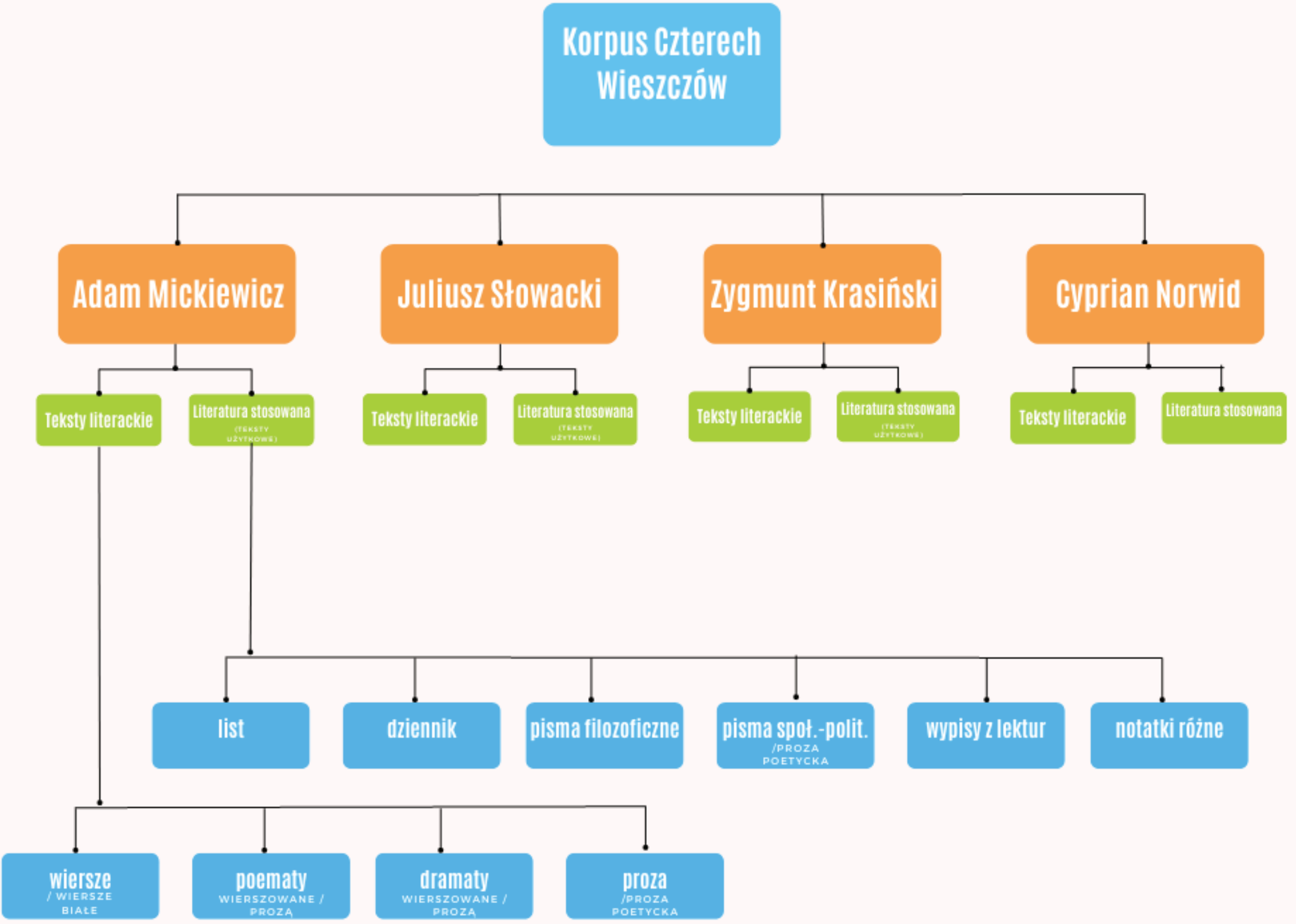
Korpus Czterech Wieszczów
Adam Mickiewicz
Juliusz Słowacki
Zygmunt Krasiński
Cyprian Norwid
Teksty literackie
Literatura stosowana (TEKSTY UŻYTKOWE)
Teksty literackie
Literatura stosowana (TEKSTY UŻYTKOWE)
Teksty literackie
Literatura stosowana (TEKSTY UŻYTKOWE)
Teksty literackie
Literatura stosowana
list
dziennik
pisma filozoficzne
pisma społ.-polit. /PROZA POETYCKA
wypisy z lektur
notatki różne
wiersze / WIERSZE BIAŁE
poematy WIERSZOWANE / PROZĄ
dramaty WIERSZOWANE / PROZĄ
proza /PROZA POETYCKA

DOI: http://dx.doi.org/10.12775/SE.2023.0011

Anna Mędrzecka-Stefańska

Między korpusem a edycją cyfrową – na przykładzie projektu *Korpus Czterech Wieszczów*

Instytut Badań Literackich PAN, kontakt: anna.medrzecka@ibl.waw.pl,
ORCID ID: 0000-0003-1165-5793

Truizmem jest stwierdzenie o niezwykłej wadze twórczości czterech wieszczów[1] – Adama Mickiewicza, Juliusza Słowackiego, Zygmunta Krasińskiego i Cypriana Norwida – dla rozwoju polskiej literatury, języka, a także dla polskiego dziedzictwa kulturowego. Teksty wszystkich tych autorów są nieustająco obiektem zainteresowania kolejnych pokoleń badaczy, a ponadto są obecne w kanonie lektur szkolnych, w języku debaty publicznej i jako podstawa licznych zwrotów wykorzystywanych w codziennym życiu pozostają niezbywalnym punktem odniesienia dla wszystkich użytkowników języka. Mimo to w erze powszechnej cyfryzacji dostępność ich utworów w sieci jest zadziwiająco niepełna.

Pilny czytelnik znajdzie zapewne przechowywane w repozytoriach, np. w Polonie, zeskanowane wersje dawnych wydań[2] (w tym *Pisma* Zygmunta Krasińskiego w opracowaniu Jana Czubka czy *Pisma wszystkie* Cypriana Norwida w edycji przygotowanej przez Wiktora Gomulickiego), po które można w ograniczonym stopniu sięgnąć w pracy badawczej. Materiały te jednak są trudno przeszukiwalne i wyposażone jedynie w podstawowy (i niepozbawiony błędów) OCR zeskanowanych stron. Korzystanie z niego nie przynosi znaczących pożytków w stosunku do wersji papierowej, nie licząc oczywiście większej dostępności wersji internetowej (co jest jej

Sztuka Edycji 1/2023
ISSN 2084-7963 (print)
ISSN 2391-7903 (online)
s. 113–120

główną zaletą i choćby z tego powodu istnienie repozytoriów i bibliotek cyfrowych jest zjawiskiem w najwyższym stopniu pożądanym).

Drugim rodzajem materiałów, które można znaleźć w sieci, są serwisy takie jak Wolne Lektury, oferujące łatwy dostęp do wybranych dzieł. Z uwagi na przeznaczenie podobnych portali – powstających przede wszystkim z myślą o użytku szkolnym – ich zasoby składają się głównie z tekstów objętych kanonem lektur i niewielkiego zbioru innych utworów. Nie ma tu miejsca na szczegółowy przegląd dzieł wieszczów dostępnych w takich serwisach; dla zilustrowania problemu warto jednak przywołać kilka przykładów.

Przede wszystkim edycje dostępne w zasobach internetowych podobnych do Wolnych Lektur czy Wikimedii nie spełniają podstawowych wymogów rzetelności naukowej. Osoby odpowiedzialne za przygotowanie danego dzieła często są anonimowe (jak np. „Wikiskrybowie"). Serwis Wolne Lektury podaje wprawdzie informację o podstawie tekstowej oraz nazwisko osoby odpowiedzialnej za opracowanie jej w wersji cyfrowej, lecz wciąż można mu wiele zarzucić, biorąc pod uwagę rzetelność w przygotowywaniu edycji. Wybór podstaw opiera się wyłącznie na kryterium dostępności, bez weryfikacji ich wiarygodności. Zazwyczaj za podstawę utworu przyjmowane są wydania nie tyle nieaktualne, ile archaiczne, nawet dziewiętnastowieczne (zob. np. A. Mickiewicz, *Poezye*, Nakład Księgarni G. Gebethnera i Spółki, Kraków 1899, na podstawie tej edycji podano wiersz *Do*** na Alpach w Splügen*). Aparat krytyczny jest w nich ograniczony do minimum, w dodatku trudno zweryfikować jego poprawność. Trzeba mieć też świadomość, że teksty udostępniane w serwisie Wolne Lektury mogą być modyfikowane przez przypadkowych „edytorów". Podsumowując, tak podane i opracowane utwory w żadnym wypadku nie mogą być wykorzystywane w badaniach naukowych.

W Polsce brakuje też naukowych edycji cyfrowych dzieł wielkich romantyków (dyskusję nad tym, czy Norwida wypada zaliczać do romantyków, pozostawiam na marginesie tych rozważań)[3]. Przygotowana zgodnie ze standardami obowiązującymi tego typu wydania[4], naukowa edycja cyfrowa prezentowałaby rzetelnie opracowane teksty literackie, mogące stać się podstawą dalszych badań. Mielibyśmy w tym przypadku do czynienia z istotną wartością dodaną, jaką jest chociażby anotowanie tekstu systemem znaczników (tagów) grupujących określone przez edytora typy słów i rodzaje ich użycia w kategorie, które można następnie przeszukiwać i analizować. Wykorzystanie do tego celu standardu TEI (Text Encoding Initiative) pozwoliłoby na opracowanie utworów różnych twórców w sposób spójny i porównywalny, a także umożliwiłoby śledzenie powiązań wewnątrz poszczególnych utworów i czytelne przedstawienie zależności między nimi.

Edycja cyfrowa pozwala też na przedstawienie „genetyki tekstu", ukazując jego dynamiczny i płynny charakter[5], co w przypadku opracowywanych utworów ma znaczenie niebagatelne. Około 80% spuścizny Juliusza Słowackiego to teksty niewydane za życia poety, pozostawione w rękopisie; *Nie-Boska komedia* Zygmunta Krasińskiego ma dwa alternatywne zakończenia, a edytorzy znajdują przekonujące argumenty na rzecz każdej z nich i muszą podejmować arbitralne decyzje dotyczące podstawy wydania[6]. Dzięki edycji cyfrowej można pokazać dzieje tekstu z perspektywy zarówno genetyki, jak i recepcji, w sposób przewyższający możliwości edycji papierowej. Ponadto naukowa edycja cyfrowa wpływa na poszerzenie doświadczenia czytelniczego dzięki umożliwieniu odbiorcy, także niespecjaliście, takiego rodzaju zaangażowania w pracę z tekstem, które współtworzy jego lekturowe doświadczenie, dając większą swobodę – oczywiście w stopniu ograniczonym i zaplanowanym przez edytora cyfrowego – wobec wyroków, rozstrzygnięć i narzuceń edytorskich[7].

> W Polsce brakuje naukowych edycji cyfrowych dzieł wielkich romantyków

Sytuację, w której brakuje naukowej edycji cyfrowej dzieł omawianych poetów[8], do pewnego stopnia może naprawić *Korpus Czterech Wieszczów* tworzony przez zespół pod kierownictwem prof. Marka Troszyńskiego w ramach współpracy z CLARIN-PL[9]. Powstający korpus nie stanowi jednak edycji cyfrowej w wyżej opisanym znaczeniu, nie jest też zwykłym repozytorium tekstów o charakterze archiwum, jak np. Polona. Teksty zostaną przygotowane i opracowane w sposób umożliwiający wykorzystanie dotychczas niedostępnych metod badania dziedzictwa polskich romantyków.

Korpus tekstów – narzędzie poznania języka

Językoznawcy już kilkadziesiąt lat temu zainteresowali się możliwością wykorzystania komputerów w badaniach statystycznych nad językiem[10], czego efektem było powstanie językoznawstwa korpusowego. Dzisiaj już wiemy, że praca językoznawców, leksykografów czy nawet tłumaczy

jest utrudniona, jeśli nie korzystają oni z korpusów. Dzięki tworzeniu coraz większych i bardziej zróżnicowanych zbiorów odpowiednio opisanych (anotowanych) tekstów możliwe są zaawansowane analizy; korpusy mogą też być pomocne w codziennej pracy językoznawców jako narzędzie szybkiego i wygodnego pozyskiwania informacji statystycznych z ogromnych zbiorów danych, których opracowanie ręczne byłoby praktycznie niemożliwe. Odpowiednio przygotowany korpus może dostarczyć informacji na temat typowego użycia słów, kolokacji, konstrukcji gramatycznych, statystyk dotyczących częstotliwości występowania określonych form językowych itp. Korpusy wspierają badania leksykograficzne, lingwistyczne, są doskonałym narzędziem ułatwiającym pracę tłumaczy czy edytorów. Potrafią przy tym osiągać objętość miliardów jednostek i są wyposażone w zaawansowane wyszukiwanie[11].

Warto zaznaczyć, że o ile w badaniach *stricte* językoznawczych wykorzystanie korpusów ma już długą tradycję, o tyle literaturoznawcy wciąż nie sięgają po ten rodzaj zasobów. Tymczasem *distant reading*, metoda analizy tekstu, która swoją nazwę wzięła od klasycznej pracy Franca Morettiego[12], nie dzięki uważnej lekturze i skupieniu na szczególe, ale właśnie dzięki analizie zbiorów danych zbyt obszernych, by możliwe było ich opanowanie i opracowanie przez człowieka, może wprowadzić nową jakość do badań literaturoznawczych. Metoda ta umożliwia wydobycie zobiektywizowanych danych i poddanie ich interpretacji przez badacza dysponującego odpowiednią wiedzą i intuicją, które są niezbędne do formułowania wniosków na podstawie zgromadzonego materiału[13]. Na gruncie polskim pionierską pracą wykorzystującą analizy korpusowe w badaniach literackich jest rozprawa doktorska autorki prezentowanego artykułu powstająca w IBL PAN pod kierunkiem prof. Marka Troszyńskiego. Analizie poddano w niej kategorię ironii w poematach Juliusza Słowackiego[14] i zbadano obecność określonych struktur gramatycznych i stylistycznych w utworach uwzględnionych w korpusie. Zaproponowaną metodę badawczą polegającą na wydobyciu z tekstu struktur i wykładników językowych charakterystycznych dla danego zjawiska literackiego, a następnie prześledzeniu ich występowania w badanym materiale, można z powodzeniem zastosować także w trakcie analizy innych zabiegów stylistycznych, elementów konstrukcji tekstu, obecności w nich określonych motywów literackich itp. Prawdopodobnie istotnym powodem, dla którego badania korpusowe nie są wykorzystywane przez polskich literaturoznawców w analizach stylu wybranych autorów, wspieraniu hipotez interpretacyjnych danymi statystycznymi czy po prostu w sprawnym przeszukiwaniu i anotowaniu (oznaczaniu) tekstów literackich, jest brak znaczących, ogólnodostępnych korpusów.

W Polsce jak dotąd opracowano korpusy polszczyzny ogólnej (chodzi przede wszystkim o *Narodowy Korpus Języka Polskiego*, nkjp.pl) oraz korpusy polszczyzny historycznej (jak np. *Korpus tekstów staropolskich*[15] czy *Korpus polszczyzny XVI wieku*[16], a także *Elektroniczny Korpus Tekstów Polskich z XVII i XVIII wieku*[17]). Badacz romantyzmu powinien być zainteresowany zwłaszcza korpusem dziewiętnastowiecznej polszczyzny, powstałym w ramach projektu *Automatycznej analizy fleksyjnej tekstów polskich z lat 1830–1918*. Jego powstaniu przyświecały dwa cele:

– opis systemowych zmian w zakresie odmiany polszczyzny pisanej w latach 1830–1918,
– stworzenie słownika fleksyjnego ukazującego ewolucję danej odmiany[18].

Korpus ten składa się z około miliona segmentów i zawiera (w formie próbek stworzonych na podstawie zasady zachowania całych wypowiedzeń, nie zaś uzyskania fragmentów o równej liczbie słów) prace popularnonaukowe, wiadomości prasowe, teksty publicystyczne, prozę i utwory dramatyczne. Możliwe jest przeszukiwanie korpusu według kryteriów fleksyjnych i metatekstowych.

Zasób jest nieocenioną pomocą dla osób zajmujących się badaniem polszczyzny XIX wieku, ale z punktu widzenia literaturoznawcy to jedynie narzędzie referencyjne. Opracowanie pełnego anotowanego, czyli opatrzonego zestawem komentarzy, korpusu dzieł najważniejszych polskich romantyków pozwoli na prowadzenie pogłębionych analiz nad ich idiolektem – rozpoczętych już wiele lat temu przez wybitnych badaczy, niedysponujących jednak wówczas narzędziami komputerowymi umożliwiającymi zautomatyzowanie pracy[19].

Korpus Czterech Wieszczów powstaje z wykorzystaniem narzędzi informatycznych wypracowanych m.in. w odniesieniu do słownika polszczyzny XIX wieku (analizator morfosyntaktyczny Morfeusz 2)[20], a także rozbudowanej infrastruktury CLARIN-PL[21]. Warto podkreślić, że z perspektywy badań korpusowych mamy do czynienia nie z opracowanym edytorsko tekstem, ale z zasobem pozwalającym przede wszystkim na badanie słownictwa poszczególnych twórców, a dzięki funkcji porównania podkorpusów możliwe jest zestawienie zasobu leksykalnego i stylu różnych autorów.

Dobór zasobów

Korpus Czterech Wieszczów z założenia uwzględnia cały zachowany piśmienny dorobek Adama Mickiewicza, Juliusza Słowackiego, Zygmunta Krasińskiego i Cypriana Norwida. Oznacza to, że oprócz tekstów tradycyjnie uznawanych za dzieła literackie (jak wiersze, dramaty, poematy czy utwory prozatorskie) w obręb korpusu zostały włączone także różnego rodzaju szkice, notatki, inedita czy listy, które zwłaszcza w przypadku Zygmunta Krasińskiego mają znaczący udział w całości podkorpusu.

Ponieważ zadaniem zespołu nie jest przygotowanie nowej edycji, omawiany zasób nie obejmuje nowych odczytań ani rozstrzygnięć i nie powstaje w odniesieniu do rękopisów i pierwodruków[22]. Zgromadzony materiał opiera się na istniejących już, cenionych wydaniach krytycznych. Członkowie zespołu dokonali wyboru podstaw dla tekstów zebranych w poszczególnych podkorpusach według następujących kryteriów: kompletność edycji, jej wiarygodność i zgodność z najnowszymi ustaleniami badawczymi i tendencjami edytorskimi, powszechność wykorzystywania w badaniach naukowych oraz funkcjonowania w obiegu czytelniczym[23].

Ponadto trzeba zaznaczyć, że w korpusie uwzględniono również wydawane później inedita, teksty pozostające w rękopisach, a także poprawiono oczywiste błędy podstawy. W każdym przypadku osoba odpowiedzialna za dany podkorpus musiała podjąć właściwą decyzję.

Przyjęte w dawnych wydaniach rozwiązania są anachroniczne i nie przystają do wymogów współczesnego edytorstwa, ponadto różnią się między sobą zarówno strategią przyjętą przez wydawców, zakresem modernizacji pisowni, jak i stosunkiem do podstaw tekstowych. W przypadku starszych wydań (tak rzecz się ma z *Pismami* Krasińskiego pod redakcją Jana Czubka) zdarza się, że zapis wymaga uwspółcześnienia. Dlatego aby stworzyć korpus, w którym możliwe będzie porównywanie ze sobą dzieł poszczególnych autorów, konieczne są wszelkiego rodzaju działania ujednolicające – przede wszystkim należy ustalić wspólne dla wszystkich podkorpusów zasady modernizacji pisowni, trzeba także ujednolicić zapisy liczb i dat. Zespół opracował też jednolite zasady postępowania w przypadku słownictwa, którego zaliczenie w poczet leksyki danego autora może budzić wątpliwości bądź które nie zawsze było uwzględniane przez edytorów – chodzi m.in. o wyrażenia obcojęzyczne, zapisy brulionowe, porzucone wersje tekstu[24].

Zespół zgodnie stwierdził, że ze względu na niewystarczającą jakość cyfrowych materiałów dostępnych w repozytoriach internetowych oraz niespełniający kryteriów naukowości charakter tekstów zamieszonych w serwisach typu Wolne Lektury, w przypadku żadnego z autorów nie jest możliwe oparcie się na źródłach online. Dlatego konieczne było zdigitalizowanie wybranych podstaw, a następnie przekonwertowanie plików graficznych do plików tekstowych i poddanie ich korekcie. Teksty zostały oczyszczone z komentarza edytorskiego, usunięto więc z nich wszystkie pozaautorskie uzupełnienia, takie jak pochodzące od wydawców tytuły dzieł, nazwy postaci dramatu czy wątpliwe rekonstrukcje brakujących fragmentów. Pozostawiono zaś rozwinięcia zaznaczonych przez autora nazw i wyrazów, a także poprawne zapisy wyrazów, które w oryginale były niepełne, co wiąże się chociażby z uszkodzeniami mechanicznymi (*casus* zniszczonego rękopisu).

Literatura cyfrowymi oczami

Tradycyjnie edycja – zarówno wydanie papierowe, jak i cyfrowe – ma na celu wypracowanie właściwej formy podania tekstu źródłowego i, zależnie od przyjętych założeń, ułatwienie jego lektury bądź pokazanie go tak, by wskazać wszystkie trudności, jakie napotyka czytelnik przyzwyczajony do linearnego poznawania tekstu[25]. Korpus powstaje w odniesieniu do zupełnie innych założeń – choć jest zbiorem tekstów, jego przeznaczeniem nie jest lektura linearna. Jest on traktowany jako zestaw danych – przede wszystkim językowych, lecz to, jakie dokładnie informacje będzie można wydobyć z korpusu, zależy od tego, jak został on przygotowany i przetworzony za pomocą narzędzi komputerowych.

Opracowanie korpusu, oprócz wyboru odpowiednich podstaw i ich korekt, obejmuje też sprowadzenie danych do formy, która będzie możliwa do rozpoznania przez dostępne narzędzia informatyczne. Pierwszym krokiem jest przeprowadzenie lematyzacji oraz tagowania korpusu. W wypadku *Korpusu Czterech Wieszczów* wykorzystuje się do tego narzędzie LEM[26], które dzieli tekst na tokeny, czyli najmniejsze wydzielane jednostki znaczące – najczęściej odpowiedniki wyrazów tekstowych, a następnie przypisuje każdemu z nich odpowiednią interpretację morfosyntaktyczną: przyporządkowuje do lematu, a więc odpowiedniej formy hasłowej, oraz opisuje za pomocą tagu, czyli uporządkowanego zestawu wartości w określonych klasach i kategoriach

gramatycznych. Klasy gramatyczne to szersze zbiory uporządkowane pod względem cech morfologicznych. Każdej klasie przyporządkowany jest zestaw kategorii gramatycznych, które z kolei mogą przyjmować określone wartości; szczegóły dotyczące kategorii gramatycznych i ich wartości są opisane w tagsecie[27]. W efekcie tagowania tekst zostaje podzielony na oznakowane segmenty. Przykładowo zdanie otwierające *Podróż do Ziemi Świętej z Neapolu* Juliusza Słowackiego: „Muzo mdlejąca z romantycznych cierpień!", może być podzielone na tokeny:

[Muzo][mdlejąca][z][romantycznych][cierpień],

które w procesie lematyzacji narzędzie przyporządkuje do lematów: Muza, mdleć, z, romantyczny, cierpienie. Następnie w procesie tagowania morfosyntaktycznego każdemu słowu zostanie przypisany zestaw tagów opisujących jego formę gramatyczną. Efekt tych działań można przedstawić tak:

Wyraz tekstowy	Leksem	Tag morfosyntaktyczny
Muzo	Muza	subst:sg:voc:m1
mdlejąca	mdleć	pact:sg:nom:f:imperf:aff
z	z	prep:gen:nwok
romantycznych	romantyczny	adj:pl:gen:n:pos
cierpień	cierpienie	subst:pl:gen:n
!	!	interp

Lematyzacja i tagowanie pozwolą na przeprowadzenie wstępnych badań korpusowych, takich jak tworzenie list frekwencyjnych czy przeszukiwanie korpusu. Dzięki otagowaniu całości korpusu możliwe jest wyszukiwanie lematów reprezentowanych przez dowolną formę wyrazową. Wszystkie wystąpienia lematu „poeta" zostaną więc wyszukane i zgrupowane w ramach jednego wyniku, niezależnie od tego, w jakim przypadku czy w jakiej liczbie słowo będzie użyte w tekście. Dotyczy to także lematów, które w różnych formach gramatycznych mają brzmienie znacząco różne od formy podstawowej (jak chociażby czasownik „być"). Możliwe jest też wyszukiwanie według tagów, tj. wskazanie wszystkich leksemów występujących w określonej formie gramatycznej. Otagowanie korpusu pozwala na przeprowadzanie następnych działań wymagających uprzedniej segmentacji materiału. Ponieważ materiały wchodzące w skład podkorpusów *Korpusu Czterech Wieszczów* zostaną ujednolicone pod względem modernizacji pisowni, mogą być analizowane za pomocą tego samego tagera. Wyniki takiej analizy dla poszczególnych podkorpusów będzie można ze sobą porównywać.

Przygotowane teksty zostaną umieszczone w systemie Inforex umożliwiającym opatrzenie tekstu bogatym zestawem metadanych i anotacji. Dzięki temu dostępna będzie funkcja zaawansowanego przeszukiwania korpusu, obejmującego wyszukiwanie określonych form i leksemów, lecz także filtrowanie wyników wyszukiwania do jednostek spełniających określone kryteria (np. tylko czasowniki w rodzaju żeńskim występujące w poematach z 1832 roku, tylko inedita, przymiotniki w rodzaju męskim występujące w tekstach podpisanych pseudonimem itp.). Będzie też możliwe wydobywanie statystyk dotyczących warstwy językowej całego korpusu, takich jak lista frekwencyjna (lista słów występujących w korpusie, poszczególnych podkorpusach, a także w pojedynczych tekstach, uporządkowana w kolejności od najczęściej do najrzadziej występujących), opis cech gramatycznych czy inne podobne właściwości tekstów, korpusów i podkorpusów.

Gdy teksty zostaną udostępnione szerokiemu gronu badaczy, będzie też możliwe analizowanie ich za pomocą innych narzędzi z infrastruktury CLARIN-PL oraz pobranie zasobu i dalsze jego wykorzystanie. Zaawansowane opcje wyszukiwania pozwolą na weryfikowanie hipotez interpretacyjnych przez śledzenie kontekstów występowania słów czy całych fraz (także w różnych wariantach i odmianach), analizowanie występowania określonych struktur językowych czy innych wyznaczników stylu.

Anotacje i metadane

Istotnym elementem w naszym projekcie, pozwalającym na uzyskanie porównywalności wyników analiz poszczególnych podkorpusów, jest stworzenie jednolitego zestawu anotacji i metadanych. Na anotacje składają się tagi i komentarze stanowiące warstwę opisu tekstu na poziomie pojedynczych słów, zdań bądź fragmentów. Mogą być one nanoszone automatycznie, półautomatycznie lub ręcznie. Nazwy własne w ramach *Korpusu Czterech Wieszczów* są np. anotowane automatycznie, ponieważ w obrębie serwisu LEM infrastruktura CLARIN-PL daje taką właśnie możliwość. O ile jednak podczas opisywania tekstu za pomocą znaczników TEI możemy stworzyć odnośnik do opisu oznaczanej jednostki hasłowej (w przypadku TEI Panorama nazwanej bytem), takiej jak osoba, miejsce czy organizacja, a także zebrać w jednym miejscu wszystkie wystąpienia danego bytu, przejrzeć szablony użycia i prześledzić sieć powiązań[28], o tyle w omawianym

systemie anotacji dostępnym w serwisie Inforex oznaczeniu podlega jedynie klasyfikacja określonego wyrazu do danej kategorii. Zatem we fragmencie listu Juliusza Słowackiego zapis „Pornik" w oznaczaniu TEI może odsyłać do bytu o nazwie „Pornic", w opisie którego zostaną zamieszczone szczegółowe informacje o tej miejscowości i jej roli w biografii Juliusza Słowackiego i w polskiej kulturze, a także odnośnik do pozostałych dotyczących jej wzmianek, niekoniecznie z wykorzystaniem nazwy. Tymczasem w opisie korpusowym pojawi się jedynie informacja, że jest to nazwa własna miejscowa, a prześledzenie jej wystąpień w tekście będzie możliwe tylko dzięki wyszukaniu różnych jej form i zapisów. Anotacje nazw własnych będą nanoszone automatycznie, a następnie weryfikowane przez członków zespołu.

Zestawy anotacji są tworzone przez badaczy i możliwy jest taki ich dobór, który będzie odpowiadał przewidywanym potrzebom przyszłych odbiorców i użytkowników korpusu. Anotacje w *Korpusie Czterech Wieszczów* zostały podzielone na kilka warstw wydobywających z tekstów różne istotne informacje.

Aby uporządkować pliki w relacji do papierowych podstaw tekstów, została opracowana lokalizacyjna warstwa anotacji. Jednostką w korpusie nie jest strona, a plik. Z przyczyn technicznych pliki mają swoją maksymalną objętość, jednak rzadko zdarza się, by fizyczne ograniczenia wielkości pliku wymusiły na zespole opracowującym korpus zastosowanie sztucznego podziału w tekście. Nadrzędną zasadą, stosowaną tam, gdzie to tylko możliwe, jest zasada 1 tekst = 1 plik. Podział większego utworu na pliki, jeśli jest konieczny, odbywa się z poszanowaniem wewnętrznego podziału (w dramatach jest to podział na akty i sceny, w poematach na pieśni, w tekstach prozatorskich na rozdziały itp.). Dlatego aby umożliwić badaczom korzystającym z korpusu porównanie tekstu z podstawą czy odniesienie się do wersji papierowej wydania, wprowadzono warstwę anotacji podającą lokalizację każdego fragmentu w podstawie. Anotacje przekształcają się w informację dotyczącą każdego pojedynczego słowa, tak by badacz mógł precyzyjnie zlokalizować wszystkie wyrazy.

Druga warstwa anotacji opisuje działania modernizacyjne podjęte przez twórców korpusu. Jak już wspomniano, aby ułatwić, a czasami nawet umożliwić pracę narzędzia, a także by nie tworzyć wrażenia sztucznego zwiększenia objętości zasobu leksykalnego, uwspółcześniono dawne zapisy i formy oraz ujednolicono je między korpusami. Jednak aby korzystający z korpusu badacze mieli dostęp do wersji tekstu zgodnej z podstawą, w warstwie anotacyjnej zostaną wskazane wszystkie zabiegi modernizacyjne.

W ramach kolejnej warstwy anotacji oznaczono pozaautorskie uzupełnienia i wtrącenia, a także teksty nieznanej proweniencji, jak tłumaczenia, cytaty, streszczenia, wypisy itp. Ostatnia warstwa zawiera informacje o poprawionych błędach (literówki, błędne odczytania, które zostały już wcześniej skorygowane, błędy odnotowane w erracie itp.).

Dzięki takiemu systemowi anotacji poruszanie się po korpusie jest o wiele łatwiejsze i oferuje użytkownikowi zestaw narzędzi umożliwiających tworzenie nawet skomplikowanych statystyk. Jednocześnie warto ponownie podkreślić, że mamy tu do czynienia przede wszystkim ze zbiorem danych służącym badaniom zasobu leksykalnego, a nie z edycją tekstów literackich. Stąd zarówno zasób oznaczanych informacji, jak i sposób ich oznaczania dostosowano do założonego wcześniej celu. Także w przypadku korpusu subiektywne decyzje badacza mają istotny wpływ na ostateczny kształt zasobu.

Druga warstwa anotacji opisuje działania modernizacyjne

Aby możliwe było porównanie podkorpusów pod kątem informacji zewnętrznych wobec tekstów, stworzono spójny i przejrzysty zestaw metadanych, czyli pozatekstowych informacji dotyczących danego pliku lub tekstu. Odpowiednio dobrany system metadanych pozwala na uporządkowanie całego zasobu chociażby pod względem chronologii powstawania tekstów, ich pierwodruków, rękopisów, odpisów itp. Metadane będą również zawierać wskazówki genologiczne oraz np. informację o tym, czy tekst został wydany za życia autora lub czy został podpisany własnym nazwiskiem.

Opisanie tekstów spójnym zestawem metadanych pozwoli więc wydobyć statystyki dotyczące m.in. okresów bardziej lub mniej intensywnej pisarskiej aktywności, liczby tekstów reprezentujących różne gatunki, a połączenie ze sobą informacji pochodzących z metadanych, anotacji i tagowania umożliwi przeprowadzenie skomplikowanych operacji opartych na systemie wielu zmiennych.

Podsumowanie

Jak wynika z powyższych rozważań, cele tworzenia korpusów są inne niż cele tworzenia edycji cyfrowych lub repozytoriów. Korpus służy uzyskiwaniu informacji statystycznych na poziomie tekstu, poszczególnych podkorpusów i całego korpusu. Dzięki przyjęciu spójnych zasad opracowania wszystkich czterech podkorpusów możliwe będzie porównanie języka i stylu

Czterech Wieszczów Romantycznych – a do tej pory było to niemożliwe ze względu na różne rozwiązania edytorskie stosowane przez wydawców, a także przez autorów istniejących słowników (chodzi o *Słownik języka Adama Mickiewicza*, *Słownik rymów Juliusza Słowackiego*, *Internetowy słownik języka Cypriana Norwida*)[29].

Choć *Korpus Czterech Wieszczów* – podkreślmy to raz jeszcze – nie jest edycją cyfrową, może wypełnić lukę w dostępności w internecie rzetelnie przygotowanego cyfrowego zasobu twórczości poetów i stać się nieodzownym narzędziem pracy badaczy polskiej literatury i kultury. W przyszłości możliwe będzie przygotowanie korpusów dzieł kolejnych twórców według tych samych zasad, które przyjęto podczas tworzenia *Korpusu Czterech Wieszczów*, co dodatkowo poszerzy pole do badań porównawczych. Ukończony korpus zostanie umieszczony w wolnym dostępie, aby mogli z niego skorzystać nie tylko specjaliści, ale wszyscy zainteresowani użytkownicy.

Key Words: corpus, natural language processing, Polish Romanticism, Adam Mickiewicz, Juliusz Słowacki, Zygmunt Krasiński, Cyprian Norwid, idiolect, vocabulary

Abstract: The text discusses the bases of the *Corpus of The Great Four* (*Korpus Czterech Wieszczów*) project, the aim of which is to create a database of all texts by Juliusz Słowacki, Adam Mickiewicz, Zygmunt Krasiński and Cyprian Norwid. The article, above all, discusses the most important differences between the corpus approach and the digital scholarly edition. The fundamental difference is in the purpose – the aim of editors' work is to provide the reader with a text that is prepared and fit to read, while the corpus is primarily used for research and analysis. Thanks to the unified rules for developing the sub-corpora of all four poets, it will also be possible to compare them with each other. The introduction of a set of annotations will allow for better organization of the material and enable new research directions.

[1] W ramach projektu używamy tego określenia z całą świadomością uproszczeń, które z nim się wiążą. Zdecydowaliśmy się na nie, mając na uwadze zwiększenie rozpoznawalności przedsięwzięcia, jako że jego odbiorcami mają być nie tylko literaturoznawcy. Projekt jest skierowany także do odbiorców niespecjalistów.

[2] Osobnym wątkiem, wykraczającym poza tematykę tego artykułu, jest zróżnicowana obecność edycji dzieł wymienionych poetów w formacie tradycyjnym. Zadaniem członków zespołu odpowiedzialnych za poszczególne podkorpusy było przeprowadzenie specjalnej kwerendy związanej z dostępnością wydań spełniających określone w założeniach projektu kryteria i wybór odpowiednich podstaw tekstowych.

[3] Brak naukowych edycji cyfrowych dotyczy nie tylko dzieł literatury romantycznej. Więcej na ten temat w artykule Bartłomieja Szleszyńskiego *Krajobraz naukowego edytorstwa cyfrowego – możliwości i wyzwania* zamieszczonym w tym numerze „Sztuki Edycji".

[4] P. Sahle, *What Is a Scholarly Digital Edition?*, w: *Digital Scholarly Editing: Theories and Practices*, eds. M. Driscoll, E. Pierazzo, Cambridge 2016, https://www.google.com/url?q=https://books.openedition.org/obp/3397&sa=D&source=docs&ust=1677859569335276&usg=AOvVaw2AR-9LkwZLCrmogRHQ8O61 (dostęp: 3.03.2023). Także w języku polskim istnieje spora już literatura dotycząca naukowych edycji cyfrowych rozpatrywanych z perspektywy teoretycznej, jak i rozwiązań praktycznych. Więcej na ten temat w: K. Niciński, *O strukturalizmie w działaniu, czyli TEI i naukowa edycja cyfrowa z perspektywy praktyki edytorskiej* (artykuł zamieszczony w prezentowanym numerze „Sztuki Edycji"). Istotne pod tym względem są publikacje wydawane w ramach serii „Filologia XXI", w tym szczególnie: J. Bryant, *Płynny tekst*, tłum. Ł. Cybulski, Warszawa 2020; J. McGann, *Nowa Republica litteraria. Pamięć i nauka w wieku cyfryzacji*, tłum. P. Bem et al., Warszawa 2016; *Od Gutenberga do Google'a. Elektroniczne reprezentacje tekstów literackich*, tłum. P. Bem, red. nauk. A. P. Lesiakowski, Warszawa 2020. Zob. też: B. Hojdis, A. Cankudis, *Cyfrowe edycje literackie i naukowe jako element cyfrowej humanistyki*, „Sztuka Edycji. Studia Tekstologiczne i Edytorskie" 2020, nr 1 (17): *Tekstografie. Edytorskie przestrzenie tekstu i obrazu*, pod red. M. Czerwińskiego, I. Kropidło i O. Taranek-Wolańskiej, s. 7–15. Milena M. Śliwińska pisała o naukowej edycji cyfrowej dzieł Zygmunta Krasińskiego (*Dzieła zebrane Zygmunta Krasińskiego – elektroniczna edycja naukowa*, „Sztuka Edycji. Studia Tekstologiczne i Edytorskie" 2014, nr 1–2 (6): *Zygmunt Krasiński – edycje i interpretacje*, pod red. M. Strzyżewskiego, s. 69–75), a w Centrum Humanistyki Cyfrowej IBL PAN zespół w składzie Marek Troszyński, Ewa Mirkowska i Anna Mędrzecka-Stefańska opracowuje edycję cyfrową *Samuela Zborowskiego*, która ma być postawą do przygotowania kompletnej edycji cyfrowej dzieł Juliusza Słowackiego.

[5] J. Bryant, op. cit.

[6] M. Bizior-Dombrowska, *Wstęp*, w: Z. Krasiński, *Nie-Boska komedia. Wydanie krytyczne pierwodruku z 1835 roku*, Toruń 2015.

[7] DSE „is about giving readers the freedom to engage with texts in a meaningful and creative way"; Ch. Ohge, *Publishing Scholarly Editions. Archives, Computing, and Experience*, Cambridge 2021, s. 121.

[8] O edycji cyfrowej *Samuela Zborowskiego* zob. A. Mędrzecka-Stefańska, E. Mirkowska, *Słowacki cyfrowy – perspektywy i szanse edycji cyfrowej*, „Sztuka Edycji. Studia Tekstologiczne i Edytorskie" 2023, nr 2 (24): *Edycje romantyków – interpretacje romantyzmu*, pod red. M. Strzyżewskiego i A. Markuszewskiej (w druku).

[9] W skład zespołu wchodzą też: Tomasz Korpysz, Ewa Mirkowska i Anna Mędrzecka-Stefańska, wsparcia informatycznego udziela nam zespół Politechniki Wrocławskiej pod kierunkiem Marcina Oleksego; T. Korpysz et al., *„Korpus Czterech Wieszczów" – cyfrowy wymiar dziedzictwa narodowego*, „Poradnik Językowy" 2022, nr 7, s. 67–78.

[10] H. Kucera, W. N. Francis, *Computational Analysis of Present-Day American English*, Providence 1967. W Polsce dość wcześnie pojawiło się zainteresowanie badaniami ilościowymi na materiale literackim (zob. *Poetyka i matematyka*, pod red. M. R. Mayenowej, Warszawa 1965; prace Jerzego Woronczaka i Władysława Kuraszkiewicza czy przełomową książkę Jadwigi Sambor pt. *Badania statystyczne nad słownictwem (na materiale „Pana Tadeusza")*, Wrocław–Warszawa–Kraków 1969 oraz późniejsze publikacje tej badaczki), jednak na przejście do badań komputerowych trzeba było długo czekać; A. Pawłowski, *Lingwistyka kwantytatywna a humanistyka cyfrowa: ciągłość czy zmiana?*, w: *Polszczyzna w dobie cyfryzacji*, pod red. A. Hąci, K. Kłosińskiej i P. Zbróga, Warszawa 2020.

[11] *Narodowy Korpus Języka Polskiego*, pod red. A. Przepiórkowskiego et al., Warszawa 2012. Literatura dotycząca tworzenia i wykorzystywania korpusów jest obszerna, warto w tym miejscu przywołać kilka podstawowych pozycji, interesujących zwłaszcza z punktu widzenia problemów poruszanych w artykule: *Podstawy językoznawstwa korpusowego*, pod red. B. Lewandowskiej-Tomaszczy, Łódź 2005; D. Biber, S. Conrad, R. Reppen, *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge 2008; A. Lüdeling, M. Kytö, *Corpus Linguistics: An International Handbook*, Berlin 2010; T. McEnery, A. Hardie, *Corpus Linguistics: Method, Theory and Practice*, Cambridge 2011; M. Piasecki, *Cele i zadania lingwistyki informatycznej*, w: *Metodologie językoznawstwa. Współczesne tendencje i kontrowersje*, pod red. P. Stalmaszczyka, Kraków 2008; M. Świdziński, *Lingwistyka korpusowa w Polsce – źródła, stan, perspektywy*, „LingVaria" 2006, nr 1, s. 23–34.

[12] F. Moretti, *Distant Reading*, London 2013.

[13] Należy przy tym mieć się na baczności, aby nie ulec pokusie nadmiernego zaufania do obiektywizmu danych uzyskanych w efekcie analiz korpusów; M. Eder, *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii*, „Teksty Drugie" 2014, nr 2, s. 90–105.

[14] Projekt był prezentowany podczas CLARIN Annual Conference 2020; *The Literary Irony in the Works of Juliusz Słowacki*, w: *Selected Papers from the CLARIN Annual Conference 2020*, https://ecp.ep.liu.se/index.php/clarin/article/view/16 (dostęp: 14.04.2023).

[15] Zob. https://ijp.pan.pl/publikacje-i-materialy/zasoby/korpus-tekstow-staropolskich/ (dostęp: 14.04.2023).

[16] Zob. http://spxvi.edu.pl/korpus/ (dostęp: 14.04.2023).

[17] Zob. https://korba.edu.pl/query_corpus/ (dostęp: 14.04.2023).

[18] Zob. http://www.f19.uw.edu.pl/o-projekcie/ (dostęp: 14.04.2023).

[19] Zob. przede wszystkim pionierską pracę J. Sambor, *Badania statystyczne nad słownictwem (na materiale „Pana Tadeusza")*; eadem, *Słowa i liczby*, Wrocław 1972. Badania statystyczne nad językiem pisarzy są wciąż istotnym elementem refleksji naukowej; oprócz opracowań odnoszących się do języka pisarza w ogólności powstają też opracowania cząstkowe; M. Wojtyńska-Nowotka, *Słownik języka Maurycego Mochnackiego (na podstawie „Rozpraw literackich")*, Warszawa 2020. Warto zaznaczyć, że Milena Wojtyńska-Nowotka nie korzystała z metod korpusowych.

[20] Zob. http://morfeusz.sgjp.pl/ (dostęp: 14.04.2023).

[21] Zob. https://clarin-pl.eu/index.php/uslugi/ (dostęp: 14.04.2023).

[22] Od tej zasady odstąpiono jednak w przypadku utworów Juliusza Słowackiego, które w ostatnim czasie zostały wydane na podstawie rękopisów, zdecydowano się więc na wykorzystanie najnowszych edycji. *Samuel Zborowski* został podany według wydania: M. Troszyński, *Alchemia rękopisu. „Samuel Zborowski" Juliusza Słowackiego*, Warszawa 2017. Utwory, których rękopisy znalazły się w raptularzu wschodnim, podano za: *Raptularz Wschodni Juliusza Słowackiego*, t. 2: *Edycja – komentarz – objaśnienia*, oprac. M. Kalinowska (poemat), Z. Przychodniak (wiersze), M. Troszyński (proza), pod red. Z. Przychodniaka, Warszawa 2019.

[23] Szczegółowo o kryteriach doboru podstaw w: T. Korpysz et al., op. cit., s. 70–72.

[24] Zasady te opisano w: ibidem, s. 72–74.

[25] J. Bryant, op. cit.; J. McGann, op. cit.; Ch. Ohge, op. cit. Rola edycji krytycznych i stawiane im wymogi były tematem wielu publikacji na przestrzeni dziesięcioleci; zob. np. R. Loth, *Podstawowe problemy i pojęcia tekstologii i edytorstwa naukowego*, Warszawa 2006; teksty zawarte w „Pamiętniku Literackim" (2020, z. 4), zwłaszcza: D. Pawelec, *Edytorstwo jako krytyka literacka*, s. 5–18; P. Bem, *Tekst „edytorski" czy „wieloautorski". Kilka uwag o heterogeniczności edytorstwa naukowego*, s. 19–30; M. Prussak, *Zmierzch edycji krytycznych?*, s. 31–43; M. Strzyżewski, *O szaleństwie edytorów w perspektywie „ratio" i „praxis"*, s. 45–56. Zob. też: G. P. Bąbiak, *Pisma zebrane a edycje krytyczne polskiej literatury. Zarys zjawiska*, „Sztuka Edycji. Studia Tekstologiczne i Edytorskie" 2021, nr 1 (19): *Tradycja a wymogi współczesności w pracy edytora*, pod red. M. Strzyżewskiego, s. 26–44.

[26] M. Piasecki, T. Walkowiak, M. Maryl, *Literary Exploration Machine: A New Tool for Distant Readers of Polish Literature*, 2017, https://ws.clarin-pl.eu/lem# (dostęp: 11.03.2023).

[27] Najczęściej wykorzystywanym przez narzędzia badające język polski tagsetem jest ten opracowany na potrzeby *Narodowego Korpusu Języka Polskiego*, http://nkjp.pl/poliqarp/help/ense2.html (dostęp: 14.04.2023).

[28] Zob. https://tei.nplp.pl (dostęp: 14.04.2023).

[29] *Słownik języka Adama Mickiewicza*, t. 1–11, pod red. K. Górskiego i S. Hrabca, Wrocław 1962–1983; M. Jeżowski, *Słownik rymów Juliusza Słowackiego*, Warszawa 2002; https://www.slownikjezykanorwida.uw.edu.pl/ (dostęp: 5.05.2023).