Idiolekty

**2 (2025)** • https://doi.org/10.21697/idiolekty.2.15 • strony 231–235

**Tomasz Korpysz**

Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie

ORCID 0000-0001-6578-5839

**Anna Mędrzecka-Stefańska**

Instytut Badań Literackich PAN

ORCID 0000-0003-1165-5793

# O projekcie „Korpus Czterech Wieszczów”

„Korpus Czterech Wieszczów” to projekt realizowany w ramach CLARIN-PL przez zespół pod kierownictwem prof. Marka Troszyńskiego[1]. Celem tego projektu jest stworzenie kompletnego, ogólnodostępnego, dającego się łatwo i wielostronnie przeszukiwać anotowanego korpusu[2] polskojęzycznych tekstów autorstwa Adama Mickiewicza, Juliusza Słowackiego, Zygmunta Krasińskiego i Cypriana Norwida[3].

[1] Punktem wyjścia projektu były prace nad korpusem dzieł Juliusza Słowackiego, prowadzone w CHC IBL PAN pod kierownictwem prof. Marka Troszyńskiego, kontynuowane i poszerzane następnie dzięki współpracy z CLARIN-PL i osobistemu zaangażowaniu prof. Macieja Piaseckiego. Opracowując założenia korpusu, wykorzystano też wcześniejsze doświadczenia zespołu w zakresie szeroko rozumianej humanistyki cyfrowej, np. prace nad edycją cyfrową *Samuela Zborowskiego*, która ma być podstawą do przygotowania kompletnej edycji cyfrowej dzieł Juliusza Słowackiego (Marek Troszyński, Ewa Mirkowska i Anna Mędrzecka-Stefańska), przeprowadzoną metodami korpusowymi analizę kategorii ironii w poematach Juliusza Słowackiego (Anna Mędrzecka-Stefańska) czy opracowywanie *Internetowego słownika języka Cypriana Norwida* (Tomasz Korpysz). Głównymi wykonawcami projektu są (w kolejności alfabetycznej): Tomasz Korpysz, Anna Mędrzecka-Stefańska, Ewa Mirkowska i Marek Troszyński, których wspierają badacze z Politechniki Wrocławskiej – przede wszystkim dr Marcin Oleksy i dr Tomasz Bernaś.

[2] Planowany efekt końcowy projektu nie będzie zatem tylko korpusem: wszystkie zgromadzone w nim teksty mają zostać w sposób jednolity opracowane i opatrzone różnego rodzaju anotacjami: tekstologicznymi, gramatycznymi i genologicznymi. Szerzej zob. Mędrzecka-Stefańska 2023.

[3] Na temat wstępnych założeń projektu zob. Korpysz, Mędrzecka, Mirkowska, Troszyński 2022.

Koncepcja korpusu zrodziła się w wyniku obserwacji, iż w erze powszechnej cyfryzacji oraz dostępności coraz większej liczby różnego typu źródeł w wersji cyfrowej dorobek czwórki polskich poetów, których dzieło i myśl mają kluczowe znaczenie dla polskiej kultury, jest rażąco nieobecny w przestrzeni internetowej: nie ma w niej dostępnego wiarygodnego kompletu tekstów żadnego z przywoływanych autorów. Niektóre ich dzieła można wprawdzie znaleźć w ogólnie dostępnych repozytoriach czy na rozmaitych portalach i stronach internetowych, jednak zwykle oparte są one na dawnych wydaniach (niepozbawionych różnego rodzaju błędów i niezgodnych ze współczesnymi standardami edytorskimi), których wykorzystywanie nie tylko w pracy badawczej, lecz także w zwykłej lekturze jest utrudnione. Ponadto najczęściej są one trudno przeszukiwalne i nieprzystosowane do wymagań nowoczesnego językoznawstwa czy literaturoznawstwa cyfrowego.

Tworzony korpus ma zawierać opracowane w jednolity sposób cztery podkorpusy, z których każdy obejmować będzie całą polskojęzyczną spuściznę danego twórcy, a więc nie tylko teksty tradycyjnie klasyfikowane jako literackie, lecz także korespondencję oraz różnego rodzaju teksty dyskursywne, użytkowe, szkice, notatki itp. Konsekwencją takiej decyzji okazało się to, że w przypadku żadnego autora nie można było oprzeć się wyłącznie na jednym wydaniu jako podstawie, należało wybrać teksty lub bloki tekstów z różnych edycji uznanych przez zespół za najbardziej odpowiednie do osiągnięcia założonego celu[4]. Tak wybrane podstawy zostały zdygitalizowane, a następnie pliki graficzne przekonwertowano do plików tekstowych i poddano je wstępnej korekcie (przede wszystkim usunięto wszystkie elementy pochodzące od wydawców). Kolejnym etapem będzie dokładna korekta, a następnie dalsze przetwarzanie tekstów z użyciem narzędzi komputerowych, przeprowadzone z wykorzystaniem serwisu LEM (zob. Piasecki, Walkowiak, Maryl 2017): lematyzacja oraz tagowanie morfosyntaktyczne, w którego efekcie każdemu słowu zostanie przypisany zestaw tagów opisujących jego formę gramatyczną[5].

Ze względu na dążenie do spójności korpusu i porównywalności wyników materiał w podkorpusach zostanie poddany modernizacji (przede wszystkim w zakresie pisowni) i różnego rodzaju ujednoliceniom (np. sprowadzanie do jednej postaci odmiennych zapisów danego słowa pochodzących z różnych edycji podstaw

[4] Szerzej na temat kryteriów doboru podstaw tekstowych zob. Korpysz, Mędrzecka, Mirkowska, Troszyński 2022: 70–73.

[5] Najważniejsze etapy przetwarzania tekstów z użyciem narzędzi do przetwarzania języka naturalnego (Natural Language Processing, NLP) przedstawiono w: Mędrzecka-Stefańska 2023 (wersja 1.), 2024 (wersja 2.).

tekstowych). Zabiegi te stosowane będą oszczędnie, aby nie zagubić indywidualnych cech różniących idiolekty badanych autorów. Przygotowane w ten sposób teksty zostaną umieszczone w systemie Inforex (zob. Marcińczuk, Oleksy, Kocoń 2017), który pozwala na wprowadzenie bogatego zestawu metadanych i anotacji. Obejmuje on np. typ tekstu (poezja, proza, wiersz, dramat itp.), jego status (tekst autorski, cytat, tłumaczenie itp.), datę powstania, lokalizację, wydobywa nazwy własne czy np. didaskalia. Dodatkowo specjalna warstwa anotacji pozwoli zlokalizować wszystkie zmiany wprowadzone przez badaczy i w razie potrzeby zrekonstruować kształt językowy podstawy. Dzięki tego typu anotacjom i metadanym w przyszłości możliwe będzie zaawansowane przeszukiwanie korpusu, pozwalające na wyszukiwanie określonych form i leksemów oraz filtrowanie wyników wyszukiwania do jednostek spełniających określone kryteria (chronologiczne, genologiczne, gramatyczne itp.).

Projekt „Korpusu Czterech Wieszczów" jest obecnie w pierwszym etapie realizacji. Jak już wspomniano wyżej, wybrano podstawy źródłowe, przygotowano pliki tekstowe obejmujące całość spuścizny poszczególnych twórców, wypracowano też spójne zasady modernizacji, tagowania, anotowania oraz strukturyzacji tekstów. Całościowemu opracowaniu poddano jak dotąd wszystkie wiersze, a także wybrane poematy oraz niewielkie próbki innych tekstów: korespondencji, utworów dramatycznych i fragmentów prozy – ich analiza pozwoliła zweryfikować niektóre wstępne założenia dotyczące np. strukturyzacji tekstów, która w wierszach nie stanowi tak istotnego problemu jak choćby w utworach dramatycznych.

Podstawowe statystyki dotyczące podkorpusów wierszy przedstawiono w tabeli 1. Zawiera ona dane dotyczące podkorpusów wierszy poszczególnych autorów: Adama Mickiewicza, Cypriana Norwida, Juliusza Słowackiego i Zygmunta Krasińskiego. Dane obejmują: liczbę dokumentów, liczbę słów, liczbę znaków i liczbę tokenów. Największą liczbę dokumentów zawiera podkorpus tekstów Mickiewicza, natomiast największą liczbę słów, znaków i tokenów – podkorpus Norwida.

**Tab. 1**

| Podkorpus (autor) | Liczba dokumentów | Liczba słów | Liczba znaków | Liczba tokenów |
|---|---|---|---|---|
| Adam Mickiewicz | 365 | 66 665 | 310 797 | 68 485 |
| Cyprian Norwid | 314 | 72 011 | 331 386 | 78 004 |
| Juliusz Słowacki | 255 | 48 065 | 218 762 | 49 764 |
| Zygmunt Krasiński | 142 | 36 882 | 158 590 | 36 586 |
| **Łącznie** | **1076** | **223 623** | **1 019 535** | **232 839** |

Obecnie trwają prace nad upublicznieniem przygotowanych materiałów – przede wszystkim tych dotyczących wierszy. Należy podkreślić, że będą one ogólnodostępne w formatach zintegrowanych z infrastrukturą CLARIN-PL, aby wszyscy zainteresowani mogli korzystać z nich za pomocą przystępnego interfejsu, bez konieczności samodzielnego przygotowania odpowiednich narzędzi. Badacze zainteresowani wykorzystaniem tych zasobów w innych programach będą mogli pobrać wybrane przez siebie pliki korpusu do dalszego przetwarzania.

Na koniec wypada wyrazić nadzieję, że prace nad „Korpusem Czterech Wieszczów" będą mogły być kontynuowane, dzięki czemu w przyszłości użytkownicy zyskają łatwy dostęp do rzetelnie przygotowanej cyfrowej bazy obejmującej wszystkie polskojęzyczne teksty najwybitniejszych twórców polskiego romantyzmu, co niewątpliwie pozwoli na ich bardzo różnorodne i wielopoziomowe analizy.

## Bibliografia

**Marcińczuk** M., **Oleksy** M., **Kocoń** J., 2017, *Inforex – A collaborative system for text corpora annotation and analysis*, w: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, s. 473–482.

**Mędrzecka-Stefańska** A., 2013, *Między korpusem a edycją cyfrową – na przykładzie projektu „Korpusu Czterech Wieszczów"*, „Sztuka Edycji" 1: *Praktyki edycji cyfrowych*, s. 113–120.

**Mędrzecka-Stefańska** A., 2023 (wersja 1.), 2024 (wersja 2.), *Ilościowa analiza tekstów. NLP*, w: *Panorama literaturoznawstwa cyfrowego*, red. M. Maryl, B. Szleszyński, T. Umerle, M. Błaszczyńska, Warszawa, nplp.pl/panorama-literaturoznawstwa-cyfrowego/ilosciowa-analiza-tekstow-nlp/ (digital born, DOI: 10.18318/978-83-67957-03-8.2.2).

**Korpysz** T., **Mędrzecka** A., **Mirkowska** E., **Troszyński** M., 2022, *„Korpus czterech wieszczów" – cyfrowy wymiar dziedzictwa narodowego. Założenia projektu*, „Poradnik Językowy" 7, s. 67–78.

**Piasecki** M., **Walkowiak** T., **Maryl** M., 2017, *Literary Exploration Machine: A Web-Based Application for Textual Scholars*, https://ws.clarin-pl.eu/lem#.

**Puzynina** J., **Korpysz** T., *Internetowy słownik języka Cypriana Norwida*, współpraca merytoryczna J. Chojak, współpraca techniczna J. Miernik, M. Żółtak, http://www.slownikjezykanorwida.uw.edu.pl.

## Abstract

## On the “Corpus of the Four Bards” project

This article outlines the main objectives and the current stage of work on the “Corpus of the Four Bards” project, carried out within the CLARIN-PL infrastructure by a team led by Professor Marek Troszyński. The aim of the project is to create a comprehensive, openly accessible, annotated corpus of Polish-language texts authored by the four most eminent representatives of Polish Romanticism: Adam Mickiewicz, Juliusz Słowacki, Zygmunt Krasiński, and Cyprian Norwid. The corpus will be fully searchable in multiple ways and will consist of four sub-corpora. These will be made available in formats integrated with the CLARIN-PL infrastructure, enabling all users to access and explore the corpus through a user-friendly interface, without the need to develop or configure specialised tools independently.

**Keywords:** Adam Mickiewicz, Juliusz Słowacki, Zygmunt Krasiński, Cyprian Norwid, “Corpus of the Four Bards”, CLARIN-PL, idiolect