Korpus tekstowy jako narzędzie literaturoznawcze

Agnieszka Karlińska, Paulina Czwordon-Lis, Maciej Maryl

TEKSTY DRUGIE 2023, NR 6, S. 294–319

DOI: 10.18318/td.2023.6.16 | ORCID: Agnieszka Karlińska: 0000-0002-4846-7086
Paulina Czwordon-Lis: 0000-0001-5136-4590
Maciej Maryl: 0000-0002-2639-041X

Praca częściowo finansowana z projektu *Cyfrowa infrastruktura badawcza dla humanistyki i nauk o sztuce* DARIAH-PL, Program Operacyjny Inteligentny Rozwój 2014-2020 #POIR.04.02.00-00-D006/20.

Wprowadzenie

W książce *Distant Horizons: Digital Evidence and Literary Change* Ted Underwood[1] stawia pytania o proces historycznoliteracki w perspektywie długiego trwania. Jego projektowi przyświeca cel nad wyraz ambitny, jakim jest dokonanie nowego otwarcia w cyfrowych badaniach literackich (Digital Literary Studies), które ma polegać na zaproponowaniu przejrzystej, zrozumiałej i powtarzalnej metodologii badań opartych na dobrze określonych i dostępnych publicznie materiałach, a przy tym odpowiadających na pytania istotne z punktu widzenia dyscypliny. W centrum tego projektu pozostaje nowe spojrzenie na diachronię.

Zgodnie z główną tezą Underwooda analiza procesu historycznoliterackiego wymaga spojrzenia o charakterze ewolucyjnym, wykraczającego poza jedną epokę (czy serię obserwacji). Wiedza historycznoliteracka, powiada Underwood,

1 T. Underwood, *Distant Horizons: Digital Evidence and Literary Change*, Chicago University Press, Chicago 2019.

Agnieszka Karlińska – mgr, Zakład Inżynierii Lingwistycznej i Analizy Tekstu, NASK PIB. Kontakt: agnieszka.karlinska@nask.pl.

Paulina Czwordon-Lis – dr, Pracownia Bibliografii Bieżącej IBL PAN. Kontakt: paulina.czwordon-lis@ibl.waw.pl.

Maciej Maryl – dr, adiunkt, kierownik Centrum Humanistyki Cyfrowej IBL PAN, WWW: maryl.org. Kontakt: Maciej.Maryl@ibl.waw.pl.

> jest niezwykle bogata w odniesieniu do pewnych ram czasowych. Dobrze nam idzie wykorzystywanie fragmentarycznych wydarzeń do charakteryzowania autorów, kierunków czy okresów. Gdybyśmy jednak próbowali połączyć te obrazki ze sobą, by pokazać szersze ramy czasowe, konsensus stanie się trudny[2].

W kolejnych rozdziałach badacz zbiera argumenty na rzecz przejścia od dyskretnego (tj. pokawałkowanego) pojmowania historii literatury do patrzenia na procesy w perspektywie ciągłej, wykraczającej poza ramy okresów i epok[3].

Dobrym przykładem zastosowania takiego podejścia są badania Ryana Heusera i Longa Le-Khaca[4], które Underwood omawia w pierwszym rozdziale książki[5]. Na przykładzie analizy słownictwa tekstów z trzech ostatnich stuleci Heuser i Le-Khac przedstawiają stopniowe odróżnianie się dyskursu autobiograficznego od powieściowego jako przejście od „mówienia do pokazywania". Świadczyć ma o tym coraz częstsze sięganie po określone kategorie słów, takie jak nazwy kolorów czy części ciała, czasowniki odnoszące się do aktywności oraz opis fizyczny. Korpus tekstów obejmujący duży zakres czasowy pozwala śledzić rozwój języka literackiego – a na tej podstawie ewolucję gatunków – w perspektywie ciągłej, wykraczającej poza ramy kolejnych epok. I tu wracamy do głównej tezy Underwooda – na takie spojrzenie pozwalają metody ilościowe, gdyż perspektywa indywidualna siłą rzeczy jest bardziej ograniczona zakresem materiału możliwym do opanowania przez jednostkę i przez to przywiązana do periodyzacji.

Przykład Underwooda pokazuje ciekawą zależność pomiędzy pytaniami badawczymi, metodami i materiałem. Badanie w historii literatury procesów długotrwałych przy użyciu metod zapożyczonych z lingwistyki komputerowej wymaga solidnej podstawy materiałowej do prowadzenia badań – czyli korpusów.

---

2 Tamże, s. 8.

3 Por. omówienie w: M. Maryl, *Computational Monograph: Reading and Writing Distant Horizons*, „Journal of Literary Theory online" 2020, nr 14 (2), http://www.jltonline.de/index.php/reviews/article/view/1090 (7.06.2022).

4 R. Heuser, L. Le-Khac, *A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method*, Literary Lab, Stanford University, Stanford 2012, http://litlab.stanford.edu/LiteraryLabPamphlet4.pdf (7.06.2022).

5 T. Underwood, *Distant Horizons*, s. 12–16.

Korpus to zbiór tekstów pozwalający zastosować metody ilościowe i na ich podstawie wyciągnąć wnioski na temat szerszej populacji, której ów korpus stanowi próbę. Hewitt i inni posługują się metaforą matrioszki, by pokazać korpus jako wytwór produkcji literackiej zawierający się w zbiorze tekstów dostępnych, nazywanych przez nich archiwum (*archive*), które to archiwum z kolei zawiera się w zbiorze wszystkich tekstów opublikowanych, czyli właśnie populacji[6]. Istotne jest też odróżnienie pojęcia kanonu, które zakresowo mogłoby się pokrywać korpusem, choć zwykle obejmuje raczej teksty ważne z perspektywy kolejnych pokoleń, a nie reprezentatywne dla danego okresu, co może utrudnić wyciąganie wniosków o szerokiej populacji[7]. Dobrze to widać na przykładzie tak zwanego problemu wielkich nieczytanych (*the great unread*), czyli ogromnego zasobu produkcji literackiej, która ze względu na niską istotność czy wtórność wobec prac kanonicznych pozostaje poza zakresem zainteresowania literaturoznawstwa.

Cyfrowe literaturoznawstwo potrzebuje korpusów, by wykorzystać wypracowane już metody i poszerzyć nasze rozumienie takich procesów, jak ewolucja gatunków, pojęć czy stylów. Konstrukcja takiego zbioru wymaga jednak precyzyjnej metodologii i rozstrzygnięcia szeregu kwestii, które mają wpływ na charakter wniosków, jakie można wyciągać na podstawie zebranego w ten sposób materiału. Dlatego nasze rozważania rozpoczynamy od ogólnej prezentacji korpusów jako zbiorów danych badawczych w różnych dyscyplinach, o dość dobrze określonej metodologii ich tworzenia. Następnie skupiamy się na literaturoznawstwie, przyglądając się swoistym protokorpusom, czyli antologiom, które zdawały się pełnić funkcję podobną, choć w innym kontekście metodologicznym. Te rozpoznania prowadzą nas do przeglądu i analizy wybranych korpusów w badaniach literackich. Prezentowane tu rozważania wieńczą etap przygotowawczy prac nad Korpusem Dyskursu Literaturoznawczego (KDL), którego założenia – wraz z opisem wyzwań, przed jakimi stanął zespół – prezentujemy w konkluzjach.

6 M. Algee-Hewitt, S. Allison, M. Gemma, R. Heuser, F. Moretti, H. Walser, *Canon/Archive: Large-Scale Dynamics in the Literary Field*, 2016, http://litlab.stanford.edu/LiteraryLabPamphlet11.pdf (7.06.2022).

7 Por. M. Algee-Hewitt, M. McGurl, *Between Canon and Corpus: Six Perspectives on 20th-Century Novels. Pamphlets of the Stanford Literary Lab, Pamphlet 8*, Stanford Literary Lab, Stanford 2015.

Projektowanie korpusów[8]

Korpus jest zazwyczaj definiowany jako możliwie reprezentatywny dla danego języka lub jego odmiany (typu dyskursu) zbiór tekstów zapisanych w formie elektronicznej, które można przetwarzać za pomocą specjalistycznego oprogramowania[9]. Chociaż zdarza się, że obejmuje on cały zbiór (populację) tekstów stanowiących przedmiot badań (np. wszystkie powieści danej autorki), w większości przypadków konieczne jest dokonanie selekcji[10]. Korpus ma zwykle charakter skończony – zakres materiału jest ściśle określony i nie powinno się dodawać do niego nowych tekstów[11]. Wyjątek stanowią tak zwane korpusy monitorujące[12], które mają strukturę otwartą i są stale uzupełniane o nowe próbki, co umożliwia śledzenie zmian i bieżących trendów w języku[13]. W procesie projektowania korpusu kluczową rolę odgrywa z jednej strony jego przeznaczenie, czyli dostosowanie do konkretnych pytań badawczych, z drugiej zaś metodologia doboru danych, czyli zagadnienia wielkości, struktury oraz reprezentatywności[14].

Chociaż analizy korpusowe prowadzono z powodzeniem również na niewielkich zbiorach tekstów, na przykład pojedynczych listach[15], przyjmuje

8 W podrozdziale wykorzystano niepublikowane wcześniej fragmenty rozprawy doktorskiej Agnieszki Karlińskiej *Psychiatria na wokandzie. Strategie dyskursywne w opiniowaniu sądowo-psychiatrycznym*, napisanej pod kierunkiem Mirosławy Marody i Macieja Maryla na Wydziale Socjologii Uniwersytetu Warszawskiego.

9 T. McEnery, A. Wilson, *Corpus Linguistics: An Introduction*, Edinburgh University Press, Edinburgh 2001; T. McEnery, R. Xiao, Y. Tono, *Corpus-Based Language Studies: An Advanced Resource Book* (Routledge Applied Linguistics), Routledge, London–New York 2006.

10 S. Hunston, *Collection Strategies and Design Decisions*, w: *Corpus Linguistics: An International Handbook*, red. A. Lüdeling, M. Kytö, De Gruyter, Berlin 2008, s. 154-168.

11 G. Wiedemann, *Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences*, „Forum Qualitative Sozialforschung / Forum: Qualitative Social Research" 2013, nr 14 (2).

12 J. Sinclair, *Corpus, Concordance, Collocation*, Oxford University Press, Oxford 1991; W. Teubert, *A Province of a Federal Superstate, Ruled by an Unelected Bureaucracy. Keywords of the Euro-sceptic Discourse in Britain*, w: *Attitudes Towards Europe*, red. A. Musolff, C. Good, P. Points, R. Wittlinger, Ashgate, Aldershot 2001.

13 P. Pęzik, *Budowa i zastosowania korpusu monitorującego MoncoPL*, „Forum Lingwistyczne" 2020, nr 7, s. 133.

14 W.N. Francis, *Problems of Assembling and Computerizing Large Corpora*, w: *Corpus Linguistics. Critical Concepts in Linguistics*, red. W. Teubert, R. Krishnamurthy, Routlege, Abington–New York 2007; S. Hunston, *Collection Strategies and Design Decisions*.

15 M. Stubbs, *Text and Corpus Analysis. Computer-Assisted Studies of Language and Culture*, Blackwell, Oxford 1996.

się, że „korpus powinien być wystarczająco duży, aby pokazać częstotliwości występowania pewnych zjawisk językowych, dzięki czemu badacze mogą ustalić, co jest typowe, a co rzadkie w języku"[16]. Jego docelowa objętość zależy od trzech podstawowych czynników: 1) aspektu języka stanowiącego przedmiot badań (zbiór służący do analiz leksykograficznych powinien być na przykład znacznie większy niż zbiór wykorzystywany do badań nad prozodią czy gramatyką), 2) stopnia zróżnicowania badanej odmiany języka (im jest większy, tym większy powinien być korpus) oraz 3) stopnia powtarzalności tekstów w obrębie określonego typu lub gatunku[17].

W literaturze wskazuje się, że korpus powinien być reprezentatywny dla danego języka, jego odmiany lub badanego tematu. Należy w tym miejscu poczynić za Rafałem Górskim[18] istotne zastrzeżenie: korpus reprezentuje nie tyle język jako taki, ile określoną populację tekstów, i to właśnie teksty powinny stanowić punkt odniesienia przy określaniu modelu reprezentatywności: „[Język] jest po części bytem abstrakcyjnym, dyspozycją psychiczną. Korpus nie reprezentuje bezpośrednio kompetencji językowej czy saussurowskiego *langue*. Korpus jest zbiorem tekstów, a więc reprezentuje *parole*"[19].

Douglas Biber[20] wskazuje, że reprezentatywność może być rozpatrywana z perspektywy sytuacyjnej i językowej pod kątem zakresu typów tekstów w danym języku lub jego odmianie oraz pod kątem ich zróżnicowania językowego, czyli dystrybucji określonych cech językowych. Obok reprezentatywności zewnętrznej, rozumianej jako reprezentacja docelowej domeny dyskursu, wyróżnia się więc reprezentatywność wewnętrzną[21], definiowaną zwykle jako stopień, w jakim korpus reprezentuje zakres zmienności języko-

16 P. Baker, *Corpus Linguistics*, w: *Research Methods in Linguistics*, red. L. Litosseleti, Continuum, London 2010, s. 95.

17 G. Kennedy, *An Introduction to Corpus Linguistics*, Longman, London 1998; D. Biber, *Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation*, „Literary and Linguistic Computing" 1990, nr 5 (4); D. Biber, *Representativeness in Corpus Design*, „Literary and Linguistic Computing" 1993, nr 8; B. Lewandowska-Tomaszczyk, *Podstawy językoznawstwa korpusowego*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2005.

18 R.L. Górski, *Metody korpusowe i kwantytatywne w językoznawstwie historycznym*, w: *Metodologie językoznawstwa. Od diachronii do panchronii*, red. P. Stalmaszczyk, Wydawnictwo UŁ, Łódź 2018.

19 Tamże, s. 117.

20 D. Biber, *Representativeness in Corpus Design*.

21 T. McEnery, R. Xiao, Y. Tono, *Corpus-Based Language Studies.*

wej w populacji[22]. Ocena reprezentatywności zewnętrznej wymaga wiedzy o świecie zewnętrznym (populacji tekstów), aby ustalić, w jakim stopniu dokumenty włączone do korpusu oddają pełne spektrum tekstów w badanej domenie dyskursu[23]. W przypadku reprezentatywności wewnętrznej punktem odniesienia jest sam korpus – jednym ze sposobów oceny reprezentatywności językowej jest podzielenie go na podkorpusy i sprawdzenie, czy wyniki analizy ilościowej poszczególnych zbiorów się pokrywają. W praktyce przy ocenie korpusów reprezentatywność wewnętrzną bierze się pod uwagę stosunkowo rzadko: „korpusy są zwykle tworzone w celu reprezentowania określonych odmian języka (reprezentatywność zewnętrzna), a mniej uwagi poświęca się temu, do jakich konkretnych językowych pytań badawczych będą odpowiednie (reprezentatywność wewnętrzna)"[24] .

Drugą kluczową cechą korpusu, bezpośrednio związaną z reprezentatywnością, a niekiedy nawet z nią utożsamianą[25], jest zrównoważenie[26]. Podczas gdy reprezentatywność ująć można jako występowanie wszystkich elementów badanej odmiany języka w korpusie, zrównoważenie jest zwykle definiowane jako zachowanie odpowiednich proporcji między reprezentacją poszczególnych elementów danej odmiany[27] czy też jako „dbałość o taką budowę korpusu, żeby żaden składnik na żadnym z poziomów nie dominował nad innym"[28]. Warto przy tym uwzględnić nie tylko liczbę, lecz także długość tekstów. W celu poprawy zrównoważenia korpusu zaleca się niekiedy próbkowanie poszczególnych publikacji, czyli włączanie do korpusu fragmentów, które stanowią pewne całości (np. rozdziałów), albo wyimków o określonej długości[29].

---

22 D. Biber, *Representativeness in Corpus Design*, s. 243.

23 D. Miller, D. Biber, *Evaluating Reliability in Quantitative Vocabulary Studies: The Influence of Corpus Design and Composition*, „International Journal of Corpus Linguistics" 2015, nr 20 (1).

24 Tamże, s. 35.

25 R.L. Górski, M. Łaziński, *Reprezentatywność Narodowego Korpusu Języka Polskiego*, w: *Narodowy Korpus Języka Polskiego*, red. A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk, Wydawnictwo Naukowe PWN, Warszawa 2012.

26 J. Sinclair, *Corpus to Corpus: A Study of Translation Equivalence*, „International Journal of Lexicography" 1996, nr 9 (3).

27 S.T. Gries, *Quantitative Corpus Linguistics with RA Practical Introduction*, Routledge, New York 2009.

28 R.L. Górski, M. Łaziński, *Reprezentatywność Narodowego Korpusu Języka Polskiego*, s. 26

29 Tamże, s. 35.

Dotyczy to zwłaszcza korpusów mniejszych, w których dłuższe teksty mogą mieć nieproporcjonalnie duży wpływ na uzyskane wyniki[30]. Próbkowanie ma zapewniać jednorodność i porównywalność danych, stwarzać możliwość prowadzenia analiz porównawczych niezależnie od wielkości badanego zbioru[31].

Pierwszym krokiem w procesie budowy korpusu jest zwykle zdefiniowanie populacji, którą korpus ma odzwierciedlać, a więc określenie zbioru możliwych tekstów[32]. Krokiem drugim jest natomiast dobór próby z tak zdefiniowanej populacji. Krokiem pośrednim może być opracowanie operatu losowania (*sampling frame*), czyli uporządkowanej listy elementów danego zbioru (zazwyczaj publikacji z określonego roku lub dekady), z której następnie losuje się określoną liczbę tekstów[33]. Twórcy korpusu Browna dokonali na przykład losowego doboru publikacji na podstawie listy książek i czasopism dostępnych w wybranych bibliotekach[34].

Jak zauważa Jadwiga Sambor[35], w korpusie reprezentatywnym teksty powinny tworzyć zbiorowość jednorodną (warunek homogeniczności próby) przy jednoczesnym zachowaniu pewnej różnorodności, na przykład w ramach rejestrów (warunek heterogeniczności próby). W celu zrównoważenia korpusu należy dokonać podziału przestrzeni, na podstawie której zostaną dobrane teksty, na jednolite klasy i określić ich procentowy udział w korpusie[36]. Korpus taki można potraktować jako zbiór podkorpusów, z których każdy powinien być stosunkowo jednolity[37]. Na przykład w wyniku podziału przestrzeni językowej ze względu na typ tekstu (artykuły naukowe i popularnonaukowe) oraz dziedzinę (humanistyka i nauki społeczne) korpus składałby się z czterech podkorpusów: artykułów naukowych z zakresu humanistyki, artykułów naukowych z zakresu nauk społecznych, artykułów popularno-

30 S. Hunston, *Collection Strategies and Design Decisions*.

31 D. Biber, *Representativeness in Corpus Design*.

32 S. Titscher, M. Meyer, R. Wodak, E. Vetter, *Methods of Text and Discourse Analysis*, Sage, London 2000.

33 D. Biber, *Representativeness in Corpus Design*.

34 E. Hajnicz, *Najbardziej znane korpusy tekstów. Opracowanie przeglądowe*, Prace Instytutu Podstaw Informatyki Polskiej Akademii Nauk 1021, Instytut Podstaw Informatyki PAN, Warszawa 2011.

35 J. Sambor, *Słowa i liczby*, Zakład Narodowy im. Ossolińskich, Wrocław 1972.

36 E. Hajnicz, *Najbardziej znane korpusy tekstów*.

37 S. Hunston, *Collection Strategies and Design Decisions*.

naukowych z zakresu humanistyki oraz artykułów popularnonaukowych z zakresu nauk społecznych. Znaczna część dyskusji na temat projektowania korpusów skupia się na tym, jak określić kryteria podziału oraz jakie powinny być relacje pomiędzy poszczególnymi podkorpusami[38].

Za Górskim[39] można wyróżnić kilka podejść do projektowania reprezentatywnego i zrównoważonego korpusu. Jednym z nich jest określenie typów tekstów, jakie mają znaleźć się w korpusie, i dobranie dla każdego z nich dokumentów o jednakowej łącznej objętości. Tak zbudowany korpus będzie niewątpliwie zrównoważony, ale pewne typy tekstów, na przykład gatunki bardzo popularne, będą miały w nim za małą reprezentację, a inne, na przykład niezwykle rzadkie – za dużą. Podejściem umożliwiającym uniknięcie tego problemu i w ocenie Górskiego znacznie lepiej ugruntowanym metodologicznie jest dobieranie dokumentów o łącznej objętości proporcjonalnej do poziomu występowania danego typu populacji. Uzyskany w ten sposób korpus będzie reprezentatywny dla ogólnej populacji tekstów, ale już niekoniecznie zrównoważony. Jeszcze innym sposobem jest odtworzenie recepcji tekstów w danej społeczności językowej i uwzględnienie wag odzwierciedlających zróżnicowanie recepcji poszczególnych tekstów (miarą recepcji jest najczęściej nakład).

Korpusy są często dzielone na ogólne, duże, obejmujące zróżnicowany zestaw gatunków i reprezentatywne dla danego języka jako całości[40] oraz specjalistyczne, z reguły znacznie mniejsze, zawierające teksty powstałe w określonej społeczności językowej[41]. W tym drugim przypadku punktem odniesienia dla oceny reprezentatywności i zrównoważenia jest zwykle wybrana odmiana języka (np. język prawniczy), gatunek (np. artykuły prasowe) albo konkretny temat (np. zmiany klimatyczne). Korpusy można klasyfikować ze względu na to, czy odzwierciedlają język lub jego odmianę w danym momencie (synchroniczne), czy zawierają teksty publikowane w wybranym okresie i odzwierciedlają rozwój historyczny języka (diachroniczne) oraz

38 Tamże.

39 R.L. Górski, *Charakterystyka chronologiczna i stylistyczna korpusu dla „Wielkiego słownika języka polskiego"*, w: *Nowe studia leksykograficzne*, red. P. Żmigrodzki, R. Przybylska, Lexis, Kraków 2008; zob. też R.L. Górski, M. Łaziński, *Reprezentatywność Narodowego Korpusu Języka Polskiego*.

40 Takim korpusem jest m.in. Narodowy Korpus Języka Polskiego, zrównoważony gatunkowo i tematycznie (Przepiórkowski i in., 2012).

41 P. Baker, *Using Corpora in Discourse Analysis*, Continuum, London 2006.

czy zawierają teksty w całości, co pozwala na uwzględnienie w analizie ich struktury, czy też fragmenty o określonej wielkości (korpusy próbkowane). Korpusy mogą się także różnić pod względem zawartości informacji metajęzykowych (metadanych), mogą na przykład wskazywać na przynależność gatunkową, datę powstania danego tekstu czy autorstwo. Dodatkowo mogą być poddane anotacji, czyli procesowi przypisania do określonych fragmentów tekstu dodatkowych informacji, takich jak znaczniki części mowy, informacje prozodyczne lub semantyczne[42].

Reprezentatywność i zrównoważenie są traktowane w literaturze jako cechy, do których warto, a w przypadku korpusów ogólnych, opracowywanych przede wszystkim na potrzeby badań językoznawczych, nawet trzeba dążyć, ale które w praktyce są bardzo trudne do osiągnięcia[43]. Wskazywano na przykład, że zrównoważenie nigdy nie będzie pełne – w projekcie korpusu nie da się uwzględnić wszystkich klas, na jakie można podzielić daną przestrzeń języka – a ocena reprezentatywności, jak już zauważyliśmy, wymaga znajomości całej populacji tekstów. W przypadku niektórych odmian języka czy typów dyskursu zakres dostępnych informacji jest niepełny, a wyznaczenie granic i ścisłe zdefiniowanie populacji bywa niemożliwe[44]. Próby przypisywania tekstom wagi odzwierciedlającej ich znaczenie kulturowe mogą prowadzić do lawinowego przyrostu liczby przyjmowanych założeń, na przykład dotyczących proporcji tekstów z różnych rejestrów stylistycznych[45]. Przeszkodą w konstruowaniu dużego i zrównoważonego korpusu są wreszcie często względy praktyczne, takie jak niewielka dostępność danego typu tekstów w formie elektronicznej lub brak możliwości ich digitalizacji czy prawa autorskie[46].

Za szczególnie trudne uchodzi stworzenie korpusu zawierającego materiał historyczny. Realizacja postulatu zrównoważonego doboru źródeł trafia tu na znacznie więcej przeszkód teoretycznych i praktycznych niż w przypadku korpusów tekstów współczesnych. Podstawowym i nierozwiązywalnym problemem jest ograniczona wiedza o piśmiennictwie danej epoki – można

42 Tamże; por. R.L. Górski, *Metody korpusowe i kwantytatywne*…

43 D. Biber, *Representativeness in Corpus Design*.

44 R.L. Górski, *Charakterystyka chronologiczna i stylistyczna*…

45 T. Underwood, *The Real Problem with Distant Reading* (blog), 2016, https://tedunderwood.com/2016/05/29/the-real-problem-with-distant-reading/ (14.06.2022).

46 S. Hunston, *Collection Strategies and Design Decisions*; P. Baker, *Corpus Linguistics*…

odtworzyć główne tendencje czy wskazać popularne typy i gatunki, ale informacje o strukturze dokumentów funkcjonujących w danej epoce nigdy nie będą pełne[47]. Przyjmuje się, że poszczególne podkorpusy korpusu diachronicznego powinny mieć podobną (a w wariancie optymalnym identyczną) budowę, Górski jednak wskazuje, że jest to warunek niemożliwy do spełnienia ze względu na ogromne dysproporcje między liczbą zachowanych tekstów z epok dawniejszych a liczbą tekstów współczesnych i ze względu na przekształcenia w obrębie populacji, na przykład pojawienie się nowych typów i marginalizacja typów niegdyś popularnych[48].

Przedstawione wyżej ograniczenia mogą prowadzić do budowy korpusu „oportunistycznego", który składać się będzie ze wszystkich dostępnych tekstów i z którego dopiero w kolejnym kroku utworzony zostanie korpus zrównoważony[49]. Co istotne, zbiory tekstów zawierające wszelkie dostępne dane językowe, niezgodne z rygorystycznymi zasadami doboru próby lub niekompletne, również mogą stanowić podstawę do wyciągnięcia istotnych wniosków na temat danej populacji, pod warunkiem jednak że badacz lub badaczka będą w stanie przedstawić źródła i konsekwencje tej niekompletności[50].

Protokorpusy w badaniach literackich

Choć praktyka tworzenia korpusów nie zadomowiła się jeszcze na dobre w badaniach literackich – o czym piszemy szerzej w kolejnej części artykułu – warto przyjrzeć się praktykom towarzyszącym podobnemu pod wieloma względami gatunkowi wypowiedzi naukowej. Antologie, bo o nich mowa, można uznać za swoiste protokorpusy, czyli zbiory tekstów tworzone często według kryteriów zbliżonych do opisywanych wyżej, ale służące nieco innym celom badawczym. Zbiory te interesują nas zatem zarówno pod względem metodologicznym, tj. kryteriów wyboru, które można zastosować do

47 W. Gruszczyński, D. Adamiec, R. Bronikowska, A. Wieczorek, *Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. – problemy teoretyczne i warsztatowe*, „Poradnik Językowy" 2020, nr 8.

48 R.L. Górski, *Metody korpusowe i kwantytatywne…*, s. 68.

49 E. Hajnicz, N*ajbardziej znane korpusy tekstów*.

50 J.Grimmer, M.E. Roberts, B.M. Stewart, *Text as Data: A New Framework for Machine Learning and the Social Sciences*, Princeton University Press, New Jersey–Oxford 2022; S. Goźdź-Roszkowski, *Corpus Linguistics in Legal Discourse*, „International Journal for the Semiotics of Law" 2021, nr 34; M. Zaśko-Zielińska, *The Linguistic Analysis of Suicide Notes*, w: V. Guillén-Nieto, D.Stein, *Language as Evidence. Doing Forensic Linguistics*, Palgrave, Cham 2021.

współczesnych korpusów, jak i pragmatycznym – jako potencjalne źródło tekstów już opracowanych do wykorzystania w korpusach. Jednak posłużenie się tymi materiałami należy poprzedzić pogłębioną analizą i refleksją nad metodologią ich powstania.

Na potrzeby prac nad Korpusem Dyskursu Literaturoznawczego przeprowadziliśmy przegląd antologii gromadzących polskojęzyczne teksty z zakresu teorii i historii literatury, a także krytykę literacką, manifesty i wypowiedzi programowe. Zidentyfikowaliśmy 29 takich pozycji wydanych w latach 1959--2020. Szczegółowej analizie poddaliśmy 21 artykułów wprowadzających, w których redaktorzy opisywali przyjęte przez siebie kryteria czy strategie doboru tekstów (zob. aneks 1). Wstępy te mają zróżnicowany charakter: od obszernych uzasadnień obecność poszczególnych tekstów po lakoniczne wzmianki w przypisach lub w nocie edytorskiej sugerujące, że ich dobór jest raczej bezdyskusyjny. Osobną grupę stanowią tu antologie z serii Biblioteki Narodowej, do których wstępy są w istocie rozprawami historycznoliterackimi. Ich autorzy rzadko poświęcają miejsce selekcji tekstów, ponieważ kanoniczność dobranego materiału stanowi istotę serii.

O ile samo pojęcie antologii ma w literaturze i literaturoznawstwie długą historię, o tyle określanie tym mianem książki zbiorowej zawierającej teksty naukowe jest zjawiskiem znacznie młodszym. Dla przykładu w Polskiej Bibliografii Literackiej[51], której struktura od zarania pozostaje w zasadzie niezmienna, w dziale *Antologie i zbiory* znaleźć można głównie antologie literackie, a także zbiory o charakterze mieszanym, uwzględniające obok utworów literackich na przykład manifesty, wystąpienia programowe lub krytykę literacką. Jest to zgodne z większością słownikowych definicji antologii.

W przypadku ukazujących się po drugiej wojnie światowej książek gromadzących polskie teksty literaturoznawcze, zwłaszcza teoretycznoliterackie, słowo „antologia" z reguły nie trafiało do tytułu – najczęstsze określenie to „wybór tekstów", a także „wypisy". Inaczej rzecz ma się z tytułami zbiorów przekładów literaturoznawczych, na przykład *Współczesna teoria badań literackich za granicą. Antologia* z lat 1976-1996 czy *Teorie literatury XX wieku* z 2006 roku. Pojęciem antologii posługują się jednak niemal wszyscy autorzy analizowanych wstępów, począwszy od *Polskiej krytyki literackiej 1800-1918* i *Stylistyki polskiej*, i tę cząstkową autoidentyfikację traktować można jako wyraz nowszej świadomości. Najczęściej termin ten pojawia się w tytułach zbiorów gromadzących polską krytykę literacką oraz programy i dyskusje czy ogólniej, teksty,

51 https://pbl.ibl.poznan.pl (4.12.2023).

których autorami mogą być – i często są – pisarze, co potwierdza szczególną bliskość między literaturą a krytyką na tle innych działów literaturoznawstwa (dokładnie na pograniczu znajdują się eseje pisarzy poświęcone literaturze). Według Danuty Ulickiej rozszerzenie pojęcia antologii na książki zawierające teksty naukowe wiązałoby się przede wszystkim z charakterystycznym dla ponowoczesności „zalegalizowaniem usunięcia" granic między sztuką a nauką[52]. Sama częstotliwość takiej autoidentyfikacji we wstępach do antologii z lat 1959-2020 zdaje się dodatkowo potwierdzać, że wskazane przez badaczkę przesunięcie definicji zachodziło w polskim literaturoznawstwie właśnie w tym czasie.

Przyjrzyjmy się teraz kryteriom wyboru stosowanym w tego typu publikacjach. Na podstawie analizy antologii naukowych Ulicka zwraca uwagę na swoiste napięcie między obiektywizmem a umownością. Choć teksty wybierane są zwykle „spośród innych możliwych na podstawie kryteriów, które nie powinny budzić kontrowersji", to jednak „zarówno wybór dokonywany na podważalnej podstawie (niepodważalnej jakoby wartości lub reprezentatywności), jak i narzucony układ pozostają umowne"[53]. Analizując kryteria wyboru wskazywane przez autorów wstępów do antologii tekstów literaturoznawczych, można wyłonić pewne ich grupy: 1) kryteria wynikające ze specyfiki antologii lub zgromadzonego w niej materiału, 2) kryteria uwzględniające nieostrość różnych granic (epok czy literaturoznawstwa jako dziedziny), 3) kwestie praktyczne (niedostępność tekstów lub przydatność dydaktyczna), 4) dążenie do utrzymania wielogłosowości (dialogiczności) i równowagi (np. typów tekstu, głosów w sporach, postaw metodologicznych) i wreszcie 5) kryteria związane z wartościowaniem.

W analizowanych przez nas pracach najczęściej wskazywana jest pierwsza grupa zagadnień, czyli kryteria pozostające w ścisłym związku ze specyfiką dziedziny lub profilem konkretnej antologii: charakterem epoki lub prądu literackiego. I tak „drobne, nieważne artykuliki oddające ducha skandalu" i charakterystyczne materiały literackie dołączone do *Antologii polskiego futuryzmu i nowej sztuki* mają oddać sprawiedliwość niezwykłości tego nurtu, koncentracja na dynamice sporu stanowi istotę strategii doboru wypowiedzi w *Walce klasyków z romantykami*, a konieczność uwzględnienia różnych szkół i doktryn to ważny aspekt kompozycyjny książki *Teoretyczne tematy i problemy*.

52 D. Ulicka, *Siła antologii*, w: *Wiek teorii. Sto lat nowoczesnego literaturoznawstwa polskiego. Antologia*, red. D. Ulicka, t. 1, Wydawnictwo IBL PAN, Warszawa 2020, s. 11.

53 Tamże.

Obok kwestii wyznaczanych umownie ram czasowych ważny okazał się także wyłaniający się w pewnych punktach związanych z charakterem opracowanego materiału problem nieostrości granic między dziedzinami wiedzy, a ściślej: granic literaturoznawstwa. Na przykład kłopot z wyznaczeniem ostrej granicy między krytyką literacką a teatralną czy artystyczną sygnalizuje Zygmunt Jakubowski we wstępie do *Polskiej krytyki literackiej*. W przypadku rozprawy w jakiejś mierze pogranicznej, kontrowersyjnej lub włączonej na prawach wyjątku we wstępie znalazło się dodatkowe uzasadnienie wyboru autora lub samego tekstu (dotyczy to także umieszczenia listu Strzemińskiego w *Wieku teorii*).

Kryteria doboru przywoływane przez antologistów często mają też wymiar praktyczny. Przekłada się to na decyzje o włączaniu do zbiorów tekstów trudniej dostępnych (*Awangarda poetycka*, *Teoria badań literackich – wypisy*, *Polska krytyka literacka*, *Stylistyka polska*, *„Chamuły", „gnidy", „przemilczacze"*) w celu wydobycia ich z zapomnienia lub scalenia głosów rozproszonych przy jednoczesnej rezygnacji z tekstów niedawno wznawianych. Równie istotnym kryterium okazuje się przydatność dydaktyczna (*Problemy teorii literatury seria 1-4*, *Teoretyczne tematy i problemy*, *Stylistyka polska*, *Genologia polska*, *Kartografowie dziwnych podróży*) lub komplementarność z podręcznikiem, kompendium czy monografią (*Stylistyka polska*, *Genologia polska*, *Teoretyczne tematy i problemy*, *Wiek teorii*).

Kolejnym deklarowanym w antologiach celem selekcji jest utrzymanie możliwie dużej wielogłosowości, a zarazem równowagi, co współgra z dążeniem do zrównoważenia korpusu omówionym w początkowej części tekstu. To podejście cechuje z jednej strony próba doboru tekstów oddających swoistość cech danej epoki, nurtu lub dziedziny a z drugiej – utrzymania równowagi między różnymi typami wypowiedzi na poziomie konstrukcji. Jako warte zachowania autorzy antologii wymieniają: występującą w całości materiału różnorodność (*Antologia polskiej krytyki na emigracji*), wielowątkowość i interdyscyplinarność (*Polska myśl przekładoznawcza*), „tam, gdzie to możliwe, więcej niż jedno ujęcie tego samego problemu" (*Stylistyka polska*) i „zróżnicowanie postaw metodologicznych", a w przypadku braku miejsca na przedstawienie dyskusji i sporów nawet żywych i ważnych, przynajmniej zasygnalizowanie punktów spornych (*Genologia polska*). Do konstrukcji odnosi się postulat zrównoważenia kompozycji antologii rozumianej jako zbiór zbliżonej liczby tekstów o różnej formie (różnorodne pamflety w *„Chamułach"*…), określonym typie (cztery grupy tekstów w *Antologii polskiej krytyki na emigracji*: recenzje, szkice biograficzno-krytyczne, prace o ambicjach historycznoliterackich i teksty z pogranicza teorii literatury), a nawet podejmujących różne

tematy (równowaga między krytyką reagującą na aktualne wydarzenia w literaturze a tą, która odnosi się do przeszłości literatury, w zbiorze *Polska krytyka literacka*).

Wreszcie najbardziej złożone są kryteria doboru tekstów oparte na ocenie ich wartości wyrażonej w mniej lub bardziej otwarty sposób. Wśród tych kryteriów należałoby wymienić: a) kanoniczność („najważniejszość"), czyli rangę (historycznoliteracką): *Problemy teorii literatury*, *Programy i dyskusje lwowskiej krytyki literackiej 1896-1914*, *Antologia polskiej krytyki na emigracji*; b) reprezentatywność (znamienność), to znaczy charakterystyczność dla epoki: *Programy i dyskusje lwowskiej krytyki literackiej 1896-1914*, *Zapomniane głosy*, *Polska genologia literacka*, *„Chamuły"…*; c) nowatorskość i pionierskość: *Teoretyczne tematy i problemy*, *Wiek teorii*; oraz d) żywotność i aktualność, czyli wartość potwierdzona przez cytowania, wzmianki w innych tekstach, powroty czy „zbieżność z kwestiami poruszanymi we współczesności": *Teoretyczne tematy i problemy*, *Problemy teorii literatury*, *Wiek teorii*, *Antologia polskiej krytyki na emigracji*. Interesującym kryterium wyboru są: otwarcie deklarowana subiektywność, osobiste upodobanie, a także inspiracje wskazywane zwłaszcza przez redaktorów – nierzadko również krytyków – antologii obejmujących najnowszą krytykę literacką (*Kartografowie dziwnych podróży*, *Była sobie krytyka*). Trzeba zaznaczyć, że wymienione kategorie niekiedy zachodzą na siebie i bywają traktowane łącznie: we wstępie do *Teorii badań literackich – wypisów* Henryk Markiewicz deklaruje, że zbiór gromadzi teksty najważniejsze, czyli przede wszystkim najwybitniejsze, lecz uzupełnione tekstami pionierskimi lub historycznie znamiennymi.

Podmiotem wartościującym może być przy tym zarówno instancja zewnętrzna – opinia autorytetów na temat tekstu, świadectwo szerokiej recepcji czy ranga źródeł, z których teksty pochodzą (np. czasopism), również zadekretowana wcześniej przez specjalistów (*Antologia polskiej krytyki na emigracji*), jak i wewnętrzna. W tym drugim przypadku sami antologiści wskazują na wybitność, na przykład dojrzałość interpretacyjną i oryginalność (*Antologia polskiej krytyki na emigracji*) tekstu nowego (*Kartografowie dziwnych podróży*, *Była sobie krytyka*), zapomnianego lub niedocenionego w swojej epoce oraz w czasie późniejszym (*Programy i dyskusje lwowskiej krytyki literackiej*).

Warto zauważyć, że pewne dylematy artykułowane przez badaczy komponujących antologie literaturoznawcze mogą zbiegać się z rozterkami konstruktorów korpusu literaturoznawczego. Dotyczy to choćby takich wyzwań, jak konieczność dokonywania „wyboru z wszystkiego" przy istotnej niemożności uzyskania wiedzy o całej populacji tekstów literaturoznawczych (problem

niezachowanych tekstów z dawnych epok i przytłaczającej przewagi tekstów współczesnych), mgliste niekiedy granice dyscypliny lub ograniczenia związane z prawami autorskimi. Wielość kryteriów przywoływana przez redaktorów różnych antologii ujawnia przy tym pewien horyzont utopijny: niektóre wymogi mogą się wykluczać, ale mogą się też sumować. Słowem, najlepiej, gdyby wyłonione teksty spełniały je wszystkie: odpowiadały dokładnie specyfice tematu antologii, spełniały warunki największej wagi i przydatności, stwarzały poczucie wielogłosowości i nie naruszały przy tym wrażenia uczciwości oraz obiektywności wyboru. Nierozwiązywalność tych dylematów i niemożność skomponowania antologii idealnej o niepodważalnych kryteriach wyboru nie blokują jednak kolejnych ważnych inicjatyw wydawniczych.

Wydaje się, że w przypadku antologii wyraźniejsze konsekwencje ma prymat celu, jakim jest dostarczanie bezpośredniej wiedzy na określony temat w formie treści do bliskiego czytania (*close reading*). Również z tego powodu – inaczej niż w przypadku korpusów – ważniejsze od wymogów zachowania odpowiednich proporcji względem populacji tekstów czy konieczności oddania wszystkich rodzajów tekstów bywają tu warunki bardziej arbitralne, czy może bardziej autorskie – kryteria rzeczowe i aksjologiczne.

Wrażenie to pogłębiane jest przez fakt, że ze względu na ograniczenia wydawnicze selekcja z całej populacji musi być w przypadku antologii o wiele ostrzejsza. Przy bardzo dużej obfitości materiału antologistom pozostaje wybór fragmentów (*Polska krytyka literacka*, *Wiek teorii*) lub zastrzeżenie niepełności. Innym wyjściem jest nastawienie na przekaz wiedzy najbardziej podstawowej i jedynie zasygnalizowanie całości (*Genologia polska*), na przykład przez przedstawienie tylko najwyrazistszych propozycji w charakterze punktów orientacyjnych (*Polska myśl przekładoznawcza*). W przypadku korpusów dopuszczalnej liczby tekstów nie obejmie żadne wydanie książkowe, dlatego rozterki antologistów dotyczące tego, czy i kiedy warto dążyć, by wybór tekstów był jak najobszerniejszy, nie ma racji bytu.

Korpusy w badaniach literackich

Praktyka tworzenia korpusów na potrzeby badań literackich jest stosunkowo nowa i nie wykształciły się jeszcze wzorcowe procedury projektowania zbiorów tekstów literackich i literaturoznawczych[54]. Do niedawna

54 E. Gius, K. Krüger, C. Sökefeld, *Korpuserstellung als literaturwissenschaftliche Aufgabe*, w: *DHd 2019 Digital Humanities: multimedial & multimodal Konferenzabstracts. 6. Tagung des Verbands*,

w analizach z zakresu humanistyki cyfrowej stosowano najczęściej korpusy oportunistyczne, zawierające teksty łatwo dostępne w formie cyfrowej[55]. Omówione wcześniej zasady wypracowane zostały na gruncie językoznawstwa, w którym badania korpusowe są dziś najmocniej rozwinięte. Niewielka obecność tego rodzaju analiz w badaniach literackich wynikać może przede wszystkim z dwóch czynników: niedostosowania metodologii do literaturoznawstwa oraz nakładu pracy niezbędnego do zgromadzenia materiału zgodnie z wytycznymi wyłożonymi w początkowej części tekstu. Wydaje się, że między tymi przyczynami zachodzi sprzężenie zwrotne – brak korpusów nie sprzyja rozwojowi metod i vice versa, bez metod trudno uzasadnić przydatność prac korpusowych. Niemniej jednak na tym polu możemy odnotować istotne inicjatywy. Zmierzają one przede wszystkim do utworzenia korpusów literackich obejmujących materiał historyczny.

Jednym z najnowszych, a zarazem najambitniejszych tego rodzaju przedsięwzięć jest European Literary Text Collection (ELTeC)[56], wielojęzyczna kolekcja korpusów czy też korpus składający się z podkorpusów, z których każdy zawiera powieści z danej narodowej (a właściwie językowej) tradycji literackiej, opracowywany w ramach akcji COST „Distant reading for European literary history"[57]. Celem jego twórców było oddanie różnorodności produkcji literackiej przy jednoczesnym zapewnieniu porównywalności tekstów i podkorpusów[58]. Dlatego przyjęto, że każdy podkorpus powinien spełniać te same kryteria kompozycyjne z zachowaniem pewnej elastyczności, pozwalającej na uwzględnienie kontekstów lokalnych. ELTeC został zaprojektowany jako korpus monitorujący, do którego będzie można stopniowo dodawać teksty w różnych językach i z różnych okresów, i jest tworzony w sposób iteracyjny. Obecnie, w ramach tak zwanego ELTeC-core, udostępniono jedenaście podkorpusów spełniających kryteria przyjęte w projekcie, z których każdy składa się ze 100 utworów napisanych oryginalnie w danym

„Digital Humanities im deutschsprachigen Raum" (DHd 2019), Frankfurt am Main und Mainz. https://doi.org/10.5281/zenodo.4622112 (4.12.2023), s. 165.

55 Np. J. Rybicki, *Pierwszy rzut oka na stylometryczną mapę literatury polskiej*, „Teksty Drugie" 2014, nr 2, s. 106-128; M. Eder, *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii*, „Teksty Drugie" 2014, nr 2.

56 https://www.distant-reading.net/eltec/ (4.12.2023).

57 „Distant reading for European literary history", COST Action CA16204.

58 Ch. Schöch, R. Patras, T. Erjavec, D. Santos, *Creating the European Literary Text Collection (ELTeC). Challenges and Perspectives*, „Modern Languages Open" 2021, nr 1, s. 25.

języku i opublikowanych w latach 1840-1920[59]. Korpus jest stopniowo rozszerzany – zwiększa się liczbę tekstów w poszczególnych językach, uwzględniane są także utwory opublikowane przed rokiem 1840[60].

ELTeC jest zrównoważony pod względem języka i dat publikacji tekstów, a zarazem jednorodny pod względem gatunkowym. Do podkorpusów włączono powieści definiowane jako fikcyjne narracyjne teksty prozatorskie o długości nie mniejszej niż 10 tysięcy słów. Nie uwzględniono przekładów i powieści w odcinkach, chyba że w danej tradycji literackiej dominowały utwory wydawane w takiej właśnie formie (ELTeC 2018). Zdecydowano, że zbiór powinien zawierać zarówno teksty, które można uznać za część współczesnego kanonu, jak i utwory już w dużej mierze zapomniane. Jako miarę kanoniczności przyjęto liczbę wznowień danej publikacji[61]. Proces selekcji powieści miał dwa etapy: najpierw zidentyfikowano utwory, które spełniały przyjęte kryteria, a następnie zrównoważono poszczególne podkorpusy. Teksty podzielono na klasy ze względu na przedział czasowy (wydzielono cztery dwudziestoletnie podokresy), stopień kanoniczności (a więc liczbę przedruków), płeć autorów, długość oraz liczbę tekstów danego autora lub autorki w zbiorze. Dążono do zachowania możliwie równych proporcji między klasami, ale ponieważ nie we wszystkich przypadkach było to możliwe, określono minimalny i maksymalny udział danej klasy tekstów w podkorpusie (np. co najmniej 10% i maksymalnie 50% tytułów powinno mieć autora płci żeńskiej).

Pod względem dążenia do wielojęzyczności podobnym projektem jest kolekcja D r a C o r [62] (Drama Corpora), która stanowi wielojęzyczny zbiór dramatów opracowanych w formacie TEI, mający służyć do komparatystycznych

59 Wybór takiego a nie innego okresu był podyktowany względami czysto praktycznymi. Po 1840 roku w wielu językach europejskich opublikowano wystarczającą dla celów projektu liczbę powieści. Zbieranie tekstów wydanych nie później niż w 1920 r. wynikało natomiast z dążenia, by uwzględnić wyłącznie utwory znajdujące się w domenie publicznej, które można swobodnie wykorzystywać i udostępniać; por. tamże. Należy jednak odnotować, że w niektórych podkorpusach, na przykład w podkorpusie polskojęzycznym, znajdują się także teksty wydane po 1920 r.

60 F. Frontini, C. Brando, J. Byszuk, I. Galleron, D. Santos, R. Stanković, *Named Entity Recognition for Distant Reading in ELTeC*, w: *CLARIN Annual Conference Proceedings*, 5-7 października 2020, red. C. Navarretta, M. Eskevich, France 2020.

61 Ch. Schöch, R. Patras, T. Erjavec, D. Santos, *Creating the European Literary Text Collection (ELTeC)*.

62 https://dracor.org/ (4.12.2023).

badań literackich[63]. Obecnie platforma zawiera korpusy w różnych językach (baszkirski, francuski, grecki, hiszpański, niemiecki, rosyjski, szwedzki, tatarski, ukraiński, węgierski, włoski), a także kolekcje (korpusy Szekspira i Calderona czy korpus Szekspira w języku niemieckim). Projekt nie dąży do reprezentatywności ani zrównoważenia, rozwija raczej różnojęzyczne kolekcje dramatów, pozwalając użytkownikom na pracę z poszczególnymi utworami lub samodzielnie zdefiniowanymi zbiorami, także w podziale na tekst główny, didaskalia czy wypowiedzi konkretnej postaci. Twórcom projektu przyświeca idea „korpusów programowalnych" (*programmable corpora*), czyli możliwości różnorodnych użyć badawczych dzięki API, które pozwala badaczkom pobierać konkretne teksty czy dane w potrzebnych formatach[64]. DraCor automatycznie generuje sieci postaci dramatów, które można eksplorować na stronie projektu lub pobierać i wykorzystywać do analizy w innych programach.

W projekcie korpusu języka francuskiego F R A N T E X T[65], opracowanym przez Analyse et Traitement Informatique de la Langue Française jako baza słownika *Trésor de la langue française*, skupiono się na kwestii recepcji tekstów i dążono do rekonstrukcji kanonu literackiego[66]. W doborze utworów kierowano się, jak to określono, „zasadą autorytetu"[67]: sięgnięto do kilku uznanych syntez historii literatury francuskiej XIX i XX wieku i sporządzono listę wszystkich wymienionych w nich utworów. Do korpusu włączono publikacje, które zostały wspomniane co najmniej czterokrotnie (w przypadku tekstów dwudziestowiecznych) lub pięciokrotnie (w przypadku utworów dziewiętnastowiecznych). Decyzje o włączeniu utworów, o których wspomniano mniej niż cztery lub pięć razy, ale więcej niż dwa, podejmowała komisja złożona z ekspertów dziedzinowych.

Również w korpusie K O L I M O (Corpus of Literary Modernism)[68], zaprojektowanym na Uniwersytecie w Getyndze i zawierającym teksty literackie

63 F. Fischer, I. Börner, M. Göbel, A. Hechtl, Ch. Kittel, C. Milling, P. Trilcke, *Programmable Corpora. Introducing DraCor, an Infrastructure for the Research on European Drama*, lipiec 2019, s. 1, https://doi.org/10.5281/ZENODO.4284002 (1.10.2023).

64 Tamże, s. 5.

65 https://www.frantext.fr/ (4.12.2023).

66 A. Grieve-Smith, *FRANTEXT's Corpus of Nineteenth-Century French*, w: *Building a Representative Theater Corpus*, red. A. Grieve-Smith, Palgrave Pivot, Cham 2019.

67 P. Imbs, *Trésor de la langue française*, CNRS, Paris 1971, s. XXIII.

68 https://kolimo.uni-goettingen.de (4.12.2023).

z okresu niemieckiego modernizmu (lata 1880-1930), do zdefiniowania populacji wykorzystano literaturę przedmiotu, najprawdopodobniej podręczniki akademickie. Dodatkowo uwzględniono dane pochodzące rejestrów bibliograficznych. Najpierw dokonano wyboru autorów i autorek – na podstawie tego, ile razy zostali określeni jako moderniści, a także na podstawie praktyk czytelniczych i historii recepcji ich dorobku. Następnie oceny istotności poszczególnych tekstów dokonali eksperci z danej dziedziny[69]. W ten sposób uzyskano zbiór o objętości około 40 tysięcy tekstów (600 milionów słów), napisanych przez 1800 osób, w tym 1426 mężczyzn. Korpus nie został ukończony – zaplanowano dalsze prace nad zwiększeniem jego reprezentatywności i normalizacją tekstów.

W ramach projektu „Gender and illness" na bazie KOLIMO utworzono korpus d-Prose 1870-1920[70]. Jako kryteria doboru przyjęto datę pierwodruku, język, gatunek i długość tekstu. Do korpusu włączono utwory opublikowane w latach 1870-1920 o minimalnej długości 1000 słów, usunięto teksty nienarracyjne i napisane w języku innym niż niemiecki, w szczególności tłumaczenia, wyeliminowano duplikaty, a także – korzystając z repozytoriów i encyklopedii literackich – zweryfikowano i uzupełniono metadane KOLIMO. Utworzony w ten sposób zbiór składa się z 2511 tekstów 334 autorów reprezentujących trzy różne ruchy literackie (naturalizm, realizm i modernizm). Zawiera zbliżone proporcje tekstów długich (powieści) oraz krótszych form prozatorskich.

Na gruncie polskim za punkt odniesienia przy projektowaniu korpusów zawierających materiał historyczny można uznać KorBa[71], elektroniczny korpus tekstów polskich z XVII i XVIII wieku (do roku 1772) opracowany przez Pracownię Historii Języka Polskiego XVII i XVIII wieku Instytutu Języka Polskiego Polskiej Akademii Nauk we współpracy z Zespołem Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN. Korpus liczy około 13,5 miliona segmentów (w uproszczeniu: wyrazów) i obejmuje nie tylko charakterystyczne dla epoki baroku teksty literackie, lecz także teksty użytkowe oraz religijne. Twórcy i twórczynie KorBa dążyli do osiągnięcia reprezentatywności i zrównoważenia, przy założeniu jednak, że realizacja

69 B. Herrmann, G. Lauer, *KOLIMO – A Corpus of Literary Modernism for Comparative Analysis*, 2017; https://kolimo.uni-goettingen.de/about (1.06.2022).

70 E. Gius, S. Guhr, I. Uglanova, *„d-Prose 1870-1920" a Collection of German Prose Texts from 1870 to 1920*, „Journal of Open Humanities Data" 2021, nr 7, s. 11.

71 https://korba.edu.pl (4.12.2023).

tego postulatu nie będzie pełna[72]. Ograniczony dostęp do materiału wymusił uwzględnienie zróżnicowanych typów źródeł, począwszy od starodruków i rękopisów, poprzez wydania pochodzące z czasów późniejszych, aż po współczesne publikacje w formie cyfrowej. Do budowy zbioru wykorzystano także bazę materiałową korpusu utworzonego w ramach międzynarodowego projektu IMPACT (Improving Access to Texts), a więc pierwszego korpusu dawnych tekstów polskich spełniającego współczesne standardy konstrukcyjne, liczącego około 1,8 miliona segmentów. Przy podziale tekstów na klasy uwzględniono chronologię (wyróżniono cztery podokresy), geografię, gatunek oraz tematykę. Dążono do zrównoważonej ilościowo reprezentacji czterech podokresów, lecz kryterium to było w trakcie prac nad korpusem wielokrotnie modyfikowane. Do zbioru włączono publikacje pochodzące z regionów wyróżnianych zwykle w badaniach historycznych tego okresu, ich liczba jest jednak zróżnicowana i odzwierciedla aktywność wydawniczą i piśmienniczą poszczególnych ośrodków.

Najnowszym korpusem polskojęzycznym jest 19/20MetaPNC (Metadata-enriched Polish Novel Corpus from the 19th and 20th centuries)[73], tworzony przez zespół badaczy i badaczek z IBL PAN, Wydziału Matematyki i Informatyki UAM, NASK PIB oraz PWr. 19/20MetaPNC ma w zamyśle stanowić referencyjny korpus prozy polskiej drugiej połowy XIX i pierwszej połowy XX wieku. Korpus jest zrównoważony pod względem historycznym i geograficznym[74]. Powstaje w sposób iteracyjny. Pierwsza wersja 19/20MetaPNC składa się z 1000 powieści napisanych oryginalnie w języku polskim i wydanych po raz pierwszy w formie książkowej w latach 1864-1939, z wyłączeniem powieści historycznych. Bazą materiałową były utwory pozyskane z polskiego podkorpusu ELTeC (100 powieści) oraz z serwisów Polona, Wikisource i Wolnelektury.pl. Ze względu na niemożność precyzyjnego określenia populacji tekstów oraz brak danych na temat produkcji i recepcji literackiej w interesującym autorów okresie ich wysiłki koncentrowały się na

72 W. Gruszczyński, D. Adamiec, R. Bronikowska, A. Wieczorek, *Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w.*

73 https://github.com/CHC-Computations/19-20MetaPNC (4.12.2023).

74 A. Karlińska, C. Rosiński, J. Wieczorek, P. Hubar, J. Kocoń, M. Kubis, S. Woźniak, A. Margraf, W. Walentynowicz, *Towards a Contextualised Spatial-diachronic History of Literature: Mapping Emotional Representations of the City and the Country in Polish Fiction from 1864 to 1939*, w: *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature,* red. S. Degaetano, A. Kazantseva, N. Reiter, S. Szpakowicz, International Conference on Computational Linguistics, Gyeongju, Republic of Korea 2022.

precyzyjnym opisie korpusu metadanymi i kwestii jego zrównoważenia. Jako kryteria równoważenia przyjęto przynależności powieści do jednej z trzech wyróżnianych w polskim literaturoznawstwie epok – pozytywizmu, Młodej Polski i dwudziestolecia międzywojennego – zgodnie z datą pierwodruku, zabór, w którym powieść ukazała się po raz pierwszy, płeć jej autora oraz poziom recepcji mierzony liczbą wznowień. Podobnie jak w ELTeC określono minimalny i maksymalny udział danej klasy tekstów w korpusie. Proces wzbogacania metadanych obejmował procedury ręczne i automatyczne, w których wykorzystano zasoby baz danych VIAF, Wikidata i Geonames. Poza standardowymi metadanymi takimi jak rok pierwodruku uwzględniono między innymi miejsce urodzenia autora wraz ze współrzędnymi geograficznymi, informację o przynależności terytorialnej miejsca wydania (zabór rosyjski, austriacki i pruski lub zagranica), liczbę przedruków oraz segmentów. W kolejnych wersjach opis metadanymi będzie jeszcze szerszy, a korpus obejmie powieści historyczne oraz opowiadania.

Omówione wyżej korpusy można krytykować za pewną arbitralność kryteriów doboru tekstów i ich ahistoryczność. Ze względu na brak wiarygodnych danych o populacji, zwłaszcza w przypadku epok dawniejszych, decyzje podejmowane przy tworzeniu korpusów literackich są często oparte na spekulacjach i założeniach teoretycznych[75]. W części opisanych projektów spekulatywność i ahistoryczność starano się ograniczać przez sięganie po dane dotyczące produkcji i recepcji tekstów pochodzące z rejestrów bibliograficznych (np. liczba wznowień) czy próbę rekonstrukcji praktyk czytelniczych w danej epoce. W innych opierano się wyłącznie na współczesnej wiedzy eksperckiej, co prowadziło do preferowania przede wszystkim tekstów kanonicznych, ważnych z dzisiejszego punktu widzenia. Arbitralność kryteriów kompozycji korpusów była pochodną dążenia do zrównoważenia danych i tym samym zapewnienia porównywalności tekstów i podkorpusów. Istotną trudnością, z którą muszą się zmierzyć twórcy i twórczynie korpusów obejmujących materiał historyczny, jest nierównomierność produkcji w ramach klas wyróżnionych na podstawie kategorii takich jak płeć, gatunek czy epoka literacka. Konieczne jest rozstrzygnięcie, czy próba ma oddawać różnorodność tekstów, czy też ich rzeczywiste proporcje w populacji. Na przykład przy wyborze 100 powieści z populacji utworów opublikowanych w ciągu 20 dekad można albo włączyć do korpusu po 5 tekstów z każdej dekady, albo ustalić, ile utworów zostało opublikowanych w danej dekadzie, i dostosować proporcje

75 T. Underwood, *The Real Problem with Distant Reading*.

do wyników tych ustaleń – ścisłe zastosowanie drugiego z tych podejść może oznaczać, że z niektórych dekad nie zostanie wybrany żaden tekst.

Konkluzje: w stronę Korpusu Dyskursu Literaturoznawczego

Z dotychczasowych rozważań jasno wynika, że korpus w pełni reprezentatywny i zrównoważony to byt idealny, pewien nieosiągalny w praktyce konstrukt. Najistotniejsze jest zatem opracowanie założeń, które pozwolą zbliżyć się do tego ideału i dobrze udokumentować podjęte decyzje, tak by korzystający z korpusu wiedzieli, jakie są jego ograniczenia, i mogli dostosować hipotezy badawcze do specyfiki zebranego materiału[76].

Jak już zostało powiedziane, prezentowane tu analizy przeprowadzono w ramach prac przygotowawczych do Korpusu Dyskursu Literaturoznawczego (KDL), który obejmie teksty polskich literaturoznawców i literaturoznawczyń z lat 1822-2022. KDL ma stanowić z jednej strony antologię istotnych utworów polskiego literaturoznawstwa, z drugiej – korpus przeznaczony do badań ilościowych, pozwalających na uchwycenie różnego rodzaju przemian.

Wedle naszej wiedzy nie ma obecnie ogólnodostępnego korpusu o profilu podobnym do KDL. Zbiorem o najbardziej zbliżonym charakterze, choć znacznie węższym zakresie zarówno czasowym, jak i tematycznym, jest T e o r i a / L i t e r a t u r a[77]. Pomysł na jego opracowanie narodził się w ramach prac nad projektem „Wiek teorii"[78], którego zwieńczeniem była trzytomowa antologia poświęcona stuleciu polskiej teorii literatury. Podstawą korpusu ma być efekt kwerendy przeprowadzonej przez zespół projektu „Wiek teorii" w 176 czasopismach publikowanych od 1900 roku, czyli łącznie 2749 pozycji bibliograficznych. Projekt korpusu zawiera też komponent badawczy, wykorzystujący narzędzia do przetwarzania języka naturalnego, między innymi analizę siatek pojęciowych czy modelowanie tematyczne. Prace nad korpusem są obecnie zawieszone.

---

76 R. Poos, R. Simpson, *Cross-disciplinary Comparisons of Hedging: Some Findings from the Michigan Corpus of Academic Spoken English*, w: *Using Corpora to Explore Linguistic Variation*, red. R. Reppen, S. Fitzmaurice, D. Biber, Benjamins, Amsterdam 2002.

77 M. Mrugalski, *Teoria/Literatura as a Mise En Abyme of Digital Research on Literary Studies: The Corpus of Polish Literary Theory Between Mathematical Intuitionism and Formalism*, „Russian Literature" 2021, nr 122-123

78 „Wiek teorii. Sto lat polskiej myśli teoretycznoliterackiej", Narodowe Centrum Nauki, grant nr 2014/13/B/HS2/00310, kier. prof. D. Ulicka.

W trakcie prac nad KDL stanęliśmy przed dwoma kluczowymi wyzwaniami. Pierwszym było precyzyjne zdefiniowanie populacji zarówno wszystkich wydanych tekstów, jak i ich twórców. Nie dysponujemy pełną listą publikacji z zakresu literaturoznawstwa, a listy, które można opracować na podstawie dostępnych danych bibliograficznych, nigdy nie będą wyczerpujące. Jednym z potencjalnych źródeł takich danych jest Biblioteka Narodowa. Jej zadania statutowe obejmują jednak przede wszystkim wieczyste przechowywanie polskiego piśmiennictwa w szerokim ujęciu, a selekcji publikacji dokonują poszczególne działy. Celem bibliotek narodowych nie jest pełne oddanie każdej z dziedzin, ale rejestrowanie wszystkich w pewnym stopniu – ich zasoby są z założenia wielo- czy też wszystkodziedzinowe i w związku z tym trudno je uznać za optymalne źródło wiedzy na temat węższej domeny dyskursu. Dane na temat publikacji literaturoznawczych można także pozyskać z bibliografii dziedzinowych, w szczególności z „Nowego Korbuta" i Bibliografii Bara. Pozycje te stanowią gruntowny rejestr dziewiętnastowiecznej literatury przedmiotu, ale już nie literatury dwudziestowiecznej. Nie dysponujemy analogicznymi źródłami pozwalających na rekonstrukcję listy publikacji z drugiej połowy XX wieku oraz prac najnowszych. Wykorzystanie wskazanych bibliografii prowadziłoby więc do przyjęcia nieheterogenicznych kryteriów doboru tekstów do korpusu. Ze względu na brak pełnych wykazów bibliograficznych nie mogliśmy zastosować losowego doboru próby.

Drugie wyzwanie wiąże się z nierównomierną dostępnością publikacji w formie cyfrowej, w szczególności niewielką dostępnością tekstów wydanych w latach 1920-1990. W przypadku prac w formie książkowej ograniczenie to wynika z prawa autorskiego, a w przypadku artykułów naukowych – z opóźnień w opracowywaniu publikacji archiwalnych przez redakcje czasopism naukowych. Chociaż zdecydowaliśmy się sięgnąć przede wszystkim po teksty już obecne w obiegu cyfrowym, podobnie jak twórcy ELTeC przyjęliśmy, że dana publikacja nie powinna być wyłączona z korpusu tylko dlatego, że nie została jeszcze zdygitalizowana. Dlatego w projekcie przewidzieliśmy prowadzenie prac dygitalizacyjnych oraz próby pozyskania tekstów bezpośrednio od wydawnictw i redakcji czasopism.

KDL będzie zawierać teksty literaturoznawcze, choć zdajemy sobie sprawę, iż kryteria zaliczenia utworu do tej dyscypliny ewoluowały przez dwieście lat. Bierzemy pod uwagę teksty poświęcone literaturze lub ukazujące się w czasopismach czy antologiach o profilu literaturoznawczym. Przy wyborze tekstów, podobnie jak w przypadku korpusu FRANTEXT, kierujemy się zasadą autorytetu i polegamy przede wszystkim na zastanej

wiedzy eksperckiej, zawartej w kluczowych zestawieniach, takich jak syntezy historycznoliterackie z materiałami pomocniczymi (wzmianka o utworze jako istotnym dla dyskusji literackich danego okresu), antologie i wybory pism danych twórców (obecność utworu jako świadectwo jego istotności) czy sylabusy zajęć z zakresu literaturoznawstwa, świadczące o ugruntowanej pozycji danego tekstu. Kolejny tekst przybliży metodologię KDL i ukaże się wraz samym korpusem.

Aneks 1. Lista antologii, z których wstępy zostały poddane analizie (porządek chronologiczny)

Polska krytyka literacka 1800-1918. Materiały, t. 1-5, red. J.Z. Jakubowski, J. Kulczycka-Saloni i in., PWN, Warszawa 1959.

Teoria badań literackich w Polsce: wypisy, t. 1-2, oprac. H. Markiewicz, Wydawnictwo Literackie, Kraków 1960.

Walka klasyków z romantykami, oprac. S. Kawyn, Zakład Narodowy im. Ossolińskich, Wrocław 1963.

Polska awangarda poetycka. Programy lat 1917-1923, t. 2: *Manifesty i protesty. Antologia*, oprac. A. Lam, Wydawnictwo Literackie, Kraków 1969.

Stylistyka polska: wybór tekstów, oprac. E. Miodońska-Brookes, A. Kulawik, M. Tatara, PWN, Warszawa 1973.

Programy i dyskusje literackie okresu Młodej Polski, oprac. M. Podraza-Kwiatkowska, Zakład Narodowy im. Ossolińskich, Wrocław 1973.

Antologia polskiego futuryzmu i nowej sztuki, oprac. Z. Jarosiński, H. Zaworska, Zakład Narodowy im. Ossolińskich, Wrocław 1978.

Genologia polska. Wybór tekstów, oprac. E. Miodońska-Brookes, A. Kulawik, M. Tatara, PWN, Warszawa 1983.

Programy i dyskusje literackie okresu pozytywizmu, oprac. J. Kulczycka-Saloni, Zakład Narodowy im. Ossolińskich, Wrocław 1985.

Problemy teorii literatury, wyb. H. Markiewicz, seria 1-4, Zakład Narodowy im. Ossolińskich, Wrocław 1976-1998.

Idee programowe romantyków polskich. Antologia, oprac. A. Kowalczykowa, Zakład Narodowy im. Ossolińskich, Wrocław 1991.

Antologia polskiej krytyki literackiej na emigracji 1945-1985, oprac. J. Dąbała, Redakcja Wydawnictw Katolickiego Uniwersytetu Lubelskiego, Lublin 1992.

Teoretycznoliterackie tematy i problemy, wyb. D. Ulicka, Wydział Polonistyki Uniwersytetu Warszawskiego, Warszawa 2003.

Była sobie krytyka… Wybór tekstów z lat dziewięćdziesiątych i pierwszych, oprac. D. Nowacki, K. Uniłowski, Wydawnictwo Uniwersytetu Śląskiego, Katowice 2003.

„Kartografowie dziwnych podróży". Wypisy z polskiej krytyki literackiej XX wieku, oprac. M. Wyka, K. Biedrzycki i in., Universitas, Kraków 2004.

Zapomniane głosy. Krytyka literacka kobiet 1894-1918, t. 1: *Wybór tekstów*, oprac. A. Wydrycka, Wydawnictwo Uniwersytetu w Białymstoku, Białystok 2006.

Polska genologia literacka, red. D. Ostaszewska, R. Cudak, Wydawnictwo Naukowe PWN, Warszawa 2007.

„Chamuły", „gnidy", „przemilczacze". Antologia dwudziestowiecznego pamfletu polskiego, oprac. D. Kozicka, Universitas, Kraków 2010.

Polska myśl przekładoznawcza. Antologia, red. P. Bończa Bukowski, M. Heydel, Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków 2013.

Programy i dyskusje lwowskiej krytyki literackiej 1896-1914. Antologia, oprac. K. Sadkowska, Wydział Polonistyki UW, Warszawa 2015.

Wiek teorii. Sto lat nowoczesnego literaturoznawstwa polskiego. Antologia, red. D. Ulicka, t. 1-2, Wydawnictwo IBL PAN, Warszawa 2020.

Abstract

Agnieszka Karlińska
NASK – NATIONAL RESEARCH INSTITUTE
Paulina Czwordon-Lis
INSTITUTE OF LITERARY RESEARCH OF THE POLISH ACADEMY OF SCIENCES
Maciej Maryl
INSTITUTE OF LITERARY RESEARCH OF THE POLISH ACADEMY OF SCIENCES

Text Corpus as a Tool for Literary Studies

The article discusses the methodology of creating corpora (collections of texts) that allow for the application of quantitative methods and drawing conclusions about the broader sampled population. In particular, the discussed corpora provide a snapshot of the discourse of a specific period, which can be used in diachronic analysis or the study of processes in the history of literature. Digital literary studies need corpora to apply its methods and to expand our understanding of such processes as the evolution of genres, literary concepts, or styles. We begin our discussion with a general overview of corpora as research datasets in various disciplines. We then focus on literary studies by scrutinizing anthologies qua proto-corpora, which seem to have served a similar function, albeit in a different methodological context. These explorations lead us to the review of selected corpora in literary studies. In conclusion, we present the main challenges of the work on the Corpus of Literary Studies Discourse (Korpus Dyskursu Literaturoznawczego; KDL).

Keywords

corpus, anthology, natural language processing, digital humanities, literary studies, digital literary studies