DOI: 10.33896/PorJ.2022.7.5

Tomasz Korpysz

(Instytut Językoznawstwa
Uniwersytet Kardynała Stefana Wyszyńskiego, Warszawa,
t.korpysz@uksw.edu.pl)

ORCID: 0000-0001-6578-5839

Anna Mędrzecka

(Centrum Humanistyki Cyfrowej
Instytut Badań Literackich Polskiej Akademii Nauk,
anna.medrzecka@ibl.waw.pl)

ORCID: 0000-0003-1165-5793

Ewa Mirkowska

(Centrum Humanistyki Cyfrowej
Instytut Badań Literackich Polskiej Akademii Nauk,
ewa.mirkowska-treugutt@ibl.waw.pl)

ORCID: 0000-0002-3921-7325

Marek Troszyński

(Centrum Humanistyki Cyfrowej
Instytut Badań Literackich Polskiej Akademii Nauk,
marek.troszynski@ibl.waw.pl)

ORCID: 0000-0002-1825-0513

„KORPUS CZTERECH WIESZCZÓW” – CYFROWY WYMIAR DZIEDZICTWA NARODOWEGO. ZAŁOŻENIA PROJEKTU

„Korpus Czterech Wieszczów” jest roboczą nazwą projektu realizowanego przez CLARIN-PL (Common Language Resources & Technology Infrastructure) we współpracy z Instytutem Badań Literackich Polskiej Akademii Nauk[1]. Cyfrowy korpus

[1] Po wielu latach wstępnych prac cząstkowych i wspólnych działań (przede wszystkim związanych z tekstami Juliusza Słowackiego) projekt został afiliowany przy CLARIN-PL dzięki zaangażowaniu koordynatora konsorcjum dr. hab. inż. Macieja Piaseckiego, który dostrzegł i docenił jego potencjał także jako nowego wyzwania dla infrastruktury cyfrowej w humanistyce. Autorem koncepcji „Korpusu Czterech Wieszczów” jest Marek Troszyński, a szczegółowe zasady ujednolicenia danych poszczególnych podkorpusów wypracowuje zespół w składzie: Marek Troszyński, Tomasz Korpysz, Ewa Mirkowska,

wieszczów będzie zbiorem obejmującym polskojęzyczne teksty twórców polskiego tzw. wielkiego romantyzmu: Adama Mickiewicza (1798–1855), Juliusza Słowackiego (1809–1849), Zygmunta Krasińskiego (1812–1859) i Cypriana Norwida (1821–1883)[2]. Chronologiczny zasięg korpusu wyznaczają graniczne daty: 1818 – debiut najstarszego z twórców, Mickiewicza, a z drugiej strony 1883 – śmierć najmłodszego, Norwida.

Urodzeni między 1798 a 1822 rokiem pisarze, określeni tu jako „czterej wieszczowie", działali w zbliżonych środowiskach, łączyły ich powiązania towarzyskie, a ich twórczość zawiera liczne wzajemne odniesienia. Twórczość ta oraz idiolekty wymienionych autorów bardzo silnie i różnorako wpłynęły na polską kulturę (zarówno wysoką, jak i popularną) oraz samą polszczyznę i są w nich wyraźnie obecne do dziś. Mimo to jak dotąd brak jest opracowań i baz danych, które pozwalałyby na w pełni wiarygodne analizy tych autorskich języków i ujmowały je we wzajemnych odniesieniach. „Korpus Czterech Wieszczów" nie tylko umożliwi takie badania, lecz także pozwoli na weryfikację utrwalonych w polskiej kulturze literackich i pozaliterackich odwołań do tekstów wymienionych twórców. Przede wszystkim jednak pozwoli na uzyskiwanie wielopoziomowych danych na temat słownictwa poszczególnych autorów i ich dzieł[3].

Niezależnie od znaczenia projektu jako korpusu czterech pisarzy będzie on także dawał wgląd w polszczyznę dziewiętnastowieczną z okresu pozbawionego obszer-

---

Anna Mędrzecka. Ze strony CLARIN-PL w pracach bieżących za wsparcie informatyczne odpowiada Marcin Oleksy, logistyką zaś i sprawami administracyjnymi zajmuje się Tomasz Bernaś. Cyfrowy format projektu gwarantuje, że w ciągu następnych lat wyniki działań będą stopniowo umieszczane w wolnym dostępie *online*.

[2] W związku z tym, że projekt ma wymiar nie tylko naukowy, lecz także popularyzatorski, a jego efekt będzie dostępny również dla niespecjalistów, używamy terminu *romantyzm* w jego ogólnie przyjętym rozumieniu i abstrahujemy od toczącej się od lat dyskusji o tym, czy Cyprian Norwid był twórcą romantycznym czy też nie. Najnowszymi głosami w tej dyskusji są: artykuł Piotra Chlebowskiego (zob. Chlebowski 2021: 127–168) oraz fragmenty książki Magdaleny Siwiec (zob. Siwiec 2021: 55–131).

[3] Różnorakie analizy leksyki to najczęściej podejmowane zagadnienie w obrębie bogatego nurtu polskich badań języka pisarzy. Słownictwo było głównym przedmiotem opisu już pierwszych prac z tego zakresu, powstałych w drugiej połowie XIX wieku (por. Gosiewska 1949), to właśnie wybieranego według różnorakich kryteriów zasobu leksykalnego dotyczy większość szczegółowych opracowań materiałowych, różnorakie analizy leksyki zwykle stanową też zasadniczą część monografii poświęconych językowi wybranych twórców (por. np. Sławkowa 2009, 2011; Bobrowski 2015; Kozłowska 2018). W obszernej literaturze przedmiotu wielokrotnie podnoszono potrzebę tworzenia całościowych słowników idiolektalnych oraz różnorakich indeksów, konkordancji, list frekwencyjnych. Obecnie istnieją jedynie słowniki języka Adama Mickiewicza (zob. Górski, Grabec red. 1962–1983), Jana Chryzostoma Paska (zob. Koneczna, Doroszewski red. 1965–1973) i Jana Kochanowskiego (zob. Kucała red. 1994–2012). W przygotowaniu jest internetowy słownik języka Cypriana Norwida (zob. Puzynina, Korpysz 2009–). Powstają też opracowania cząstkowe, jak np. *Słownik osobliwości leksykalnych Stanisława Wyspiańskiego na materiale utworów dramatycznych* (zob. Śliwiński red. 2016) czy *Słownik języka Maurycego Mochnackiego (na podstawie „Rozpraw literackich")* (zob. Wojtyńska-Nowotka 2020). Szerzej zob. Korpysz 2010; Sobolewska 2010; Kozłowska 2013.

niejszych prac leksykograficznych. Pomiędzy sześciotomowym *Słownikiem języka polskiego* Samuela Bogumiła Lindego (1807–1814) a wileńskim *Słownikiem języka polskiego*, wydanym staraniem Orgelbranda (1861) istnieje wyraźna luka leksykograficzna[4], tymczasem okres ten – podobnie jak cały wiek XIX[5] – jest czasem intensywnego rozwoju polszczyzny, w tym zwłaszcza różnorakich zmian jej zasobu leksykalnego (zob. np. Bajerowa 1986, 1992, 2000; Kwapień 2010). Opracowany korpus, wobec braku słownika z równoległego okresu[6], w jakimś stopniu umożliwi obserwacje zasobu leksykalnego polszczyzny ogólnej, stanie się też ważnym punktem odniesienia dla powstającego tzw. gronowego mikrokorpusu polszczyzny XIX wieku (zob. Derwojedowa 2020).

„Korpus Czterech Wieszczów" oraz zintegrowane z nim narzędzia informatyczne pozwolą na łatwe i wielopoziomowe wyszukiwanie różnorodnych danych dotyczących języka poszczególnych twórców, a także na porównywanie danych dotyczących ich idiolektów (zwłaszcza leksyki), co będzie mogło stać się podstawą wielorakich badań nie tylko z zakresu historii literatury, teorii literatury, edytorstwa i językoznawstwa (historii języka, idiolektologii[7], stylistyki), ale też np. kulturoznawstwa, filozofii, historii idei czy teologii. System autorskich podkorpusów powiązany będzie ze słownikiem form hasłowych (lematów) oraz zostanie wyposażony w narzędzie umożliwiające wielopoziomową analizę warstwy leksykalnej wraz z kolokacjami[8].

---

[4] Słowniki te, zwłaszcza leksykon Lindego, nie mogą być przy tym traktowane jako w pełni wiarygodne źródła wiedzy o polszczyźnie ogólnej.

[5] Abstrahujemy tu od problemu, czy wiek XIX można traktować jako wyodrębniający się okres dziejów polszczyzny – por. np. Dubisz 2012.

[6] Warto zaznaczyć, że zdaniem wielu badaczy problem jest szerszy i „mimo istnienia słowników ogólnych języka polskiego oraz wielu monografii – brakuje naukowego opracowania leksyki ogólnopolskiej w XIX wieku i to zarówno w postaci monografii, jak i słownika czy też leksykonu, który obejmowałby leksykę typową dla tego okresu (ogólnopolską, standardową)" (Kwapień 2014: 260). Jedynie w małej części warunek ten spełniają także nowsze opracowania – zob. przede wszystkim Kwapień 2013; Wawrzyńczyk 2019.

[7] Termin *idiolektologia*, stosowany przede wszystkim przez badaczy języka osobniczego ze środowiska językoznawczego UKSW, został zaproponowany w pracy: Kozłowska 2018; szerzej zob. Korpysz, Kozłowska [w druku].

[8] Tworzenie przeszukiwalnych zasobów językowych jest współcześnie jednym z ważnych elementów badań lingwistycznych na całym świecie. Publicznie dostępne *online* są już liczne korpusy i edycje cyfrowe, część z nich wyposażona przy tym została w rozbudowane narzędzia umożliwiające analizę udostępnionego materiału. Niemożliwe jest w tym miejscu omówienie wszelkich dostępnych źródeł danych tekstowych, jednak jako interesujące przykłady można wymienić chociażby korpus dzieł wszystkich Shakespeare'a (https://www.opensourceshakespeare.org/), dający dostęp do wyszukiwarki korpusowej oraz zestawień konkordancji, a także repozytorium korpusów tekstów dramatycznych w różnych językach DraCor (https://dracor.org/), dające dostęp do analiz kontekstowych oraz umożliwiające pobranie odpowiednich danych w różnych formatach. Bardzo ciekawym zasobem jest archiwum korespondencji Bruckhardta (https://burckhardtsource.org/), pozwalające na wyświetlanie i przeszukiwanie listów pod różnym kątem (archiwum opracowane jest w standardzie TEI). Szeroki zasób narzędzi do przeszukiwania oraz przeprowadzania analiz oferuje korpus posiedzeń sądowych OldBailey (https://www.oldbaileyonline.org/) – nie jest to wprawdzie zasób tekstów o charakterze literackim,

Docelowo korpus składać się będzie z czterech podkorpusów, obejmujących całość leksyki danego autora, użytej w jego dostępnych tekstach. Oprócz dzieł literackich jako źródło danych uwzględnione zostaną także wszystkie inne zachowane teksty (korespondencja, rozprawy, odezwy, przemówienia, recenzje, notatki, wypisy, dedykacje, napisy na rysunkach itp.). Wyłączone zostaną jedynie obszerne pisemne wypowiedzi w językach obcych, jednak już obcojęzyczne pojedyncze wyrazy, wyrażenia czy nawet zdania spójnie wplecione w tok polskiego tekstu zostaną uwzględnione. Zgodnie z najnowszymi tendencjami do korpusów w pewnym zakresie włączone zostaną także teksty o niepewnym statusie, np. nieautoryzowane wypowiedzi spisane przez słuchaczy. Dotyczyć to będzie jednakże tylko wypadków, które utrwalono w obiegu czytelniczym jako teksty danego autora. Najbardziej znaczącym i objętościowo pokaźnym przykładem jest polska wersja prelekcji paryskich Adama Mickiewicza, skompilowana na podstawie w większości nieautoryzowanych przekładów. Leksyka pochodząca z tego typu tekstów zostanie w specjalny sposób oznaczona, tak aby zainteresowany użytkownik mógł z niej korzystać lub wyłączyć ją z ogólnego zbioru danych poddawanych analizie.

Najważniejszym kryterium doboru edycji, które mają stać się podstawą korpusu, jest ich kompletność lub względna kompletność. Zastrzeżenie co do „względnej kompletności" wynika z tego, że w wypadku Juliusza Słowackiego wciąż jeszcze istnieją obszerniejsze teksty niedrukowane oraz teksty odnalezione po zamknięciu dotychczasowych wydań zbiorowych (zob. Słowacki 1952–1975, 1959; Sawrymowicz oprac. 1962–1963) i drukowane osobno (zob. Słowacki 1996; Kalinowska, Makowska, Przychodniak, Troszyński, Kaja oprac. 2019), a w wypadku wszystkich czterech twórców odnajdowane są *inedita* lub też podstawy (rękopisy, pierwodruki, autoryzowane odpisy itp.) uznawane dotychczas za zaginione. Takie źródła z oczywistych względów nie mogły zostać uwzględnione nawet w tych istniejących wydaniach, które w zamierzeniu edytorów miały być kompletne. Z pewnymi zastrzeżeniami można uznać, że w wypadku dwóch autorów, czyli Adama Mickiewicza i Cypriana Norwida, takie wydania istnieją (odpowiednio: tzw. Wydanie Rocznicowe – zob. Mickiewicz 1993–2005 oraz *Pisma wszystkie*, opracowane przez Juliusza Wiktora Gomulickiego – zob. Norwid 1971–1976) i po koniecznych weryfikacjach oraz

ale narzędzia do analizy korpusu wykorzystane w wypadku tego zasobu są inspirujące także z perspektywy projektu „Korpusu Czterech Wieszczów". Również w Polsce przygotowywane są kolejne korpusy i bazy danych – zarówno dotyczące polszczyzny ogólnej (w tym także XIX-wiecznej; przykładem jest tzw. mikrokorpus gronowy polszczyzny 1830–1918 – zob. Derwojedowa 2020, por. Derwojedowa, Kieraś, Skowrońska 2014), jak i wybranych typów tekstów czy autorów. Opisy kilku powstających korpusów, analizy wybranych szczegółowych zagadnień lingwistyki korpusowej, jak również informacje o wielu innych oraz bogatą literaturę przedmiotu znaleźć można m.in. w monograficznych numerach „Prac Filologicznych" (t. 67/2015) i „Poradnika Językowego" (z. 8/2020); zob. też np. Górski 2003; Twardzik 2003; Świdziński 2006; Hebal-Jezierska red. 2014; Klapper, Kołodziej 2014; Bronikowska, Przyborska-Szulc 2018; Majdak 2018.

uzupełnieniach mogą one stanowić punkt wyjścia do tworzenia korpusu. Z kolei w wypadku Juliusza Słowackiego i Zygmunta Krasińskiego odpowiednich edycji brak. Najpełniejsze (czyli zawierające największą liczbę tekstów uwzględnianych w korpusie) są odpowiednio: wydanie opracowane przez Juliusza Kleinera (zob. Słowacki 1952–1975) oraz edycja toruńska pod red. Mirosława Strzyżewskiego (zob. Krasiński 2017) (ta jednak nie zawiera listów). Podstawą wyboru mogą być zatem wskazane wyżej edycje, nie muszą one jednak ostatecznie być traktowane jako „kanoniczne" i uznane za źródło danego podkorpusu ze względu na to, że omawiane kryterium, choć główne i najważniejsze, nie jest jedyne i nie zawsze może być uznane za rozstrzygające.

Kolejnym ważnym kryterium doboru podstaw tekstowych jest wiarygodność edycji, jej charakter, zgodność z najnowszymi ustaleniami badawczymi oraz tendencjami w edytorstwie. Z tego względu co do zasady pierwszeństwo należy przyznać edycjom naukowym, krytycznym, a nie popularnym (choćby te drugie były obszerniejsze). Jako podstawa zostały zatem wybrane wydania oparte na rękopisach i pierwodrukach, opatrzone komentarzami edytorskimi, uwzględniające autorskie poprawki, krytycznie prezentujące historię wydania danego tekstu itp. Zasady tej nie można było przyjąć bezwyjątkowo, ponieważ w długiej historii edycji tekstów czterech wieszczów zdarzają się wydania krytyczne, które były jedynie cząstkowe, a które później stały się podstawą wydań popularnych. Włączano do nich niekiedy teksty opracowane wcześniej na potrzeby całościowego lub cząstkowego wydania krytycznego (przykładem mogą być choćby *Dzieła zebrane* Cypriana Norwida w opracowaniu Juliusza Wiktora Gomulickiego (zob. Norwid 1966) i opublikowane później przez tego samego edytora *Pisma wszystkie* (zob. Norwid 1971–1976) czy też *Dzieła wszystkie* Adama Mickiewicza pod red. Konrada Górskiego (zob. Mickiewicz 1969–) oraz późniejsze tzw. Wydanie Rocznicowe (zob. Mickiewicz 1993–2005).

Z drugiej strony wydania krytyczne niekiedy mogą być przestarzałe z punktu widzenia dzisiejszego stanu badań, nowoczesnych metodologii i współczesnych tendencji edytorskich, a nowsze wydania, niemające charakteru i ambicji wydań krytycznych, bywają pod tym względem bardziej odpowiednie na potrzeby projektu (jak np. PIW-owskie edycje listów Zygmunta Krasińskiego – zob. Krasiński 1963, 1965, 1970, 1971, 1975, 1977, 1979, 1980, 1988, 1991a, 1991b).

Powyższe kryteria uzasadniają to, dlaczego należy odrzucić obecne w Internecie edycje i „paraedycje" cyfrowe oraz teksty na różnych zasadach i przez różne podmioty udostępniane w wolnym dostępie. Coraz więcej tekstów (a nawet całe ich obszerne bloki) znaleźć można na różnorakich stronach internetowych, w bibliotekach internetowych, w serwisach typu *Wolne lektury* itp. Żadna jednak z takich publikacji nie spełnia kryteriów poprawności naukowej i weryfikowalności, żadnej nie można uznać za edycję krytyczną (żadna też nie jest kompletna) i w konsekwencji żadna nie może zostać wykorzystana w projekcie. Na jego potrzeby wszystkie wyda-

nia uznane za wiarygodne podstawy tekstów zostaną więc zeskanowane i cyfrowo przygotowane.

Ostatnie kryterium odwołuje się do tzw. *social approach*, a więc do szeroko rozumianej tradycji wykorzystywania danej edycji w pracach naukowych, do jej zakorzenienia i funkcjonowania w obiegu czytelniczym, do jej utrwalenia jako podstawy kolejnych popularnych wydań, do jej stosowania jako źródła tekstów wykorzystywanych np. w podręcznikach itp. Kryterium to jest niezwykle istotne, ponieważ końcowy efekt projektu ma być dostępny dla wszystkich – nie tylko dla wąskiej grupy specjalistów dobrze znających literaturę przedmiotu i swobodnie się w niej poruszających. W konsekwencji przyjęcie jako najważniejszej podstawy jakiegoś wydania poprawnego edytorsko, ale bardzo specjalistycznego, niszowego, trudno dostępnego itp. sprawiłoby, że przyszli użytkownicy nie mogliby odpowiednio wykorzystać, a nawet zweryfikować otrzymanych danych, ponieważ trudno byłoby im sięgnąć do tekstu źródłowego i choćby zlokalizować cytat czy też umiejscowić dane słowo w szerszym kontekście. Tymczasem istnieją edycje, które mimo krytycznych głosów są – także przez specjalistów – stale wykorzystywane w pracy naukowej, a przede wszystkim funkcjonują w szerokim obiegu czytelniczym (zob. Norwid 2009–).

Wymienione wyżej trzy główne kryteria wyboru edycji jako podstawy korpusu w odniesieniu do każdego twórcy wymagają głębokiego namysłu i wieloaspektowej oceny dostępnych źródeł. Efektem tego procesu będzie wskazanie konkretnego wydania jako podstawowego, „kanonicznego". W wypadku żadnego z autorów takie wydanie nie może zostać uznane za podstawę jedyną. Po pierwsze, obok niego muszą zostać uwzględnione wydawane później (czy też równolegle) *inedita* oraz ewentualne teksty pozostające w rękopisach. Po drugie, nierzadko konkretne teksty lub ich fragmenty zostały w danej edycji – zweryfikowanej pod kątem kompletności, poprawności i utrwalenia w świadomości czytelniczej, a w konsekwencji przyjętej jako podstawa korpusu – opracowane lub po prostu odczytane błędnie, co poprawiono w późniejszych opracowaniach szczegółowych lub innych wydaniach. Jeśli tego typu zmiany są istotne dla zasobu leksykalnego danego twórcy, to jako podstawę dla takiego konkretnego tekstu traktuje się edycję osobną czy też inną niż „kanoniczna". W konsekwencji w bazie korpusowej mogą się znaleźć wydania konkretnych tekstów spoza takiej edycji. Stworzenie bazy edycji źródłowych jest zatem pracą twórczą, wymagającą znajomości poszczególnych edycji oraz szerzej: literatury przedmiotu dotyczącej danego twórcy. Teksty i leksemy oraz ich użycia pochodzące z tekstów spoza wydania „kanonicznego" będą miały przypisany system akronimów wyraźnie wskazujących na ich lokalizacje, żeby użytkownik mógł je łatwo odszukać we właściwej podstawie.

Zgodnie z założeniami celem projektu „Korpus Czterech Wieszczów" jest opracowanie wiarygodnej, weryfikowalnej bazy, która umożliwi nie tylko pozyskiwanie różnorakich danych dotyczących języka poszczególnych twórców czy też konkret-

nych, wybranych dzieł oraz wyodrębnionych z nich zgodnie z potrzebami użytkownika typów czy zbiorów słów, lecz także – a może przede wszystkim – badania kontrastywne. Unikatowy zbiór danych wyposażony w odpowiednie narzędzia informatyczne pozwoli na porównywanie zasobu leksykalnego wszystkich autorów w rozmaicie wybranych przekrojach[9]. Aby ten cel osiągnąć, w trakcie kolejnych etapów projektu zostaną wypracowane spójne szczegółowe zasady tworzenia poszczególnych podkorpusów oraz wykorzystania określonych narzędzi informatycznych[10].

Niezależnie od wcześniej omówionych głównych założeń dotychczasowa praca nad tekstami ujawniła konieczność uzgodnienia konkretnych procedur w wypadkach szczególnych. Dotyczy to na przykład postępowania wobec tekstów stworzonych w całości w językach obcych. Leksyka z tych tekstów, jak już o tym była mowa wyżej, nie zostanie włączona do korpusu (chyba że pojawią się w niej pojedyncze słowa polskie – te zostaną uwzględnione), ale same takie teksty zostaną skatalogowane, w specjalny sposób oznaczone oraz pod pewnymi względami scharakteryzowane (język, objętość, lokalizacja). Uzgodnienia wymagało też postępowanie w odniesieniu do pojedynczych słów zapożyczonych i wtrąceń obcojęzycznych wplecionych w tekst pisany w języku polskim. Ze względu na to, że wszyscy twórcy uwzględnieni w projekcie dużą część życia spędzili na emigracji, tego typu elementy obcojęzyczne stanowią istotny element charakterystyki ich języka osobniczego, dlatego też leksyka taka zostanie włączona do poszczególnych podkorpusów i w specjalny sposób oznaczona[11].

Osobną kategorię stanowią słowa zawarte w brulionach i brudnopisach, słowa skreślone, różne warianty i odmiany tekstów, występujące w tekstach nieukończonych i nieostatecznych, a także pojedyncze słowa niepowiązane składniowo. Tego typu elementy, w wydaniach popularnych pomijane i zwykle niefunkcjonujące w obiegu czytelniczym, nie zawsze były też uwzględniane w edycjach „kanonicznych". Współcześnie często są one umieszczane w wydaniach krytycznych i naukowych opracowaniach – nie tylko takich, które mają charakter edycji źródłowej (także np. w szczegółowych opracowaniach o charakterze edytorskim). Takie elementy językowe, jeśli zostały wiarygodnie opracowane i opublikowane, zostaną uwzględnione w poszczególnych podkorpusach i w specjalny sposób oznaczone.

Dla wszystkich podkorpusów przyjęto też jednolity sposób postępowania wobec tekstów o obcej proweniencji: tłumaczeń, cytatów, parafraz, streszczeń, wypisów

[9] Ze względu na powstający *Internetowy słownik języka Cypriana Norwida* (zob. Puzynina, Korpysz 2009–) obecnie jest to możliwe (tylko w ograniczonym zakresie) jedynie w odniesieniu do tekstów Norwida – zob. Korpysz 2014a, 2014b.

[10] Na temat wykorzystania współczesnych narzędzi leksykograficznych do analiz tekstów dawniejszych zob. np. Derwojedowa, Kieraś, Skowrońska, Wołosz 2014.

[11] Szczegółowe rozwiązania dotyczące tego, co i w jaki sposób będzie oznaczane, zostaną wypracowane w kolejnych etapach prac nad projektem.

itp. Leksyka z takich tekstów zostanie włączona do korpusu i odpowiednio, w jednolity sposób oznaczona.

Niezależnie od spójnych kryteriów wyboru podstaw źródłowych (edycji „kanonicznych" i uzupełniających) istnieją między nimi różnorakie odmienności wynikające z zasad przyjętych przez poszczególnych edytorów. Zakładana i oczekiwana porównywalność wyników obliguje do zastosowania takiej metodologii, która w rezultatach poszukiwań zniweluje różnice zapisu wynikające z autografu lub reguł modernizacji zastosowanych przez poszczególnych edytorów. Dotyczy to przede wszystkim uwzględniania lub pomijania realnych wariantów fonetycznych (niekiedy o charakterze regionalizmów) oraz zapisów autorskich będących wynikiem pośpiechu czy niedbałości (np. problem oznaczania ścieśnionych samogłosek, zapisy typu *król – krol*, *kobieta – kobiéta – kobita*, *kulbakę – kulbake – kulbakie*, *mieszać – mięszać*, *pędzel – pęzel*, *posełać – posyłać*, *tłomaczyć – tłumaczyć*, *wziąć – wziąść* itp.). Tego typu warianty autorskie oraz odmienności wynikające z różniących się pomiędzy edycjami zasad modernizacji będą sprowadzone do porównywalnych form podstawowych, a ewentualne oboczności pisowniane będą sygnalizowane dzięki specjalnym oznaczeniom.

Autorskie warianty ortograficzne i zapisy niezgodne ze współczesną normą w tym zakresie (np. *xsiążę*, *xsiądz*), uwzględniane w wydaniach źródłowych, zostaną sprowadzone do porównywalnych współczesnych form podstawowych, a ewentualne oboczności pisowni będą sygnalizowane dzięki specjalnym oznaczeniom. Takie ujednolicenie pozwoli uniknąć sztucznego multiplikowania liczby leksemów i umożliwi realne porównywanie wyników z poszczególnych podkorpusów.

Najistotniejszym elementem pozwalającym osiągnąć porównywalność wyników jest wprowadzenie we wszystkich podkorpusach jednolitego systemu wielopoziomowych metadanych i anotacji automatycznych i półautomatycznych, pozwalających na oznaczenie, a następnie wyselekcjonowanie z korpusu różnorakich danych[12]. Będą one dotyczyły zarówno np. chronologii powstania danego tekstu, jego istniejącej podstawy (rękopis, pierwodruk, odpis, autoryzowany zapis itp.), jego literackości bądź przynależności do nurtu dokumentu osobistego (notatki, wypisy z lektur itp.), aspektów genologicznych (tekst poetycki a tekst prozatorski, wiersz, poemat, dramat, list, notatka itp.), jak i wybranych informacji językoznawczych (ich zakres nie jest jeszcze na tym etapie prac sprecyzowany). Taki spójny system anotacji umożliwi nie tylko pokazanie podobieństw, lecz także wydobycie różnic między poszczególnymi podkorpusami, a w konsekwencji – idiolektami wybranych twórców. Dotychczasowy brak w pełni wiarygodnych badań porównawczych wynikał m.in. z faktu, że nieliczne istniejące źródła o charakterze językoznawczym i leksykograficznym,

[12] Szczegółowy zestaw i forma metadanych, a także sposoby ich wprowadzania i przyszła dostępność dla użytkownika zostaną ustalone w dalszym toku prac całego zespołu.

jak np. *Słownik języka Adama Mickiewicza* (zob. Górski, Hrabec red. 1962–1983) i powstający *Internetowy słownik języka Cypriana Norwida* (zob. Puzynina, Korpysz 2009–) były tworzone według różnych zasad (szerzej zob. Korpysz 2010).

Cyfrowy „Korpus Czterech Wieszczów", oparty na tekstach autorów mających zasadnicze znaczenie dla kulturowej tożsamości Polaków oraz znaczący wpływ na kształt polszczyzny, opracowany zgodnie z opisanymi wyżej procedurami, zostanie umieszczony w Internecie w wolnym dostępie i udostępniony badaczom oraz wszystkim zainteresowanym użytkownikom.

Bibliografia

Bajerowa, I. 1986. *Polski język ogólny XIX wieku. Stan i ewolucja*, t. 1: *Ortografia, fonologia z fonetyką, morfonologia*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Bajerowa, I. 1992. *Polski język ogólny XIX wieku. Stan i ewolucja*, t. 2: *Fleksja*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Bajerowa, I. 2000. *Polski język ogólny XIX wieku. Stan i ewolucja*, t. 3: *Składnia, synteza*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Bobrowski, J. 2015. Badania nad językiem pisarzy w Polsce po 1989 roku – zarys problematyki. *Język Polski* 1–2, s. 145–153.

Bronikowska, R., Przyborowska-Szulc, A. 2018. *Elektroniczny korpus tekstów XVII i XVIII wieku (do 1772 roku)*. W: *Historia języka w XXI wieku. Stan i perspektywy*, red. M. Pastuch, M. Siuciak, s. 129–135. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Chlebowski, P. 2021. „Lista jednego, ni ząbeczka w liściu". Norwid – poza romantyzmem. *Studia Norwidiana* 39: tom specjalny w 200. rocznicę urodzin poety, s. 127–168.

Dubisz, S. 2012. *Rozwój polszczyzny w XIX wieku*. W: tenże, *Język – Historia – Kultura (wykłady, rozprawy, rozważania)*, t. 3. Warszawa: Wydawnictwo Wydziału Polonistyki UW.

Derwojedowa, M. 2020. Mikrokorpus gronowy polszczyzny 1830–1918. *Poradnik Językowy* z. 8, s. 52–65.

Derwojedowa, M., Kieraś, W., Skowrońska, D. 2014. Korpus polszczyzny XIX wieku – od mikrokorpusu do korpusu średniej wielkości. *Prace Filologiczne* 65, s. 249–254.

Derwojedowa, M., Kieraś, W., Skowrońska, D., Wołosz, R. 2014. Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych. *Polonica* 34, s. 21–28.

Gosiewska, Z. 1949. Z historii badań nad językiem i stylem autorów (szkic informacyjny). *Poradnik Językowy* z. 4, s. 16–23.

Górski, R.L. 2003. *Korpus współczesnego języka polskiego IJP PAN, tzw. korpus krakowski*. W: *Językoznawstwo w Polsce. Stan i perspektywy*, red. S. Gajda, s. 158–161. Opole: Wydawnictwo Uniwersytetu Opolskiego.

Górski, K., Hrabec, S. red. 1962–1983. *Słownik języka Adama Mickiewicza*, t. 1–11. Wrocław: Zakład Narodowy im. Ossolińskich.

Hebal-Jezierska, M. red. 2014. *Praktyczny przewodnik po korpusach języków słowiańskich*. Warszawa: Wydawnictwo Wydziału Polonistyki UW.

Kalinowska, M., Makowska, U., Przychodniak, Z., Troszyński, M., Kaja D. oprac. 2019. *Raptularz wschodni Juliusza Słowackiego*, *Edycja – studia – komentarze*, t. 1–3. Warszawa: Wydawnictwo DiG.

Klapper, M., Kołodziej, D. 2014. Elektroniczny korpus tekstów staropolskich do 1500 r. Perspektywy i problemy. *Prace Filologiczne* 65, s. 205–212.

Koneczna, H., Doroszewski, W. red. 1965–1973. *Słownik języka Jana Chryzostoma Paska*, t. 1–2. Wrocław: Zakład Narodowy im. Ossolińskich.

Korpysz, T. 2010. Słowniki języka autorów jako typ opracowań leksykograficznych. *Poradnik Językowy* z. 4. s. 51–71.

Korpysz, T. 2014a. *Język autora w sieci. O „Internetowym słowniku języka Cypriana Norwida"*. W: *Znaczenie. Tekst. Kultura. Prace ofiarowane Profesor Elżbiecie Janus*, red. A. Kozłowska, A. Świątek, s. 71–83. Warszawa: Wydawnictwo Naukowe UKSW.

Korpysz, T. 2014b. *O „Internetowym słowniku języka Cypriana Norwida" i możliwościach wykorzystania go w badaniach (nie tylko) norwidologicznych*. W: *Tekst artystyczny w badaniach lingwistycznych*, red. E. Skorupska-Raczyńska, M. Maczel, J. Żurawska-Chaszczewska, s. 45–64. Gorzów Wielkopolski: Wydawnictwo Akademia im. Jakuba z Paradyża.

Korpysz, T., Kozłowska, A. [w druku]. *Program badawczy idiolektologii bielańskiej*.

Kozłowska, A. 2013. Indeks leksyki pisarza jako typ opracowania leksykograficznego. *Prace Filologiczne* 64, cz. 1, s. 147–158.

Kozłowska, A. 2018. *Polskie badania nad językiem pisarzy*. W: *Z polskich studiów slawistycznych. Językoznawstwo. Prace na XVI Międzynarodowy Kongres Slawistów w Belgradzie*, seria 13, t. 2, red. Z. Greń, s. 145–154. Poznań: Wydawnictwo Naukowe UAM.

Krasiński, Z. 1963. *Listy do ojca*, wstęp i oprac. S. Pigoń. Warszawa: Państwowy Instytut Wydawniczy.

Krasiński, Z. 1965. *Listy do Jerzego Lubomirskiego*, wstęp i oprac. Z. Sudolski. Warszawa: Państwowy Instytut Wydawniczy.

Krasiński, Z. 1970. *Listy do Adama Sołtana*, wstęp i oprac. Z. Sudolski. Warszawa: Państwowy Instytut Wydawniczy.

Krasiński, Z. 1971. *Listy do Konstantego Gaszyńskiego*, wstęp i oprac. Z. Sudolski. Warszawa: Państwowy Instytut Wydawniczy.

Krasiński, Z. 1975. *Listy do Delfiny Potockiej*, t. 1–3, wstęp i oprac. Z. Sudolski. Warszawa: Państwowy Instytut Wydawniczy.

Krasiński, Z. 1977. *Listy do Koźmianów*, wstęp i oprac. Z. Sudolski. Warszawa: Państwowy Instytut Wydawniczy.

Krasiński, Z. 1979. *Listy do Stanisława Małachowskiego*, wstęp i oprac. Z. Sudolski. Warszawa: Państwowy Instytut Wydawniczy.

Krasiński, Z. 1980. *Listy do Henryka Reeve*, t. 1–2, wstęp i oprac. P. Hertz, tłum. A. Olędzka-Frybesowa. Warszawa: Państwowy Instytut Wydawniczy.

Krasiński, Z. 1980. *Listy do A. Cieszkowskiego, E. Jaroszyńskiego i B. Trentowskiego*, t. 1–2, wstęp i oprac. Z. Sudolski. Warszawa: Państwowy Instytut Wydawniczy.

Krasiński, Z. 1991a. *Listy do plenipotenta i oficjalistów*, wstęp i oprac. Z. Sudolski. Warszawa: Państwowy Instytut Wydawniczy.

Krasiński, Z. 1991b. *Listy do różnych adresatów*, t. 1–2, wstęp i oprac. Z. Sudolski. Warszawa: Państwowy Instytut Wydawniczy.

Krasiński, Z. 2017. *Dzieła zebrane. Nowe wydanie*, t. 1–8. Toruń: Wydawnictwo Naukowe UMK.

Kucała, M. red. 1994–2012. *Słownik polszczyzny Jana Kochanowskiego*, t. 1–5. Kraków: Instytut Języka Polskiego PAN.

Kwapień, E. 2010. *Kształtowanie się zasobu leksykalnego polszczyzny XIX wieku – rzeczowniki (na podstawie danych leksykograficznych)*. Warszawa: Wydawnictwo Wydziału Polonistyki UW.

Kwapień, E. 2014. „Leksykon polszczyzny XIX wieku" – potrzebny, pożyteczny, realny*?*. *Prace Filologiczne* 65, s. 257–268.

Majdak, M. 2018. *Elektroniczny słownik języka polskiego XVII i XVIII wieku IJP PAN*. W: *Historia języka w XXI wieku. Stan i perspektywy*, red. M. Pastuch, M. Siuciak, s. 176–182. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Mickiewicz, A. 1969–. *Dzieła wszystkie*, red. K. Górski. Wrocław: Zakład Narodowy im. Ossolińskich.

Mickiewicz, A. 1993–2005. *Dzieła*, red. Z.J. Nowak, M. Prussak, Z. Stefanowska, Cz. Zgorzelski, t. 1–17. Warszawa: Spółdzielnia Wydawnicza „Czytelnik".

Norwid, C. 1966. *Dzieła zebrane*, oprac. J.W. Gomulicki, t. 1: wiersze, tekst, tom 2: wiersze, dodatek krytyczny. Warszawa: Państwowy Instytut Wydawniczy.

Norwid, C. 1971–1976. *Pisma wszystkie*, zebrał, tekst ustalił, wstępem i uwagami krytycznymi opatrzył J.W. Gomulicki, t. I–XI. Warszawa: Państwowy Instytut Wydawniczy.

Norwid, C. 2009–. *Dzieła wszystkie*, red. naczelny S. Sawicki, t. 3: *Poematy* cz. I, oprac. S. Sawicki, A. Cedro. Lublin: Towarzystwo Naukowe KUL 2009; t. 4: *Poematy* cz. II, oprac. S. Sawicki, P. Chlebowski. Lublin 2011; t. 5: *Dramaty* cz. I, oprac. J. Maślanka Lublin 2015; t. 6: *Dramaty* cz. II, oprac. J. Maślanka. Lublin 2013; t. 7: *Proza* cz. I, oprac. R. Skręt. Lublin 2007; t. 10: *Listy* cz. I, oprac. J. Rudnicka. Lublin 2008; t. 11: *Listy* cz. II, oprac. J. Rudnicka. Lublin 2016; t. 12: *Listy* cz. III, oprac. J. Rudnicka, uzup. E. Lijewska. Lublin 2019.

Puzynina, J., Korpysz, T. 2009–. *Internetowy słownik języka Cypriana Norwida*. Online: www.slownikjezykanorwida.uw.edu.pl

Sawrymowicz, E. oprac. 1962–1963. *Korespondencja Juliusza Słowackiego*, t. 1–2. Wrocław: Zakład Narodowy im. Ossolińskich.

Siwiec, M. 2021. *Sytuacja Norwida, sytuacja Baudelaire'a. Paralele nowoczesności*. Kraków: Towarzystwo Autorów i Wydawców Prac Naukowych „Universitas".

Sławkowa, E. 2009. *O różnych sposobach językoznawczej refleksji nad językiem artystycznym*. W: *Język pisarzy jako problem lingwistyki*, red. T. Korpysz, A. Kozłowska, s. 25–44. Warszawa: Wydawnictwo Naukowe UKSW.

Sławkowa, E. 2011. *Kierunki badań nad słownictwem pisarzy*. W: *Język pisarzy: problemy słownictwa*, red. T. Korpysz, A. Kozłowska, s. 13–28. Warszawa: Wydawnictwo Naukowe UKSW.

Słowacki, J. 1952–1975. *Dzieła wszystkie*, red. J. Kleiner przy współudziale W. Floryana, t. 1–17. Wrocław: Zakład Narodowy im. Ossolińskich.

Słowacki, J. 1959. *Dzieła*, red. J. Krzyżanowski, t. 1–14, wyd. 3. Wrocław: Zakład Narodowy im. Ossolińskich.

Słowacki, J. 1996. *Raptularz 1843–1849*, oprac. M. Troszyński. Warszawa: Wydawnictwo „Topos".

Sobolewska, K. 2010. Leksykograf jako partner historyka literatury. *Prace Filologiczne* 58, s. 379–390.

Śliwiński, W. red. 2016. *Słownik osobliwości leksykalnych Stanisława Wyspiańskiego na materiale utworów dramatycznych*. Warszawa: Wydawnictwo Libron.

Świdziński, M. 2006. Lingwistyka korpusowa w Polsce – źródła, stan, perspektywy. *LingVaria* 1, s. 23–32.

Twardzik, W. 2003. *Korpus staropolski Instytutu Języka Polskiego PAN w Krakowie*. W: *Językoznawstwo w Polsce. Stan i perspektywy*, red. S. Gajda, s. 155–157. Opole: Wydawnictwo Uniwersytetu Opolskiego.

Wojtyńska-Nowotka, M. 2020. *Słownik języka Maurycego Mochnackiego (na podstawie „Rozpraw literackich")*. Warszawa: Dom Wydawniczy „Elipsa".

"Corpus of the Four Poets": the digital dimension of the national heritage. Project assumptions

Summary

This paper presents the preliminary assumptions of the "Corpus of the Four Poets" project, carried out in cooperation between Polish philologists (IBL PAN, UKSW) and IT specialists (CLARIN-PL, Wrocław University of Technology). The aim of the project is to create a digital corpus containing all Polish-language texts by the authors of the so-called "Great Romanticism": Adam Mickiewicz, Juliusz Słowacki, Zygmunt Krasiński, and Cyprian Norwid, as well as to develop IT tools integrated with the corpus. The final result will be an online open access database with software. The texts for the corpus have been selected from the print sources that are well-known to readers, which provides the most complete and accurate form of the texts and a relative completeness of the lexical base. The set of metadata and annotations developed in the subsequent stages of the project will enable retrieval of the data on the language of individual authors and their objectivised comparison.

Keywords: Zygmunt Krasiński– Adam Mickiewicz – Cyprian Norwid – Juliusz Słowacki – poet–prophet – corpus – vocabulary – idiolect.

Adj. Monika Czarnecka