FOLIA 383

Annales Universitatis Paedagogicae Cracoviensis

Studia ad Bibliothecarum Scientiam Pertinentia 21 (2023)

ISSN 2081-1861

DOI 10.24917/20811861.21.25

Adam Pawłowski
Uniwersytet Wrocławski
ORCID 0000-0002-0804-5698

Korpus prasy polskiej ChronoPress jako infrastruktura i narzędzie badań medioznawczych

Wprowadzenie

Prasa na nośniku papierowym nie jest dziś tak użytecznym przedmiotem badań, jakim była przez dziesięciolecia poprzedzające nastanie ery cyfrowej. Jednym z głównych tego powodów jest postępująca zmiana praktyki badawczej, polegająca na tym, że badacze i studenci, szczególnie pokolenia młodego i średniego, preferują korzystanie z wersji cyfrowych. Powoływanie się na badania empiryczne tego zjawiska wydaje się zbyteczne – jest to bowiem powszechny obraz rzeczywistości akademickiej i całego rynku wydawniczego. W zaawansowanych technologicznie systemach informacyjnych podejście takie nie utrudnia badań lub, patrząc szerzej, procesów heurystycznych, a wręcz je ułatwia. Niewątpliwym zaletom lektury elektronicznej – a należą do nich dostęp do ogromnych baz danych, łatwość przeszukiwania tekstu, możliwość powiększania grafiki, szybkie generowanie cytatów itd. – nie towarzyszy wada, jaką jest niewielki stopień cyfrowego pokrycia materiału drukowanego. Niestety w przypadku prasy polskiej z lat 1945–1989 sytuacja jest inna. Mimo nakładów i wysiłków bibliotek w magazynach pozostaje ogrom tekstów w formie niezdygitalizowanej bądź też niedostępnej online poza czytelnią z uwagi na ochronę praw autorskich[1]. Sam poziom opracowania i wzbogacenia metadanymi cyfrowych zasobów prasowych jest także niski. Powodem tego jest niespójna struktura typograficzna i informacyjna dawnych gazet i czasopism. Mają one nieregularną i nieciągłą formę edytorską, która utrudnia opracowanie komputerowe, a w szczególności automatyczną segmentację – a w efekcie skuteczną recepcję w środowisku cyfrowym. Można wprawdzie nałożyć przeszukiwalne warstwy tekstowe na poziom graficzny – taki poziom opracowania występuje właśnie w większości zasobów – ale maszynowe wyodrębnienie artykułów jako spójnych całości jest w zasadzie niemożliwe. A właśnie takie całości, wzbogacone metadanymi, nadają się do dalszego przetwarzania metodami big data. Przypomnijmy, że

1 Kwestia praw autorskich jest z kilku powodów dyskusyjna w przypadku prasy okresu peerelowskiego. Periodyki ówczesne były częścią majątku społecznego – a więc wspólnego, natomiast dzisiejsi właściciele instytucjonalni dawnych tytułów nie istnieli w okresie peerelowskim. W warstwie edycyjnej warto zauważyć, że wiele artykułów nie było podpisanych albo podpisanych pseudonimami – de facto są to więc teksty anonimowe.

dzielenie tekstów na zapowiedzi z pierwszej strony i treść umieszczoną w środku numeru było w erze gutenbergowskiej bardzo częstą praktyką. Badania, jakie w ramach budowy korpusy ChronoPresss prowadziłem na dynamicznych układach typograficznych takich peerelowskich periodyków, jak „Express Wieczorny" czy „Przegląd Sportowy", ukazały bezradność technologii w tej kwestii – tylko człowiek był w stanie powiązać podzielone teksty poszczególnych artykułów w całości, a następnie przypisać im podstawowe metadane.

Ponadto częste w latach peerelowskich było publikowanie artykułów niesygnowanych, co utrudnia tworzenie indeksów autorów, powszechnie stosowanych w większości innych form wydawniczych (na przykład naukowych), a współcześnie także w prasie wydawanej cyfrowo. Z kolei nacisk na aktualność prasy skutkuje tym, że teksty są pisane szybko i schematycznie – rzadko kiedy reprezentują więc wartości wykraczające poza kontekst konkretnego czasu i miejsca. Do tego należy dodać niski społeczny prestiż gazet i części czasopism w stosunku do innych nośników treści (przede wszystkim książek). O ile książkę zwyczajowo eksponuje się w przestrzeni publicznej i prywatnej jako symbol wykształcenia, statusu społecznego i intelektu właściciela[2], o ile trzyma się ją przez lata jako swoistą wartość kolekcjonerską, powracając do niej raz po raz, o tyle gazeta, nazajutrz po wydaniu i przeczytaniu lub przejrzeniu, staje się niemal bezwartościowym materiałem wtórnym. Ma to związek z imperatywem systematycznego zastępowania starych numerów gazet i czasopism nowymi i wymusza regularną utylizację wydań wcześniejszych, charakterystyczną dla społeczeństwa konsumpcyjnego. Należy wreszcie dodać, że prasa okresu peerelowskiego wydawana była przez aparat propagandy państwa totalitarnego, pod kontrolą cenzury. Z dzisiejszej perspektywy należałoby więc raczej mówić o czytaniu z komentarzami, objaśnieniami i interpretacji. Cechy powyższe – nietrwałość, nierzetelność informacji, szablonowy skład i szybką dezaktualizację – wzmacnia ogólnie niska jakość nośnika, czyli tani papier i brak oprawy.

Jednak mniemanie, iż tysiące stron i miliony wyrazów wypełniających szpalty gazet, wyrzucanych każdego dnia przez prasy drukarskie, nie niosły ze sobą żadnej trwałej wartości informacyjnej, nie odpowiada prawdzie. Niezależnie od rejestracji jednostkowych faktów, potrzebnej w badaniach z zakresu na przykład historii społecznej lub antropologii kultury, prasa ujęta *en masse* jako korpus tekstów kryje w sobie treści ogólniejsze, wartościowe, nieulegające dezaktualizacji, dające uniwersalny obraz rzeczywistości konkretnego czasu i miejsca. Odnosi się to także do sytuacji państw totalitarnych, gdzie cenzura sprawuje kontrolę nad procesami komunikacji bieżącej, ale nie dostrzega treści i zjawisk ukrytych w tle. Otóż treści takie można odkrywać i badać narzędziami humanistyki cyfrowej w ramach dyscypliny określanej jako **kulturomika** (ang. *culturomics*). Jej przedmiotem nie są pojedyncze komunikaty tekstowe kierowane do czytelników, formułowane intencjonalnie przez autorów, lecz relacje i/lub procesy zachodzące w dłuższych

2 Wysoce pouczająca w tym względzie jest semiotyka sztuki. W malarstwie i rzeźbie książka czytana lub trzymana przez konkretne osoby była zawsze swoistym komunikatem, który otoczeniu uświadamiał, że przedstawiona postać reprezentuje wysoki poziom intelektualny (jest na przykład pisarzem lub uczonym), cieszy się prestiżem społecznym i/lub pełni odpowiedzialną funkcję.

okresach, ukryte w wielkim strumieniu danych[3]. Co do swej istoty – praktyki *big data* nie są dla przedstawicieli humanistyki i nauk społecznych zjawiskiem zupełnie nowym. Jednak z uwagi na wysoki koszt i nakład czasu do nadejścia ery cyfrowej były bardzo rzadkie.

Dotychczasowe badania

Omawiając literaturę przedmiotu, warto zauważyć, że natura humanistyki cyfrowej nie przystaje do obowiązującej w tradycji piśmiennictwa koncepcji „stanu badań". Ta z natury dynamiczna i osadzona w przestrzeni Internetu dyscyplina jest zaprzeczeniem „stanu", kojarzącego się ze stałością i stabilnością. Przedstawia sobą proces, zmianę i potencjalność wyników generowanych przez infrastruktury cyfrowe, a więc w jakimś sensie odcina się od humanistycznych praktyk poznawczych epoki druku. Właśnie wielość perspektyw i nacisk na swobodne przetwarzanie wielkich korpusów danych przez użytkownika zamiast podawania odbiorcy gotowego wyniku są szczególnie charakterystyczne dla cyfrowego medioznawstwa i znacznej części humanistyki cyfrowej (więcej na ten temat w pracy Pawłowski 2023a). Swoboda oznacza tutaj wybór „ścieżki heurystycznej" spośród dostępnych funkcjonalności danej aplikacji, dostosowany do potrzeb użytkownika. Nowość obecnej sytuacji objawia się także na poziomie metod przetwarzania danych i ich wizualizacji. W mniejszym stopniu zmienia się natomiast cel postępowania naukowego (odkrywanie ukrytych relacji i związków przyczynowo-skutkowych) oraz natura materiału empirycznego. Prasa drukowana minionych dziesięcioleci, obecnie dostępna coraz powszechniej w postaci elektronicznych korpusów tekstu, jest przykładem nowego podejścia do „starych" danych, które dzięki technologii zyskują jakby drugie życie.

Aby powyższe tezy zilustrować, wystarczy zauważyć, że najlepsze (na ogół także najaktualniejsze) kompendia digital humanities umieszczano najpierw w sieci, a dopiero później, jeśli zachodziła potrzeba, drukowano. Przykładem tego są liczne tutoriale, a także podręczniki, na przykład *Introduction To Digital Humanities*, dostępny początkowo na stronach WWW University of California, a dziś w innych repozytoriach[4] lub *A Companion to Digital Literary Studies* i *Companion to Digital Humanities* pod redakcją innych autorów, posadowiony w zasobach Alliance of Digital Humanities Organizations (por. Schreibman et al. 2013, Siemens, Schreibman 2008)[5]. Jeżeli natomiast poszukuje się informacji o projektach naukowych z zakresu humanistyki cyfrowej, obejmujących oczywiście także badania prasy, pomocą jest baza poszerzonych abstraktów Digital Humanities Conference, czyli konferencji organizowanych corocznie przez ADHO[6]. Zawsze należy

3 Otwarte pozostaje pytanie o relację między tymi warstwami informacji. Można jednak przyjąć, że autorzy lub czytelnicy, śledzący na bieżąco konkretne treści w prasie, nie są w stanie samodzielnie odtwarzać zjawisk w skali makro, odkrywanych przez algorytmy operujące na wielkich danych i długich okresach.

4 [On-line:] https://searchworks.stanford.edu/view/11649226 – 2.02.2024.

5 [On-line:] http://www.digitalhumanities.org/companionDLS – 2.02.2024.

6 [On-line:] http://digitalhumanities.org/dh-abstracts/search – 2.02.2024, warto jednak zauważyć, że baza ta nie obejmuje wszystkich projektów, prezentowanych na konferencjach afiliowanych przy ADHO.

jednak pamiętać, że sieć jest dynamiczna i zdarzają się przepływy zasobów – nawet między prestiżowymi i stabilnymi instytucjami, czego skutkiem są zmiany adresów URL.

Warto także dodać, że przełom cyfrowy w jakimś sensie ustanowił cezurę metodologiczną, symbolicznie odcinając i spychając w zapomnienie ogromny dorobek badaczy, posługujących się metodami typowymi dla ery gutenbergowskiej. Nie oznacza to, że zapomniana została, na przykład, metoda analizy zawartości – przydatna, ale w dzisiejszych warunkach kosztowna i obarczona subiektywizmem anotatorów (por. Pisarek 1972, 1983). Nie oznacza to także, że na skutek przełomu „obalono" metody wnioskowania statystycznego o całej populacji na podstawie małych prób. Zmiana polegała raczej na tym, że badacze zyskali dostęp do wielkich korpusów danych, które nie wymagają już odwoływania się do metod, opartych na indukcji, a część procesów zautomatyzowano. Na przykład wspomniana wyżej analiza zawartości ma swój częściowy odpowiednik w technice *topic modelling* i w generowaniu słów kluczy. Ponadto lingwiści i medioznawcy uświadomili sobie, że populacje tekstowe nie są tak jednorodne jak na przykład produkty przemysłowe czy dane biometryczne, dlatego wnioskowania uogólniające na temat na przykład struktury słownictwa w hipotetycznej „populacji ogólnej" można uznać za co najmniej dyskusyjne, ponieważ taki byt empiryczny jak populacja ogólna języka po prostu nie istnieje. Katalizatorem przełomu stały się także nowoczesne techniki NLP (Natural Language Processing), wykorzystujące podejście oparte na uczeniu maszynowym.

Wszystko to sprawia, że dyskutując nad kwestiami wcześniejszych badań tekstów prasowych, należałoby przede wszystkim wskazać publikacje przełomowe, ustanawiające nowe standardy pracy. Za tego rodzaju kamień milowy można uznać pracę specjalistów związanych z firmą Google, którzy poddali analizom ogrom tekstów z bazy Google Books (Michel et al. 2011), a także kolejne badania wielkich korpusów, wywodzące się z pogranicza humanistyki, inżynierii i medioznawstwa, realizowane wspólnie przez uczonych i specjalistów z koncernów medialnych (Flaounas et al. 2013). Należy także odnotować inną ważną zmianę, jaka zaszła w sposobie uprawiania nauki, polegającą na tym, że celem badaczy przestała być sama publikacja, a stało się nią stworzenie działającej infrastruktury cyfrowej. Współczesne infrastruktury są w jakimś sensie kontynuacją tych dawnych – bibliotek, archiwów, katalogów itd. – ale zawierają materiały cyfrowe wzbogacone niezwykle sprawnym „wkładem algorytmicznym", czyli narzędziami analizy danych (por. Pawłowski 2023). Warto na koniec nadmienić, że dzięki pracom literaturoznawcy Franco Morettiego zaczęto stosować pojęcie lektury wspomaganej technologią – tzw. *distant reading*, przeciwstawione czytaniu tradycyjnemu i powolnemu, określanemu jako *close reading*. O ile jednak *close reading* po prostu czytaniem jest, o tyle w przypadku *distant reading* chodzi o analizę wielkich danych, przetworzonych technologią cyfrową do postaci zagregowanej i dostępnej w formie wykresów, tabel i innych infografik. Przykładem infrastruktury cyfrowej spełniającej powyższe warunki jest właśnie korpus prasy polskiej ChronoPress, realizowany w ramach konsorcjum CLARIN-PL.

Charakterystyka korpusu[7]

Korpus Polskich Tekstów Prasowych ChronoPress zawiera dziś reprezentację strumienia prasy z okresu 1945–1966. W kolejnych etapach ChronoPress będzie rozbudowany do roku 1990, tak aby pokryć czas przynależności Polski (od 1952 Polskiej Rzeczpospolitej Ludowej) do bloku wschodniego. Docelowo ChronoPress ma w przyszłości objąć symboliczne sto lat dziejów Rzeczpospolitej, czyli okres 1918–2018. W związku z wysokim kosztem digitalizacji całych numerów, a także potencjalnym zagrożeniem roszczeniami z tytułu praw autorskich przyjęto zasadę pracy na reprezentatywnych próbkach traktowanych na prawach cytatu[8]. Podejście takie, uznające relewantność reprezentacji dużego zbioru przez mniejszy, jest zgodne z zasadami indukcji naukowej, jest też stosowane w badaniach humanistycznych i społecznych. Poprawna „mikrofotografia" wielkiego korpusu wystarcza, by ukazać najważniejsze relacje, tendencje i zjawiska zawarte w treści, nie jest natomiast wyszukiwarką służącą do przeglądania pełnych numerów. Pewien paradoks tej sytuacji, w świetle wcześniejszych stwierdzeń, polega na tym, że korpus ChronoPress jest reprezentacją w stosunku do całego strumienia informacji prasowej lat peerelowskich, ale jednocześnie spełnia już minimalne warunki, stawiane „wielkim danym" – co pokazują przedstawione dalej przykłady i scenariusze użycia.

Długość pojedynczej próbki wynosi w przybliżeniu 300 wyrazów tekstowych (tzw. słowoform lub tokenów), ale jest zmienna, ponieważ przyjęto zasadę nieucinania zdań, a nawet większych całości tematycznych (ułatwia to m.in. analizę składniową). Tak krótkie odcinki tekstu funkcjonują, jak wyżej wspomniano, na prawach cytatu, ponieważ stanowią fragmenty większych całości. Każdy rok jest reprezentowany przez ok. 6000 próbek, czyli średnio półtora miliona wyrazów tekstowych. Korpus w obecnej postaci obejmuje lata 1945–1966 i ma całkowitą objętość sięgającą czterdziestu milionów wyrazów tekstowych.

Jednym z istotnych etapów tworzenia korpusu był dobór periodyków. Zgodnie z dobrze znaną zasadą programowania i analizy danych *rubbish in, rubbish out*, o jakości zasobu nie decyduje wyłącznie jego objętość, ale także właściwa selekcja elementów składowych. Nawet wielki zbiór danych tekstowych da wyniki o niskiej wartości poznawczej, jeżeli dobór tekstów będzie w jakiś sposób niezrównoważony. W przypadku bazy korpusu ChronoPress selekcja tytułów opierała się na zasadach wywiedzionych z dominującej w tamtym czasie doktryny i praktyki marksizmu-leninizmu. Jako twórca projektu na wstępie przyjąłem, że rynek czytelniczy składał się z głównych klas społecznych, tak jak je definiowała doktryna maksizmu-leninizmu, oraz innych grup, które z powodów ideologicznych i pragmatycznych władza uważała za istotne. Konkretnym grupom czytelniczym przypisałem następnie tytuły prasy ogólnopolskiej i liczby próbek. Zgodnie z tą zasadą podstawowymi grupami odbiorczymi byli robotnicy i mieszkańcy wsi (w terminologii marksistowskiej klasa robotnicza i chłopi). Do tego należało uwzględnić tzw. inteligencję pracującą[9]. Warto przypomnieć, że przymiotnik „pracujący" w odniesieniu do inteligencji

7 Wszystkie dotychczasowe prace nad korpusem ChronoPress prowadzone były w ramach konsorcjum CLARIN-PL (http://clarin-pl.eu).

8 W rozumieniu Ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych, Dz.U. z 2019 r. poz. 1231.

9 W pewnym uproszczeniu inteligencja w Rosji Sowieckiej i w państwach Europy Środkowej okresu totalitarnego może być uważana za odpowiednik klasy średniej w państwach

miał w nowomowie PRL charakter „odwrotnego dowartościowania", co polegało na stygmatyzacji inteligencji przedwojennej, rzekomo odrzucającej nowy porządek i cieszącej się nieuzasadnionymi przywilejami klasy próżniaczej, a następnie przeciwstawieniu tej grupie inteligentów „nowych" – wychowanych w socjalizmie, aprobujących ten system, uczestniczących w jego budowie, pracujących podobnie jak chłop i robotnik[10].

Rysunek 1. Przykład zastosowania metody reprezentacyjnej do skanu gazety

CZERWONY SZTANDAR

ROZKAZ
Wodza Naczelnego
do Dowódcy Wojsk 1 Białoruskiego Frontu
Marszałka Związku Radzieckiego ŻUKOWA
do Naczelnika Sztabu Frontu
Generała-pułkownika MALININA

Wódz Naczelny
Marszałek Związku Radzieckiego J. STALIN

Lwów będzie wielkim przemysłowym centrum

Radzieckie Biuro Informacyjne

Oprócz tych trzech dominujących grup wyróżniono gazety i czasopisma adresowane do młodzieży, wojska, kobiet i katolików. W przypadku młodzieży był to odbiorca potencjalnie najbardziej podatny na indoktrynację i przyszłościowy. Wojsko stanowiło niezbędne zaplecze siłowe nowej władzy i dysponowało dobrze finansowanym aparatem medialnym (przede wszystkim „Polska Zbrojna", a następnie

demokratycznych. Jednak tak pojmowana inteligencja jest wyróżniana na podstawie wykształcenia i domniemanego poczucia odpowiedzialności za kraj, podczas gdy klasa średnia definiowana jest przede wszystkim na podstawie kryteriów ekonomicznych.

10 Swoistym komentarzem do powyższego stwierdzenia może być fakt, iż w „Roczniku Statystycznym" z 1956 roku inteligencja jako kategoria zawodowa lub grupa społeczna w ogóle nie występuje, natomiast znaleźć można wiele odwołań do chłopstwa i robotników.

„Żołnierz Wolności"). Kobiety jako segment czytelniczy były traktowane w sposób szczególny w związku z tym, że ich emancypacja była jednym z bardzo niewielu postulatów socjalizmu dziewiętnastowiecznego, które dyktatorskie reżimy podporządkowane ZSRR utrzymały w swoich programach po 1945 roku. Wreszcie katolicy reprezentowali znakomitą większość społeczeństwa polskiego[11] i ze względów pragmatycznych musieli być tolerowani. Objawiało się to zgodą na publikowanie „Tygodnika Powszechnego" – jedynego periodyku „niesocjalistycznego" o zasięgu teoretycznie ogólnokrajowym, dystrybuowanego jednak głównie kanałami kościelnymi, prywatnymi lub pocztą. Lata następujące po pierwszym przełomie politycznym w ZSRR i innych państwach bloku wschodniego, oznaczającym koniec ortodoksyjnego stalinizmu (1956, tzw. odwilż), wzbogaciły korpus o próbki prasy popularnej, na którą zawsze jest duże zapotrzebowanie, sportowej i kolorowej (kobiecej, podróżniczej).

Rysunek 2. Przykład oznaczonej próbki w korpusie

```
<?xml v"rsi"n="1.0"?>
<sample>
    <title_newspaper>Pokolenie</title_newspaper>
    <title_article> W kopalniach Wałbrzycha </title_article>
<authors>
    <author>Jerzy Wroński</author>
</authors>
<language>pl</language>
<style>press</style>
<year>1950</year>
<month>01</month>
<day>08</day>
<date>1950-01-08</date>
<period>w</period>
<status>1_obieg</status>
<support>paper</support>
<exposition>1</exposition>
<text>
<![CDATA[ Poszli od razu za wezwaniem. Ich kolega Ogrodowczyk wstał na odprawie aktywu robotniczego i powiedział: "Najbardziej doświadczeni młodzi górnicy powinni pokazać swoim kolegom, jak należy pracować". Dziesięciu najlepszych górników, młodszych rębaczy: [...] Przyszli do kopalni wałbrzyskich z dalekich sztolni północnej Francji, z zielonej kotliny Francji, z górniczych osad Belgii, z chałup wiejskich rzeszowskiego. [...] Ogrodowczyk, Drzazga i Zając tłumaczą, jak potrafią — aby zwyciężyła zespołowa praca. ]]>
</text>
</sample>
```

11 Jak pisze Bartłomiej Noszczak: „Polska stała się państwem monoetnicznym i monowyznaniowym, z dominującą liczbą osób wyznania rzymskokatolickiego, szacowaną po 1945 r. na około 23 mln, co stanowiło 97,7 proc. ludności" (Noszczak s.140).

Próbki są anotowane metainformacją w formacie XML, a sam tekst jest przetworzony parserem morfosyntaktycznym WCRF (Radziszewski 2013), który pozwala na sprowadzenie form odmienianych do lematów, co ułatwia wyszukiwanie (polski jest językiem silnie fleksyjnym). Niestety stosowany parser nie ma funkcji dezambiguacji semantycznej, co oznacza, że na etapie wyszukiwania nie odróżnia się form homograficznych. Dlatego nie jest na przykład możliwe rozróżnienie leksemów „Nysa" w znaczeniu nazwy miasta, dwóch rzek (Nysa Kłodzka i Łużycka) oraz nazwy samochodu. Nie zawsze rozróżniane są także przypadki homografii typu Niemiec (dopełniacz od Niemcy i mieszkaniec tegoż państwa). Parser dobiera jednak klasy gramatyczne, co w wielu przypadkach ułatwia wyszukiwanie (na przykład wyszukiwany leksem „czerwony" można oznaczyć jako przymiotnik lub rzeczownik). ChronoPress oferuje użytkownikowi możliwość przeglądania próbek, generowania konkordancji i list frekwencyjnych. Z punktu widzenia eksploracji danych najważniejsze są trzy moduły: szeregów czasowych, profilowania leksemów przez kontekst, mierzenia podobieństwa leksemów oraz generowania mapy[12]. Zaletą funkcjonalności szeregów czasowych jest możliwość przechodzenia od grafiki do tekstu w postaci konkordancji leksemu, powiązanej z danym punktem na osi czasu.

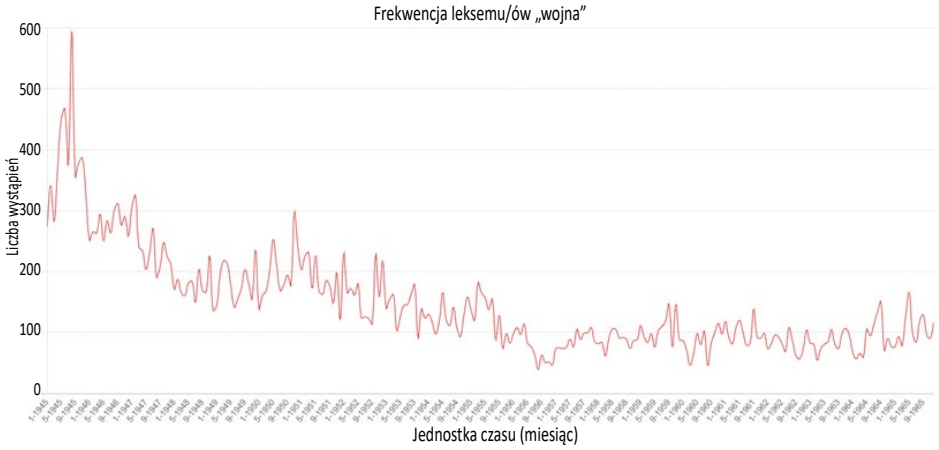
O ile technologia jest najbardziej dynamicznym elementem bazy i całego systemu, będzie więc podlegać częstym zmianom i ulepszeniom, sprawą dyskusyjną może być sam pomysł prezentowania prasy z okresu, kiedy Polska była państwem totalitarnym i niesuwerennym. Można w szczególności zadać pytanie, co wyrażają generowane wykresy i inne dane pochodne, skoro wolność prasy nie istniała i każde słowo w przestrzeni publicznej podlegało kontroli cenzury. Otóż w przypadku lektury tradycyjnej, określonej wcześniej jako *close reading*, tekstów prasowych praktycznie nigdy, nawet w państwach demokratycznych, nie należy utożsamiać z prawdą naukową. Natomiast w państwie totalitarnym teksty takie są dodatkowo swoistą kreacją propagandową, mówiącą wiele o intencjach nadawcy, ale mało o opisywanej rzeczywistości. Zupełnie inaczej wygląda sytuacja analizy danych, jaką umożliwia infrastruktura cyfrowa (taką formę obcowania z tekstem można określić jako *distant reading*). Dane o frekwencji i dystrybucji leksemów ujawniają relacje niewidoczne dla cenzora i ukazujące w zupełnie innym świetle uniwersum medialne epoki. Oczywiście ich interpretacja wymaga, jak każdy proces naukowy, pewnej znajomości realiów danego okresu historycznego. Poniżej przedstawiam przykładowe wyniki analiz prowadzonych na infrastrukturze ChronoPress.

A. Analiza szeregów czasowych

Z systemu wygenerowano szereg czasowy leksemu „wojna" (Rys. 3). Jak widać, krzywa pokazuje wyraźnie, że temat ten był w polskiej prasie bardzo często przywoływany tuż po zakończeniu drugiej wojny światowej. Tendencję tę zmienił dopiero przełom roku 1956. Nieregularny kształt krzywej, a w szczególności jej odbicie w latach 1950–1951, wskazuje, że pojawił się temat nowej wojny lub wojny w innych częściach świata (w tym przypadku w Wietnamie). Krzywa szeregu czasowego ukazuje więc pewien proces, który badacz mediów lub historii współczesnej może w swojej dyscyplinie zinterpretować.

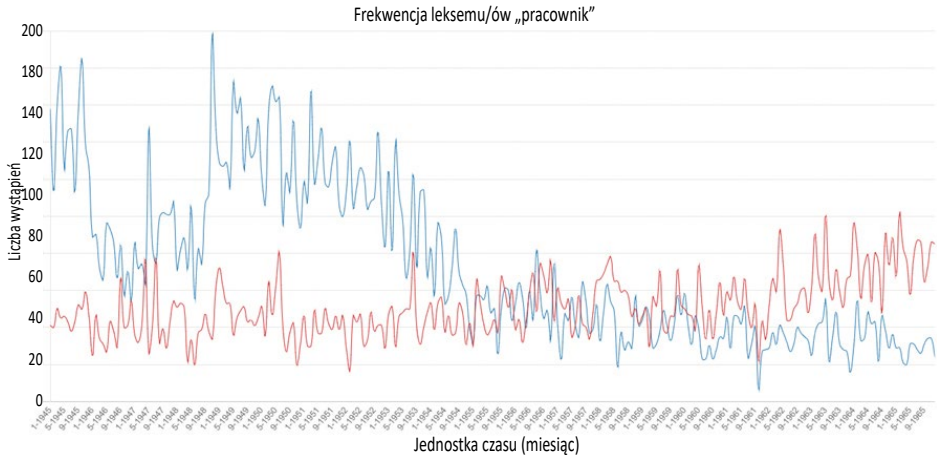
12 W celu rozpoznania bytów nazwanych zastosowano moduł Liner 2 (Marcińczuk et al. 2013). Moduł mapy jest w chwili składania tego tekstu w przygotowaniu.

Rysunek 3. Szereg czasowy leksemu „wojna” w prasie polskiej (1945–1965)



Inny przypadek analizy trendów pokazuje, jak zmieniał się dyskurs oficjalny w okresie powojennym (Rys. 4). Porównanie leksemów „robotnik” i „pracownik” wskazuje na stopniowe odchodzenie od marksistowskiej terminologii w odniesieniu do świata pracy. Leksem „robotnik” po roku 1948 (ma wtedy miejsce likwidacja resztek swobód demokratycznych i systemu wielopartyjnego, powstaje PZPR i „Trybuna Ludu”) uzyskuje bardzo wysokie frekwencje, ale w miarę upływu lat postrzegany jest jako coraz bardziej zideologizowany i nieatrakcyjny. Rośnie natomiast średnia frekwencja leksemu „pracownik”, bardziej uniwersalnego, lepiej dostosowanego do zmian w gospodarce i nieobciążonego skojarzeniami z doktryną marksistowską.

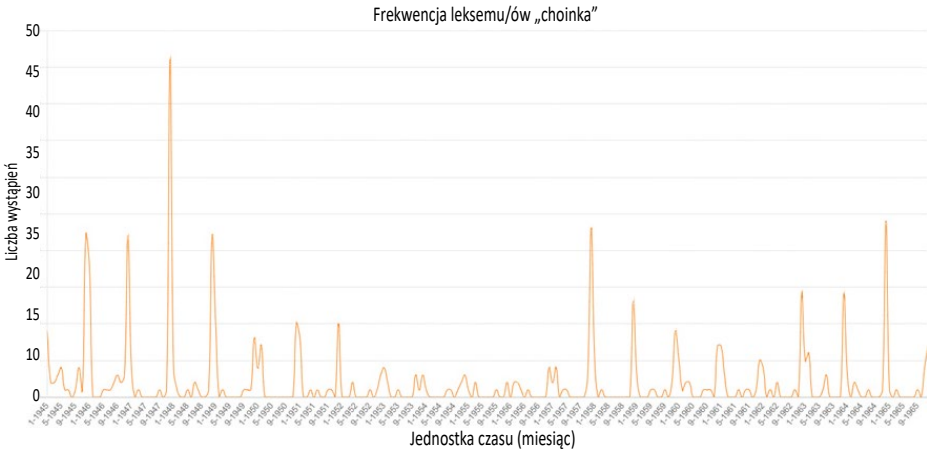
Rysunek 4. Szereg czasowy leksemów „robotnik” (kolor niebieski) i „pracownik” (kolor czerwony) w prasie polskiej (1945–1965)



Kolejnemu badaniu poddano ewolucję częstości sumy leksemów „wigilia” i „choinka” (Rys. 5). Obraz, jaki się tutaj rysuje, ujawnia dwa procesy, przy czym

pierwszy ma podłoże ideologiczne, drugi kulturowe. Oba atrybuty świąt religijnych osiągają znaczące frekwencje tylko w miesiącach grudniowych, są więc typowym świadectwem rytmu kulturowego. Ale jednocześnie ich obecność niemal zanika w czasach stalinizmu (1950–1955), kiedy władze postawiły sobie za cel rozprawienie się z polską religijnością. Cecha ta ma więc podłoże ideologiczne i jest jednorazowym efektem czynnika zewnętrznego, który zaburzył naturalny rytm kultury.

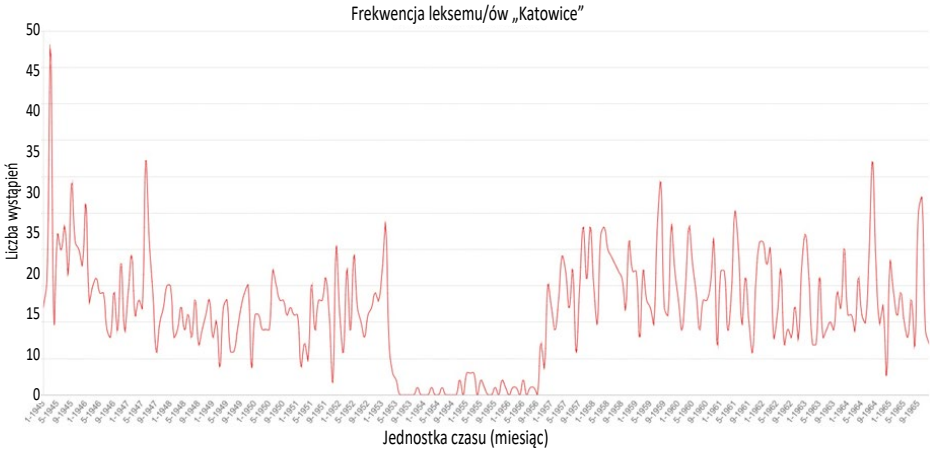
Rysunek 5. Szereg czasowy sumy frekwencji leksemów „wigilia" i „choinka" w prasie polskiej (1945–1965)



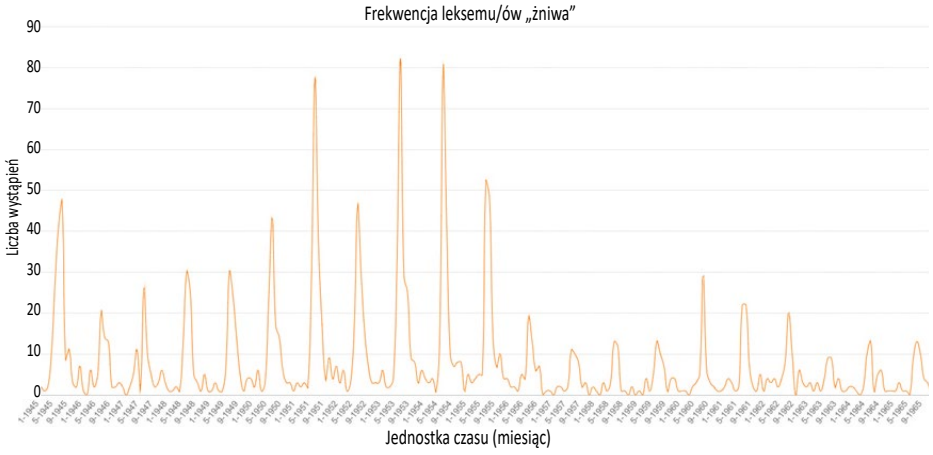
Następna analiza dotyczy frekwencji nazwy miasta Katowice, która na kilka lat zniknęła z obiegu medialnego, ponieważ zmieniono ją na Stalinogród (Rys. 6). Zasada tworzenia patronimicznych nazw miast, honorujących wielkich komunistów, była rozpowszechniona w ZSRR, a Sowieci próbowali ją także wprowadzić w satelickich państwach bloku wschodniego (Leningrad, Kaliningrad, Stalingrad, bułgarskie Stalino, węgierski Sztálinváros, Karl-Marx-Stadt, Titograd – por. Belei, Sojka-Masztalerz 2018). Wykres ukazuje właśnie tę przerwę w normalnej egzystencji Katowic w dyskursie medialnym. Nieliczne użycia „normalnej" nazwy w latach 1953–1956 pojawiają się jedynie w kontekstach historycznych lub jako tytuły książek wydanych przed rokiem 1953.

Czwartej analizie poddano leksem „żniwa", który wyraźnie ukazuje tzw. rytmy astronomiczne, porządkujące ludzką aktywność w kontekście rolniczym (Rys. 7). Ekstrema są tutaj regularne, pojawiają się mniej więcej w tych samych miesiącach roku. W szczególności widzimy je latem oraz, z o wiele mniejszym natężeniem, jesienią, kiedy odbywał się zbiór buraków lub innych płodów rolnych. Widoczna na Rys. 7 linia szeregu czasowego przypomina swoim kształtem piłę i jest charakterystyczna dla mediatyzacji większości zjawisk osadzonych w czasie astronomicznym (prace na roli, choroby sezonowe itp.) lub kulturowym (święta, rocznice – por. Rys. 5). Spadek intensywności ekstremów krzywej leksemu „żniwa" po roku 1956 świadczy o złagodzeniu przez propagandę PRL modelu kampanijnego. Łatwo sprawdzić, że bardzo podobnie zachowują w czasie się inne leksemy związane z cyklami natury lub kultury (w tym także polityki – por. „plenum", „zjazd" lub „wybory").

Rysunek 6. Szereg czasowy nazwy „Katowice" w prasie polskiej (1945–1965)



Rysunek 7. Szereg czasowy nazwy leksemu „żniwa" w prasie polskiej (1945–1965)



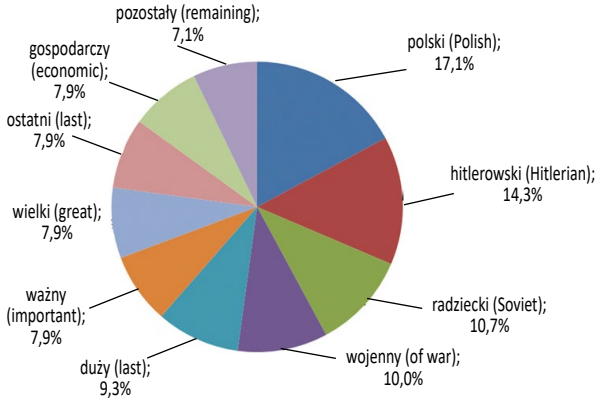
B. Profilowanie pojęć

Profilowanie leksemu polega na analizie jego bezpośredniego lub pośredniego otoczenia w wielu tekstach. Jeżeli na przykład w polskich tekstach z okresu socjalizmu analizuje się rzeczownik „partia", bardzo prawdopodobne w jego otoczeniu będą leksemy „socjalistyczny" i „robotniczy". Profilować można albo w oparciu o proste frekwencje leksemów współwystępujących, albo wykorzystując bardziej złożone miary statystyczne (np. log-likelihood ratio). Te ostanie wskazują na odstępstwo od losowego prawdopodobieństwa pojawienia się danego leksemu w otoczeniu innego, przy założeniu równomiernego rozkładu jednostek w linii tekstu. Współwystępowania zbyt częste w stosunku do statystycznie równomiernych są traktowane jako znaczące. Ciekawym problemem jest ewolucja profili leksykalnych w czasie. Właśnie taki eksperyment można przeprowadzić na leksemie „Niemcy", porównując dwa wybrane okresy (1945–1946 i 1956–1958). Okazuje się, że nawet przy zastosowaniu liczb bezwzględnych różnice między tymi okresami są

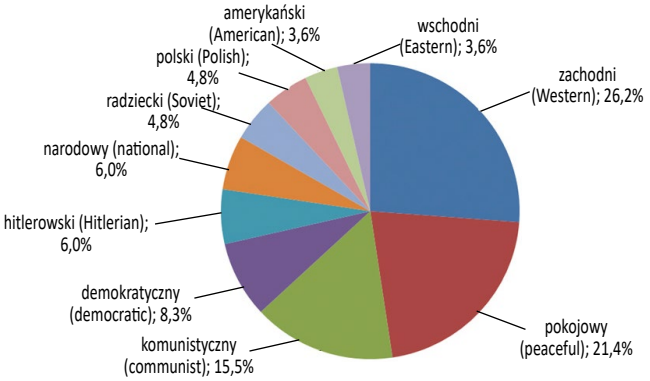
wyraźne. Wykres 8 pokazuje, że Niemcy w latach powojennych postrzegane były przez pryzmat wojny, o czym świadczą przymiotniki „hitlerowski" i „radziecki". Przymiotnik „polski" występuje w otoczeniu leksemu „Niemcy" w związku z tym, że druga wojna światowa była w propagandzie konceptualizowana jako starcie przede wszystkim polsko-niemieckie.

Sytuacja jest inna w latach pięćdziesiątych (Rys. 9). Kontekst wojenny, chociaż dalej obecny, zostaje zastąpiony ofensywą propagandową, której celem jest wspieranie rozbrojenia i demilitaryzacji. Dominuje więc przymiotnik „pokojowy", natomiast „hitlerowski" przesuwa się na dalsze miejsce. Ponadto, co bardzo istotne, pojawiają się ślady podziału Niemiec na części wschodnią i zachodnią. Ponieważ nie było możliwe automatyczne rozróżnienie Niemiec Wschodnich i Zachodnich, należy się domyślać, że najprawdopodobniej przymiotniki „demokratyczny" i „komunistyczny" były związane właśnie z dawną sowiecką strefą okupacyjną, która od roku 1949 istniała jako NRD.

Rysunek 8. Profil semantyczny leksemu „Niemcy" w latach 1945–1947[13]



Rysunek 9. Profil semantyczny leksemu „Niemcy" w latach 1956–1958[14]



13 Grafikę sporządzono na podstawie danych wyeksportowanych z ChronoPressu.

14 Grafikę sporządzono na podstawie danych wyeksportowanych z ChronoPressu.

Wnioski

Dotychczasowe doświadczenia retrospektywnych badań prasy z użyciem korpusu ChronoPress są pozytywne i obiecujące. Potencjał poznawczy zdigitalizowanej prasy polskiej z lat 1945–1966 (a w perspektywie okresu 1918–2018), eksplorowanej przy pomocy narzędzi humanistyki cyfrowej jest ogromny, o czym świadczą powyższe przykłady, a także pozytywne reakcje specjalistów z zakresu historii współczesnej, antropologii kultury i lingwistyki, którym wyniki badań opartych na infrastrukturze ChronoPressu były prezentowane podczas konferencji lub warsztatów. Badania te pokazują nowe oblicze i potencjał naukowy zasobów polskiej prasy drukowanej, o ile tylko zostanie ona włączona do infrastruktury korpusu ChronoPress. Korpus ChronoPress nie jest biblioteką cyfrową i z oczywistych względów nie spełni oczekiwań uzasadnionych w przypadku pracy na pełnych wersjach periodyków. Jednak podejście zgodne z metodyką humanistyki cyfrowej pozwala na agregowanie danych i tworzenie uogólnień, które nie byłyby osiągalne przy korzystaniu ze skanów z warstwą tekstową. W kontekście wielkich procesów digitalizacji (m.in. projektów Patrimonium[15] i Omnis[16]) doświadczenia ChronoPressu są więc bardzo cenne: pozwalają na przygotowanie przez biblioteki ewentualnej koncepcji nowych, ale już przetestowanych narzędzi eksploracji danych.

Przedstawione wyniki potwierdzają relewantność wykorzystania wielkich zbiorów danych tekstowych i narzędzi NLP w badaniach prasy. Badania takie, zaliczane do nurtu kulturomiki i humanistyki cyfrowej, mają znaczny potencjał poznawczy, wynikający z tego, że ujawniają ukryte pod warstwą sztampowych tekstów prasy codziennej zjawiska, trendy rozwojowe, powiązania zjawisk i hierarchie, których obecność badacze mogli przeczuwać, ale bez zasobów i narzędzi cyfrowych nie mogli badać metodami empirycznymi. Jednym z efektów działania portalu ChronoPress jest potwierdzenie ważności tzw. teorii długiego trwania (*longue durée*), która kojarzona jest z grupą badaczy skupionych wokół francuskiego czasopisma „Annales" (Marc Bloch, Lucien Febvre, później Immanuel Wallerstein). ChronoPress nie pokazuje wprawdzie danych reprezentujących setki lat, ale ujawnia dane leksykostatystyczne, które z dużym prawdopodobieństwem można uznać za część takich trendów. Dotychczasowe badania w nurcie szkoły Annales prowadzone były na danych typowo statystycznych (pomiary antropometryczne, wskaźniki gospodarcze itd.), jednak materiał tekstowy, o ile zostanie należycie dobrany i przetworzony, ukazuje podobne procesy. Teksty są przecież swoistym odzwierciedleniem procesów ekonomicznych, politycznych czy kulturowych – także w państwach totalitarnych.

Na zakończenie warto wspomnieć o perspektywach rozwoju portalu ChronoPress. Kolejne lata powinny przynieść jego rozwój horyzontalny, prowadzący do pokrycia coraz dłuższego odcinka czasowego. Proces ten jest powolny i kosztowny, ale stosunkowo przewidywalny. Drugi obszar rozwoju to wzbogacenie portalu o dane nowe pod względem genologicznym i funkcjonalnym (m.in. transkrybowane ścieżki głosowe programów informacyjnych). Zadanie to jest wbrew pozorom bardziej skomplikowane i kosztowniejsze od dodawania tekstów prasowych, ponieważ wymaga m.in. konwersji dźwięku na tekst i korekty. Niezależnie od powyższych

15 [On-line:] https://www.bn.org.pl/projekty/patrimonium – 2.02.2024.

16 [On-line:] https://www.bn.org.pl/projekty/omnis – 2.02.2024.

ograniczeń korpus ChronoPress i jego funkcjonalności można uznać za nową jakość w metodologii badań prasy.

Bibliografia

A Companion to Digital Humanities, red. S. Schreibman, R.G. Siemens, J. Unsworth John, John Wiley & Sons, Blackwell Companions to Literature and Culture, New York 2008.

A Companion to Digital Literary Studies, red. R.G. Siemens, S. Schreibman Susan, John Wiley, Blackwell Companions to Literature and Culture, New York 2013.

Belej O., Sojka-Masztalerz H., *Ojkonimia ukraińska w kontekście procesów dekomunizacyjnych Europy środkowo-wschodniej końca XX – początku XXI wieku*, „Onomastica" 2018, nr LXII, s. 259–272.

Flaounas I., Ali O., Lansdall-Welfare T., De Bie T., Mosdell N., Lewis J., Cristianini N., *Research Methods in the Age of Digital Journalism*, "Digital Journalism" 2013, nr 1:1, s. 102–116.

Marcińczuk M., Kocoń J., Janicki M., *Liner2 – A Customizable Framework for Proper Names Recognition for Polish*, [w:] *Intelligent Tools for Building a Scientific Information Platform*, red. R. Bembenik, Ł. Skonieczny, H. Rybiński, M. Niezgódka, Springer, Berlin–Heidelberg 2013, s. 231–253.

Michel J.-B., Shen Y.K., Aiden A.P., Veres A., Gray M.K., The Google Books Team, Pickett J.P., Hoiberg D., Clancy D., Norvig P., Orwant J., Pinker S., Nowak M.A., Lieberman A.E., *Quantitative Analysis of Culture Using Millions of Digitized Books*, "Science" 2011, nr 331, s. 176–182.

Noszczak B., *Polityka państwa wobec Kościoła rzymskokatolickiego w Polsce w latach 1944–1956*, [w:] *Polski wiek XX*, t. 3, red. K. Persak, P. Machcewicz, Bellona i Muzeum Historii Polski, Warszawa 2010, s. 137–166.

Od Gutenberga do Zuckerberga. Wstęp do humanistyki cyfrowej, red. A. Pawłowski, Universitas, Kraków 2023.

Pawłowski A., *Humanistyka cyfrowa: geneza, zakres, cele i perspektywy*, [w:] idem, *Od Gutenberga do Zuckerberga. Wstęp do humanistyki cyfrowej*, Universitas, Kraków 2023a. Pisarek W., *Frekwencja wyrazów w prasie: wiadomości – komentarze – reportaże*, Ośrodek Badań Prasoznawczych, Kraków 1972.

Pisarek W., *Analiza zawartości prasy*, Ośrodek Badań Prasoznawczych, Kraków 1983.

Radziszewski A., *A Tiered CRF Tagger for Polish*, [w:] *Intelligent Tools for Building a Scientific Information Platform*, red. R. Bembenik, Ł. Skonieczny, H. Rybiński, M. Niezgódka, Springer, Berlin–Heidelberg 2013, s. 215–230.

Streszczenie

Przedmiotem artykułu jest omówienie nowych metod badań prasy drukowanej, wykorzystujących narzędzia humanistyki cyfrowej. Ich celem jest powtórne wprowadzenie wartościowych poznawczo zasobów prasy cyfrowej do obiegu naukowego. Jako jedno z rozwiązań praktycznych artykuł proponuje korpus ChronoPress, zawierający reprezentację prasy PRL i bogaty wybór narzędzi analitycznych. Artykuł omawia genezę, zasady konstrukcji oraz główne cechy ChronoPressu. Przedstawia też kilka scenariuszy użycia tej infrastruktury, ukazując jej możliwości eksploracyjne. Szczególnie ważne są tutaj leksykalne szeregi czasowe, które pozwalają odkrywać dynamikę procesów rozwijających się w czasie, a w aplikacji

ChronoPress widoczne są w postaci interaktywnych wykresów, uruchamiających generator konkordancji leksemów w konkretnych punktach osi chronologicznej. Korpus jest anotowany morfosyntaktycznie i semantycznie (kategorie podstawowe), obejmuje obecnie lata 1945–1966.

Słowa kluczowe: prasa, humanistyka cyfrowa, szeregi czasowe, PRL, ChronoPress

ChronoPress Polish press corpus as infrastructure and tool of media studies

Abstract

The aim of this article was to discuss new methods of researching the printed press using digital humanities tools. The goal of these methods is to reintroduce the cognitively valuable resources of the digital press into scientific circulation. As one practical solution, the authors of this article proposed the ChronoPress corpus, which includes a representation press of the communist period in Poland and a rich selection of analytical tools. The genesis, design principles, and main features of ChronoPress were discussed. Moreover, several scenarios for the use of this infrastructure demonstrating its exploratory capabilities were presented. Particularly important here are lexical time series, which allow us to discover the dynamics of processes developing over time. In the ChronoPress application, it is also possible to generate concordances by clicking on a point in the time series graph. The corpus is annotated morphosyntactically and semantically (basic categories) and currently covers the years 1945–1966.

Keywords: press, digital humanities, time series, Poland under communism, ChronoPress