu do czynienia z łaciną nowożytną, zawierającą liczne oboczności od norm klasycznych, co może wpływać na precyzję automatycznej anotacji, zwłaszcza w zakresie tagowania morfosyntaktycznego.

Wśród dostępnych rozwiązań najwłaściwszym wyborem dla niniejszego projektu okazała się Stanza – uniwersalna, w pełni neuronowa biblioteka NLP oparta na architekturze modularnej, obsługująca ponad 60 języków, w tym również łacinę[45]. Oferuje ona kompletny *pipeline* przetwarzania tekstu – od tokenizacji i segmentacji zdań, przez lematyzację i tagowanie morfosyntaktyczne, aż po analizę składniową (*dependency parsing*) i rozpoznawanie nazw własnych (NER). Model łaciny w Stanza został wytrenowany na danych z repozytorium Universal Dependencies, co czyni go narzędziem zgodnym z nowoczesnymi standardami anotacyjnymi. Dodatkowym atutem było to, że ten sam pakiet można zastosować także do przetwarzania języka ruskiego, co pozwoliło uprościć *workflow* i zachować spójność anotacji w całym korpusie[46].

Sojka, Aleš Horák, Ivan Kopeček, Karel Pala, t. 11107 (Cham, Switzerland: Springer International Publishing, 2018), 188–96.

[43] https://chronofleks.nlp.ipipan.waw.pl/mod_ujedn/ [dostęp: 22.05.2025].

[44] Te narzędzia zostały także wykorzystane w aplikacji tego samego autora o nazwie Tribunica – program ten umożliwia segmentację plików PDF i TXT, przypisywanie nagłówków za pomocą wyrażeń regularnych, zapis danych w bazie SQLite oraz automatyczną anotację morfosyntaktyczną z wykorzystaniem Morfeusza i Concrafta. Gotowy korpus może zostać wyeksportowany w formacie TEI XML.

[45] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, Christopher D. Manning, „Stanza: A Python Natural Language Processing Toolkit for Many Human Languages", w *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–8 (Online, 2020), https://doi.org/10.48550/arXiv.2003.07082.

[46] Dostępne są również inne narzędzia do przetwarzania łaciny, m.in.: Classical Language Toolkit (CLTK), spaCy z modelem łacińskim, UDPipe czy model CRACOVIA.

Najwięcej problemów w analizie językowej sprawił język ruski w wariancie używanym na terenie Wielkiego Księstwa Litewskiego. Pomimo rosnącej liczby zasobów dla historycznych języków wschodniosłowiańskich dostępne modele oferowane m.in. w ramach pakietu Stanza nie przyniosły zadowalających rezultatów przy przetwarzaniu fragmentów anotowanych tekstów sejmikowych. Główna przyczyna tego stanu rzeczy wiąże się z ograniczeniami korpusu, na którym bazuje model TOROT Treebank, czyli Tromsø Old Russian and OCS Treebank. Obejmuje on około 145 000 tokenów tekstów staroruskich i średniowielkoruskich, jednak jego zawartość opiera się w dużej mierze na tekstach religijnych, kronikarskich i literackich z obszaru Państwa Moskiewskiego, głównie w języku staro-cerkiewno-słowiańskim[47]. Z tego względu modele oparte na TOROT – mimo wysokiej jakości anotacji lingwistycznej – nie są bezpośrednio dostosowane do wariantu języka używanego w źródłach z WKL.

Wobec takiego problemu, w ramach eksperymentu, zdecydowano się na wytrenowanie własnego modelu językowego dla języka ruskiego. Zrobiono to za pomocą Stanza, bazując na zasobie językowym The Ruthenian UD treebank, który zawiera teksty napisane na terenach WKL i Ukrainy z lat ok. 1300–1700[48]. Wytrenowany model osiągnął następującą skuteczność dla poszczególnych podstawowych tagów. Należy przy tym zaznaczyć, że użyty *treebank*[49] nie zawiera kolumny XPOS (językowo specyficznych tagów części mowy), co oznacza, że ocena nie obejmuje tego typu anotacji[50]:

– Tokenizacja (dzielenie tekstu na mniejsze jednostki, zwane tokenami[51]) – 97,55%,

– UPOS (uniwersalne części mowy) – 96,89%,

– UFeats (cechy morfologiczne) – 87,7%,

– AllTags (poprawność wszystkich komponentów) – 50,81%,

– Lematyzacja (przypisywanie każdemu słowu jego formy podstawowej) – 91,81%.

Poszczególne narzędzia różnią się sposobem pracy z danymi. Stanza działa w pełnym *pipeline*: sama segmentuje tekst na zdania i tokeny, a następnie zapisuje

[47] Helene Andreassen, Hanne Grønnestad, „The Tromsø Old Russian and OCS Treebank (TOROT)", *Scripta & e-Scripta* nr 14–15 (2015): 9–25.

[48] https://github.com/UniversalDependencies/UD_Old_East_Slavic-Ruthenian [dostęp: 22.05.2025].

[49] Korpus językowy, w którym każdy tekst został opatrzony bogatą anotacją lingwistyczną – najczęściej dotyczącą struktury składniowej i morfosyntaktycznej zdań.

[50] https://universaldependencies.org/treebanks/orv_ruthenian/index.html [dostęp: 22.05.2025]. Mimo to Stanza domyślnie i tak trenuje tag XPOS, który w tym wypadku osiągnął skuteczność na poziomie 55,76%.

[51] Najmniejsza jednostka podziału tekstu w procesie komputerowego przetwarzania języka, zwykle odpowiadająca pojedynczemu słowu, liczbie lub znakowi interpunkcyjnemu.

wyniki w formacie CoNLL-U, w którym każdy token ma zestaw atrybutów (lemat, UPOS, cechy morfologiczne, relacje składniowe). Morfeusz natomiast przyjmuje jako wejście pojedyncze słowo (ciąg znaków, np. „sprawować") i zwraca listę możliwych analiz morfologicznych, ale nie wykonuje segmentacji ani nie łączy informacji w strukturę zdania. W praktyce oznacza to konieczność wcześniejszej tokenizacji tekstu, aby wyniki z Morfeusza mogły być porównane lub scalone z danymi ze Stanzy.

W kontekście większości analizowanych dokumentów segmenty morfologiczne w praktyce odpowiadają tokenom, np. forma *bydź* otrzymuje zestaw cech [VERB, Inf, Pres], co odpowiada jednej jednostce tokenizacyjnej. Różnice w sposobie tokenizacji (np. traktowanie znaków interpunkcyjnych jako odrębnych jednostek) mogą jednak wpływać na wynikową strukturę pliku wyjściowego i wymagają wcześniejszego ujednolicenia. W zależności od potrzeb badań informację o segmentacji można zapisać w dodatkowych kolumnach formatu CoNLL-U (np. w polu MISC), co pozwala przechowywać dane o podziale bez komplikowania głównej warstwy tokenów. W prostszych analizach wystarcza operowanie wyłącznie na poziomie tokenów.

Mimo tych rozbieżności dostępność otwartych narzędzi programistycznych – zwłaszcza w środowisku Python – pozwala na opracowanie skryptów konwersji danych wejściowych i wyjściowych, które integrują różne komponenty w spójną ścieżkę przetwarzania. Dzięki temu możliwe jest wykorzystanie kilku narzędzi w ramach jednego *workflow*, bez konieczności ręcznej adaptacji każdego etapu[52].

Istotnym problemem pozostaje natomiast brak jednolitości w zakresie stosowanych zestawów tagów morfosyntaktycznych. Poszczególne biblioteki korzystają z własnych konwencji anotacyjnych, co utrudnia porównywalność i dalszą analizę wyników. Z tego względu postulatem na przyszłość pozostaje ujednolicenie tagsetu, najlepiej w oparciu o standard *Universal Dependencies*, który obecnie stanowi najpowszechniej stosowany model anotacji morfosyntaktycznej i składniowej w wielojęzycznych projektach językoznawczych.

*

Przetwarzanie języków historycznych stawia wiele wyzwań, które sytuują je w grupie tzw. języków o ograniczonych zasobach (*low-resource languages*).

[52] Niestety wymaga to od użytkownika lepszej znajomości programowania, ponieważ tworzy dość skomplikowany *workflow*.

Choć termin ten często kojarzony jest z językami mniejszościowymi, w przypadku języków dawnych – takich jak polszczyzna barokowa, ruski w wariancie WKL, łacina w wariancie nowożytnym – ograniczenia te są jeszcze bardziej dotkliwe. Brakuje nie tylko wielu zasobów językowych, lecz również narzędzi, które mogłyby wspierać ich analizę w sposób zautomatyzowany.

Analiza wielojęzycznych tekstów historycznych, takich jak akta sejmikowe z terenu Wielkiego Księstwa Litewskiego pokazuje, że skuteczne przetwarzanie tego typu materiału wymaga starannego przygotowania danych i świadomego doboru narzędzi. Mimo ograniczeń zasobów językowych możliwe jest opracowanie efektywnego *pipeline'u* przetwarzania, obejmującego ekstrakcję, czyszczenie, anotację i wstępną analizę morfosyntaktyczną tekstu.

Istotnym wyzwaniem pozostaje brak jednolitych standardów ortografii i notacji w edycjach źródłowych, co ogranicza automatyzację procesu i wymaga dostosowywania narzędzi do specyfiki materiału. Brak pełnej analizy historycznej ortografii sprawia, że teksty przygotowywane do NLP są kompromisem między wiernym odtworzeniem języka źródłowego a potrzebą standaryzacji danych.

Podjęte eksperymenty dowodzą, że nawet przy ograniczonych zasobach możliwe jest prowadzenie badań nad historycznymi językami polityki w sposób systematyczny i półzautumatyzowany lub zautomatyzowany[53]. Jednocześnie napotkane problemy wskazują na pilną potrzebę dalszej rozbudowy narzędzi, w szczególności w kierunku ujednolicania anotacji w standardzie Universal Dependencies, co otworzyłoby nowe możliwości dla badań ilościowych i jakościowych nad dawnymi odmianami języków.

Ze względu na ograniczenia praw autorskich nie mogę udostępnić efektów pracy w otwartym repozytorium. Rezultatem jest korpus tekstów opatrzony szczegółową anotacją morfosyntaktyczną[54] i językową, który stanowi podstawę moich dalszych badań doktorskich. Zasady jego wykorzystania oraz przykłady analiz zostaną przedstawione w przygotowywanej dysertacji.

[53] Obecnie prowadzę eksperymenty – przy wsparciu zespołu Instytutu Podstaw Informatyki PAN – nad wykorzystaniem jednego narzędzia do analizy wielojęzycznych tekstów, w których dochodzi do przełączania kodów językowych. Eksperyment obejmuje zarówno tagowanie części mowy, jak i parsowanie zależności składniowych.

[54] Za przykład poznakowania morfosyntaktycznego może posłużyć przytoczony wyżej Korpus sejmikowy lub Korpus barokowy.

BIBLIOGRAFIA

Źródła wydane

Акты, издаваемые Археографическою Коммиссиею. Т. 1–4, 7–8, 13. Вильно: Виленская Археографическая Комиссия, 1865–1886. [*Akty, izdavaemye Arkheograficheskoiu Kommissieiu*. T. 1–4, 7–8, 13. Vil'no: Vilenskaia Arkheograficheskaia Komissiia, 1865–1886.]

Bondyra, Wiesław, Jerzy Ternes, oprac. *Akta sejmikowe województwa lubelskiego: 1697–1763*. Lublin–Radom: Uniwersytet Marii Curie-Skłodowskiej, 2024.

Bondyra, Wiesław, Henryk Gmiterek, Jerzy Ternes, oprac. *Akta sejmikowe ziemi chełmskiej 1572–1668*. Lublin: Archiwum Państwowe w Lublinie–Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, 2013.

Gmiterek, Henryk, oprac. *Akta sejmikowe województwa lubelskiego 1572–1632*. Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, 2016.

Kempa, Tomasz, Tomasz Ambroziak, oprac. *Akta sejmikowe województwa wileńskiego (1566–1655).* Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, 2024.

Kraszewski, Igor, Michał Zwierzykowski, Robert Kołodziej, Andrzej Kamieński, oprac. *Akta sejmikowe województw poznańskiego i kaliskiego. Lata 1656–1668.* Poznań: Grupa Wydawnicza FNCE–Wydawnictwo Naukowe FNCE, 2024.

Lulewicz, Henryk, oprac. *Akta zjazdów stanów Wielkiego Księstwa Litewskiego.* T. 1: *Okresy bezkrólewi (1572–1576, 1586–1587, 1632, 1648, 1696–1697, 1706–1709, 1733–1735, 1763–1764)*. T. 2: *Okresy panowań królów elekcyjnych XVI–XVII wiek.* Warszawa: Neriton, 2006–2009.

Nagielski, Mirosław, red. nauk., Michał Bąk i in., oprac. *Lauda ziemi rawskiej i województwa rawskiego 1583–1793*. Warszawa: Neriton, 2017.

Saviščevas, Eugenijus, Jonas Drungilas, oprac. *Źródła do dziejów Żmudzi (1522–1648)*, red. Tomasz Kempa. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, 2023.

Stolicki, Jarosław, oprac. *Akta sejmiku podolskiego in hostico 1672–1698*. Kraków: Towarzystwo Naukowe „Societas Vistulana", 2002.

Zwierzykowski, Michał, Robert Kołodziej, oprac. *Akta sejmikowe województwa bełskiego. Lata 1572–1655*. Kraków: Księgarnia Akademicka, 2020.

Zwierzykowski, Michał, Robert Kołodziej, Andrzej Kamieński, oprac. *Akta sejmikowe województwa bełskiego: lata 1656–1695*. Kraków: Księgarnia Akademicka, 2023.

Opracowania

Ambroziak, Tomasz. „Jak wydawać cyrylickie akta sejmikowe? Analiza rosyjskich, ukraińskich i białoruskich współczesnych zasad wydawniczych oraz wybranej praktyki edytorskiej", cz. 1. *Miscellanea Historico-Iuridica* 21, nr 1 (2022): 321–45.

Andreassen, Helene, Hanne Grønnestad. „The Tromsø Old Russian and OCS Treebank (TOROT)". *Scripta & e-Scripta* nr 14–15 (2015): 9–25. https:// munin.uit.no/handle/10037/22366.

Augustyniak, Urszula. *Koncepcje narodu i społeczeństwa w literaturze plebejskiej od końca XVI do końca XVII w.* Warszawa: Państwowe Wydawnictwo Naukowe, 1989.

Augustyniak, Urszula. „Polska i łacińska terminologia ustrojowa w publicystyce politycznej epoki Wazów". W *Łacina jako język elit*, red. Jerzy Axer, 33-71. Warszawa: Wydawnictwo DiG, 2004.

Augustyniak, Urszula. „Wolność szlachcica w Rzeczypospolitej XVII w. Propozycje badawcze". *Przegląd Historyczny* 114 (2023): 23–49.

Хорошкевич, Анна Леонидовна, Сергей Михайлович Каштанов, ред. *Методические рекомендации по изданию и описанию материалов Литовской метрики*. Москва, Вильнюс, 1985. [Khoroshkyevich, Anna Leonidovna, Sergey Mikhailovich Kashtanov, red. *Metodicheskie rekomendatsii po izdaniyu i opisaniyu materialov Litovskoy metriky*. Moskva–Vil'nyus, 1985.]

Gruszczyński, Włodzimierz, Dorota Adamiec, Maciej Ogrodniczuk. „Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.)". *Polonica* 33 (2013): 311–8.

Johnson, Kyle P., Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, William J. B. Mattingly. „The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages". W *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, red. Heng Ji Jong, C. Park, Rui Xia, 20–9. Online. Association for Computational Linguistics, 2021. https://doi.org/ 10.18653/v1/2021.acl-demo.3.

Jusupović, Monika, red. *Akta sejmiku kowieńskiego z lat 1733–1795*. Warszawa: Neriton, 2019.

Lepszy, Kazimierz, oprac. *Instrukcja wydawnicza dla źródeł historycznych od XVI do połowy XIX wieku*. Wrocław: Ossolineum, 1953.

Maciszewski, Jarema. „Kultura polityczna Polski «złotego wieku»". W *Dzieje kultury politycznej w Polsce*, red. Jerzy A. Gierowski, 11–5. Warszawa: Państwowe Wydawnictwo Naukowe, 1977.

Mazur, Karol, *W stronę integracji z Koroną. Sejmiki Wołynia i Ukrainy w latach 1569–1648*. Warszawa: Wydawnictwo Neriton, 2006.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, Christopher D. Manning. „Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". W *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–8. Online (2020). https://doi.org/10.48550 /arXiv.2003.07082.

Piotrowski, Michael. *Natural Language Processing for Historical Texts*. San Rafael: Morgan & Claypool Publishers, 2012.

Rachuba, Andrzej. „Edycje akt sejmikowych z terenu Wielkiego Księstwa Litewskiego". *Miscellanea Historico-Iuridica* 21, nr 1 (2022): 347–63.

Woliński, Marcin. „Morfeusz Reloaded". W *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, red. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, 1106–11. Reykjavík: European Language Resources Association (ELRA).

Waszczuk, Jakub, Witold Kieraś, Marcin Woliński. „Morphosyntactic Disambiguation and Segmentation for Historical Polish With Graph-Based Conditional Random Fields". W *Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11–14, 2018, Proceedings*, red. Petr Sojka, Aleš Horák, Ivan Kopeček, Karel Pala, vol. 11107, 188–96. Cham, Switzerland: Springer International Publishing, 2018.

STRONY INTERNETOWE

Biblioteka pypdfium2. https://pypi.org/project/pypdfium2/ [dostęp: 27.04.2025].

Biblioteka PyTesseract. https://pypi.org/project/pytesseract/ [dostęp: 27.04.2025].

ChronoFleks. https://chronofleks.nlp.ipipan.waw.pl/mod_ujedn/ [dostęp: 27.04.2025].

Elektroniczny korpus tekstów polskich z XVII i XVIII wieku (do 1772 roku). „Korba – Korpus barokowy". https://korba.edu.pl/query_corpus/ [dostęp: 27.04.2025].

Korpus dokumentów sejmikowych Rzeczypospolitej przedrozbiorowej. „O korpusie”. https://sejmiki.nlp.ipipan.waw.pl/overview [dostęp: 27.04.2025].

Lingua Language Detector. https://pypi.org/project/lingua-language-detector/1.1.0/ [dostęp: 27.04.2025].

Repozytorium The Ruthenian UD Treebank. https://github.com/ UniversalDependencies/UD_Old_East_Slavic-Ruthenian [dostęp: 27.04.2025].

Tribunica – projekt zgromadzony w repozytorium GitLab. https://gitlab.com/ szmer/tribunicia/-/tree/master?ref_type=heads [dostęp: 27.04.2025].

Universal Depedencies, UD Old East Slavic Ruthenian. https://universaldependencies. org/treebanks/orv_ruthenian/index.html [dostęp: 22.05.2025].

Wyrażenia regularne – źródło. https://swtch.com/~rsc/regexp/regexp1.html [dostęp: 27.04.2025].