Academic Journal of Modern Philology

Jakub Maciej Łubocki
Uniwersytet Wrocławski

ISSN 2299–7164
Vol. 13 (2021)
Special Issue
s. 191–201

Google Books Ngram Viewer jako narzędzie weryfikacji istnienia ekwiwalentów obcego terminu (na przykładzie rosyjskiego terminu *библиографоведение*)

Google Books Ngram Viewer as a Tool to Verify the Existence of Equivalents of a Foreign Term (on the Basis of the Russian Term *библиографоведение*)

Abstract

In the 1970s in the Soviet Union (by the standard: GOST 16448-70 *Bibliografiâ. Terminy i opredeleniâ*, Moskva 1971) the top-down term 'библиографоведение' was introduced to denote a scientific discipline that is part of the complex of bibliological disciplines, researching and developing issues of the theory, history and methodology of bibliographies as well as methodology, technology and organization of bibliographic activity. After the term became popular in the Soviet Union, there were suggestions in several sources in the years 1974–1999 that its equivalents also began to function in other languages – Polish, English, German, French. The analysis of data from Google Books Ngram Viewer, carried out for the other article (see: Jakub Maciej Łubocki, *O problemie wieloznaczności terminu 'bibliografia' kolejny raz – o terminie 'bibliografoznawstwo' po raz pierwszy*, „Roczniki Biblioteczne", 2018 (R. 62), s. 135–178; doi.org/10.19195/0080-3626.62.9), allowed for a proper (mainly negative) assessment of these suggestions. This article highlights the advantages and disadvantages of conducting an analysis with the use of this tool. It also allows us to shed light on the problem of – sometimes too careless – equivalence of terms (and thus improper equivalence) and the possibility of its verification in the absence of data in the available dictionaries.

Keywords: bibliography, bibliografoznawstwo, terminology analysis, terminological equivalence, Google Books Ngram Viewer

Niniejszy krótki przyczynek powstał na marginesie pracy (Łubocki 2018), dla potrzeb której autor tych słów zapoznał się z funkcjonalnościami Google Books Ngram Viewer (NV) pod kątem ich przydatności w analizie terminologicznej. Celem tamtej pracy były rozważania nad możliwością usunięcia wieloznaczności terminu *bibliografia* poprzez wprowadzenie neoterminu *bibliografoznawstwo* do polskiego systemu

doi:10.34616/ajmp.2021.13

terminologicznego, analogicznie do sytuacji z lat 70. XX wieku na terenie Związku Radzieckiego, gdzie taki neologizm wprowadzono, a następnie w pełni zaakceptowano. Referowanie sporu, jaki wówczas rozgorzał za wschodnią granicą Polski, przyniosło między innymi pytanie o występowanie ekwiwalentów tego terminu w innych językach, które jakoby już wówczas miały się pojawiać. Brak przypisów poświadczających źródło tych rewelacji utrudnił odpowiedź na pytanie, co dało asumpt ich autorom do takich twierdzeń. Analiza leksykograficzna także ich nie potwierdziła. Stąd autor rozważań zwrócił uwagę na nowe możliwości, jakie potencjalnie daje NV, ale jednocześnie zapoznał się pojawiającymi się już w literaturze zastrzeżeniami przed pochopnym wyciąganiem wniosków na tak uzyskanej podstawie. Ponieważ w literaturze prace obiektywnie oceniające NV nie są jeszcze zbyt obfite, a prac na temat wykorzystania NV w analizie terminologicznej nie stwierdzono w ogóle – postanowiono przedstawić zwięźle wnioski, jakie płyną z doświadczenia poszukiwania hipotetycznych ekwiwalentów terminu *библиографоведение* w innych językach w korpusach NV. Nie jest to jednak pierwszy artykuł, który eksponuje problemy w naukowym wyzyskiwaniu zasobów NV (zob. np.: Gooding 2012; Pechenick, Danforth & Dodds 2015; Pettit 2016).

*

1.07.1971 roku[1] w Związku Radzieckim ustanowiono normę GOST 16448-70 *Bibliografia. Terminy i definicje* (GOST 16788-70 1970; więcej o samej normie – Rešetinskij 1970) zawierającą 77 terminów. W rozdziale 1. normy – *Pojęcia ogólne* – umieszczono 4 pojęcia: 'bibliografia', 'produkcja bibliograficzna', 'działalność bibliograficzna' i 'bibliografoznawstwo'. Choć sam termin *bibliografoznawstwo* w rozumieniu *teoria bibliografii* został zaproponowany już w 1948 roku (w: Markov 1948), to dopiero rok 1970 wprowadził go do powszechnej świadomości. Norma ta była pierwszą, jak się później okazało skuteczną, próbą przezwyciężenia problemu wieloznaczności terminu *bibliografia*. Świadomość tego problemu sięgała w Związku Radzieckim co najmniej 1936 roku i dotychczasowe próby rozwiązania go – np. w słownikach (Šamurin 1958; Čavdarov *et al.* 1966) – nie przyniosły oczekiwanych efektów (zob.: Rešetinskij 1969). Norma *Bibliografia. Terminy i definicje* wywołała burzliwy spór oraz rezon poza granicami Związku Radzieckiego (zob.: Eychlerowa 1978), jednak upływ czasu pokazał, że jej rozwiązania zostały zaakceptowane i obecnie są powszechnie stosowane: termin *библиографоведение* odnajdujemy w artykułach hasłowych słowników oraz encyklopedii ogólnych i specjalnych, tytułach czasopism czy nagłówkach klasyfikacji. Wyodrębnienie terminu ułatwiło także wyodrębnienie się „teorii bibliografii" jako samodzielnej dyscypliny naukowej, stąd rosyjskie *библиографоведение*, ukraińskie *бібліографознавство* czy białoruskie *бібліяграфазнаўства* odnajdujemy także w tytułach słowników, bibliografii, podręczników i monografii poświęconych wyłącznie temu zakresowi. Miejsce to trwale sankcjonują także urzędowe wykazy dziedzin i dyscyplin naukowych Rosji, Ukrainy i Białorusi (więcej na ten temat zob.: Łubocki 2018: 144–153).

Pojawienie się nowego terminu w jednym języku zawsze wywołuje problem jego przekładu na inny język. W tym przypadku mamy do czynienia z osobliwą odmianą tego problemu: otóż co jakiś czas pojawiały się informacje, że termin *библиографоведение* jest już terminem znanym w innych językach. Bywały o to podejrzewane język polski, angielski, niemiecki i/lub francuski. Poniżej przedstawiono trzy ilustracje

1 Norma ta została wprawdzie ustanowiona dopiero w połowie 1971 roku, jednak Komitet Norm, Miar i Przyrządów Mierniczych przy Radzie Ministrów ZSRR (ros. Комитет стандартов, мер и измерительных приборов при Совете Министров СССР) zatwierdził jej projekt już 26 grudnia 1969 roku, a gotową normę opublikowano w roku 1970.

tego przypadku – z roku 1974 (Barsuk 1974: 18; ryc. 1), 1984 (Migoń 1984: 122; ryc. 2) i 1999 (GOST 7.0-99 1999: pkt 3.1.38; ryc. 3).

«Библиографоведение» фигурирует как одно из ключевых понятий в подготавливаемой к печати энциклопедии книговедения. Не сомневаемся, что в специальной литературе тех стран, где еще нет соответствующего эквивалента, он очень скоро появится (эпизодически мы уже встречаемся в польских, немецких, английских изданиях с терминами «bibliografoznawstwo»; «Bibliographiewissenschaft», «— wesen», «— lehre»; «bibliographical science» или «science of bibliography» и т. д.).

Rycina 1. Fragment w języku rosyjskim z 1974 roku sugerujący występowanie ekwiwalentów terminu *библиографоведение* w języku polskim, niemieckim i angielskim
Źródło: Barsuk 1974: 18.
Tłum. własne: „Bibliografoznawstwo pojawia się jako jedno z kluczowych pojęć w przygotowywanej do druku encyklopedii księgoznawczej. Nie wątpię, że w literaturze specjalistycznej tych krajów, gdzie jeszcze nie ma [dla niego – JMŁ] odpowiedniego ekwiwalentu, wkrótce on się pojawi (sporadycznie już spotykamy się w polskich, niemieckich, angielskich wydawnictwach z terminami «bibliografoznawstwo»; «Bibliographiewissenschaft», «-wesen», «-lehre»; «bibliographical science» czy «science of bibliography» itd.)".

bibliotekoznawstwo i bibliografię (właściwie 'bibliografoznawstwo'; por. angielski termin 'bibliographology', niemiecki 'Bibliographienkunde' i rosyjski 'bibliografowiedienije'[31]) wraz z innymi formami informacji o książce.

Rycina 2. Fragment w języku polskim z 1984 roku sugerujący występowanie ekwiwalentów terminu *библиографоведение* w języku angielskim, niemieckim i rosyjskim
Źródło: Migoń 1984: 122.

3.1.38 библиографоведение: Научная дисциплина, изучающая теорию, историю, методологию. технологию, методику, организацию библиографии

en Science of bibliography
fr Science de bibliographie

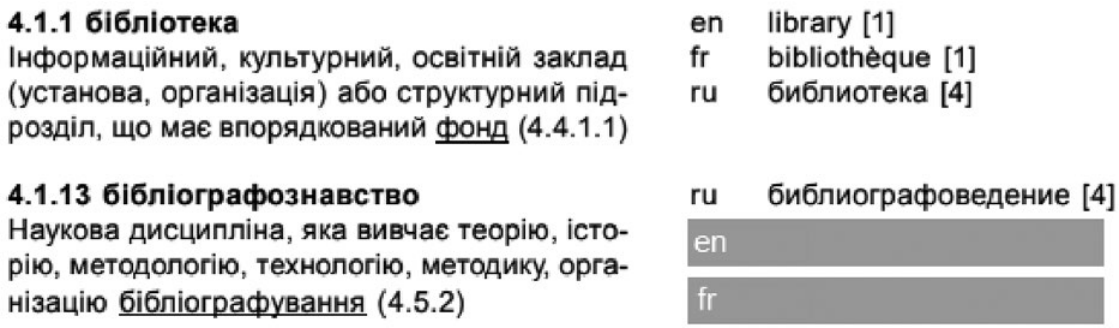
Rycina 3. Fragment w języku rosyjskim z 1999 roku sugerujący występowanie ekwiwalentów terminu *библиографоведение* w języku angielskim i francuskim
Źródło: GOST 7.0-99 1999: pkt 3.1.38.

Tym ilustracjom należy się komentarz. Do ustępu z 1974, którego autorem jest Abram Barsuk broniący rozstrzygnięć normy z 1970 roku, Barbara Eychlerowa (1977: 306) już trzy lata później odniosła się sceptycznie: zaprzeczyła, jakoby w języku polskim termin *bibliografoznawstwo* istniał. Z kolei ustęp z 1999 roku, pochodzący z rosyjskiej normy GOST 7.0–99, obok definicji terminu *библиографоведение* podaje także jego ekwiwalenty w języku angielskim i francuskim, kiedy równoległa norma ukraińska DSTU 7448:2013 (2014: pkt 4.1.13; ryc. 4) wykazuje więcej powściągliwości – w miejscu przeznaczonym na ekwiwalenty podano jedynie odpowiednik rosyjski, mimo że przy innych terminach, o ile to możliwe, podawano także odpowiedniki francuskie i angielskie. Wynika to z faktu, że ekwiwalenty przejmowano

wyłącznie z istniejących już słowników (czego dowodzą przypisy w nawiasach kwadratowych kierujące do bibliografii załącznikowej normy), a więc najwyraźniej nie odnaleziono w nich angielskiego i francuskiego odpowiednika terminu *библиографоведение*.

Na końcu należy odnotować, że po dziś dzień w tekstach pisanych przez – przykładowo – naukowców rosyjskich lecz w języku angielskim, autorzy na własną rękę czasem mechanicznie tłumaczą znany sobie termin (stąd formalnie można spotkać w piśmiennictwie określenie np. „bibliography science" – jak np. w: Oparina 2012). Jest to jednak zbyt daleko idąca swoboda.

Brak jakichkolwiek przypisów we wszystkich tych przypadkach uniemożliwia odpowiedzi na pytanie, na podstawie jakich źródeł wskazano te – trzeba przyznać zaskakujące w świetle zgromadzonych niżej faktów – ekwiwalenty, których nie udało się odnaleźć także w słownikach i encyklopediach napisanych w językach narodowych (nie będących źródłami informacji o charakterze przekładowym). W tej sytuacji



Rycina 4. Fragment w języku ukraińskim z 2014 roku, paralelny do fragmentu z ryciny 3, jednak uczciwie pokazujący, że w źródłach leksykograficznych nie odnaleziono właściwych ekwiwalentów terminu *библиографоведение* w języku angielskim i francuskim

Źródło: opracowanie własne na podstawie: DSTU 7448:2013 2014: pkt 4.1.1., pkt 4.1.13.

Rycina pokazuje, oprócz fragmentu paralelnego z terminem *бібліографознавство*, także pobliski fragment z terminem *бібліотека*, dla którego podano ekwiwalenty. Dla podkreślenia różnicy rycinę zmanipulowano, dodając zaznaczenie na szaro, uwypuklające brak ekwiwalentów.

podano w wątpliwość, czy w ogóle one istnieją. Aby ją rozwiać, spróbowano zdobyć bardziej obiektywny dowód w tej sprawie.

*

Takiej dowodowej poszlaki dostarczyło diachroniczne badanie występowania terminu *библиографоведение* oraz jego inkryminowanych ekwiwalentów przy użyciu NV, który agreguje i wizualizuje dane uzyskane z projektu Google Books (GB) w postaci wykresów. Analiza została przeprowadzona 21.02.2017 roku, zasięg chronologiczny ustalono na lata 1800–2008, a wygładzenie na poziomie 0. Po szczegółowe wyniki analizy należy odesłać do tekstu pierwotnego (Łubocki 2018: 162–166), tu skupiając się wyłącznie na samym narzędziu, bowiem oprócz niewątpliwych korzyści, jakie oferuje NV[2], trzeba mieć na względzie także trudności, jakie są związane z tym narzędziem (uwagi prezentują stan na dzień 17.08.2020).

2 W toku analizy wykazano, że w korpusie języka angielskiego nie wystąpiło ani razu słowo **bibliographology*, w korpusie języka niemieckiego – **Bibliographiekunde* oraz **Bibliographienwissenschaft*; udało się także ustalić stopień wykorzystywania na przestrzeni czasu istniejących ekwiwalentów niemieckich oraz terminu rosyjskiego.

1. Przede wszystkim dotkliwie brakuje syntetycznego tekstu w języku polskim na temat historii, funkcjonowania i możliwości analitycznych tego narzędzia, jego wad i zalet, a także wskazówek omawiających, jak uniknąć błędów w interpretacji (poza krótkim fragmentem w: Wilkowski 2013: 68–70). Symboliczne, że nie ma nawet polskiego artykułu hasłowego na Wikipedii poświęconego NV[3]. Prawdopodobnie wynika to z faktu, że do dziś NV nie oferuje pracy na polskim korpusie językowym, stąd niemożliwe jest wykonanie analiz dla języka polskiego. Wobec powyższego można posiłkować się nielicznymi publikacjami ogólnymi (Younes & Reips 2019; Sparavigna & Marazzato [2015]; Zaharov & Masevič 2014; Hayes 2011) lub rozproszonymi po różnych artykułach uwagami poczynionymi na marginesie badań empirycznych wykorzystujących NV (np.: Zięba 2018; Ophir 2016). Wciąż warto, mimo upływu czasu i poszerzenia możliwości badawczych, sięgać także po tekst „założycielski" (Michel *et al.* 2011), wprowadzający korpus językowy NV do obiegu naukowego, przedstawiający pierwsze odkrycia dokonane w pracy nad nim oraz zapowiadający samo upublicznienie NV. Do lektury tych tekstów należy gorąco zachęcić, bowiem płynie z nich pewien obraz pracy z NV. Na końcu pewną wyręką są informacje podane na stronie samego NV (books.google.com/ngrams/info) oraz anglojęzycznej Wikipedii (en wikipedia.org/wiki/Google_Ngram_Viewer).
2. Praca w NV odbywa się na ograniczonym zasobie tekstów – ok. 5 milionów, a więc szczuplejszym niż samo GB, które według informacji z 2013 roku zawiera ok. 15 milionów, a to z kolei ma jakoby odpowiadać 12% wszystkich książek opublikowanych drukiem od momentu jego wynalezienia (GB szacuje liczbę tych książek na 129 864 880) – w maju 2014 roku było to już ponad 30 milionów tj. 24% (Wilkowski 2013: 69; Weiss 2015: 183). Na marginesie warto zaznaczyć jeszcze jeden problem tych liczb. Teoretycznie wiadomo, na jakiej podstawie uzyskano tak dokładny szacunek, a także co GB mniej więcej uznaje za „książkę" (zob.: Jackson 2010), jednak z punktu widzenia sporu o definicję książki (trwającego praktycznie od zarania dziejów książki) oraz problemu „niekompletność vs. redundancja" źródeł stanowiących podstawę szacunków należy tę arytmetyczną akrybię wziąć w nawias i uznać za wysoce hurraoptymistyczną.
3. Wyniki prezentowane na wykresach podawane są w ułamkach procentów, a nie bezwzględnych liczbach, co niejednokrotnie utrudnia wnioskowanie (wprawdzie istnieje także możliwość pobrania surowych danych, jednak do ich użycia i interpretacji konieczne są wysokie kompetencje informatyczne). Wynika to z faktu, że NV pracuje na jednostkach zwanych ngramami. N-gram w ogólności to jakiś model językowy stosowany w rozpoznawaniu mowy. W przypadku NV ngram oznacza ciąg od jednego do pięciu wyrazów (jeden wyraz to 1 gram, fraza składająca się z dwóch wyrazów to 2 gramy itd.). Wykres, który na osi rzędnych przedstawia częstość występowania danego ngrama względem upływu czasu (przedstawionego na osi odciętych), generowany jest wyłącznie dla fraz 5-wyrazowych lub krótszych, występujących w korpusie co najmniej 40 razy i tylko w jednym wybranym korpusie (dla danego języka, a nie całej bazy tekstów zebranych w GB, zatem badanie występowania jednego wyrazu w kilku korpusach językowych jest utrudnione). Częstotliwość występowania ngrama w danym roku przedstawiona jest w postaci wartości

3 Istnieją artykuły w języku angielskim, hiszpańskim, francuskim, portugalskim, rumuńskim i rosyjskim.

względnej jako pomnożony przez 100% iloraz liczby wystąpień danego ngrama w danym roku podzielonej przez całkowitą liczbę wyrazów w korpusie w tym samym roku. Przykładowo (według: Zaharov & Masevič 2014: 308; za: Michel *et al.* 2011) w korpusie języka angielskiego (wersja z 2009 roku[4]) wyraz *slavery* (niewolnictwo) w 1861 roku wystąpił 21 460 razy na 11 687 stronach w 1208 książkach. Ten korpus dla roku 1861 notuje łącznie 386 434 758 wyrazów, co daje względną częstotliwość występowania ngrama wynosi 0,0055533307%.

4. W interpretacji kształtu samego wykresu uzyskanego w NV należy uważać na opcję wygładzenia („smoothing"). Wygładzenie odwołuje się do pojęcia średniej ruchomej, a jego wartość oznacza liczbę lat sąsiadujących z rokiem, dla którego wskazano wynik, które są brane do obliczenia średniej. Przykładowo wygładzenie na poziomie 1 oznacza, że do obliczenia średniej zostaną wzięte wyniki po 1 roku granicznym, stąd wynik ukazany dla roku 1950 to w istocie średnia wyniku dla 1950 roku, wyniku dla 1949 roku i wyniku dla 1951 roku podzielona przez 3 (gdyż po 1 roku sąsiadującym z 1950 to właśnie lata 1949 i 1951). Przy wygładzeniu na poziomie 5 wynik dla 1950 roku to w istocie średnia z lat 1945–1955, bowiem ten okres dzieli 5 lat liczonych od i do roku 1950. Wygładzanie służy uwypuklaniu mniej oczywistych tendencji, z drugiej jednak strony łatwo wówczas popaść w manipulację w imię „podkręcania" wyników. Stąd, jeśli „gładkość 0 oznacza brak wygładzania: tylko surowe dane"[5], należy w badaniach naukowych korzystać wyłącznie z tej opcji.
5. Niektórych wskazań nie udaje powiązać się z faktycznymi jednostkami tekstu. Nie można zatem uznać takich wskazań za wiarygodne, a tym samym – brać pod uwagę w analizie. Tak było w przypadku wyrazu *Bibliographienkunde*, które na wykresie NV występuje już w 1953 roku, ale dla którego nie podano powiązania z konkretnym dokumentem w GB. Stąd za pierwsze udokumentowane wystąpienie należy uznać dopiero to z 1969 roku. Podobnie wystąpienia wyrazu *библиографоведение* dla lat 1925 i 1933 w korpusie języka rosyjskiego nie prowadzą do żadnego dokumentu w GB.
6. Datowanie niektórych wskazań jest błędne, a więc ich też nie można brać pod uwagę w analizie. W korpusie języka rosyjskiego w roku 1929 wskazano wystąpienie wyrazu *библиографоведение*. Jest to ewidentna pomyłka popełniona w trakcie opracowywania metadanych[6], bowiem dokument w GB, do którego prowadzi wskazanie, jest zeszytem rosyjskiego czasopisma „Библиография" z 2001 roku („Bibliografiâ" 2001, z. 4 (315)).
7. Możliwość błędnego OCR-owania (automatycznego optycznego rozpoznawania znaków) oryginalnych tekstów może prowadzić do utraty niektórych wskazań lub podawać wskazania fałszywe. Przykładowo poszukiwanie wyrazu *Bibliographiewissenschaft* zaowocowało pierwszym wystąpieniem w 1924 roku na wykresie NV. Byłaby to rewolucyjna

4 NV dla języka angielskiego oferuje kilka kolejnych wersji korpusów (z 2009, 2012 lub 2019 roku), a także korpusy wyspecjalizowane: dla języka angielskiego brytyjskiego, języka angielskiego amerykańskiego czy języka angielskiego w beletrystyce.

5 Tłum. JMŁ. Oryg.: „a smoothing of 0 means no smoothing at all: just raw data" – books.google.com/ngrams/info [data dostępu: 17.8.2020]).

6 Wynikła być może z faktu, że stronie tytułowej tego czasopisma do końca 2014 roku (tj. do momentu zmiany tytułu na „Bibliografiâ i knigovedenie") widniał dopisek „издаётся с марта 1929" („wychodzi od marca 1929 [roku]").

informacja w świetle tego, że spodziewano się takiego wystąpienia dopiero w okolicach roku 1970. Niestety, przeprowadzona autopsja 1. rocznika czasopisma „Zeitschrift für Buchkunde" za rok 1924 (do którego odwoływało się wskazanie), nie potwierdziła, że występuje w nim wyraz *Bibliographiewissenschaft*. Najprawdopodobniej więc był to błąd OCR, a pierwsze uwierzytelnione wystąpienie tego wyrazu pojawiło się na otrzymanym wykresie NV w 1968 roku.

8. Oprócz błędów doświadczonych osobiście, warto jeszcze za Zaharovem i Masevičem (2014) wskazać, że uzyskane wyniki mogą oprócz tego wypaczać: skróty słów (tożsame graficznie z samodzielnymi wyrazami, przykładowo: *ul* jako skrót od wyrazu *ulica* oraz *ul* jako „dom dla pszczół"), dzielenie wyrazów (które często jest rozpoznawane przez OCR jako dwa oddzielne wyrazy) oraz homonimia (której to algorytm OCR nie jest w stanie wychwycić). Ale przede wszystkim – zmienna ortografia (która ewoluowała od czasów wynalezienia druku w niemal każdym języku), a nawet sama grafia (kształt) niektórych liter (by sięgnąć tylko do klasycznego przypadku litery ſ/s). Jak twierdzi Brian Hayes, „do obecnych celów [badawczych] szum powodowany przez proces OCR-owania jest niewielkim zakłóceniem, które możemy spokojnie zignorować" (Hayes 2011: 194) i gdyby w „magiczny sposób" oczyścić bazę NV ze wszelkich błędów i nieporozumień[7], wykresy niewiele by się zmieniły. Jednak w przypadku właśnie badań terminologicznych nie można się z tym zgodzić. Wystarczy spojrzeć na rycinę dołączoną do artykułu B. Hayesa (ryc. 5), aby przekonać się, ile umknie w analizie słowa *separate* (oddzielać) z powodu błędów pozostających w zbiorze NV.

B. Hayes przeprowadził to studium przypadku znajdując w zasobie NV wszystkie słowa, które zostały wyszukane w bazie poprzez zmianę jednej litery w ciągu liter składających się na wyraz *separate*. Dało to wynik 65 wariantów, pogrupowanych według miejsca zamiany litery oraz rodzaju błędu (wymiana litery na wersalik; pomyłka wynikająca z błędnego odczytu ſ; literówka; inne poprawne słowo w języku angielskim; literówka w innym słowie języka angielskiego; słowo w innym języku; mieszana przyczyna; błąd OCR), których częstotliwość została odzwierciedlona na rycinie rozmiarem stopnia pisma, proporcjonalnym do wszystkich znalezionych wystąpień („separate" wystąpiło 27 528 661 razy, najrzadsza forma – 42). Z analizy tej okazało się, że choć wprawdzie „seperate" jest popularniejszą formą omyłki, to „sepatate" i „scparate" są bardziej prawdopodobne jako błąd maszynowy OCR, a ponad 60 000 błędów spowodowało pomylenie ſ z f lub inną literą. Jednak najważniejsze spostrzeżenie jest takie, że niektóre warianty wcale nie muszą być błędami: „separase" (separyna) to nazwa enzymu, a „separare" (rozdzielić) to czasownik w języku włoskim. W konkluzji B. Hayes zauważa, że „poprawienie wszystkich błędów miałoby niewielki wpływ na częstotliwość występowania wyrazu *separate*, z drugiej jednak strony wyczyściłoby bazę danych z ponad 40 fałszywych słów" (Hayes 2011: 194).

W przypadku badań nad terminologią, która z natury wiąże się przede wszystkim z obszarami nauki i techniki, jest to zastrzeżenie mniej istotne, ale, zwłaszcza przy badaniach o nachyleniu kulturowym (w tym literaturoznawczym), należy mieć świadomość tego że w GB stwierdza się nadreprezentację literatury naukowej w stosunku do innych typów piśmiennictwa, a to może znacząco wpływać na uzyskane

7 Które sam szacuje na ok. 15%, choć błędy wynikające z samego OCR mają stanowić ok. 1%, bowiem „istnieje tylko jeden sposób, aby odczytać słowo poprawnie, natomiast jest niezliczona liczba sposobów, by się w tym pomylić", a do tej grupy należy jeszcze zaliczyć błędy zawinione nie przez OCR, lecz przez autorów i drukarzy.

Rycina 5. Warianty błędnego odczytania wyrazu *separate* w języku angielskim
Źródło: Hayes 2011: 194.

wyniki. Podobnym w charakterze zastrzeżeniem (mniej istotnym w badaniach terminologicznych, jeśli mają one charakter frekwencyjny) jest fakt, że tekst o małej wadze lub o ulotnej treści jest równy tekstowi poczytnemu i istotnemu, gdyż każdy tekst jest reprezentowany jednostkowo. To zastrzeżenie jest podnoszone głównie w trakcie badań nad rozwojem, popularnością i znaczeniem pewnych idei. O tym zastrzeżeniu, choć luźno związanym z samym terminoznawstwem, należy wspomnieć, bowiem NV zostało stworzone z myślą o rozwoju badań kulturomicznych, a sam termin *kulturomika* oznacza tę odmianę leksykologii komputerowej, która bada zachowania i kulturę ludzką poprzez analizę ilościową zdigitalizowanych tekstów. Termin ten jest neologizmem utworzonym na gruncie języka angielskiego amerykańskiego przez analogię do terminu *genomika* i został po raz pierwszy wprowadzony i opisany właśnie w owym „założycielskim" dla projektu NV artykule (Michel *et al*. 2011: 181–182). Zatem związek NV i badań nad kulturą ludzką jest nierozerwalny.

*

Niektórym z bolączek wskazanym powyżej potrafimy już w pewnym stopniu zapobiegać. Nadja Younes i Ulf-Dietrich Reips opracowali 5 procedur zwiększających wiarygodność wyników uzyskanych za pomocą NV oraz podali sposoby ich łączenia. Są to stosowanie kilku korpusów językowych, badanie krzyżowe przeprowadzone na różnych korpusach tego samego języka, zwrócenie uwagi na fleksję słów, badanie częstotliwości występowania synonimów oraz standaryzacja częstości występowania wyrazów. Szczegółowy ich opis i zastosowanie zamieścili w osobnej pracy (Younes, Reips 2019) poświęconej temu problemowi.

Alexander Koplenig (2017) słusznie wskazuje, że podstawą wszystkich tych problemów jest brak metadanych, które „nie są jedynie ładnym dodatkiem, lecz potężnym źródłem informacji w humanistyce cyfrowej" (Koplenig 2017: 169). W związku z tym na podstawie analiz w NV nie można mówić o obrazowaniu przemian językowych czy kulturowych w ogólności, a jedynie o reprezentacji tych przemian w zbiorach danych NV. Tego rodzaju problemem nie są obarczone inne, duże zasoby danych – bibliograficzne bazy danych i archiwa czasopism, których zasoby można wykorzystać także do analizy częstotliwości pojawiania się określonych pojęć w publikacjach naukowych, mimo że nie były w tym celu tworzone. Na tę możliwość wskazują Jason Chumtong i David Kaldewey (2017), którzy z jednej strony odżegnują się od krytyki NV, z drugiej – analizy wykonane przy jego pomocy określają jako „quick and dirty" (Chumtong & Kaldewey 2017: 6). Należy pamiętać także o kolejnym upośledzeniu tego rodzaju analizy – jest to wyłącznie analiza reprezentacji słownej jakiegoś zjawiska, a nie samego zjawiska, w związku z czym wykresy NV, według B. Hayesa, to „konkurs popularności między słowami, a nie pojęciami, które oznaczają" (Hayes 2011: 191).

W podsumowaniu należy zatem stwierdzić, że uzyskiwane wykresy mogą w pewnym stopniu wspierać proces analiz terminologicznych. Na pewno nie można na ich podstawie jednoznacznie wyznaczyć np. daty pierwszego wystąpienia danego słowa w piśmiennictwie czy też jego znaczenia lub zasięgu stosowania. Natomiast trendy, które zarysowują się na wykresach NV, można – z pewnymi zastrzeżeniami – uznawać za prawdopodobne, choć w wielu przypadkach wymagać będą dalszych potwierdzeń i wyjaśnień, szczególnie w obszarze danych pochodzących ze źródeł obcych dla anglosaskiego obszaru kulturowego.

Bibliografia

Barsuk, Abram Il'ič (1974) „Standartizaciâ bibliografičeskoj terminologii – trebovanie vremeni". [W:] *Sovetskaâ Bibliografiâ*. Vol. 6 (148); 15–28.

Čavdarov, S. *et al*. (red.) (1966) *Terminologičen rečnik po naučna informaciâ*. Moskva: Sovet Ekonomičeskoj Vzaimopomošči. Postoânnaâ komissiâ po koordinacii naučnyh i tehničeckih issledovanij.

Chumtong, Jason, David Kaldewey (2017) *Beyond the Google Ngram Viewer: Bibliographic Databases and Journal Archives as Tools for the Quantitative Analysis of Scientific and Meta-scientific Concepts.* Bonn: Rheinische Friedrich-Wilhelms-Universität Bonn, Forum Internationale Wissenschaft. Dostępny online: hdl.handle.net/20.500.11811/1150 [data dostępu: 19.8.2020].

DSTU 7448:2013 (2014) *Ìnformacìâ ta dokumentacìâ. Bìblìotečno-ìnformacìjna dìâlnìst'. Termìni ta viznačennâ ponât'*. Kiïv: Mìnekonomrozvitku Ukraïni.

Eychlerowa, Barbara (1978) „Termin *bibliografia* w radzieckiej normie „Bibliografja. Tierminy i opriedielenja"". [W:] *Przegląd Biblioteczny*. Vol. 46 (3); 301–311.

Gooding, Paul (2012) "Mass Digitization and the Garbage Dump: the Conflicting Needs of Quantitative and Qualitative Methods". [W:] *Literary and Linguistic Computing*. Vol. 28 (3); 425–431.

GOST 16448-70 (1970) *Bibliografiâ. Terminy i opredeleniâ*. Moskva: Gosudarstvennyj komitet standartov Soveta Ministrov SSSR.

GOST 7.0-99 (1999) *Informacionno-bibliotečnaâ deâtel'nost', bibliografiâ. Terminy i opredeleniâ*. Moskva: IPK Izdatel'stvo standartov.

Hayes, Brian (2011) "Bit Lit. With Digitized Text from Five Million Books, One is Never at a Loss for Words". [W:] *American Scientist*. Vol. 3 (99); 190–194.

Jackson, Joab (2010) "Google: 129 Million Different Books have been Published". [W:] *PCWorld*. 6.08.2010. Dostępny online: pcworld.com/article/202803/google_129_million_different_books_have_been_published.html [data dostępu: 17.8.2020].

Koplenig, Alexander (2017) "The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets – Reconstructing the Composition of the German Corpus in Times of WWII".[W:] *Digital Scholarship in the Humanities*. Vol. 32 (1); 169–188.

Łubocki, Jakub Maciej (2018) „O problemie wieloznaczności terminu 'bibliografia' kolejny raz – o terminie 'bibliografoznawstwo' po raz pierwszy". [W:] *Roczniki Biblioteczne*. Vol. 62; 135–178.

Markov, I.G. (1948) „O predmete i metode bibliografii (opyt postanovki voprosa)". [W:] *Trudy MGBI*. Vol. 4; 101–134.

Michel, Jean-Baptiste *et al.* (2011) "Quantitative Analysis of Culture Using Millions of Digitized Books Science". [W:] *Science*. Vol. 331 (6014); 176–182.

Migoń, Krzysztof (1984) *Nauka o książce. Zarys problematyki*, Wrocław: Zakład Narodowy im. Ossolińskich. Wydawnictwo.

Oparina, O.D. (2012) "The Interdisciplinary Aspects of the Interaction of Bibliography with Social Sciences and Humanities". [W:] *Scientific and Technical Information Processing*. Vol. 39 (1); 42–46.

Ophir, Shai (2016) "Big Data for the Humanities Using Google Ngrams: Discovering Hidden Patterns of Conceptual Trends". [W:] *First Monday*. Vol. 21 (7). Dostępny online: dx.doi.org/10.5210/fm.v21i7.5567 [data dostępu: 17.8.2020].

Pechenick, Eitan Adam, Christopher M. Danforth, Peter Sheridan Dodds (2015) "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-cultural and Linguistic Evolution". [W:] *PloS One*. Vol. 10 (10). Dostępny online: doi.org/10.1371/journal.pone.0137041 [data dostępu: 2.9.2020].

Pettit, Michael (2016) "Historical Time in the Age of Big Data: Cultural Psychology, Historical Change, and the Google Books Ngram Viewer". [W:] *History of Psychology*. Vol. 19 (2); 141–153.

Rešetinskij, I.I. (1969) "Sovremennoe sostoânie bibliografičeskoj terminologii i zadači ee standartizacii". [W:] *Sovetskaâ Bibliografiâ*. Vol. 2 (114); 19–32.

Rešetinskij, I.I. (1970) "Pervyj gosudarstvennyj standart na bibliografičeskuû terminologiû". [W:] *Sovetskaâ Bibliografiâ*. Vol. 3 (121); 11–16.

Šamurin, Evgenij Ivanovič (1958) *Slovar' knigovedčeskih terminov. Dlâ bibliotekarej, bibliografov, rabotnikov pečati i knižnoj torgovli*. Moskva: Sovetskaâ Rossiâ.

Sparavigna, Amelia C., Roberto Marazzato (2015) *Using Google Ngram Viewer for Scientific Referencing and History of Science*. Dostępny online: https://arxiv.org/abs/1512.01364 [data dostępu: 17.8.2020].

Weiss, Andrew (2015) "Google Ngram Viewer". [W:] Carol Smallwood (red.) *The Complete Guide to Using Google in Libraries, vol. 1.: Instruction, Administration, and Staff Productivity.* Lanham: Rowman & Littlefield; 183–189.

Wilkowski, Marcin (2013) *Wprowadzenie do historii cyfrowej*. Gdańsk: Instytut Kultury Miejskiej.

Younes, Nadja, Ulf-Dietrich Reips (2019) "Guideline for Improving the Reliability of Google Ngram Studies: Evidence from Religious Terms". [W:] *PloS One*. Vol. 14 (3). Dostępny online: doi.org/10.1371/journal.pone.0213554 [data dostępu: 17.8.2020].

Zaharov, Viktor Pavlovič, Andrej Cezarevič Masevič (2014) "Diahroničeskie issledovaniâ na osnove korpusa russkih tekstov Google Books Ngram Viewer". [W:] *Strukturnaâ i Prikladnaâ Lingvistika*. Vol. 10; 303–327.

Zięba, Anna (2018) "Google Books Ngram Viewer in Socio-Cultural Research". [W:] *Research in Language*. Vol. 16 (3); 357–376.