

31 maja 2017



*P*rzegląd
*S*ocjologii
*J*akościowej

Tom XIII

Numer 2

Big Data i CAQDAS
w badaniach jakościowych

pod redakcją

Macieja Brosza, Grzegorza Brydy
oraz Piotra Siudy

www.przegladsocjologiijakosciowej.org

REDAKTOR NACZELNY

Krzysztof Tomasz Konecki, UŁ

REDAKTORZY PROWADZĄCY

Jakub Niedbalski, UŁ

Izabela Ślęzak, UŁ

REDAKTORZY TEMATYCZNI

Waldemar Dymarczyk, UŁ

Marek Gorzko, US

Anna Kacperczyk, UŁ

Sławomir Magala, Erasmus

University

Łukasz T. Marciniak, UŁ

REDAKTOR

DZIAŁU RECENZJI

Dominika Byczkowska-Owczarek,

UŁ

REDAKTOR JĘZYKOWY

Aleksandra Chudzik (j. polski)

Jonathan Lilly (j. angielski)

AUDYTOR WEWNĘTRZNY

Anna Kubczak, UŁ

KOREKTA

Magdalena Chudzik-Duczmańska

Magdalena Wojciechowska, UŁ

SKŁAD

Magdalena Chudzik-Duczmańska

PROJEKT OKŁADKI

Anna Kacperczyk, UŁ

WWW

Edyta Mianowska, UZ

ADRES REDAKCJI

Uniwersytet Łódzki

Wydział Ekonomiczno-Socjologiczny

Instytut Socjologii

Katedra Socjologii

Organizacji i Zarządzania

ul. Rewolucji 1905 r. 41/43

90-214 Łódź

redakcja.psj@gmail.com

Przegląd
Socjologii
Jakościowej

Tom XIII
Numer 2

Big Data i CAQDAS
w badaniach jakościowych

pod redakcją
Macieja Brosza, Grzegorza Brydy
oraz Piotra Siudy

Prawa autorskie

Czasopismo oraz wszystkie zamieszczone w nim artykuły stanowią dorobek współczesnej socjologii. Mogą zostać wykorzystane bez specjalnej zgody dla celów naukowych, edukacyjnych, poznawczych i niekomercyjnych z podaniem źródła, z którego zostały zaczerpnięte.

Wykorzystywanie ogólnodostępnych zasobów zawartych w naszym piśmie dla celów komercyjnych lub marketingowych wymaga uzyskania specjalnej zgody od wydawcy. Pobieranie opłat za dostęp do informacji lub artykułów zawartych w naszym piśmie lub jakiegokolwiek ograniczanie do niego dostępu jest zabronione. Autorzy nadsyłanych artykułów ponoszą odpowiedzialność za uzyskanie zezwoleń na publikowanie materiałów, do których prawa autorskie są w posiadaniu osób trzecich.

Logotyp, szata graficzna strony oraz nazwa *Przeglądu Socjologii Jakościowej* (*Qualitative Sociology Review*) znajdują się w wyłącznym posiadaniu wydawcy. Wszystkie pozostałe obiekty graficzne, znaki handlowe, nazwy czy logotypy zamieszczone na tej stronie stanowią własności ich poszczególnych posiadaczy.

RADA NAUKOWA

Jan K. Coetzee

University of the Free State, Bloemfontein, South Africa

Markieta Domecka

University of Surrey, UK

Aleksandra Galasińska

University of Wolverhampton, UK

Piotr Gliński

Uniwersytet Białostocki

Marek Kamiński

New York University, USA

Michał Krzyżanowski

Lancaster University, UK

Anna Matuchniak-Krasuska

Uniwersytet Łódzki

Barbara Misztal

University of Leicester, UK

Janusz Mucha

Akademia Górniczo-Hutnicza

Sławomir Partycki

Katolicki Uniwersytet Lubelski

Anssi Perakyla

University of Helsinki, Finland

Robert Prus

University of Waterloo, Canada

Marek Szczepański

Uniwersytet Śląski

Piotr Sztompka

Uniwersytet Jagielloński

SPIS TREŚCI

Od redaktorów

Maciej Brosz, Grzegorz Bryda, Piotr Siuda

Big Data i CAQDAS a procedury badawcze w polu socjologii jakościowej 6

Artykuły

Mariusz Dzieglewski

CAQDAS w badaniach digitalizacji i odbioru dziedzictwa kulturowego. Korzyści i ograniczenia 24

Jakub Niedbalski

From a qualitative researcher's workshop—the characteristics of applying computer software in studies based on the grounded theory methodology 46

Marek Troszyński, Aleksander Wawer

Czy komputer rozpozna hejtera? Wykorzystanie uczenia maszynowego (ML) w jakościowej analizie danych 62

Agnieszka Kwiatkowska

„Hańba w Sejmie” – zastosowanie modeli generatywnych do analizy debat parlamentarnych 82

Alicja Zawistowska, Małgorzata Skowrońska

Dobra zmiana czy Polska w ruinie? 110

Krzysztof Tomanek

Metodyka dla analizy treści w projektach stosujących techniki text mining i rozwiązania CAQDAS piątej generacji 128

Konkurs fotograficzny PSJ

Wyniki VII edycji konkursu fotograficznego „Przeglądu Socjologii Jakościowej” 144

Maciej Brosz
Uniwersytet Gdański

Grzegorz Bryda
Uniwersytet Jagielloński

Piotr Siuda
Uniwersytet Kazimierza Wielkiego

Od redaktorów: *Big Data* i CAQDAS a procedury badawcze w polu socjologii jakościowej

Abstrakt Świat życia codziennego zalewany jest ogromną liczbą różnych skwantyfikowanych i zdigitalizowanych danych. Mogą one być przetwarzane i traktowane jako materiał badawczy – również jakościowy. Zastosowanie w badaniach jakościowych wielkich zbiorów danych (*Big Data*) modyfikuje postępowanie na wszystkich etapach procesu badawczego: od projektowania badania aż po formułowanie wniosków końcowych. Czy osadzenie w badaniach jakościowych modelu *Big Data* prowadzi ku ateoretyczności badań? Jakie skutki niesie rezygnacja z próby losowej na rzecz kompletności danych? Celem artykułu jest zasygnalizowanie owych zmian i krótkie ich scharakteryzowanie z uwzględnieniem znaczącej roli różnego typu oprogramowania (zwłaszcza CAQDAS), a co za tym idzie – analiz, które można przeprowadzać.

Słowa kluczowe *Big Data*, CAQDAS, komputerowa analiza danych jakościowych, danetyzacja, przetwarzanie danych, badania jakościowe

Maciej Brosz, dr, socjolog, adiunkt w Zakładzie Socjologii Stosowanej w Instytucie Filozofii, Socjologii i Dziennikarstwa Uniwersytetu Gdańskiego, właściciel firmy Q&Q Zakład Realizacji Badań Społecznych. Zainteresowania naukowe: socjologia zamieszkiwania, jakość życia w środowisku miejskim, rewitalizacja zdegradowanych obszarów miejskich, komputerowe wspomaganie badań jakościowych, programowanie w języku R.

Adres kontaktowy:

Uniwersytet Gdański
Instytut Filozofii, Socjologii i Dziennikarstwa
ul. Jana Bażyńskiego 4, 80-309 Gdańsk
e-mail: maciej.brosz@ug.edu.pl

Grzegorz Bryda, dr, adiunkt w Instytucie Socjologii UJ, w Zakładzie Socjologii Komunikacji Społecznej, Kierownik Pracowni analiz i badań jakościowych CAQDAS TM Lab. Zainteresowania: teoria socjologiczna, kogniwytyka, *big data*, metodologia badań jakościowych, informatyka społeczna, CAQDAS, TextMining i NLP w analizie danych jakościowych, modelowanie procesów społecznych. Współpracuje z instytucjami publicznymi i prywatnymi w zakresie metodologii badań społecznych oraz wielowymiarowej analizy danych ilościowych i jakościowych.

Adres kontaktowy:

Instytut Socjologii, Uniwersytet Jagielloński
ul. Grodzka 52, 31-044 Kraków
e-mail: grzegorz.bryda@uj.edu.pl; pracownia.caqdastm@gmail.com

Piotr Siuda, doktor socjologii, prodziekan ds. nauki oraz adiunkt na Wydziale Administracji i Nauk Społecznych Uniwersytetu Kazimierza Wielkiego w Bydgoszczy. Interesuje się społecznymi aspektami Internetu oraz metodami badań socjologicznych. Autor kilku monografii naukowych; publikował w wielu periodykach naukowych, w tym w wysoko punktowanych pismach, posiadających *impact factor* („European Journal of Cultural Studies”, „International Journal of Cultural Studies”). Koordynator szeregu projektów badawczych, w tym zadań: „Dzieci sieci”, „Dzieci sieci 2.0”, „Prosumpcjonizm pop-

-przemysłów”. Członek The Association of Internet Researchers oraz Polskiego Towarzystwa Socjologicznego. Prowadzi sympozja i szkolenia z pisania artykułów naukowych. Strona domowa: <http://www.piotrsiuda.com>; blog: <http://www.piotrsiuda.pl>.

Adres kontaktowy:

Uniwersytet Kazimierza Wielkiego
Wydział Administracji i Nauk Społecznych
ul. Ogińskiego 16, 85-092 Bydgoszcz
e-mail: piotr.siuda@ukw.edu.pl

Współcześnie twierdzi się o nastaniu ery *Big Data* (por. Chen, Zhang 2014) – obserwujemy gwałtowny przyrost liczby informacji gromadzonych w formie elektronicznej, jak i rozwój technologii dotyczących pozyskiwania danych, ich zapisu oraz magazynowania w postaci repozytoriów, hurtowni, archiwów czy też różnorodnych korpusów danych tekstowych. Zmianie ulega także sposób wykorzystywania danych, tak ilościowych, jak i jakościowych. Współcześnie dane te w formie zdigitalizowanej można spotkać w każdym obszarze życia codziennego, począwszy od baz danych dotyczących transakcji bankowych, informacji z kas fiskalnych, rejestrów użycia kart kredytowych, zestawień rozmów telefonicznych, przez statystyki urzędowe, archiwa danych statystycznych, danych społecznych (sondazowych, jakościowych), aż po rejestry medyczne, biologiczne i tym podobne¹. Warto podkreślić, iż źródłem *Big Data* jest nie tylko Internet, kojarzony zwykle z tej skali zasobami. Wielkie zbiory pozostają w domenach archiwów

¹ W systemach *Big Data* pozyskujemy i analizujemy zarówno dane miękkie (opinie, sądy, komentarze), jak i dane twarde (zdarzenia, fakty, transakcje, zachowania).

państwowych, instytucji samorządowych, korporacji, fabryk, przedsiębiorstw, szpitali, czyli wszędzie tam, gdzie rejestrowane są ślady ludzkiej działalności. Wraz z rozwojem *Big Data* doskonalone są narzędzia ich przetwarzania oraz analizowania (specjalne miejsce w tym artykule poświęcone zostanie narzędziom CAQDAS, czyli oprogramowaniu wspomagającemu proces analizy danych jakościowych).

Pojęcie *Big Data* ukuto w takich naukach jak astronomia czy genetyka, które jako pierwsze w XXI wieku zmierzyły się z olbrzymią ilością danych. Obecnie koncepcja *Big Data* przenosi się na wszelkie obszary działalności ludzkiej. Termin ten nie ma jednoznacznej definicji. Początkowo oznaczał taki wzrost liczby informacji, że pamięci komputerów służących do ich przetwarzania stały się niewystarczające, co zmusiło do zaprojektowania nowych narzędzi. Obecnie uważa się, że termin *Big Data* obejmuje wszystko to, co może być realizowane badawczo w dużej skali celem zyskania nowej wiedzy lub wartości w znaczący sposób zmieniających nasze życie codzienne, choć nie ma ostatecznej zgody co

do zasięgu owej skali. Na pewno natomiast *Big Data* stanowi wyzwanie dla naszego sposobu życia i interakcji ze światem.

Wielkie dane są także wyzwaniem dla badaczy jakościowych, gdyż ich zaistnienie modyfikuje postępowanie na wszystkich etapach procesu badawczego. Celem przedstawianego tutaj artykułu jest zasygnalizowanie owych zmian i krótkie ich scharakteryzowanie z uwzględnieniem znaczącej roli różnego typu oprogramowania (zwłaszcza typu CAQDAS). W kolejnych częściach tekstu zajmiemy się kolejnymi etapami badawczymi. Zaczniemy od projektowania badań, starając się odpowiedzieć na pytanie, czy *Big Data* i inne nowe technologie prowadzą do powstania nowego paradygmatu badawczego rzutującego na to, jak przystępuje się do badań w polu socjologii jakościowej. Następnie pokrótce omówimy narzędzia i techniki badań jakościowych opartych na *Big Data* i różnego typu oprogramowaniu, po czym przejdziemy do etapu analizy danych gromadzonych z ich wykorzystaniem.

Warto podkreślić, że przedstawiany artykuł należy traktować w kategoriach wstępu do całego prezentowanego numeru „Przeglądu Socjologii Jakościowej”. Zawarte w tym numerze publikacje są pokłosiem zorganizowania jednej z grup tematycznych w trakcie XVI Ogólnopolskiego Zjazdu Socjologicznego PTS, „Solidarność w czasach nieufności”, w Gdańsku we wrześniu 2016 roku (grupa „Big Data, CAQDAS i nowe technologie w polu socjologii jakościowej”). Właśnie dlatego artykuł zamyka krótkie streszczenie wszystkich tekstów z tego numeru.

Proces projektowania badań – kilka słów o nowym paradygmacie badawczym

Czy pojawienie się na badawczym horyzoncie socjologicznych dociekań jakościowych tak zwanego *Big Data* oznacza inne podejście do konstruowania badań, czy drastycznie zmienia się proces ich projektowania? Jakie istotne przekształcenia wiążą się z koniecznością wykorzystywania odpowiedniego oprogramowania (np. typu CAQDAS) umożliwiającego przetwarzanie ogromnych ilości danych? Czy użycie wspomnianych nowych technologii powoduje pojawienie się zupełnie nowego paradygmatu badawczego?

Odpowiedź jest pozytywna – *Big Data* i nowe technologie oznaczają nowe epistemologiczne spojrzenie na proces projektowania naukowych dociekań, co wynika przede wszystkim z faktu, że wiedza generowana w ich toku ma pochodzić nie z testowania różnych teorii przez odwoływanie się do odpowiednich danych empirycznych. Bezpośrednim źródłem poznania stają się same dane – to z nich wypływać ma wiedza o świecie społecznym. Zmianę tę doskonale ujął Rob Kitchin (2014), odnosząc się nie do samych badań jakościowych, ale do nauki jako takiej. Jego zdaniem mamy do czynienia z nowym paradygmatem o zakresie multidyscyplinarnym, przy czym paradygmat ujęty został tutaj nie w kategoriach Kuhnowskich, czyli jako powszechnie obowiązujący sposób postrzegania tego, w jaki sposób uprawiać naukę. Paradygmaty naukowe powstawać mają w wyniku zmiany formatów danych, z którymi badacze mają do czynienia. Rewolucje naukowe nie wynikają z wyczerpywania się starych paradygmatów, ich niemożności odpowiadania na kluczowe pytania nurtujące akademików – rewolucje są wy-

nikiem powstawania nowych metod analitycznych (por. Hey, Tansley, Tolle 2009).

Ujmując sprawę w ten sposób, nie można zapominać, że samo pojęcie paradygmatu naukowego podlega sporej krytyce – zwłaszcza w naukach społecznych, a więc w polu mocno zróżnicowanym pod względem podejść badawczych (por. Masterman 1970). Patrzenie na owe nauki, w tym socjologię, jako determinowane przez jakiś jeden nadrzędny paradygmat zdaje się nieuprawnionym upraszczaniem bardzo zróżnicowanego obszaru, sztucznym nakładaniem ram niepasujących do wielości sposobów patrzenia na świat społeczny. Mamy zatem do czynienia z rozbieganiem się teoretycznego ujmowania nauki z tym, jak wyglądają badania naukowe w praktyce – w wypadku terminu paradygmatu nie uwzględnia się w należyty sposób procesów ciągłej ewolucji różnych nurtów akademickich.

Mimo omawianych kontrowersji, zaryzykować można stwierdzenie, że, jeśli chodzi o zaistnienie *Big Data* oraz wykorzystanie różnego typu oprogramowania, można mówić o tak znaczącej zmianie, że zastosowanie pojęcia paradygmatu jest na miejscu. Termin ten ma służyć podkreśleniu skali zachodzącej zmiany. Nowa rewolucja, którą obserwujemy, ma być rewolucją danych oraz metod analitycznych i skutkować znaczącą zmianą praktyk badawczych w ramach nauk społecznych, w tym zmianą sposobów projektowania badań.

„Śmierć teorii” czy „wzbogacenie” obecnych procedur

Bardziej radykalni zwolennicy wykorzystania nowych technologii ogłosili już nawet „śmierć teorii”

(por. Anderson 2008; Pigliucci 2009; Kitchin 2014; Parks 2014). Ciągłe wzrastająca liczba danych, a także idące za owym wzrostem doskonalenie technik ich analizowania (lepsze, wydajniejsze oprogramowanie) czynią teorię zbędną w odkrywaniu praw rządzących społeczeństwami ludzkimi. *Big Data* automatycznie produkują znaczącą wiedzę na temat skomplikowanych zjawisk społecznych – niepotrzebne jest stosowanie się do procedur procesu badawczego, łącznie z formułowaniem hipotez. Dane mówią same za siebie, wolne od teoretycznych ograniczeń, co powoduje, że badacze społeczni uwolnić się powinni od obsesji poszukiwania przyczynowości na rzecz poszukiwania korelacji – mniej istotne staje się odpowiadanie na pytanie „dlaczego?”, na znaczeniu natomiast zyskuje szukanie odpowiedzi na pytanie „co?”. Celem ma być nie tyle odkrycie przyczyn zjawisk i procesów, lecz powiązań, relacji między nimi. Zbędne jest konstruowanie teoretycznych modeli – nauki społeczne powinny poświęcić się zwiększaniu jakości zbieranych (wielkich) danych oraz doskonaleniu narzędzi ich analizowania. Warto podkreślić, że tego rodzaju podejście dominować zaczyna przede wszystkim w świecie biznesu i badaczy zachowań konsumenckich² – kiedy przeniesiemy

² Dzieje się tak z prostego powodu – patrzenie na dane jako „mówiące same za siebie” stanowi dla biznesu świetną podstawę sprzedaży swoich produktów. *Big Data* mają oferować uzyskanie dogłębnej, obiektywnej i przynoszącej zyski informacji bez zaangażowania nauki oraz naukowców. Dobrym przykładem są systemy polecania produktów klientom sklepów internetowych. Weźmy chociażby internetowe księgarnie, gdzie poszczególnemu użytkownikowi „podsuwa się” książki nie w oparciu o czynniki kulturowe czy konwenanse związane z czytaniem, ale w oparciu o wzorce zakupowe wszystkich konsumentów danej e-księgarni. To, czy danej osobie X spodoba się konkretna pozycja książkowa domniemywa się w oparciu o śledzenie zwyczajów zakupowych innych osób kupujących podobne (lub takie same) książki jak ta osoba. Stwierdzenie takich nabywczych uwarunkowań jest w tym wypadku celem – zupełnie zbędna jest wiedza na temat powodów występowania takich, a nie innych zależności.

je na pole rozważań akademickich, w tym na pole socjologii jakościowej, wydaje się ono wiązać z wieloma uproszczeniami.

Wielkie dane nie powstają przecież znikąd – są zawsze wynikiem działań ludzkich ukierunkowanych na zdobywanie konkretnych informacji, a używane sposoby analizowania, a także wykorzystywane algorytmy zależą od decyzji konkretnych badaczy. Wykrywanie zależności widocznych w wypadku konkretnych danych nie zachodzi zatem w próżni i zawsze jest wynikiem wcześniejszych odkryć, teorii, ale też doświadczeń czy umiejętności poszczególnych osób. Dane (nie tylko te wielkie) nigdy nie mówią same za siebie, nie mogą się „oswobodzić” od interpretacyjnych ram nakładanych na nie przez badaczy. Interpretacje zależą zaś od indywidualnych predyspozycji, przekonań czy postaw naukowców i nawet jeśli proces zbierania oraz analizowania danych jest mocno zautomatyzowany, jest on zawsze „osadzony” w konkretnych wartościach i kontekstualizowany w obrębie danego pola badawczego. Pozyskiwanie *Big Data* i wykorzystywanie różnych narzędzi służących ich „obróbce” nie jest procesem tak obiektywnym, jak moglibyśmy sądzić. Interpretacje są przecież także wynikiem decyzji dotyczących tego, jak owe dane zbierać i analizować, jakie nowe technologie wykorzystywać i tym podobne. Ze względów opisanych wyżej powinniśmy patrzeć na *Big Data* zupełnie inaczej niż proponują zwolennicy tezy o „śmierci teorii”. Jak przekonuje wspomniany wcześniej Kitchin (2014), zmiana dotycząca nowego paradygmatu rzeczywiście polega na uzyskiwaniu zrozumienia danego zjawiska w oparciu o dane, a nie w oparciu o teorię, ale nie

traci się jej z oczu. Do projektowania badań podchodzi się w sposób indukcyjny (od szczegółu do ogółu), choć wyjaśnianie przez indukcję nie jest końcem procesu badawczego. Można powiedzieć, że wykorzystanie *Big Data* jest dopiero wstępem do formułowania hipotez i wdrażania metody dedukcyjnej (od ogółu do szczegółu)³. *Big Data* służą identyfikowaniu potencjalnych pytań badawczych, mających być potem weryfikowanymi w toku dalszych badań. Dużą rolę gra w tym wypadku sama teoria – to jak dane są generowane i jak zostaną użyte wynika z przyjęcia pewnych założeń podpartych wiedzą teoretyczną. To ona podpowiada, jak należy podejść do konkretnych danych, aby uzyskać wartościowe informacje. Podsumowując, można powiedzieć, że mamy do czynienia ze zmodyfikowaniem tradycyjnego procesu badawczego w taki sposób, aby uwzględnił on nową drogę budowania teorii – drogę uwzględniającą wykorzystanie *Big Data*.

Przy okazji warto zaznaczyć, że badacz jakościowy podejmujący ową drogę i projektujący dane badanie musi zdawać sobie sprawę z konieczności „otworzenia się” na inne dyscypliny naukowe. Wydobywanie wartościowych informacji z wielkich danych ze zrozumiałych względów wymaga interdyscyplinarnego podejścia w sferze metodologii. Socjolog pracujący z *Big Data* może nie być świadomy możliwości tkwiących w różnego typu oprogramowaniu – z pewnością przydatna jest w tym wypadku wiedza informatyczna czy statystyczna. Potrzebę interdyscyplinarności widać zresztą także na poziomie teoretycznym – z racji tego, że wielkie dane bardzo

³ Można zatem w tym względzie zauważyć pewne analogie do klasycznej teorii ugruntowanej.

często są tak bogate w szczegóły, że ich analiza i interpretacja odwoływać się musi do teoretycznych doświadczeń wielu dyscyplin.

„Jasne” strony nowego paradygmatu

Przy tym wszystkim przy projektowaniu badań wykorzystujących *Big Data* w polu socjologii jakościowej należy być świadomym, że dane te mogą mieć bardzo duże znaczenie dla rozwiązania powszechnie znanego dylematu metodologicznego. Chodzi o zapewnienie realizmu badawczego z jednej strony, a z drugiej strony o zachowanie kontroli nad warunkami przeprowadzania badania (por. Chang, Kauffman, Kwon 2013). Zwykle wybór konkretnej metody badawczej wiąże się z „opowiedzeniem się” albo za realizmem, albo za ściślejszą kontrolą. Na przykład obserwacja socjologiczna o charakterze uczestniczącym niejawnym (badacz staje się pełnoprawnym członkiem obserwowanej grupy, nie informując jej członków o tym, że są przedmiotem obserwacji) zapewnia wysoki realizm badania, natomiast zupełnie niemożliwe staje się kontrolowanie warunków, w jakich ono się odbywa. Rzeczy mają się odwrotnie, jeśli chodzi o, na przykład, tradycyjne eksperymentalne badania laboratoryjne, gdzie dąży się do wyeliminowania przypadkowości oraz zredukowania wpływu czynników zewnętrznych mogących zniekształcać otrzymane wyniki. Wykorzystanie *Big Data* oraz różnego oprogramowania wspierającego ich analizę umożliwia zażegnanie opisywanego konfliktu. Wielkie dane można zbierać w oparciu o wcześniej ustalone założenia teoretyczne, skupiać się na różnych ich aspektach, manipulować zmiennymi tak, aby uzyskać dane, na których akurat zależy bada-

czom. Można zatem tak zaprojektować badanie, aby znaleźć dane spełniające wcześniej sformułowane założenia eksperymentalne, a więc uzyskać sporą dozę kontroli. Jednocześnie w wypadku badań z wykorzystaniem *Big Data* nie ma potrzeby „kopiowania” rzeczywistego świata społecznego w sztucznym otoczeniu. Możemy „obserwować” zjawiska przebiegające w naturalny sposób, zbierać informacje reprezentujące ludzkie działania i interakcje, gromadzić cyfrowe ślady ludzkiej działalności – na przykład tweety, opinie internautów, kliknięcia na aukcjach sieciowych i tak dalej. Wykorzystanie *Big Data* pozwala uzyskać dogłębną wiedzę o jednostkach czy społecznościach; możliwe staje się także uchwycenie dynamiki wielu różnych zjawisk społecznych⁴. Patrząc ogólnie na jakościowe badania socjologiczne oparte na *Big Data*, warto zaznaczyć, że badacze uzyskują dostęp do danych zupełnie nowego rodzaju, a także korzystają z rozlicznych ułatwień w dostępie do danych do tej pory trudno osiągalnych. Ponadto można nimi łatwo zarządzać przy pomocy różnych narzędzi, na przykład oprogramowania typu CAQDAS. Niejednokrotnie wykorzystanie omawianych

⁴ Warto podkreślić, że o przydatności *Big Data* można mówić na trzech poziomach socjologicznych dociekań – na poziomie makro, mezo oraz mikro, chociaż dla badacza jakościowego istotne są w tej mierze dwa ostatnie poziomy. Jeśli chodzi o pierwszy, najszerzy w swoim zasięgu, wielkie dane mogą pomóc chociażby w odkrywaniu wzorów międzynarodowego przepływu ludności (na przykład migrantów), ale też zależności między państwami, gospodarkami czy też różnymi sektorami przemysłu. Analizy na poziomie mezo wykorzystujące *Big Data* mogą skupiać się na śledzeniu zachowań poszczególnych jednostek używających urządzeń mobilnych na danym obszarze geograficznym; innym przykładem jest zbieranie danych na temat komunikacji i zachowań użytkowników poszczególnych portali typu *socialnetworking* (np. Facebook). Mikroanalizy natomiast obejmować mogą zagadnienia dotyczące aktywności sieciowej poszczególnych internautów (np. tekstualne wzorce wykrywane we wpisach na danym blogu).

nowych technologii w polu badań jakościowych uzupełnia i dopełnia bardziej tradycyjne podejścia badawcze, umożliwia konstruowanie badań o znacznie szerszym zasięgu oraz takich, które udzielają odpowiedzi na pytania, na które odpowiedzi nie udałoby się uzyskać bez pomocy owych technologii (por. Shah, Cappella, Neuman 2015: 9).

Warto zaznaczyć, że mimo zasygnalizowanych wyżej niewątpliwych zalet *Big Data* wielu badaczy ma negatywne nastawienie, jeśli chodzi o możliwość ich wykorzystania w badaniach jakościowych (por. Ramsay 2010). Trzeba bowiem pamiętać, że *Big Data* mogą mieć charakter redukcjonistyczny. Rezultatem ich wykorzystania mogą być analizy ignorujące szerszy kontekst społeczny, na przykład uwarunkowania kulturowe konkretnych zjawisk społecznych. Mamy zatem do czynienia ze „słabą”, jedynie powierzchowną analizą, zamiast z wnikliwym poznaniem danego zagadnienia – głębia zrozumienia ma być zastąpiona skalą danych redukujących skomplikowane i wielowymiarowe struktury społeczne do liczb. Dobry przykład podał w swoim artykule wspomniany wcześniej Kitchen (2014), kiedy wspominał o projekcie analizującym język używany przez użytkowników serwisu internetowego Twitter. Stworzona przez badaczy swoista językowa mapa Twittera pokazała wzorce geograficznej koncentracji różnych społeczności etnicznych w mieście Nowy Jork. Dociekania nie były jednak w stanie odkryć sposobów, w jaki powstają takie zgromadzenia, a także tego, jakie są konsekwencje ich istnienia. Rozstrzygnięcie tych kwestii wymaga przecież oparcia się na teorii oraz „głębokiej” wiedzy kontekstualnej. Podany przykład ma potwierdzać, że zjawiska społeczne są

zbyt skomplikowane oraz przypadkowe, aby dało się je zredukować do praw i formuł. Ludzie często nie zachowują się w sposób racjonalny – ich życie pełne jest sprzeczności, paradoksów oraz nieprzewidywanych wydarzeń. Dodatkowo ogromne zróżnicowanie stylów życia czy kultur powoduje, że redukcjonowanie zróżnicowanego świata społecznego do uniwersalnych modeli nie jest uprawnione.

Przyznając takiemu podejściu rację, należy od razu zaznaczyć, że nie można twierdzić, że analizy dokonywane w oparciu o *Big Data* są pozbawione wartości. Jest wręcz odwrotnie, choć rzeczywiście należy uznać, że wytworzone informacje są bardzo specyficzne. Potrzebują osadzenia w kontekście społecznym, a także tego, o czym mowa była wcześniej, czyli oparcia w teorii. Potrzebują być może także wsparcia się na informacjach wytworzonych w toku badań prowadzonych zgodnie ze starym paradygmatem, czyli tych bazujących na danych gromadzonych w znacznie mniejszej skali. Należy przecież pamiętać, że *Big Data*, CAQDAS oraz wszelkie nowe technologie nie stanowią o radykalnym zerwaniu z przeszłością badań jakościowych. Można tutaj raczej mówić o metodologicznym postępie, nad charakterem którego wciąż trzeba się zastanowić. Cały czas bowiem zachodzi potrzeba szerszej krytycznej refleksji nad epistemologicznymi konsekwencjami użycia *Big Data*, zwłaszcza w polu badań jakościowych. Poza wszelką wątpliwością jest tylko to, że omawiany nowy paradygmat opiera się na dostępności nowych narzędzi i rozwiązań technologicznych umożliwiających proces analizy danych – to właśnie tymi narzędziami i rozwiązaniami zajmiemy się bliżej w kolejnej części tekstu.

Metody i techniki prowadzenia badań z wykorzystaniem *Big Data* i innych nowych technologii

O metodach i technikach prowadzenia badań z wykorzystaniem *Big Data*, CAQDAS i innych nowych technologii z pewnością można orzec, iż jest to obszar zróżnicowania oraz przenikających się tradycji i nowych rozwiązań lub idei. Stąd też próba uchwycenia stanu obecnego wobec stanu poprzedniego obejmuje równoległe istniejące rozwiązania i praktyki użytkowe – te nowoczesne i złożone, jak i wykorzystywane od kilkunastu lat, prostsze i mniej skomplikowane. Obok możliwości stwarzanych przez współczesne narzędzia wspomagające proces analityczny należy też uwzględnić umiejętności samych użytkowników. Cóż z tego, że dysponujemy zaawansowanymi narzędziami analizy danych – programami, pakietami i algorytmami, skoro poziom kultury informatycznej socjologów jest dalece niewystarczający. Socjologowie, jak i przedstawiciele innych nauk społecznych, sięgający po nowoczesne oprogramowanie są przygotowani do jego wykorzystywania w stopniu bardzo ograniczonym. Umiejętności pisania skryptów, prostych programów czy posługiwania się konsolą, wierszem poleceń nie są tymi, na które kładzie się nacisk w ramach studiów kierunkowych. Socjolog to wszak nie informatyk. Jednakże faktem jest, iż w orbicie zainteresowań socjologów znalazły się narzędzia wymagające takich właśnie kompetencji.

Analityczny przegląd metod i technik prowadzenia badań z wykorzystaniem omawianych nowych technologii napotka wiele trudności. Próby ich dokonania polegają na przyjmowaniu konwencji typologicznych.

Techniki uchwycone w prezentowanym tu krótkim przeglądzie procedur jakościowej (pytanie o ich ewentualną niejakościowość celowo zostaje pominięte i ujęte w nawias) analizy danych wywodzą się z dwóch nurtów. Wskazanie pierwszego (**ujęcie przedmiotowe**) polega na zdekodowaniu akronimu CAQDAS, odnoszącego się do bogatej i rozwijającej się grupy programów użytkowych wykorzystywanych na wiele sposobów w analizach danych jakościowych. W tym miejscu wiele mówi sama nazwa: oprogramowanie wspomagające proces analizy danych jakościowych. Drugi nurt (**ujęcie funkcjonalne**) zdefiniować można poprzez zidentyfikowanie typów przeprowadzanych analiz lub ich celu z uwzględnieniem rozmiaru opracowywanego materiału. Przyjęta na potrzeby dokonywanego przeglądu stosowanych metod i technik prowadzenia badań logika prezentacji odpowiada procesowi poznawania narzędzi CAQDAS i w tym sensie nawiązuje do diachronicznej formuły opisu procesu stawania się użytkownikiem tego typu rozwiązań.

Przetwarzanie danych: podejście manualne

Wśród sposobów wykorzystywania oprogramowania w badaniach jakościowych poczesne miejsce zajmuje ten polegający na funkcjonalnym wyeliminowaniu tradycyjnych nośników informacji: notatnika, papieru, odręcznego kodowania z wykorzystaniem specjalnych arkuszy. Podstawową funkcjonalnością omawianego tu oprogramowania jest klasyfikowanie informacji i nadawanie im zgodnego z preferencjami użytkownika oznaczenia; funkcję tę określa się mianem *code and retrieve*. Zastosowanie oprogramowania CAQDAS pozwala

na wprowadzenie do procesu obróbki i przetwarzania danych charakterystycznego dla techniki komputerowej porządku. Przetwarzane dane porządkowane są w osobliwą budowlę przypominającą drzewo katalogów i podkatalogów, żywcem przypominającą strukturę UNIXopodobnego systemu operacyjnego. Ten typ operacji można wykonać w każdym z dostępnych programów – od najprostszych na przykład OpenCode, RQDA, QDA Miner lite, po te o znacznie większych możliwościach: NVivo, Atlas.ti, QDA Miner i inne. Wyróżnikiem najprostszego modelu używania oprogramowania wspomagającego proces przetwarzania i analizy danych jest manualny tryb kodowania, obejmujący lekturę analizowanego materiału oraz odręczne (choć zapośredniczone interfejsem programu) kodowanie. Ten typ pracy z danymi jest możliwy w sytuacji, gdy objętość zgromadzonego materiału nie przekreśla szans na ich uważne, kilkukrotne przeczytanie. Mowa zatem o nie więcej niż kilkuset stronach tekstu.

Celem procesu kodowania jest zidentyfikowanie kluczowych elementów treści oraz powiązań między nimi. Rozbudowane pole poszukiwań staje się doskonałym obszarem zastosowań dla podstawowych operacji, które człowiekowi zajęłyby godziny, dni i tygodnie, zaś komputerowi sekundy. Użytkownik oprogramowania szybko identyfikuje udogodnienia związane z wyszukiwaniem, porównywaniem, odpytywaniem zbudowanego zbioru danych. Opisany moment przybliży kolejny model wykorzystania oprogramowania związany z wprowadzeniem wspomaganego procesu kodowania i oparcia go o częściowo automatyczne procesy.

Przetwarzanie danych: semiautomatyka

Rozwój metod automatycznego lub półautomatycznego przetwarzania materiału znajduje swe przyczyny w zwiększającej się objętości materiału poddawanej analizie lub ograniczeniach czasowych, uniemożliwiających dokładną lekturę materiału. Jeden i drugi powód w równym stopniu przyczyniają się do powstawania udogodnień pozwalających stworzyć odpowiedni zbiór danych do planowanych analiz.

Proces obróbki danych wiąże się z wykorzystaniem narzędzi wyszukiwania, budowania zapytań, kwerend w obrębie bazy danych z uwzględnieniem szczególnych warunków definiowanych przez analityka. Podejście to pozwala bez konieczności całościowej lektury zidentyfikować fragmenty tekstu spełniające warunki zapytania. Podstawowa implementacja semiautomatycznego przetwarzania tekstu pozwala na częściową redukcję materiału, który należałoby opracować manualnie. Podejście to pozwala jednak na coś więcej niż tylko na wspomnianą redukcję. Wprawne posługiwanie się językiem bazodanowych zapytań oraz operatorów logicznych (znanych każdemu humaniście) w połączeniu z podstawowymi algorytmami maszynowego uczenia się (*machine learning*) umożliwia stworzenie na podstawie definiowanych przez analityka pomiarów wejściowych (np. ręczne zakodowanie kilku słów występujących w określonej relacji) reguł, za pomocą których program przeprowadzi dalszy proces obróbki danych. Programy, które są wyposażone w tak działające moduły, to między innymi QDA Miner z modułami WordStat, SimStat, w nieco mniejszym wymiarze Atlas.ti oraz NVivo. Wśród

wymienianych narzędzi można także wymienić program (choć trafniejszym określeniem jest język programowania) R z zainstalowanym pakietem *tm* (*text mining*) lub ekwiwalentnym.

Przetwarzanie danych: podejście Big Data

Jak już wspomnieliśmy we wcześniejszej części artykułu, obserwowany w ostatnim dziesięcioleciu dynamiczny rozwój cyfrowych technik generowania, przetwarzania i gromadzenia informacji spowodował pojawienie się w rzeczywistości społecznej wielkich zbiorów danych. Przetwarzanie *Big Data* eliminuje, jak już wspomniano, możliwość tradycyjnego oznaczania, indeksowania czy kodowania. Sam proces wstępnego rozpoznania wzorca, pomijając czasochłonność, niósłby ze sobą względu na skalę zbioru danych zbyt duże ryzyko nieuprawnionego selektywnego odczytania, bowiem rekordy bazy danych można liczyć w dziesiątkach, setkach tysięcy czy wręcz w milionach. Identyfikacji wzorców dokonuje się przy pełnej automatyzacji w oparciu o generowane na potrzeby badań lub istniejące słowniki pozwalające rozpoznawać określone sekwencje danych. Równoległe do wspomnianego podejścia rozpoznawania wzorców wykorzystuje się też analizy o charakterze statystycznym. Uwzględnienie częstości występowania słów czy indeksów oraz prawdopodobieństwa wystąpień określonych struktur językowych pozwala na wysnuwanie wniosków nie tylko odnośnie profilu syntaktycznego tekstu, ale także semantycznego. Stąd też dla analiz określanych połączonym mianem CAQDAS i *Big Data* fundamentalne znaczenie ma wywodzący się z nauk informatycznych zespół praktyk programistycznych i analitycznych określany jako

przetwarzanie języka naturalnego (NLP – *natural language processing*).

Wskazanie programów zagospodarowujących poszczególne techniki *Big Data* jest nie lada wyzwaniem. Wiąże się to z faktem, iż przedsięwzięcia badawcze tego rodzaju obejmują wiele procedur. Złożoność procesu wyjawia już samo wskazanie, iż chodzi o pozyskanie danych (np. *webscraping*, *webcrawling/onlinecrawling*), ich wstępne odczytanie i przygotowanie (dzielenie na kolumny, strony, dekodowanie znaków – analiza składniowa, tzw. *parsing*), identyfikację wzorców (model nadzorowany lub nienadzorowany, z wykorzystaniem algorytmów, np. LDA, STM, lub z wykorzystaniem algorytmów klastrowania), wizualizację zagregowanych danych w formie chmur tagów, dendrogramów czy z wykorzystaniem map różnego rodzaju. Środowiskiem programistycznym, do którego sięga wielu badaczy, jest to związane z językiem programowania R. Różnorodność pakietów uzupełniających funkcjonalność programu pozwala na zbudowanie narzędzia dostosowanego do potrzeb analityka.

Celem przyświecającym zaprezentowanemu tu krótkiemu przeglądowi metod i technik wykorzystywanych w badaniach jakościowych ze wspomaganie komputerowym nie była kompleksowa systematyka istniejących rozwiązań. Przegląd ten spełnić miał zadanie wyznaczenia punktu lub punktów zaczepienia dla badacza poszukującego sposobu na pokonanie trudności przewidywanych w planowanych działaniach badawczych. Tego typu odbiorcy należy się też pewna odpowiedź. Otóż do tej pory nie stworzono takiego programu, który wszystko wykonałby za badacza. Droga, jaką trzeba pokonać

pomiędzy pytaniem, hipotezą, falsyfikacją lub ich odpowiednikami w różnych tradycjach metodologicznych, nadal pozostaje domeną socjologa-badacza. Elementem owej drogi jest oczywiście również analizowanie zebranych danych – zagadnieniu temu poświęcimy kolejną część artykułu.

Big Data – ich źródło oraz proces analizy

Jak zauważyliśmy już w prezentowanym artykule, pojawienie się *Big Data* oznacza zmianę metod pozyskiwania, gromadzenia, zapisywania i analizowania informacji, które wpływają na sposób rozumienia i organizacji społeczeństwa. Świat życia codziennego zalewany jest ogromną liczbą różnych skwantyfikowanych i zdigitalizowanych danych społecznych, ekonomicznych, kulturowych, geograficznych, technologicznych, których liczba rośnie codziennie lawinowo. Informacja staje się wartością wtedy, gdy możemy ją przekuć w formę danych, by potem poddać szczegółowej obróbce i analizie. Koncepcję tę określa się mianem danetyzacji, czyli takiego sposobu przetwarzania informacji, który umożliwia ich późniejsze wykorzystanie i dalszą analizę danych (Mayer-Schoberger, Cukier 2014: 103–132). Doskonałym przykładem danetyzacji są rozwijane w naukach humanistycznych i społecznych korpusy tekstowe wykorzystywane do analiz lingwistycznych, a także zbiory danych tekstowych pochodzące z mediów społecznościowych (Facebook, Twitter), umożliwiające na przykład analizę dyskursu czy analizę sentymentu. Przekształcanie danych tekstowych: książek, dokumentów, zdań czy słów w dane pozwala na rozwijanie różnorodnych sposobów ich użycia, analizowania oraz tworzenia modeli teoretycznych i predykcyjnych. Bez cyfryza-

cji i danetyzacji nie byłoby *Big Data*, a także analiz typu CAQDAS w badaniach jakościowych. Jednakże *Big Data* to coś więcej niż proces digitalizacji danych.

Big Data jest zaliczane do działu informatyki nazywanej sztuczną inteligencją, a dokładniej tak zwanych systemów uczących się, jednakże w praktyce podejście to nie polega na uczeniu komputerów tak, by myślały jak ludzie. Sednem *Big Data* jest zdolność do przewidywania, szacowania prawdopodobieństwa wystąpienia określonych zdarzeń, zachowań lub sytuacji dzięki zastosowaniu matematyki do przetwarzania ogromnej liczby danych ustrukturyzowanych i nieustrukturyzowanych. Jednocześnie systemy analityczne *Big Data* są tak zaprojektowane, by rozwijać własne modele predykcyjne przez monitorowanie zdarzeń i procesów dostarczających nowych danych z otoczenia. Przykładowo Amazon zaleca idealną książkę, Google wyszuka odpowiednią stronę, Facebook wie, kogo/co lubimy, LinkedIn odgadnie, kogo znamy. *Big Data* oznacza zdolność do korzystania z informacji zawartej w danych w nowatorski sposób, który ułatwia lepsze zrozumienie rzeczywistości oraz tworzenie dóbr i usług o znacznej wartości, a także przewidywanie, na podstawie modeli analitycznych, ścieżek przebiegu procesów społecznych i biznesowych.

Opierając się na koncepcji *Big Data*, możemy przetwarzać wszystkie dane dotyczące badanego zjawiska/procesu, jakie tylko jesteśmy w stanie o tym zjawisku/procesie zebrać. Obecnie ograniczenia w gromadzeniu i analizie danych nie są już takim problemem. Analizie podlegają wszystkie z/groma-

zione (zdigitalizowane) dane bez potrzeby dobierania próby losowej i szacowania błędu ekstrapolacji wyników. Użycie wszystkich danych pozwala dostrzec szczegóły, z których wcześniej nie zdawaliśmy sobie sprawy, ponieważ byliśmy przyzwyczajeni do redukcji liczby danych. Koncepcja *Big Data* nie wymaga schematu doboru próby, zarówno losowej, jak i nielosowej. Większe spektrum danych nie wymaga także zachowania dużej dokładności⁵. Gdy możliwości pomiaru są ograniczone, skupiamy się na tym, co istotne. Wzrost skali danetyzacji sprzyja niedokładności pomiaru, ale w konsekwencji zwiększa się możliwość wnioskowania. Dokładność, precyzja „pomiaru” wymaga dobrze przygotowanych danych, co sprawdza się raczej w przypadku małej liczby danych. W koncepcji *Big Data* rezygnujemy ze sztywnej precyzji na rzecz ogólnej tendencji, poznania kierunku rozwoju jakiegoś zjawiska czy procesu. Nie oznacza to jednak rezygnacji z precyzji, ale nie jest ona priorytetem. W koncepcji *Big Data* rezygnujemy z niedoskonałości próby losowej na rzecz kompletności danych. Tracimy na dokładności, ale zyskujemy lepsze zrozumienie, wgląd w określone zjawisko. W *Big Data* istotne jest bowiem poszukiwanie i rozumienie związków między danymi, których do tej pory nie byliśmy w stanie pojąć. Zmiana skali dostępności danych spowodowała zmianę ich statusu. Zmiana ilości doprowadziła do zmiany jakości, prób zrozumienia istoty badanego zjawiska czy procesu, struktury danych, wydobycia wiedzy zawartej w danych bez potrzeby sięgania w głąb, poznawania szczegółowych cech czy istoty jakie-

⁵ Dzięki mniejszej liczbie błędów wynikających z doboru próby losowej możemy zaakceptować większą liczbę błędów pomiaru, a tym samym mniejszą dokładność pomiaru.

goś zjawiska lub procesu (choć – zgodnie z założeniami nowego paradygmatu opisywanego na początku tekstu – z uwzględnieniem informacji kontekstowych oraz teorii). *Big Data* wymaga od badacza interakcji z danymi. Kiedy pozwalamy „przemówić danym”, odkrywamy powiązania, których istnienia nie podejrzewaliśmy. W erze analogowej zbieranie i analiza danych pochłaniały zazwyczaj dużo czasu, nowe pytania badawcze wymagały konieczności ponownego odtworzenia procesu zbierania danych i analizy. W erze *Big Data* wraz z digitalizacją danych i możliwością przetwarzania analogowych informacji w sposób zrozumiały dla komputerów z jednej strony nastąpił postęp w dziedzinie zarządzania danymi, ich analizy i tworzenia modeli analitycznych, zaś z drugiej, jeśli chcemy analizować duże ilości danych, musimy się pogodzić z niedokładnością i niepewnością naszego wnioskowania.

CAQDAS, Data/Text Mining a proces analizy danych jakościowych

Przywołanemu wcześniej zjawisku danetyzacji świata życia codziennego towarzyszy rozwój nowych algorytmów, technik analitycznych oraz technologii informatycznych w zakresie przetwarzania i analizy danych, w tym lingwistyki komputerowej i sztucznej inteligencji. Danetyzacja wymaga metodologii służących rozwijaniu modeli umożliwiających kompleksową analizę zjawisk lub procesów. Kluczowe znaczenie odgrywa w tym rozwoju eksploracja danych (ang. *Data Mining*), określana także jako drążenie danych, pozyskiwanie wiedzy, wydobycie danych, ekstrakcja wiedzy zawartej w danych. *Data Mining* to podstawowy etap procesu

odkrywania wiedzy w bazach danych (ang. KDD, *Knowledge Discovery in Databases*). Logika KDD zawiera się w sekwencji następujących etapów: zrozumienia danych, wyboru danych do analizy, wstępnego przetworzenia danych, przekształcenia danych do analizy, przeprowadzenia eksploracji w celu odkrycia struktury wzorców i zależności, konstruowania modeli analitycznych, oceny stopnia dopasowania modeli do danych, a następnie oceny i interpretacji wyników pod kątem uzyskanej wiedzy. Nie ma jednoznacznej, ogólnie przyjętej definicji eksploracji danych. Większość definicji zwraca jednak uwagę na trzy rzeczy: analizę dużych zbiorów danych (*Big Data*), poszukiwanie struktury zależności między danymi i wizualizację jako formę reprezentacji wyników. Dane w koncepcji *Big Data* nie są traktowane jako coś statycznego, jako takie, których przydatność kończy się wraz z ich zgromadzeniem, lecz jako struktury dynamiczne, dlatego też mogą być wielokrotnie wykorzystywane w inteligentny sposób z użyciem zaawansowanych algorytmów i technik analitycznych odnoszących się właśnie do metod i technik eksploracji oraz klasyfikacji. Z racji tego, że analizie poddaje się wszystkie dane, to wszelkiego rodzaju odstępstwa od normy, przypadki nietypowe stają się ważnymi informacjami w zrozumieniu istniejących w zbiorze danych zależności.

W analizach typu *Big Data* kluczową rolę odgrywa wspomniany już w pierwszej części artykułu zwrot w kierunku poszukiwania korelacji, przy jednoczesnej rezygnacji z poszukiwania przyczynowości. Dotychczasowe analizy w badaniach jakościowych ograniczały się do sprawdzania niewielkiej liczby hipotez, które formułowane były zgodnie z logiką

dedukcyjną, przed zebraniem danych⁶ lub indukcyjną, rodziły się w trakcie procesu analizy danych (tak jak na przykład w metodologii teorii ugruntowanej). W analizach typu *Big Data* (także w badaniach jakościowych) nie musimy się skupiać na poszukiwaniu zależności przyczynowo-skutkowych, naszym celem jest odkrywanie relacji między zdarzeniami, faktami, sądami, zachowaniami i tym podobnymi, które umożliwiają poznanie określonego zjawiska lub procesu. Skupienie się na poszukiwaniu korelacji między danymi nie prowadzi do wyjaśnienia dlaczego coś się dzieje, ale pozwala stwierdzić, co się dzieje, z czym mamy do czynienia, jaka jest skala zjawiska. Punktem wyjścia w analizie jest przede wszystkim zrozumienie danych i relacji między nimi (odkrywanie struktury relacji w procesie analizy), a dopiero w konsekwencji poszukiwanie wyjaśnienia zależności między nimi. To ostatnie jest oczywiście ważne przy wspomnianym na początku artykułu odrzuceniu tezy o „śmierci teorii”. Opisany proces zrozumienia danych powinien być następnie wsparty poszukiwaniem informacji „kontekstowych” lub/i znaleźć podbudowę w teorii.

W podejściu *Big Data* wykorzystuje się między innymi techniki statystyczne (statystyki opisowe, tabele kontyngencji, analizę czynnikową, dyskryminacyjną, hierarchiczną analizę skupień, regresję logistyczną itp.), techniki uczenia maszynowego, sieci neuronowe, algorytmy indukcyjne, genetyczne czy drzewa klasyfikacyjne w celu odkrywania wiedzy zawartej w danych i tworzenia wielowymiaro-

⁶ „Hipotezy” powstają w drodze eksploracji danych, jako efekty identyfikacji systematycznych relacji pomiędzy zmiennymi w sytuacji, gdy natura tych relacji nie jest z góry określona. Stąd drażnienie danych utożsamia się zazwyczaj z podejściem indukcyjnym do odkrywania wiedzy.

wych modeli predykcyjnych. Współcześnie procesy eksploracji danych znajdują na przykład zastosowanie w analizie danych o ruchu internetowym (analiza logów), rozpoznawaniu sygnałów obrazu, mowy, pisma, sensu wyrazów i zdań, struktur chemicznych, stanu zdrowia człowieka, wspomaganie diagnostyki medycznej, biologii i badaniach genetycznych, analizie operacji bankowych, prognozowaniu wskaźników ekonomicznych, pogody, plam na Słońcu, aż po zagadnienia z zakresu kognitywistyki, doświadczeń psychologicznych, analizy sposobu rozumowania i kategoryzacji, poruszania się i planowania i tym podobne.

Jak wcześniej wspominaliśmy, *Big Data* dotyczy trzech zmian w podejściu do analizy informacji, które uzupełniają się i wzmacniają wzajemnie: możliwości analizowania dużej liczby danych z określonej dziedziny, braku konieczności ograniczania się do mniejszych zbiorów (stosowania doboru próby), gotowości do zajmowania się nieuporządkowanymi danymi płynącymi z rzeczywistego świata i nieprzywiązywania zbyt dużej wagi do ich dokładności. Dlatego w analizach typu *Big Data* eksploracja danych poprzedza eksplanację, a zrozumienie tego, co tkwi w danych, potrzebę poszukiwania relacji przyczynowych. Logika tego podejścia wydaje w pełni odpowiadać myśleniu badaczy jakościowych, z racji tego, że w badaniach jakościowych mamy najczęściej do czynienia z dużą swobodą pozyskiwania danych, a same dane jakościowe są zwykle danymi nieustrukturyzowanymi. Niestety wciąż brakuje w środowisku analityków i badaczy jakościowych w Polsce pogłębionej refleksji nad analizami *Big Data*, a także możliwościami wykorzystywania metod i technik eksploracji danych

jakościowych oraz odkrywania wiedzy w obszarze badań jakościowych.

Podobnie rzecz ma się z CAQDAS. W ciągu ostatnich dwóch dekadach, wraz z rozwojem technologii informatycznych, zwiększa się świadomość badaczy jakościowych dotycząca korzystania z oprogramowania CAQDAS, szczególnie w analizie wywiadów socjologicznych (Bryda 2014a). Pomimo że rdzeń współczesnej analizy danych jakościowych stanowią wciąż procedury teorii ugruntowanej, zaimplementowane w wielu programach CAQDAS⁷, to dzięki procesowi digitalizacji danych i danetyzacji samych badań jakościowych, czego przykładem jest tworzenie korpusów dokumentów tekstowych czy archiwów danych jakościowych, większe znaczenie w procesie analiz jakościowych zaczęła odgrywać analiza treści (Berelson 1952; Holsti 1969; Brent 1984; Weber 1990; Krippendorff 2004), wzbogacona o najnowsze osiągnięcia w dziedzinie lingwistyki komputerowej. Jeśli prześledzimy pojawianie się nowych funkcjonalności w programach CAQDAS na przestrzeni ostatnich kilkudziesięciu lat, to zobaczymy, że rozwój wspomaganej komputerowo analizy danych jakościowych w kierunku *DataMining* czy *TextMining* (Wiedemann 2013; Bryda 2014b) nie byłby możliwy bez rozwoju technik ilościowej i jakościowej analizy treści, metod mieszanych (Tashakkori, Teddlie 2003), a także metodologii eksploracji danych tekstowych i odkrywania wiedzy (Hand, Mannila, Smyth 2005; Larose 2006; 2008). W badaniach jakościowych procesowi temu towarzyszy wyraźny zwrot metodologiczny w kierunku

⁷ Teoria ugruntowana wytyczyła nie tylko wzorce i procedury przeprowadzania analiz jakościowych, ale jej założenia metodologiczne stały u podstaw rozwoju wielu obecnych funkcjonalności programów CAQDAS.

paradygmatu *mixed-methods*. Jego wyrazem jest przechodzenie od klasycznej analizy danych jakościowych (*Qualitative Analysis*), przez *Qualitative Content Analysis*, w kierunku pogłębionej eksploracji danych jakościowych (Bryda 2014b) i *TextMining* wykorzystującej techniki statystyczne i algorytmy z dziedziny inteligencji komputerowej czy przetwarzania języka naturalnego (Bryda, Tomanek 2014). *TextMining* ma korzenie w rozwijającej się od kilkunastu lat metodologii *Data Mining*, ale obecnie staje się podstawą wielu analiz jakościowych i rozwoju funkcjonalności we wspomaganej komputerowo analizie danych jakościowych (Ho Yu, Jannasch-Pennell, DiGangi 2011). Rozwój CAQDAS w kierunku wykorzystania zaawansowanych metod eksploracji i odkrywania wiedzy w danych (głównie tekstowych) jest możliwy nie tylko dzięki zastosowaniu nowych technologii informatycznych, ale przede wszystkim dzięki ewolucji świadomości analitycznej badaczy jakościowych i metodologii prowadzenia analizy danych jakościowych w kierunku *Big Data*, gdzie główną rolę odgrywa poszukiwanie korelacji i prawdopodobieństwo.

Big Data, CAQDAS w praktyce badawczej

Wpływ *Big Data*, CAQDAS i nowych technologii na proces badań jakościowych, sposób zbierania i analizy danych staje się coraz bardziej widoczny. Również w Polsce mamy do czynienia z rosnącym zainteresowaniem świata akademickiego, jak też podmiotów rynkowych problematyką *Big Data* oraz możliwościami wykorzystywania oprogramowania CAQDAS w projektowaniu i prowadzeniu badań, a także analizie danych jakościowych. Jak dotąd za pomocą programów CAQDAS analizuje się głównie dane tekstowe, takie jak transkrypcje wywiadów, teksty prasowe

czy notatki z obserwacji. *Big Data* wnosi jednak nowe rodzaje nieustrukturyzowanych danych dotyczących interakcji (Facebook, Twitter), a także inny niż dotychczas sposób myślenia o samych danych i sposobie ich analizowania.

Nawiązując do wcześniejszych rozważań, jako redaktorzy tego tomu, chcielibyśmy przybliżyć zagadnienie *Big Data* i CAQDAS w praktyce. Oddajemy do rąk czytelników tom „Przeglądu Socjologii Jakościowej” w całości poświęcony tej problematyce. Publikacja zawiera teksty przygotowane przez badaczy i praktyków, których kompetencje w zakresie *Big Data* i pracy z programami CAQDAS oparte są na połączeniu rzetelnej wiedzy i doświadczenia.

Publikację rozpoczyna tekst Mariusza Dziegłewskiego dotyczący korzyści i ograniczeń w wykorzystywaniu oprogramowania CAQDAS w badaniach digitalizacji i odbioru dziedzictwa kulturowego. Autor poddaje refleksji problem łączenia i przenikania się tradycyjnych metod badania i procedur analizy ze wspomaganą komputerowo analizą danych jakościowych. Opisuje on rolę, jaką w projektowaniu badań odgrywa CAQDAS, sposób, w jaki wpływa to oprogramowanie na percepcję problemu badawczego oraz interpretację wyników badania, a także problematykę przenikania się różnych podejść metodologicznych i analitycznych na różnych etapach projektu badawczego: budowania bazy, kodowania danych, analizy, wizualizacji i interpretacji wyników. Analizując możliwości oraz ograniczenia wynikające ze stosowania CAQDAS dla analizy treści dokumentów prawnych i transkrypcji wywiadów pogłębionych, autor poszukuje optymalnego połączenia tradycyjnych i nowoczesnych metod badania oraz analizy danych, które pozwoliłoby na

uniknięcie „pułapek” związanych z wykorzystaniem nowych technologii w badaniach społecznych.

Kolejny artykuł – Jakuba Niedbalskiego – ma charakter poglądowy i edukacyjny. Autor stawia sobie za cel zapoznanie czytelników z możliwościami NVivo, narzędzia należącego do rodziny CAQDAS oraz jego faktycznym zastosowaniem w projektach realizowanych zgodnie z założeniami metodologii teorii ugruntowanej. Autor pokazuje, w jaki sposób można wykorzystać narzędzia komputerowego wspomaganie analizy danych jakościowych w praktyce badawczej. Na przykładzie konkretnego projektu badawczego przybliża etapy pracy w programie NVivo zgodnie z procedurami metodologii teorii ugruntowanej, wskazując na istniejące udogodnienia i potencjalne trudności związane ze stosowaniem oprogramowania komputerowego jako elementu warsztatu badacza jakościowego.

W artykule dotyczącym „mowy nienawiści” i wykorzystania algorytmów uczenia maszynowego w analizie danych jakościowych Marek Troszyński zajmuje się procesem automatyzacji kodowania (anotacji i tagowania) danych tekstowych pochodzących z forów internetowych w oparciu o znaczenia zawarte w tekście. Wdrożenie tego procesu pozwala na ilościowe analizy korpusów danych tekstowych liczących setki tysięcy tekstów. Autor skupia uwagę na procesie konceptualizacji i operacjonalizacji „mowy nienawiści”, przygotowaniu dokładnej instrukcji kodowej oraz treningu zespołu kodującego w celu uzyskania wysokiego współczynnika zgodności między kodami. Następnie przedstawia zastosowane metody kodowania automatycznego, wskazując czynniki, które są kluczowe dla procesu badawczego wykorzystującego uczenie maszynowe.

Problematyka radykalizacji i brutalizacji języka, nadużywania słów nacechowanych negatywnie w dyskursie politycznym i o polityce, a także degradacji znaczenia tych słów stanowi przedmiot zainteresowania Agnieszki Kwiatkowskiej, która przedstawia możliwości zastosowania modeli generatywnych do analizy debat parlamentarnych. W artykule analizuje ona zbiór przemówień sejmowych z lat 1991–2016 odnoszących się do idei hańby, zdrady, niesławy i skandalu. W tym celu wykorzystuje nienadzorowane algorytmy przeszukiwania korpusów tekstów oraz analizy ukrytych tematów, w tym generatywny model tematyczny, metodę ukrytej alokacji Dirichleta i jej rozszerzenie – strukturalny model tematyczny jako metodę ekstrakcji tematów w dużych korpusach danych tekstowych.

W kolejnym artykule: *Dobra zmiana czy Polska w ruinie?* Alicja Zawistowska i Małgorzata Skowrońska przeprowadziły analizę ewolucji znaczeniowej wpisów opatrzonych hashtagami #dobrazmiana i #polskawruinie zamieszczonych w serwisie społecznościowym Twitter. Autorki poddają analizie wpisy, które pojawiły się w latach 2015–2016. Celem tej analizy jest ukazanie dynamiki zabarwienia emocjonalnego obu haseł, a także ukazanie wpływu dominującego na Twitterze stylu komunikacji na wspomnianą ewolucję znaczenia tych wpisów. W artykule przedstawiono również podstawowe problemy metodologiczne związane z zastosowaniem analizy treści w mediach społecznościowych. Publikację kończy artykuł Krzysztofa Tomanka dotyczący metodyki analizy treści w projektach stosujących techniki *TextMining* i oprogramowanie CAQDAS. Autor wskazuje przykładowe dylematy metodologiczne występujące w trakcie pracy z dużymi wolumenami

danych tekstowych pochodzących z różnych źródeł i zapisanych w różnorodnych formatach, zwracając uwagę w szczególności na problem jakości danych nieustrukturyzowanych typu *quan* i *qual*. Na przykładzie własnego projektu przedstawia zastosowanie metody analizy danych wykorzystującej różnorodne narzędzia CAQDAS do (pół)automatycznej klasyfikacji wypowiedzi pisanych wtedy, gdy mamy do czynienia z danymi o różnorodnej jakości. Próbuje również pokazać, kiedy klasyfikacja (pół)automatyczna jest przydatna, a kiedy nie ma szans powodzenia oraz momenty, w których badacz jakościowy wykorzystuje wiedzę z innych dziedzin: przetwarzanie języka naturalnego czy uczenie maszynowe w procesie analizy danych.

Bibliografia

Anderson Chris (2008) *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. „Wired”, 16 lipca [dostęp 26 kwietnia 2017 r.]. Dostępny w Internecie: <http://www.uvm.edu/~cmplsys/wordpress/wp-content/uploads/reading-group/pdfs/2008/anderson2008.pdf>.

Berelson Bernard (1952) *Content Analysis in Communication Research*. Glencoe, IL: Free Press.

Brent Edward E. (1984) *Qualitative Computing: Approaches and Issues*. „Qualitative Sociology”, vol. 7 (1/2), s. 36–60.

Bryda Grzegorz (2014a) *CAQDAS a badania jakościowe w praktyce*. „Przegląd Socjologii Jakościowej”, t. 10, nr 2, s. 12–38. Dostępny w Internecie: www.przegladsocjologiijakosciowej.org.

Bryda Grzegorz (2014b) *Caqdas, Data Mining i odkrywanie wiedzy w danych jakościowych* [w:] Jakub Niedbalski, red., *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*. Łódź: Wydawnictwo UŁ, s. 13–40.

Wśród zagadnień, którym autorzy poświęcili swoje teksty, na szczególną uwagę zasługują: podjęcie dyskusji nad zgodnością zasad, na jakich funkcjonuje oprogramowanie CAQDAS, z regułami oraz procedurami metodologii badań jakościowych; możliwości zastosowania narzędzi CAQDAS w realizacji projektów badawczych opartych na różnych metodach jakościowych i w ramach różnych podejść analitycznych; zgodności „architektury oprogramowania” z procedurami wybranych metod i technik badawczych; wpływu *Big Data* i nowych technologii na proces badawczy, implementacji nowych algorytmów i technik; wpływu rozwiązań wykorzystywanych w innych dziedzinach nauki na proces analizy i badań opartych na metodach jakościowych.

Bryda Grzegorz, Tomanek Krzysztof (2014) *Od CAQDAS do TextMiningu. Nowe techniki w analizie danych jakościowych* [w:] Jakub Niedbalski, red., *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*. Łódź: Wydawnictwo UŁ, s. 191–218.

Chang Ray M., Kauffman Robert J., Kwon Young Ok (2013) *Understanding the Paradigm Shift to Computational Social Science in the Presence of Big Data*. „Decision Support Systems”, vol. 63, s. 67–80.

Chen C. L. Philip, Zhang Chun-Yang (2014) *Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data*. „Information Sciences”, vol. 275, s. 314–347.

Hand David, Mannila Heikki, Smyth Padhraic (2005) *Eksploracja danych*. Przełożyła Agnieszka Chądzyńska. Warszawa: WNT.

Hey Tony, Tansley Steward, Tolle Kristin (2009) *Jim Gray on eScience: A Transformed Scientific Method* [w:] Hey Tony, Tan-

sley Steward, Tolle Kristin, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft Research, s. xvii–xxxii.

Ho Yu Chong, Jannasch-Pennell Angel, DiGangi Samuel (2011) *Compatibility between Text Mining and Qualitative Research in the Perspectives of Grounded Theory, Content Analysis, and Reliability*. „The Qualitative Report”, vol. 16, no. 3, s. 730–744.

Holsti Ole R. (1969) *Content Analysis for the Social Sciences and the Humanities*. Reading, MA: Addison-Wesley.

Kitchin Rob (2014) *Big Data, New Epistemologies and Paradigm Shifts*. „Big Data & Society”, April-June, s. 1–12.

Krippendorff Klaus (2004) *Content Analysis. An Introduction to Its Methodology*. Thousand Oaks, CA: Sage.

Larose Daniel T. (2006) *Odkrywanie wiedzy z danych: wprowadzenie do eksploracji*. Przełożyła Anna Wilbik. Warszawa: PWN.

Larose Daniel T. (2008) *Metody i modele eksploracji danych*. Przełożyła Anna Wilbik. Warszawa: PWN.

Masterman Margaret (1970) *The Nature of a Paradigm* [w:] Imre Lakatos, Alan. E. Musgrave, eds., *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press, s. 59–90.

Cytowanie

Brosz Maciej, Bryda Grzegorz, Siuda Piotr (2017) *Od redaktorów: Big Data i CAQDAS a procedury badawcze w polu socjologii jakościowej*. „Przegląd Socjologii Jakościowej”, t. 13, nr 2, s. 6–23 [dostęp dzień, miesiąc, rok]. Dostępny w Internecie: www.przegladsocjologiijakosciowej.org.

Big Data, CAQDAS and research procedure in the field of qualitative research

Abstract: The reality of everyday life is covered by huge amounts of various quantified and digitized data. The quantity of data grows everyday enormously. These data can be processed and treated as research material, also qualitative. The application of Big Data in qualitative research modifies the procedure on every step of research process: from research design up to conclusion. Does implementing Big Data strategy into qualitative research lead to atheoretical approach? What are the consequences of using the complete data sets instead of random sample technique? The purpose of this article is to indicate this changes and their brief characteristics considering the significant role of different kind of software (especially CAQDAS), and so the analysis that can be conducted.

Keywords: big data, CAQDAS, computer-aided qualitative data analysis, data processing, datafication, qualitative data

Mariusz Dziegłowski
Uniwersytet Pedagogiczny im. KEN

CAQDAS w badaniach digitalizacji i odbioru dziedzictwa kulturowego. Korzyści i ograniczenia

Abstrakt Celem artykułu jest refleksja na temat korzyści i ograniczeń płynących z połączenia i wzajemnego przenikania się tradycyjnych metod badania i analizy ze wspomaganą komputerowo analizą danych jakościowych. Artykuł zawiera krytyczną analizę tego procesu przeprowadzoną na przykładzie projektu badawczego pod tytułem „Cyfrowe praktyki i strategie upowszechniania i odbioru dziedzictwa kulturowego w Polsce w latach 2004–2014”. W pierwszej części artykułu omówione zostały założenia projektu ze szczególnym uwzględnieniem roli i sposobu łączenia tych podejść na etapie organizacji badań. Autor artykułu opisuje rolę, jaką w projektowaniu badań odgrywa CAQDAS i sposób, w jaki wpływa ona na percepcję problemu badawczego, oraz późniejszą interpretację wyników badania. W drugiej części artykułu przedstawiono poszczególne etapy badania, ze szczególnym uwzględnieniem przenikania się różnych podejść do badań i analizy na etapie budowania bazy danych, kodowania danych tekstowych, analizy, wizualizacji danych i interpretacji wyników. W sposób szczególny omówiono korzyści i ograniczenia zastosowania CAQDAS do analizy treści i zawartości dokumentów prawnych jako danych zastanych oraz transkrypcji wywiadów pogłębionych. Głównym celem metaanalizy przebiegu badania jest propozycja optymalnego połączenia tradycyjnych i nowoczesnych metod badania i analizy danych w złożonych projektach, które pozwoliłoby na wykorzystanie nowych możliwości technologicznych w badaniach społecznych, a jednocześnie – na uniknięcie „pułapek” i ograniczeń z nimi związanych.

Słowa kluczowe CAQDAS, wywiady pogłębione, analiza treści, digitalizacja, dziedzictwo kulturowe

Mariusz Dziegłowski, dr, socjolog, adiunkt w Instytucie Filozofii i Socjologii Uniwersytetu Pedagogicznego im. KEN w Krakowie. Jego zainteresowania naukowe koncentrują się wokół socjologii kultury, procesów migracji zagranicznych oraz komunikowania społecznego. Jest autorem książki *Reportaże Ryszarda Kapuścińskiego – źródło poznania społeczeństwa i kultur* (Lublin 2009). Dotychczas opublikował kilkadziesiąt artykułów naukowych, raportów, prac pod redakcją z zakresu różnych aspektów migracji międzynarodowych, tożsamości społeczno-kulturowej, dziedzictwa kulturowego, socjologii wizualnej, narracji i wizerunku w mediach. Publikuje w czasopiśmie zagranicznych takich jak

„Media, Culture & Society” oraz w czasopiśmie polskich, jak: „Studia Migracyjne – Przegląd Polonijny”, „Polish Sociological Review”, „Studia Socjologiczne”. W swojej pracy badawczej wykorzystuje narzędzia CAQDAS, w szczególności do analizy danych zastanych (dokumenty, materiały wizualne), analizy treści (artykuły prasowe, seriale) oraz transkrypcji wywiadów pogłębionych.

Dane kontaktowe:

Uniwersytet Pedagogiczny im. KEN, Instytut Filozofii i Socjologii
ul. Podchorążych 2, 30-084 Kraków
e-mail: mdziegłowski@wp.pl

Oprogramowanie wspierające analizę danych jakościowych (CAQDAS) jest coraz częściej wykorzystywane w polskich badaniach socjologicznych. Jednocześnie – w odróżnieniu od krajów, gdzie pierwsze aplikacje wykorzystywano już w latach 60. i opisywano w literaturze akademickiej w latach 80. i 90. – polscy badacze mają stosunkowo niewielki „staż” w ich zastosowaniu. Ponadto literaturę na temat wykorzystania nowych technologii w praktyce badawczej w języku polskim również uznaje się za skromną. Nie dziwi zatem fakt, iż wokół CAQDAS narosło wiele nieuzasadnionych uogólnień. Konsekwencją tych generalizacji jest widoczny podział na zwolenników i przeciwników komputerowego wsparcia w analizie danych jakościowych. Grzegorz Bryda zwraca uwagę na argumenty, które najczęściej podają zwolennicy zastosowania CAQDAS, takie jak: efektywność, możliwość analizy dużej liczby obserwacji i konstrukcji „gęstego” klucza kategoryzacyjnego, łatwe zarządzanie danymi i ich archiwizacja. Przeciwnicy komputerowego wsparcia analizy danych jakościowych wskazują na dekontekstualizację kodowanych komputerowo treści, analizę skrojoną według schematu badań ilościowych i brak przestrzeni na nieustrukturyzowaną, intuicyjną refleksję badacza (Bryda 2014). Stopniowo jednak coraz częściej pojawiają się grupy tematyczne w ramach konferencji ogólnokrajowych, w ramach których badacze dzielą się swoimi doświadczeniami z wykorzystania CAQDAS w konkretnych projektach badawczych. Dzięki takiej wymianie doświadczeń wzrasta świadomość korzyści i ograniczeń związanych z CAQDAS. Coraz liczniejsze publikacje naukowe w tym obszarze pozwalają lepiej zrozumieć istotę komputerowo wspomaganą analizy danych jakościowych oraz uniknąć stereotypowych generalizacji.

Anglojęzyczne opracowania związane z wykorzystaniem komputerowego wsparcia badań społecznych pojawiają się już w latach 80. i koncentrują się głównie na wspomaganiu tradycyjnych technik, takich jak wywiad czy analiza treści (Becker, Gordon, LeBailly 1984; Brent 1984; Gerson 1984; Pfaffenberger 1988). Punktem zwrotnym w krytycznej refleksji nad zastosowaniem wsparcia komputerowego w badaniach społecznych okazał się powołany do życia w 1994 roku na University of Surrey projekt „CAQDAS Network Project”. Projekt ten, poza integracją środowiska pionierów stosujących nowe rozwiązania technologiczne w badaniach socjologicznych, przyczynił się w istotny sposób do powstania kluczowych publikacji, które uwzględniły zarówno doświadczenia metodologów, badaczy-praktyków, jak i producentów oprogramowania wspierającego badania jakościowe. Czołową rolę w tym nurcie refleksji nad wsparciem komputerowym w badaniach odgrywają Raymond Lee i Nigel Fielding, którzy wprowadzili do powszechnego użycia termin CAQDAS (Fielding, Lee 1993; 1996; 1998). Badacze ci skoncentrowali się przede wszystkim na istotnych możliwościach i ograniczeniach wykorzystania CAQDAS w procesie badań jakościowych¹.

Polskie opracowania zagadnień związanych z wykorzystaniem CAQDAS pojawiły się dopiero w ostatniej dekadzie XXI wieku i wciąż są one nieliczne. Jako jedną z pierwszych prac akademickich poświęconych komputerowej analizie danych jakościowych wymienić należy artykuł Jacka

¹ Szczegółowe omówienie tych zagadnień znaleźć można w: Bryda Grzegorz (2014) *CAQDAS a badania jakościowe w praktyce*. „Przegląd Socjologii Jakościowej”, tom 10, nr 2, s. 12–37.

Bielińskiego, Katarzyny Iwińskiej i Anny Rosińskiej-Kordasiewicz (2007) na łamach czasopisma „ASK”. Kolejną ważną pracą jest publikacja Macieja Brosza dotycząca analizy danych nieustrukturyzowanych przy wsparciu oprogramowania NVivo (Brosz 2012). Wśród opracowań najczęściej cytowanych przez polskich badaczy wymienić należy prace Jakuba Niedbalskiego. Publikacje te obejmują całe spektrum zagadnień związanych z komputerowym wsparciem analizy danych jakościowych, od oprogramowania open-source (Niedbalski 2012; 2013), poprzez wykorzystanie dedykowanego oprogramowania NVivo i Atlas.ti (Niedbalski, Ślęzak 2012; Niedbalski 2014a) do szczegółowych zagadnień związanych z wykorzystaniem CAQDAS. Wysiłki badaczy zainteresowanych popularyzacją wiedzy na temat CAQDAS w środowisku polskich badaczy jakościowych zaowocowały dwiema ważnymi pracami zbiorowymi. Publikacje te zawierają artykuły dotyczące zagadnień związanych z wykorzystaniem wsparcia komputerowego na różnych etapach pracy badacza (kategoryzacja obserwacji, kodowanie, analiza, wizualizacja danych, raportowanie i interpretacja wyników badania). Pierwsza publikacja to tematyczny numer „Przeglądu Socjologii Jakościowej” (2014, tom 10, nr 2) pod redakcją Jakuba Niedbalskiego poświęcony analizie danych jakościowych wspieranych komputerowo, druga publikacja to praca zbiorowa pod redakcją tego samego autora, która opublikowana została nakładem wydawnictwa Uniwersytetu Łódzkiego w tym samym roku (Niedbalski 2014b). Obydwie prace uznać można za kompendium wiedzy przybliżające polskim badaczom tematykę CAQDAS. Wśród szczegółowych zagadnień opisanych w tych publikacjach polski czytelnik znajdzie między in-

nymi takie jak: organizacja procesu badawczego z wykorzystaniem wsparcia komputerowego (Brosz 2014; Głowacki 2014); zastosowanie CAQDAS w analizie osadzonej w różnych tradycjach teoretycznych, takich jak analiza dyskursu (Kamasa 2014) czy teoria ugruntowana (Niedbalski 2014c) oraz w analizie różnego typu materiałów źródłowych, takich jak orzeczenia Trybunału Konstytucyjnego (Winczorek 2014). W omawianych publikacjach podjęto też próbę przedstawienia zaawansowanych technik analitycznych, takich jak analiza sentymentu (Tomanek 2014), budowa słownika klasyfikacyjnego (Tomanek, Bryda 2014) czy technik *text-miningu* (Bryda, Tomanek 2014). W omawianych publikacjach przedstawiono również praktyczne aspekty komputerowego wsparcia analizy danych w oparciu o wiedzę i doświadczenie związane z zastosowaniem CAQDAS w konkretnych projektach badawczych (Brzeziński 2014; Burski, Brzeziński 2014; Ślęzak, Niedbalski 2014). Polska literatura dotycząca CAQDAS osadzona jest w znacznym stopniu w tradycji teorii ugruntowanej, znacznie mniej jest prac dotyczących badań zakorzenionych w tradycyjnej analizie zawartości i analizie treści przekazów medialnych (prasa, telewizja). Prace dotyczące CAQDAS opierają się w dużej mierze na opracowaniach anglojęzycznych, praktycznym doświadczeniu i autorefleksji badaczy.

Prezentowana w niniejszym artykule metaanaliza przebiegu dwóch modułów badawczych w złożonym projekcie realizowanym przez liczny zespół wpisuje się w ten nurt literatury poświęconej CAQDAS, który ma na celu przedstawienie korzyści i ograniczeń zastosowania oprogramowania w praktyce badawczej. Analiza taka przynosi kon-

kretnie efekty w postaci wniosków pozwalających przyszłym badaczom zoptymalizować zastosowanie wsparcia komputerowego w podobnych projektach, uniknąć „pułapek” oraz wykorzystać szerokie możliwości, jakie niesie ze sobą CAQDAS. Nasza metaanaliza pozwala ponadto uchwycić jedną ze specyficznych cech zastosowania oprogramowania w projekcie, która polega na uczeniu się w praktyce (ang. *learning in practice*). Kolejne doświadczenia badaczy w pracy z wykorzystaniem CAQDAS przynoszą bowiem wiedzę pozwalającą na reagowanie na bieżąco na efekty związane z komputerowym wsparciem analizy danych i wprowadzenie zmian w przebiegu badania.

Nasza analiza opiera się na autorefleksyjnym spojrzeniu na przenikanie się CAQDAS i tradycyjnego podejścia do badań. Celem takiej metaanalizy jest wskazanie korzyści i ograniczeń komputerowego wsparcia analizy na kilku etapach badawczych: projektowanie badania, przygotowanie bazy obserwacji, kodowanie, analiza danych i interpretacja wyników badania. Analiza przeprowadzona została w oparciu o dwa odrębne moduły projektowe: analizę zawartości i treści dokumentów prawnych (moduł 1) oraz analizę transkrypcji wywiadów pogłębionych oraz zawartości stron internetowych (moduł 2). Zgodnie z logiką analizy w pierwszej części artykułu przedstawiono główne założenia projektu oraz wpływ decyzji o zastosowaniu CAQDAS na percepcję i zakres problemu badawczego. W drugiej części opisano korzyści i ograniczenia związane z wykorzystaniem CAQDAS w analizie treści dokumentów prawnych, a następnie – w analizie transkrypcji wywiadów pogłębionych i zawartości stron internetowych. W końcowej części artyku-

łu przedstawiono wnioski w formie rekomendacji umożliwiających przyszłym badaczom optymalizację zastosowania CAQDAS w podobnym projekcie badawczym.

Założenia projektu badawczego a CAQDAS

Przedstawiona w niniejszym artykule analiza swoistej „interakcji” pomiędzy wykorzystaniem CAQDAS a tradycyjnymi metodami badań (Bryda 2014: 12) opiera się na doświadczeniach zespołu badaczy skupionych wokół projektu „Cyfrowe praktyki i strategii upowszechniania i odbioru dziedzictwa kulturowego w Polsce w latach 2004–2014”. Głównym celem szeroko zakrojonego projektu jest opis procesu przemian w upowszechnianiu i odbiorze dziedzictwa kulturowego w formie cyfrowej w Polsce w ostatniej dekadzie XXI wieku. W okresie tym obserwujemy dynamiczny rozwój digitalizacji obiektów kultury, takich jak artefakty, książki, czasopisma, fotografie, materiały audiowizualne, malarstwo, rzeźba, a także zabytkowe budowle i architektura. Obiekty te w formie cyfrowych kopii płaskich (np. fotografia) lub modeli trójwymiarowych (np. zabytki) udostępniane są coraz częściej w Internecie przez instytucje do tego powołane, takie jak muzea czy galerie, ale również przez tak zwane „archiwa społeczne”, które często są rezultatem oddolnej inicjatywy nieformalnych grup, stowarzyszeń, fundacji czy indywidualnych pasjonatów. Digitalizacja do pewnego stopnia zmienia tradycyjny porządek styczności z obiektami dziedzictwa kulturowego, a także praktyki i rytuały związane z obcowaniem z obiektami kultury. Cyberprzestrzeń w założeniu twórców projektu

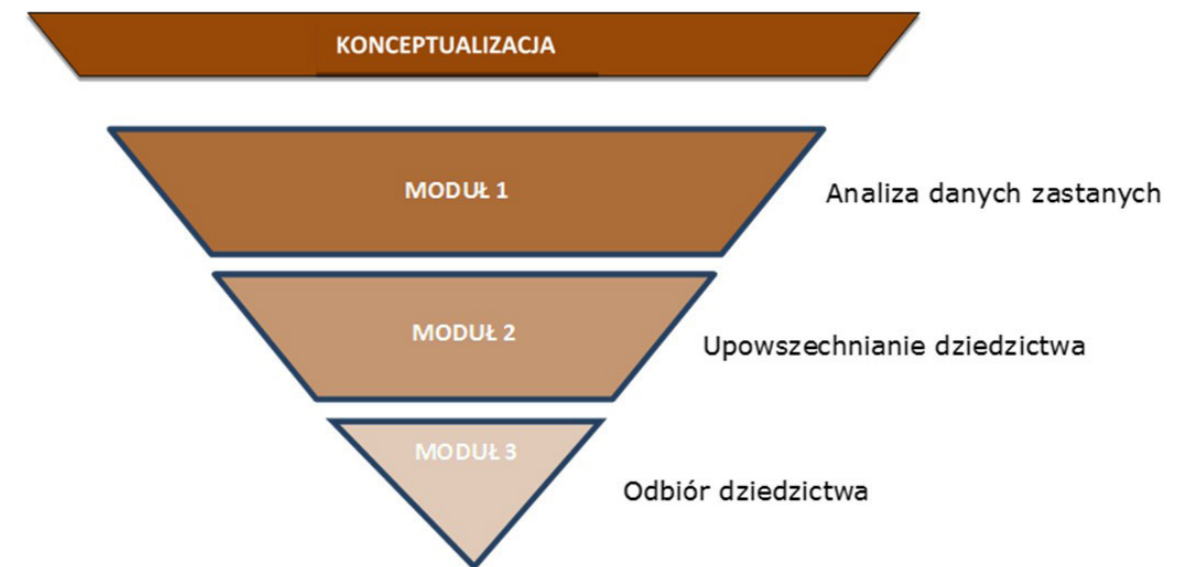
rozumiana jest jako „chmura”, w której wszyscy użytkownicy Internetu mogą znaleźć interesujące ich kopie obiektów dziedzictwa kulturowego. Jednocześnie w konceptualizacyjnej fazie projektowania badania przyjęto założenie o komplementarności praktyk offline i online, które często błędnie rozumiane są jako ucieczka od świata realnego w świat wirtualny. Teoretyczną inspiracją dla podjęcia badania były prace Manuela Castelsa, który podkreśla wagę dostępu do zasobów kulturowych jako kluczowego elementu w procesie konstruowania tożsamości jednostkowej i zbiorowej (Castels 2009). Kulturowe atrybuty krążą w przekazach medialnych i Internecie – w formie zasobów historycznych, religijnych, dokumentów pamięci społecznej – i to właśnie z tych atrybutów konstruowana jest tożsamość jednostki. W omawianym projekcie przyjęto „dyskursywną” definicję dziedzictwa kulturowego, rozumiejąc to pojęcie jako pole dyskusyjne, które powstaje w procesie interakcji pomiędzy uczestnikami kultury i w którym zawiera się rezerwuuar zasobów kulturowych i ich znaczeń. W prezentowanym projekcie położono nacisk na praktyki i strategie digitalizacji, upowszechniania i recepcji zdigitalizowanych obiektów. Digitalizację rozumiemy szeroko jako złożony proces, na który składa się „identyfikacja i selekcja dokumentów; ich przygotowanie, porządkowanie, konserwacja; zebranie podstawowych metadanych; konwersja cyfrowa, kontrola jakości kopii cyfrowych oraz metadanych; dostarczenie użytkownikowi dostępu do dokumentów (w postaci cyfrowej); utrzymanie kopii cyfrowych i metadanych, kopii zapasowych i planowanie na przyszłość” (Dzięglewski, Juza 2015: 8). Szczególną uwagę zwracamy też na rozróżnienie pojęć „udostępnianie” i „upowszechnianie”

dziedzictwa – rozumiejąc pod tym pierwszym pojęciem stworzenie technicznej możliwości dostępu do kopii obiektów kultury (np. poprzez publikację na stronie internetowej), zaś pod tym drugim pojęciem – zakrojoną na szeroką skalę popularyzację i propagowanie korzystania z zasobów kulturowych online (np. poprzez akcje reklamowe). Posługując się pojęciami „praktyki” i „strategie”, wskazujemy na istotną różnicę pomiędzy rutynowymi, codziennymi działaniami aktorów, często nieplanowanymi i podejmowanymi spontanicznie (praktyki), a działaniami refleksyjnymi, długoterminowymi, planowymi, zorientowanymi na realizację określonych celów (strategie). Tak rozumiane pojęcia określają zakres obszaru badawczego, który ogniskuje się wokół głównego pytania badawczego: *Jak przebiegał proces przemian codziennych praktyk oraz strategii udostępniania i odbioru zdigitalizowanych elementów dziedzictwa kulturowego w latach 2004–2014 w Polsce?*

Proces badawczy podzielono na trzy etapy/moduły (ryc. 1), w ramach których postawiono szczegółowe pytania badawcze oraz przyjęto metodologiczne podejście *mixed methods*, stosując zróżnicowane metody (jakościowe i ilościowe), techniki i narzędzia badawcze, w tym również narzędzia CAQDAS. Schemat przebiegu procesu badawczego oparto na zasadzie dedukcyjnej (schemat „lejka”), w ramach której wiedza pozyskana na wcześniejszych etapach badania posłużyła do projektowania kolejnych etapów badawczych.

Głównym celem pierwszego modułu badawczego była eksploracja – wstępne poznanie obszaru badawczego w oparciu o analizę danych zastanych dotyczących regulacji prawnych w zakresie digitalizacji;

Ryc. 1. Schemat przebiegu procesu badawczego



Źródło: oprac. własne.

dostępnych w Internecie repozytoriów zawierających zdigitalizowane zasoby oraz sposobu wykorzystania obiektów cyfrowych w pracach magisterskich

przez przyszłych specjalistów w dziedzinie kultury. Szczegółowe pytania badawcze i zastosowane techniki w tym module przedstawiono w tabeli 1.

Tabela 1. Pytania i techniki badawcze zastosowane podczas pierwszego modułu badawczego

Jednostka analizy	Technika	Pytanie badawcze
Dokumenty prawne (n= 102)	Bad. jakościowe Analiza treści [CAQDAS]	<i>Jakie są główne założenia polityki związanej z digitalizacją i upowszechnianiem dziedzictwa kulturowego w Internecie w prawodawstwie unijnym i polskim? Jak polityki te zmieniają się w badanym okresie?</i>
Portale internetowe (N=455)	Bad. ilościowe Analiza zawartości	<i>Jakie są główne cechy charakteryzujące repozytoria cyfrowe utworzone przez publiczne instytucje kultury i archiwa społeczne? Jak zmieniają się one w badanym okresie?</i>
Prace magisterskie (n=200)	Bad. ilościowe i jakościowe Analiza zawartości i treści	<i>W jaki sposób przyszli specjaliści w dziedzinie kultury wykorzystują zasoby internetowe w swoich pracach? W jaki sposób praktyki te zmieniają się w badanym okresie?</i>

Źródło: oprac. własne.

W pierwszym module podjęto decyzję o zastosowaniu oprogramowania wspomagającego analizę treści aktów prawnych ze względu na duży wolumen danych (102 akty prawne, ok. 600 stron).

Drugi etap badawczy dotyczył przemian w strategiach i praktykach digitalizacji, udostępniania i upowszechniania kopii obiektów kultury w Internecie w instytucjach kultury i na portalach tworzonych przez nieformalne grupy i pojedyncze osoby w ramach archiwistyki społecznej. W ramach tego etapu wybrano w sposób celowy 12 portali internetowych, które odzwierciedlają typy repozytoriów wyłonione podczas badania eksploracyjnego. Wybrane portale poddano pogłębionej, wielowymiarowej analizie zawartości i treści. Ponadto przeprowadzono 20 wywiadów pogłębionych (IDI) z twórcami tych portali. Badania posłużyły do zestawienia w ramach studium przypadku danych pochodzących z doświadczeń i praktyk twórców portali z danymi pochodzącymi z analizy zawartości samych portali. Ponadto wywiady z twórcami portali pozwoliły na wskazanie ogólnych prawidłowości, trendów i zmian w ich

codziennych praktykach związanych z digitalizacją, udostępnianiem i upowszechnianiem zasobów online. W module tym zdecydowano o zastosowaniu narzędzia CAQDAS zarówno podczas wielowymiarowej analizy portali wybranych do studium przypadku, jak też i do jakościowej analizy transkrypcji wywiadów. Szczegółowe pytania badawcze i techniki badania przedstawiono w tabeli 2.

Trzeci – ostatni etap badawczy – dotyczy użytkowników portali zawierających cyfrowe kopie obiektów kultury, których podzielono na „aktywnych uczestników kultury” i „profesjonalistów”. Wśród tych pierwszych przeprowadzono sondaż diagnostyczny, wśród drugich – pogłębione wywiady indywidualne. Celem badań było uchwycenie przemian w doświadczeniach i praktykach związanych z kontaktem z obiektami kultury (w świecie fizycznym i poprzez Internet). Na tym etapie badań zdecydowano się zastosować narzędzie CAQDAS do analizy transkrypcji wywiadów narracyjnych. Szczegółowe pytania i techniki badawcze przedstawiono w tabeli 3.

Tabela 2. Pytania i techniki badawcze zastosowane podczas drugiego modułu badawczego

Jednostka analizy	Techniki	Pytania badawcze
Repozytoria cyfrowe (n=10)	Analiza zawartości, analiza treści [CAQDAS]	Jakie są główne praktyki i strategie upowszechniania zdigitalizowanych obiektów w Internecie w świetle analizy zawartości i treści? Jak zmieniają się one w badanym okresie?
Twórcy portali (instytucje kultury / archiwa społeczne) (n=20)	IDI [CAQDAS]	Jakie są główne praktyki i strategie upowszechniania zdigitalizowanych obiektów w Internecie w doświadczeniu twórców portali? Jak zmieniają się one w badanym okresie?

Studium przypadku

Źródło: oprac. własne.

Tabela 3. Pytania i techniki badawcze zastosowane w trzecim module badawczym

Jednostka analizy	Technika	Pytanie badawcze
Aktywni uczestnicy kultury (n=500)	CAWI PAPI	Jakie są główne praktyki związane z korzystaniem / ponownym wykorzystaniem zasobów pochodzących z internetowych repozytoriów przez aktywnych uczestników kultury? Jak zmieniają się one w badanym okresie?
Profesjonalni uczestnicy kultury (muzealnicy, nauczyciele, animatorzy kultury, projektanci itd.) (n=20)	Wywiad narracyjny [CAQDAS]	Jakie są główne praktyki związane z korzystaniem / ponownym wykorzystaniem zasobów pochodzących z internetowych repozytoriów przez profesjonalnych uczestników kultury? Jak zmieniają się one w badanym okresie?

Źródło: oprac. własne.

Decyzja o zastosowaniu narzędzia CAQDAS w opisanym powyżej projekcie wynikała z kilku przesłanek. Przede wszystkim komputerowe wsparcie analizy danych jakościowych wydawało się nieodzowne ze względu na duży wolumen danych do zakodowania i analizy. W przypadku analizy dokumentów prawnych materiał źródłowy objął około 600 stron, w przypadku transkrypcji wywiadów pogłębionych – około 500 stron. Projekt, ze względu na swoją złożoność i wielowymiarowy charakter, opierał się na pracy zespołowej. W przypadku procesu kodowania obejmował on zespół od 4 do 10 koderów. Na zastosowanie narzędzia CAQDAS wpływ miało również bardzo złożone i „gęste” drzewo kodowe opracowane na etapie konstrukcji narzędzia. W przypadku wywiadów pogłębionych utworzono 35 kodów zgrupowanych w 12 kategoriach analitycznych. Zastosowanie wsparcia komputerowego wiązało się również z przekonaniem

o zwiększeniu – w porównaniu z tradycyjnymi metodami – rzetelności kodowania i, co za tym idzie, większej precyzji w analizie danych. CAQDAS był ponadto wysoko oceniany przez badaczy biorących udział w projekcie pod względem organizacji warsztatu badacza, a w szczególności zwiększenia efektywności zarządzania danymi i ich archiwizowania (w jednym pliku projektowym), znacznie szybsze i sprawniejsze kodowanie oraz operacje na zmiennych, obserwacjach i kodach (filtrowanie, wyszukiwanie, analiza ilościowa).

W projekcie zdecydowano się na zastosowanie oprogramowania QDA Miner 4.0 firmy Provalis Research. Taki wybór podyktowany był kilkoma czynnikami. Przede wszystkim oprogramowanie to powstało w oparciu o teorie związane z analizą zawartości mediów i przekazów medialnych, co wyróżnia je wśród wielu aplikacji powstałych

w oparciu o teorię ugruntowaną, która nie wpisuje się w założenia teoretyczne projektu i jego dedukcyjny charakter. Czynnikiem niezwykle ważnym była też dobrze rozwinięta funkcjonalność pracy zespołowej. QDA Miner umożliwia łatwe dzielenie pomiędzy koderów obserwacji do zakodowania, a następnie szybkie ich łączenie do postaci jednego pliku. Aplikacja umożliwia również sprawdzenie zgodności kodowania. Ponadto QDA Miner ma znacznie rozwinięte funkcjonalności umożliwiające dokonanie szeregu analiz ilościowych, umożliwia łatwy import danych do innych programów (SPSS, EXCEL) oraz szybkie wygenerowanie raportu wraz z wizualizacją danych (np. analiza korespondencji).

Już na etapie projektowania badania – w jego fazie koncepcyjnej i metodologicznej – zaobserwować można istotny wpływ decyzji o zastosowaniu CAQDAS na percepcję obszaru badań, jego zakres i specyfikę. W przypadku analizy dokumentów prawnych wsparcie komputerowe zwiększa znacznie możliwości objęcia badaniem dużej liczby obserwacji (102 akty prawne), przez co umożliwia wnikliwą i bardzo złożoną analizę procesu przemian polityk unijnych i krajowych w zakresie digitalizacji. Objęcie badaniem dużej liczby obserwacji może jednak nieść ze sobą niebezpieczeństwo zbyt skategoryzowanego i „zamkniętego” kodowania danych. Na etapie interpretacji tak zakodowanych danych nie ma zbyt wiele miejsca na intuicję badacza. Alternatywą dla zastosowania CAQDAS byłaby tradycyjna analiza treści, w której znacznemu ograniczeniu uległaby liczba obserwacji (kilkanaście najważniejszych aktów prawnych). W rezultacie problem badawczy zostaje w znacznie mniejszym stopniu poddany ujętej w sztywne kategorie analizie, którą zastępuje analiza mniej ustrukturyzowana i refleksyjna, ale zakres i zróżnicowanie

szczegółowych aspektów polityk dotyczących digitalizacji jest już znacznie ograniczony. A zatem szczegółowość, precyzja i możliwość wielowymiarowego opisu zjawiska i procesu, będące domeną CAQDAS, to elementy, które zmieniają percepcję problemu badawczego przez badaczy. Liczniejsza próba oraz wystandardyzowane procedury kodowania wiążą się też z ryzykiem zbyt ustrukturyzowanej, uporządkowanej na wzór analizy danych ilościowych analizy danych jakościowych. W efekcie zastosowania „gęstego”, szczegółowego drzewa kodowego badacz ryzykuje otrzymanie trudnego do odpowiedniej hierarchizacji strumienia wyników odnoszących się do wąskich obszarów/aspektów badanego fenomenu. Tak ustrukturyzowane wyniki mogą prowadzić do „rozproszonej” i „płaskiej” interpretacji wyników badania.

CAQDAS w analizie dokumentów prawnych

Przygotowanie bazy obserwacji

Podczas badania dokumentów prawnych, którego celem był opis przemian polityk unijnych i krajowych w zakresie digitalizacji dziedzictwa kulturowego, zastosowano celowy dobór próby. Baza aktów prawnych została sporządzona na podstawie informacji zawartych na stronie internetowej Ministerstwa Kultury i Dziedzictwa Narodowego (MKiDN) oraz bazy EurLex. Przy doborze dokumentów przyjęto następujące kryteria: data publikacji (2004–2014) i dziedzina/obszar tematyczny dokumentu (określony na podstawie słów kluczowych)². W wyniku do-

² Słowa kluczowe: dziedzictwo kulturowe, digitalizacja, dziedzictwo narodowe, upowszechnianie kultury, cyfrowe dziedzictwo.

boru próby wskazano 102 dokumenty prawne, które pobrano w formie plików PDF z internetowych baz danych (EurLex i strony MKiDN). Pliki następnie skategoryzowano, nadając im nazwy według prostego kodu umożliwiającego szybkie ich wyszukanie. Kod nazwy pliku zawierał 2 zmienne: *rok publikacji_poziom (UE/PL).pdf*. Wszystkie pliki importowano do pliku projektowego w programie QDA Miner. Już na tym etapie pojawiły się pierwsze trudności związane z importowaniem plików PDF. Pliki zawierające tabele (np. zestawienia instytucji odpowiedzialnych za różne programy digitalizacji w Polsce) nie zostały prawidłowo importowane, w związku z czym niezbędne okazało się odtworzenie tabel w innym pliku tekstowym (.doc) i ponowne importowanie obserwacji. Znacznym ułatwieniem w pracy związanej z poruszaniem się w materiale źródłowym jest przede

wszystkim możliwość umieszczenia wszystkich obserwacji w jednym pliku projektowym i możliwość skategoryzowania ich według przyjętych zmiennych.

W opisywanym badaniu wszystkie dokumenty zostały opisane za pomocą 8 zmiennych: (1) nazwa pliku, (2) kolejny numer, (3) tytuł/nazwa aktu prawnego, (4) poziom legislacyjny (UE, PL), (5) typ dokumentu prawnego, (6) rok publikacji, (7), dziedzina prawa, (8) zagadnienia poruszane w dokumencie (słowa kluczowe) (ryc. 2)

Kodowanie

Drzewo kodowe składało się z 12 kodów w pięciu kategoriach analitycznych: (1) dziedzictwo – definiowanie pojęcia i jego zakres, (2) priorytety – cele

Ryc. 2. Zmienne opisujące obserwacje w programie QDA Miner

OBSERWACJE: (Filtrowane: 14/102)	
ZMIENNE	
FILE	33_2011_PL
DOKUMENT	[DOKUMENT]
NUMER	33
NAZWA	ZARZADZENIE NR. 46 MINISTRA KULTURY I DZIEDZICTWA NARODOWEGO
POZIOM	Polska
TYP	Rozporządzenie i Zarządzenie_PL
ROK	2012
DZIEDZINA	Dziedzictwo narodowe/kulturowe
ZAGADNIENI	DZIEDZICTWO NARODOWE, KULTURA, ZABYTKI, DOTACJE

Źródło: oprac. własne.

ogólne i szczegółowe, jakie wyznacza prawodawca w zakresie digitalizacji, udostępniania i upowszechniania dziedzictwa online, (3) struktury – instytucje powołane do nadzoru działalności digitalizacyjnej oraz formuła ich finansowania, (4) zalecenia – rekomendacje dotyczące strategii i planowego działania w zakresie digitalizacji i prawa autorskiego i (5) odbiorcy – wskazania dotyczące różnych kategorii odbiorców dziedzictwa w formie cyfrowej oraz sposobów wykorzystania zasobów. Wszystkie kody dokładnie opisano – koderzy dzięki funkcjonalności oprogramowania mogli przypomnieć sobie instrukcje dotyczące danego kodu, najeżdżając myszką na jego pole (ryc. 3).

Ponadto przez cały czas pracy koderzy mieli przed sobą wydruk książki kodowej oraz szczegółowe instrukcje zawierające opis postępowania na wypadek potencjalnych trudności w kodowaniu. Koderzy uczestniczyli też w szkoleniu z zakresu obsługi interfejsu programu QDA Miner. Główną techniką podczas kodowania dokumentów było kodowanie *incident-by-incident*³. Zespół koderów był monitorowany przez koordynatora, który sprawdził 10% zakodowanego materiału. W rezultacie sprawdzenia zbieżności kodowania oraz przeglądu pracy koderów okazało się, iż niektórzy koderzy odmiennie rozumieją instrukcje kodowania, a w szczególności wielkość jednostki tekstu, która winna być zakodowana (wyraz, zdanie, fraza, akapit). Koderów cechowały też różne

³ Reguła *incident-by-incident* odnosi się do wielkości jednostki kodowania i oznacza, iż kodowano każdy fragment tekstu (niezależnie od jego wielkości), który odnosił się do danej kategorii analitycznej uwzględnionej w „drzewie kodowym”, traktując ten fragment jako jednostkę kodowania. W praktyce badawczej możliwe jest również kodowanie według reguły *word for word* (jednostką kodowania jest słowo) lub *line-by-line* (jednostką kodowania jest wers, zdanie lub fraza – w zależności od instrukcji kodowania).

kompetencje językowe. Teksty aktów prawnych napisane są bardzo sformalizowanym językiem z zastosowaniem prawniczego żargonu i specjalistycznych zwrotów. Niektórzy koderzy pomijali istotne fragmenty tekstu, jeśli nie odnosił się on wprost do informacji, które wskazano w instrukcji. W przypadku niektórych koderów znaczne partie tekstu należało zakodować jeszcze raz. Problem zgodności kodowania, a co za tym idzie – rzetelności badania, okazał się jednym z istotnych ograniczeń pracy z QDA Minerem w przypadku dużego (10 osób) zespołu koderskiego. Niemniej jednak samą funkcjonalność podziału obserwacji pomiędzy koderów, łączenia ich pracy w jeden plik projektowy i możliwość analizy zgodności kodowania uznać należy za bardzo przydatne w przypadku pracy więcej niż jednego kodera.

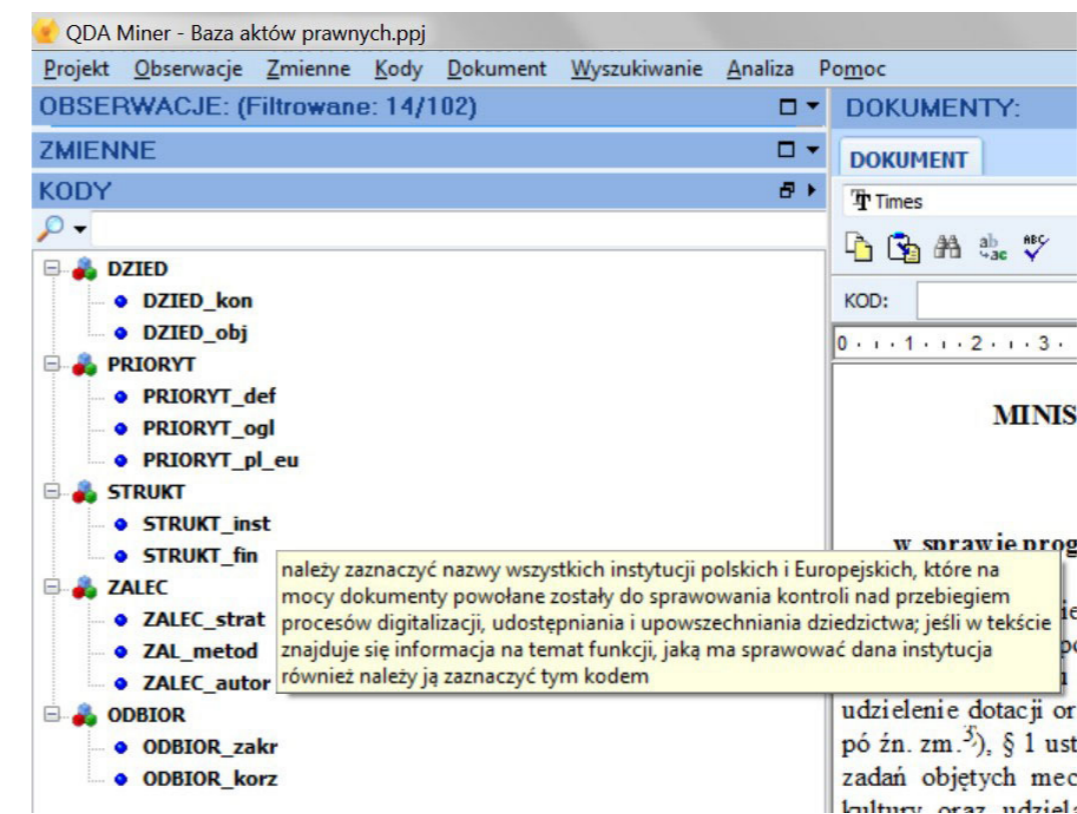
Analiza danych

Na etapie analizy wyraźnie rozróżnić należy doświadczenia związane z procedurami analizy ilościowej i jakościowej. W tym pierwszym przypadku zakres analizy na zmiennych i kodach, które oferuje QDA Miner, jest imponujący. Proste procedury analityczne na zmiennych (częstości, tabele krzyżowe) pozwoliły na wstępne rozeznanie kierunku przemian polityk związanych z digitalizacją i odpowiedź na takie pytania, jak: *Jakie zagadnienia najczęściej podejmowano w aktach prawnych w poszczególnych latach? Jaka była intensywność (częstość) publikacji aktów prawnych w poszczególnych latach? Jakie domeny prawa dominowały w prawodawstwie unijnym, a jakie w krajowym w poszczególnych latach?* Za równie użyteczne uznać należy wbudowane w aplikację proste procedury ilościowej analizy na kodach oraz tabele krzyżowe zmienne/kody. Procedury te pozwoliły

na odpowiedź na takie pytania szczegółowe, jak: *Jakie kategorie analityczne najczęściej pojawiają się w aktach prawnych? Czy / w jaki sposób zmieniał się nacisk na poszczególne kategorie (kody) w kolejnych latach? W jaki sposób zmiany te zachodziły w prawodawstwie polskim, a w jaki w europejskim?* Wszystkie procedury analizy ilościowej w aplikacji QDA Miner są intuicyjne i pozwalają na bardzo szybkie uzyskanie wyników. Przeprowadzanie takich analiz w sposób tradycyjny byłoby niezwykle czasochłonne. Dodatkową korzyścią wynikającą z zastosowania QDA Minera w analizie ilościowej, która w naszym przypadku miała tylko charakter eksploracyjny, jest możliwość automatycznej wizualizacji danych (w postaci tabeli, wykresu, wizualizacji 3D) oraz eksportowanie jej w gotowej postaci do raportu.

Znacznie więcej trudności nastęcza jakościowa analiza danych, którą w naszym przypadku przeprowadzono w szablonach Excel dla każdego kodu wyeksportowanego z programu QDA Miner. Trudności te związane były przede wszystkim ze zróżnicowanym formatem jednostek analitycznych (słowo, fraza, dłuższa wypowiedź) oraz w przypadku wielu zakodowanych fragmentów z brakiem kontekstu, w którym dana wypowiedź została osadzona. W wyniku tych trudności, aby odpowiednio zrozumieć zakodowany fragment i jego znaczenie, badacz musiał sięgać do tekstu źródłowego. Dodatkowym – obok dekontekstualizacji – ograniczeniem w analizie jakościowej jest brak możliwości prześledzenia związków pomiędzy różnymi kategoriami analitycznymi i brak osadzenia poszcze-

Ryc. 3. Drzewo kodowe wraz z instrukcją dla kodera



Źródło: oprac. własne.

gólnych jednostek analitycznych w szerszym spektrum znaczeniowym. Specyficzny sformalizowany język aktów prawnych pozwala jedynie na analizę konkretnych wytycznych, strategicznych rozwiązań oraz planów, nie pozwala na uchwycenie w pełni intencji prawodawców. W rezultacie zdanie się jedynie na jakościową analizę treści w oparciu o zakodowany w programie materiał tekstowy nie pozostawia zbyt wiele miejsca na analizę refleksyjną, intuicyjną, opartą na wiedzy kontekstowej i poza-kontekstowej badacza.

Interpretacja wyników badania

Na etapie interpretacji wyników analizy danych badacz napotyka – przewidywane już w fazie projektowania badań – trudności związane ze specyfiką komputerowego wsparcia analizy. W przypadku interpretacji wyników analizy ilościowej badacz musi zadać sobie kilka istotnych pytań. Przykładowo, w badaniu dokumentów prawnych badacz staje przed pytaniem: *Co to znaczy, że kody odnoszące się do definiowania pojęcia „digitalizacja” i określenia jego zakresu (kategoria analityczna „dziedzictwo”) występują w analizowanych aktach prawnych najczęściej, a kody dotyczące odbiorców cyfrowych kopii i sposobów ich wykorzystania (kategoria „odbiorcy”) są najmniej liczne?*

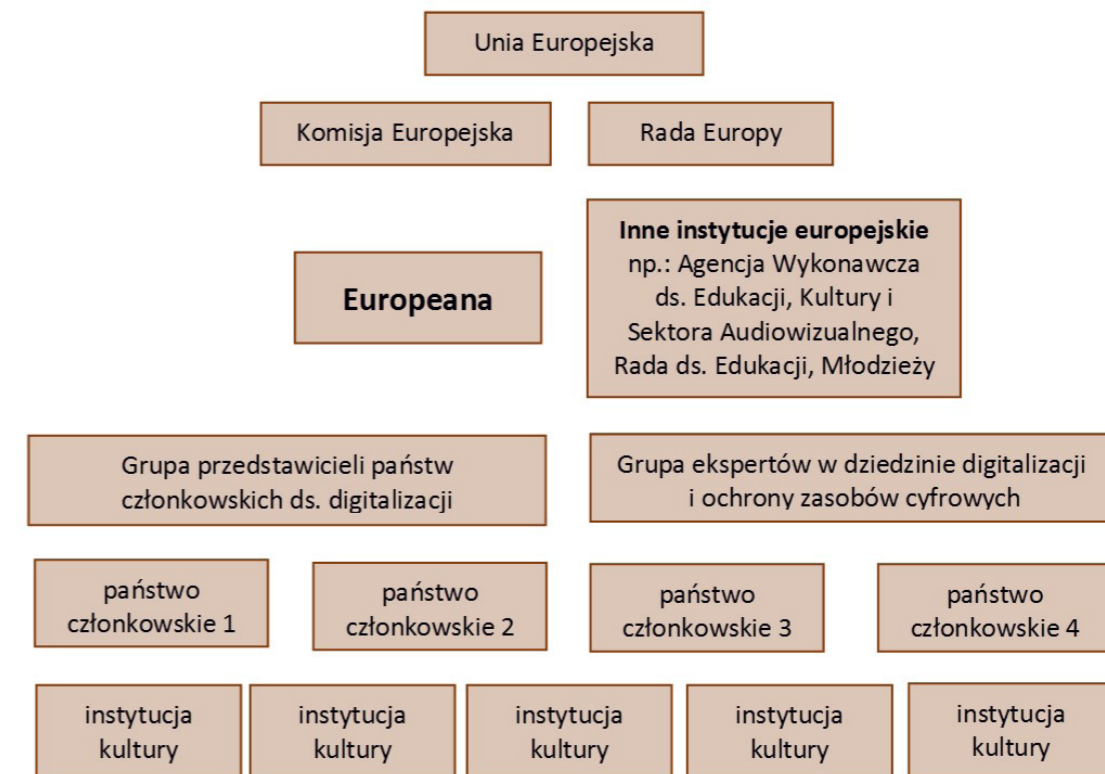
Interpretując takie wyniki analizy ilościowej, można podejrzewać, iż prawodawcy koncentrują się na precyzyjnej definicji tego, co winno być uznawane za dziedzictwo kulturowe, ignorują natomiast potencjalnych beneficjentów korzystających z cyfrowych kopii obiektów kultury dostępnych online. Taka interpretacja jest jednak nieuprawniona, ponieważ sama częstość występowania pojęcia „dziedzictwo

kulturowe” w różnych przypadkach i kontekstach nie wskazuje na większą wagę tego obszaru dla decydentów, tym bardziej, iż większość rozporządzeń dotyczących prawa polskiego opublikowana została przez Ministerstwo Kultury i Dziedzictwa Narodowego (podkr. własne).

Podobne problemy związane z uogólnianiem i interpretacją wyników badania pojawiają się w przypadku analizy danych jakościowych. W naszym przypadku analiza kategorii analitycznej „struktury” miała na celu odtworzenie schematu „struktury instytucjonalnej” procesu digitalizacji w Unii Europejskiej i Polsce na podstawie analizy instytucji zaangażowanych w ten proces, ich funkcji i wzajemnych powiązań. W wyniku tej analizy badacze opisali tę strukturę i przedstawili ją w postaci graficznej (ryc. 4).

Przeprowadzone w drugim module projektu indywidualne wywiady pogłębione z twórcami portali udostępniających kopie obiektów kultury w Internecie wskazują, iż tak odtworzona struktura nie funkcjonuje w ich codziennej praktyce związanej z digitalizacją i udostępnianiem zasobów online. Interpretując wyniki z tych dwóch badań, można stwierdzić, że wytyczne, rekomendacje, rozporządzenia sformułowane przez twórców polityk w zakresie digitalizacji nie są realizowane w praktyce. Taka interpretacja jednak (choć prawdopodobna) nie jest uzasadniona w świetle przebiegu analizy wspomaganą przez QDA Minera. Z założenia bowiem – ze względu na celowy dobór próby i brak pełnej dokumentacji prawnej – wyłonienie instytucji, określenie ich funkcji i wzajemnych zależności jest fragmentaryczne, często nie uwzględnia kon-

Ryc. 4. Struktura instytucjonalna procesu digitalizacji na szczeblu europejskim



Źródło: oprac. własne.

tekstu i nie pozwala na tego typu wnioskowanie. A zatem badacz musi wyraźnie opisać ograniczenia związane z tego typu analizą i wnioskowaniem.

CAQDAS w analizie transkrypcji wywiadów i zawartości stron internetowych

Przygotowanie obserwacji

Analiza zastosowania CAQDAS w drugim module badawczym wskazuje wyraźnie na fakt uczenia się przez badaczy efektywnego korzystania ze wsparcia komputerowego w trakcie prowadzenia badań (*learning in practice*). Zgodnie z projektem przebiegu badań – drugi moduł badawczy miał na celu pogłębiony opis praktyk i strategii digitalizacji,

udostępniania i upowszechniania zasobów online z perspektywy architektury repozytoriów cyfrowych oraz doświadczeń ich twórców. Merytoryczny obszar badania dodatkowo umiejscowiono na continuum czasowym – w latach 2004–2014. W tym celu z 455 repozytoriów wybrano 12 stron internetowych utworzonych przez publiczne instytucje kultury oraz przedstawicieli archiwistyki społecznej. Dobór ten opierał się na kilku kryteriach, które wynikały z analizy danych w pierwszym module, a w szczególności – roboczej typologii repozytoriów uwzględniającej ich podstawowe charakterystyki. Strony poddano analizie zawartości i treści w oparciu o wcześniej przygotowany klucz kategoryzacyjny. Dane w postaci uzupełnionych formatów tekstowych importowano do programu QDA Miner. Wywiady pogłębione z twórcami portali (20)

przeprowadzono przy wykorzystaniu częściowo ustrukturyzowanego scenariusza, nagrano, sporządzono transkrypcje i następnie importowano do osobnego pliku programu QDA Miner. Ponieważ wszystkie dane zapisane były w formacie DOC, badacze nie napotkali żadnych trudności z importowaniem ich do plików projektowych. Strony internetowe opisano za pomocą 10 zmiennych: (1) nazwa pliku, (2) kolejny numer, (3) nazwa portalu, (4) nazwa repozytorium, (5) nazwa organu prowadzącego portal, (6) lokalizacja podmiotu, (7) typ portalu (typologia z pierwszego modułu), (8) typ podmiotu, (9) rok powstania portalu, (10) rok powstania organu prowadzącego. Transkrypcje wywiadów pogłębionych opisano za pomocą 7 zmiennych: (1) nazwa pliku, (2) kod respondenta, (3) kategoria respondenta, (4) nazwa portalu, (5) typ podmiotu, (6) typ portalu, (7) rok utworzenia portalu.

Podobnie jak w pierwszym module badawczym zastosowanie CAQDAS jest niezastąpione w sprawnym zarządzaniu materiałem źródłowym, sortowaniu, kategoryzowaniu, wyszukiwaniu i filtrowaniu obserwacji. W przypadku przygotowania obserwacji do analizy należy jednak przemyśleć liczbę i użyteczność zastosowanych w opisie obserwacji zmiennych. W naszym przypadku zbyt duża liczba zmiennych opisujących obserwacje wprowadziła niepotrzebny „szum” podczas operacji na danych. Niektóre z tych zmiennych zostały w dalszej części badania usunięte jako zbędne.

Kodowanie

Przyjętą w badaniu techniką kodowania było kodowanie *incident-by-incident*. Tym razem zespół kodowy

ograniczono do pięciu osób z większym doświadczeniem (dwóch badaczy, dwóch studentów i koordynator). Koordynator sprawdził 10% zakodowanych fragmentów tekstu oraz zgodność kodowania. Redukcja liczebności zespołu z większym doświadczeniem oraz negocjacje pomiędzy koderami dotyczące niejednoznacznych przypadków kodowania pozwoliły uniknąć błędów z pierwszego modułu. Dzięki temu osiągnięto wysoką zgodność kodowania, a co za tym idzie – wysoki stopień rzetelności. W analizie wykorzystano też funkcjonalność tworzenia w tekście hiperlinków ze znacznikiem czasu. Dzięki tej funkcjonalności możliwe było prześledzenie interesujących badaczy procesów w porządku chronologicznym oraz wizualizacja tych procesów w postaci osi czasu. Nawet bardzo proste zestawienie danych dotyczących daty powstania analizowanych portali na osi czasu i ich wizualizacja pozwoliły na łatwą orientację w doświadczeniach twórców poszczególnych portali – wyróżnionych kolorem ze względu na typ podmiotu prowadzącego (instytucja „odgórna” – niebieski i archiwa społeczne – zielony) (ryc. 5).

W przypadku kodowania transkrypcji wywiadów pogłębionych zastosowano bardzo szczegółowe („gęste”) drzewo kodowe zawierające 35 kodów zgrupowanych w 12 kategorii analitycznych, które odnoszą się do różnych wymiarów doświadczeń, wiedzy i refleksji twórców portali udostępniających cyfrowe zasoby. Badacze położyli szczególny nacisk na precyzję instrukcji dla koderów. Z jednej strony tak rozbudowane drzewo kodowe stanowiło wyzwanie dla koderów, z drugiej jednak strony – precyzyjne instrukcje i zasada wzajemnego ustalania (negocjowania) znaczeń przez koderów przyczyniły się do sprawnego zakodowania transkrypcji, która objęła ponad

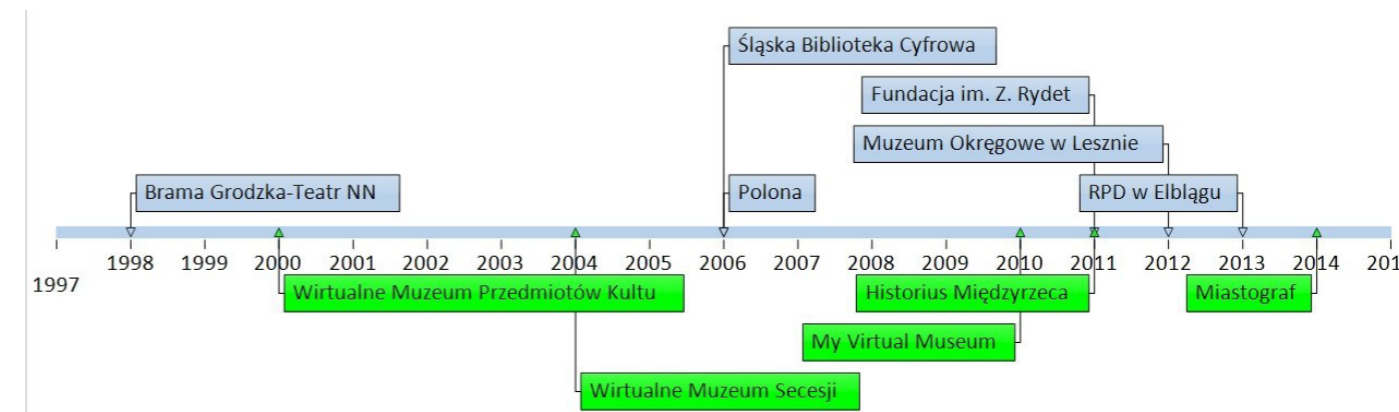
500 stron. Na efektywną pracę koderów wpłynęły ponadto dwa inne czynniki: zgodny ze scenariuszem schemat przebiegu wywiadu (kolejność kodów), codzienny język mówiony oraz literalny sens informacji przypisanych do danego kodu. Przykładowo kod „ODB_who” odnosił się do tych fragmentów tekstu, w których respondent opowiada o adresatach portalu, którego jest współtwórcą, ich profilu oraz sposobach ich identyfikacji przez respondenta (np. licznik wejść na stronę itp.). W tym przypadku zakodowanie odpowiedniego fragmentu nie wymaga wysokich kompetencji językowych, tak jak w przypadku analizy dokumentów prawnych w pierwszym module badawczym.

Analiza danych

Analiza danych w drugim module badawczym przeprowadzona została w dwóch wymiarach: „horyzontalnym” i „wertykalnym” (ryc. 6).

Analiza „horyzontalna” miała na celu uchwycenie wspólnych prawidłowości w praktykach twórców portali i ich architekturze oraz zmian w tych dwóch obszarach w badanym okresie. Analiza „wertykalna” miała na celu szczegółową, pogłębioną analizę danych pochodzących z analizy zawartości i treści badanego portalu oraz doświadczeń i praktyk jego twórców w ramach studium przypadku.

Ryc. 5. Data utworzenia portali internetowych udostępniających cyfrowe kopie obiektów kultury



Źródło: oprac. własne.

Ryc. 6. Dwa wymiary analizy (moduł drugi)

	STUDIUM PRZYPADKU 1	STUDIUM PRZYPADKU 2	STUDIUM PRZYPADKU 3	STUDIUM PRZYPADKU n	
PORTALE	PORTAL 1	PORTAL 2	PORTAL 3	PORTAL n	Horyzontalne
WYWIADY	RES 1 RES 2	RES 3 RES 4	RES 5 RES 6	RES i RES i+1	Horyzontalne
	Wertykalne				

Źródło: oprac. własne.

Zastosowanie QDA Miner podczas analizy „horyzontalnej” okazało się bardzo przydatne, znacznie wpłynęło na efektywność pracy badaczy, głównie ze względu na funkcjonalności związane z ilościową analizą danych (opisane powyżej), jak i z jakościową analizą danych. Pewnym ograniczeniem na etapie analizy jakościowej okazał się duży wolumen danych odnoszących się do bardzo szczegółowych, wąsko zakrojonych obszarów tematycznych. Taki poziom szczegółowości wynikał jednak z konstrukcji narzędzia badawczego, którego celem było zebranie danych do wielowymiarowego opisu złożonego procesu. Podczas analizy „wertykalnej” podstawowym utrudnieniem okazał się brak technicznej możliwości jednoczesnej pracy na dwóch plikach projektowych (strony internetowe i wywiady). Analizę taką można byłoby przeprowadzić przy wsparciu programu QDA Miner w przypadku utworzenia jednego pliku projektowego zawierającego dane z analizy zawartości portali internetowych i wywiadów pogłębionych. Jednak sytuacja uniemożliwiająca taką analizę (dwa osobne pliki projektowe) zmusiła badaczy do tradycyjnej analizy zebranego i zakodowanego materiału, która polegała na zestawieniu korespondujących ze sobą kategorii analitycznych w badaniach stron internetowych i ich twórców. Zestawienie takie ukazuje obszary, w których doświadczenia twórców, ich wiedza i praktyki związane z digitalizacją i upowszechnianiem dziedzictwa kulturowego w sieci znajdują swoje odbicie w architekturze tworzonych przez nich portali oraz takie obszary, w których te dwie perspektywy oddalają się od siebie. Rezygnacja z zastosowania CAQDAS w analizie „wertykalnej”, której celem były studia przypadku, otworzyła przestrzeń dla nieustrukturyzowanego,

intuicyjnego oglądu badacza i refleksyjnego opisu przenikania się dwóch perspektyw obserwacji (zawartość portalu i doświadczenia jego twórcy). Ten „moment” w projekcie szczególnie dobitnie ilustruje specyficzne przenikanie się, swoistą interakcję pomiędzy CAQDAS a tradycyjnymi metodami i technikami badań. Interakcję taką należy rozumieć w sposób symboliczny, a jednocześnie bardzo trafnie opisujący faktyczne oddziaływanie dwóch sposobów pracy badacza-analityka. Interakcję w socjologii rozumiemy jako wzajemne oddziaływanie na siebie dwóch lub więcej aktorów, którzy „orientują się” na bieżące zachowania swoich interlokutorów. Podobnie w praktyce badawczej, w której wykorzystujemy narzędzia CAQDAS, redefiniują one podejście badacza do problemu, zakres i formy analizy danych utrwalone poprzez tradycję, a jednocześnie zachowania badacza wynikające ze zinternalizowanych nawyków wpływają na sposób, w jaki CAQDAS zostaje włączone do projektu badawczego.

Interpretacja

Podczas „horyzontalnej” analizy transkrypcji wywiadów pogłębionych badacze musieli zmierzyć się z bardzo dużą liczbą kodów odnoszących się do szczegółowych aspektów związanych z praktykami digitalizacyjnymi. Aspekty te to między innymi: sposób rozumienia pojęcia „dziedzictwo”; czynniki, które wpłynęły na decyzję o podjęciu działalności digitalizacyjnej; przebieg procesu digitalizacji w danej instytucji; komunikowanie się z użytkownikami portalu i tym podobne. Analiza kodów przynosi wiedzę na temat tych poszczególnych aspektów interesującego nas fenomenu w jego dynamicznym,

procesualnym wymiarze. Z punktu widzenia interpretacji wyników badań trudno jest tylko na podstawie analizy w programie QDA Miner ustanowić wagę poszczególnych aspektów (kodów, kategorii analitycznych) oraz odnieść je do głównych pytań/problemów badawczych. Przykładowo, jeśli pytanie badawcze brzmi: *Jakie są strategie i praktyki digitalizacji dziedzictwa kulturowego?*, analizą obejmujemy dane zawarte w 9 szczegółowych kodach, między innymi:

DIGI_def: *jak respondent rozumie pojęcie „dziedzictwo kulturowe”?*

DIGI_fnc: *jaki jest cel digitalizacji w opinii respondenta?*

DIGI_efe: *jakie są koszty i korzyści digitalizacji dla obiektu dziedzictwa kulturowego?*

DIGI_zak: *jaka część zasobów gromadzonych w instytucji, którą reprezentuje respondent, została zdigitalizowana?*

DIGI_org: *jak zorganizowany jest proces digitalizacji (etapy, infrastruktura, kadra)?*

Podczas gdy szczegółowe „szatkowanie” problemu badawczego sprawia, iż trudno interpretować wyniki badań w ramach analizy „horyzontalnej”, procedura ta jest bardzo przydatna w analizie „wertykalnej”. W tym wypadku analiza służy pogłębionemu, wieloaspektowemu opisowi studium przypadku, a interpretacja nie jest głównym celem badania.

Trzeci moduł projektu nie został jeszcze zakończony. Na tym etapie planuje się zastosowanie oprogramowania QDA Miner do analizy transkrypcji wywiadów pogłębionych z „profesjonalnymi uczestnikami kultury” (animatorami, muzealnikami, nauczycielami). Wywiady mają na celu poznanie, z per-

spektywy tej wąskiej grupy odbiorców, zmian, jakie zaszły w ich pracy zawodowej na skutek nowych możliwości związanych z digitalizacją i wykorzystaniem cyfrowych kopii obiektów kultury. W przypadku analizy tych wywiadów planuje się zastosowanie funkcjonalności tworzenia hiperlinków odnoszących się do znaczników czasu. Technika ta pozwoli uporządkować chronologicznie narrację na temat życia zawodowego oraz odtworzyć prawidłowości i trendy w korzystaniu z kopii obiektów cyfrowych w pracy zawodowej respondentów.

Wnioski

Powyższa metaanaliza przebiegu procesu badawczego stanowi jednocześnie jego metodologiczną ewaluację. Wskazuje również pewien proces, któremu podlegają sami badacze podczas realizacji projektu. Proces ten określiliśmy mianem „uczenia się w praktyce” (ang. *learning in practice*). Polega on na autorefleksyjnej postawie badacza, który na podstawie doświadczeń wynikających z zastosowanych procedur CAQDAS i ich efektów wyciąga wnioski, stara się zminimalizować ograniczenia i zwiększyć możliwości analitycznego wykorzystania komputerowego wspomaganie badań. Główną obserwacją płynącą z powyższej analizy jest potwierdzenie tezy sformułowanej wcześniej przez innych badaczy (Bryda 2014) o swoistej interakcji zachodzącej pomiędzy CAQDAS a tradycyjnymi metodami i technikami badawczymi. Zastosowanie CAQDAS z jednej strony powoduje zmianę percepcji problemu badawczego, sposobu gromadzenia i porządkowania obserwacji, kodowania, analizy i interpretacji wyników badania. Z drugiej strony jednak – tradycyjne postępowanie badawcze wyznacza ramy, w

jakich CAQDAS może zostać wykorzystany w projekcie badawczym.

Szczegółowe obserwacje dotyczące korzyści i ograniczeń pracy z QDA Minerem w opisanym powyżej projekcie prowadzą do kilku rekomendacji, które mogą zapobiec „pułapkom” związanym z zastosowaniem tego oprogramowania. Badacz przystępujący do projektowania badania winien przede wszystkim odpowiedzieć na kilka istotnych pytań dotyczących zastosowania narzędzi CAQDAS. Przede wszystkim: *Co jest celem zastosowania CAQDAS w projekcie? Czy planowany projekt jest odpowiedni do jego zastosowania? Jakie oprogramowanie jest najlepsze do planowanego projektu? W jaki sposób zastosowanie CAQDAS zmieni istotę i zakres pytań badawczych?* Najczęściej pojawiającymi się argumentami za zastosowaniem CAQDAS jest efektywność, szybkość przeprowadzenia analizy na dużej liczbie obserwacji, łatwe zarządzanie danymi i ich archiwizacja w jednym pliku projektowym. Nie wszystkie projekty jednak będą odpowiednie do zastosowania wsparcia komputerowego. Wśród coraz liczniej dostępnych na rynku w Polsce aplikacji największą popularnością cieszą się takie aplikacje, jak Nvivo, MaxQDA, Atlas Ti, QDA Miner. Większość aplikacji została jednak stworzona w oparciu o teorię ugruntowaną, tylko niektóre z nich wyrastają z teorii analizy zawartości i treści tekstu stosowanej w badaniach prasoznawczych i medioznawczych. Zastosowanie odpowiedniego oprogramowania, które mieści się w teoretycznych ramach przyjętych przez badacza, z pewnością pozwoli uniknąć wielu trudności analitycznych (np. dedukcyjny vs. indukcyjny proces analizy danych). Decyzja o zastosowaniu

CAQDAS w projekcie badawczym powinna być również poprzedzona refleksją na temat zmian, jakie komputerowe wspomaganie badania „wymusza” w zakresie obszaru badawczego, i percepcji problemu badawczego (np. szerszy obszar badania, ale bardziej „płaski” charakter pozyskanych danych jakościowych).

Kluczowym etapem w pracy z CAQDAS jest kodowanie. W przypadku kodowania zespołowego – tak jak w opisanym powyżej projekcie – należy zadbać o precyzyjną instrukcję kodową oraz przeprowadzenie warsztatów analitycznych, których celem winno być zminimalizowanie zróżnicowania związanego z kompetencjami lingwistycznymi i „wrażliwością językową” koderów. Warsztaty takie winny opierać się na wspólnym, zespołowym kodowaniu określonej liczby obserwacji aż do uzyskania względnej jednomyślności. Kluczowa w tym zakresie jest również rola koordynatora tego procesu, którego zadaniem jest wyjaśnianie niejednoznacznych sytuacji związanych z kodowaniem. Dobrym sposobem, który oferuje większość narzędzi CAQDAS, jest wykorzystanie w takich sytuacjach komentarza (*memo*), który może zawierać różne uwagi i wątpliwości kodera związane z zakodowanym fragmentem tekstu.

Bardzo ważnym momentem podczas pracy z CAQDAS jest interpretacja wyników badania. Jak dowodzi przeprowadzona powyżej analiza – zbyt nie uleganie logice analizy wspomaganą komputerowo może prowadzić do nieuprawnionych uogólnień. Dlatego należy zadbać o przestrzeń na nieustrukturyzowany, refleksyjny i w dużej mierze oparty na intuicji badacza ogląd danych jako-

ściowych. Tradycyjne zapoznanie się z materiałem badawczym pozwala na uniknięcie ograniczeń, a czasem błędów związanych z dekontekstualizacją zakodowanych fragmentów. CAQDAS posiada również funkcjonalności, które wspierają osadzenie zakodowanych fragmentów w odpowiednim kontekście. Są to wspomniane wcześniej komentarze (*memos*) oraz hiperlinki, które pozwalają łączyć różne wątki w ramach jednej obserwacji, wielu obserwacji oraz pomiędzy kategoriami analitycznymi. Właściwa i uprawniona interpretacja wyników badania wymaga od badacza zestawienia wyników pochodzących z analizy wspieranej kompu-

terowo oraz tradycyjnego zapoznania się badacza z materiałem źródłowym.

Z całą pewnością narzędzia CAQDAS są bardzo istotnymi elementami warsztatu badawczego współczesnego socjologa. Nazwa tej rodziny oprogramowania wyraźnie jednak wskazuje na ich ograniczone możliwości. Narzędzia CAQDAS mają na celu komputerowe wspomaganie pracy badacza, nigdy go jednak nie zastąpią. To badacz ostatecznie decyduje o wykorzystaniu takich, a nie innych narzędzi ze swojego warsztatu – z całą świadomością konsekwencji ich zastosowania.

Bibliografia

Becker Howard S., Gordon Andrew C., LeBailly Robert K. (1984) *Fieldwork with the Computer: Criteria for Assessing Systems*. „Qualitative Sociology”, vol. 7, no. 1/2, s. 16–33.

Bieliński Jacek, Iwińska Katarzyna, Rosińska-Kordasiewicz Anna (2007) *Analiza danych jakościowych przy użyciu programów komputerowych*. „ASK”, nr 16, s. 89–114.

Brent Edward E. (1984) *Qualitative Computing: Approaches and Issues*. „Qualitative Sociology”, vol. 7, no. 1–2, s. 36–60.

Brosz Maciej (2012) *Komputerowe wspomaganie badań jakościowych. Zastosowanie pakietu NVivo w analizie materiałów nieustrukturyzowanych*. „Przegląd Socjologii Jakościowej”, t. 8, nr 1, s. 98–125 [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.przegladsocjologiijakosciowej.org>>.

Brosz Maciej (2014) *Badanie jakościowe w warunkach niedoszacowania czasu – czyli o konieczności „podążania na skrótach”*. *Organizacja procesu badawczego pod kątem zastosowania komputerowego wsparcia analizy danych jakościowych*. „Przegląd Socjologii Jakościowej”, t. 10,

nr 2, s. 40–59 [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.przegladsocjologiijakosciowej.org>>.

Bryda Grzegorz (2014) *CAQDAS a badania jakościowe w praktyce*. „Przegląd Socjologii Jakościowej”, t. 10, nr 2, s. 12–37 [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.przegladsocjologiijakosciowej.org>>.

Bryda Grzegorz, Tomanek Krzysztof (2014) *Od CAQDAS do Text Miningu. Nowe techniki w analizie danych jakościowych* [w:] Niedbalski Jakub, red., *Metody i techniki odkrywania wiedzy: narzędzia CAQDAS w procesie analizy danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 191–218.

Brzeziński Kamil (2014) *Wykorzystanie programu komputerowego QDA Miner w analizie jakościowego materiału badawczego na przykładzie pogłębionych wywiadów swobodnych z mieszkańcami łódzkich gated communities* [w:] Niedbalski Jakub, red., *Metody i techniki odkrywania wiedzy: narzędzia CAQDAS w procesie analizy danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 41–76.

Burski Jacek, Brzeziński Kamil (2014) *Próba wykorzystania programu komputerowego QDA Miner do realizacji projektu badawczego „Cztery dyskursy o nowoczesności – modernizm peryferii na przykładzie Łodzi (XIX–XX wiek)”*. „Przegląd Socjologii Jakościowej”, t. 10, nr 2, s. 82–99 [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.przehladsocjologiijakosciowej.org>>.

Castels Manuel (2009) *Siła tożsamości*. Przełożył Sebastian Szymański. Warszawa: Wydawnictwo Naukowe PWN.

Dzięglewski Mariusz, Juza Marta (2015) *Praktyki i strategie upowszechniania i odbioru dziedzictwa kulturowego. Raport metodologiczny*. Kraków: Wydawnictwo Małopolskiego Instytutu Kultury [dostęp 12 września 2016 r.]. Dostępny w Internecie: <<http://e-sklep.mik.krakow.pl/ebooks/raport-metodologiczny-cdk.pdf>>.

Fielding Nigel, Lee Raymond (1993) *Using Computers in Qualitative Research*. London: Sage Publications.

Fielding Nigel, Lee Raymond (1996) *Qualitative Data Analysis: Representations of a Technology. A Comment on Coffey, Holbrook and Atkinson*. „Sociological Research Online”, vol. 1, no. 4 [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.socresonline.org.uk/1/4/lf.html>>.

Fielding Nigel, Lee Raymond (1998) *Computer Analysis and Qualitative Research*. London: Sage Publications.

Gerson Elihu (1984) *Qualitative Research and the Computer*. „Qualitative Sociology”, vol. 7, no. 1/2, s. 61–74.

Głowacki Kamil (2014) *Organizacja procesu badawczego a oprogramowanie do organizacji wiedzy i zarządzania*. „Przegląd Socjologii Jakościowej”, t. 10, nr 2, s. 162–195 [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.przehladsocjologiijakosciowej.org>>.

Kamasa Victoria (2014) *Techniki językoznawstwa korpusowego wykorzystywane w krytycznej analizie dyskursu*. *Przegląd*. „Przegląd Socjologii Jakościowej”, t. 10, nr 2, s. 100–117 [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.przehladsocjologiijakosciowej.org>>.

Niedbalski Jakub (2012) *OpenCode – narzędzie wspomagające proces przeszukiwania i kodowania danych tekstowych w badaniach jakościowych*. „Przegląd Socjologii Jakościowej”, t. 8, nr 1, s. 220–228 [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.przehladsocjologiijakosciowej.org>>.

Niedbalski Jakub (2013) *Odkrywanie Caqdas. Wybrane bezpłatne programy komputerowe wspomagające analizę danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Niedbalski Jakub (2014a) *Komputerowe wspomaganie analizy danych jakościowych. Zastosowanie oprogramowania NVivo i Atlas.ti w projektach badawczych opartych na metodologii teorii ugruntowanej*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Niedbalski Jakub, red. (2014b) *Metody i techniki odkrywania wiedzy: narzędzia CAQDAS w procesie analizy danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Niedbalski Jakub (2014) *Zastosowanie oprogramowania Atlas.ti i NVivo w realizacji badań opartych na metodologii teorii ugruntowanej*. „Przegląd Socjologii Jakościowej”, t. 10, nr 2, s. 60–81 [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.przehladsocjologiijakosciowej.org>>.

Niedbalski Jakub, Ślęzak Izabela (2012) *Analiza danych jakościowych przy użyciu programu NVivo a zastosowanie procedur metodologii teorii ugruntowanej*. „Przegląd Socjologii Jakościowej”, t. 8, nr 1, s. 126–165 [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.przehladsocjologiijakosciowej.org>>.

Pfaffenberger Bryan (1988) *Microcomputer Applications in Qualitative Research*. Newbury Park: Sage Publications.

Ślęzak Izabela, Niedbalski Jakub (2014) *Główne funkcje programu NVivo a procedura metodologii teorii ugruntowanej, czyli jak realizować badanie oparte na MTU, korzystając z oprogramowania CAQDAS?* [w:] Niedbalski Jakub, red., *Metody i techniki odkrywania wiedzy: narzędzia CAQDAS w procesie analizy danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 77–92.

Tomanek Krzysztof (2014) *Analiza sentymentu – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych*. „Przegląd Socjologii Jakościowej”, t. 10, nr 2, s. 118–137 [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.przehladsocjologiijakosciowej.org>>.

Tomanek Krzysztof, Bryda Grzegorz (2014) *Odkrywanie wiedzy w wypowiedziach tekstowych. Metoda budowy słownika kla-*

syfikacyjnego [w:] Niedbalski Jakub, red., *Metody i techniki odkrywania wiedzy: narzędzia CAQDAS w procesie analizy danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 219–248.

Winczorek Jan (2014) *Wykorzystanie oprogramowania R i RQDA w jakościowo-ilościowej analizie treści orzeczeń Trybunału Konstytucyjnego*. „Przegląd Socjologii Jakościowej”, t. 10, nr 2, s. 138–161 [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.przehladsocjologiijakosciowej.org>>.

Cytowanie

Dzięglewski Mariusz (2017) *CAQDAS w badaniach digitalizacji i odbioru dziedzictwa kulturowego. Korzyści i ograniczenia*. „Przegląd Socjologii Jakościowej”, t. 13, nr 2, s. 24–45 [dostęp dzień, miesiąc, rok]. Dostępny w Internecie: <www.przehladsocjologiijakosciowej.org>.

CAQDAS in the Research Project on the Digitalization and Reception of Cultural Heritage. Advantages and Constraints

Abstract: The paper is aimed at the reflection on advantages and constraints emerging from the mutual diffusion of traditional research methods with computer assisted qualitative data analysis. The article includes critical analysis of this process based on the research project “Digital practices and strategies of dissemination and reception of cultural heritage in Poland between 2004–2014.” In the first part of the paper, the main assumptions of the project are discussed with the special insight into the function and fusion of those two approaches at the stage of the research project conceptualization. The author of the paper discusses the impact of CAQDAS on the perception of the research problem and latter interpretation of findings. In the second part of the article, the following stages of the research project are presented, with the special insight into the fusion of various approaches to the research at the stage of database construction, coding of textual data, analysis, data visualization, and interpretation of findings. The special attention has been paid to advantages and constraints of CAQDAS application to content analysis of legal documents and transcripts of in-depth individual interviews. The main outcome of such meta-analysis of the research process is the proposal of the most effective fusion of traditional and new methods of data analysis in complex research projects. We believe that such proposal enables one to make use of new technological opportunities in social research, avoid “pitfalls,” and overcome constraints embedded in CAQDAS.

Keywords: CAQDAS, In-Depth Interviews, Content Analysis

Jakub Niedbalski
University of Lodz

From a qualitative researcher's workshop— the characteristics of applying computer software in studies based on the grounded theory methodology

Abstract The purpose of the article is to show how computer aided qualitative data analysis tools can be utilized in research practice. Based on the example of a particular research project, I make an attempt to describe how studies are performed pursuant to the procedures of the grounded theory methodology, using NVivo software. This is a presentation of a workshop of a qualitative researcher who uses computer software to aid the research process on a daily basis. At the same time, I stress the existing improvements, but also the consequences and potential difficulties related to the application of computer software in qualitative research. The article is of a review and educational character, and it is intended to familiarize the reader with the possibilities of a tool belonging to CAQDAS and its real application in carrying out a research project based on a selected research method.

Keywords Computer aided qualitative data analysis, NVivo, grounded theory methodology, studies into sport of persons with disabilities

Jakub Niedbalski, PhD in sociology, Assistant Professor in the Department of Sociology and Organization of Management of the University of Lodz. He specializes in computer analysis of qualitative data, methods of qualitative research, notions of sociology of disability and of physical culture. He conducts research in the field of social and physical activation of the disabled. Author of the following books: *To live and work in a social welfare home. Sociological study of interactions between personnel and mentally disabled wards* (The Uni-

versity of Lodz Publishing House, Lodz 2013), *Discovering of CAQDAS. Description of the selected computer programs supporting qualitative data analysis* (The University of Lodz Publishing House, Lodz 2013) and *Computer Assisted Qualitative Data Analysis Software. Using of the NVivo and Atlas.ti in the research projects based on methodology of the grounded theory* (The University of Lodz Publishing House, Lodz 2014).

email: jakub.niedbalski@gmail.com

Introduction¹

As the author of this article, my intention was to demonstrate the manner in which NVivo – a tool that belongs to the CAQDAS family – allows me to successfully perform research that is based on the grounded theory methodology. On the basis of the analysis that I carried out, I would like to demonstrate how to carry out data analysis by combining various functions of the NVivo program, that is, moving from original materials towards increasingly more advanced stages of theoretization and work on a higher conceptual level which would result in the development of a theory (Niedbalski 2014).

In the article, I have made an attempt – based on my own research – to present how computer-aided qualitative data analysis tools can be utilized so that the process of data interpretation can be carried out in accordance with the grounded theory methodology in a correct, effective and convenient manner (Niedbalski 2014). I do not try to describe methodologically the particular options and functions implemented in the computer software, rather I document and recreate how they served to create empirical materials while carrying out research into disabled individuals.

The contents presented in the article are a result of both my personal experiences, as well as studies into the literature of the subject (cf. Lonkila 1995; Kelle

¹ This article adopts certain fragments of my book entitled "Komputerowe wspomaganie analizy danych jakościowych. Zastosowanie oprogramowania NVivo i Atlas.ti w projektach badawczych opartych na metodologii teorii ugruntowanej." [Computer-aided qualitative data analysis. NVivo and Atlas.ti software in research projects based on the grounded theory methodology], Wydawnictwo UŁ, Łódź, 2014.

2005; Gibbs, 2011), raising the issue of computer-aided qualitative data analysis software in the context of research based on the procedures of grounded theory methodology.

Preparation of data

While starting to work in the CAQDAS environment, we first of all need to pay attention to two main stages that an analyst's work may be divided into; and which correspond to particular sets of actions carried out with the mentioned software. These are: the preparatory work level (data management) and the conceptual level (of analytical work).

The first level mainly covers the actions related to preparing the materials (sources of data), the creation of a project, collecting data, and their proper sorting and grouping. This stage also covers the activities of the software user, such as: describing and transcribing data, as well as drawing up notes that play an informative role, or performing a simple search of the data in the project. Therefore, there is a whole series of actions that the user must take to carry out proper analytical and conceptual work. But at the same time, they pose a significant stage in the researcher's activity, as all of those actions provide the foundations for our interpretation of the data, that is, the process of theoretizing, including the development of the theory.

In turn, the conceptual level refers to the whole analytical process, which is based on such actions as: data coding, sorting and allocating it to particular categories, and then combining data and searching

through them in terms of connections and interdependencies. Hence, this level is inextricably linked to the development of hypotheses and establishing them in the source of further data analysis and interpretation. In this stage, one may also find helpful software functions – apart from the already mentioned coding and categorization – such as drawing up theoretical memos, the possibility to regroup and continuously modify generated data, using various kinds of search tools that are helpful in the process of not only searching through contents, codes or categories but also searching data on the basis of research queries, or finally tools that provide the possibility to visualize the data analysis process results and create models that reflect the network of connections and interdependencies that exist between elements of the project. And this is the aspect of work in the NVivo environment that the other three sections of the article are devoted to. I would like to stress that I focus here on the manner of conduct that complies with the grounded theory methodology and the arrangement of particular actions that accompany this process, and less space is devoted to the technical issues of the software itself, which can be found in other works of literature on the subject (cf. Niedbalski and Ślęzak 2012; Niedbalski 2014; Niedbalski and Ślęzak 2014).

Coding

A basic procedure of analysis in the case of the grounded theory is coding, which means the process of determining “what the data is about” (Charmaz and Mitchell 2001:340) and is referred to as “connecting data without views about them” (Coffey, Holbrook, and Atkinson 1996:27; cf. Lofland et

al. 2009:275). In other words, the coding process consists in providing selected parts of the material with particular labels that give information about the contents of a given fragment. Regarding the degree of abstraction of such a label, “material” and “theoretical” codes are differentiated (Konecki 2000:51 and further). Furthermore, coding usually takes place according to a certain scheme: open coding (initial), when many codes are allocated, with the intention of labelling a given fragment in various manners, and then there is focused coding - focused on selected, key categories with a greater theoretical significance (Lofland et al. 2009:275). Thus, following the advice given by Strauss and Corbin (1990:59), I divided the particular stages of work into: “open coding” (a certain type of data reconstruction), “focusing on category coding” (which means the reconstruction that takes place at the level of generated categories) and “selective coding” (employed in order to validate the developed hypotheses).

Open coding

I started my research by going into the field and “immersing” myself in the world of the researched individuals – in this case, disabled people who practice sport. My first action was to get acquainted with both the place where I would be observing these people and with the people themselves who were my potential interviewees within the scope of future interviews. Therefore, I treated the observation technique, on the one hand, as a manner to obtain data that are significant for the research, and on the other, as a source of “pilot” information on representatives of the researched society. When I decided that I had acquired a certain knowledge about the re-

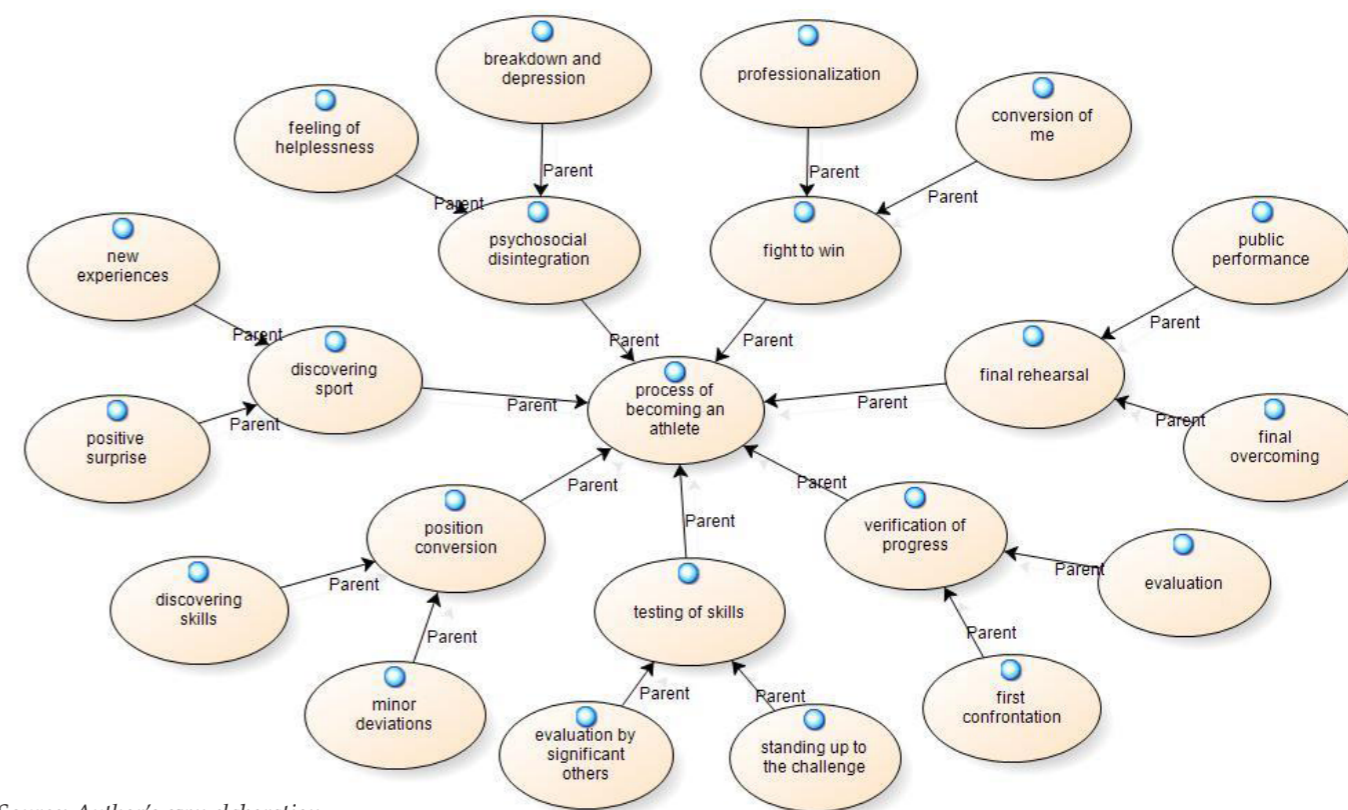
searched environment and, at the same time – based on my observations – I concluded that I had been accepted by its members to a greater extent than was the case at the beginning of my presence, and I was no longer making an unnecessary fuss among these people anymore, I started the interviews.

As research based on the grounded theory (GT) methodology is characterized by the fact that it is impossible to separate its subsequent stages – since applying GT means that they become intertwined, shifting fluidly from one into the other – I started to analyze the data which was coming from the observations, just as with the transcribed interviews, which provided me with an initial list of codes and categories. During the analysis process, I developed

and used almost 500 codes, which were gradually combined, grouped or simply excluded, ultimately providing 40 categories. The interviews and notes from the observations were coded as a “stage” in relation to the course of the story, which is composed of the following phases: “psychosocial disintegration,” “discovering sport,” “conversion of position,” “skills testing,” “progress verification,” “final rehearsal,” “fight for a win” and “specifically.” Fragments related to turning points identified on the basis of the data analysis within particular phases of the process of becoming an athlete were marked within particular stages.

During the analysis, there was a key for material codes developed for each phase in the scope of the

Figure 1. Coding scheme picturing various levels of categories



Source: Author's own elaboration.

main process, that is, the development of a sports career. Grouping codes into a hierarchical code took place from the bottom. Each of the above-mentioned phases was allocated several dozen codes which were later arranged into more general categories until seven main groups were achieved. The scheme of coding is presented in Figure 1. It includes the two highest levels of codes, corresponding to the phases and exemplary turning points included in their scope, which determined the sequence of the process of becoming a person who practices sport.

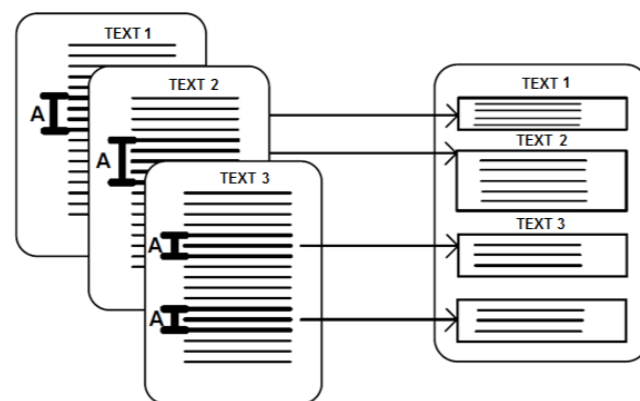
Constant comparison

The process of providing categories, searching for relations between them, and developing hypotheses would not be possible without consideration of the “constant comparison” procedure, according to which a researcher-analyst searches for differences and similarities between data fragments, codes or cases, as well as determined dimensions that may occur between them. In other words, the constant comparison method is about confronting various components of the project in order to check the similarities between them or to emphasize certain features that differentiate them. Increasingly more general categories revealing underlying uniformities are generated on this basis (Gorzko 2008:86). It provides an opportunity to gradually focus on the first of all those fragments of obtained information that start to match each other in some terms.

An example here may be the coded contents of particular interviews and memos that come from observations, regarding all the information on the significance of sport apportioned by the disabled individ-

uals. Memories and thoughts regarding the current situation of each of my interviewees were often related to mentioning various (in terms of content and form) statements, but referring to the same notion, which I finally coded as the “function” of sport. For example, the function that is called “revealing” is actually a category that was generated on the basis of 23 fragments that come from particular interviews coded originally in the following manner: “overcoming the fear of showing oneself,” “publicly revealing one’s body,” “showing oneself without a prosthesis,” or “fighting shame in front of others.” Finally, there were a dozen or so functions determined. These are: “improvement,” “separation,” “strengthening,” “compensation,” “challenge,” “escape,” “support,” “reveal,” “prove,” and “pro-integration.” Each is a separate category that was created as a result of comparing from a few to a dozen or so fragments that bear particular etiquettes (codes allocated during material coding).

Figure 2. Comparison scheme for the contents of codes (function of data “retrieval”)



Source: Based on *Computer-Assisted Analysis of Qualitative Data*. Paper prepared for the Discussion paper series of the LSE Methodology Institute by Udo Kelle, University of Bremen, August 1997.

Cases

The comparisons may also be made between various people, items, scenarios, or situations, as well as information obtained from the same people as regards places, items, or types of events (Charmaz and Mitchell 2001:165). Therefore, the constant comparison method should be related to the comparison of events and notions with other observed cases, and to the comparison of notions (Glaser 1978:49-50).

Cases may serve as a representation of particular people or organizations that pose a subject of interest to the researcher. Thanks to the numerous options of the program, particular cases may be given attributes and qualities, among the other particulars. It means that we can, for example, create separate cases that will, in turn, be described by selected attributes and their qualities, such as sex or education.

While making use of those possibilities in my research, I compared statements made by particular interviewees divided with regard to sex, in topics (or threads) which were interesting for me. From a practical perspective, I was able to search through segments of text that were limited by certain parameters (variables), which allowed me to, for example, separate information related to some topics that had come from those participants of the research that had a common feature. For example, among other things, I wanted my research to specify the scope of similarities and differences regarding the situation of disabled women and men who practice sport, respectively. Hence, I excerpted fragments from particular texts where women spoke about

the manner in which they are perceived by those around them, and how – according to them – others see their dysfunctions, and I compared them to the segments of texts raising the same topic, but coming from interviews with men. Such parameters related to the comparison of particular data allowed me to put forth several initial hypotheses related to the gender of disabled athletes. An example may be provided by the following hypothesis: “Men may count on greater approval from those around them as regards their bodies, which, despite certain dysfunctions, usually do not pose such a great barrier in relationships with the environment as they do in the case of women.”

The coding paradigm—theoretical coding

In the case of numerous research projects, comparing text segments leads to descriptive typologies and theory development. An additional tool – a procedure present in the grounded theory methodology which plays a helpful role in the process – is the so-called coding paradigm, or focused coding, in other words. Strauss (1987:27) proposes that coding should not be limited just to naming categories, but it should also consider the causes of its occurrence, intervention conditions, interactions, strategies, and tactics of action, as well as the context and consequences of actions described by a given category. These elements pose the so-called coding paradigm, and such coding is called focused category coding (Strauss and Corbin 1990:96-116; Kelle 1997a:7-8; Konecki 2000:48a). It consists of specifying categories, their development and making references between them or searching for mutual connections (Gibbs 2011:98).

Here, it is worth remembering that the main objective of focused coding according to Strauss is:

- to reveal the presence of attributes of categories and their dimensions (which starts as early as during the open coding);
- the identification of a variety of conditions, actions/interactions, and consequences related to the “phenomenon”;
- referring a category to its sub-categories and determining their connections;
- searching for data clues, which can suggest the dependencies that may be established between the main categories (Strauss 1987).

According to Strauss, the analytical procedures of focused coding allow the researcher to recognize the relationships between the structure and the process (cf. Gorzko 2008:313). Also in this stage, the NVivo program provides the researcher with support, because it has elaborate functions which support the development of the hierarchy of categories, but also enable the creation of connections between codes. It is significant because, thanks to such tools, it becomes possible to continue the analysis and to take it to a higher conceptual level (Fielding and Lee 1998). The existence of various functions allows the user to create a structure of categories, which facilitates the process of arranging the coding results, thus enabling a reflection of the “superiority-inferiority” relationship (including the specification of categories, subcategories, and their qualities). It also allows the user to determine the character and rela-

tionships connecting the generated categories more precisely, meaning that it becomes possible to use the coding paradigm in the coding process. In other words, the coding paradigm means a general theory of action, which may be used to develop the structure or “axis” of a developing theory. In the research that I described, I applied the five-element Strauss and Corbin model (1990), in order to construct an initial analytical scheme: causative conditions, intervening conditions, context, micro-actions, and consequences. After completing those actions, the purpose of the studies, which, until that point had been very general, became more concrete and precise. While using the suggestions proposed by Strauss and Corbin (1990), I constructed the following structure of notions explaining the phenomenon that I was studying:

1. The causative conditions, which in my research meant comparing factors and circumstances that caused the main action, that is, the decision to start practicing sport and the gradual move from the physical activity being treated as a form of rehabilitation to an extreme sport;
2. The phenomenon (the main category), according to my analyses, was the reconstruction of the process of becoming an athlete.
3. Intervening actions, which meant the cognitive model was related to the perception of their own corporeality and a reconstructed concept of disability;
4. The concept was brought down to the circumstances that limited or supported the main ac-

tion, relating to the behaviors and actions of people from the disabled athlete's surroundings;

5. The strategies of actions/interactions were represented by two extreme categories: independence (internal steering) or dependence (external steering) in constructing the career of the disabled athlete.

Based on such elements of paradigm coding, I also succeeded in operationalizing so-called consequences, which meant reconstructing the notion of disability and the process of role realization and, later on, the concept of a disabled sportsman's career.

Selective coding

The third phase of the analysis, according to Strauss and Corbin (1990:119) shifts to the construction of the “story” (Strauss and Corbin 1990:119) which is composed of a limited number of hypothetical statements which serve to construct the theory after being “tested.” In my research, the story was called “the reconstruction of ‘me’ from being a disabled individual to an athlete” (main category, and some of its hypothetical findings were presented in the following manner:

1. Sport allows reconstruction in the perception of the self;
2. The reconstruction of “me” depends both on external factors, as well as on the degree of motivation for the action and the willingness to search for meaning in the sport activity.

3. In the process of the reconstruction of “me” in the role of an athlete, the prevailing factor is the concept of disability felt by a given person;

4. Limitations that flow from the environment, including the behaviors and actions undertaken by others, hinder or reduce the chance for success of the disabled in the “me” reconstruction process through practicing sport.

5. The strengthening that comes from the environment, including the behaviors and actions undertaken by others, facilitate or improve the chances for success of the disabled in the “me” reconstruction process through practicing sport.

6. The consequence of the “me” reconstruction is the adoption of a certain concept for the realization of the role of the athlete.

Having drawn up the story, another step was to go back into the field in order to collect other samples of data. Therefore, I gathered further information in order to document the relevance of the above-mentioned statements, and thus to continue the basic task that, in the case of GT, should generate the theory. Research that leads to a theory is a time-consuming process that demands great amounts of work. Therefore, it is of great importance for the researcher to make the most of the information they possess, not only what they have just collected, but also the information they already had. Hence, apart from carrying out actions in the field, I also went back to my “old” notes with a new research “objective,” to confirm or reject the theses from the presented story. Each thesis must be supported with an

accurate description that complies with the facts. In other words, it must be enriched with details and episodes that show the reader that the research is solid and rich. Thus, I extended the story, I composed a more elaborate model, and I made the theory a little more “sophisticated.”

This is also the area where the NVivo program helped me, as it offers tools for drawing up theoretical memos. Here, it is worth remembering that the memos are a reflection of the analytical thoughts related to the codes, and they are adopted to make the applied categories more specific and to provide the coding process with a direction. What is more, as suggested by Gibbs (2011:68), they also provide a certain connector between two stages of the analysis – coding and writing a report. According to Gorzko (2008:101) the written notes are a kind of analytical tool, creating a certain thought and theoretical space where the researcher can conceptualize the data.

In the NVivo program, the role of theoretical notes is played by memos, that is, records of theoretical thoughts and concepts by the researcher. Memos are concise notes drawn up by the researcher, including information on concepts regarding the whole project, particular material, or issues to be discussed or interpreted in the future. The concept of memos in these programs is analogous to the procedure of generating notes in the grounded theory methodology. They are there to help the researcher move to a higher conceptual level, and they serve to generate theories as tools of theoretical coding (Konecki 2000). Therefore, the convergence of functions available in NVivo with the requirements that the

researcher is provided with by the research methodology is made visible. If used properly, that is, for successive and systematic creation, it may effectively contribute to the creation of a theory.

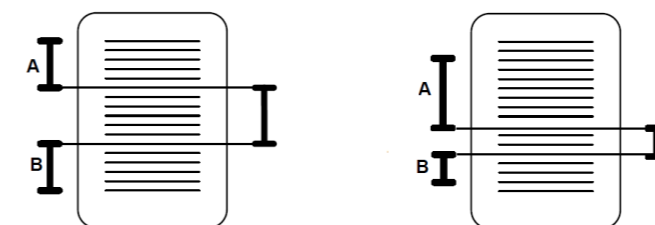
Strengthening the hypothesis

The hypotheses that were generated must be “verified.” However, it needs to be kept in mind that in the case of qualitative research, the verification of hypotheses is a process that cannot be brought down to statistical testing. Instead of “testing” and “confirming” hypotheses (Miles and Huberman 2000:262), or “verifying” them (Strauss and Corbin 1990:108) the process of reviewing them means returning to the data (i.e., reading the shorthand notes or field notes again), or moving back to the territory of research (i.e., carrying out new observations and interviews) in order to find any confirming or disqualifying evidence for the correctness of the generated hypotheses. This results from, among others, the specificity of the qualitative hypotheses, which – especially in their initial form – are usually quite loose “connections” made by the researcher.

The computer program which I used to analyze the data also had special functions which enabled me to “verify” the hypotheses through scanning parts of interviews and notes from observations. Therefore, computer assisted qualitative data analysis software may be useful to improve theoretical concepts, and to create and “establish” hypotheses. In my research, for example, I was willing to test my intuition about the interdependency between involvement in a disabled person practicing sport and the influence others have on the process. In order to research such an

initial hypothesis, the option to search data in terms of the spatial presence of codes in the source materials turned out to be useful. For example, we can introduce a proper configuration, pointing to segments of texts coded with an “x” and those coded with a “y,” which appear quite far from the first instances of them being mentioned. Therefore, the hypothesis of the relationship between sport being practiced by a disabled individual and the impact other people have on the process may be studied through searching for all elements of the text coded by “sport interest” and fragments coded with “persons from their surroundings,” located a certain distance from the first time it is mentioned (expressed by the number of paragraphs). On the basis of a search presented in such a manner, it might be agreed that interviews carried out with various individuals were coded in such a manner that there are text fragment codes that represent the distance between fragments’. If such an interdependence is of a repeatable character, and it can be observed in various interviews, the researcher may claim that the hypothesis is confirmed in the current pool of data.

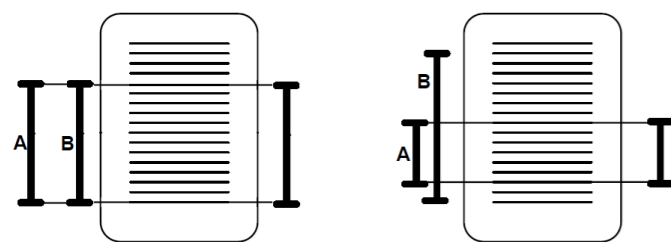
Figure 3. Search scheme for data fragments marked with selected codes in terms of distance of presence



Source: Based on *Computer-Assisted Analysis of Qualitative Data*. Paper prepared for the Discussion paper series of the LSE Methodology Institute by Udo Kelle, University of Bremen, August 1997.

A way to verify the hypotheses was also to use the code search tools adopting logic operators. As in the case of distance operators, the objective here is not the literal verification of hypothesis correctness but to download an original text encoded with particular codes and analyze it again. For example, I used two codes in my research for this purpose – “disability” and “significance of physical activity,” and two attributes – “man” and “woman.” Afterwards, while analyzing the data, I spotted a connection between such codes and attributes which gave grounds for the hypothesis: “Physical activity is a value in itself, most of all in the case of disabled men rather than in the case of disabled women.” Afterwards, I wanted to check whether the relationship between codes and attributes was confirmed in other cases, that is, in the remaining materials. Having introduced proper search parameters and adopting the “AND” operator between the “value in itself,” “physical activity” and “man” codes, and then later repeating the same procedure, but replacing the “man” attribute with “woman,” I was able to check in what documents and how often these codes occur in the same fragments of data. The use of the “and” operator meant that only those fragments of text were found that I had coded with both the first and the second code that provided the basis for the search (in other words, the program will not find data segments coded with only one code or those fragments that were not allocated any of the codes), and which at the same time were marked as words spoken by interviewees that were women or men (the scopes of codes does not need to fully overlap, it is enough when there are common parts).

Figure 4. Search scheme in terms of the co-existence of data fragments allocated with selected codes



Source: Based on *Computer-Assisted Analysis of Qualitative Data*. Paper prepared for the Discussion paper series of the LSE Methodology Institute by Udo Kelle, University of Bremen, August 1997.

Preserving a work style, as recommended by the authors of *The discovery of grounded theory*, in other words, treating the study as a whole process that is consciously directed at generating theories, leads very quickly to – as claimed by Glaser and Strauss (1967) – the formulation of a multiplicity of hypotheses. At the beginning not connected, in a short time they start to form a theoretical framework of the research (Gorzko 2008:84). Therefore, from the perspective of the processual character of the generation of theories, it seems useful that the created inquiry and interconnected search query may be updated according to subsequent changes introduced into the project. We can restate the inquiry at regular time intervals, and hence monitor the development process of our coding, and evaluate whether the recent analyses are going in a direction that is satisfactory for us.

Finally, there is a need to mention the existence of certain methodological requirements and limitations in the application of the search strategy in the

verification of hypotheses. First of all, the testing condition is that the search tools do not look through the same materials that served for the initial development of the hypothesis. Secondly, the hypotheses must be empirically verifiable, meaning that they must be precise enough and that they should have an empirical part. Thirdly, the codes employed for hypotheses testing must mark phenomena that are clearly determined (Kelle 1997a:14).

Furthermore, a search tool should not be used in an automatic manner, or the co-existence of codes should not be treated as an indicator confirming a hypothesis or not, but only as a clue that suggests the researcher's intuition complies with information that comes from particular interviewees (Kelle 1997b).

Moreover, it is worth remembering that although the search functions are one of the most effective tools available in CAQDAS, they are not without flaws and limitations. First of all, they can be used only when the text coding was carried out with special care and the particular attributes of given objects were specified. Secondly, discovering relationships between data may be reliable when the met assumptions are related to, among other things, the completeness of the recorded text and its adequate arrangement. Thirdly, it must be kept in mind that, at any given moment in a conversation, people may recall several various notions; and they remember certain matter later, often adding to their statement.

Models and diagrams

Strauss and Corbin (1990) suggest that, during the process of generating theories, apart from coding,

sorting and arranging information, when seeking patterns between data and using a system of notes, we should also turn to visual representations of connections and interdependencies between generated analytical categories. Any visualizations in the form of charts, diagrams or networks are useful to order the relationships between categories which emerge during selective coding. From the GT methodology perspective, the most significant models are those which form the basis of diagrams that integrate data. Schemes, diagrams or models are used to visually represent connections and interdependencies which exist between components of the developed theory. What is more, in contrast to linear representations of various relationships, the network distribution of those relations is closer to the human manner of perceiving reality, and it therefore becomes one of the most important interpretation processes of the analyst (Muhr and Firese 2003/2004:211).

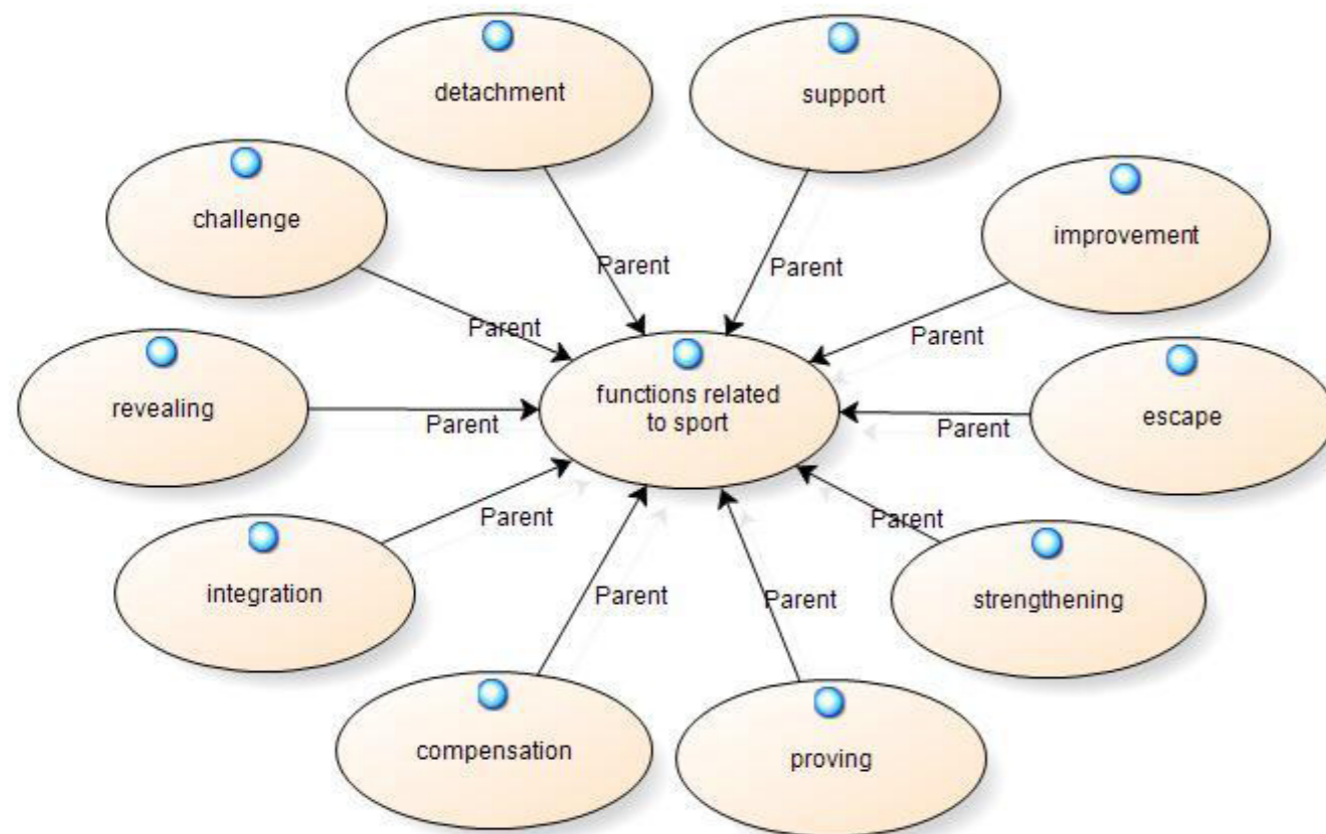
The comparisons, references and explanations formulated during data analysis form the basis for constructing and confirming the correctness of the models (Gibbs 2011:155). The model itself has a certain structure, within the scope of which the aspects of the analyzed phenomenon which were defined earlier as crucial are presented in relation to other aspects or qualities of a given situation (Gibbs 2011:154). The creation of models is of crucial significance within the whole process of data analysis, although it is especially visible when crystallizing the researcher's concept, which in turn is strictly connected with generating a central category. The visualization of categories and their relationships facilitates this process to a significant extent. Thus, the models play a crucial role during selective cod-

ing, where crystallization of the main phenomenon or process that all of the categories generated within the course of the analysis will be related to takes place. Strauss and Corbin (1990) recommend that, having refined the coding scheme, ordering the categories and comparing the cases a model should be done, specifying those elements that comply with the coding paradigm.

At the beginning of the analysis, we may use a model to plan the main research actions and draw our expectations or initial concepts. Afterwards, the models may help in presenting relationships between notions or other elements of the project.

The NVivo program, as a tool intended to support the researcher in the development of a theory, enables him to create models. These models can help to determine and review the initial concepts and idea on the questions that are interesting for the researcher, visually represent relationships between the project elements, identify emerging patterns, theories and explanations, as well as document and record subsequent stages of work over the project. Below is an example of a model that includes categories depicting functions that the researchers ascribed to disabled individuals' involvement in a sport activity.

Thanks to such a visualization of data, it has become much more convenient for me – a researcher performing analyses that lead towards the generation of theories – to compare various elements of a single project. First of all, the representation of subsequent stages of an analysis in the form of a mind map allowed for an improved observation of

Figure 5. Functions related to disabled individuals practicing sport

Source: Author's own elaboration.

the relationships and patterns of data. At the same time, applying the modeling function allowed me to develop a project draft, and a vision of my own ideas related to the development of material (Miles, and Huberman 2000; Seale 2008).

Finally, it is worth adding that when using a computer program, and creating various visual representations of the data analysis that we carried out (including the creation of an integrating diagram that somehow crowned the whole process), we can document the course of all actions undertaken by the researcher in this scope. Therefore, we – as researchers – can monitor subsequent stages of the formulation of our analytical “path,” but it also means we become

more transparent to our reader, who have a possibility to look into the development of the whole research process. Such a direct and tangible expression of this action may be a presentation of the history of the study, which we can present not only in a descriptive manner but also through the data exported from the NVivo program in the form of images presenting selected moments of the analysis that we conducted.

Conclusions

Taking into account the challenges that researchers who employ qualitative methods need to face, especially those who implement their own projects based on the procedures of the grounded theory methodol-

ogy, it may be assumed that CAQDAS tools are a good way to improve the effectiveness and convenience of the work. The presented NVivo program is certainly such a tool, as it proves to be useful both with the material and in the theoretical coding process, through the systemization of coding and the ease of seeking appropriate fragments during data development and hypotheses “verification.” The NVivo program was also helpful when searching for connections between particular categories. While working in the program and using the implemented functions, we managed to perform data analysis (of interviews, notes from observations of existing data, as well as audio and video materials) in a manner which corresponded to the requirements put forward by the procedures of the grounded theory methodology. It is also worth emphasizing that the program proved to be helpful as a tool for collecting, and at the same time controlling a significant amount of material, which can be simply processed, modified, sorted and reorganized, as well as searched through. It enables the researcher to gain greater control over the collected data. It is also accompanied by the possibility to subordinate various elements of the projects through grouping them in accordance with the preferences of the researcher (Wiltshier 2011:4). The NVivo program allows comprehensive sorting of data – both source materials and any information resulting from an analysis carried out by a researcher (Seale 2008:235). It must also be kept in mind that computer software devoted to qualitative data analysis creates the possibility to constantly modify all project elements as new data emerges (Bringer, Johnston, and Brackenridge 2006:248). The flexible manner of creating and modifying the project elements allows the researcher to follow the data, and the generated code may

be quickly modified if it is decided that it does not reflect the data content to a sufficient extent (Glaser 1978:4-5; Konecki 2000:28). At the same time, the system of analytical notes allows the smooth alteration of actions related to the collection and analysis of the data. Furthermore, it is a program equipped with tools that facilitate the introduction of data triangulation procedures, methods or other researchers into a research project (Konecki 2000:86). It allows the cooperation of numerous researchers involved in a single project thanks to such functions as: identifying team members and tracing and comparing actions undertaken by particular individuals, including the possibility to verify who added what data and when they did it, and what modifications they introduced (Seale 2008:238; Wiltshier 2011:1-2;). It is a highly valuable possibility, especially if we consider the growing internationalization of research teams.

An additional asset of the software is the fact that all stages of a researcher's analytical work are recorded in the program. It is possible to verify the codes, categories, and memos created and show the analytical thought of a user developed at every moment. It allows the researcher to present the methodological and analytical background in a clear manner (Morse and Richards 2002 as cited in Bringer, Johnston, and Brackenridge 2004:252). It is also worth emphasizing that the program architecture itself somewhat forces the researcher to continuously think about the relationships between codes and categories, to compare and modify them, that is, to carry out a systematic analysis. Hence, the problem that awaits the qualitative researcher, connected with focusing just on data gathering and avoiding its in-depth analysis, may be avoided (Hammersley and Atkinson 2000:196).

References

- Bringer, Joy D., Lynne H. Johnston, and Celia H. Brackenridge. 2004. "Maximizing Transparency in a Doctoral Thesis: The Complexities of Writing about the Use of QSR*NVIVO within a Grounded Theory Study." *Qualitative Research* 4(2):247-265.
- Bringer, Joy D., Lynne H. Johnston, and Celia H. Brackenridge. 2006. "Using Computer-Assisted Qualitative Data Analysis Software to Develop a Grounded Theory Project." *Field Methods* 18(3):245-266.
- Charmaz, Kathy and Richard G. Mitchell. 2001. "Grounded Theory in Ethnography." Pp. 160-174 in *Handbook of Ethnography*, edited by Paul Atkinson. London: Sage.
- Coffey, Amanda, Beverly Holbrook, and Paul Atkinson. 1996. "Qualitative Data Analysis: Technologies and Representations." *Sociological Research Online* 1(1). Retrieved May 06, 2015 (<http://www.socresonline.org.uk/1/1/4.html>).
- Fielding, Nigel and Raymond Lee M. 1998. *Computer Assisted Qualitative Research*. Newbury Park: Sage.
- Gibbs, Graham. 2011. *Analizowanie danych jakościowych*. Translated by Maja Brzozowska-Brywczyńska. Warszawa: PWN.
- Glaser, Barney G. 1978. *Theoretical Sensitivity*. San Francisco: The Sociology Press.
- Glaser, Barney G. and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory. Strategies for Qualitative Research*. Chicago: Aldine Publishing Company.
- Gorzko, Marek. 2008. *Procedury i emergencja. O metodologii klasycznych odmian teorii ugruntowanej*. Szczecin: Wydawnictwo Naukowe Uniwersytetu Szczecińskiego.
- Hammersley, Martyn and Atkinson Paul. 2000. *Metody badań terenowych*. Translated by Sławomir Dymczyk. Poznań: Wydawnictwo Zysk i S-ka.
- Kelle, Udo. 1997a. "Theory Building in Qualitative Research and Computer Programs for the Management of Textual Data." *Sociological Research Online* 2(2). Retrieved May 06, 2015 (<http://www.socresonline.org.uk/socresonline/2/2/1.html>).
- Kelle, Udo. 1997b. *Computer-Assisted Analysis of Qualitative Data*. Paper prepared for the Discussion paper series of the LSE Methodology Institute by University of Bremen, August 1997.
- Kelle, Udo. 2005. "Computer-Assisted Qualitative Data Analysis." Pp. 473-490 in *Qualitative Research Practis*, edited by Clive Seale et al. London, Thousand Oaks, New Delhi: Sage.
- Konecki, Krzysztof. 2000. *Studia z metodologii badań jakościowych. Teoria ugruntowana*. Warszawa: PWN.
- Lofland, John et al. 2009. *Analiza układów społecznych. Przewodnik metodologiczny po badaniach jakościowych*. Translated by Anna Kordasiewicz, Sylwia Urbańska, and Monika Żychlińska. Warszawa: Wydawnictwo Naukowe Scholar.
- Lonkila, Markku. 1995. "Grounded Theory as an Emerging Paradigm for Computer-Assisted Qualitative Data Analysis." Pp. 41-51 in *Computer-Aided Qualitative Data Analysis*, edited by Udo Kelle. London: Sage.
- Miles, Matthew B. and Michael A. Huberman. 2000. *Analiza danych jakościowych*. Translated by Stanisław Zabielski. Białystok: Transhumana.
- Morse, Janice M. and Lyn Richards. 2002. *Readme First for a User's Guide to Qualitative Methods*. Thousand Oaks, London, New Delhi: Sage.
- Muhr, Thomas and Susanne Frieze (2003/2004) *User's Manual for ATLAS.ti 5.0*. Berlin: Scientific Software Development.
- Niedbalski, Jakub. 2014. *Komputerowe wspomaganie analizy danych jakościowych. Zastosowanie oprogramowania NVivo i Atlas.ti w projektach badawczych opartych na metodologii teorii ugruntowanej*. Łódź: Wydawnictwo UŁ.
- Niedbalski, Jakub and Izabela Ślęzak. 2014. „Główne funkcje programu NVivo a procedury metodologii teorii ugruntowanej, czyli jak realizować badanie oparte na MTU, korzystając z oprogramowania CAQDA?” Pp. 77-92 in *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analizy danych jakościowych*, edited by Jakub Niedbalski. Łódź: Wydawnictwo UŁ.
- Niedbalski, Jakub and Izabela Ślęzak. 2012. „Analiza danych jakościowych przy użyciu programu NVivo a zastosowanie procedur metodologii teorii ugruntowanej.” *Przegląd Socjologii Jakościowej* 8(1):126-165. Retrieved May 06, 2015 (<http://www.przegladsocjologiijakosciowej.org>).
- Seale, Clive. 2008. „Wykorzystanie komputera w analizie danych jakościowych.” Pp. 233-256 in *Prowadzenie badań jakościowych*, edited by David Silverman. Warszawa: PWN.
- Strauss, Anselm L. 1987. *Qualitative Analysis for Social Scientists*. Cambridge: University Press Cambridge.
- Strauss, Anselm L. and Juliet Corbin. 1990. *Basics of Qualitative Research*. London, New Delhi: Sage.
- Wiltshier, Fiona. 2011. "Researching With NVivo." *Forum: Qualitative Social Research* 12(1). Retrieved May 06, 2015 (<http://www.qualitative-research.net/index.php/fqs/issue/view/36>).

Citation

Niedbalski, Jakub. 2017. "From a qualitative researcher's workshop—the characteristics of applying computer software in studies based on the grounded theory methodology." *Przegląd Socjologii Jakościowej*, 13(2):46-61. Retrieved Month, Year (<http://www.przegladsocjologiijakosciowej.org>).

Z warsztatu badacza jakościowego – charakterystyka wykorzystania oprogramowania komputerowego w badaniach opartych na metodologii teorii ugruntowanej

Abstrakt: Celem artykułu jest ukazanie, w jaki sposób można wykorzystać narzędzia komputerowego wspomaganie analizy danych jakościowych w praktyce badawczej. Na przykładzie konkretnego projektu badawczego staram się przybliżyć, jak realizować badania zgodnie z procedurami metodologii teorii ugruntowanej, korzystając z programu NVivo. W ten sposób ukazany zostaje warsztat badacza jakościowego stosującego na co dzień oprogramowanie komputerowe wspomagające proces badawczy. Jednocześnie wskazuję na istniejące udogodnienia, ale także konsekwencje i potencjalne trudności związane ze stosowaniem programu komputerowego w badaniach jakościowych. Artykuł ma charakter poglądowy i edukacyjny, który ma za zadanie zapoznać czytelników z możliwościami narzędzia należącego do rodziny CAQDAS oraz jego faktycznym zastosowaniem w realizacji projektu badawczego opartego na wybranej metodzie badawczej.

Słowa kluczowe: komputerowa analiza danych jakościowych, NVivo, metodologia teorii ugruntowanej (MTU), badania nad sportem osób niepełnosprawnych

Marek Troszyński
Collegium Civitas

Aleksander Wawer
Instytut Podstaw Informatyki PAN

Czy komputer rozpozna hejtera? Wykorzystanie uczenia maszynowego (ML) w jakościowej analizie danych

Abstrakt Celem artykułu jest przedstawienie procesu automatyzacji kodowania tekstów pochodzących z mediów społecznościowych. Wdrożenie tego procesu pozwala na ilościowe potraktowanie jakościowych metod analizy treści. W efekcie otrzymujemy możliwość przeprowadzenia analizy na korpusach liczących setki tysięcy tekstów, które są kodowane w oparciu o ich znaczenia. Jest to możliwe dzięki wykorzystaniu algorytmów uczenia maszynowego (ML).

Omawianą metodę kodowania prezentujemy na przykładzie projektu oznaczania „mowy nienawiści” w tekstach pochodzących z polskich forów internetowych. Kluczowym problemem jest precyzyjna konceptualizacja i operacjonalizacja tej kategorii. Pozwala to na przygotowanie dokładnej instrukcji kodowej oraz przeprowadzenie treningu zespołu kodującego. Efektem jest podwyższenie współczynnika zgodności kodujących. Oznaczone teksty zostaną wykorzystane jako dane treningowe dla metod automatycznej kategoryzacji opartych o algorytmy uczenia maszynowego. W dalszej części artykułu opisujemy zastosowane metody kodowania automatycznego. Tekst kończy podsumowanie wskazujące na czynniki, które są kluczowe dla procesu badawczego wykorzystującego uczenie maszynowe.

Słowa kluczowe jakościowa analiza treści, uczenie maszynowe, mowa nienawiści, zgodność kodujących

Marek Troszyński, doktor socjologii, kierownik Obserwatorium Cywilizacji Cyfrowej Collegium Civitas, adiunkt tamże. Zainteresowania naukowe: socjologia kultury, wykorzystanie metod automatycznej analizy języka naturalnego (NLP) w socjologicznych badaniach nad dyskursem.

Adres kontaktowy:

Collegium Civitas
Plac Defilad 1
00-901 Warszawa
e-mail: mtroszynski@civitas.edu.pl

Aleksander Wawer, doktor nauk technicznych w kierunku informatyka, absolwent socjologii i informatyki. Adiunkt w Zespole Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki PAN. Zainteresowania naukowe obejmują wybrane problemy przetwarzania języka naturalnego, w szczególności analizę wydźwięku, ekstrakcję relacji oraz głębokie uczenie maszynowe.

Adres kontaktowy:

Instytut Podstaw Informatyki PAN
ul. Jana Kazimierza 5, 01-248 Warszawa
e-mail: axw@ipipan.waw.pl

Gwałtowny rozwój komunikacji zapośredniczonej przez komputer, a w szczególności tak zwanych mediów społecznościowych, sprawia, że w dziedzinie badań nad treścią przekazów medialnych obserwujemy znaczące rozszerzenie dostępnego obszaru badawczego. W przestrzeni publicznej pojawił się zbiór tekstów pisanych przez tradycyjnie rozumianych „odbiorców” (tzw. *User Generated Content* – UGC). Badacze stanęli wobec istotnego problemu – jak analizować teksty tworzone przez nieprofesjonalistów, które charakteryzują się różnorodnością języka, stylów wypowiedzi, konwencji, wykorzystywanych socjolektów, gwary czy sformułowań potocznych. Co ważniejsze, są to wypowiedzi liczone w setkach tysięcy czy milionach tekstów.

W odpowiedzi na ten problem chcemy przedstawić metodę analizy danych (analizy treści), która, trwając przy jakościowym podejściu do kodowania (przypisywanie kodów w oparciu o znaczenia zawarte w tekście), pozwala jednocześnie na ilościowy opis dużego korpusu różnorodnych tekstów (np. zapisu dyskursu z mediów społecznościowych). Metoda została wypracowana w trakcie projektu badawczego, którego celem było oznaczenie mowy nienawiści wobec mniejszości (narodowych, etnicznych, seksualnych) w polskim Internecie.

We wspomnianym projekcie, chcąc monitorować mowę nienawiści, musieliśmy przygotować narzędzia badawcze, które pozwolą na skuteczne oznaczanie wielkiej liczby tekstów. Ponieważ „mowa nienawiści” jest złożonym konstruktem teoretycznym, a jej reprezentacje językowe mogą przyjmować niezliczone formy, nie sposób przeprowadzić

analiz ilościowych bazujących na frekwencjach i współwystępowaniu poszczególnych słów. Analiza jakościowa zakłada mozolne kodowanie znaczeń (w tym ładunku emocjonalnego), co w przypadku zbioru kilku tysięcy tekstów również każe nam stawiać pytanie o wykonalność tego zadania. Rozwiązaniem jest połączenie jakościowego kodowania z narzędziami udostępnionymi przez lingwistykę komputerową i metody automatycznego przetwarzania języka naturalnego (*Natural Language Processing* – NLP).

NLP a nauki społeczne

Narzędzia komputerowej, jakościowej (również ilościowej) analizy tekstów wykorzystywane są w socjologii od kilku dziesięcioleci. Początki sięgają pionierskich programów komputerowych dla języka angielskiego, korzystających ze słowników (General Inquirer [Stone i in. 1966] czy LIWC [2007]). W języku polskim wymienić można w tym kontekście słowniki wydźwięku stworzone dla języka polskiego (Wawer, Rogozińska, 2012).

Współczesne podejście jest nieco inne. Zamiast tworzyć słowniki, które operacjonalizują pożądane kategorie badawcze, tworzy się raczej kolekcje tekstów z oznaczeniami (kodowaniem) wybranych zjawisk. Mogą to być oznaczenia na poziomie wyrazów, wielowyrzowych fraz, zdań lub większych fragmentów tekstu – odpowiednio do tego, na jakim poziomie leksykalnym i składniowym dane zjawisko występuje. Następnie, wykorzystując owe kolekcje oznaczonych w tekstach problemów, próbuje się stworzyć komputerowe modele tych zjawisk, korzystając z narzędzi uczenia maszynowego (ML)

(Bishop 2007). Skuteczność (przykładowo: precyzja) narzędzi tworzonych w ten sposób jest wyższa niż metod słownikowych. Pozwalają one na uchwycenie nie tylko koncepcji i znaczeń kodowanych na poziomie słów, ale także ciągów słów, fraz, zdań i całych dokumentów. Wszystkie te metody zgrupowane są pod zbiorczą nazwą uczenia maszynowego z nadzorem.

W przypadku większości metod wymagane jest wstępne przetwarzanie tekstu w celu otrzymania odpowiedniej reprezentacji. Liczba kroków i poziom przetworzenia zależą od algorytmu rozpoznającego pożądane zjawisko. Z reguły jest to podział na słowa (tokenizacja), rozpoznawanie części mowy oraz przetwarzanie składniowe, płytkie lub głębokie. Zestawy narzędzi tego typu zawiera pakiet Stanford NLP (Manning i in. 2014) oraz dla języka polskiego – Multiserwis, utrzymywany w Instytucie Podstaw Informatyki PAN (Ogrodniczuk, Lenart 2013).

Analiza na poziomie frazowym przeprowadzana jest zazwyczaj z wykorzystaniem metody warunkowych pól losowych (ang. *Conditional Random Fields*) (Lafferty, McCallum, Pereira 2001), ostatnio także głębokich sieci neuronowych (Lample i in. 2016). W metodach tych jest brana pod uwagę kolejność występowania słów w tekście oraz kontekst zdaniowy. Są one do pewnego stopnia odporne na zmienny szyk składniowy. Metody te są z sukcesem używane w takich zastosowaniach jak rozpoznawanie dłuższych niż jeden wyraz fragmentów tekstu określonego typu, jak przykładowo frazy zawierające wydźwięk (ang. *sentiment*). Istnieje możliwość zastosowania tych metod również do innych problemów związanych z jakościową analizą tek-

stu, zwłaszcza takich, w których kodowane przez badaczy (potencjalnie lingwistów) zjawiska i treści mają charakter wielowyrazowy, silnie zależny od kontekstu. Sprawdzają się tam, gdzie nie jest możliwe zastosowanie prostego podejścia słownikowego, czyli takiego, w którym wszystkie znane z góry wielowyrazowe ciągi wpisane są na stałe do słownika, a każde ich wystąpienie w tekście jest tożsame z wystąpieniem badanego zjawiska.

Do rozpoznawania zjawisk na poziomie zdań oraz większych fragmentów tekstu stosowane są metody klasyfikacji z nadzorem oparte na reprezentacji *bag-of-words*, czyli niebiorącej pod uwagę kolejności słów w tekście, a tylko sam fakt ich wystąpienia. Algorytmy tego typu to przykładowo maszyny wektorów podpierających (ang. *Support Vector Machines*, SVM) (Cortes, Vapnik 1995) lub metody oparte na drzewach lub lasach drzew decyzyjnych (Breiman 2001). Przykładowe zastosowania obejmują klasyfikację tematyczną tekstów (czyli przykładowo stwierdzenie, czy tekst dotyczy sportu czy może polityki), analizy stylometryczne (identyfikacja różnego typu cech psychologicznych lub demograficznych piszącej osoby).

Istnieje wiele obszarów badawczych na gruncie szeroko rozumianych nauk społecznych, w których wykorzystywane są komputerowe narzędzia przetwarzania języka naturalnego. W przypadku języka angielskiego, którego dotyczy większość prowadzonych badań, można mówić o takich generalnych kierunkach jak przewidywanie atrybutów osób piszących teksty (przykładowo ich emocje, płeć, wiek, przekonania polityczne i system wartości) czy przewidywanie zjawisk społecznych na podstawie ze-

branych tekstów (przykładowo wyniki wyborów, epidemie chorób, notowania giełdowe).

Interesującym obszarem zastosowania komputerowego przetwarzania tekstu są badania literackie i stylometryczne. W paradygmacie tym metody komputerowe i algorytmy służą jako narzędzie poznawcze, agregujące i syntetyzujące informacje z dużych zbiorów tekstowych danych literackich. Używane są między innymi informacje statystyczne i frekwencyjne, o współwystępowaniu pojęć oraz określonego typu słów. Analiza taka pozwala na poznanie struktur charakterystycznych dla określonego okresu lub gatunku literackiego i uzyskanie wglądów zupełnie innych niż lektura poszczególnych pozycji. Możliwa jest zwłaszcza eksploracja trendów w czasie lub różnic geograficznych (Jockers 2013). W analizach tego typu wykorzystywane są też wizualizacje współwystąpień słów i sieci pojęć (Moretti 2013).

Wykorzystanie NLP nie zwalnia badaczy społecznych z odpowiedzialności za proces badawczy, w szczególności za jakość kodowania tekstu. Na przykładzie badań nad mową nienawiści chcemy pokazać szerszy problem metodologiczny, który zamyka się w pytaniu – jak można zoptymalizować działania zespołu badawczego (osób odpowiedzialnych za kodowanie tekstów), aby próbować wykorzystać efekty takiego kodowania do „nauczenia” narzędzia (skryptu), które z określoną precyzją oznaczy w ten sposób dowolnie duży zbiór tekstów.

W trakcie trwających od 2012 roku projektów badawczych testowaliśmy różne metody automatyzacji kodowania tekstów zawierających mowę

nienawiści, by oszacować zasięg zjawiska i tematy poszczególnych wypowiedzi (Troszyński 2015). Ostatecznie przyjęliśmy metodę określoną przez nas jako półautomatyczna – przeszkolony zespół koderów/anotatorów¹ kodował fragmenty tekstu (przypisując im określone kategorie, np. poziom negatywnych emocji na skali 0–4), by przejść do budowania algorytmu, który automatyzuje tę czynność, wykorzystując techniki uczenia maszynowego. Efekt działań algorytmów był ponownie weryfikowany przez zespół kodujący.

Przedmiot badania – mowa nienawiści

Kluczowy dla zrozumienia złożoności prezentowanego zagadnienia jest przedmiot badań – mowa nienawiści. Mamy tu konstrukt, który nie dość, że jest wytworem określonej teorii, to dodatkowo jest opisywany w różny sposób w różnych dyskursach kulturowych. Spójrzmy na wybrane definicje tej kategorii pisane z perspektywy prawa, językoznawstwa i socjologii.

Najwięcej uwagi kategorii „mowy nienawiści” poświęca się w obrębie nauk prawnych (Wieruszewski i in. 2010; Siwicki 2011; Bychawska-Siniarska, Głowacka 2013; Heinze 2016). W polskim systemie prawnym nie istnieje powszechnie przyjęta definicja tej kategorii. Jednym z dokumentów obecnych w obiegu prawnym jest raport Biura Studiów i Ekspertyz Kancelarii Sejmu RP, według którego „mowa nienawiści to wypowiedzi ustne i pisemne oraz

¹ Prezentowany tekst łączy perspektywy dwóch dyscyplin – lingwistyki i socjologii. Dlatego będziemy posługiwać się podwójną nazwą kluczowej dla tekstu funkcji w procesie badawczym – osoby, która oznacza teksty, przypisując im odpowiednie kategorie z instrukcji kodowej – anotator/koder.

przedstawienia ikonizujące, oskarżające, wyszydające i poniżające grupy i jednostki z powodów po części od nich niezależnych – takich jak przynależność rasowa, etniczna i religijna, a także płeć, preferencje seksualne, kalectwo (...). Jest to upubliczniona przemoc werbalna, wyraz nienawiści kolektywnej (...)" (Łodziński 2003: 5).

Próby podsumowania definicji mowy nienawiści z perspektywy systemu prawnego podjęły się Dominika Bychawska-Siniarska i Aleksandra Gliszczyńska-Grabias: „Mowa nienawiści, w potocznym rozumieniu, to słowa, wypowiedzi czy publikacje wyrażające **skrajnie negatywne, nienawistne uczucia i poglądy wobec określonych jednostek lub grup** [wyróżnienie własne]. Najczęściej wypowiedzi takie kojarzymy z rasizmem, ksenofobią, nienawiścią wobec mniejszości seksualnych lub nienawiścią wobec mniejszości religijnych” (2016: 4). Co ważniejsze, mowa nienawiści rozpowszechniana w Internecie nosi „pewne znamiona, których nie posiadają wypowiedzi w tzw. realu, takie jak trwałość, długoterminowość, powtarzalność, anonimowość oraz transgraniczny charakter” (Bychawska-Siniarska, Gliszczyńska-Grabias 2016: 5).

Pojęcie mowy nienawiści jest opisywane w perspektywie językoznawczej. Warto zwrócić uwagę na koncepcję Jadwigi Linde-Usiekiewicz (2015), która do opisu mowy nienawiści wykorzystuje teorię relewancji Sperbera i Wilsona (2011). Efektem jest definicja: „Mową nienawiści jest taka wypowiedź, dla której optymalnego przetwarzania niezbędne są przesłanki (czyli przekonania lub reprezentacje przekonań) dotyczące tego, że jakaś **grupa**, w tej wypowiedzi przywołana i zdefiniowana

przez cechy uznawane za stałe, **jest gorsza pod jakimś względem, a przez to zasługuje na gorsze traktowanie** [wyróżnienie własne], i że osoba lub osoby, do których ta wypowiedź się odnosi, jest członkiem tej grupy” (Linde-Usiekiewicz 2015). Podobnie jak w przypadku refleksji prawniczej tak i w tej perspektywie podnoszona jest kwestia przeniesienia cech grupy na przynależące do niej jednostki.

I wreszcie perspektywa nauk społecznych. Temat ten podejmuje Lech Nijakowski, który tak określa to zjawisko: „mowa nienawiści polega na **przypisywaniu szczególnie negatywnych cech i/lub wzywaniu do dyskryminujących działań** [wyróżnienie własne], wymierzonych w pewną kategorię społeczną, przede wszystkim taką, do której przynależność jest postrzegana jako «naturalna» (przypisana), a nie z wyboru” (2008: 132). Podobnie jak w przypadku cytowanych wyżej definicji główny nacisk położony jest na przynależność do grupy oraz „przypisywanie negatywnych cech”.

Próbując uwspólnić powyższe definicje, przyjęliśmy rozstrzygnięcia, które pozwoliły dookreślić przyjęty obszar badawczy:

1. wyodrębniliśmy dwie kategorie w obszarze szeroko pojmowanej mowy nienawiści: „mowę nienawiści” oraz „język wrogości”. Utworzenie szerszej kategorii „języka wrogości” uznaliśmy za konieczne, by wyjść poza prawnokarne konotacje „mowy nienawiści”,
2. przedmiotem analiz uczynimy mowę nienawiści skierowaną przeciw mniejszościom i człon-

kom mniejszości (przynależność do określonej grupy). Oznacza to, że w badaniu pominęliśmy mowę nienawiści wobec polityków, grup większościowych, osób prywatnych.

Konceptualizując powyższe kategorie, przyjęliśmy, że **mowa nienawiści (w wąskim sensie)** to treści zawierające jawne wezwanie do działania oraz skrajnie negatywne emocje (wzmocnione często przez wulgaryzmy) skierowane przeciwko określonej grupie społecznej. W badaniu posługujemy się również pojęciem **język wrogości, który jest kategorią szerszą niż mowa nienawiści**. Określamy w ten sposób wszelkie treści odnoszące się negatywnie, nieprzychylnie lub odwołujące się do negatywnych emocji wobec określonych mniejszości. Mogą to być zarówno wypowiedzi prawdziwe, jak i fałszywe. Kluczowym elementem jest wykorzystanie w tekście mechanizmu stereotypu, który działa poprzez przypisywanie określonych cech i zachowań zbiorowości wszystkim jej członkom.

Pierwsze próby kodowania automatycznego – podejście słownikowe

Pierwszym zadaniem badawczym była budowa korpusu tekstów, które zostaną poddane analizie. W tym celu zostały zrealizowane następujące działania:

- wybraliśmy (w oparciu o dane o liczbie realnych użytkowników) portale horyzontalne, na których zostały wyodrębnione fora internetowe oraz komentarze, jako źródła tekstów do badania (gazeta.pl, wp.pl, onet.pl oraz interia.pl),

- wytypowaliśmy słowa kluczowe – określenia wybranych mniejszości (badanie dotyczyło mniejszości: muzułmańskiej, żydowskiej, LGBT, czeczeńskiej, romskiej, niemieckiej, rosyjskiej, ukraińskiej), w tym również określenia uznawane za obraźliwe.

Następnym krokiem było napisanie crawlera – narzędzia, które skanowało treści postów internetowych na wskazanych portalach i kopiowało do bazy danych te, które zawierały przynajmniej jedną nazwę mniejszości. Zebrane teksty oczyszczono poprzez usunięcie znaczników html i korektę pisowni. W tak opracowanych treściach zidentyfikowano fragmenty zawierające słowa klucze (określenia poszczególnych mniejszości) – zazwyczaj były to zdania lub mniejsze fragmenty wypowiedzi. Te fragmenty poddano dalszemu przetwarzaniu.

Przygotowaliśmy 2 korpusy tekstów²:

- 1500 tekstów pochodzących z postów zamieszczanych na forach internetowych, wybranych arbitralnie przez zespół kodujących, uwzględniając różne poziomy natężenia emocji wobec mniejszości – korpus ten wykorzystywany był jako korpus treningowy,
- 11 176 tekstów zebranych automatycznie przez crawler z forów internetowych; jego oznaczenie było głównym zadaniem dla ze-

² Przedmiotem analizy były „teksty” zbierane w sposób automatyczny – w pierwszej wersji przez dedykowany crawler, w kolejnych wersjach wykorzystując komercyjne narzędzia. Różne były metody określania długości relewantnego tekstu (odległość od słów kluczowych, spójność gramatyczna). Efektem były wypowiedzi zbliżone w formie i długości do zdania.

społu koderów/anotatorów biorącego udział w projekcie.

Oba korpusy bazowały na tekstach zamieszczonych w sieci w 2012 roku.

Pierwsza wersja narzędzia, którego celem było wykrywanie wśród postów języka wrogości, stworzona była na bazie hipotezy mówiącej, że język wrogości jest w istocie podzbiorem języka negatywnie nacechowanego (ang. *negative sentiment*). Aby zbudować to narzędzie, skorzystaliśmy z dostępnej wówczas wersji słownika wyrażen negatywnie nacechowanych, stworzonego w Instytucie Podstaw Informatyki PAN (<http://zil.ipipan.waw.pl/SlownikWydzwieku>). Słownik ten został następnie zmodyfikowany w taki sposób, aby zawęzić go do podzbioru słów (w dużej części były to przymiotniki) i wyrażen, który jest nacechowany zarówno negatywnymi, jak i wrogimi emocjami. Dzięki analizom korpusu okazało się, że konieczne jest poszerzenie tego zbioru negatywnych słów o słowa i wyrażenia w nim niewystępujące, a charakterystyczne dla języka polskiego Internetu. Dążyliśmy do tego, aby ten zbiór skonstruowany był w taki sposób, aby wystąpienie przynajmniej jednego elementu w danym tekście (poście) pozwalało na skategoryzowanie go jako zawierającego język wrogości.

Dla stworzenia narzędzia, które pozwoli na oznaczanie mowy nienawiści, zaczęliśmy od spisania wyczerpującego zbioru słów (ze względu na charakter szukanej treści są to przeważnie czasowniki oraz wszelkie formy wulgaryzmów), wyrażen i reguł gramatycznych związanych z nimi, nawołujących do podjęcia działań wymierzonych przeciwko

mniejszościom. Każdy z nich był opisywany przez analityka jako zawierający (1) lub niezawierający (0) wezwanie do działania. Następnie po analizie postów oznaczonych jako „1” wypisano słowa i wyrażenia, na podstawie których koderzy/anotatorzy kwalifikowali dany wpis jako mowę nienawiści. Po zebraniu zbioru słów i wyrażen członkowie zespołu badawczego analizowali formy gramatyczne, w jakich występują poszczególne części wypowiedzi w postach uznanych wcześniej za zawierające wezwanie do działania. Zauważono, że o ile wyrażenia i słowa różne od czasowników występują zazwyczaj w stałych formach (np. „do gazu”, „won”, „wynocha”, „Polska dla Polaków”), o tyle czasowniki przyjmują postać bezokoliczników (np. „zabić”, „wykastrować”) lub trybu rozkazującego (np. „wybijmy”, „odizolujcie”). Wyrażenia wielowyrazowe (pojedyncze słowa są dostępne w zbiorze słów) oraz zdefiniowane gramatyczne formy czasowników zapisano w zbiorze reguł. Dzięki temu w ściśle zdefiniowanych sytuacjach syntaktycznych automatyczny analizator doda bądź usunie znacznik wezwania do działania z określonego fragmentu postu. Ponadto do mowy nienawiści włączane są posty zawierające wulgaryzmy.

Szukając sposobów na odejście od subiektywnych przekonań dotyczących występowania w poszczególnych tekstach negatywnych emocji (związanych z treściami obraźliwymi dla określonych mniejszości), przeprowadziliśmy sondaż, którego celem była ocena wybranych wypowiedzi przez większą grupę respondentów. Badanie było przeprowadzone przez Interaktywny Instytut Badań Rynkowych na próbie reprezentatywnej dla dorosłych internautów (N=800). Respondenci oceniali wyświetlane

fragmenty tekstów (po 30 fragmentów dla każdego respondenta) pod kątem ich obraźliwości, wykorzystując skalę 0–4 (brak obraźliwości, bardzo obraźliwe)³ (por. Troszyński 2015: 207). Analiza wyników badania miała pozwolić nam na intersubiektywizację oceny natężenia negatywnych emocji. Jednak zebrane dane wskazały na bardzo dużą rozbieżność w ocenianiu stopnia obraźliwości poszczególnych wypowiedzi. Pokazuje to, że celem tego typu badania powinno być raczej opisanie zmiennych (cech społeczno-demograficznych, przekonań politycznych itp.), które wpływają na „wrażliwość” na mowę nienawiści. Dzięki temu wyraźnie widzimy czynniki, które wpływają na postrzeganie mowy nienawiści, a co za tym idzie – na interpretację tekstu przez kodujących. W naszym projekcie zdecydowaliśmy się na stosowanie maksymalnie szerokiego rozumienia mowy nienawiści / języka wrogości poprzez bezpośrednie wpisanie do instrukcji kodowej (patrz niżej) postulatu podmiany przedmiotu wypowiedzi na własną grupę (np. „Żydzi won z tego kraju” koder/anotator narodowości polskiej oceniał jak „Polacy won z tego kraju”). Pozwoliło to na częściowe przynajmniej wyrównanie wśród koderów/anotatorów poziomu oceny obraźliwości tekstu.

Pierwsze próby polegające na przybliżeniu wyników automatycznej analizy treści do wyników wzmiankowanego badania (czyli rozkładu odpowiedzi w korpusie 1500 postów) zakończone zo-

³ Wyświetlanie kolejnych fragmentów tekstów poprzedzone było poleceniem: „Przedstawione teraz zostanie P. kilkadziesiąt autentycznych wypowiedzi polskich internautów. Proszę wskazać, jak bardzo P. zdaniem są one obraźliwe w stosunku do osób, których dotyczą. Odpowiadając, proszę posłużyć się skalą od 0 do 4, gdzie 0 oznacza zdanie niezawierające treści obraźliwych lub neutralne, a 4 oznacza zdanie bardzo obraźliwe. Pozostałe cyfry służą do wyrażenia opinii pośrednich. Odpowiadając, proszę posłużyć się własną oceną”.

stały niepowodzeniem. Jakość przewidywań (czyli kategoryzacji postów jako zawierających język nienawiści) uzyskana tą metodą była niska w stopniu nierokującym na poprawę.

Problem ten, jak się okazało, ma naturę bardziej skomplikowaną niż proste słownikowe podejście, opisane powyżej. Okazało się, że kluczowe jest uchwycenie współwystępowania wielu słów i sensów, a co za tym idzie – wypisanie ich w formie słownika „słów, które uznajemy za wrogię” nie rozwiązuje problemu. Co więcej, wrogość w znaczeniu pewnych słów pojawić się może tylko w wybranych kontekstach leksykalnych, których rozpoznawanie wymaga analizy całego dostępnego tekstu (dyskusji na forum internetowym).

Biorąc pod uwagę wszystkie te doświadczenia, zdecydowaliśmy się na pracę nad wyższym poziomem przeszkolenia zespołu koderów/anotatorów oraz oznaczanie nie tylko w obrębie niewielkiego (1500 tekstów) korpusu treningowego. Korpus, na którym pracowaliśmy, w kolejnym etapie liczył już 11 176 tekstów, z których każdy kategoryzowany był przez jedną osobę. W miejsce rozwijania list słów i wyrazów w słownikach przyjęliśmy inną metodę, polegającą na wykorzystaniu metod uczenia maszynowego.

Budowa instrukcji kodowej

Punktem wyjścia do tej części badania było przekonanie (wynikające z praktyki badawczej), że w procesie badawczym, który uwzględnia metodologię jakościową, jakością kodowania zebranego materiału jest jednym z kluczowych czynników wpływających na cały proces analizy danych. Aby wykorzystać narzę-

dzia NLP bazujące na uczeniu maszynowym, proces kodowania musi być tak bardzo ujednoznaczony jak to tylko (w określonej sytuacji badawczej, warunkowanej harmonogramem i budżetem) możliwe. Rozumiemy przez to pozostawienie jak najmniejszego obszaru na swobodne decyzje koderów/anotatorów, a w konsekwencji nacisk na budowanie jak największej zgodności pomiędzy nimi.

Zgodność między koderami/anotatorami zależy od jakości instrukcji dla anotujących (im bardziej szczegółowe i precyzyjne, tym wyższa zgodność) oraz złożoności semantycznej i syntaktycznej tekstu. Jednak zazwyczaj instrukcje nie są wystarczające: w praktyce okazuje się, że istnieje wiele przypadków brzegowych, noszących cechy kilku kategorii, potencjalnie możliwych do zaklasyfikowania na różne sposoby. Rozwiązaniem tego problemu jest opracowanie (i wspólne szczegółowe omówienie) określonej liczby przypadków tego typu, zidentyfikowanych jako problematyczne, a także przyjęcie wzorcowych rozstrzygnięć, nawet arbitralnych. Dlatego rozwiązaniem jest trening koderów/anotatorów, czyli realizacja w obrębie zespołu badaczy kolejnych zadań polegających na oznaczaniu takich samych partii tekstu i porównywaniu wyników pracy poszczególnych analityków. Istotą treningu nie jest mierzenie zgodności, ale dyskusja i „flumaczenie się” kodujących z podjętych decyzji. Tylko takie postępowanie może doprowadzić do uwspólnienia rozumienia instrukcji kodowej. W lingwistyce powszechnie przyjętym sposobem rozstrzygania niezgodności jest schemat kodowania, w którym oznaczenia (anotacje) wykonywane są przez dwie osoby o podobnych kompetencjach, a konflikty między nimi rozstrzygane są przez trzecią osobę o najwyższych umiejętnościach

i wiedzy. W realizowanym przez nas treningu zadaniem stawianym przed analitykami było osiągnięcie konsensusu w trakcie dyskusji. Celem było zastąpienie arbitralnej decyzji jednego z badaczy merytoryczną argumentacją na rzecz przyjęcia jednego z rozwiązań (przypisania danego kodu).

W analizie treści podstawowe pytanie brzmi: czy badacze posługują się tym samym zestawem pojęć (kategorii języka naturalnego)? Brak zgodności w kodowaniu wskazuje na różnice w rozumieniu pojęć. W przypadku lingwistyki i problemów składniowych niezgodność wynika z nieostrych granic między opisywanymi zjawiskami składniowymi lub braku kompetencji koderów/anotatorów. W przypadku semantyki (również w analizie treści w naukach społecznych) brak zgodności wiąże się także z różnicami w wyznawanych wartościach. Konieczne jest zatem „odkrycie” tych wartości, jasne pokazanie założeń, na których opiera się proces interpretacji.

W kontekście uczenia maszynowego i aplikacji przetwarzających teksty kluczowy jest aspekt rozumienia samego problemu: zgodność między kodującymi jako „górną granicą”, którą mogą osiągnąć narzędzia automatyczne. Inaczej rozumiemy sukces algorytmów maszynowych, jeśli zgodność między wytrenowanymi anotatorami nie przekracza 0.6 (np. kappa), a inaczej, gdy osiąga 0.95. Zgodność pokazuje, jak bardzo skomplikowane jest badane zjawisko. Z drugiej strony negatywnym punktem odniesienia jest czysto losowy wybór anotacji (oznaczeń). Od jakości anotacji wykonanych przez zespół koderów/anotatorów zależy także zdolność algorytmów do „wyuczenia się” charakterystyki zjawiska. Spójne i konsekwentne oznaczanie bada-

nych znaczeń w tekstach jest kluczowe: algorytmy uczące nie są odporne na sprzeczności lub nieciągłości w danych treningowych.

Aby osiągnąć jak największą zgodność kodujących, a co za tym idzie – kodowanie automatyczne o wysokiej precyzji, należy zrealizować dwa kroki procedury:

1. stworzyć wyczerpujący i kompletny klucz kodowy (opisujący znaczenia i sposoby użycia poszczególnych kodów), przedyskutować znaczenie poszczególnych kategorii z członkami zespołu badawczego.
2. przeprowadzić wieloetapowe szkolenie członków zespołu ankierskiego, na każdym etapie porównując rezultaty kodowania, i co ważniejsze – wymusić „obronę” danego sposobu kodowania poprzez werbalizację zasad językowych, które pozwoliły koderowi/anotatorowi zastosować określony kod.

Poniżej opisujemy stosowane w omawianym projekcie kategorie kodowe i ich znaczenia. Część tych kategorii to kategorie „techniczne”, przygotowane na potrzeby automatyzacji procesu kodowania, zatem ich omówienie ma na celu zaprezentowanie specyfiki omawianego tu procesu. Te właśnie elementy odróżniają nasz klucz kodowy od typowego narzędzia jakościowego. Trudność polegała na konieczności całkowitej algorytmizacji działań koderów. Nie tylko powinni oni kodować „tak samo” (z jak największą zgodnością), ale również powinni być w stanie opisać mechanizmy, które skutkują przypisaniem konkretnego kodu do wypowiedzi. To jest kluczowy czynnik – wyjście poza implicite przyjmowane zasa-

dy nadawania sensów bazujące na kompetencji językowej kodujących. Jeśli chcemy skutecznie wyuczyć automat przyjętych przez nas zasad kodowania, konieczna jest ich werbalizacja i algorytmizacja.

Zestawienie użytych kodów:

1. Sensowność (oznaczana jako: S1 lub S0). Teksty, które zostały zakwalifikowane do badania, zostały zapisane w bazie w sposób automatyczny (jako wynik działania skryptu). Co oznacza, że pewna część z nich to fragmenty, których nie można sensownie zinterpretować. Są to wycinki wypowiedzi, pojedyncze litery lub słowa, fragmenty, którym kodujący nie potrafi przypisać jednoznacznego sensu. Zmienną tę kodowaliśmy zero-jedynkowo (1, 0). Była to zmienna, która pozwalała na odfiltrowanie nierelevantnych tekstów. Dzięki temu mogliśmy znacząco zmniejszyć poziom szumów w analizowanych wypowiedziach.
2. Temat wypowiedzi (TW_nazwa). Z perspektywy celów badawczych to jedna z najważniejszych zmiennych. Pozwoli nam oszacować, o czym „mówią” zebrane teksty. Przyjęliśmy zestawienie tematów wzorowane na General Inquirer, odwołujące się do dużych systemów społecznych: edukacja, nauka, sztuka, ekonomia, sport, prawo, militaria, polityka, społeczeństwo, religia (por. Stone i in. 1966).
3. Mniejszość (M_nazwa): przypisanie tekstu do określonej mniejszości. Poza nazwami mniejszości konieczne było zakodowanie całej gamy określeń wskazujących na przynależność do danej grupy, zarówno tych neutralnych, jak i negatywnych.

Przypadki, gdy sama nazwa mniejszości może być użyta w negatywnym sensie, znacząco ułatwiały właściwe kodowanie sensów (np. „czarnuch”).

4. Autoopis (A0 lub A1): określenie, czy autor wypowiedzi mówi sam o sobie. Zmienna ważna dla przypisania właściwego sensu do wypowiedzi.

5. Negatywne emocje (Neg_wartość): określenie poziomu występowania w tekście negatywnych emocji wobec mniejszości. Przyjęliśmy skalę 0–4, gdzie: 0 oznacza brak negatywnych emocji (treść jest albo neutralna, albo pozytywna wobec mniejszości); 1 – wartość pośrednia, 2 – oznacza obecność negatywnych emocji, ale nie są one bardzo silne; 3 – wartość pośrednia; 4 – oznacza obecność skrajnie negatywnych emocji (często występuje razem z wulgaryzmami).

6. Pozytywne emocje (Poz_wartość) – analogicznie kodowaliśmy pozytywne emocje w tekście.

7. Wezwanie do działania negatywnego (W_neg): nawoływanie innych do podjęcia negatywnie nacechowanych działań wobec danej mniejszości.

8. Wezwanie do działania pozytywnego (W_poz): analogicznie, jak powyżej, ale działania o pozytywnych skutkach.

9. Ironia/sarkazm (IS): określenie, czy w danej wypowiedzi znajdują się elementy ironiczne bądź sarkastyczne.

10. Typ ramki (Typ_nazwa): określenie źródła naszej wiedzy o tym, czy fragment zawiera negatywne

emocje wobec mniejszości, gdzie: „Leksykalna” oznacza, że wiemy to ze słów użytych w treści (możemy wskazać słowo/słowa, które jednoznacznie odsyłają do negatywnych emocji, np. debile); „Kontekstowa” – oznacza, że „znak” (rodzaj emocji/nastawienia) rozpoznajemy nie z konkretnych słów, a z szerszego kontekstu, zrozumienia wypowiedzi. W pierwszej kolejności sprawdzamy, czy są słowa lub frazy, które są obraźliwe. Jeśli istnieją takie, to oznaczamy ramkę jako „leksykalna” (i kończymy tu kodowanie), jeśli nie – jako „kontekstowa”.

11. Kontekstowość (Kon_nazwa): jeśli typ ramki zostanie określony jako kontekstowy, to określamy źródło kontekstu; „Wewnętrzna” oznacza, że kontekst znajduje się w treści analizowanego fragmentu (np. „bo skończycie tak jak oni” [a wcześniej jest napisane, jak skończyli oni]); „Zewnętrzna” – znajomość kontekstu wynika z wiedzy, którą mamy spoza tekstu (np. „(..) bo skończy się jak w ‘68”).

Przyjęliśmy, że aby zakwalifikować tekst jako mowę nienawiści, to musiał on być opisany kodami „Neg_4” oraz „W_neg”. Teksty kwalifikowane jako język wrogości to te oznaczone kodami „Neg_1”, „Neg_2”, „Neg_3”, co kwalifikowało nam znacznie większy zbiór tekstów do dalszych analiz. Pozostałe oznaczenia kodowe pełniły funkcję zmiennych niezależnych w badaniu, co pozwalało na pogłębienie analizy zjawiska.

Poniżej przedstawiamy przykłady kodowania z wykorzystaniem tak zbudowanej instrukcji kodowej. Są to wybrane rekordy z bazy danych, które przywołujemy w tym miejscu tekstu, by stanowiły ilustrację naszych rozważań.

Tabela 1.

ID	ID mniejsz	fragment do zakodowania	kody
5243	51	długość ich jest krótsza o 20 lat od długości przeciętnego Kanadyjczyka pederasta i lesbijki częściej popełniają samobójstwa 14 razy częściej niż pozostali przedstawiciele populacji trzykrotnie częściej nadużywają nikotyny częściej popadają w alkoholizm	S1; TW_zdrowie; M_homoseks; A0, neg2; poz0; tr_leks
7439	71	W omawianej publicystyce spotkać się można ze stwierdzeniem aczkolwiek chyba nie powtarzającym tak często i z takim naciskiem jak wspomniane poprzednio Hitler doszedł w roku 1933 do władzy przede wszystkim dlatego zakulisowo wsparli go Żydzi bez ich aproba	S1; tw_historia; m_żydzi; a0; neg1, tr_konteks, kon_wewn
7026	72	polacy są wybitnie inteligentnym narodem potrafią w tak ekstremalnych warunkach jakie są w Polsce bo Niemiec jakby do nas przyjechał i miał z tego co zarabia i opłacić mieszkanie samochód nowy z salonu kupiony na raty to by się zes akał w majty	S1, tw_ekonomia; m_niemcy; a0; neg1; tr_konteks; kon_zewn
7472	71	Można zatem głosić Żydzi skazali się sami na to co ich spotkało bo są pozbawieni wszelkiej społecznej solidarności utrzymując zarazem stanowią jedną wielką siłę solidarnie dbającą wyłącznie o swoje własne interesy i działającą na niekorzyść innych przed	S1; tw_obyczaje; m_żydzi; a0; neg3; tr_leks

Źródło: badania własne.

Dla zrozumienia koncepcji „ramki kontekstowej” warto zwrócić uwagę na 2 teksty:

- tekst nr 7026: jest to wypowiedź o Niemcach, zaklasyfikowana jako negatywna (1). Ale sens tej negatywności zbudowany jest na przekonaniu o powszechnej znajomości faktów dotyczących różnic w sytuacji ekonomicznej Polaków i Niemców. To oznacza, że nie możemy uzasadnić tego kodu (negatywności) poprzez wskazanie konkretnego słowa lub grupy słów;

- tekst nr 7439: wypowiedź o Żydach, zaklasyfikowana jako negatywna (1). Jej sens jest jednoznaczny pod warunkiem, że rozpoznajemy postać historyczną Adolfa Hitlera wraz z wszystkimi jej konotacjami.

Trening zespołu kodującego

Na potrzeby projektu zebraliśmy czteroosobowy zespół koderów/anotatorów, w którym uczestniczyli studenci socjologii zainteresowani analizą treści.

Pierwszym etapem szkolenia było szczegółowe omówienie poszczególnych kategorii kodowych. Po takim wprowadzeniu zespół rozpoczął trwający 4 tygodnie trening. Polegał on na cyklicznych spotkaniach, w trakcie których omawialiśmy wyniki ich pracy, szczególną uwagę poświęcając fragmentom tekstów, które były niejednoznaczne dla koderów. Każde takie spotkanie kończyło się przydzieleniem nowego zbioru tekstów do kodowania. W czasie treningu anotatorzy posługiwali się przygotowanym na użytek tego procesu arkuszem MS Excel (z wpisanymi w poszczególne wiersze fragmentami tekstu i predefiniowanymi kategoriami kodowymi w kolejnych kolumnach).

Postępy na tym etapie kodowania mierzyliśmy na oznaczonym równoległe przez wielu kodujących zbiorze 1500 tekstów miarą zgodności – alfą Krippendorffa (Krippendorf 1980). Zbiorem, na którym przeprowadzaliśmy obliczenia, były oznaczone fragmenty tekstów, a wartościami zmiennych – etykiety oznaczonych zjawisk (kodów). Na początku procesu niewytrenowany zespół koderów osiągał dla wszystkich anotowanych kodów zgodność na poziomie $\alpha < 0,2$. Efektem treningów było zbliżenie się zespołu na większości kodów do zgodności na poziomie $\alpha = 0,6$. Taki poziom uznaliśmy za wystarczający, by zakończyć szkolenie i przejść do kodowania całości materiału. Rzecz jasna trening należałoby kontynuować, by dążyć do osiągnięcia jeszcze większej zgodności, ale nie pozwolił nam na to harmonogram prac. Wartość $\alpha = 0,6$ znajduje swe uzasadnienie w literaturze przedmiotu: „Za Lombardem i wsp. (2004) można przyjąć, iż zmienna została rzetelnie stworzona, jeśli wartość wskaźnika przekracza poziom (...) 0,5” (Krejtz, Krejtz 2005: 249).

Dla usprawnienia właściwego kodowania (przygotowania zbioru tekstów jako próby uczącej do kodowania maszynowego) zostało zbudowane narzędzie online pozwalające, po zalogowaniu się, na przypisywanie kodów do fragmentów tekstów i bezpośrednie zapisywanie całości w bazie danych. Narzędzie charakteryzowało się maksymalnie uproszczonym interfejsem, w którym koderzy/anotatorzy, korzystając z list rozwijanych, oznaczali kolejne fragmenty tekstu. Efektem pracy zespołu było ręczne zakodowanie około 12 000 fragmentów tekstów. Interfejs umożliwiał zarządzanie tekstami, szybkie odnajdywanie nieoznaczonych jeszcze fragmentów, edycję wcześniejszych zmian.

Próby automatyzacji kodowania

W tej części opisujemy wyniki eksperymentów, w których do procesu automatycznego kodowania tekstów stosujemy metody komputerowe, określane jako algorytmy uczenia z nadzorem (ang. *supervised learning*). W przypadku tej klasy metod decyzja na temat kodowania tekstu podejmowana jest przez model uczenia maszynowego (ML), wytrenowany uprzednio na podstawie zbioru przykładów, dzięki którym został on nauczony interesującego nas zjawiska. Algorytm testowany przez nas w niniejszym eksperymencie to maszyny wektorów podpierających (SVM), jedna z lepszych i bardziej uznanych metod stosowanych między innymi do automatycznej klasyfikacji tematyki tekstów.

W eksperymentach korzystaliśmy z pakietu oprogramowania ML o nazwie scikit-learn (Pedregosa i in. 2011). Oferuje on nie tylko możliwość elastycznego trenowania i ewaluacji modeli ML, ale także

wykonywania innych czynności, takich jak przygotowanie przestrzeni cech (ang. *feature space*), selekcji cech (ang. *feature selection*). Pakiet scikit-learn zawiera między innymi oba algorytmy klasyfikacyjne wspomniane w poprzedniej części artykułu, czyli zarówno maszyny wektorów podpierających (SVM), jak i lasy losowe. W przeprowadzonych eksperymentach korzystaliśmy (za pośrednictwem scikit-learn) z implementacji SVM opartej o bibliotekę liblinear, korzystającą z liniowej funkcji jądrowej (ang. *linear kernel*). Według dostępnej literatury oraz zgodnie z doświadczeniem autorów niniejszego artykułu jest to optymalna funkcja jądrowa dla klasyfikacji danych tekstowych. Głównym powodem jest jej wydajność obliczeniowa przy pracy z rzadkimi macierzami i wysokowymiarowymi danymi. Przestrzeń cech dla danych tekstowych konstruowana jest poprzez wektoryzację (ang. *vectorization*) tekstów, czyli przypisanie poszczególnym słowom tekstu liczbowych identyfikatorów, wykorzystywanych po znormalizowaniu i konwersji TF-IDF jako cechy wejściowe (ang. *input features*) modeli klasyfikujących ML. Reprezentacja tego typu nosi nazwę przestrzeni wektorowej (ang. *vector space*) i efektem jej zastosowania jest właśnie rzadka, wysokowymiarowa macierz cech, której przetwarzanie algorytmem SVM możliwe jest w praktyce tylko za pomocą liniowych funkcji jądrowych. Warto także nadmienić, że korzystaliśmy z domyślnych wartości parametrów wywołania metody SVM pakietu scikit-learn (w szczególności domyślnej wartości stałej C oraz domyślnej regularyzacji).

Algorytmy klasyfikujące działają w dwóch „trybach”: uczącym i klasyfikującym. W tym pierwszym trybie budowane są modele interesujących

nas zjawisk. Do tego celu konieczne jest „pokazanie” algorytmom dwóch zbiorów przykładowych tekstów, przykładowo obraźliwych i nieobraźliwych (neutralnych). Na podstawie wystąpień słów i kombinacji słów w tekstach lub fragmentach tekstów algorytmy „uczą się” dystynktywnych cech obraźliwego języka. Dzięki pewnym własnościom języka oraz samych algorytmów możemy mówić o zdolnościach generalizacyjnych modeli:

- nieco uproszczonym przykładem takiego generalizującego mechanizmu językowego jest synonimia: wystarczy „pokazać” uczącym się modelom słowo drań, a dzięki synonimii również jest uwzględniane: kreatura, bydlak, łobuz i tak dalej,
- generalizujące własności algorytmów dobrze opisuje w tym przypadku intuicja podobieństwa, mająca swoje matematyczne sformułowanie. Dzięki niemu możemy uznawać za obraźliwe teksty tylko częściowo przypominające, nawet w odległy sposób, teksty już widziane. W trybie uczenia algorytm ustala między innymi próg podobieństwa, czyli jak bardzo „odległe” teksty (względem znanych) można nadal uznawać za obraźliwe i w jakich kontekstach.

W drugim trybie, klasyfikującym, zbudowane uprzednio modele działają analogicznie do anotatorów – ludzi: na podstawie analizy tekstów (nawet niewidzianych w trybie uczącym) podejmują decyzję o ich obraźliwości. Modele te mogą skutecznie rozpoznawać teksty *trochę* podobne do tych oglądanych w trybie uczącym.

W obszarze kodowania tekstów można powiedzieć, że model ML analizuje (tryb uczenia) oraz próbuje powielać (tryb oznaczania lub tagowania) decyzje, które analityk podejmował podczas procesu kodowania. ML w decyzjach tych identyfikuje to, co powtarzalne. Stosuje się tu zatem kryterium frekwencyjne. Celem tego procesu jest zbudowanie modelu opartego na zidentyfikowanych kryteriach decyzyjnych analityka. Jakość korpusu tekstów oznaczonych przez analityków, związana z metodologią procesu kodowania tekstów, jest ściśle związana z jakością wyników uzyskiwanych później przez modele uczenia maszynowego.

Typowymi miarami oceny jakości automatycznego kodowania są precyzja (ang. *precision*) oraz pełność (ang. *recall*). Precyzja związana jest z odsetkiem błędów popełnianych przez model na skutek błędnej identyfikacji fragmentu tekstu jako przynależącego do danej klasy. W naszym przypadku precyzja mówi o tym, jak dużo tekstów zostało poprawnie rozpoznanych przez nasz model (np. jako obraźliwe). Pełność natomiast opisuje, jak wiele tekstów z ich całkowitej liczby (np. tekstów obraźliwych) zostało oznaczonych przez model.

Jakość modeli ML trenowanych na pierwszej wersji korpusu, stworzonej na podstawie początkowej, niezbyt rozbudowanej wersji instrukcji kodowania oraz bez przeprowadzenia intensywnych treningów dla anotujących, nie była zadowalająca. Nieznacznie tylko przekraczały one czysto losowy wybór. Dobrym przykładem ilustrującym to zjawisko były eksperymenty z modelem uczenia maszynowego przewidującym, czy dany fragment tekstu zawiera treści obraźliwe. Mamy tu do czynienia z

przewidywaniem 2-wartościowej zmiennej o równej liczebności obu klas (zbiór tekstów, 1500 oznaczonych fragmentów – zazwyczaj zdań – zawierał dokładnie po połowie tekstów obraźliwych i nieobraźliwych, co przy losowym wyborze, np. rzucając monetą, daje precyzję 0.5). Wyniki, jakie uzyskaliśmy, można opisać następująco:

- Algorytmy trenowane na tekstach oznaczonych przez niewytrenowanych anotatorów: precyzja równa 0,57.
- Algorytmy trenowane na tekstach oznaczonych przez wytrenowanych anotatorów: precyzja równa 0,80.

Jak widać, poprawa instrukcji kodowania oraz przeprowadzenie serii treningów anotujących osób znalazło odzwierciedlenie w poprawie jakości kodowania automatycznego o 23 pp, mierzonej liczbą błędów popełnianych przez model ML. Wzrost tego typu niezmiernie rzadko można uzyskać za pomocą metod informatycznych, takich jak na przykład zastosowanie lepszego algorytmu.

Innym pytaniem, jakie należy postawić w tym miejscu, jest kwestia przyczyn błędów popełnianych przez modele uczenia maszynowego przewidujących obraźliwość tekstów, stworzonych na korpusach oznaczonych przez wytrenowanych anotatorów. Przeprowadzona przez nas analiza błędnie rozpoznanych przypadków (precyzja, jak pamiętamy, 0.8) pokazała, że najczęstsze źródła błędów związane są z następującymi powodami:

- Obraźliwość zakodowana na poziomie leksy-

kalnym, ale przy wykorzystaniu nieznanymi dotychczas słów. Błąd ten polegał na pojawieniu się słów i fraz „nieoglądanych” przez algorytm w fazie treningowej. Można zatem założyć, że liczebność tego typu błędów można w pewnym stopniu ograniczyć, zwiększając rozmiar danych treningowych.

- Obraźliwość zakodowana na poziomie ponadleksykalnym:
 - Język figuratywny (metafory), niekompozycyjność znaczeń (znaczenie frazy wykracza poza sumę znaczeń słów). Rozpoznawanie struktur tego typu jest dość trudnym zadaniem, od niedawna jednak podejmowanym z umiarkowanym sukcesem przez obiecujący nurt badań związanych z semantyką dystrybucyjno-kompozycyjną (Gutierrez i in. 2016). Zaadresowanie tego typu zjawisk wymaga specjalnych zbiorów treningowych i dedykowanych, dość złożonych rozwiązań algorytmicznych.
 - Odniesienia do kontekstu społecznego i rzeczywistości pozatekstowej. W przypadku tego typu błędów nie istnieją obecnie automatyczne metody przynoszące zadowalające efekty. W dużym stopniu jest to kwestia doświadczeń, indywidualnej wrażliwości i orientacji podmiotu kodującego.

Generalną konkluzją płynącą z tych eksperymentów jest stwierdzenie, że tworzenie skutecznych

narzędzi automatycznych, rozpoznających wybrane aspekty znaczenia tekstów, takich jak ich obraźliwość względem mniejszości, wymaga zapewnienia, że badacze posługują się tym samym zestawem pojęć. Wysoka niezgodność pokazuje na różnice w rozumieniu pojęć, co odbija się bezpośrednio na jakości automatycznych modeli komputerowych. Niezgodność ta często wynika z nieostrych granic między opisywanymi zjawiskami lub brakiem kompetencji językowych osób kodujących. W analizie treści w naukach społecznych brak zgodności wiąże się także z różnicami w wartościach między anotującymi.

Wszystkie te obserwacje prowadzą nas do stwierdzenia, że kodowanie automatyczne nie jest metodą samodzielną; jest propozycją przypisania określonych kategorii, która powinna być ponownie weryfikowana przez badaczy.

Podsumowanie

W prezentowanym tekście staraliśmy się prześledzić kolejne kroki na drodze do owocnej współpracy pomiędzy naukami społecznymi a lingwistyką komputerową. Analizowany projekt był realizowany według metody typowej dla jakościowej analizy treści, znaczącą zmianą jest zastosowanie uczenia maszynowego, by zastąpić kodowanie przez grupę koderów/anotatorów kodowaniem maszynowym.

Wracając do postawionego powyżej pytania: jak zoptymalizować działania zespołu badawczego pod kątem przygotowania zakodowanych tekstów, które staną się próbą uczącą dla algorytmów uczenia maszynowego?

Bazując na doświadczeniach przy realizacji monitoringu mowy nienawiści, wskazujemy następujące elementy optymalizacji procesu badawczego:

1. Cały proces musi być przygotowany wyjątkowo starannie, konieczna jest werbalizacja jak największej liczby założeń (przyjmowanych najczęściej implicite przez członków zespołu), szczegółowa algorytmizacja kolejnych działań przewidzianych w poszczególnych etapach.
2. Konieczna jest precyzyjna konceptualizacja i operacjonalizacja kategorii teoretycznych oraz pokazanie na przykładach brzegowych (trudnych, niejednoznacznych), jak te kategorie stosować w praktyce.
3. Konieczne jest, przy budowie korpusu, uwzględnienie na poziomie słów kluczowych zarówno określeń neutralnych, pozytywnych, jak i negatywnych (obraźliwych).
4. Szczególnej uwagi wymaga proces kodowania zebranych fragmentów tekstu. Dla maksymalizacji zgodności kodujących konieczne jest:
 - a. wprowadzenie rozbudowanej instrukcji kodowej – przygotowanej przez doświadczonych badaczy, którzy w pełni rozumieją cele badania oraz brali udział we wszystkich jego etapach (konceptualizacji i operacjonalizacji);
 - b. stworzenie zespołu koderów/annotatorów – osób o zbliżonych kompetencjach językowych, które gotowe są przyjąć i przyswoić sposób interpretacji tekstu narzucony przez badaczy;

- c. przeprowadzenie treningu koderów/annotatorów – „próbne” kodowanie tekstów wybranych (wylosowanych) z korpusu; proces ten powinien być oparty o kilka iteracji. Kluczowa jest werbalizacja przesłanek stojących za wyborem poszczególnych kodów, a następnie ujednoczenie tych przesłanek dla całego zespołu, osiągnięcie konsensusu w procesie racjonalnej dyskusji;
- d. „kody techniczne” – konieczne jest wprowadzenie kodów, które pozwolą rozdzielić różne sposoby nadawania sensu wypowiedzi. W klasycznym kodowaniu te rozróżnienia są elementem kompetencji językowej kodujących i pozostają najczęściej niezwerbalizowane.

Poprawa jakości danych, będąca skutkiem zastosowania zbioru niniejszych zaleceń, ma widoczny i znaczący efekt w postaci lepszej jakości modeli uczenia maszynowego. Oczywiście nie wyczerpuje się w ten sposób możliwości popełnienia zarówno pomyłek ludzkich – na etapie tworzenia danych treningowych – oraz pomyłek maszyny, na etapie testowania modeli algorytmów. Tematyka automatycznego kodowania treści wymaga dalszych badań, kierunek wyznacza rozwijająca się infrastruktura Clarin (narzędzia NLP na potrzeby nauk społecznych i humanistycznych, <http://clarin-pl.eu/pl/czym-jest-clarin/>).

W ocenie autorów niniejszego tekstu dalszy rozwój metod przetwarzania języka naturalnego umożliwi komputerowe rozpoznawanie coraz bardziej złożonych pojęć, wykraczających poza obszar lingwistyki. Pojęć społecznie konstruowanych, które tradycyjnie stanowiły domenę jakościowych badań nad tekstem. Z technicznego punktu widzenia stanie

się to możliwe dzięki coraz lepszym metodom reprezentacji znaczenia słów i tekstu (ang. *embeddings*) oraz głębokiemu uczeniu maszynowemu.

To przybliży nas do zrozumienia celu opisywanego procesu – skutecznej współpracy pomiędzy przedstawicielami NLP a nauk społecznych. Wydaje się, że na obecnym etapie specjalizacji poszczególnych dyscyplin konieczne jest budowanie multidyscy-

plinarnych zespołów naukowych. Dzięki temu badacze społeczni zyskają narzędzia pozwalające skutecznie analizować duże korpusy tekstów, co między innymi przybliży nas do zrozumienia zjawisk społecznych reprezentowanych przez dyskurs internetowy, a lingwiści komputerowi otrzymują możliwość zderzenia wypracowanych algorytmów i metod analizy danych z realnymi, nietrywialnymi problemami społecznymi.

Bibliografia

- Bishop Christopher (2006) *Pattern Recognition and Machine Learning*. Secaucus: Springer-Verlag.
- Breiman Leon (2001) *Random Forests*. „Machine Learning”, vol. 45, no. 1, s. 5–32.
- Bychawska-Siniarska Dominika, Głowacka Dorota, red., (2013) *Mowa nienawiści w internecie: jak z nią walczyć*. Warszawa: Helsińska Fundacja Praw Człowieka.
- Bychawska-Siniarska Dominika, Gliszczyńska-Grabias Aleksandra (2016) *W stronę sieci tolerancji. Prawnomiędzynarodowe instrumenty walki z mową nienawiści* [dostęp 14 maja 2017 r.]. Dostępny w Internecie <<http://www.siecieterancji.pl/aktualnosci/w-strone-sieci-tolerancji-publicacja-w-module-prawnym>>.
- Cortes Corinna, Vapnik Vladimir (1995) *Support-Vector Networks*. „Machine Learning”, vol. 20, no. 3, s. 273–297.
- Gutierrez Dario i in. (2016) *Literal and Metaphorical Senses in Compositional Distributional Semantic Models*. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, (ACL) 2016, August 7-12, 2016, Berlin, Germany, vol. 1 [dostęp 14 maja 2017 r.]. Dostępny w Internecie: <<http://aclweb.org/anthology/P/P16/P16-1018.pdf>>.
- Heinze Eric (2016) *Hate Speech and Democratic Citizenship*. Oxford: Oxford University Press.

- Jockers Matthew (2013) *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.
- Krejtz Izabela, Krejtz Krzysztof (2005) *Wybrane statystyki zgodności między sędziami w analizie treści* [w:] Katarzyna Stemplewska-Żakowicz, Krzysztof Krejtz, red., *Wywiad psychologiczny. Wywiad jako postępowanie badawcze*. Warszawa: Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego, s. 231–249.
- Krippendorff Klaus (1980) *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA: Sage.
- Lafferty John D., McCallum Andrew, Pereira Fernando C. N. (2001) *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01), San Francisco, USA, Morgan Kaufmann Publishers Inc., s. 282–289.
- Lample Guillaume i in. (2016) *Neural Architectures for Named Entity Recognition*. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. The Association for Computational Linguistics, s. 260–270.
- Linde-Usiekiewicz Jadwiga (2015) *Teoria relewancji jako narzędzie opisu mowy nienawiści*. „Studia Pragmalingwistyczne”, t. 7, s. 53–68.
- Lombard Matthew, Snyder-Duch Jennifer, Bracken Cheryl Campanella (2004) *A Call for Standardization in Content Ana-*

lysis Reliability. „Human Communication Research”, vol. 30, s. 434–437.

Łodziński Sławomir (2003) *Problemy dyskryminacji osób należących do mniejszości narodowych i etnicznych w Polsce*. Warszawa: Kancelaria Sejmu, Biuro Studiów i Ekspertyz.

Manning Christopher D. i in. (2014) *The Stanford CoreNLP Natural Language Processing Toolkit*. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. The Association for Computational Linguistics. ACL, System Demonstrations.

Moretti Franco (2013) *Distant Reading*. London: Verso Books.

Nijakowski Lech (2008) *Mowa nienawiści w świetle teorii dyskursu* [w:] Anna Horolets, red., *Analiza dyskursu w socjologii i dla socjologii*. Warszawa: Wydawnictwo Adam Marszałek, s. 113–133.

Ogrodniczuk Maciej, Lenart Michał (2013) *A Multi-Purpose Online Toolset for NLP Applications*. Proceedings of the 18th International Conference on Applications of Natural Language to Information Systems, vol. 7934 of Lecture Notes in Computer Science, Springer-Verlag. Springer Berlin Heidelberg, s. 392–395.

Pedregosa Fabian i in. (2011) *Scikit-Learn: Machine Learning in Python*. „Journal of Machine Learning Research”, vol. 12, s. 2825–2830.

Siwicki Maciej (2011) *Nielegalna i szkodliwa treść w Internecie. Aspekty prawnokarne*. Warszawa: Oficyna Wolters Kluwer.

Sperber Dan, Wilson Deidre (2011) *Relevancja. Komunikacja i poznanie*. Przełożyły Magdalena Charzyńska i n.. Kraków: Wydawnictwo Tertium.

Stone Philip J. i in. (1966) *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: MIT Press.

Troszyński Marek (2015) *Hate Speech. Towards a Research Standard* [w:] Jacek Sobczak, Jędrzej Skrzypczak, red., *Professionalism in Journalism in the Era of New Media*. Berlin: Logos, s. 199–208.

Wawer Aleksander, Rogozińska Dominika (2012) *How much supervision? Corpus-based lexeme sentiment estimation*. IEEE 12th International Conference on Data Mining Workshops (SENTIRE 2012), Los Alamitos, USA, IEEE Computer Society, s. 724–730

Wieruszewski Roman i in., red., (2010) *Mowa nienawiści a wolność słowa. Aspekty prawne i społeczne*. Warszawa: Wolters Kluwer.

Cytowanie

Troszyński Marek, Wawer Aleksander (2017) *Czy komputer rozpozna hejtera? Wykorzystanie uczenia maszynowego (ML) w jakościowej analizie danych*. „Przegląd Socjologii Jakościowej”, t. 13, nr 2, s. 62–80 [dostęp dzień, miesiąc, rok]. Dostępny w Internecie: <www.przegladsocjologiijakosciowej.org>.

Can a Computer Recognize Hate Speech? Machine Learning (ML) in Qualitative Data Analysis

Abstract: The purpose of this article is to present the process of automatic tagging of hate speech in social media. The implementation of this process allows for quantitative treatment of qualitative methods: analysis on the corpora of hundreds thousands of texts based on their meaning. The process is possible through algorithms of machine learning (ML).

The example of the hate speech designation project in texts from Polish online forums is presented. The key issue is the precise of conceptualization and operationalization of category “hate speech.” This allows for preparing specific instructions and conducting the training code unit. As a result we get higher rates of inter-coder agreement. Marked texts will be used as training data for automated categorization methods based on ML algorithms. Then we describe the course of machine coding. This article also seeks to establish problems associated with automatic coding of hate speech and propose solutions. In summary, we point the factors that are crucial to the research process that uses machine learning.

Keywords: machine learning, qualitative data analysis, hate speech, intercoder agreement

Agnieszka Kwiatkowska
Uniwersytet SWPS, Warszawa

„Hańba w Sejmie” – zastosowanie modeli generatywnych do analizy debat parlamentarnych

Abstrakt W ciągu ostatniego dziesięciolecia nastąpił wyraźny wzrost obecności skrajnej retoryki politycznej w debacie parlamentarnej w Polsce. Jednym z aspektów postępującej radykalizacji języka jest nadużywanie słów o wysokim negatywnym natężeniu emocjonalnym w celu opisanie zdarzeń występujących w codziennej polityce, co prowadzi do degradacji ich znaczenia. Jak wskazują badania, brutalizacja języka debaty wywołuje negatywne konsekwencje na poziomie elit politycznych i elektoratu, prowadząc do zmniejszonej efektywności działania parlamentu, zaostrzenia konfliktów między partiami i ich zwolennikami oraz do alienacji politycznej obywateli.

W artykule przedstawiam możliwości wykorzystania stenogramów parlamentarnych do analizy przemian dyskursu politycznego, w tym wykorzystane algorytmy przeszukiwania korpusów tekstów oraz analizy ukrytych tematów. Jako przykład wykorzystuję zbiór przemówień sejmowych z lat 1991–2016 odnoszących się do idei hańby, zdrady, niesławy i skandalu. Opieram się głównie na generatywnym modelu tematycznym, wykorzystującym metodę ukrytej alokacji Dirichleta i jej rozszerzeniu – strukturalny model tematyczny, będących nienadzorowanymi metodami ekstrakcji tematów z dużych korpusów tekstów. Wyniki przeprowadzonych analiz potwierdzają wzrost radykalnego słownictwa w czasie, a także określają charakterystyki jego użytkowników oraz najczęściej występujące konteksty (polityka historyczna, kontrola nad mediami, polityka zagraniczna).

Słowa kluczowe Polska, debaty parlamentarne, elity polityczne, ilościowa analiza treści, *text mining*

Agnieszka Kwiatkowska, socjolog, politolog, adiunkt w Instytucie Nauk Społecznych Uniwersytetu SWPS, członek zespołu Polskiego Generalnego Studium Wyborczego. Zainteresowania naukowe: zachowania wyborcze, preferencje ideologiczne, elity polityczne, dyskurs publiczny.

Adres kontaktowy:

Centrum Studiów nad Demokracją
Uniwersytet SWPS
ul. Chodakowska 19/31, 03-815 Warszawa
e-mail: agn.kwiatkowska@swps.edu.pl

Język polskiej polityki jest postrzegany negatywnie zarówno przez językoznawców, jak i społeczeństwo. Pomimo że jest to język elit, to przypisywane mu są potocznie „same złe cechy: manipulacyjność, mętniactwo, wulgarność, agresywność” (Bralczyk 2003: 7). Charakterystyczne dla niego jest zastępowanie merytorycznych argumentów retoryką moralistyczną, brak nakierowania na kompromis, autoprezentacyjne monologi, rytualizm, a także brak poszanowania dla oponentów – partnerów

w dyskursie (Piotrowski 1997). Wprowadzanie do codziennej polityki silnych emocji i zachowań będących ich ekspresją, często nieuzasadnionych, a stanowiących jedynie mechanizm walki politycznej, utrudnia prowadzenie rzetelnej debaty publicznej. W ciągu ostatniej dekady te negatywne cechy jeszcze bardziej się wyostrzyły.

Obserwując polską debatę polityczną – toczoną na forum parlamentu i w mediach – zauważyć można wyraźny wzrost zbrutalizowanej retoryki politycznej, przejawiający się między innymi w częstym wykorzystywaniu skrajnie negatywnie nacechowanych emocjonalnie określeń. Takie słowa jak „hańba”, „zaprzaństwo” czy „zdrada”, wraz ze wzrostem częstotliwości ich wykorzystywania w polityce, przestały opisywać wyłącznie szokujące, na długo zapadające w pamięć wydarzenia historyczne i zdewaluowały się, zmieniając się w niebudzące sprzeciwu otoczenia określenia działań, których nie aprobujemy. Nadużywanie skrajnych słów sprawiło, że ich autorzy i odbiorcy uodpornili się na ich moc i przestały one spełniać swoją karzącą funkcję.

Wzrost skrajnej retoryki w polityce nie jest, oczywiście, zjawiskiem charakterystycznym wyłącznie dla Polski. Badając przemówienia w Kongresie Stanów Zjednoczonych, Uslaner (1993) zauważa, że kryzys, który osłabił tradycyjne normy wytyczające zakres dopuszczalnych zachowań w debacie parlamentarnej, takich jak grzeczność, uprzejmość, taktowność, rozpoczął się już w latach siedemdziesiątych i osiemdziesiątych jako odzwierciedlenie zwiększonej polaryzacji partyjnej po udziale Stanów Zjednoczonych w wojnie w Wietnamie oraz po aferze Wa-

tergate. Również w innych krajach zaobserwowano brutalizację języka polityki (Semetko, Schoenbach 2003; Ilie 2004; Spary 2010; Bates i in. 2014; Murphy 2014).

Postępującą brutalizację języka możemy opisać jako wzrost częstości wykorzystywania słów o wysokim natężeniu emocjonalnym oraz pojęć, w których konotacja dominuje nad denotacją (Bralczyk 2016). Wykorzystywanie słów ekstremalnie negatywnie nasyconych emocjonalnie na określenie zdarzeń zachodzących w ramach zwykłej, codziennej polityki prowadzi do degradacji ich znaczenia. Uczestnicy i odbiorcy debaty politycznej przyzwyczajają się z czasem do najgorszych epitetów, jeżeli są regularnie powtarzane, co inspiruje dalszą brutalizację języka. Wykorzystana w tytule artykułu „hańba”, czyli termin określający według *Słownika języka polskiego PWN* „niegodne postępowanie przynoszące ogólne potępienie i pogardę” oraz „ujmę na honorze” (*Słownik języka polskiego PWN*), należała kiedyś do słów uznawanych za nieakceptowalne w standardowej dyskusji politycznej. W potransformacyjnej retoryce politycznej pojawia się nie tylko w oskarżeniach o czyny największego kalibru, ale nawet jako wyraz niezadowolenia z decyzji administracyjnej oponenta na poziomie samorządu gminy (por. Urzykowski 2016).

Badania empiryczne dotyczące społecznej akceptacji zachowań politycznych wykraczających poza normy kulturowe prowadzone są przede wszystkim w okresach nasilenia rywalizacji politycznej – czyli w okresie kampanii wyborczej. Wnioski z nich płynące obejmujące opłacalność takich zachowań są niejednoznaczne. Fridkin i Kenney

(2004), dokonując przeglądu literatury w zakresie skuteczności kampanii negatywnej, wskazują, że – zgodnie z rezultatami większości badań – elementy skrajnie negatywne przyciągają uwagę odbiorcy, powodując wzrost jego zainteresowania mówcą i przedmiotem dyskusji i zwiększają jego emocjonalny stosunek do polityki, a także wpływają na zwiększenie atrakcyjności stosujących je kandydatów kosztem ich konkurentów, odnotowują jednak szereg badań wskazujących na brak efektu takich kampanii. Metaanalizy przeprowadzone przez Lau, Sigelman, Heldman i Babbitt (1999) oraz Lau, Sigelman i Rovner (2007), obejmujące badania nad wpływem kampanii negatywnych na ocenę kandydatów, wskazują, że kampanie negatywne nie są bardziej skuteczne niż pozytywne. Późniejsze prace Krupnikov (2011; 2012) sugerują jednak, że wpływ kampanii negatywnych widoczny jest zarówno w wyborze kandydata, jak i poziomie demobilizacji politycznej, jednak wyłącznie przy spełnieniu określonych warunków dotyczących chronologii procesu podejmowania decyzji oraz specyficznych form kampanii.

Skromniejszą, chociaż rosnącą w ostatnich latach grupę pozycji teoretycznych stanowią badania obejmujące wykorzystanie skrajnego słownictwa w codziennym dyskursie parlamentarnym. Prace w tej kategorii dotyczą głównie Kongresu Stanów Zjednoczonych (Uslaner 2000; Dodd, Schraufnagel 2012; Wolf, Strachan, Shea 2012) i parlamentu Wielkiej Brytanii (Harris 2001; Jamieson 2001; Ilie 2014; Murphy 2014). Istnieją również nieliczne prace podejmujące ten problem w odniesieniu do polskiego Sejmu (Kamińska-Szmaj 2001; Laskowska 2004; Batko-Tokarz 2008; Piniarski 2011; Polkowska 2015).

Wykorzystywanie negatywnego słownictwa jest kuszące dla polityków. Wyrzistość wypowiedzi, przede wszystkim ich brutalizacja, pozwala wybić się spośród natłoku informacji przekazywanych w mediach i szybko zyskać rozpoznawalność. Dodatkowo opisywanie przeciwników politycznych w kategoriach absolutnego dobra i zła, a nie ze względu na ich kompetencje pozwala na uniknięcie konieczności merytorycznej konfrontacji z ich argumentami. Moralna stygmatyzacja przeciwnika ma służyć jego dyskredytacji jako osoby niezdolnej do prowadzenia polityki bez względu na jego wcześniejsze osiągnięcia.

Język polityki, łączący funkcję środka komunikacji i narzędzia rywalizacji, jest szczególnie podatny na ekstremizację. Rozwój Internetu i poszerzenie dostępu do materiałów z debat parlamentarnych sprawia, że skrajne, brutalne wypowiedzi szybko rozchodzą się w społeczeństwie. Prosty, wyrazisty i radykalny język wydaje się być w opinii części wyborców bardziej zbliżony do prawdziwego opisu rzeczywistości, a jego użytkownicy „mówią, jak jest”. Dodatkowo agresja, niski poziom dyskusji, inwektywy i poniżanie przeciwnika podsycają antyelitarystyczną satysfakcję wyborców. Co więcej, media, w pogoni za zwiększaniem oglądalności, chętniej skupiają się na emocjonalnych, zwłaszcza w negatywnym znaczeniu, częściach debaty parlamentarnej niż na elementach merytorycznych, nużących z punktu widzenia większości widzów, czytelników, słuchaczy.

Brutalizacja języka polityki to nie tylko kwestia estetyki. Jakość debaty parlamentarnej, w tym również poziom brutalizacji języka, ma skutki na poziomie

elit politycznych i elektoratów. Ekstremalna retoryka na obu tych poziomach prowadzi do wzrostu polaryzacji, trudności z zawieraniem kompromisu, postrzegania oponentów politycznych jako aktywnie chcących nam zaszkodzić wrogów, a nie osób z odmiennymi poglądami (Jamieson 2001; Wolf, Strachan, Shea 2012; Maisel 2012). Obowiązujące reguły zachowania w debacie parlamentarnej mają na celu załagodzenie rywalizacyjności i umożliwienie dojścia do kompromisu w podejmowanych decyzjach pomimo rozbieżnych pozycji ideologicznych. Jednakże język oparty na oczernianiu oponenta i przypisywaniu mu najgorszych motywacji polaryzacyjność tę wzmacnia. Agresja słowna w polityce tworzy wrażenie wszechobecnego konfliktu, w którym nie można zająć neutralnych stanowisk. Nadmierna emocjonalizacja debaty parlamentarnej jest również niekorzystna pod kątem efektywności polityk państwowych – zastępuje rzeczową dyskusję miałą kłótnią. Badania prowadzone nad pracami Kongresu Stanów wykazały, że liczba wyzwisk w debacie parlamentarnej jest pozytywnie skorelowana z nieefektywnością prac: dłuższym czasem wymaganym do osiągnięcia konsensusu i niższą liczbą podejmowanych decyzji (Jamieson 2001).

Język debaty parlamentarnej wpływa na styl komunikacji publicznej. Ma charakter normotwórczy, ponieważ ludzie przejmują zwyczaje językowe elit politycznych, z którymi stykają się za pośrednictwem mediów (Bralczyk 2003: 8). Co więcej, badania Kalmoe (2014) wykazały, że naładowanie agresją retoryki politycznej wpływa u wyborców z dyspozycją do zachowań agresywnych na wzrost poparcia dla aktów przemocy politycznej. Dodatkowo skrajna retoryka w polityce, jeżeli stosowana jest nagmin-

nie, w długim okresie wpływa na zniechęcenie polityką, cyniczne interpretowanie świata polityki, spadek poczucia wpływu i afiliacji politycznej, co w perspektywie długoterminowej przekłada się na niższą partycypację polityczną (Fridkin, Kenney 2004; Mutz 2015).

Jeżeli członkowie wszystkich ugrupowań są oskarżani o zdradę, pohańbienie państwa – prowadzi to do delegitymizacji społecznej całej elity. Na podstawie badań eksperymentalnych nad wyborami do Senatu Stanów Zjednoczonych Fridkin i Kenney (2004) stwierdzili, że negatywne komunikaty padające ze strony konkurujących kandydatów pod adresem przeciwnika są skuteczne jako narzędzie walki politycznej, gdy skupiają się na kwestiach merytorycznych i dotyczą proponowanych polityk, a nie osoby kandydata. Jednak wówczas, kiedy negatywne komunikaty sformułowane są w zbyt skrajnej formie i koncentrują się na niemerytorycznych aspektach, wpływają na pogorszenie oceny obu kandydatów biorących udział w kampanii – atakującego i atakowanego.

Wykorzystywanie skrajnej retoryki w nowych demokracjach ułatwia brak doświadczenia publicznego dużej części elity politycznej. Słabo rozwinięta kultura polityczna sprawia, że używanie silnie negatywnie nacechowanego słownictwa rządziej jest interpretowane jako naruszenie norm społecznych dotyczących zachowania posła. Wynikający z krótkiego okresu zdobywania demokratycznego doświadczenia niski poziom wiedzy politycznej – na poziomie elit i elektoratów – przyczynia się do faworyzowania kwestii symbolicznych, określanych też mianem kwestii łatwych, dotyczących przede

wszystkim dylematów tożsamościowych i światopoglądowych, względem wymagających większego wyrafinowania politycznego kwestii trudnych, obejmujących merytoryczne aspekty polityk sektorowych (Carmines, Stimson 1986). Kwestie symboliczne i tożsamościowe stanowią z kolei podatny grunt dla konfliktu i retoryki politycznej opartej na negatywnych emocjach i słownym stygmatyzowaniu przeciwnika.

Zakres negatywnych skutków w obszarze współdziałania politycznego i zniechęcenia wyborców do polityki decyduje o wadze problemu badawczego. W dalszych analizach postaram się zweryfikować główną hipotezę artykułu, zgodnie z którą następuje wzrost częstości wykorzystania radykalnego słownictwa w debacie parlamentarnej, a także odpowiedzieć na szereg szczegółowych pytań badawczych: Jak często słyszymy o hańbie w parlamencie? W jakich kontekstach tematycznych najczęściej pojawiają się zarzuty hańby, skandalu, zdrady? Jakie partie najczęściej posługują się silnie negatywnie nacechowanym słownictwem? Czy partie opozycyjne robią to częściej niż partie rządowe? Czy tego typu słowa częściej padają z ust wyznaczonego przez marszałka mówcy, czy z ław sejmowych, jako forma pozaregulaminowego dyskursu? Wreszcie, czy następuje zmiana języka debaty parlamentarnej, a także zmiana kontekstów, a więc tematów, w których pojawiają się zarzuty hańby i zdrady?

Równorzędnym celem artykułu jest przedstawienie metody modelowania tematycznego jako podejścia alternatywnego wobec słownikowych metod kategoryzowania danych jakościowych.

Metodologia i źródła danych

Stenogramy z posiedzeń parlamentarnych są bogatym, a z rzadka wykorzystywanym w polskiej politologii źródłem danych o pozycjach ideologicznych partii. Szczególnie widoczny jest brak pogłębionych analiz ilościowych (omówienie polskich prac poświęconych dyskursowi parlamentarnemu przedstawia Polkowska [2015: 13–17]). Tymczasem dane stenograficzne udostępniane na bieżąco przez Sejm Rzeczypospolitej Polskiej są wielowymiarowym zasobem tekstowym pozwalającym na diachroniczne analizy dynamiki rywalizacji partyjnej, wewnątrzpartyjnych konfliktów, a także stylu i jakości debaty parlamentarnej.

Język debat parlamentarnych stanowi specyficzną formę języka polityki, który zdefiniować możemy jako funkcjonalną odmianę języka ogólnego, stosowaną w tekstach wytwarzanych przez środowisko polityków i innych osób związanych ze sferą polityki, skierowaną intencjonalnie do ogółu społeczeństwa i charakteryzującą się dominacją funkcji perswazyjnej (Walczak 1994: 20). W debacie parlamentarnej, po pierwsze, podstawowym adresatem wypowiedzi są przede wszystkim obecni na sali posłowie, chociaż mówcy formułują swoje przemówienia ze świadomością, że przez obecność mediów na sali ich słowa – a przynajmniej część – trafi potencjalnie do szerokiego grona odbiorców. Zdaniem części badaczy to nieobecne, ale potencjalne audytorium jest dla mówców głównym partnerem w dyskusji, a nie znajdujący się na sali sojusznicy i oponenty, co sprawia, że uczestnicy nie komunikują się ze sobą, a zorientowani są przede wszystkim na wytworzenie zestawu uzasadnień skutecz-

nie oddziałujących na opinię publiczną (Tyszka 1996).

Drugą cechą wyróżniającą ten typ języka jest strukturyzowany charakter dyskursu, polegający na: sekwencyjnych wypowiedziach pojedynczych posłów, co uniemożliwia oficjalnie możliwość natychmiastowej repliki (jak wskażemy dalej w artykule, rekompensuje to wzrastająca liczba wtrąceń z sali), które artykułowane są w kolejności zarządzanej przez marszałka lub jego zastępcę, oraz są ograniczone czasowo, a także podlegają konwencjonalizacji i etykietce parlamentarnej. Wypowiedzi posłów należą w dużej części do specjalnej kategorii „tekstów pisanych, aby je powiedzieć” (*texts written-to-be-spoken*), przez co sytuują się między pisaną a mówioną formą języka.

Kolejną cechą, która decyduje o specyfice języka parlamentarnego, jest jego funkcjonalna dwoistość. Z jednej strony jego podstawowym zadaniem jest doprowadzenie do powstania dokumentów prawnych, ich modyfikacji lub uchylecia, przez co znaczna część dyskursu to teksty prawne, ich fragmenty lub odniesienia do nich. Z drugiej strony wypowiedzi posłów pełnią funkcję ekspresyjną, nakierowaną na retoryczne pokonanie przeciwnika – bezpośrednio na sali lub za pośrednictwem środków masowego przekazu – w oczach opinii publicznej.

Powyższe cechy języka debaty parlamentarnej sprawiają, że bardzo dobrze nadaje się on do analiz treści. Dzięki silniejszemu ustrukturyzowaniu w stosunku do języka naturalnej rozmowy lub sporu, charakterystycznemu dla debat politycznych odbywających się w mediach masowych, a także dzięki

formalnej strukturze debaty, umożliwiającej mówcy przekazanie całości wypowiedzi – o ile przestrzega wyznaczonego czasu – bez zabierania mu głosu, uzyskujemy wypowiedzi pogłębione, dotyczące szerokiego zakresu kwestii. Stenogramy zawierają wypowiedzi wielu przedstawicieli ugrupowań posiadających reprezentację parlamentarną, a nie tylko ich liderów, przez co wypowiedzi są zindywidualizowane pod kątem stylu, dobieranych argumentów, w przeciwieństwie do oficjalnych programów i stanowisk partyjnych. Pozwala to na prowadzenie rzetelnej obserwacji przemian języka dyskursu politycznego.

Zautomatyzowane analizy korpusów wystąpień i głosowań parlamentarnych w Sejmie RP pozwalają na wyodrębnienie kwestii istotnych w dyskursie parlamentarnym oraz prześledzenie ich zmian w czasie. W artykule, poza wprowadzającymi statystykami opisowymi, wykorzystuję metodę alokacji ukrytej zmiennej Dirichleta (LDA; Latent Dirichlet Allocation [Blei, Ng, Jordan 2003]), która jest nienadzorowaną (*unsupervised*) metodą automatycznej ekstrakcji podtematów z dużych korpusów tekstów niezależnie od podejmowanej przez nie problematyki. Tak więc można ją zastosować do dowolnych zbiorów dokumentów politycznych. Wśród wciąż stosunkowo rzadkich zastosowań LDA i jej rozszerzeń można wymienić między innymi analizę determinant powodzenia reform wyborczych w Wielkiej Brytanii (Moser, Reeves 2014), automatyczną systematyzację publicznie dostępnych dokumentów w szwedzkim dyskursie politycznym, w tym stenogramów posiedzeń parlamentarnych (Hägglöf 2014), czy predykcję przydziału ministerstw do poszczególnych członków

koalicji na podstawie analizy programów partyjnych i umowy koalicyjnej (Zirn, Stuckenschmidt 2014).

Nienadzorowane metody analizy danych tekstowych można widzieć jako zautomatyzowaną implementację założeń teorii ugruntowanej (Glaser, Strauss 2009) – nie narzucają one odgórnych interpretacji badacza poprzez wcześniej opracowany model teoretyczny, a uzyskiwane wyniki powstają na skutek działania modelu generatywnego. W metodzie tej zakłada się, że każdy dokument (wypowiedź) jest mieszaniną tematów, a każde słowo w dokumencie ma określone prawdopodobieństwo pojawienia się w konkretnym temacie. Tak więc tematy są reprezentowane jako rozkłady prawdopodobieństwa na słowach. Celem LDA jest rekonstrukcja tematów podejmowanych w analizowanym zbiorze tekstów w oparciu o współwystępowanie poszczególnych słów i fraz w wypowiedziach.

W przedstawionym w artykule badaniu zastosowany został nowy wariant tej metody o nazwie STM (strukturalny model tematyczny; *Structural Topic Model*) (Roberts i in. 2014). Wariant ten stanowi rozszerzenie LDA pozwalające na włączenie do modelu metadanych (*metadata*) opisujących dokumenty. Wyróżniającą cechą modelu jest to, że w STM dla każdego dokumentu aprioryczny rozkład prawdopodobieństwa przynależności słów do tematów jest zależny od metadanych. Dobór metadanych w badaniu wynika z ich dostępności na stronach internetowych Sejmu RP i obejmuje: imię i nazwisko mówcy, jego przynależność klubową, datę wypowiedzi oraz temat debaty (punkt

porządku dziennego obrad w danym dniu, na ogół zawierający tytuł diskutowanego projektu ustawy). Włączenie czasu i zmiennych politycznych do metadanych modelu umożliwia oszacowanie ich wpływu na rozkład wyodrębnionych tematów, a także przekłada się na lepsze dopasowanie modelu i niższy stopień błędu.

Wszystkie analizy przedstawione w artykule wykonane zostały za pomocą języka R (R Core Team 2014) w wersji 3.3.2 w środowisku R Studio (RStudio Team 2015) w wersji 0.99.903. Z podstawami pracy z językiem R zapoznać można się w pracy Fischettiego (2015), a z przykładowym zastosowaniem języka w analizie danych tekstowych w naukach społecznych w artykule Winczorka (2014).

Materiałem badawczym jest pełny zbiór wystąpień parlamentarnych w kadencjach I–VII i jednej czwartej kadencji VIII Sejmu RP (do końca 28. posiedzenia, czyli do 21 października 2016 włącznie) obejmujący łącznie około 125 milionów słów w 292 582 przemówieniach. Korpus przemówień sejmowych został pobrany ze strony Sejmu Rzeczypospolitej Polskiej (zob. <http://sejm.gov.pl>), połączony z dostępnymi również na tej stronie metadanymi dotyczącymi mówców, a pozyskane dane zostały oczyszczone i pozbawione błędów. Teksty przemówień zostały poddane analizie morfologicznej i sprowadzeniu każdego ze słów do podstawowej formy gramatycznej.

Następnie ze zbioru słów usunięte zostały tak zwane *stop words*, czyli często występujące słowa nieistotne z punktu widzenia analizy (np. spójniki) oraz słowa popularne i nieróżnicujące wypo-

wiedzi, w tym także słowa pojawiające się w neutralnych kontekstach niemal w każdej wypowiedzi parlamentarnej (tak jak „pan”, „marszałek”), związane ze specyfiką procedowania w Sejmie. Słowa te nie posiadają wartości informacyjnej dla prowadzonych analiz, a ich bardzo częste występowanie utrudnia zbudowanie modelu, stąd też usuwanie ich jest standardową praktyką. Teksty zostały również pozbawione interpunkcji.

Do celów postawionych przed badaniem z danych wejściowych wyselekcjonowany został podkorpus przemówień odnoszących się tematycznie do idei hańby, zdrady, niesławy i skandalu. Ze względu na to, że dopasowanie modeli generatywnych jest lepsze dla dużych zbiorów tekstów, rozszerzyłam podkorpus będący podstawą, na której budowany jest model, o słowa silnie negatywnie nacechowane emocjonalnie, stanowiące specyficzny podzbiór radykalnego języka polityki, mianowicie: słowa brzmiące archaicznie i rzadko używane w codziennych rozmowach, nacechowane patosem i ujemnym wartościowaniem moralnym, mające na celu dyskredytację osoby, zachowania lub wydarzenia: hańba, niesława, ośmieszenie, poruta, wstyd, podłość, dyshonor, skandal, łotrństwo, niegodziwość, zaprzaństwo, targowica, infamia, awantura, hucpa, zdrada oraz inne formy gramatyczne współdzielące ten sam rdzeń. Technicznie ujmując, wybrane zostały wszystkie i tylko te przemówienia sejmowe, w których co najmniej raz użyte zostało przynajmniej jedno ze słów wyszukanych za pomocą wyrażenia regularnego „hańb|hanieb|niesław|ośmiesz|wstyd|podł|y|e|łość|dyshonor|skandal|łotr|niegodziw|zapra|targowic|infam|awantur|hucp|zdrad|zdra-

jc”, a więc wszelkich możliwych form gramatycznych powstałych na bazie wyszukiwanych rdzeni lub ich rozszerzeń precyzujących, zawężających obszar wyszukiwania do pożądanego zakresu tematycznego.

Ostatecznie wybranych zostało 11 519 przemówień z 16 865 wystąpieniami słów kluczowych, co przedstawia tabela 1.

Analizy ilościowe

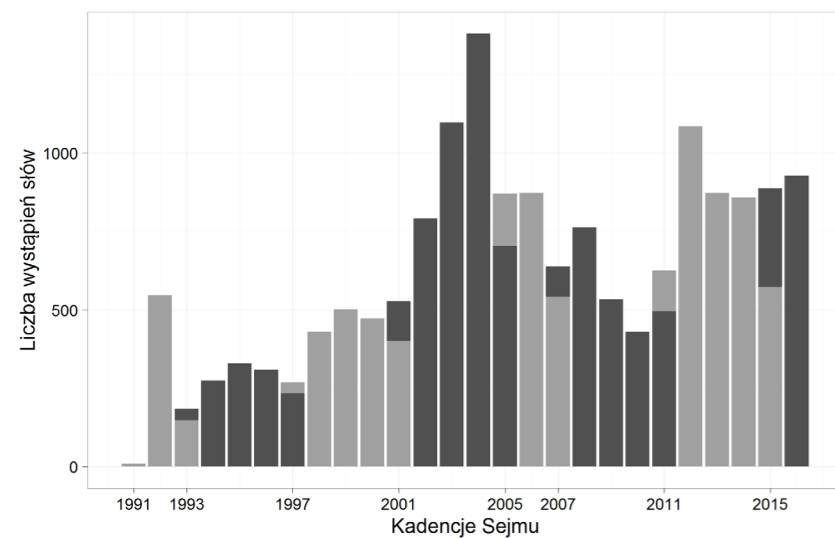
Zebrane dane wskazują na to, że przemówienia, w których używane jest słownictwo nacechowane ekstremalnie negatywnie, są nierównomiernie rozłożone w czasie (rys. 1). Po początkowym natężeniu w roku 1992 nastąpił spadek i okres względnego spokoju aż do roku 2001 włącznie (a zwłaszcza w pierwszej połowie tego okresu, w czasie kadencji 1993–1997 z rządzącą koalicją SLD-PSL). W latach 2002–2004 i pierwszej połowie 2005 roku (rządy SLD-UP, do 3 marca 2003 roku w koalicji z PSL) miał miejsce dramatyczny przyrost liczby przemówień wykorzystujących słownictwo ze zdefiniowanego powyżej zbioru, z rokiem 2004 stanowiącym dotychczasowe maksimum. W kolejnych dwóch kadencjach zaobserwować można spadek negatywnej retoryki – szczególnie w latach 2007–2011 w czasie pierwszego rządu Tuska (koalicja PO-PSL). Tendencja spadkowa odwróciła się od 2012 roku, od którego oskarżenia o zdradę, pohańbienie i zdradę utrzymują się na bardzo wysokim poziomie (zwłaszcza biorąc pod uwagę, że zamieszczone na wykresie liczebności nie uwzględniają końcówki 2016 roku).

Tabela 1. Występowanie wyszukiwanych słów w stenogramach Sejmu RP 1991–2016

WYSZUKIWANA FRAZA	PRZYKŁADY WYSTĘPOWANIA	LICZBA PRZEMÓWIEŃ	LICZBA WYSTĄPIEŃ SŁOWA
hańb hanieb	hańba, hańby, haniebny, hańbić, pohańbiony	1659	2159
niesław	niesława, niesławny, zniesławić, zniesławiony	284	389
ośmiesz	ośmieszyć, ośmieszenie, ośmieszył, ośmieszony	811	950
wstyd	wstyd, wstydić, zawstydza, bezwstydny	3612	4715
podł(y)łość	podły, podłe, podłość, podłemu, podłością	220	279
dyszonor	dyszonor, dyshonorem, dyshonoru	21	22
skandal	skandal, skandalu, skandaliczny, skandalista	3845	4881
łotr	łotr, łotrem, łotrze, łotrostwo	24	31
niegodziw	niegodziwy, niegodziwiec, niegodziwość	188	223
zaprza	zaprzaństwo, zaprzaństwami, zaprzaniec	36	39
targowic	targowica, targowicy, targowiczanie	131	187
infami	infamia, infamii, infamię	41	55
awantur	awantura, awanturnik, awanturować	696	862
hucp	hucpa, hucpa, hucpę, hucp	178	216
zdrad	zdrada, zdradzić, zdradzeni, zdradziecki	1108	1450
zdrajc	zdrajca, zdrajcy, zdrajcom	309	407
Razem		11519	16865

Uwaga: Liczba przemówień ogółem jest niższa od sumy składowych ze względu na współwystępowanie słów w przemówieniu. Wyszukiwanie na danych sprowadzonych do podstawowej formy gramatycznej. Odfiltrowane zostały również wystąpienia słowa „zdrada” jako nazwiska posła Jerzego Zdrady.

Rys. 1. Wykorzystanie radykalnego słownictwa w przemówieniach sejmowych w latach 1991–2016



Źródło: obliczenia własne. Kolorem wyróżnione zostały naprzemiennie kadencje Sejmu.

Uwzględnione w badaniu słowa najczęściej występują podczas zbiorczych debat sejmowych: w Oświadczeniach, Pytaniach w sprawach bieżących, Interpelacjach i zapytaniach, a także w trakcie mającej miejsce przynajmniej raz w trakcie kadencji debacie Przedstawienie przez Prezesa Rady Ministrów programu działania Rady Ministrów z wnioskiem o udzielenie jej wotum zaufania. Jeśli weźmiemy pod uwagę wyłącznie te debaty, podczas których dyskutowano indywidualne ustawy, to najczęściej skrajne słownictwo użyte zostało w debatach: Przedstawiony przez Radę Ministrów „Raport o stanie spraw publicznych i instytucji państwowych na dzień zakończenia rządów koalicji PO-PSL (2007–2015)” (45 przemówień z użyciem badanych zwrotów), Pierwsze czytania poselskich projektów uchwał w sprawie skrócenia kadencji Sejmu Rzeczypospolitej Polskiej (druki nr 3713, 3714 i 3955) (32 przemówienia) i Sprawozdanie Komisji Konstytucyjnej Zgromadzenia Narodowego o projekcie Konstytucji Rzeczypospolitej Polskiej – drugie czytanie.

Spośród samych przemówień najwięcej odniesień do zdrady, hańby i zbliżonych semantycznie określeń w jednym przemówieniu wykorzystanych zostało przez posła Zdzisława Jankowskiego (niezależny), który wygłaszając w debacie pod tytułem „Oświadczenia 26 listopada 2003 r.” przemówienie, użył 15 takich zwrotów. Poniżej przytaczam wybrane fragmenty tego oświadczenia:

To jest zdrada stanu, panie i panowie posłowie. Mijamy świadomość tego, że po 1989 r. w Polsce do władzy dorwali się sprzedawczyki i zdrajcy, którzy postawili sobie za cel zniszczenie państwa polskiego i zniewolenie narodu polskiego. (...) To jest hańba i zdrada pol-

skiej racji stanu. (...) Tej zdrady nie da się wytłumażyć, tę zdradę należy potępić. (...) Przyszłe pokolenie Polaków i Polek będzie nas przeklinało i złorzeczyło za tę monstrualną i haniebną zdradę dokonaną przez rezydentów brukselskich. (Jankowski 2003)

W dalszej kolejności najwięcej analizowanych zwrotów wykorzystał Antoni Stryjewski (KRN, IV kadencja) – 13 zwrotów w przemówieniu z 29 lipca 2004 dotyczącym uchwały Senatu w sprawie ustawy o uposażeniu posłów do Parlamentu Europejskiego:

Izba wyższa polskiego parlamentu jest także obecnie miejscem zdrady narodowej. (...) Senatorowie koalicji zdrady, podobnie jak posłowie tejże samej koalicji, używają Wysoką Izbę – jej autorytet i majestat – do uwiarygodnienia zdrady Rzeczypospolitej, do oszukania Polaków i wreszcie do służby poddańczej na rzecz obcej i wrogiej nam instytucji międzynarodowej – aspirującej i ewoluującej w kierunku superpaństwa pod hegemonią niemiecką – jaką jest Unia Europejska. (...) Dziś Polakom przychodzi płacić haniebną wolą Sejmu i Senatu Rzeczypospolitej wybrańcom, przebierańcom, funkcjonariuszom Unii Europejskiej z kasy polskiego państwa. (Stryjewski 2004)

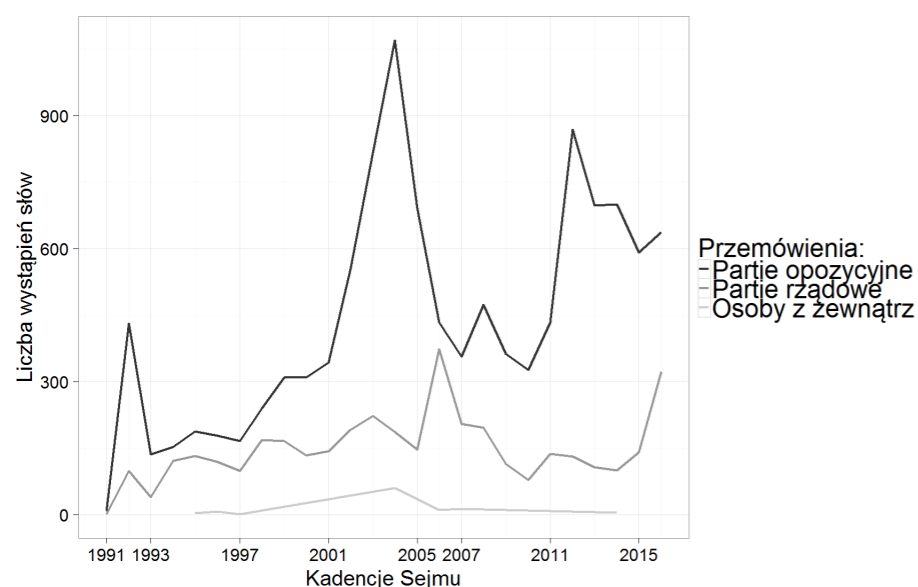
A następnie poseł Jarosław Kaczyński (PiS, V kadencja) – 12 zwrotów w przemówieniu z 9 lipca 2014 w debacie na temat wniosku o wyrażenie wotum nieufności Radzie Ministrów kierowanej przez Donalda Tuska i wybranie Piotra Glińskiego na prezesa Rady Ministrów:

poparcie dla rządu, który jest zły, niesprawny, zajmuje się socjotechnicznymi gramami, propagandowymi gramami ze społeczeństwem, który za nic ma prawo, konstytucję,

demokrację, który dla swojej obrony jest gotów sięgnąć do środków policyjnych, ukrywa prawdę, który jest przy tym doszczętnie skompromitowany i ośmieszony. (...) To przegraliśmy w sposób haniebny, ośmieszający. (...) sprawa Lasów Państwowych, to jest skandal, sprawa SKW i umowy z Rosją, to też jest skandal. (Kaczyński 2014).

Ekstremalne słownictwo to element retoryki przede wszystkim ugrupowań opozycyjnych, które wykorzystują ją znacznie częściej niż partie rządowe. Na rysunku 2 przedstawiono, że to właśnie partie opozycyjne są odpowiedzialne za szczytowe natężenia negatywnej retoryki, które wystąpiły w trakcie V i VII kadencji.

Rys. 2. Wykorzystanie radykalnego słownictwa w przemówieniach sejmowych w latach 1991–2016 w rozbiciu na partie rządowe i opozycyjne



Źródło: obliczenia własne. Wykres z danych po usunięciu ze zbioru głosów z sali.

Sumarycznie ekstremalnym słownictwem posługują się najczęściej zbiorczo potraktowane mniejsze kluby (rys. 3). Podczas pierwszego wzrostu wykorzystania ekstremalnego słownictwa (w 1992) to właśnie liczne grono nieskonsolidowanych partii, przede wszystkim prawicowych, w rozdrobionym politycznie Sejmie głównie odpowiadało za radykalizację języka, przede wszystkim (kolejno): Konfederacja Polski Niepodległej, Unia Demokratyczna, Zjednoczenie Chrześcijańsko-Narodowe, Solidarność i Porozumie-

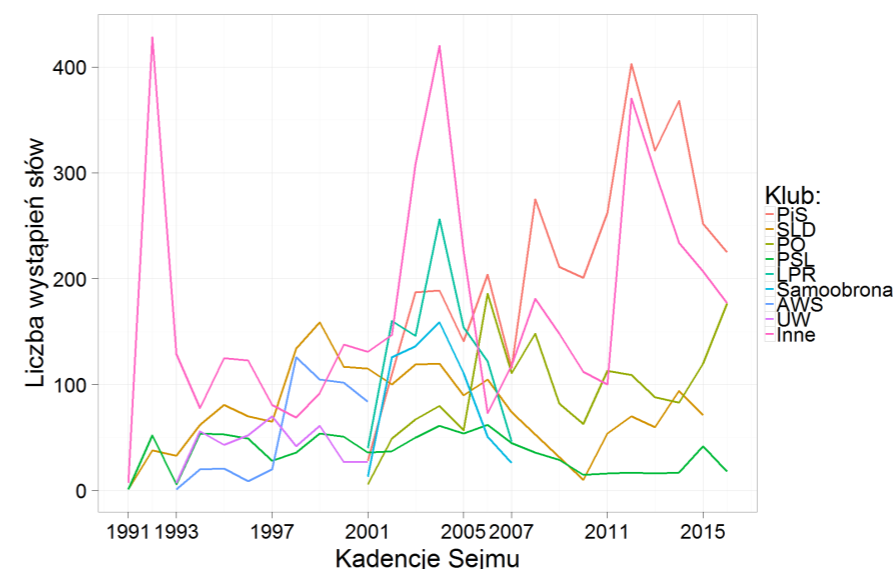
nie Centrum. Również podczas kolejnego okresu natężenia w czasie rządów lewicy 2001–2004, ze szczytem w 2004 roku, ugrupowaniami najczęściej uciekającymi się do oskarżeń o hańbę czy zdradę były (kolejno dla 2004 roku): Ruch Katolicko-Narodowy (a więc mikropartia wniesiona do Sejmu na listach LPR, której klub liczył w najlepszym momencie pięć osób), Socjaldemokracja Polska, Unia Pracy i Dom Ojczysty. Dostrzec jednak można w tym okresie silną pozycję trzech ugrupowań, które w następnej kaden-

cji Sejmu utworzyły razem koalicję rządową: LPR, PiS i Samoobrony. Pomiedzy tymi okresami największa liczba przemówień wykorzystujących słowa z analizowanego zbioru charakteryzowała główną partię opozycyjną w kadencji 1997–2001, czyli Sojusz Lewicy Demokratycznej.

Sytuacja zmieniła się jednak w połowie kadencji 2005–2007, czyli wówczas gdy do mniejszościowego rządu PiS dołączyli koalicjanci w postaci Samoobrony (w kwietniu 2006 r.) i LPR (maj 2006 r.). Od

tego roku Prawo i Sprawiedliwość jako główna partia rządowa stało się klubem najczęściej korzystającym z ekstremalnej retoryki i zachowało tę pozycję aż do 2015 roku. Szczególnie wyróżniało się ze względu na liczbę przemówień z analizowanymi słowami w latach 2008–2011. W 2016 roku, w czasie rządów Prawa i Sprawiedliwości, obserwujemy radykalizację zarówno mniejszych klubów (przede wszystkim Nowoczesnej), jak i niespotykaną wcześniej na tę skalę ekstremizację języka Platformy Obywatelskiej.

Rys. 3. Wykorzystanie radykalnego słownictwa w przemówieniach sejmowych w latach 1991–2016 w rozbiciu na kluby parlamentarne



Źródło: obliczenia własne. Wykres z danych po usunięciu ze zbioru głosów z sali.

Rok 2016 przyniósł również inne – choć związane z opisywanymi wcześniej – interesujące zjawisko, jakim jest nienotowany wcześniej udział retoryki ekstremalnej wykorzystywanej w głosach z sali, czyli parlamentarnych komentarzach. Komentarze w stenogramach sejmowych są pisane w sposób

wyróżniający je od głównego tekstu debaty i mają za zadanie odtworzenie wypowiedzi innych osób niż wyznaczeni regulaminem mówcy, a także atmosfery na sali, specyficznych zachowań posłów. Przykłady różnorodnych form komentarzy przedstawiam poniżej:

(Głos z sali: Lipa, lipa!)

(Poruszenie na sali)

(Oklaski)

(Głos z sali: Niewdzięcznicy, żeby mi to było ostatni raz!)

(Głos z sali: O Jezulu!)

(Głosy z sali: Hańba! Hańba!) (Oklaski)

(Poseł Ryszard Zbrzyzny: A gdzie komisje są, komisje śledcze?)

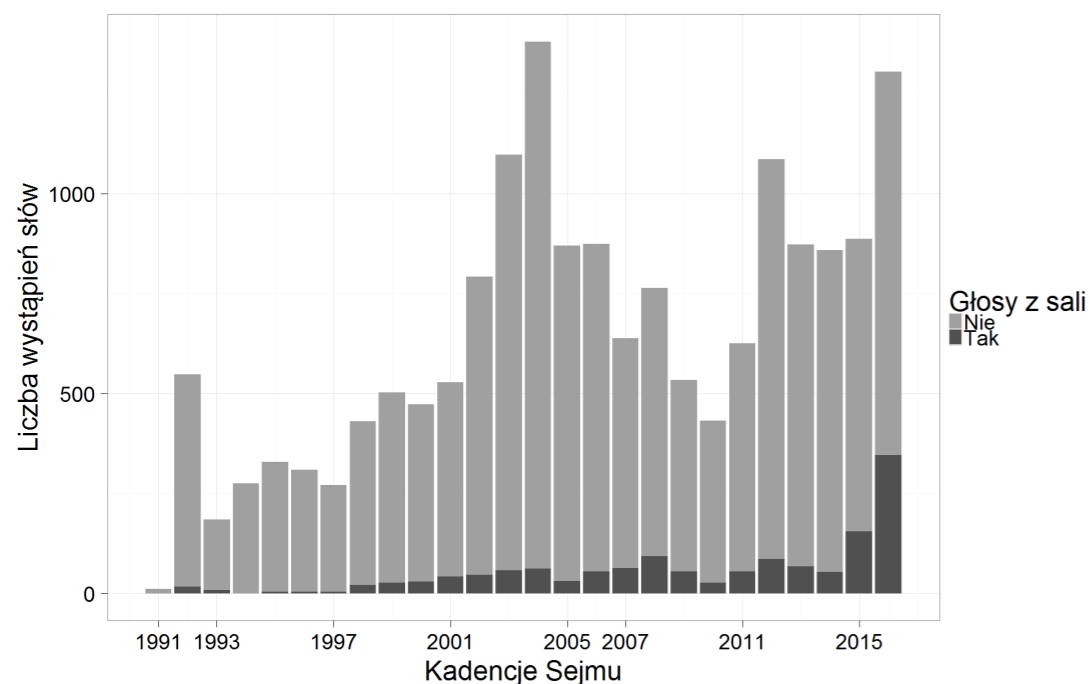
(Poseł Cezary Grabarczyk: Ty nas moralności nie ucz.)

(Poseł Iwona Arent: Jakie ty? Jakie ty, ty z Łodzi.)

nacechowane emocjonalnie, przede wszystkim stanowiąc mniej lub bardziej kulturalną polemikę z mówcą. W większości przypadków nie jest określone, komu przypisać autorstwo głosu z sali (stąd wykorzystywana konwencja oznaczania tych wtrąceń jako „głos z sali” lub „głosy z sali”), co powoduje niemożność przypisania głosów do klubów i kół parlamentarnych i ogranicza możliwości wykorzystania ich w dalszych analizach (można natomiast określić reprezentantom jakich ugrupowań najczęściej przerywają osoby na sali). Niektórzy mówcy odnoszą się w swoich przemówieniach do wtrąceń z sali, inni je ignorują. Czasami, jak w ostatnim przykładzie powyżej, wtrącenie prowadzi się w dyskusję z innymi posłami lub posłankami na sali.

Większość tekstu pobocznego to zapisane wtrącenia z sali w formie dopowiedzeń, sprostowań lub okrzyków. Często są one silnie negatywnie

Rys. 4. Wykorzystanie radykalnego słownictwa w przemówieniach sejmowych w latach 1991–2016 w rozbiu na przemówienia regulaminowe i głosy z sali

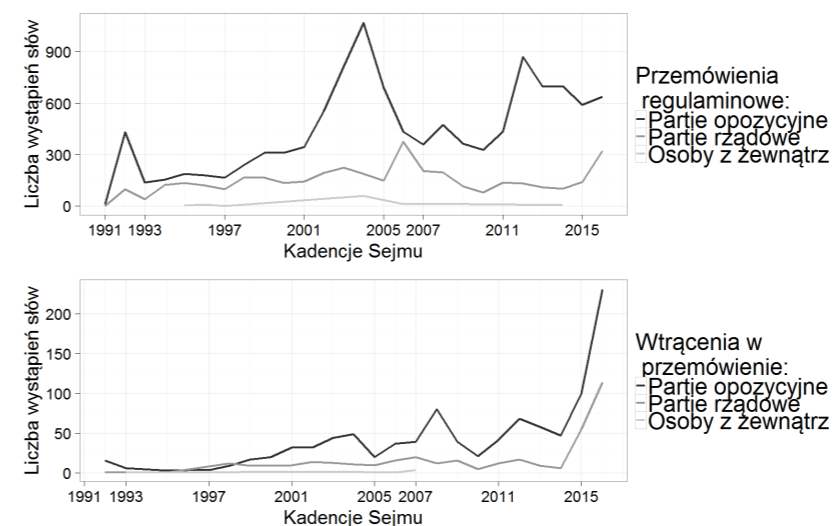


Źródło: obliczenia własne.

Liczba przemówień, w trakcie których padały z sali słowa z analizowanego zbioru, stanowi nieco ponad 9% ogółu analizowanych przemówień z negatywną retoryką (łącznie z wtrąceniami z sali). Jednakże w roku 2016 odsetek wtrąceń w przemówienie regulaminowe mówcy, zawierających przynajmniej jedno z tych słów, wzrasta do 26%. Jak obrazuje to rys. 5, wtrącenia z sali przerywające wypowiedź

mówcy dotyczą jednak dwukrotnie częściej posłów z ugrupowań opozycyjnych niż rządowych. Tak więc opozycja znacznie częściej niż siły rządowe wykorzystuje negatywną retorykę w swoich przemówieniach, jednak to również jej przemówienia są częściej przerywane z wykorzystaniem negatywnej retoryki niż przemówienia posłów i posłanek ugrupowań wchodzących w skład rządu.

Rys. 5. Wykorzystanie radykalnego słownictwa w przemówieniach sejmowych posłów partii rządzących i opozycyjnych w latach 1991–2016 w rozbiu na przemówienia regulaminowe i głosy z sali



Źródło: obliczenia własne.

Modelowanie tematyczne

Przechodząc do dalszej części analiz opartych o strukturalne modelowanie tematyczne, na zbudowanym podkorpusie do analizowanego zbioru wystąpień dopasowano szereg modeli STM w celu wydobycia ukrytych tematów, po czym wybrano i przeanalizowano model o zadanej liczbie $K = 18$ tematów, charakteryzujący się największą interpretowalnością wyników i wysokim stopniem dopasowania modelu, zgodnie z parametrami spójności

semantycznej (*semantic coherence*) i wyłączności (*exclusivity*) (Roberts, Stewart, Tingley 2016).

Tematy otrzymane w wyniku analizy można podzielić na trzy grupy. Pierwszą z nich są te, które można określić jako „techniczne”. Tematy takie składają się ze słów, które mogą wystąpić w każdym kontekście, choć nie zostały uwzględnione na liście *stop words*. Chociaż mogą plasować się wysoko na liście najczęściej występujących tematów w debatach, to częstotliwość ich pojawiania się nie zmienia

się istotnie w czasie ani pomiędzy partiami. Uznają je więc za nieistotne dla badanej kwestii. Dla tematów w tej grupie charakterystyczne są słowa takie jak: *marszałek, oklaski, wypowiedź, wysoki, mieć, swój, odpowiedź*. Wybrany do analizy model STM zwrócił trzy tematy techniczne odnoszące się kolejno do: procedowania sejmowego, tworzonego prawa oraz głosowań (patrz tabela 2a).

Drugą grupę stanowią tematy, które można określić jako efemeryczne. Ich cechą charakterystyczną jest częste występowanie nazw własnych, zwłaszcza imion i nazwisk posłów, oraz sformułowań odnoszących się do wydarzeń bieżących, istotnych i zajmujących znaczące miejsce w dyskursie publicznym w chwili wygłaszania przemówienia, jednak niemających charakteru trwałego.

Wreszcie tematy w trzeciej grupie, najistotniejszej z punktu widzenia badania, bezpośrednio przekładają się na konteksty najsilniej związane z ideą hań-

by, zdrady i skandalu. Określone zostały w związku z tym jako tematy merytoryczne. Przedstawiono je, wraz z częstością ich występowania, w tabeli 2b. Dla każdego tematu podano:

- jego krótką nazwę (która jest jedyną cechą narzuconą przez autora, nieprzypisaną automatycznie przez algorytm);
- listę dziesięciu słów wybieranych z tego tematu z największym prawdopodobieństwem (PROB);
- listę dziesięciu słów o najwyższym współczynniku częstości – wyłączności dla tego tematu (*Frequency–Exclusivity score*; FREX [Bischof, Airoidi 2012]), to znaczy takich słów, które pojawiają się często w danym temacie, a jednocześnie relatywnie rzadko w innych tematach;
- częstość pojawiania się słów wybranych z tego tematu w całej debacie (FREQ);
- względną częstość, to znaczy częstość wyłącznie wśród wybranych siedmiu istotnych tematów (SUMFREQ).

Tabela 2a. Tematy techniczne uzyskane z modelu STM

Temat	PROB	FREX	FREQ	SUMFREQ
Posiedzenia	ustawa, komisja, sejm, projekt, klub, sprawa, sprawozdanie, polski, zmiana, nad	sprawozdanie, uchwała, posiedzenie, projekt, poselski, czytanie, druk, klub, numer, sejm	4,63%	8,65%
Prawo	ustawa, projekt, praca, zmiana, dotyczy, prawo, osoba, wysoki, rozwiązanie, izba	zapis, nowelizacja, regulacja, podkomisja, legislacyjny, rozwiązanie, pracodawca, uregulować, inspekcja, regulować	7,58%	14,15%
Głosowania	przyjąć, głosowanie, artykuł, odrzucić, głosować, komisja, senat, sejm, przystępować, ręka	zechcieć, głosować, nieuzyskanie, senat, odrzucić, stwierdzać, podnieść, bezwzględny, wnosić, ręka	1,39%	2,59%

Źródło: obliczenia własne.

Tabela 2b. Tematy merytoryczne uzyskane z modelu STM

Temat	PROB	FREX	FREQ	SUMFREQ
Polityka historyczna	polski, naród, Polak, narodowy, rok, Polska, swój, Rzeczpospolita, wojna, historia	sowiecki, wykląć, Mikołajczyk, Pilecki, niepodległościowy, kombatancki, armia, krzyż, eksterminacja, wojenny	8,03%	14,99%
Media	publiczny, polityczny, telewizja, rada, wyborczy, polski, media, krajowy, wysoki, wybory	abonament, telewizja, radio, radiofonia, media, nadawca, wybory, telewizyjny, wyborca, wyborczy	6,59%	12,30%
Polityka zagraniczna	polski, europejski, unia, polityka, kraj, sprawa, zagraniczny, państwo, Polska	traktat, NATO, integracja, zboże, prezydencja, dyplomacja, Rosja, MSZ, europejski, członkostwo	6,35%	11,86%
Budżet	złoty, milion, podatek, finanse, środki, wydatek, fundusz, miliard, kwota, wzrost	wydatek, dochody, inflacja, deficyt, składka, ZUS, wzrost, ulga, oszczędność, rezerwa	5,91%	11,04%
Służby specjalne	sprawa, służba, osoba, komisja, policja, prokuratura, bezpieczeństwo, działanie, państwo, prokurator	prokuratura, policjant, śledztwo, policja, lustracja, funkcjonariusz, lustracyjny, śledczy, zeznanie, lew	5,06%	9,44%
Sądownictwo	prawo, konstytucyjny, sąd, konstytucja, trybunał, artykuł, państwo, karny, sprawa, prawny	sędzia, konstytucyjny, trybunał, orzeczenie, konstytucja, kara, sąd, karny, kodeks, orzekać	4,97%	9,28%
Służba zdrowia	dziecko, rodzina, zdrowie, osoba, życie, mieć, prawo, pacjent, społeczny, lekarz	vitro, pacjent, lek, leczenie, aborcja, przemoc, szpital, lekarz, dziecko, leczyć	4,88%	9,11%
Prywatyzacja	polski, bank, państwo, prywatyzacja, spółka, skarb, firma, majątek, przedsiębiorstwo	cukrownia, prywatyzować, stocznia, prywatyzacja, bank, prywatyzacyjny, kopalnia, giełda, węgiel, PZU	4,41%	8,23%
Edukacja	szkoła, samorząd, gmina, nauczyciel, edukacja, województwo, miasto, kultura, mieć, środowisko	uczelnia, nauczyciel, szkoła, oświatowy, burmistrz, zabytek, uczeń, powiat, impreza, las	3,82%	7,13%
Infrastruktura	droga, program, zostać, budowa, środki, mieć, autostrada, milion, ministerstwo, inwestycja	autostrada, PKP, fotoradar, odcinek, kilometr, winieta, kolejowy, kierowca, pociąg, drogowy	3,54%	6,61%

Źródło: obliczenia własne.

Z modelu STM uzyskano dziesięć znaczących tematów, wszystkie o wysokim stopniu interpretowalności. Poniżej omawiam, ze względu na ograniczenia objętościowe artykułu, wyłącznie sześć najczęstszych tematów, które wykorzystane zostały do dalszej analizy.

Najczęstszym kontekstem dla oskarżeń o hańbę, zdradę, skandal lub pokrewne okazała się polityka historyczna i związane z nią kwestie niepodległości, tożsamości narodowej i wojny. Najbardziej charakterystyczne przemówienia dla tego tematu, a więc takie, w których słowa mają najwyższe średnie prawdopodobieństwo przynależenia do tego tematu, to dwa przemówienia posła Jana Ziobry z klubu Solidarnej Polski (Oświadczenie z okazji 74. rocznicy wybuchu II wojny światowej wygłoszone 30 sierpnia 2013 r. oraz Oświadczenie z okazji rocznicy agresji ZSRR na II Rzeczpospolitą 17 września 1939 r. wygłoszone 26 września 2013 r.), przemówienie posła Lecha Kołakowskiego z PiS o Dniu Flagi Polskiej z 27 maja 2015 roku oraz przemówienie Marka Polaka z PiS z dnia 3 marca 2015 r.

Dnia 17 września 1939 r. dodatkowo na Polskę ruszyli zdradziecko ze wschodu Sowieci. (Ziobro 2013a)

Był to zdradziecki atak Sowietów na Polskę, która odparowała w tym czasie agresję ze strony hitlerowskich Niemiec. (Ziobro 2013b)

Jednak zdradzona i rozdarta rozbiorami Rzeczpospolita na długo zniknęła z mapy Europy. (Kołakowski 2015)

W hołdzie Żołnierzom Wyklętym – do końca wiernym bohaterom podziemia antykomunistycznego

AK, WiN, KWP, NZW, NSZ walczącym z hitlerowskim i sowieckim okupantem oraz ich komunistycznymi sojusznikami – zdrajcami narodu polskiego. Cześć i chwała Bohaterom. (Polak 2015)

Warto zauważyć, że wykorzystanie ekstremalnego słownictwa w tym temacie może być usprawiedliwione najmocniej spośród wszystkich tematów, ponieważ odwołuje się do wielkich wydarzeń historycznych, które na stałe wpisały się w świadomość kulturową Polaków, a opisywane wydarzenia i czyny postaci historycznych długoterminowo zaważyły na losach kraju. Stąd też rozpatrywanie w kategorii hańby lub zdrady takich wydarzeń jak rozbiory Polski, atak Związku Radzieckiego na Polskę w 1939 roku czy nawet współpraca z władzami komunistycznymi po wojnie może być sensowniej rozpatrywane według absolutnych kryteriów moralnych niż wydarzenia codziennej polityki, które znajdziemy w pozostałych tematach.

Kolejnym jeśli chodzi o częstość kontekstową hańby i zdrady tematem jest zarządzanie i regulacje dotyczące mediów. Podejmowane w nim zagadnienia to kwestie misji telewizji publicznej, upolitycznienia przekazu mediów publicznych oraz procesu nominowania na kluczowe stanowiska w tym sektorze, wykorzystywania telewizji do prowadzenia walki politycznej w trakcie kampanii wyborczych oraz nieprzejrzystego trybu przyznawania koncesji (ten problem pojawia się głównie w odniesieniu do Telewizji Niepokalanów i Telewizji Trwam). Przemówienia w tym temacie często pojawiają się przy okazji dyskusji nad rocznymi sprawozdaniami Krajowej Rady Radiofonii i Telewizji. Poniżej zamieszczam fragmenty z trzech najbardziej cha-

rakterystycznych przemówień: Anny Elżbiety Sobeckiej (PiS) z 22 października 2013 roku z debaty nad wnioskiem o pociągnięcie do odpowiedzialności konstytucyjnej przed Trybunałem Stanu członków KRRiTV; Stanisława Papieża z 26 maja 2004 roku z debaty w sprawie zlecenia Najwyższej Izbie Kontroli skontrolowania postępowań koncesyjnych KRRiTV; oraz Elżbiety Kruk z 27 czerwca 2012 roku z debaty nad sprawozdaniem KRRiTV za 2011 rok.

Odmówienie Telewizji Trwam prawa do nadawania na cyfrowym multipleksie było skandalem, bo przeczyło wolności mediów i ich pluralizmowi, a właśnie nad przestrzeganiem tych zasad ma czuwać Krajowa Rada Radiofonii i Telewizji. (Sobecka 2013)

Dlatego jest wielkim skandalem, że w Polsce w mediach elektronicznych, zdominowanych przez zagraniczne koncerny medialne, choć mówi się w języku polskim, nie odzwierciedla się ducha i aspiracji narodu polskiego, co najwyraźniej widać w eliminowaniu patriotycznego myślenia, programów historycznych, manipulowaniu informacjami i w nieobiektywnej publicystyce oraz w epatowaniu widza i słuchacza obcymi wzorcami i złudnymi iluzjami. (Papież 2004)

Skandaliczne jest w szczególności to, że udział w tym biorą media publiczne odpowiedzialne za budowę polskiej wspólnoty. (...) Wszak media czynią wiele, by ośmieszyć katolickie, narodowe czy patriotyczne wartości. (Kruk 2012)

Polityka zagraniczna staje się kontekstem ekstremalnej retoryki przede wszystkim w debatach poprzedzonych przemówieniami ministra spraw zagranicznych – przede wszystkim corocznej In-

formacji ministra spraw zagranicznych o głównych kierunkach polityki zagranicznej Polski w danym roku, ale również informacji o stanie negocjacji z Unią Europejską lub o stanie prac nad Traktatem Europejskim. Można wyróżnić dwa aspekty, w których oskarżenia o podłość lub zdradę pojawiają się najczęściej. Pierwszym z nich są odniesienia do przeszłych działań w polityce zagranicznej Polski, podejmowanych przez władze komunistyczne, które negatywnie wpłynęły na pozycję międzynarodową Polski i sytuację wewnętrzną kraju. Drugi aspekt odnosi się do procesu akcesji Polski do Unii Europejskiej, związanych z nim przygotowań oraz przemian polityki zagranicznej. Oskarżenia padają tutaj zarówno na poziomie bardzo ogólnym i dotyczą potencjalnej utraty suwerenności kraju, jak i przy okazji kwestii szczegółowych, przykładowo wynagrodzeń polskich europarlamentarzystów. Autorami poniższych, charakterystycznych dla tego tematu wypowiedzi są: Bronisław Geremek (UW, przemówienie z 5 marca 1998 r.), Hanna Mierzejewska (PiS, 16 czerwca 2004 r.) i Dariusz Lipiński (PO, 16 marca 2011 r.).

Były to działania motywowane podłością i głupotą, wyrażały brak szacunku dla Polski. (Geremek 1998)

Czy nie uważa pan minister, że zgoda na przyjęcie traktatu europejskiego, który ma być aktem nadrzędnym wobec Konstytucji Rzeczypospolitej Polskiej, w jego aktualnym brzmieniu będzie niczym innym jak tylko zdradą stanu? (Mierzejewska 2004)

Dlatego chciałbym apelować do wszystkich sił politycznych o współdziałanie w tym, aby polska prezydencja okazała się sukcesem, choć po wystąpieniu

pana przewodniczącego Błaszczaka, a zwłaszcza po haniebnym wystąpieniu pana posła Girzyńskiego wiem, że nie wszyscy na tej sali są zdolni do wzniesienia się ponad obelgi i partyjniackie zacietrzewienie. (Lipiński 2011)

Kwestie budżetowe, stanowiące następny temat, są kontekstem dla skrajnego słownictwa zwłaszcza przy omawianiu projektu ustawy budżetowej oraz jej realizacji. Posłowie wyrażają swoje oburzenie wobec konstrukcji budżetu i niesprawiedliwego lub niekorzystnego dla kraju, ich zdaniem, rozdysponowania środków znajdujących się w budżecie, a także niegospodarności ministrów w administrowaniu tymi środkami. Pojedynczą kwestią najsilniej pobudzającą skrajną retorykę są kwestie polityki emerytalnej państwa (w tym zabezpieczenia środków na przyszłe emerytury, zasady ich waloryzacji i kwestia wieku emerytalnego). Do tej kwestii właśnie odnoszą się trzy najbardziej charakterystyczne dla tego tematu przemówienia – autorstwa Tomasza Pacholskiego (RP) z 23 października 2012 roku, Włodzimierza Czechowskiego (Samoobrona) z 29 lipca 2005 roku oraz Waldemara Andzela (PiS) z 25 lipca 2012 roku – z których wyimki zaprezentowane zostały poniżej:

Jest to całkowicie skandaliczne, bo w Funduszu Rezerwy Demograficznej powinno dzisiaj znajdować się 31 mld zł, tymczasem jest tam tylko 17 mld zł. (Pacholski 2012)

Najbardziej haniebna była ostatnia nowelizacja ustawy emerytalnej przeprowadzona 16 września 2004 r. likwidująca coroczną waloryzację emerytur i rent i 20-procentowy dodatek wzrostu płac i zastąpienie jej

waloryzacją okresową co drugi lub trzeci rok, kiedy skumulowany wskaźnik inflacji średniorocznej osiągnie 105%. (Czechowski 2005)

Pieniądzy szukacie tylko poprzez wzrost podatków pobieranych od obywateli i skandaliczne podnoszenie wieku emerytalnego do 67 lat. (Andzel 2012)

Następnym w kolejności procentowego udziału w przemówieniach odwołujących się do idei hańby i zdrady jest temat służb specjalnych. Przemówienia zawierające oskarżenia padają w tym temacie w kontekście różnorodnych spraw, w których udział brały służby specjalne, między innymi tak zwanych spraw: moskiewskiej pożyczki, afery Olina, afery starachowickiej, porwania Krzysztofa Olewnika. Odnoszą się do nich również trzy charakterystyczne wypowiedzi z tego tematu, które wygłosili kolejno: Zbigniew Nowak (niezależny) 10 lipca 2003 roku, Sebastian Karpiniuk (PO) 11 lutego 2009 roku, Konstanty Miodowicz (PO) 23 października 2003 roku.

Czy prawdą jest to, że Henryk Długosz uzyskał jako pierwszy dostęp do zdradzonej tajemnicy państwowej, a następnie przekazał ją posłowi Jagielle? (Nowak 2003)

Stajemy dziś przed koniecznością wyjaśnienia okoliczności porwania i uprowadzenia, okoliczności wyjątkowo brutalnego morderstwa, ale – co równie bulwersujące – stajemy przed koniecznością wyjaśnienia, czy czasami przerażające i haniebne zaniedbania organów ścigania od końca 2001 r. wynikają jedynie z prymitywnej indolencji i ignorancji, czy może były one wynikiem – co zabrzmia przerażająco – umyślnego działania. (Karpiniuk 2009)

W haniebnym procederze złamania tajemnicy państwowej i służbowej, zdaniem prowadzącej śledztwo prokuratury, współpartycypowali poseł Henryk Długosz i były wiceminister spraw wewnętrznych pan Zbigniew Sobotka – koledzy partyjni Andrzeja Jagiełły. Lecz na tym nie koniec. Sprawa starachowicka zatacza coraz szersze kręgi. (Miodowicz 2003)

Temat sądownictwo jako kontekst występowania w przemówieniach słownictwa z analizowanego zbioru obejmuje szereg kwestii związanych z regulacjami prawnymi dotyczącymi trzeciej władzy. Obejmują one: tworzenie ustroju sądów różnego szczebla, statusu sędziego, postępowania sądowego, ale także rozliczenia osób, które naruszyły podstawowe standardy wykonywania zawodu w okresie 1944–1989. W trakcie ósmej kadencji Sejmu kwestią szczególnie wywołującą silnie negatywne emocje, wyrażane w radykalnym słownictwie, stał się Trybunał Konstytucyjny, w tym wybór niektórych sędziów Trybunału w sposób niezgodny z konstytucją, a także projekt ustawy o statusie sędziów Trybunału Konstytucyjnego. Przykłady przemówień charakterystycznych dla tego tematu to: przemówienie Krystyny Pawłowicz (PiS) z 17 marca 2016 roku w sprawie opinii Komisji Weneckiej dotyczącej nieprzestrzegania przez Sejm uchwał TK, wypowiedź Kazimierza Michała Ujazdowskiego z 3 lipca 1998 roku dotyczącą informacji o orzecznictwie TK w 1997 roku, a także przemówienie Ewy Lieder (.N) z 5 października 2016 roku podczas pierwszego czytania poselskiego projektu ustawy o statusie sędziów Trybunału Konstytucyjnego.

Komisja Wenecka akceptuje skandaliczne prawnie uchylanie się przez Trybunał Konstytucyjny od stosowania tej zasady, akceptując absurdalny, w cudzoślowie, standard odmienny, anarchizujący system prawny w Polsce, mianowicie taki, że prawo stanowione przez Sejm w Polsce jest z gruntu nielegalne od chwili uchwalenia i jest sprzeczne z konstytucją do czasu, aż łaskawie, dowolnie Trybunał Konstytucyjny uzna zgodność z konstytucją. (Pawłowicz 2016)

Rozumiem powody, dla których bronicie państwo dekretu o stanie wojennym, ale zważywszy na autorytet prawa i autorytet wymiaru sprawiedliwości sytuacja, w której polskie sądy wydają orzeczenia na podstawie dekretu o stanie wojennym, jest skandalem. (Ujazdowski 1998)

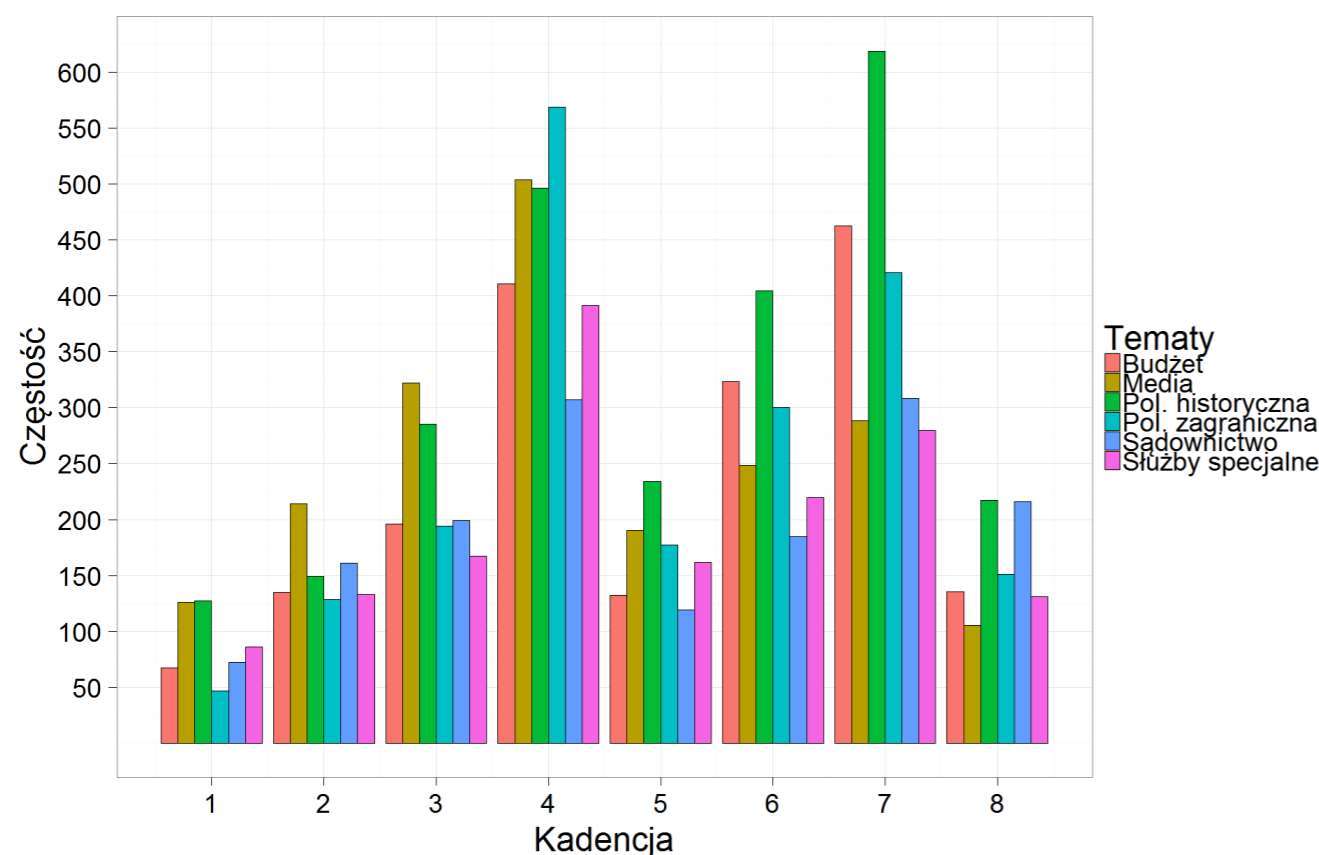
Szanowny Panie Marszałku! Wysoka Izbo! Po pierwsze, Nowoczesna składa wniosek o odrzucenie tej skandalicznej ustawy. (Lieder 2016)

Pozostałe, rzadziej występujące tematy odnoszą się do: prywatyzacji, publicznej służby zdrowia, edukacji oraz infrastruktury.

Dynamika kontekstów ekstremalnej retoryki w Sejmie

Częstość pojawiania się ekstremalnej retoryki w debatach parlamentarnych w poszczególnych kontekstach tematycznych jest zróżnicowana w czasie. W przypadku opisanych powyżej sześciu najczęściej występujących tematów częstość ich występowania zaprezentowana została na rysunku 6.

Rys. 6. Częstość występowania poszczególnych tematów jako kontekstów radykalnego słownictwa



Źródło: obliczenia własne. Wysokości słupków na wykresie dla poszczególnych kadencji zostały zrelatywizowane do liczebności wystąpień słów z badanego zbioru (wykres 1). Tak jak w przypadku poprzednich wykresów, obliczenia obejmują tylko ¼ VIII kadencji Sejmu.

Pierwsze trzy kadencje charakteryzują się dominacją dwóch tematów: mediów i polityki historycznej. Regulacja publicznych mediów, przede wszystkim telewizji, jako kontekst tematyczny pojawiania się retoryki ekstremalnej zwiększała swój udział w debatach sejmowych aż do IV kadencji, w której znalazła kumulację w postaci tak zwanej afery Rywina. Propozycja korupcyjna, którą Lew Rywin złożył Adamowi Michnikowi podczas prowadzonych w 2002 roku prac nad nowelizacją ustawy o radiofonii i telewizji, ujawniona przez prasę, spowodowała powołanie w tej sprawie komisji śledczej w styczniu 2003 roku i była wielokrotnie omawiana w Sejmie

przez ponad półtora roku. Raport komisji śledczej, przyjęty wbrew woli jej przewodniczącego, a także przegłosowanie na posiedzeniu Sejmu 28 maja 2004 roku całkowicie odmiennego raportu mniejszości autorstwa Zbigniewa Ziobry stało się źródłem skrajnych emocji wśród parlamentarzystów. W późniejszych kadencjach temat mediów pojawiał się z mniejszą częstotliwością, a w ostatniej kadencji stał się najrzadszym spośród analizowanych kontekstów.

Polityka historyczna oraz polityka zagraniczna swoje maksima uzyskały w kadencjach IV i VII

(choć pierwsza z nich tematem występującym najczęściej jako kontekst hańby była w całym okresie 2005–2015). W IV kadencji paliwem dla oskarżeń o niegodne zachowania na wielką skalę stał się proces akcesji Polski do Unii Europejskiej i związane z nim niezbędne modyfikacje prawa krajowego, przede wszystkim konieczność przekazania organom wspólnotowym części kompetencji władz krajowych. Część prawniczych ugrupowań parlamentarnych interpretowała akcesję jako utratę suwerenności i podporządkowanie się decyzjom kolejnego bloku, który porównywano z dawnym blokiem sowieckim.

Z kolei w VII kadencji wzrost częstości występowania polityki historycznej wynikał przede wszystkim z nasilającego się konfliktu między Prawem i Sprawiedliwością a Platformą Obywatelską, wszechogarniającego polską scenę polityczną i aktywizującego się silnie w debatach, w których ścierały się różne wizje interpretacji przeszłości. Natomiast polityka zagraniczna wzbudzała gwałtowne emocje w VII kadencji między innymi w związku z kryzysem uchodźczym w Unii i odgórną decyzją nakazującą krajom członkowskim przyjęcie ustalonych kwot uchodźców. Pomimo faktu, że wymóg wobec Polski był niski w porównaniu z innymi krajami członkowskimi, powszechna w społeczeństwie polskim niechęć do obcych etnicznie (Jasińska-Kania, Łodziński 2009) w połączeniu ze sprzeciwem wobec „samowoli Brukseli”, podsycanym, zwłaszcza w okresie kampanii wyborczej, przez większość liczących się ugrupowań, spowodowała nasilenie skrajnie negatywnych nastrojów w tym temacie w Sejmie.

Pozostałe tematy pojawiają się rzadziej jako konteksty hańby. Są one obecne w każdej kadencji, jednak w niektórych z nich pojawiają się z większą częstością. Temat służb specjalnych wywoływał negatywną retorykę przede wszystkim w okresie 2001–2005, ze względu na ujawnienie wówczas kluczowych afer politycznych, wymienionych powyżej. Tematy budżetowe i sądownictwa szczególnie często przyciągały ekstremalne słownictwo w trakcie IV i VII kadencji.

Zmianom kontekstu, a więc tematów, w których pojawiają się zarzuty hańby i zdrady, towarzyszą zmiany dystansu ideologicznego między partiami. Przyjmuję tu założenie, że partie posługujące się zbrutalizowaną retoryką w podobnych kontekstach tematycznych będą charakteryzowały się bliższymi ideologicznie stanowiskami niż partie, które używają jej w całkowicie odmiennych kontekstach. Korzystając z rezultatów modelu STM, jakimi są rozkłady prawdopodobieństwa na zbiorze słów określające każdorazowo ich szansę przynależenia do konkretnego tematu, definiuję odległość ideologiczną między partiami jako miarę różnicy między dwoma dyskretnymi rozkładami prawdopodobieństwa.

Miarą, z której korzystam, jest odległość Bhattacharyi (Bhattacharyya 1943), zgodnie z badaniami Ponti i Tagarelli (2009) dająca trafne wyniki dla modeli opartych na LDA. Zawarte w wykorzystywanym pakiecie STM rozkłady prawdopodobieństwa *a posteriori* wszystkich tematów dla każdej wypowiedzi ograniczone zostały wyłącznie do wybranych tematów, a następnie przeskalowane i uśrednione dla wszystkich tematów dla każdej

interesujący z tego powodu, że partia, po początkowym okresie radykalnych zmian pozycji ideologicznej, w kadencjach VI–VIII zachowuje niespotykaną wśród pozostałych ugrupowań spójność.

Widzimy również, że brutalizacja języka się rozszerza. O ile przez zdecydowaną większość analizowanego okresu dotyczyła ona przede wszystkim radykalnych partii prawicowych, to w latach 2015–2016 zwiększa się częstość wykorzystywania radykalnego słownictwa przez posłów Platformy Obywatelskiej.

Zaprezentowana w artykule analiza przemian kontekstów tematycznych radykalnego słownictwa w debatach parlamentarnych zbudowana została indukcyjnie na podstawie danych empirycznych stanowiących kompletną populację. Metoda, oparta na modelu generatywnym, nie wymaga tworzenia klucza kategoryzacyjnego ani kodowania tekstów ze względu na to, że kategorie budowane są automatycznie wyłącznie na podstawie dwóch elementów: częstości współwystępowania wyrazów ze sobą w poszczególnych dokumentach oraz metadanych dotyczących wypowiedzi i mówcy. W miejsce przypisywania słów do kategorii wykorzystywane jest przypisywanie słowom prawdopodobieństw przynależności do poszczególnych kategorii. Model budowany jest w sposób iteracyjny, a budowa zostaje zakończona, gdy kolejne porównania z danymi nie przynoszą już istotnych zmian w przypisanych słowom prawdopodobieństwach.

Tym samym podział na kategorie nie wymaga decyzji badacza określających powiązania między słowami, problematycznych ze względu na wieloznaczność sformułowań, często występujące zależ-

ności między kategoriami i zróżnicowanie odbioru przekazu w różnych grupach społecznych. Modelowanie tematyczne rozwiązuje zatem istotną bolączkę analizy treści, jaką jest silny wpływ ustalonych kategorii oraz ich składu na rezultaty. Poprzez brak pierwotnie ustalanych kategorii modelowanie tematyczne wypełnia warunek ograniczenia prekonceptualizacji badań na każdym etapie tworzenia modelu. Kategorie wyłaniają się z danych, nie podlegając wstępnemu teoretyzowaniu.

Techniki komputerowej pogłębionej eksploracji danych tekstowych (*text mining*) pozwalają na znacznie szybsze, oparte na dużych próbach i mniej podatne na błędy i wpływ badacza analizy. Wydajność obliczeniowa algorytmu sprawia, że analiza przeprowadzona została na całym zbiorze wypowiedzi zawierających badane słowa, a nie na wybranych przemówieniach, co z kolei likwiduje problemy z doбором próby. Analizę z wykorzystaniem strukturalnego modelu tematycznego charakteryzuje również wysoka rzetelność, przejawiająca się w stabilności rezultatów, powtarzalności wyników i dokładności pomiaru (Krippendorff 2004: 215–216). Model z zadanymi niezmiennymi parametrami, włącznie z ziarnem pseudolosowym (*seed*) generującym deterministycznie ustawienia inicjalizacji modelu, zwróci identyczne rezultaty. Jednocześnie rezultaty różnorodnych modeli tematycznych wygenerowanych na tym samym zbiorze danych wykazują wysoką zgodność rozbicia na kategorie. Z opisanych powodów modelowanie tematyczne stanowić może szybszą i bardziej rzetelną alternatywę wobec słownikowych metod kategoryzowania danych tekstowych lub też służyć do weryfikacji kategorii uzyskanych w sposób tradycyjny.

Podziękowania

Korpus przemówień sejmowych wykorzystany w artykule został opracowany jako część projektu autorki pt. „Przemiany rywalizacji partyjnej w Polsce 1991–2015 – analiza dyskusji i głosowań”, sfinansowanego w ramach

Bibliografia

Bates Stephen i in. (2014) *Questions to the Prime Minister: A Comparative Study of PMQs from Thatcher to Cameron*. „Parliamentary Affairs”, vol. 67, no. 2, s. 253–280.

Batko-Tokarz Barbara (2008) *Perswazja w dyskursie sejmowym*. Kraków: Scriptum.

Bhattacharyya Anil (1943) *On a Measure of Divergence between Two Statistical Populations Defined by Their Probability Distributions*. „Bulletin of the Calcutta Mathematical Society”, vol. 35, s. 99–109.

Bischof Jonathan M., Airoldi Edoardo M. (2012) *Summarizing Topical Content with Word Frequency and Exclusivity*. „Proceedings of the 29th International Conference on Machine Learning”, s. 201–208.

Blei David M., Ng Andrew Y., Jordan Michael I. (2003) *Latent Dirichlet Allocation*. „Journal of Machine Learning Research”, vol. 3, s. 993–1022.

Bralczyk Jerzy (2003) *O języku polskiej polityki lat osiemdziesiątych i dziewięćdziesiątych*. Warszawa: Trio.

Bralczyk Jerzy (2016) *Powszednia nienawiść. Czy radykalizacja języka polityki w Polsce może wyzwolić agresję zwykłych ludzi?* Rozm. przepr. Karolina Wigura. „Kultura Liberalna”, 20.04.2016 r.

Carmines Edward G., Stimson James A. (1986) *On the Structure and Sequence of Issue Evolution*. „American Political Science Review”, vol. 80, no. 3, s. 901–920.

Badań Statutowych Uniwersytetu SWPS 2014–2015. Chciałabym również podziękować za cenne uwagi uczestnikom panelu *Big Data, CAQDAS i nowe technologie w polu socjologii jakościowej* na XVI Ogólnopolskim Zjeździe Socjologicznym w Gdańsku, 14–17.09.2016, a także dwóm anonimowym recenzentom.

Dodd Lawrence C., Schraufnagel Scott (2012) *Congress and the Polarity Paradox, Party Polarization, Member Incivility and Enactment of Landmark Legislation*. „Congress & The Presidency”, vol. 39, s. 109–132.

Fischetti Tony (2015) *Data Analysis with R*. Birmingham: Packt Publishing.

Fridkin Kim L., Kenney Patrick J. (2004) *Do Negative Messages Work? The Impact of Negativity on Citizens' Evaluations of Candidates*. „American Politics Research”, vol. 32, s. 570–605.

Glaser Barney G., Strauss Anselm L. (2009) *Odkrywanie teorii ugruntowanej: strategie badania jakościowego*. Przełożyli Krzysztof Konecki i Marek Gorzko. Kraków: Nomos.

Hägglöf Hillevi (2014) *Automatic Organization of Online Swedish Political Discourse*. Rozprawa doktorska [dostęp 30 listopada 2016 r.]. Dostępny w Internecie: <<http://stp.lingfil.uu.se/exarb/master/hagglof2014.pdf>>.

Hańba [hasło w:] *Słownik języka polskiego PWN* [online] [dostęp 30 listopada 2016 r.]. Dostępny w Internecie: <<http://sjp.pwn.pl/slowniki/hanba.html>>.

Harris Sandra (2001) *Being Politically Impolite: Extending Politeness Theory to Adversarial Political Discourse*. „Discourse & Society”, vol. 12, no. 4, s. 451–472.

Ilie Cornelia (2004) *Insulting as (Un) Parliamentary Practice in British and Swedish Parliaments*. „Cross-Cultural Perspectives on Parliamentary Discourse”, vol. 26, no. 2, s. 45–68.

Jamieson Kathleen H. (2001) *Civility in the House of Representatives: The 106th Congress*. Report Series The Annenberg Public Policy Center of the University of Pennsylvania [dostęp 30 listopada 2016 r.]. Dostępny w Internecie: <http://www.annenbergpublicpolicycenter.org/Downloads/Civility/Old-reports/2001_civility106th.pdf>.

Jasińska-Kania Aleksandra, Łodziński Sławomir (2009) *Obszary i formy wykluczenia etnicznego w Polsce. Mniejszości narodowe, imigranci, uchodźcy*. Warszawa: Scholar.

Kalmoe Nathan P. (2014) *Fueling the Fire: Violent Metaphors, Trait Aggression, and Support for Political Violence*. „Political Communication”, vol. 31, no. 4, s. 545–563.

Kamińska-Szmaj Irena (2001) *Słowa na wolności: język polityki po 1989 roku: wypowiedzi, dowcip polityczny, słownik inwektyw*. Wrocław: Europa.

Krippendorff Klaus (2004) *Content Analysis. An Introduction to its Methodology*. Thousand Oaks: Sage.

Krupnikov Yanna (2011) *When Does Negativity Demobilize? Tracing the Conditional Effect of Negative Campaigning on Voter Turnout*. „American Journal of Political Science”, vol. 55, no. 4, s. 797–813.

Krupnikov Yanna (2012) *Negative Advertising and Voter Choice: The Role of Ads in Candidate Selection*. „Political Communication”, vol. 29, no. 4, s. 387–413.

Laskowska Elżbieta (2004) *Dyskurs parlamentarny w ujęciu komunikacyjnym*. Bydgoszcz: Wydawnictwo Akademii Bydgoskiej.

Lau Richard R. i in. (1999) *The Effects of Negative Political Advertisements: A Meta-Analytic Assessment*. „American Political Science Review”, vol. 93, no. 4, s. 851–875.

Lau Richard R., Sigelman Lee, Rovner Ivy B. (2007) *The Effects of Negative Political Campaigns: A Meta-Analytic Reassessment*. „Journal of Politics”, vol. 69, no. 4, s. 1176–1209.

Maisel Sandy L. (2012) *The Negative Consequences of Uncivil Political Discourse*. „PS: Political Science & Politics”, vol. 45, no. 3, s. 405–411.

Moser Scott, Reeves Andrew (2014) *Taking the Leap: Voting, Rhetoric, and the Determinants of Electoral Reform*. „Legislative Studies Quarterly”, vol. 39, no. 4, s. 467–502.

Murphy James (2014) *(Im)Politeness during Prime Ministers Questions in the UK Parliament*. „Pragmatics and Society”, vol. 5, no. 1, s. 76–104.

Mutz Diana C. (2015) *In-Your-Face Politics. The Consequences of Uncivil Media*. Princeton: Princeton University Press.

Piniarski Arkadiusz (2011) *Język polskiej debaty parlamentarnej*. Rozprawa doktorska [dostęp 30 listopada 2016 r.]. Dostępny w Internecie: <<https://repozytorium.amu.edu.pl/bitstream/10593/2705/2/Arkadiusz%20Piniarski%20J%20C4%99zyk%20polskiej%20debaty%20parlamentarnej%20T1.pdf>>.

Piotrowski Andrzej (1997) *Wprowadzenie* [w:] Marek Czyżewski, Sergiusz Kowalski, Andrzej Piotrowski, red., *Rytualny chaos. Studium dyskursu publicznego*. Kraków: Wydawnictwo Aureus, s. 28–38.

Polkowska Laura (2015) *Język prawicy*. Warszawa: Wydawnictwo UKSW.

Ponti Giovanni, Tagarelli Andrea (2009) *Topic-Based Hard Clustering of Documents Using Generative Models* [w:] Zbigniew Raś, Setsuo Ohsuga, eds., *Proceedings of the 18th International Symposium on Foundations of Intelligent Systems*. Berlin, Heidelberg: Springer-Verlag, s. 231–240.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. [dostęp 30 listopada 2016 r.]. Dostępny w Internecie <<http://www.R-project.org>>.

Roberts Margaret E. i in. (2014) *Structural Topic Models for Open-Ended Survey Responses*. „American Journal of Political Science”, vol. 58, no. 4, s. 1064–1082.

Roberts Margaret E., Stewart Brandon M., Tingley Dustin (2016) *Package ‘stm’. Estimation of the Structural Topic Model Version 1.1.3* [dostęp 30 listopada 2016 r.]. Dostępny w Internecie: <<https://cran.r-project.org/web/packages/stm/stm.pdf>>.

RStudio Team (2015) *RStudio: Integrated Development for R*. Boston: RStudio, Inc. [dostęp 30 listopada 2016 r.]. Dostępny w Internecie <<http://www.rstudio.com>>.

Semetko Holli A., Schoenbach Klaus (2003) *News and Elections: German Bundestag Campaigns in the Bild, 1990-2002*. „Harvard International Journal of Press/Politics”, vol. 8, no. 3, s. 54–69.

Sparo Carole (2010) *Disrupting Rituals of Debate in the Indian Parliament*. „The Journal of Legislative Studies”, vol. 16, no. 3, s. 338–351.

Tyszka Tadeusz (1996) *Debaty sejmowe: dochodzenie do decyzji czy konkurs uzasadnień?* „Studia Psychologiczne”, t. 34, nr 2, s. 67–83.

Urzykowski Tomasz (2016) *„Hańba” na obchodach rocznicy bitwy pod Olszynką Grochowską*. „Gazeta.pl”, 27.11.2016 r. [dostęp 30 listopada 2016 r.]. Dostępny w Internecie: <<http://wyborcza.pl/1,75248,21038911,hanba-na-obchodach-rocznicy-bitwy-pod-olszynka-grochowska.html>>.

Uslaner Eric M. (1993) *The Decline of Comity in Congress*. Ann Arbor: University of Michigan Press.

Uslaner Eric M. (2000) *Is the Senate More Civil than the House?* [w:] Burdette Loomis, ed., *Esteemed Colleagues: Civility*

and Deliberation in the U. S. Senate. Washington: Brookings, s. 32–55.

Walczak Bogdan (1994) *Co to jest język polityki?* [w:] Janusz Anusiewicz, Bogdan Siciński, red., *Język a kultura*, t. 11, *Język polityki a współczesna kultura polityczna*. Wrocław: TPPW, s. 15–20.

Winczorek Jan (2014) *Wykorzystanie oprogramowania R i RQDA w jakościowo-ilościowej analizie treści orzeczeń Trybunału Konstytucyjnego*. „Przegląd Socjologii Jakościowej”, t. 10, nr 2, s. 138–161.

Wolf Michael R., Strachan Cherie J., Shea Daniel M. (2012) *In-civility and Standing Firm: A Second Layer of Partisan Division*. „PS: Political Science and Politics”, vol. 45, no. 3, s. 428–434.

Zirn Cäcilia, Stuckenschmidt Heiner (2014) *Multidimensional Topic Analysis in Political Texts*. „Data and Knowledge Engineering”, vol. 90, s. 38–53.

Cytowanie

Kwiatkowska Agnieszka (2017) „Hańba w Sejmie” – zastosowanie modeli generatywnych do analizy debat parlamentarnych. „Przegląd Socjologii Jakościowej”, t. 13, nr 2, s. 82–109 [dostęp dzień, miesiąc, rok]. Dostępny w Internecie: <www.przegladsocjologiijakoosciowej.org>.

“Disgrace in the Polish Parliament”—analysis of parliamentary debates using generative models

Abstract: Over the last decade there has been a clear increase in extreme political rhetoric in the parliamentary debate in Poland. One aspect of the progressive radicalization of the language is overuse of words having a high negative emotional intensity to describe events in everyday politics, which leads to the degradation of their importance. As research shows, brutalization of the language of the debate has negative consequences at the levels of the political elite and the electorate, leading to reduced efficiency of the parliament, the exacerbation of conflicts between parties and among their supporters, and the political alienation of citizens.

In this article, I present the possibilities of using the transcripts of parliamentary discourse to analyze the changes of the political discourse, including the algorithms used for searching text corpora and analysis of the latent topics. As an example, I use a set of parliamentary speeches from the years 1991-2016 related to the idea of shame, betrayal, disgrace and scandal. I base my analyses on generative topic modeling employing on the method of latent Dirichlet allocation and its extension—Structural Topic Model, both being unsupervised methods of extracting topics from large text corpora. The results of the analysis confirm the increase in usage of radical vocabulary in time, and also describe the characteristics of its users and the most common contexts the extreme vocabulary tend to appear (identity politics, control over the media, foreign policy).

Keywords: Poland, Parliamentary Debates, Political Elites, Quantitative Content Analysis, Text Mining

Alicja Zawistowska, Małgorzata Skowrońska Uniwersytet w Białymstoku

Dobra zmiana czy Polska w ruinie?

Abstrakt Celem artykułu jest analiza ewolucji znaczeniowej wpisów opatrzonych hashtagami #dobrazmiana i #polskawruinie zamieszczonych w serwisie społecznościowym Twitter. Oba hasła pojawiły się w kampanii parlamentarnej PiS w roku 2015 i szybko zyskały popularność, ale ich późniejsze kariery były odmienne. Analiza 494 wpisów z czterech okresów z lat 2015–2016 pokazuje, jak zmieniało się zabarwienie emocjonalne obu haseł. Hashtag #dobrazmiana tracił stopniowo swoje początkowe – pozytywne znaczenie nadane mu w kampanii parlamentarnej. Z upływem czasu nadawano mu ironiczny, prześmiewczy kontekst. Hasło „Polska w ruinie” funkcjonowało krócej, a jego pierwotny sens został przez twitterowiczów błyskawicznie odwrócony. Analiza prowadzi do wniosku, że hasła, które początkowo miały jasno zdefiniowany sens, zostały przez użytkowników portalu przekształcone zgodnie z panującym na Twitterze stylem komunikacji – szybkim i mocnym przekazem. W konsekwencji konteksty używania haseł w pewnym stopniu różnicowały się, ale służyć zaczęły przede wszystkim do wyrażania niezadowolenia. Przeprowadzone badanie pozwoliło także zidentyfikować kilka metodologicznych problemów związanych z analizą treści w mediach społecznościowych.

Słowa kluczowe Twitter, hashtag, dobra zmiana, Polska w ruinie, emocje

Alicja Zawistowska, dr, pracuje w Instytucie Socjologii i Kognitywistyki Uniwersytetu w Białymstoku. Interesuje się metodologią badań socjologicznych i edukacją.

Małgorzata Skowrońska, mgr, pracuje w Instytucie Socjologii i Kognitywistyki Uniwersytetu w Białymstoku. Jej zainteresowania badawcze to badania dystansu społecznego i socjologia kultury.

Adres kontaktowy:

Wydział Historyczno-Socjologiczny
Instytut Socjologii i Kognitywistyki
Uniwersytet w Białymstoku
ul. Plac Uniwersytecki 1, 15-403 Białystok
e-mail: zawistowska.alicja@gmail.com

Twitter jest portalem społecznościowym, który umożliwia dzielenie się krótkimi informacjami z innymi użytkownikami Internetu. Jego historia jest krótka. Pierwszy wpis na Twitterze został wysłany przez jednego z twórców portalu, Jacka Dorsey, 21 marca 2006 roku. Od tamtego czasu liczba użytkowników Twittera systematycznie wzrasta. W lipcu 2016 roku korzystało z niego 313 milionów osób, z których niemal 80% pochodziło spoza USA (Twitter 2016). Twitter stanowi przestrzeń, w której – zgodnie z założeniami twórców – użytkownicy dzielą się swoimi przemyśleniami, a wiadomości te rozprzestrzeniają się w niezwykle szybkim tempie. Jedynym ograniczeniem jest objętość wpisu, który nie może przekraczać 140 znaków. Wężłowość for-

my nie przeszkadza jednak wirtualnej społeczności na całym świecie wysyłać ponad 500 milionów wiadomości dziennie i około 6000 na sekundę (Twitter 2016). Z tego między innymi powodu Twitter jest mikroblokiem, na którym chętnie publikują swoje wpisy i zdobywają popularność wpływowi publicyści, politycy czy celebryci. Przepływające w sieci informacje uporządkowane za pomocą hashtagów, czyli krótkich haseł poprzedzonych symbolem #. Funkcjonują tysiące takich etykiet, którymi użytkownicy oznaczają swoje wpisy. Niektóre z nich uzyskują ogromną popularność, a wraz z nimi rozgłosu nabierają oznaczone nimi wydarzenia lub osoby, inne z kolei szybko są zapominane. Kategoryzacja treści za pomocą hashtagów potrafi jednak z czasem zmieniać swój sens tak, że treści, które się za nią kryją, podlegają dużej zmianie. Celem niniejszego artykułu jest przedstawienie ewolucji znaczeniowej wpisów zamieszczonych w serwisie społecznościowym Twitter oznaczonych hashtagami: #dobrazmiana i #polskawruinie w czterech wybranych okresach z lat 2015–2016.

#ThankYouChrisMessina

Jako pierwszy hashtagi wykorzystywać zaczął Facebook, ale nawyk oznaczania i segregowania treści umieszczanych w sieci w największym stopniu wykształcił się wśród użytkowników Twittera. Z Twittera właśnie pochodzi pierwszy wpis opatrzony krzyżykiem, a umieścił go tam w sierpniu 2007 roku związany z Google Chris Messina. Messina rozesłał do swoich przyjaciół wiadomość opatrzoną niepozornym krzyżykiem z doklejonym wyrazem-kategorią, sugerując, że taki sposób formułowania treści na Twitterze ułatwia ich szybkie rozpoznawanie,

przeoglądanie i sortowanie według wątków. Trzy lata później sposób ten był już na tyle rozpowszechniony, że założyciel Social Media Marketing Institute, Ginger Wilcox, ogłosił rok 2010 rokiem hashtagu (Parker 2011). Dzisiaj hashtag jest formułą na tyle rozpoznawaną, że masowo korzystają z niej agencje marketingowe, oznaczając w sieci swoje produkty czy usługi.

Używanie hashtagów stało się tak popularne wśród „zwykłych” twitterowiczów dlatego, że obok funkcji porządkującej, którą pełni kratka ze słowem-kluczem, znacząco wpływa ona także na dynamikę dyskusji w sieci; na szybki i żywiołowy przepływ opinii. Użytkowników Twittera zainteresowanych publiczną debatą nad aktualnymi sporami politycznymi, wydarzeniami sportowymi czy kulturalnymi hashtag wciąga do dyskusji, ponieważ umożliwia „równoległe” komentowanie tego, co właśnie oglądane jest za pośrednictwem innych mediów. Użytkownicy Twittera, by wyrazić swoją opinię, mają do dyspozycji 140 znaków, które, jak twierdzi blogerka *New Yorkera*, Susan Orlean (2010), właśnie dzięki hashtagom mogą stanowić kompleksową i bogatą wiadomość wysyłaną do świata. Twitterowy „hash” pozwala nadać tej wiadomości określony wydźwięk. Pozwala także wiele kwestii pozostawić w sferze domysłu i niedopowiedzenia. Z biegiem czasu hashtagi różnicowały swoje funkcje. Z jednej strony nastąpiło pewne ich uszczegółowienie, czyli sytuacja, kiedy użytkownicy przy pomocy hashtagów precyzyjnie i na bieżąco rejestrują zmiany, nowe konteksty opisywanych wydarzeń. Z drugiej zaś strony obok precyzyjnych „szufladek” pojawiły się także szerokie kategorie znaczeniowe, często o abstrakcyjnym charakterze, umożliwiające tworzenie

wielowarstwowych skojarzeń. Na przykład kiedy odbywały się igrzyska w Soczi, pierwsze wpisy dotyczące zimowej olimpiady oznaczane były hashtagem #sochi2014. Kiedy jednak na jaw wyszły problemy organizacyjne związane z imprezą, popularność zyskał hashtag #sochiproblems, którym oznaczano wiele wpisów umieszczanych długo po zakończeniu igrzysk, tylko pośrednio związanych z tym wydarzeniem. Na przykład wysoki rachunek za telefon otrzymany przez jednego z twitterowiczów doczekał się wpisu: „Przyszedł rachunek za komórkę z pobytu w Soczi – 127 funtów” (Chaciński 2016). Należy podkreślić również, że stosowanie hashtagów jako pewnych ciągów skojarzeń luźno powiązanych z treścią postu dosyć szybko uruchomiło komiczny i (auto)ironiczny potencjał symbolicznej kratki. Dobrze pokazuje to przykład niezwykle dziś popularnego określenia #firstworldproblems używanego przez twitterowiczów chcących podzielić się ze światem trywialnymi spostrzeżeniami. Ów hashtag pozwala im czynić to bez obaw i oskarżeń o próżność. Przykładają przecież odpowiednią miarę do sytuacji, którą zdecydowali się opisać, i informują o dystansie do siebie. Właśnie owo drugie dno, którego dostarczają wpisom rozmaite hashtagi, wydaje się najbardziej interesujące, kiedy myślimy o analizie treści tweetów. Jednocześnie ten żartobliwy, ironiczny czy abstrakcyjny charakter hashtagów następcza badaczom największej trudności interpretacyjnych.

Inną funkcją hashtagów jest tworzenie tymczasowych wspólnot skupionych wokół pewnych wydarzeń. Jako przykład podać można hasła #JeSuisCharlie czy #PrayForParis, które błyskawicznie spopularyzowały się na Twitterze tuż po zamachach terro-

rystycznych w Paryżu w 2015 roku (zob. Adamska 2015). Rekordowe pod względem popularności hashtagi, takie jak ten użyty po ataku na siedzibę francuskiej gazety, który w krótkim czasie został użyty ponad 5 milionów razy, zarezerwowane są jednak tylko dla wydarzeń skupiających uwagę dużej części międzynarodowej opinii publicznej. Globalne zainteresowanie jest bowiem jedną rzeczą, która połączyła tak popularne, a zarazem rozbieżne znaczeniowo hashtagi jak #WorldCup, którym oznaczano mistrzostwa świat w piłce nożnej i #RIPPrince, który w dniu śmierci artysty pojawił się 6 milionów razy (Twitter 2016).

Cechą charakterystyczną wspomnianych hashtagów jest ich samorzutne powstawanie – tworzone były bowiem spontanicznie i stanowiły reakcję na wydarzenia mające miejsce „tu i teraz”. Oczywiście jest jednak, że w przypadku hashtagów tworzonych na przykład przez organizacje promujące pewne idee czy produkty lub partie polityczne stopień ich rozprzestrzeniania i żywotności jest ściśle uzależniony od działań marketingowych. Wśród najbardziej chwytliwych haseł wyborczych stanowiących element działań z zakresu *public relations* wymienić można choćby sławne #YesWeCan, które na Twitterze pojawiło się po raz pierwszy w kwietniu 2008 roku w trakcie wyborów prezydenckich w USA. Uważa się zresztą, że wygrana kampania Baracka Obamy była pierwszą, która wykorzystwała potencjał mediów społecznościowych, a samo hasło okazało się niezwykle nośne i przez długi czas używane było w różnych kontekstach niezwiązanych z polityką. Właśnie możliwość zastosowania hasła w odniesieniu do zróżnicowanych spraw może być ważnym czynnikiem wpływającym na utrzymanie jego popularności.

Sieci społecznościowe, hashtagi i nastroje

Podobnie jak w przypadku innych kanałów komunikacji, wymiana informacji na Twitterze również kieruje się pewnymi prawidłowościami. Do interesujących wniosków doszły badaczki Sarita Yardi i Danah Boyd (2010), które analizowały reakcje emocjonalne użytkowników amerykańskiego Twittera na zabójstwo lekarza specjalizującego się w tak zwanych późnych aborcjach, George’a Tiller. W maju 2009 roku został on zastrzelony w stanie Kansas przez aktywistkę ruchu antyaborcyjnego Scotta Roedera. Badaczki zgromadziły 30 000 wpisów dotyczących tego zdarzenia, pojawiających się w serwisie w ciągu doby od zabójstwa i wyodrębniły spośród nich 1 447 postów do analizy. Wybrane przez nie tweety posiadać musiały chociaż jedną odpowiedź-reakcję innego twitterowicza. Jak ustaliły, na posty reagowały głównie 2 typy użytkowników: zwolennicy prawa do dokonywania aborcji (często oznaczający posty hashtagiem #prolife) oraz przeciwnicy takiego rozwiązania (oznaczający swoje wpisy jako #prochoice). Przede wszystkim ustalono, że użytkownicy znacznie chętniej reagovali na posty zgodne z ich przekonaniami niż na te, które stoją z ich poglądami w sprzeczności. Dowodzić to może tego – jak sugerują same autorki badania – że w sytuacji, kiedy na Twitterze debatuje się nad problemami społecznymi, na przykład aborcją, często mamy do czynienia z interakcjami osób „podobnie myślących” (Yardi, Boyd 2010: 13). Rzadziej pojawiały się tu interakcje pomiędzy oponentami, chociaż ten typ konwersacji także angażował użytkowników. Kiedy do tego rodzaju wymiany zdań dochodziło, w przeważającej mierze koncentrowała się ona na dwóch wątkach. Użytkownicy #prochoice podkreślali sprzeczność pomiędzy prezentowaniem postawy *pro-*

life, zgodnie z którą cenić się powinno każdy przejaw życia, i zastrzeleniem człowieka. Z kolei użytkownicy #prolife swoje posty poświęcali głównie wyjaśnianiu przekonań i precyzowaniu wartości towarzyszących postawie *prolife* oraz udowadnianiu, że zabójca takiej postawy nie reprezentował.

Analizowana debata dotycząca aborcji pokazuje pewną specyfikę Twittera. Większa chęć podejmowania debaty z tymi, których punkt widzenia podzielamy, sprawia, że w ramach serwisu tworzą się enklawy zgadzających się ze sobą. Taka sytuacja może wprawdzie, jak twierdzą badaczki, przynosić pewne profity – chociażby zmniejszać poziom agresji wyrażanej przez tweetujących, niemniej może jednocześnie produkować „duszną jednorodność”, która w miarę upływu czasu może antagonizować relacje między myślącymi odmiennie.

Badanie przeprowadzone przez Yardi i Danah Boyd (2010) pokazuje, że zabójstwo George’a Tiller skłoniło użytkowników Twittera do ekspresji zróżnicowanych emocji – od złości, przez smutek po empatię. Jak ustaliły badaczki, poziom smutku pozostawał relatywnie niezmienny, podczas gdy poziom pozostałych emocji zmieniał się w czasie. Jak piszą autorki, Twittera częściej używa się do wyrażania niezadowolenia czy złości niż do wyrażania wsparcia, w związku z tym nie może zaskakiwać fakt, że analizowane wpisy zawierały znacznie więcej wiadomości o negatywnym ładunku emocjonalnym. Jak wykazały badaczki, poziom emocji negatywnych wzrastał wraz z upływem czasu. Znamienne, że wystarczyły im 24 godzinny następujące po zabójstwie, aby zebrać materiał wystarczający do analizy zmiennej temperatury emocji.

Badaczki dzielą się w swoim artykule także ciekawym spostrzeżeniem dotyczącym specyfiki samego Twittera. Twierdzą bowiem, że fakt angażowania się użytkowników w konwersacje spójne ideologicznie i skupione wokół wspólnych przekonań oraz rezygnacja z dyskusji z ideologicznymi oponentami wynikać może z pewnych ograniczeń serwisu. Autorki wskazują przykłady tweetów, w ramach których użytkownicy wyraźnie sfrustrowani są ograniczeniami technologicznymi tego medium (głównie liczbą znaków w pojedynczym poście oraz tendencyjnym obrazem zdarzenia, którego posty dotyczą), ponieważ uniemożliwia im to rzeczową dyskusję. Tweety wymagają bowiem skrótów myślowych, uogólnień, prostych komunikatów i tym podobnych. Badaczki sugerują więc, że sam Twitter jako medium może wpływać na formowanie się zwartych, jednorodnych enklaw zwolenników poszczególnych przekonań. Pozornie bowiem serwis umożliwia zetknięcie się z różnymi punktami widzenia i opiniami, daje jednak niewiele możliwości na prowadzenie konwersacji deliberacyjnych – szczegółowych, problematyzujących omawiane zdarzenie, ponieważ preferuje pośpiech i emocje (Yardi, Boyd 2010: 25).

Powyższe badanie stanowi też ilustrację sposobu rozprzestrzeniania się emocji w serwisach społecznościowych. Interesujące wnioski na ten temat płyną z eksperymentu przeprowadzonego przez zespół pod kierunkiem Kramer (2014), przeprowadzonego wśród 700 000 użytkowników Facebooka. Badanie polegało na filtrowaniu treści widocznych na tablicy użytkownika poprzez usuwanie lub dodawanie wpisów o różnym zabarwieniu emocjonalnym i śledzeniu reakcji na te manipulacje. Mechanizm spo-

łecznego zarażania się emocjami okazał się niezwykle prosty. Z badania wynikało, że ludzie, którzy na swoich tablicach wiedzieli mniej wpisów, zdjęć czy filmów o pozytywnym wydźwięku, sami wytwarzali mniej pozytywnych treści, a częściej tworzyli wpisy nacechowane negatywnie. Natomiast, gdy na skutek celowej „cenzury” użytkownik widział mniejszą liczbę wpisów o negatywnym zabarwieniu, sam również tworzył takich wpisów mniej. Wyniki tego eksperymentu dowodzą, w jak łatwy sposób użytkownicy Internetu mogą być emocjonalnie „programowani” przez prosty zabieg selekcji treści zamieszczonych na ich własnych tablicach, które stwarzają pozór „prywatnej” przestrzeni. Takie możliwości środków masowego przekazu nie są oczywiście niczym nowym – przypomina o tym choćby panika wywołana przez słynną audycję radiową o ataku kosmitów z 1937 roku. Media społecznościowe okazały się po prostu kolejnym kanałem, w którym sprawdza się ta sama metoda.

Po przekroczeniu pewnego progu rozpowszechnienia treści ludzie „zarażają się” nie tylko nastrojem, ale są również bardziej skłonni eksponować własne poglądy. Zachowanie takie zostało uchwycone w badaniu społecznościowej akcji poparcia dla małżeństw homoseksualnych zainicjowanej w 2013 roku na Facebooku. Polegała ona między innymi na zmianie zdjęcia profilowego na jeden z istniejących wtedy symboli poparcia. State i Adamic (2015), którzy przeprowadzali owe badanie, twierdzą, że zanim użytkownicy zmienili swoje zdjęcie profilowe, najpierw obserwowali, czy zrobią to ich znajomi. Analizy pokazały, że wystarczyło maksymalnie 6 znajomych z nowym zdjęciem, aby badany użytkownik również dokonał zmiany (State, Ada-

mic 2015). Na zagęszczenie sieci społecznych jako czynnika zwiększającego popularność niektórych hashtagów wskazali też Ma, Sun i Cong (2013). Argumentowali, że szybciej będą się rozprzestrzeniały hasła, które cyrkulują wśród powiązanych ze sobą użytkowników tworzących subspołeczności (Ma, Sun, Cong 2013). Przykładem może być choćby różnica między wspomnianym wcześniej #firstworldproblems, który wymaga posiadania wiedzy o specyficznej konwencji, w jakiej występuje to hasło, a uniwersalnym #rio odnoszącym się do Olimpiady w tym mieście.

Użytkownicy Twittera

Analiza treści tweetów w oczywisty sposób wiąże się też z koniecznością przyjrzenia się samym użytkownikom Twittera. Częstkowej charakterystyki polskich twitterowiczów dostarcza Megapanel PBI/Gemius. W opracowaniu badań zamieszczonym przez wirtualnemedial.pl czytamy, że w badanych latach Twitter cieszył się rosnącym zainteresowaniem Polaków (Wirtualne Media 2016). W okresie od 2010 do 2013 roku liczba użytkowników serwisu wzrosła trzykrotnie – w listopadzie 2013 roku wynosiła 2,64 miliona. Dane z roku 2015 (Gemius 2015) ponownie pokazują wzrost społeczności polskich twitterowiczów – liczy ona wówczas już 3,3 miliona osób (4 mln wg Sotrender 2015b). W roku 2015 najliczniejszą kategorię użytkowników stanowiły osoby w wieku od 25 do 34 lat, choć zanotowano także nieznaczny wzrost wśród najstarszych użytkowników (50+). Z Twittera nieco chętniej korzystali mężczyźni, którzy stanowili 55% użytkowników. Jeśli szukać specyfiki zawodowej, to największy odsetek twitterowiczów stanowią dziennikarze, publi-

cyści, szefowie korporacji i przedstawiciele wolnych zawodów oraz uczniowie i studenci. Korzystanie z serwisu deklarują najczęściej osoby z wyższym wykształceniem (Sotrender 2015b).

O zainteresowaniach polskich twitterowiczów pośrednio wnioskować można natomiast na podstawie liczby najbardziej popularnych użytkowników. Wśród nich w 2015 roku byli muzycy popowi, prawnicy dziennikarz sportowy, a zaraz po nich prawnicy publicyści tacy jak Katarzyna czy Samuel Pereira (Sotrender 2015a). Wśród najczęściej wspomnianych użytkowników byli natomiast, oprócz wymienionych osób, Andrzej Duda oraz stacja TVN24. Największe profile – wyrażone liczbą obserwujących – należały natomiast do stacji muzycznej MTV oraz TVN24, a wśród najbardziej poczytnych polityków znajdował się Radosław Sikorski. Jeśli potraktować te dane jako wskaźniki treści najbardziej przyciągających uwagę twitterowiczów, to należałyby do nich polityka, popkultura oraz sport. Prawdopodobnie odzwierciedla to nie tylko zainteresowania użytkowników, ale również ich strukturę wieku. Ostrożnie należy wypowiadać się jednak o bardziej szczegółowych preferencjach i gustach całej społeczności twitterowej, bo te zmieniają się w zależności od bieżących wydarzeń oraz aktualnie popularnych newsów. Przykładowo, w okresie ważnych meczów piłki nożnej rosnącą popularnością cieszą się profile konkretnych piłkarzy.

Oczywistym jest także, że nie wszyscy użytkownicy polskiego Twittera mają tę samą rangę. Sam Twitter jest raczej tubą przydatną do ekspresji przekonań czy opinii i nie pełni funkcji medium stricte informacyjnego. Można wyznaczyć trzy sfery

rozprzestrzeniania się tych opinii, a z poszczególną sferą rozpowszechniania powiązać typy użytkowników. Sfera pierwsza obejmuje przede wszystkim publikowanie informacji i powiązać z nią można oficjalne profile partii politycznych, a także profile redakcji telewizyjnych, radiowych i prasowych. Tweety użytkowników tego typu charakteryzuje nieco mniejszy ładunek emocjonalny. Jeśli idzie o treść, ich wpisy są zwykle spójne z linią ideową partii czy redakcji. Sfera druga związana jest już raczej z „zasiewaniem” opinii, a nie informowaniem. Obejmuje ona użytkowników spersonalizowanych, występujących pod imieniem i nazwiskiem lub ewentualnie rozpoznawalnym pseudonimem. Wpisy takich użytkowników są w większym stopniu nacechowane emocjonalnie, znacznie częściej zawierają krytyczny komentarz do przedstawianych zdarzeń i artykułują wprost poglądy autorów. Użytkownicy spersonalizowani to dziennikarze – zarówno ci związani z poszczególnymi redakcjami, jak i tak zwani dziennikarze obywatelscy, publicyści, blogerzy, komentatorzy życia publicznego, ale także celebryci – osoby kojarzone z show-biznesem. Ostatnia sfera obejmuje przede wszystkim przechwytywanie opinii i związana jest ze „zwyczajnymi” twitterowiczami. Taki typ użytkownika włącza się do dyskusji poprzez kontestowanie lub potwierdzanie opinii przechwyconej od użytkowników spersonalizowanych. Sam jednak opinii nie publikuje. Część zwyczajnych użytkowników pozostaje zupełnie bierna, to jest nie umieszcza autorskich tweetów, a jedynie odnosi się do wpisów użytkowników oficjalnych i spersonalizowanych czy też wyłącznie ich obserwuje. Zakładamy, że użytkowników spersonalizowanych uznawać można za twitterowych

liderów opinii i że to właśnie oni w największym stopniu wpływają na nastroje polityczne zwyczajnych użytkowników. Zakładamy także, że użytkownicy spersonalizowani są najaktywniejszymi uczestnikami debaty i w największym stopniu wpływają na jej dynamikę, choćby z uwagi na fakt cyklicznej aktywności w sieci, liczby umieszczanych wpisów oraz reakcji na tę aktywność, mierzoną liczbą polubień i udostępnień.

Twitter jest też miejscem, z którego chętnie korzystają politycy, szczególnie w okresie wyborczym. Z analiz przeprowadzonych w okresie wyborów parlamentarnych w 2015 roku wynika, że konta na tym portalu mieli najczęściej kandydaci PO (105 osób), potem PiS (61), a następnie przedstawiciele Zjednoczonej Lewicy (46) i PSL (25) (Deryło, Makarenko 2015). Twitter staje się zatem popularnym środkiem komunikacji politycznej, ale tylko między niewielką frakcją polityków i ich elektoratem.

Metodologia badania

Badanie emocjonalnego zabarwienia treści zamieszczanych na portalach społecznościowych stanowi dla badaczy metodologiczne wyzwanie. Najwięcej kłopotu badaczom nastęrcza wielowymiarowość języka naturalnego, obecność gry słów, ironii, aluzji, niedopowiedzeń komunikatów wrażeń nie wprost. Masowość danych tekstowych skłoniła badaczy do opracowania metod, które w sposób automatyczny mogłyby wykrywać te niuanse w języku. Najprostszym przykładem może być metoda używana przez wyszukiwarkę Twittera, która jako pozytywne klasyfikuje wpisy, którym towarzyszy emotikon przedstawiający uśmiech. Inne, bardziej

zaawansowane metody polegają na klasyfikacji wypowiedzi w oparciu o specjalne słowniki (np. SentiWordNet). Kwestią dyskusyjną jest jednak trafność klasyfikacji zabarwienia emocjonalnego dokonywana przy użyciu tych metod. Analizy porównawcze wskazują, że niektóre z nich dokładniej rozpoznają emocjonalny wydźwięk tekstu, inne potrzebują do tego dodatkowych modyfikacji (Tomanek 2014). Uogólniając tę kwestię, można powiedzieć, że za sprawą zróżnicowania różnych form wypowiedzi pojawiających się w sferze komunikacji w Internecie dylematy badaczy zajmujących się analizą treści namnożyły się. Wpisy użytkowników portali społecznościowych są zbyt krótkie i powierzchowne, aby można było je potraktować jako typowy materiał jakościowy, ale pozostawiają też zbyt wiele w sferze domysłów, aby bez utraty istotnych niuansów można je było analizować w sposób masowy. Z problemem tym zetknęliśmy się również w naszym badaniu. W artykule skupimy się na wpisach opatrzonych hashtagami „dobra zmiana” oraz „Polska w ruinie”. Hasła te pojawiły się po raz pierwszy w trakcie kampanii prezydenckiej PiS w 2015 roku, a potem spopularyzowane zostały przez sztab tej samej partii w okresie poprzedzającym wybory parlamentarne. Interesować nas będzie „kariera” obu haseł, a więc to, jak przebiegały trajektorie ich popularności oraz emocjonalny wydźwięk nadawany im przez użytkowników Twittera w wybranych okresach 2015 roku.

Hasło „dobra zmiana” pojawiło się po raz pierwszy w dyskursie publicznym podczas konwencji kończącej kampanię prezydencją Andrzeja Dudy 20 maja 2015 roku. Zwrot „Polska w ruinie” nie był natomiast oficjalnym hasłem wyborczym, ale

został dostrzeżony przez opinię publiczną w okolicy wyborczych wypowiedziach polityków PiS. Publicystyczne analizy „karier” obu haseł pokazują, że „dobra zmiana” okazała się bardziej chwytliwa i miała cieplejszy wizerunek. Przykładowo, według badania przeprowadzonego przez firmę Sotrender ¾ z 680 tysięcy wpisów opatrzonych tym hasłem miało wydźwięk neutralny na trzystopniowej skali (neutralny, pozytywny, negatywny), choć obserwowano tu pewne fluktuacje spowodowane bieżącymi wydarzeniami (Sotrender 2015a). Internauci zareagowali natomiast zupełnie inaczej na hasło „Polska w ruinie”. Agnieszka Kublik i Iwona Szpała (2015) opisują, jak zostało ono „przechwycone” przez internautów, którzy stworzyli na ten temat serię przesmiewczych memów, wyśmiewając jego pierwotny sens. W naszym artykule chcielibyśmy dokładniej prześledzić różnicującą się z czasem „temperaturę” wpisów opatrzonych tymi hashtagami, które pojawiały się w czasie istotnych wydarzeń politycznych w 2015 roku.

Założyliśmy, że hashtagi zyskują najintensywniejsze zabarwienie emocjonalne w czasie ważnych wydarzeń społecznych lub politycznych i z tego powodu zdecydowaliśmy się obserwować je tylko w pewnych punktach czasu. Do analizy wybrałyśmy 4 okresy: (1) wybory parlamentarne (od 11 października 2015 do 8 listopada 2015 roku); (2) początek konfliktu o Trybunał Konstytucyjny (1–2 grudnia 2015 r.); (3) manifestacje zorganizowane 12 grudnia przez Komitet Obrony Demokracji i 13 grudnia 2015 roku przez Prawo i Sprawiedliwość; (4) moment wprowadzenia świadczeń z programu 500+ oraz afery wokół stadniny koni w Janowie Podlaskim z 2 kwietnia 2016 roku.

Wpisy z interesującymi nas hashtagami odnajdywałyśmy przy użyciu wyszukiwarki Twittera¹, a następnie kopiowałyśmy je do utworzonego zbioru danych. Zebrałyśmy wszystkie wpisy z tych okresów, które miały co najmniej jedno „polubienie” lub choć raz były przesłane innym użytkownikom. Wpisy nie spełniające tych kryteriów nie zostały włączone do analizy. Łączenie zebrałyśmy 494 wpisy (307 z hashtagiem #dobrazmiana i 197 #polskawruinie). Następnie uzyskane w ten sposób wpisy kodowałyśmy zgodnie z czteropunktowym, stworzonym przez nas kluczem. Do pierwszej kategorii należały wpisy krytyczne, a więc takie, w których wyrażono negatywną ocenę opisywanego zjawiska, a krytyka wyrażana była wprost, na przykład: „Zawistni, samotni ludzie codziennie truci jadem #PiS i #RadioMaryja taki jest efekt #dobrazmiana”. Druga kategoria to wpisy krytyczno-prześmiewcze. Pojawiała się tu krytyka przeplatana drwiną czy ironią. W tego rodzaju wpisach częściej miałyśmy do czynienia z przewrotnym wobec pierwotnego sensu rozumieniem zjawiska lub pozorną aprobatą dla dobrej zmiany czy Polski w ruinie. Na przykład: „Lepiej być gorszego sortu niż ekipę mieć z resortu #polskawruinie” lub „Więc mówicie, że ta pani od słodzenia o. Rydzkowi na praśnoerotycznym przydechu decyduje teraz o publikacji wyroków TK? #dobrazmiana”. W ramach kolejnej kategorii opisaną jako „inny kontekst” zgrupowałyśmy wszystkie te wpisy, w ramach których użytkownicy odnosili się do kontekstów innych niż polityczny. Znalazły się tu także wpisy związane z prywatnymi, osobistymi doświadczeniami twitterowiczów, na przykład: „zgódź się na pakiet Start+ z pełnym sportem, strać nspport i jedyne niesportowe kanały

jakie oglądałeś. Mam swoją prywatną #dobrazmiana” lub „Pralka przestała się tłuc przy wirowaniu #DobraZmiana #Codziennie3Pozytywy”. W ramach ostatniej kategorii zgromadziłyśmy wpisy o pozytywnym wydźwięku, odnoszące się do opisywanych zjawisk w sposób afirmatywny. W przypadku wpisów z hashtagiem „Polska w ruinie” do ostatniej kategorii zaliczyłyśmy wpisy, z których wynikało, że aprobowały zły stan państwa. Warto w tym miejscu zaznaczyć, że nasza analiza nie odróżnia wpisów dokonywanych przez „zwykłych” użytkowników od tych, które były tworzone przez osoby aktywnie zaangażowane w rozpowszechnianie pewnych idei. W niektórych sytuacjach o preferencjach politycznych czy światopoglądzie można było co prawda wnioskować po nazwie użytkownika, ale tylko znikoma część postów opatrzonych #dobrazmiana w naszym zbiorze danych została napisana przez administratorów profili konkretnych partii lub polityków. Nie oznacza to oczywiście, że pewne poglądy nie są rozpowszechniane przez adherentów partii politycznych ukrywających się pod neutralnymi nickami. Warto jednak dodać, że zastosowane przez nas kryterium włączenia do analizy (jedno polubienie lub przesłanie) wykluczyło dużą liczbę wpisów, którymi w ogóle nie zainteresowali się użytkownicy. O wielkości zbioru, na którym przeprowadziłyśmy analizę, zdecydowała więc aktywność samych użytkowników oraz wybrane przez nas ramy czasowe. Pewnym utrudnieniem okazała się natomiast automatyzacja zbierania wpisów. O ile bowiem istnieją łatwo dostępne narzędzia umożliwiające pobieranie wpisów z Twittera tworzonych na bieżąco (np. pakiet TwitterR w środowisku R), to możliwość automatycznego pobierania wpisów archiwalnych okazała się trudniejsza.

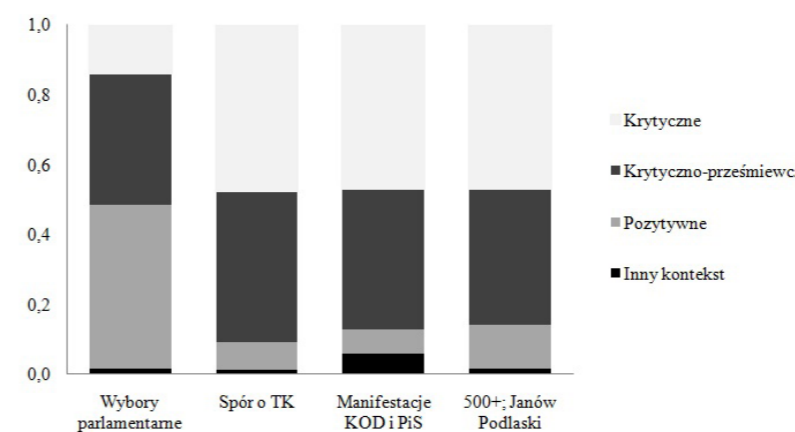
Ewolucja znaczeniowa hasła #dobrazmiana

Wykresy 1–4 obrazują proporcję wpisów zakwalifikowanych do poszczególnych kategorii (wykresy słupkowe) oraz częstość występowania hashtagów w wybranych do analizy okresach (wykresy liniowe).

Hasło „dobra zmiana” przeszło swojego rodzaju znaczeniową ewolucję. Na próbcie zebranego przez nas materiału widać bowiem, jak hashtag ten tracił stopniowo swoje początkowe, pozytywne znaczenie nadane mu w kampanii parlamentarnej PiS. W tam-

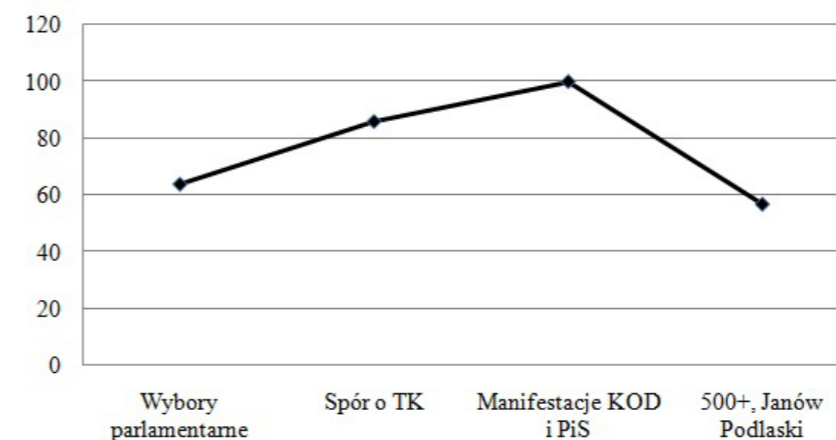
tym czasie #dobrazmiana w rozumieniu pozytywnym najczęściej wyrażała nadzieję na poprawienie sytuacji Polski i Polaków. Kilka wpisów towarzyszących temu hashtagowi zawierało odwołania do obietnic wyborczych, między innymi programu 500+. Jako przykład podać można następujący wpis: „Zamiast produkcji ludzi na szkle i mrożenia ich w azocie – 500 zł na dziecko #dobrazmiana”. Inne wpisy z kolei wyrażały krytykę wobec obecnej, złej sytuacji w kraju, za którą w domyśle odpowiadają politycy PO: „Prezesi spółek z udziałem Skarbu Państwa zarabiają po 10k PLN DZIENNIE! #dobrazmiana”.

Wykres 1. Klasyfikacja wpisów opatrzonych hashtagiem #dobrazmiana.



Źródło: opracowanie własne.

Wykres 2. Liczba wpisów opatrzonych hashtagiem #dobrazmiana w wybranych okresach.



Źródło: opracowanie własne.

¹ Zob. <https://twitter.com/search-advanced?lang=en>.

Pozytywny wydźwięk znaczeniowy hashtaga #dobrazmiana nasilił się w dniu wyborów i tuż po nich. Wyraźnie we wpisach osób usatysfakcjonowanych wynikami *exit poll* i szacunkami PKW można było odczuć entuzjazm: „Lubię ten stan euforia i zadowolenie, aż żyć się chce” czy też „Po pracy do późna w nocy w komisji, aby wyniki były prawdziwe, dopiero teraz wyrażam moją wielką radość, że przyszedł czas na #dobrazmiana”.

Jednak wpisy o pozytywnym wydźwięku odnoszące się do „dobrej zmiany” nawet w czasie wyborów stanowiły tylko około połowy wszystkich pojawiających się wątków. Pozostałe miały charakter krytyczny lub krytyczno-prześmiewczy. Niejednokrotnie wymierzone były bezpośrednio w działania zwycięskiej partii i wyrażały obawę o skutki realizacji reform wyborczych: „Czekamy na opodatkowanie banków przez @pisorgpl Od nowego roku wzrosną opłaty za konto - #DobraZmiana”. Oprócz pesymistycznych oczekiwań dotyczących funkcjonowania instytucji wśród powyborczych emocji dało się również zauważyć obawy co do funkcjonowania zwykłych ludzi. Reprezentuje to następujący wpis: „Posprzątałam generalnie pokój. Na wypadek, gdyby rację mieli ci, co głoszą, że w każdej chwili może wpaść do domu Macierewicz. #dobrazmiana”.

„Dobra zmiana” zaczęła mocniej tracić swój pozytywny wydźwięk, ale jednocześnie zyskiwać na popularności na początku grudnia 2015 roku. W tym okresie uwaga twitterowiczów skupiona była przede wszystkim na sprawie wyboru sędziów do Trybunału Konstytucyjnego oraz „nocnych głosowaniach” Sejmu. W związku z tym hashtag #dobra zmiana pojawiał się w towarzystwie haseł

takich jak #nocnazmiana albo #pakietdemokratyczny. Przykładem może być choćby ten wpis: „Polska polityka to #DobraZmiana czy już #NocnaZmiana?” albo „No i co? Zero jawności i zero debaty to #dobrazmiana?”. Innym dość częstym nawiązaniem była rządowa zapowiedź zakończenia finansowania programu leczenia niepłodności *in vitro*, wobec której również wypowiadano się krytycznie. Zauważyć zatem można, że hashtag ten powszechniej zaczął być używany w znaczeniu prześmiewczym – stał się łatwym sposobem na wyrażenie ironii. Wpisom z #dobrazmiana towarzyszył nieraz znak zapytania, który również był sposobem podważenia jego oryginalnego sensu. Zróżnicowaniu uległy też konteksty użycia tego hasła. W omawianym okresie obejmowały one zarówno sprawy społeczne, jak i polityczne, choć te ostatnie ze względu na ogromne emocje, jakie wywołała sprawa wokół TK, były dominujące. Hasło to okazało się jednak na tyle elastyczne, że można nim było również oznaczać wpisy wyrażające pozytywne reakcje wobec zmian w składzie TK: „Naród głosował za #DobraZmiana więc zmiany muszą być zdecydowane. Dacie radę :) @BeataSzydlo i @AndrzejDuda”. Zauważyć można więc, że hasło to stało się sposobem na wyrażenie skrajnie odmiennych sensów.

Rosnącą krytykę wobec „dobrej zmiany” formułowano w czasie, gdy odbywała się manifestacja Komitetu Obrony Demokracji (12 grudnia) oraz tak zwana manifestacja poparcia dla rządu, mająca miejsce w rocznicę wprowadzenia stanu wojennego. Krytykę wzmocniał również rozwijający się spór o Trybunał Konstytucyjny oraz bieżące posunięcia polityczne partii rządzącej. W badanym okresie prześmiewcza krytyka „dobrej zmiany” wyrażana

była między innymi we wpisach dotyczących manifestacji KOD. Chętnie stosowano tu zwrot „najgorszy sort Polaków”, przejęty z wypowiedzi Jarosława Kaczyńskiego (11.12.2016 r.), który najczęściej traktowany był jako żartobliwa autoidentyfikacja. Przykładem tego są następujące wpisy: „Na ulicę wyszli «najgorszego sortu». Jestem z nich dumny, damy radę. #zKODpodTK @Konstytucja #DobraZmiana #TujestPolska” lub „Jestem dumnym członkiem organizacji NSP (Najgorszego Sortu Polaków) #dobrazmiana @BeataSzydlo”.

Konsekwentnie formułowano również krytykę wokół działań partii rządzącej, związanych z Trybunałem Konstytucyjnym. W tego rodzaju wpisach częściej niż w pozostałych odnoszono się bezpośrednio do konkretnych polityków PiSu lub do prezydenta Andrzeja Dudy. Przykładem może być wpis: „Przemowa JK atakuje Trybunał czyli rządy prawa, dzieli Polaków na my - oni, i świadczy że przestraszył się wczorajszego marszu #dobrazmiana?” albo „@PremierRP Skandal, pani Kempa idzie prostą drogą pod Trybunał Stanu, gdzie już czekają na @AndrzejDuda i @BeataSzydlo. Taka #dobrazmiana”.

Część wpisów krytycznych oraz krytyczno-prześmiewczych w badanym okresie dotyczyła wyboru na stanowisko nowego prezesa PKP, Bogusława Kowalskiego, który, jak donoszono poprzez tradycyjne media, w poprzednim ustroju współpracował ze służbą bezpieczeństwa. Ta decyzja rządzących nasiliła nie tylko krytykę wyrażaną wprost: „#dobrazmiana Posłowie PZPR (Piotrowicz) w Sejmie i SB-ccy kapusie w spółkach Skarbu Państwa. W PiS same szumowiny” czy „Kolejny TW, tym razem w PKP... #dobrazmiana”. Krytykowano tak-

że ironicznie: „Czekam kiedy nowo wybrany przez @pisorgpl prezes PKP S.A. powie, że «ktoś musiał być agentem Służby Bezpieczeństwa». #dobrazmiana #PKP #PRL”. Powołanie Kowalskiego pobudziło również twitterowiczów do porównywania praktyk obecnej władzy z posunięciami poprzedniego rządu. Przykładem jest wpis: „Jak PO obsadza urzędy i stanowiska swoimi to jest to nepotyzm i sitwa. Gdy robi to #PiS to jest to #DobraZmiana? #PZU #Orlen #PGNiG #PKP”.

W badanym okresie wyróżnić można też nieliczne wpisy pozytywnie odnoszące się do „dobrej zmiany” bądź będące wyrazem wsparcia dla tych, którzy zdecydowali się ją wprowadzać. Przykładem może być tu wpis „adresowany” do Andrzeja Dudy: „@prezydentpl @AndrzejDuda Panie Prezydencie, proszę się nie przejmować i działać dla dobra Polski:))) Wszystko przed nami #dobrazmiana”.

Wpisy z połowy grudnia 2015 roku pokazują zatem dalsze pogłębianie krytyki „dobrej zmiany”. Wpisy pozytywne stanowią tu nieco mniej niż 10%. Wzrasta zaś liczba wpisów krytyczno-prześmiewczych oraz krytycznych – utrzymuje się ona na poziomie ponad 40%. Popularność hashtaga #dobrazmiana wpływa również na jego dalsze rozpowszechnianie się w kontekstach innych niż polityczny. Przykładem mogą być tu wpisy odnoszące się do sportu, na przykład „To chyba ich ostatni taniec? Zimą w Legii #dobrazmiana?” lub „Już nawet logo Jagiellonii zmienili! @TomaszMarzec #dobrazmiana”.

W ostatnim z analizowanych przez nas okresów negatywny wydźwięk hasła „dobra zmiana” utrzymywał się, choć zauważyć też można nieznaczne jego

„ocieplenie”. Głównym tego powodem było uruchomienie programu 500+ i pozytywne wpisy związane czy to z łatwością przygotowania wniosku, czy satysfakcją twitterowiczów płynącą z wypełnienia obietnicy wyborczej: „Wczoraj wypełniłem uproszczoną wersję wniosku #500+. Trwało to dokładnie 3min. Skrajnie proste! Bardzo #dobrazmiana. UPO dostałem po 1min”. Hasło programu prorodzinnego używane było również do skonstrastowania działań obecnego i poprzedniego rządu: „PO miało Amber Gold, a PiS ma pięćset plus. #dobrazmiana”.

W tym okresie dominujące jednak były wpisy związane z sytuacją w stadninie koni w Janowie Podlaskim. To właśnie krytyka związana ze śmiercią koni i zmianami personalnymi na stanowisku prezesa stadniny nadały ostatecznie negatywny wydźwięk „dobrej zmianie”. Znowu – jak w poprzednich okresach – hasło to używane było przede wszystkim ironicznie lub krytycznie: „Tylko koni żal #dobrazmiana #stadniny” albo „W Janowie z powodu #dobrazmiana umierają kolejne konie. Czekam aż zaczną umierać ludzie”. Sprawa ta niejako przykryła pozytywny wydźwięk hasła, którego w tym czasie nadały mu doświadczenia niektórych twitterowiczów aplikujących o świadczenia rodzinne.

Krótką „kariera” hashtaga #Polskawruinie

Inaczej przedstawiała się kariera hasła „Polska w ruinie” (wykresy 3 i 4). Przede wszystkim była znacznie krótsza. Analizując częstość pojawiania się tego hashtaga w sposób bardziej szczegółowy niż w przypadku #dobrejzmiany, widać, że był tu tylko jeden wzrost popularności, właśnie w oko-

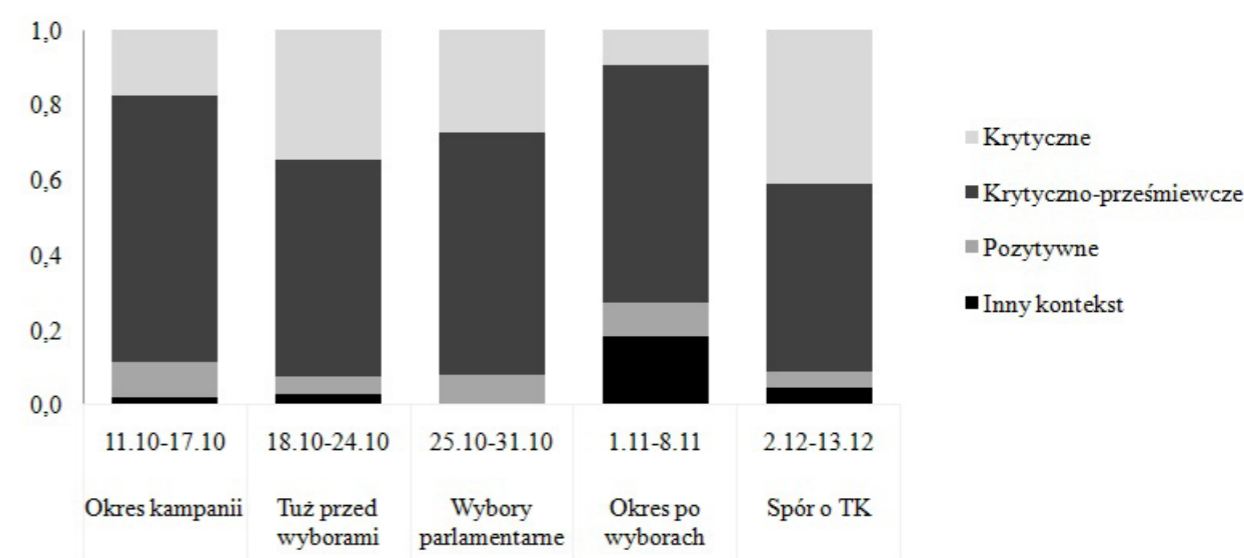
licy wyborów. W tym czasie hasło to używane było do wyrażenia ostrej krytyki wobec zwycięskiej partii: „#PolskaWybiera regres, rozdawnictwo, prowincjonalizm, patrzenie wstecz, państwo wyznaniowe. #CzasDecyzji #PolskaWRuinie” lub „Obawiam się ze #polskawruinie to nie było hasło, a raczej zapowiedź #wybory2015 #bazarek”. Już wtedy hasło to miało przede wszystkim zabarwienie krytyczno-prześmiewcze i używane było do podkreślenia kontrastu między jego pesymistycznym wydźwiękiem a optymistycznym spojrzeniem twitterowiczów: „Mamy Orliki, mamy stadiony, mamy mnóstwo szkółek piłkarskich i mamy sukces narodowej drużyny! #PolskawRuinie” albo: „Na 3 kompy w kawiarni 3 to MacBooki. ;) #Handlujz tym #polskawruinie”. Twitterowicze utworzyli wiele podobnych wpisów – spora część z nich odnosiła się do dobrej w ich mniemaniu sytuacji ekonomicznej państwa, niskiego bezrobocia czy dobrego poziomu życia. Na Twitterze dominującym przekazem w tym okresie były jednak nie wpisy zawierające tekst, ale zdjęcia przedstawiające polskie miasta jako nowoczesne metropolie – mieniące się światłami wieżowce centralnej Warszawy czy malownicze starówki mniejszych miast stały się kontrapunktem oryginalnego znaczenia tego hasła. Komentarz do tych zdjęć stanowił nie raz sam hashtag #polskawruinie albo inne zdanie wyrażające ironię. Wobec tak jednoznacznych nastrojów hasło to pojawiało się w wypowiedziach polityków PiS coraz rzadziej (w kwietniu w ogóle nie pojawiło się na Twitterze), a jego dramatyczny wydźwięk był łagodzony.

Jeśli natomiast chodzi o wpisy zgodne z prawdziwym sensem tego hasła, to odwoływały się one do

różnych kontekstów związanych ze stanem państwa i złym funkcjonowaniem jego instytucji. Przykładem mogą być też wpisy: „Złoty wśród trzech najgorszych walut miesiąca #PIS #polskaWRuinie” albo „#Bieda jest wszechobecna: w skrajnym ubóstwie żyje już 10% #dzieci w #Polska!”. Jeszcze inny użytkownik użył go do opisanie swoich doświadczeń w ten sposób: „Iść do okulisty z nr 11, od 8 do 10 wszedł 1 pa-

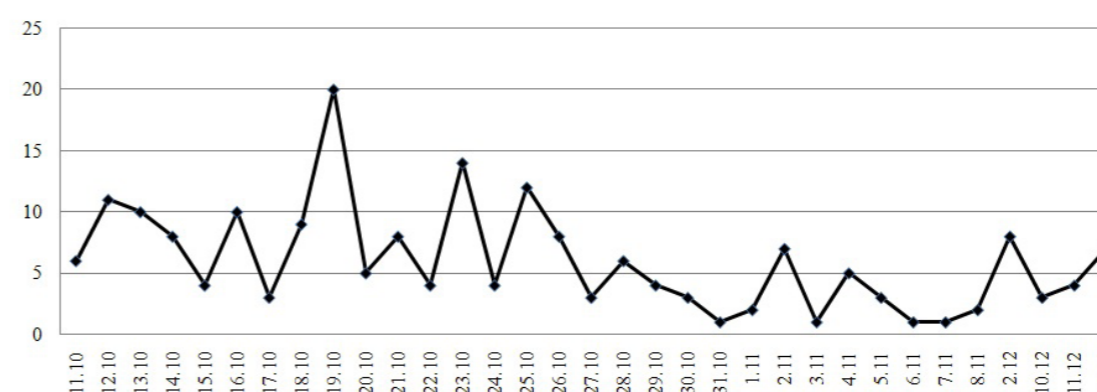
cjent - pytanie za ile godzin wejdzie 11 :/ #polskawruinie”. Znacznie rzadziej niż „dobra zmiana” hasło to używane było do oznaczenia sytuacji wykraczających poza sferę polityki i spraw socjalnych. Rzadkim przykładem może być wpis-obszernia poczyniona przez jednego z twitterowiczów: „Nastolatka siedząca w autobusie kończy swoje wypowiedzi zwrotem «iks de» #polskawruinie”.

Wykres 3. Klasyfikacja wpisów opatrzonych hashtagiem #polskawruinie na tle wydarzeń z 2015 roku.



Źródło: opracowanie własne.

Wykres 4. Liczba wpisów opatrzonych hashtagiem #polskawruinie od 11 października do 13 grudnia 2015 roku.



Źródło: opracowanie własne.

Analizowane przez nas hasła wprowadzono do języka polityki, opatrując je dość konkretnymi definicjami albo przynajmniej ograniczoną liczbą kontekstów, do których miały się odnosić. Jednak przebieg ich „kariery” w sieci był dosyć zaskakujący i odmienny. W czasie trwania kampanii wyborczej poprzedzającej wybory parlamentarne hasło „Polska w ruinie” symbolizowało krytyczną ocenę działań ówczesnej władzy, rozliczne jej zaniechania czy wręcz celowe działania godzące w dobro obywateli i państwa. Z kolei „dobra zmiana” definiowana była jako remedium na opłakany stan kraju, a symbolizowała miała nadchodzącą nową jakość polskiej polityki oraz skuteczność proponowanych rozwiązań czy reform. Oba hasła po wygranych przez PiS wyborach zostały w pewien sposób przez internautów (w tym twitterowiczów) przejęte i przedefiniowane. Jednak ich żywotność i elastyczność okazały się znacząco różne. Hasło „Polska w ruinie” w zamysle spin doktorów Prawa i Sprawiedliwości odnosić się miało do wszelkich działań Platformy Obywatelskiej, która w mniemaniu ówczesnej opozycji doprowadziła do ruiny. Wobec tego internauci w czasie trwania kampanii parlamentarnej albo aprobowali zaproponowaną definicję, albo ją krytykowali. Krytycy nie tyle nie zgadzali się z definicją, co zgadzali się z nią pozornie, zestawiając hasło na przykład ze zmianami infrastrukturalnymi powstałymi dzięki funduszom europejskim. Słowem, próbowano tu pokazywać niedorzeczność sformułowania „ruina”. Co ciekawe, po wyborach w zasadzie powrócono do pierwotnej definicji, z tym jednak zastrzeżeniem, że w roli rządzonych obsadzono już aktualną władzę.

Hashtag #polskawruinie okazał się znacznie mniej nośny i żywotny. Jego znacznie mniejsza popular-

ność (w zestawieniu z „dobrą zmianą”) mogła mieć kilka przyczyn. Po pierwsze jest to hasło jednowymiarowe i punktowe. W związku z tym trudniej łączy się z kontekstami innymi niż polityczny, podczas gdy „dobra zmiana” naturalnie przylega na przykład do wydarzeń sportowych (choćby z tego powodu, że w ramach wielu sportów zespołowych posługujemy się sformułowaniem „dokonać zmiany” na boisku). Po drugie #polskawruinie to sformułowanie znacznie mniej abstrakcyjne, co wpływa na możliwości produkcji skojarzeń z takim hashtagiem w tle. Po trzecie wreszcie #polskawruinie mogła tracić na popularności z uwagi na swój jednoznacznie negatywny wydźwięk, podczas gdy pojemna #dobrazmiana, także w okresie po wyborach, rozbudzała zniuansowane emocje i budowała rozmaite nowe konteksty. Warto też podkreślić, że „dobra zmiana” okazała bardzo sprawnie działającym sloganem wyborczym, który w trakcie kampanii trudno było skutecznie ośmieszyć. Jej charakter określić można jako włączający, inkluzyjny. Prawdopodobnie z tego powodu krytyka tego hasła w sieci była nieco opóźniona. #polskawruinie okazała się zbyt siermiężna i mało elastyczna, co jej natychmiastową krytykę znacznie ułatwiło.

Dokonując uogólnień, można powiedzieć, że kariera „dobrej zmiany” rozwijała się w kilku etapach. Początkowo było to hasło wyborcze. Jako takie niosło ze sobą duży ładunek pozytywnych emocji. Na poziomie dosłownym nawoływało do faktycznej zmiany partii rządzącej, a na poziomie metaforycznym niosło ze sobą nadzieję na lepsze czasy. Jednak już niedługo później – my obserwujemy to już w grudniu – hasło to używane było w sensie ironicznym, kwestionującym czy też podającym w wątpliwość

owe korzystne dla Polski skutki. Przejawiało się to w nasileniu wpisów o charakterze krytycznym oraz krytyczno-prześmiewczym. Wyraźnie hasło to stało się znacznie bardziej popularne wśród krytyków rządów PiS, którzy na różne sposoby przekształcali jego sens, posiłkując się bieżącymi kontekstami. Jednocześnie spektrum znaczeniowe „dobrej zmiany” systematycznie się poszerzało. Hasło to zaczęło być stosowane w najróżniejszych kontekstach niezwiązanych z polityką. Przejął je świat sportu czy mody i używał do opisanego sytuacji, w których nastąpiła jakakolwiek, nawet banalna zmiana. Na tym etapie hasło nieodwracalnie oderwało się od swojego pierwotnego, ściśle politycznego kontekstu. Ze względu na swoje szerokie i elastyczne pole znaczeniowe, możliwość stosowania go w różnych sytuacjach hasło to okazało się znacznie bardziej nośne niż zło-wieszczko brzmiąca „Polska w ruinie”.

Podsumowanie

Przedstawiony w artykule *case study* stanowi ilustrację ewolucji znaczeniowej dwóch hashtagów. Pokazuje między innymi, że kategoryzacja treści za pomocą wspólnej etykiety potrafi być niespójna i zmieniać się w zależności do szerszego tła społecznego. W tym przypadku tło stanowiły istotne w badanym przez nas okresie wydarzenia na scenie politycznej. Hasła, które początkowo miały jasno zdefiniowany sens, zostały przez użytkowników portalu przekształcone zgodnie z panującym w nim stylem komunikacji – szybkim i mocnym przekazem. W konsekwencji konteksty używania haseł w pewnym stopniu różnicowały się, ale służyć zaczęły przede wszystkim jako sposób wyrażania niezadowolenia czy rozczarowania. użyt-

kownicy zamiast opisywać swoje stany emocjonalne czy przemyślenia w bardziej obszerny sposób kwitowali wypowiedzi przy użyciu prostego #dobrazmiana.

Przeprowadzone badanie pozwala także sformułować kilka praktycznych wskazówek dotyczących analizy wpisów pochodzących z Twittera i ogólniej – mediów społecznościowych. Pierwszą rzeczą, z jaką styka się badacz, przystępując do analizy takiego materiału, jest jego powierzchowność, lapidarność i nieuporządkowanie. Większa część użytkowników nie zbliżała się nawet do twitterowego limitu 140 znaków. Być może właśnie ta lakoniczność wypowiedzi skłania użytkowników do stosowania pozatekstowych metod, które pozwalają uwypuklić odczuwane emocje czy wyrażane sensy. Są to przeróżne znaki interpunkcyjne, emotikony, dodatkowe hashtagi oraz zdjęcia, które często towarzyszyły zebrany przez nas wpisom. Stosując wielokropki, wykrzykniki czy pytajniki ludzie potrafili nie tylko wyjaskrawić swoje stanowisko, ale, jak często to obserwowaliśmy w przypadku obu hashtagów, zmienić ich sens. Ta specyfika twitterowych wpisów sprawia, że nadają się one zarówno do analizy ilościowej (chodzi o ich masową liczbę), ale poprzez zawarte w nich odcienie znaczeń również jakościowej. Bez względu na przyjętą strategię analizy danych, media społecznościowe stanowią pewien warty uwagi rejestr nastrojów istniejących w danej chwili. Pozwalają zrekonstruować emocjonalny „cykl życia” pewnych wydarzeń, uchwycić bieżące trendy czy spontaniczne reakcje. Tej właściwości nie mają klasyczne badania społeczne, w których wspomnienia o przeszłych stanach emocjonalnych, już przygaśnięte, wywołuje

się przy użyciu pytań. Dane pochodzące z Twittera mają jednak swoje oczywiste ograniczenia. Przede wszystkim pochodzą od osób bardziej zaangażowanych w sprawy polityki w porównaniu do użytkowników innych mediów społecznościowych oraz osób, które wcale z nich nie korzystają.

Ciągle jest to populacja, o której wiadomo bardzo niewiele. Poza tym na Twitterze, przynajmniej na razie, znacznie częściej można spotkać się z reakcjami zabarwionymi szyderstwem, ironią, drwiną niż wpisami wyrażającymi szczere poparcie czy akceptację jakiejś idei czy poglądów.

Bibliografia

Adamska Karolina (2015) *Hashtag, czyli komunikat? Rola i funkcje hashtagów na Twitterze*. „Studia Medioznawcze”, t. 3, nr 62, s. 61–70.

Chaciński Bartek (2016) *Hashtag wychodzi z Internetu* [dostęp 9 sierpnia 2016 r.]. Dostępny w Internecie: <<http://www.polityka.pl/tygodnikpolityka/ludzieistyle/1575293,1,hashtag-wychodzi-z-internetu.read>>.

Deryło Aleksander, Makarenko Vadim (2015) *Kandydaci do Sejmu w mediach społecznościowych. Twitter nie jest bardziej prawicowy niż Facebook* [dostęp 4 maja 2017 r.]. Dostępny w Internecie: <<http://biqdata.wyborcza.pl/kandydaci-do-sejmu-w-mediach-spoecznościowych-twitter-facebook>>.

Gemius (2015) *Kto w Polsce korzysta z Twittera* [dostęp 15 sierpnia 2016 r.]. Dostępny w Internecie: <<http://antyweb.pl/gemius-kto-w-polsce-korzysta-z-facebook-i-twittera/>>.

Kublik Agnieszka, Szpala Iwona (2015) *Polska w ruinie:-) Internauci wyśmiali hasło, Szydło łagodzi przekaz PiS*, Gazeta Wyborcza [dostęp: 4 maja 2017 r.]. Dostępny w Internecie: <<http://wyborcza.pl/1,75398,18526299,polska-w-ruinie.html>>.

Kramer D. I. Adam, Guillory Jamie E., Hancock Jeffrey T. (2014) *Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks*. „PNAS”, t. 111, nr 24, s. 8788–8790.

Ma Zongyang, Sun Aixin, Cong Gao (2013) *On Predicting the Popularity of Newly Emerging Hashtags in Twitter*. „Journal of the American Society for Information Science and Technology”, vol. 64, no. 7, s. 1399–1410.

Orlean Susan (2010) *Hash. „New Yorker”* [dostęp 29 listopada 2016 r.]. Dostępny w Internecie: <<http://www.newyorker.com/culture/susan-orlean/hash>>.

Parker Ashley (2011) *Twitter’s Secret Handshake* [dostęp 9 sierpnia 2016 r.]. Dostępny w Internecie: <http://www.nytimes.com/2011/06/12/fashion/hashtags-a-new-way-for-tweets-cultural-studies.html?_r=1>.

Sotrender (2015a) *Twitter Trends Październik 2015 – miesiąc wyborów i piłki nożnej* [dostęp 29 listopada 2016 r.]. Dostępny w Internecie: <<https://blog.sotrender.com/pl/2015/11/twitter-trends-pazdziernik-2015/#close>>.

Sotrender (2015b) *Twitter w Polsce – podsumowanie* [dostęp 29 listopada 2016 r.]. Dostępny w Internecie: <blog.sotrender.com/pl/2016/01/twitter-w-polsce-podsumowanie-2015-r-infografika/>.

State Bogdan, Adamic Lada (2015) *The Diffusion of Support in an Online Social Movement: Evidence from the Adoption of Equal-Sign Profile Picture*. „CSCW ’15Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing” [dostęp 29 listopada 2015 r.]. Dostępny w Internecie: <<http://dl.acm.org/citation.cfm?id=2675290>>.

Tomanek Krzysztof (2014) *„Analiza sentymentu” – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych*. „Przegląd Socjologii Jakościowej”, t. 10, nr 2, s. 118–136 [dostęp 15 sierpnia 2016 r.]. Dostępny w Internecie: <www.przegladsocjologiijakosciowej.org>.

Twitter (2016) *Twitter Usage / Company Facts* [dostęp 29 listopada 2016 r.]. Dostępny w Internecie: <<https://about.twitter.com/company>>.

Wirtualne Media (2016) *Twitter rekordowo popularny w Polsce. Jacy są jego użytkownicy?* [dostęp 15 sierpnia 2016 r.]. Dostępny

w Internecie: <<http://www.wirtualnemedi.pl/artukul/twitter-rekordowo-popularny-w-polsce-jacy-sa-jego-uzytkownicy>>.

Yardi Sarita, Boyd Danah (2010) *Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter*. „Science, Technology & Society”, vol. 30, no. 5, s. 316–327.

Cytowanie

Zawistowska Alicja, Skowrońska Małgorzata (2017) *Dobra zmiana czy Polska w ruinie?* „Przegląd Socjologii Jakościowej”, t. 13, nr 2, s. 110–127 [dostęp dzień, miesiąc, rok]. Dostępny w Internecie: <www.przegladsocjologiijakosciowej.org>.

Good Change or Poland in Ruins?

Abstract: The paper addresses the evolution of the messages tagged with hashtags “good change” and “Poland in ruins” posted on Twitter. Both slogans have been introduced by Polish political party Law and Justice during parliamentary campaign in 2015. Although slogans quickly gained attention of public opinion, their later careers were different. An analysis of 494 posts from 4 periods of time in 2015–2016 shows that #goodchange gradually lose its original, positive meaning given in the parliamentary campaign. Instead, over the time, it gained more ironic connotations. The career of #polandinruins was shorter and Twitter users immediately changed it sense into irony. The context of the slogans has been differentiated, but served primarily to express dissatisfaction. The study also raises a methodological issue concerning validity of coding in the face of evolution of connotations.

Keywords: Twitter, Hashtag, Good Change, Poland in Ruins, Emotions

Krzysztof Tomanek
Uniwersytet Jagielloński

Metodyka dla analizy treści w projektach stosujących techniki text mining i rozwiązania CAQDAS piątej generacji

Abstrakt Projekty, w których przychodzi nam pracować z dużymi wolumenami danych tekstowych, pochodzących z różnych źródeł i zapisanych w różnorodnych formatach, rodzą wiele dylematów natury metodologicznej, wymagają często niestandardowych decyzji i rozwiązań. W szczególności zadanie polegające na opracowaniu danych o różnorodnej jakości, nieustrukturyzowanych typu *quan* i *qual* wymagać może pracy, w której dynamicznie zmieniają się strategie analizy danych, sposoby przekształcania danych tekstowych. Artykuł opisuje przykład takiej właśnie „dynamicznej” metodyki. Wykazała ona swoją wartość w zadaniu polegającym na klasyfikacji wypowiedzi pisanych. W tak zarysowanym kontekście autor artykułu mierzy się z następującymi celami: (a) czy można zastosować oprogramowanie klasy CAQDAS do pracy półautomatycznej lub automatycznej zastępującej część manualnej pracy nad klasyfikacją wypowiedzi? (b) jak skonstruować metodykę klasyfikacji dla danych o różnorodnej jakości? (c) kiedy klasyfikacja automatyczna jest przydatna, a kiedy nie ma szans powodzenia? W artykule zaznaczone zostaną momenty, w których analityk sięga po wiedzę typową dla analiz danych jakościowych oraz te, kiedy wiedza z tego obszaru nie jest już wystarczająca do realizacji wskazanych celów (*natural language processing*, uczenie maszynowe). Przykład projektu będący tłem artykułu wymusił zastosowanie kilku narzędzi i języków wspierających pracę na danych. Praca nad transformacją, klasyfikacją oraz wizualizacją wyników wymagała zastosowania bazy MySQL oraz programów: R, QDA Miner, Wordstat, QlikSense. Roli i ograniczeniom narzędzi klasy CAQDAS poświęconych zostało także kilka uwag.

Słowa kluczowe analiza treści, *Mixed Methods*, *Big Data*, techniki *text mining*, CAQDAS

Krzysztof Tomanek, napisał doktorat z zakresu nauk społecznych na Uniwersytecie Jagiellońskim. Główne obszary zainteresowania i rozwoju naukowego: metodologia nauk społecznych, *big data*, *data mining*, *text mining*, badania z zakresu zaufania i lojalności, wizualizacja danych oraz interaktywne wizualizacje danych. Autor artykułów naukowych i popularnonaukowych z zakresu

praktycznych zastosowań analiz tekstu, *sentiment analysis*, CAQDAS.

Adres kontaktowy:

Uniwersytet Jagielloński, Instytut Socjologii
ul. Grodzka 52, 31-044 Kraków
e-mail: k_tomanek@wp.pl

Analiza treści stosowana jest wobec zróżnicowanych typów danych tekstowych. Innych strategii analitycznych wymaga praca z tekstami opublikowanymi na blogach czy wypowiedziami zapisanymi na forach dyskusyjnych, a innych praca z tekstami pamiętników czy transkrypcjami pochodzącymi z wywiadów grupowych. Dlatego też analiza treści tożbiór technik, które stosowane są do rozwiązywania różnorodnych problemów badawczych. Przykłady takich obszarów badawczych, które dynamizują rozwój technik analizy treści, to: *culturomics*, analiza opinii, analiza sentymentu.

Jedną z najczęściej cytowanych w ostatnich latach w obszarze *culturomics* jest publikacja pokazująca, jak zmieniała się literatura XX wieku pod względem zawartego w niej „ładunku emocjonalnego” – czyli, innymi słowy, pod względem stosowania słów nacechowanych emocjonalnie (Acerbi i in. 2013). Praca zespołu Alberto Acerbi’ego mieści się w obszarze analiz sentymentu. Sformułowanie „analiza sentymentu” (SA) bywa używane zamiennie (Liu 2012) z bardziej ogólnym – „analiza opinii” (AO). W obu przypadkach pierwszy człon frazy dotyczy automatycznych i półautomatycznych metod analizy treści. Celem tych metod jest identyfikowanie i klasyfikowanie wypowiedzi ze względu na pojawiające się w nich słowa i frazy kluczowe. W przypadku SA są to słowa nacechowane emocjonalnie, a w AO zakres poszukiwań obejmuje nie tylko opinie identyfikowane ze względu na temat czy formę wypowiedzi, ale także ze względu na obiekt, którego opinia dotyczy, oraz na przykład styl, w jakim wypowiedź została sformułowana. Analiza opinii ma zatem szerszy zakres przedmiotowy, a analiza sentymentu jest jednym z elementów tego obszaru (Liu 2012).

Jeszcze inny, stosunkowo nowy obszar, w którym analiza treści jest stosowana w sposób półautomatyczny i automatyczny, to stylometria (Rybicki, Heydel 2013; Eder 2014). Ten rodzaj analiz treści za cel stawia sobie charakterystykę sposobu posługiwania się językiem przez autora wypowiedzi. Celem takiej analizy może być zatem zbudowanie profilu autora tekstów, rozpoznawanie autorstwa tekstów (Zaśko-Zielińska 2014; Rosa 2015). Systemy dokonujące automatycznych stylometrycznych budowane są także dla języka polskiego (Eder, Piasiecki 2015).

Cechą wspólną wspomnianych powyżej obszarów, w których stosowane są techniki analizy treści, jest fakt, że rozwijają się wraz z rozwojem technologii komputerowych. Analizując doniesienia z badań nad tekstami, z niewielką dozą ryzyka można stwierdzić, że analizy tekstu wspierane komputerowo coraz częściej służą naukowcom z różnorodnych dziedzin (Lieberman i in. 2007; Baptiste i in. 2011; Taboada i in. 2011). Powszechnie zautomatyzowane lub półautomatyczne analizy tekstu stosują naukowcy zajmujący się analizami kulturowymi (Baptiste i in. 2011), lingwiści (Lieberman i in. 2007: 713–716), historycy (Pagel, Atkinson, Meade 2007: 717–720), czy zajmujący się antropologią (DeWall i in. 2011: 200–207). Coraz częściej też analiza treści wspierana komputerowo stosowana jest przez socjologów (Niedbalski 2014).

W każdym wspomnianym powyżej badaniu analizy treści realizowane były przy wsparciu narzędzi z obszaru CAQDAS (*Computer-Assisted-Qualitative-Data-Analysis-Software*). Wśród dostępnych istnieją takie rozwiązania CAQDAS, które dysponują

algorytmami kodującymi materiał tekstowy automatycznie. Niektóre z takich technik kodowania działają niczym czarne skrzynki. Analityk nie zna ich budowy ani sposobu, w jaki przetwarzają one dane. Co prawda oprogramowanie takie pozwala na przykład zdefiniować jednostkę analizy; określić słowa kluczowe; wskazać fragment tekstu, który wykorzystywany będzie jako wzorzec do kodowania (np. QDA Miner, R), ale nie daje pełnej wiedzy o sposobie działania techniki analitycznej. Istnieją też takie rozwiązania CAQDAS, które oferują transparentne metody i techniki wspierające pracę z kodowaniem tekstów. Taka sytuacja oznacza dla analityka możliwość nie tylko zapoznania się z definicją algorytmu, ale także jego modyfikację. Mamy więc do czynienia z sytuacją, w której oprogramowanie niesie ze sobą możliwość samodzielnego budowania systemów uczących się kodowania tekstów (R, Qualrus). Taka sytuacja sprzyja rozwojowi metod analiz danych jakościowych. CAQDAS, które pozwala projektować metody i algorytmy, to niemal „nieograniczone” środowisko dla wyobraźni analityka.

Omawiana w niniejszym artykule analiza treści sięga po narzędzia CAQDAS, które pozwalają na projektowanie technik, algorytmów w sposób, dla którego ograniczeniem jest jedynie wyobraźnia analityka. Fakt ten pozwala rozszerzyć propozycję Jakuba Niedbalskiego dotyczącą klasyfikacji CAQDAS (Niedbalski 2013: 153–166). Zastosowanie narzędzi otwartych programistycznie otwiera rozdział piątej generacji CAQDAS. Ten nowy etap ewolucji oprogramowania wspierającego analizy danych jakościowych poza możliwościami analitycznymi, jakie daje, posiada jeszcze jedną istotną

cechę. Spełnia mianowicie istotne – z punktów widzenia poznania naukowego – kryterium transparentności metodologicznej w prowadzonych analizach.

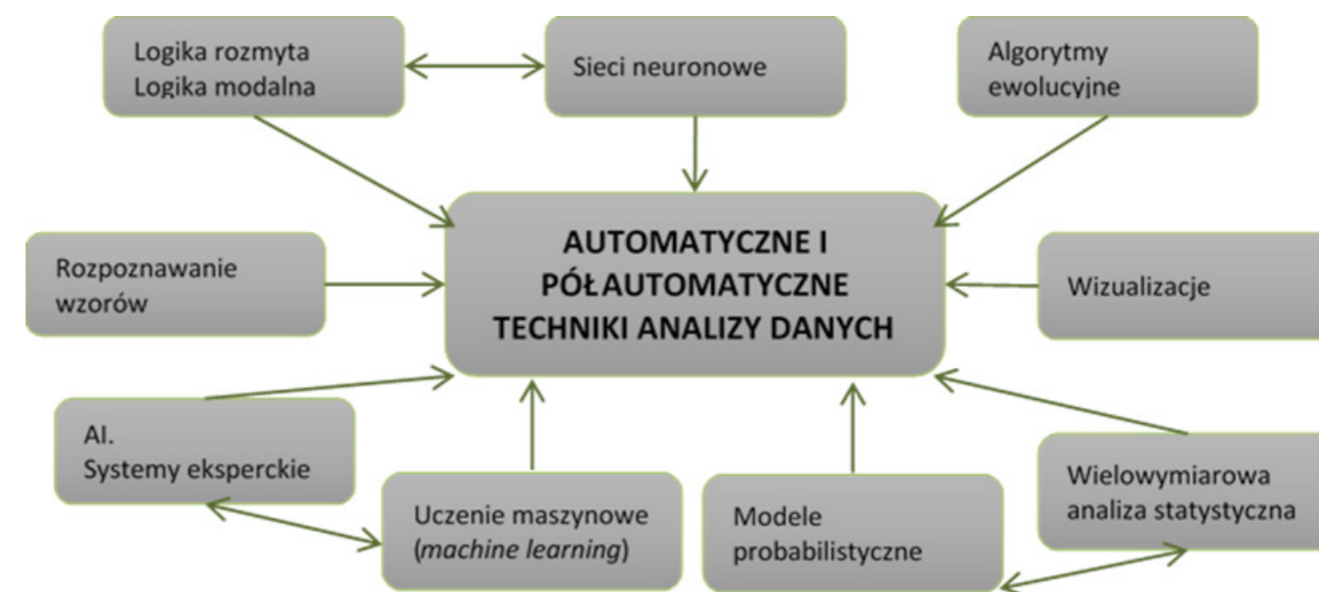
Słowo o zastosowaniu technik półautomatycznych i automatycznych

Wzrost mocy obliczeniowej komputerów osobistych oraz pojemności ich pamięci stworzył w ostatnich latach możliwości zarówno gromadzenia olbrzymich ilości informacji, jak i ich szybkiego przetwarzania. W konsekwencji obserwujemy szybki rozwój różnorodnych automatycznych i półautomatycznych metod analizy danych, technik odkrywania wiedzy również w danych tekstowych. Dla uproszczenia na ilustracji 1 pokazuję obszary tych rozwiązań.

Automatyczne i półautomatyczne metody analiz stosowane są w bardzo wielu dziedzinach: od badań medycznych do przewidywania kursów giełdowych, od przemysłu do gier komputerowych. Stosowane są również w analizie treści. Oto kilka przykładowych zastosowań dla różnorodnych problemów badawczych:

- a. rozpoznawanie metod argumentacji (Tomanek 2016),
- b. wykrywanie wypowiedzi prowadzących do napięć i konfliktów podczas FGI (Jordan i in. 2007),
- c. identyfikacja wypowiedzi tekstowych nacechowanych emocjonalnie i wywołujących emocje (Duggleby 2005),
- d. identyfikacja form perswazji w tekstach prasowych (Appel, Richter 2010),

Diagram 1. Obszary, z których korzystają i w których rozwijane są automatyczne i półautomatyczne metody analizy danych



Źródło: opracowanie własne na podstawie Duch (1997), Bolc, Cytowski (1989–1991), Cichosz (2000).

- e. rozpoznawanie typowych wypowiedzi w wywiadach IDI,
- f. analiza schematów poznawczych w komentarzach oceniających (Kidd, Parshall 2000),
- g. identyfikacja problemów drażliwych w wywiadach FGI (Kaplowitz 2000),
- h. klasyfikacja dużych zbiorów artykułów prasowych w oparciu o schematy kodowania (Schanie, Pino-Foltz, Logsdon 2008),
- i. diagnostyka specyfiki wypowiedzi zwiększających popularność postów na forach dyskusyjnych (Tomanek 2015),
- j. identyfikacja liderów opinii, grup opiniotwórczych na forach internetowych (Smith i in. 2009),
- k. rozwój tematycznych słowników analitycznych w oparciu o reguły leksykalne (Tomanek, Bryda 2015).

Charakterystyka projektu

Techniki półautomatycznej i automatycznej analizy treści opisane w tym artykule zastosowane zostały wobec dużego zbioru wypowiedzi i opinii pozyskanych w trakcie badań ankietowych. Badania, o których mowa, realizowane były w tradycyjnej formule (wywiady *face to face*) oraz z zastosowaniem podejścia CAWI. Zadania respondentów biorących udział w badaniu polegały na zapisaniu skojarzeń związanych z markami producentów i produktami konsumenckimi; uzasadnień oceny marki i produktu¹.

¹ Ze względu na prośbę zleceniodawcy projektu nazwa badań zostanie pominięta, a sam opis badania z konieczności pozostaje skrócony do minimum.

Ankieta zrealizowana została wśród względnie licznych grup respondentów (średnio rocznie uzyskano 10 000 odpowiedzi). Zbiór danych składał się z informacji liczbowych, jak i tekstowych. Ten drugi typ danych to wypowiedzi pisane będące odpowiedziami na pytania otwarte. Projekt, który realizowany był w cyklu rocznym (w latach 2012–2015), zwiększał liczbę respondentów i gromadził coraz większą liczbę wypowiedzi tekstowych. Pierwsza edycja badań przyniosła stosunkowo dużą liczbę 34 453 opinii². W kolejnych latach liczba danych tekstowych przyrasta o ponad 10, 5 i 27 tysięcy. Między 2012 i 2015 rokiem korpus liczy 207 907 wypowiedzi (przyrost pomiędzy pierwszym i ostatnim rokiem badań to 44%³). W każdej edycji projektu analizy tekstów mają na celu wychwycenie wypowiedzi „cennych” z punktu widzenia prezentacji danych w raporcie końcowym (liczba wybieranych do raportu wypowiedzi oznaczona jest w tabeli jako „wybrane”). Szczegółowe dane prezentuje tabela 1.

Przyrost danych w opisanym powyżej tempie ma wpływ na kilka aspektów realizacji projektu. Istotnej zmianie ulegają: czas opracowania danych (Cole i in. 1996), etap analizy oraz wizualizacji, a także koszt realizacji projektu. W takiej sytuacji naturalnymi wydają się pytania:

- Czy część najbardziej czasochłonnej pracy wykonywanej przy analizie tekstów można w jakimś stopniu zautomatyzować?
- Czy w wyniku automatyzacji uzyskamy dane o takiej samej jakości jak w przypadku analiz wykonywanych manualnie? Czy automatyzacja opracowania danych da w efekcie trafne rezultaty?
- Czy metoda i narzędzie wykonujące pracę automatycznie mogą korzystać z wiedzy badaczy, koderów?

Tabela 1. Liczba zebranych wypowiedzi tekstowych w badaniach realizowanych w latach 2012–2015

	Skojarzenia		Uzasadnienia		Łączna liczba wypowiedzi			
	Łącznie w zbiorze	Wybranych	Łącznie w zbiorze	Wybranych	Łącznie w zbiorze	Wybranych	Wybranych %	
II etap analizy I etap	I edycja projektu rok 2012	22 431	450	12 022	450	34 453	900	2,6%
	I edycja projektu rok 2013	27 682	630	17 145	630	44 827	1 260	2,8%
	I edycja projektu rok 2014	31 820	1 200	18 685	1 600	50 505	2 800	5,5%
	I edycja projektu rok 2015	45 506	1 452	32 616	1 473	78 122	2 925	3,7%
	Łącznie	127 439	3 732	80 468	4 153	207 907	7 885	3,8%

Źródło: opracowanie własne.

² W tabeli 1 jest to kategoria oznaczona jako „Łączenie w zbiorze” – ta kategoria odnosi się osobno do skojarzeń, jak i uzasadnień, jest też zsumowana w części tabeli opisanej jako „Łączna liczba wypowiedzi”.

³ Każda edycja badania oparta jest na realizacji wywiadów ankietowych: metodą tradycyjną (spotkanie ankietera z respondentem i rozmowa twarzą w twarz) oraz poprzez kwestionariusz dostępny online). Proporcja wyników w badaniach tradycyjnych i online zmieniała się z roku na rok z 25% do 10% na rzecz badań online.

Dalsza część rozważań poświęcona będzie poszukiwaniu odpowiedzi na sformułowane powyżej pytania. Kontekstem i przykładem dla tych dociekań będą dane zgromadzone podczas realizacji opisanego projektu. Najogólniej cel analityczny dotyczący pracy na wypowiedziach tekstowych brzmiał nastę-

pująco: w jaki sposób możliwy jest wybór wypowiedzi „wartościowych”/„cennych”? Cel ten pierwotnie realizowany był manualnie przez zespół koderów. Ich praca przyniosła wiele wzorców wypowiedzi, które uznane były za wartościowe i prezentowane były w raporcie końcowym z przeprowadzonych badań. Wyniki pracy koderów stały się podstawą do zaprojektowania zbioru algorytmów, które posłużyły do klasyfikacji opinii respondentów. Zadanie to realizowane było z zastosowaniem metod półautomatycznych, jak i automatycznych. W dalszej części rozważań opisane zostaną dwie wykorzystane w praktyce strategie analiz (w tabeli nr 1 – I strategia to „I etap analiz”, a II strategia to „II etap analiz”). Obie zakończyły się budową dwóch odrębnych klasyfikatorów treści.

Metodologia

Podczas czterech lat realizacji projektu zastosowanych zostało kilka strategii analiz oraz przetestowanych zostało kilka narzędzi analitycznych. Pierwszą i podstawową w początkowych edycjach projektu była strategia rozumiejącego czytania tekstów, manualnego kodowania i klasyfikacji wypowiedzi (tradycyjna strategia analizy treści). Wraz z przyrostem danych tekstowych stosowane były metody półautomatycznego i automatycznego kodowania tekstów. Etap klasyfikacji półautomatycznej oznaczał, iż wyniki pracy algorytmu weryfikowane były pod kątem ich trafności przez koderów. Etap weryfikacji za każdym razem przynosił uwagi krytyczne, które następnie formułowane były jako dodatkowe warunki logiczne, zgodnie z którymi działał automatyczny klasyfikator treści. Zastosowanie takiej logiki działania (automatyczna klasyfikacja →

weryfikacja → poprawa klasyfikacji automatycznej) było niezbędne do prowadzenia eksperymentów metodologicznych, które miały na celu odpowiedź na pytanie sformułowane w tej pracy (por. wcześniejsza strona, pytania: a, b, c).

Warto dodać jednak, iż w trakcie realizacji projektu strategii metodologiczne ewoluowały w każdym niemal obszarze związanym z opracowaniem danych tekstowych:

- Od manualnej pracy nad wyborem tekstów o wysokiej jakości do automatycznej eliminacji wypowiedzi mało wartościowych (kryteria eliminacji wypowiedzi omówione zostaną w dalszej części tekstu).
- Od strategii kodowania otwartego jednoetapowego do kodowania dwustopniowego (wstępnej klasyfikacji, a następnie kodowania zogniskowanego).
- Od pracy manualnej przy kodowaniu i klasyfikacji tekstów do klasyfikacji automatycznej i do metod półautomatycznych weryfikowanych przez koderów (I etap analizy oparty na danych z lat 2014–2015; I etap analizy oparty na danych z lat 2012–2015).
- Od pracy zorganizowanej liniowo, gdzie analizy manualne poprzedzają automatyczne do pracy przebiegającej symultanicznie na kilku zadaniach analitycznych równocześnie (podejście specyficzne dla tak zwanych metody zwinnych).
- Od stosowania zamkniętych programistycznie narzędzi CAQDAS do stosowania narzędzi

pozwalających na samodzielne projektowanie: reguł klasyfikacji, funkcji ważenia, miar określających jakość uzyskanych rezultatów – tak zwana 5 generacja narzędzi CAQDAS (Tomanek 2014b).

Opisane powyżej kierunki zmian strategii były wynikiem dwóch obserwacji. Po pierwsze, już po zakończeniu pierwszej edycji projektu wiadome było, że dwa analizowane typy wypowiedzi różnią się w sposób znaczący. Uzasadnienia osiągają w najlepszych przypadkach kilkanaście słów (przeciętna długość zdania złożonego). Skojarzenia są natomiast wypowiedziami krótkimi w postaci jednego słowa, frazy, kilku słów (niezmiernie rzadko zdarzały się wypowiedzi zapisane w formie pełnego zdania). Te dwa typy wypowiedzi wymagają zmian zarówno w doborze algorytmów przeszukujących treści, jak i w sposobie ich wykorzystania. Te dwie lekcje wyciągnięte z procesu diagnostyki wypowiedzi zmieniają również proces prowadzenia analiz. Zmiany zastosowane w analizie opisane są poniżej.

Ad 1.

Praca nad wyborem wypowiedzi skupiała się pierwotnie na podejmowaniu decyzji opartych na czytaniu wypowiedzi. Szybko jednak można było się przekonać, że zarówno wśród skojarzeń, jak i uzasadnień pojawiają się wypowiedzi będące wraz z emocjami niezwiązanych z treścią pytania zadane w badaniu; zapisem przypadkowych ruchów palców po klawiaturze (np. ciągi typu „dsdsdssds”, „eqweweqw”). W związku z możliwością wychwycenia niektórych z tych nietypowych wypowiedzi skonstruowany został algorytm, który je elimino-

wał automatycznie. Podstawowe warunki pracy takiego algorytmu zakładały:

1. Wykluczenie wszystkich obserwacji zawierających mniej niż 5 znaków (najkrótsze spośród zidentyfikowanych słów, które okazywały się atrakcyjne, składały się z więcej niż pięciu znaków, na przykład – skojarzenie: piękna. Decyzja ta przeszła przez kilka testów, w których koder oceniał eliminowane i klasyfikowane przez algorytm wyniki.
2. Eliminacja wszystkich wypowiedzi, w których pojawił się przynajmniej jeden wulgaryzm (w tym celu zastosowane zostało tak zwane podejście słownikowe, które w tekstach wyszukuje słowa uwzględnione w słowniku) (Bolasco, Ratta-Rinaldi 2004; Tomanek 2014a).
3. Wykluczenie wszystkich obserwacji, w których nie pojawiło się przynajmniej jedno słowo możliwe do lematyzacji.
4. Eliminacja wypowiedzi, które bez spacji zawierają więcej niż 13 znaków (przykładem wypowiedzi, która bliska jest tej granicy i jest klasyfikowana może być słowo: „fantastycznie”. Przykładem ciągu znaków eliminowanych może być: sadadasdadassad.

Ad 2.

Kodowanie otwarte ma ten niewątpliwy walor, że dostarcza do dalszych analiz większej dawki tekstu wraz z kontekstem, w którym pojawia się kluczowy, najistotniejszy dla wypowiedzi fragment. O ile jednak strategia ta jest cenna w pracy nad tekstami

literackimi, zapisami indywidualnych wywiadów pogłębionych lub wywiadów grupowych, o tyle staje się kłopotliwa dla zadań skupionych na automatyzacji analiz, w szczególności dla tych, które prowadzone są na krótkich wypowiedziach w formie równoważników zdań, fraz. Kodowanie dwustopniowe pozwala na precyzyjniejszy wybór tekstów przeznaczonych do prezentacji w raporcie.

Ad 3.

Manualna praca koderów w trakcie kolejnych edycji projektu trwała coraz dłużej. Od jednego tygodnia w pierwszym projekcie do trzech w ostatniej, czwartej edycji. Zastosowanie automatycznych metody czyszczenia tekstów oraz ich wstępnej klasyfikacji (ad 1) pozwoliło na skrócenie czasu pracy bez straty jakości wyników. Bardziej ambitne zadanie polegające na identyfikacji i klasyfikacji „cennych” wypowiedzi prowadzić miało do w pełni automatycznej klasyfikacji. Zamysł ten (I etap oparty na danych z lat 2014–2015) realizowany był w następujących etapach:

1. Analiza danych z lat 2014–2015 i wyodrębnienie wypowiedzi eksperckich jako wzorcowych wypowiedzi stosowanych w procesie uczenia klasyfikatora automatycznego.
2. Budowa zbioru uczącego (do szkolenia algorytmu klasyfikacyjnego) i testowego (do testowania trafności klasyfikacji), na których prowadzone były analizy.

Wstępne prace nad algorytmem, który automatycznie „czyścił” dane tekstowe z wypowiedzi

niepożądanych, przynosiły wyniki podobne do wyborów losowych. Na 100 wybranych wypowiedzi (w kilku niezależnych losowaniach) wyniki uzyskiwały trafność na poziomie 48% do 51% (4 lub 5 wypowiedzi na 10 nadawało się do prezentacji w raporcie), a pozostałe wymagały usunięcia z analiz. Tak skonstruowana metoda ani nie dawała wartościowych wyników (lepiej niż wybór losowy), ani nie skracala czasu pracy. Stąd też konieczna była poprawka do strategii opartej na pełnej automatyzacji. Ta zmiana oparta została na rozwinięciu podejścia półautomatycznego. Polegało ono na weryfikacji nietrafnie wybranych przez algorytm wypowiedzi oraz wypowiedzi odrzuconych przez algorytm. Wyniki tej pracy pozwoliły na wprowadzenie praktycznych zmian w budowie metody klasyfikującej wypowiedzi (zmiany te opisane są w dalszej części tekstu).

Warto podkreślić, że sukces wskazanej tu strategii opiera się na konstrukcji algorytmu, który czerpie z wniosków dostarczanych przez koderów. Analiza oparta na tych wnioskach rozwijana była na próbie tekstów (zbiór uczący), a następnie wyniki uczenia algorytmu weryfikowane były na zbiorze wypowiedzi wcześniej nieanalizowanych (zbiór testowy). Weryfikacja obu tych analiz (uczenie, testowanie) dała lepszą kontrolę nad budową metody klasyfikacji, a co za tym idzie – nad jakością uzyskanych wyników.

Ad 4.

W pracy z algorytmami uczącymi się i metodami półautomatycznymi kluczowa jest możliwość realizacji badań nad tekstem w sposób symultaniczny

(z zastosowaniem różnych metod analiz na tym samym zbiorze tekstów). W związku z tym przyjęto założenie, iż prace nad automatycznym klasyfikatorem prowadzone będą z zastosowaniem równolegle kilku metod. Wśród nich znalazły się: regresja logistyczna, regresja logistyczna karana – typu Lasso, drzewa decyzyjne (CART), Support Vector Machines (C-SVM oraz One-Class SVM), Naiwny Klasyfikator Bayesowski. Praca kilku osób w jednym czasie nad tym samym zagadnieniem wymagała również sprawnej wymiany wiedzy i wzajemnego informowania się o postępach (sukcesach, jak i porażkach) w testowaniu różnych algorytmów. Tak zrodziła się potrzeba stosowania metodyk zwinnych (*agile methods*, na przykład metodyki *scrum*) (Schwaber 2013). To zwinne podejście w prowadzeniu prowadziło analityków od analiz jakościowych do analiz ilościowych, i ponownie od podejścia *quan* do *qual*.

Ad 5.

Zastosowanie automatycznej klasyfikacji wypowiedzi na te poddawane dalszej analizie i te eliminowane z analiz pozwoliło na zastosowanie podstawowych technik stosowanych w ramach *text mining*. Były to:

1. Parsowanie – unifikacja struktury tekstu: dekompozycja danych tekstowych, ilościowa reprezentacja zbioru dokumentów.
2. Transformacja i redukcja wymiarów: transformacja reprezentacji tekstu do formy ilościowej; redukcja wymiarów do zwartego formatu informacyjnego.

3. Analiza: zastosowanie algorytmów analizujących tekst i budujących reguły klasyfikacji.

Aby możliwe było zrealizowanie wskazanych prac na tekście, konieczne było stosowanie otwartych narzędzi analiz tekstu. Potrzebne było zatem oprogramowanie piątej generacji CAQDAS. W przypadku tego projektu wybór padło na program R⁴. To rozwiązanie programistyczne daje przede wszystkim możliwości rozwoju i implementacji różnorodnych algorytmów z zakresu analiz jakościowych.

Opis badań i analiz

Kiedy analiza dotyczy kilkudziesięciu tysięcy wypowiedzi, pojawia się pokusa, by zastosować metody, które wykonają zadanie analityczne „częściowo” za nas – badaczy. Wśród metod tych istnieją dwa charakterystyczne podejścia (Sołdacki 2006). Są to:

- a. Głęboka analiza tekstu (ang. *Deep Text Processing*, DTP): to podejście opiera się na komputerowej analizie lingwistycznej wielu możliwych interpretacji, powiązań między słowami, frazami, relacji gramatycznych występujących w tekście. Z powodu tych „wielu możliwych interpretacji” taka analiza nie zawsze jest potrzebna, a także możliwa do osiągnięcia w czasie, jaki mamy na nią przeznaczony, czy też w oparciu o narzędzia, jakimi dysponujemy. Z tych też powodów coraz częściej realizowana jest częściowa czy płytka analiza tekstu.
- b. Płytko analiza tekstu (ang. *Shallow Text Processing*, STP): efekt tej analizy jest niepełny w stosunku do

rezultatów DTP (Piskorski 2001). Analiza tego typu rozpoznaje: słowa i ich odmiany; zapis w liczbie mnogiej i pojedynczej; przymiotniki; nazwy własne; podmiot, do którego wypowiedź się odnosi oraz identyfikuje jego rodzaj. Pomijane są tu bardziej złożone problemy, takie jak: rozpoznawanie ironii, identyfikacja emocji, metafory.

Zadanie, które jest tu opisywane (identyfikacja „cennych” skojarzeń i uzasadnień), nie jest takim, które wymaga analizy DTP. Pożądanym wynikiem pracy nad wypowiedziami jest wyodrębnienie tych, które wnoszą „istotną” informację o podmiocie wypowiedzi. Istotnym, dla przykładu, nie jest określenie: „wydarzenie było super” (jest to wypowiedź raczej trywialna ze względu na wartość informacyjną), ale już „takie wydarzenia zawsze gromadzą liczną publiczność” wnosi treść istotną z punktu widzenia oceny wydarzenia. Do identyfikacji krótkich wypowiedzi wystarczająca jest często analiza STP, która z kolei realizowana była na dwa odmienne sposoby (Tomanek 2014c):

- a. Analiza oparta na metodach słownikowych: w tej strategii skupiamy się na identyfikacji słów (Key-Word-in-Context), fraz (Key-Phrase-in-Context), słów z określonych obszarów tematycznych (Bag-of-Words). Identyfikacji wypowiedzi istotnych służą reguły logiczne, syntaktyczne i uproszczona analiza kontekstu, w jakim słowa i frazy występują.
- b. Metody statystyczne: w tym przypadku skupiamy się na automatycznym przetwarzaniu treści w oparciu o przyjęte uprzednio wzory wypowiedzi

dzi pożądaných (czyli takich, jakich szukamy) lub też posługujemy się automatyczną klasyfikacją bezwzorcową.

Oba zarysowane podejścia znacząco ograniczają zakres analizy tekstu. Dla zadania, z jakim się tu mierzymy, jest to ich silna strona. Oba jednak mogą dostarczać znacząco odmiennych wyników. Oto kilka najważniejszych powodów pozwalających uzasadnić ten wniosek:

1. O ile analiza słownikowa może czerpać z analizy tematycznej – to znaczy bierze pod uwagę znaczenie słów (posługuje się słowami sklasyfikowanymi w ramach określonych tematów), o tyle analiza statystyczna może w ogóle nie sięgać do znaczeń. Podstawą analizy statystycznej może stać się tylko i wyłącznie liczbowa reprezentacja słowa, na przykład długość słowa albo jego unikalność (rzadkie występowanie) w różnych wypowiedziach.
2. Analiza słownikowa: sięgano wiedzę ekspercką (czyli po wypowiedzi zidentyfikowane przez koderów), ale też rozszerza ją o dodatkowe możliwe wypowiedzi. Dzieje się to przez rozbudowanie algorytmu przeszukującego tekst o frazy, które są równoważne znaczeniowo słowom kluczowym stosowanym we wcześniejszych analizach (regułą podstawową dla rozszerzania zbioru słów kluczowych jest synonimia); innym rozszerzeniem jest identyfikacja wypowiedzi biegunowo różnych (regułą pozwalającą na identyfikację słów znaczeniowo-biegunowo różnych jest antonimia).
3. Analiza słownikowa: posługuje się regułami logicznymi pozwalającymi na analizę kontekstu,

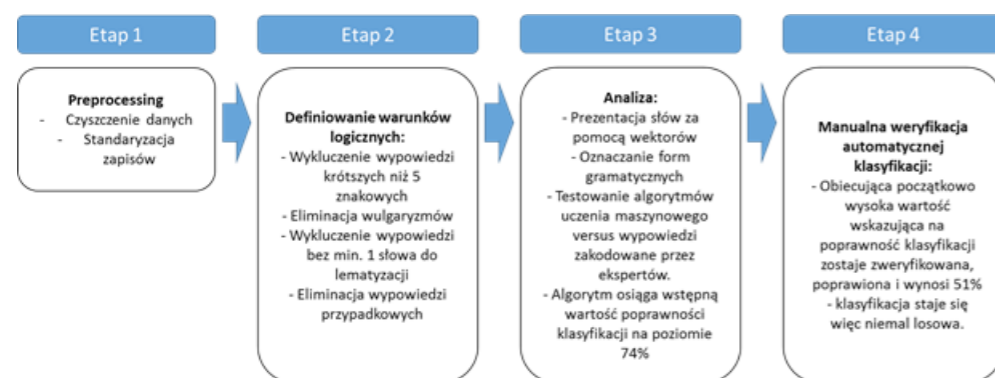
⁴ Zob. <https://www.rstudio.com/>.

w jakim słowa i frazy występują. Dzięki prostym zapisom uwzględniającym operatory logiki możemy zidentyfikować: wypowiedzi sprzeczne znaczeniowo; wypowiedzi o konkretnych miastach, osobach, wydarzeniach (Gonzalez, Dankel 1993). Analiza statystyczna, posługując się takimi miarami jak korelacja, unikalność słów, pomaga w znalezieniu pewnych wzorów mówiących o występowaniu słów w tekście, ale nie podpowiada niczego w kontekście znaczeń

analizowanych treści. W szczególności analiza statystyczna podpowiadać może zależności gramatyczne pomiędzy elementami wypowiedzi.

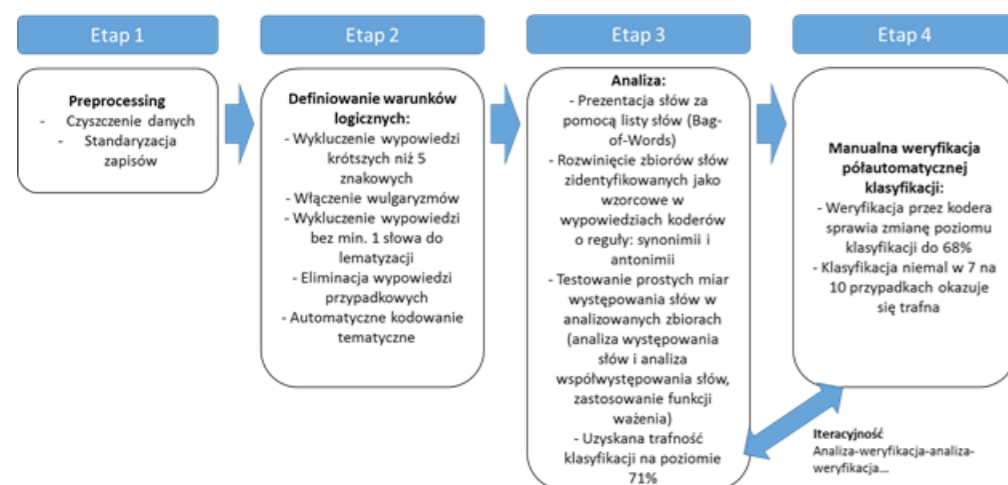
Wskazane powyżej powody stały u podstaw przeprowadzenia dwóch niezależnych analiz. Jednej – analizy statystycznej opartej na metodach wskazanych wcześniej w tekście (por. ad 4) oraz na metodzie słownikowej. Obrazowo przebieg obu analiz można zaprezentować tak, jak na diagramach 2 i 3.

Diagram 2. Przebieg procesu analiz z zastosowaniem analiz statystycznych



Źródło: opracowanie własne.

Diagram 3. Przebieg procesu analiz z zastosowaniem analizy słownikowej



Źródło: opracowanie własne.

Etapy 1, 2 i 4 opisane zostały we wcześniejszych fragmentach tekstu. Etap 3 to wielowymiarowa analiza tekstu. Na tym etapie każdej analizowanej wypowiedzi przypisane zostają wartości liczbowe, które oznaczają: liczbę słów w wypowiedzi, unikalność słowa (brak słowa we wzorcowych wypowiedziach wskazanych przez koderów), fakt wystąpienia słowa we wzorcowych wypowiedziach wskazanych przez koderów, wystąpienie w wypowiedzi wyrazu identyfikującego nazwę usługodawcy lub produktu.

Etap 3 w tej strategii – poza wskazanymi w analizie statystycznej – uwzględniał dodatkowo następujące zabiegi analityczne: zastosowanie stop listy (listy słów nieistotnych), identyfikacja wypowiedzi zawierających słowa oceniające (plus ich synonimy), identyfikacja wypowiedzi zawierających minimum dwa słowa o przeciwnym wydźwięku (wypowiedzi o zabarwieniu ambiwalentnym), identyfikacja wypowiedzi oceniających wraz z przeczeniem.

Dwie zarysowane powyżej strategie dały znacząco odmienne wyniki w kontekście poprawności klasyfikacji. W przypadku automatycznych analiz statystycznych poprawność wyboru wypowiedzi wyniosła 74% i została zweryfikowana przez koderów do 51%. Klasyfikacja stosująca metodę słownikową osiągnęła poziom poprawności wyboru równy 71%. Po weryfikacji wynik ten utrzymał się na poziomie 68%. Kluczowym w tej strategii okazała się iteracyjność przebiegu weryfikacji: praca koderów kończyła się zaleceniami wskazującymi, jak należy zmienić klasyfikator automatyczny – po zmianie dokonanej w algorytmie następowała kolejna ocena dokonywana przez koderów oraz kolejna automatyczna klasyfikacja. Strategia ta wykonana została trzy-

krotnie. Dzięki temu zabiegowi wzrosła trafność identyfikacji wypowiedzi bliskich znaczeniowo, a także biegunowo odmiennych. Ta podstawowa przewaga jednak nie daje 100% trafności klasyfikacji. Idealna klasyfikacja na poziomie 100% możliwa jest do osiągnięcia dla: języków nienaturalnych (na przykład dla języków skryptowych, języków programowania); wypowiedzi w dobrze opisanym języku (być może na przykład w łacinie); wypowiedzi silnie sformalizowanych (być może np. instrukcje obsługi urządzeń mechanicznych). Mimo iż nie udało się osiągnąć tego idealnego poziomu trafności klasyfikacji, wydaje się, że metoda półautomatyczna wykorzystująca strategię analizy słownikowej oraz weryfikację wykonaną przez koderów pozwala na redukcję prac manualnych w pracy nad tekstami – co można poczytywać za sukces tego podejścia.

Wnioski

Zastosowanie metod półautomatycznej i automatycznej w analizie treści niesie ze sobą nie tylko sukcesy poznawcze, ale także szereg porażek i problemów. Wśród korzyści zastosowania metody półautomatycznej wymienić można następujące: za podstawę przyjmuje ona opracowanie tekstu oparte o tradycyjne podejście powszechne w analizie treści (kodowanie tekstu przez człowieka) i jakościową analizę danych; daje możliwość wykonania szybkiej analizy na dużych zbiorach wypowiedzi; jest procesem iteracyjnym, co pozwala na zmniejszenie liczby błędów w regułach klasyfikacji oraz zwiększenie trafności klasyfikacji; daje możliwość stosowania istniejących narzędzi klasyfikacyjnych jako elementów inspirujących analizę lub wzbogacających klasyfikację (na przykład listy słów specyficzne dla danych

tematyk – inspiracją jest tu więc analiza tematyczna; jest punktem wyjścia do wielowymiarowych analiz treści (to w związku z możliwością liczbowej reprezentacji wypowiedzi); jest podejściem rozwijającym się w ramach programu badawczego (wtedy, gdy jeden słownik klasyfikacyjny rozwijany jest w różnorodnych projektach i testowany na różnych zbiorach tekstów) – w efekcie pozwala na ewolucję słowników klasyfikujących, które stosowane mogą być dla nieeksplorowanych jeszcze tekstów (zbiorów testowych).

Wybrane w toku analiz podejście wiąże się jednak z pewnymi problemami: wybór kategorii analitycznych trafnie odzwierciedlających analizowaną treść nie może opierać się tylko na analizie frekwencji słów i fraz, ale wymaga również zastosowania odmian słów, zapisów słów w różnych rodzajach oraz czasach – zabieg taki zwiększa czasochłonność analiz oraz wymaga wydajnych komputerów z nowoczesnymi procesorami; opracowanie tekstów zgodnie z zasadami *preprocessingu* danych tekstowych jest procesem czasochłonnym; konieczność wielokrotnego wykonywania analiz (iteracyjność) zwiększa czasochłonność procesu analitycznego.

Analiza tekstów opierająca się o metodę, jaką stanowi słownik klasyfikacyjny, daje dobre wyniki wtedy, gdy: po pierwsze, realizowana jest jako proces iteracyjny (czasochłonny, ale zwiększający trafność klasyfikacji); po drugie, wykorzystuje strategię mieszane (*bottom up, top down*); po trzecie, jest metodą nadzorowaną i opiera się na wielokrotnej weryfikacji.

Zaznaczyć należy, że przeprowadzony eksperyment i analiza obejmowały teksty stosunkowo proste językowo,

składniowo i stylistycznie. Analizowany język zapisany był w naturalnej formie (spontanicznych, nieustrukturyzowanych) wypowiedzi. Łatwo poddawał się analizie (był wycyszczony z błędów zapisu, rozpoznawane były wszystkie formy wyrazów – bez względu na odmianę, rodzaj, liczbę). Mimo osiągniętego „sukcesu poznawczego” (klasyfikacja na poziomie 68%), trudno ogłosić tu sukces metody. Nie jest bowiem możliwe ekstrapolowanie osiągniętego rezultatu na zadania, w których mielibyśmy dokonać klasyfikacji wypowiedzi dla języków niszowych, slangowych, subkulturowych, emocji czy choćby ustrukturyzowanych dłuższych lub językowo trudniejszych wypowiedzi. Takimi wypowiedziami są te zawierające żart, parafrazę, metaforę, aforyzm (Weizenbaum 2008). Dla takich wypowiedzi niezbędne byłoby opracowanie dodatkowych rozwiązań (Forslid, Wiken 2015).

Podsumowując powyższe rozważania, odpowiedzi na postanowione w artykule pytania można sformułować w sposób następujący:

a. Czy część najbardziej czasochłonnej pracy wykonywanej przy analizie tekstów można w jakimś stopniu zautomatyzować?

Wykonane eksperymenty na analizowanych wypowiedziach pokazują, że stosunkowo prostą jest eliminacja wypowiedzi niewnoszących wartości do raportów badawczych. Aby cel ten osiągnąć, niezbędnym jest wykonanie diagnostyki tekstów skupionej na: formie zapisu stosowanego języka, występowaniu nieznaczących ciągów znaków odtwarzanych przez wielokrotne przyciskanie klawiszy występujących blisko na klawiaturze. Trud-

niejszym zadaniem jest natomiast zidentyfikowanie wypowiedzi ambiwalentnych, eliminacja złożonych i długich wypowiedzi, które nie są wartościowe dla dalszych analiz. Istotnym na tym etapie jest przejrzanie korpusu wypowiedzi przez koderów, który może sformułować wstępne warunki eliminacji wypowiedzi z dalszych analiz.

b. Czy w wyniku automatyzacji uzyskamy dane o takiej samej jakości jak w przypadku analiz wykonywanych manualnie? Czy automatyzacja opracowania danych da w efekcie trafne rezultaty?

Automatyczna analiza wypowiedzi tekstowych formułowanych w języku naturalnym prowadzona bez nauczyciela (wsparcia wiedzy koderów) wydaje się być – na tym etapie rozwoju wiedzy – skazana

na błąd klasyfikacji. Dlatego też udział koderów w analizach tekstów wydaje się warunkiem *sine qua non* do osiągnięcia wysokiego poziomu trafności klasyfikacji.

c. Czy metoda i narzędzie wykonujące pracę automatycznie mogą korzystać z wiedzy badaczy, koderów?

Im więcej wiedzy ludzkich koderów wykorzystane zostanie na etapie projektowania klasyfikatorów półautomatycznych, tym większa jest trafność klasyfikacji. Koderzy stanowią także bardzo cenne źródło wiedzy na etapie weryfikacji wyników metody półautomatycznej. Pytanie o to, jaki zakres wiedzy możliwy jest do opisu za pomocą algorytmów pozostaje nadal otwartym.

Bibliografia

Acerbi Alberto i in. (2013) *The Expression of Emotions in 20th Century Books*. „PLoS ONE”, vol. 8, no. 3, s. 1–6.

Appel Markus, Richter Tobias (2010) *Transportation and Need for Affect in Narrative Persuasion: A Mediated Moderation Model*. „Media Psychology”, vol. 13, s. 101–135.

Bolasco Sergio, Ratta-Rinaldi della Francesca (2004) *Experiments on Semantic Categorisation of Texts: Analysis of Positive and Negative Dimension*. „JADT 2004: 7es Journées internationales d'Analyse statistique des Données Textuelles” [dostęp 1 maja 2014 r.]. Dostępny w Internecie: http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_018.pdf.

Bolc Leonard, Jerzy Cytowski (1989–1991) *Metody przeszukiwania heurystycznego, t. 1–2*. Warszawa: PWN.

Cichosz Paweł (2000) *Systemy uczące się*. Warszawa: WNT.

Cole Ron i in. (1996) *Survey of the State of the Art in Human Language Technology*. Cambridge University Press [dostęp 5 maja 2017 r.]. Dostępny w Internecie: <http://www.dfki.de/~hansu/HLT-Survey.pdf>.

DeWall Nathan C. i in. (2011) *Tuning in to Psychological Change: Linguistic Markers of Psychological Traits and Emotions over Time in Popular U.S. Song Lyrics*. „Psychology of Aesthetics, Creativity, and the Arts”, vol. 5, no. 3, s. 200–207.

Duch Włodzisław (1997) *Fascynujący świat programów komputerowych*. Poznań: Wydawnictwo Nakom.

Duggleby Wendy (2005) *What about Focus Group Interaction Data?* „Qualitative Health Research”, vol. 15, no. 6, s. 832–840.

Eder Maciej (2014) *Metody ścisłe w językoznawstwie i pułapki pozornego obiektywizmu. Przykład stylometrii*. „Teksty Drugie”, t. 2, s. 90–105.

EderMaciej, Piasecki Maciej (2015) *System do klasyfikacji tekstu i analizy stylometrycznej, referat wygłoszony podczas warsztatów CLARIN* [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://clarin-pl.eu/pliki/warsztaty/Stylometria%20i%20klasyfikacja%20-%20warsztaty.ppt>>.

Forslid Erik, Wiken Niklas (2015) *Automatic Irony and Sarcasm Detection in Social Media, UPPTEC F 15045 Examensarbete 30* [dostęp 30 listopada 2016 r.]. Dostępny w Internecie: <<http://uu.diva-portal.org/smash/get/diva2:852975/FULLTEXT01.pdf>>.

Gonzalez Avelino J., Dankel Douglas D. (1993) *The Engineering of Knowledge-Based Systems: Theory and Practice*. Upper Saddle River, NJ: Prentice-Hall International.

Jordan Joanne i in. (2007) *Using Focus Groups to Research Sensitive Issues: Insights from Group Interviews on Nursing in the Northern Ireland "Troubles"*. „International Journal of Qualitative Methods”, vol. 6, no. 4 [dostęp 14 kwietnia 2017 r.]. Dostępny w Internecie: <<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.533.61&rep=rep1&type=pdf>>.

Kaplowitz Michael D. (2000) *Statistical Analysis of Sensitive Topics in Group and Individual Interviews*. „Quality & Quantity”, vol. 34, s. 419–431

Kidd Pamela S., Parshall Mark B. (2000) *Getting the Focus and the Group: Enhancing Analytical Rigor in Focus Group Research*. „Qualitative Health Research”, vol. 10, no. 3, s. 293–308.

Lieberman Erez i in. (2007) *Quantifying the Evolutionary Dynamics of Language*. „Nature”, vol. 449, no. 7163, s. 713–716.

Liu Bing (2012) *Sentiment Analysis and Opinion Mining* [dostęp 1 maja 2014 r.]. Dostępny w Internecie: <www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>.

Michael Jean Baptiste (2011) *Quantitative Analysis of Culture Using Millions of Digitized Books. Program for Evolutionary Dynamics*. Cambridge: Harvard University.

Niedbalski Jakub (2013) *CAQDAS – oprogramowanie do komputerowego wspomaganie analizy danych jakościowych. Historia, ewolucja i przyszłość*. „Przegląd Socjologiczny”, t. 62, nr 1, s. 153–166.

Niedbalski Jakub, red. (2014) *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analiz danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Pagel Mark, Atkinson Quentin D., Meade Andrew (2007) *Frequency of Word-Use Predicts Rates of Lexical Evolution Throughout Indo-European History*. „Nature”, vol. 449, s. 717–720.

Piskorski Jakub (2001) *Shallow Text Processor for Information Extraction from Free-Text Business Documents*. Poznań: Akademia Ekonomiczna w Poznaniu.

Rosa Krzysztof (2015) *Listy pożegnalne niedoszłych samobójców. Analiza treści*. „Przegląd Socjologiczny”, t. 64, nr 4, s. 103–128.

Rybicki Jan, Heydel Magdalena (2013) *The Stylistics and Stylometry of Collaborative Translation: Woolf's "Night and Day" in Polish*. „Digital Humanities 2012: digital diversity: cultures, languages and methods”, vol. 28, no. nr 4, s. 708–717.

Schanie Carrie L., Pinto-Foltz Melissa D., Logsdon Cynthia M. (2008) *Analysis of Popular Press Articles Concerning Postpartum Depression: 1998-2006*. „Issues Ment. Health Nurs.”, vol. 29, no. 11, s. 1200–1216.

Schwaber Ken (2013) *Scrum Gide* [dostęp 20 marca 2017 r.]. Dostępny w Internecie: <<http://www.scrumguides.org/docs/scrumguide/v1/Scrum-Guide-PL.pdf>>.

Smith Marc in. (2009) *C&T '09: Proceedings of the Fourth International Conference on Communities and Technologies* [dostęp 5 maja 2017 r.]. Dostępny w Internecie: <<http://www.connectedaction.net/wp-content/uploads/2009/08/2009-CT-NodeXL-and-Social-Queries-a-social-media-network-analysis-toolkit.pdf>>.

Soldacki Przemysław (2006) *Zastosowanie metod płytkiej analizy tekstu do przetwarzania dokumentów w języku polskim*. Niepublikowana praca doktorska, Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych, Instytut Informatyki [dostęp 27 listopada 2017 r.]. Dostępny w Internecie: <<https://www.google.pl/url?sa=t&rc=1&source=web&cd=1&ved=0ahUKEwjl7q5lMvQAhXDFswKHfNPAPwQFggkMAA&url=https%3A%2F%2Frepo.pw.edu.pl%2Fdocstore%2Fdownload.seam%253Bsessionid%3DF74241A1317DC5E22F87A22B33BE1F6F%3Ffiled%3DWEIT-b192c072-00cc-41df-9bba-a2b0a211e9bc&usq=AFQjCNH-0laWKSvcxkvp6FNAqRhpA-HuKr0A&bv=139782543dbGg&cad=rja>>.

Taboada Maite i in. (2011) *Lexicon-Based Methods for Sentiment Analysis*. „Journal of Computational Linguistics”, vol. 37, no. 2, s. 267–307.

Tomanek Krzysztof (2014a) *Analiza sentymentu: historia i rozwój metody w ramach CAQDAS* [w:] Niedbalski Jakub, red., *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analiz danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 155–172.

Tomanek Krzysztof (2014b) *Jak nauczyć metodę samodzielności* [w:] Niedbalski Jakub, red., *Metody i techniki odkrywania wiedzy. Narzędzia CAQDAS w procesie analiz danych jakościowych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 174.

Tomanek Krzysztof (2014c) *„Analiza sentymentu” – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych*. „Przegląd Socjologii Jakościowej”, t. 10, nr 2, s. 118–136.

Tomanek Krzysztof (2015) *Spoleczność fanów nauki w świecie wirtualnym. Analiza eksploracyjna treści i aktywności społeczności inter-*

netowej skupionej wokół fanpage'a „I fucking love science”. „Edukacja Humanistyczna”, nr 1(32), s. 123–138.

Tomanek Krzysztof (2016) *Analiza argumentacji. Praktyczne implikacje zastosowania modelu argumentacji Stephena Toulmina do analiz danych tekstowych* [w:] Wojciech Doliński i in., red., *Rzeczywistość i zapis. Problemy badania tekstów w naukach społecznych i humanistycznych*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, s. 229–242.

Tomanek Krzysztof, Bryda Grzegorz (2015) *Odkrywanie postaw dydaktyków zawartych w komentarzach studenckich. Analiza treści z zastosowaniem słownika klasyfikacyjnego*. „Przegląd Socjologiczny”, t. 64, nr 4, s. 51–81.

Weizenbaum Joseph (2008) *Moglibyśmy mieć raj*. „Forum”, nr 28, s. 28–29.

Zaśko-Zielińska Monika (2014) *Ślady idiolektów w listach pożegnalnych zgromadzonych w Polskim Korpusie Listów Pożegnalnych Samobójców* [w:] R. Cieśla, red., *Dokument i jego badania*. Wrocław: Katedra Kryminalistyki Wydziału Prawa, Administracji i Ekonomii Uniwersytetu Wrocławskiego, s. 425–435.

Cytowanie

Tomanek Krzysztof (2017) *Metodyka dla analizy treści w projektach stosujących techniki textmining i rozwiązania CAQDAS piątej generacji*. „Przegląd Socjologii Jakościowej”, t. 13, nr 2, s. 128–143 [dostęp dzień, miesiąc, rok]. Dostępny w Internecie: <www.przegladsocjologiijakosciowej.org>.

Methodology for Content Analysis in Text Mining Projects and Fifth Generation of CAQDAS

Abstract: Projects which we work with—large volumes of text data that are acquired from various sources and stored in a variety of formats—rise many dilemmas of a methodological nature, often require unstandardized decisions and solutions. In particular, compiling data of various quality, unstructured types, and of quan and qual nature requires dynamic strategies, ideas, and ways of analysis. The article describes an example of this approach. It shows its value in classification of written statements. In such context, the author of the article faces the following objectives: (a) can we use CAQDAS so that semiautomatic or automatic work would replace some manual work regarding classification of the expressions; (b) how to construct a classification methodology for data of various quality; (c) when the automatic classification is useful and when there is no chance of success?

The article will be marked with moments in which the analyst reaches for knowledge typical for qualitative data analysis, and when the knowledge of this area is no longer sufficient to classify content (natural language processing, machine learning). An example of a project being the background of this article forced the use of several tools and languages to support work with the data. Work on the transformation, classification, and visualization of results required applications such as: MySQL, R, QDA Miner, WordStat, Qlik Sense. Role and limits of the computer-assisted qualitative data analysis software tools have also been noted.

Keywords: Content Analysis, Mixed Methods Approach, Big Data, Text Mining, CAQDAS

WYNIKI VII EDYCJI KONKURSU FOTOGRAFICZNEGO „PRZEGLĄDU SOCJOLOGII JAKOŚCIOWEJ”

Drodzy Czytelnicy,

z przyjemnością prezentujemy Państwu prace laureatów VII edycji konkursu fotograficznego „Przełądu Socjologii Jakościowej” odbywającego się pod hasłem „Wszystko na sprzedaż”.

Pierwsza ze zwycięskich prac pt. „Opętanie”, autorstwa Pana Michała Bieleckiego, ukazuje sytuację, w której rytualny, transowy taniec mieszkańców Jawy zostaje potraktowany jako egzotyczna pamiątka, którą warto pochwylić obiektywem aparatu i pochwalić się nią w gronie internetowych znajomych.

Drugi esej, którego twórcą jest Pan Przemysław Wrochna, pt. „Ziemia obiecana”, to nostalgiczna opowieść o tęsknocie emigrantów za utraconym światem: bliskimi, ojczyzną, dotychczasową tożsamością, poczuciem bezpieczeństwa ontologicznego.

Członkowie Jury zwrócili uwagę, że Autorzy obu zwycięskich prac w nieszablonowy sposób potraktowali tegoroczne motto konkursowe – „Wszystko na sprzedaż”. Zwrócono również uwagę na formalną stronę nagrodzonych prac.

Gratulujemy Laureatom, a wszystkim Uczestnikom dziękujemy za wzięcie udziału w siódmej edycji konkursu fotograficznego „Przełądu Socjologii Jakościowej”. Zapraszamy również Państwa do uczestnictwa w kolejnych edycjach.

Wzorem lat ubiegłych w zakładce „Konkurs” znajdującej się na głównej stronie PSJ będziemy nadal zamieszczać ogłoszenia zapraszające do kolejnych edycji, publikować na bieżąco wyniki konkursu oraz nagrodzone i wyróżnione prace.

W imieniu Jury, Waldemar Dymarczyk

Opętanie

Dyptyk fotograficzny

Zdjęcia zostały wykonane w 2015 roku w Yogyakarcie (Jawa, Indonezja). Przedstawiają rytualny taniec transowy zwany *Jathilan*, który swoimi korzeniami sięga czasów królestwa Majapahit (1293–1500). Obecnie stał się on popularną formą rozrywki dla mieszkańców Jawy. Podczas rytuału ciała tancerzy nawiedzane są przez duchy, które następnie obecni na scenie szamani, zwani *paawang*, przeganiają z pomocą swoich niezwykłych umiejętności i magicznych przedmiotów.

Człowiek z telefonem na drugim zdjęciu bardzo chciał zamieścić na popularnym serwisie społecznościowym nagranie z dobrze widocznym opętanym tancerzem. Miejsce wśród publiczności nie gwarantowało mu mocnego ujęcia, wszedł więc na scenę. Swoją zdobycz wkrótce zamieni na kilkadziesiąt lajków.

Michał Bielecki



„Ziemia obiecana”

Emigracja, będąca rodzajem migracji zewnętrznej (międzynarodowej), jest zagadnieniem od dawna rozważanym na gruncie polskiej socjologii – zarówno w odniesieniu do badań empirycznych, jak i wtedy gdy mamy na myśli refleksje *stricto* teoretyczne. Decyzje dotyczące emigracji są procesem złożonym, obejmującym wiele czynników oraz niosą ze sobą szereg konsekwencji (pozytywnych i negatywnych). Dotyczą one w głównej mierze nadziei na nowe, lepsze życie, niestety jednak wiążą się także z koniecznością zostawienia za sobą poczucia bezpieczeństwa ontologicznego i tożsamości, które dawali im bliscy, a także ojczysty kraj.

„Wszystko na sprzedaż” Andrzeja Wajdy przedstawia taki właśnie wyraz refleksyjnej opowieści, zarówno o utracie bliskiej osoby, ale i wartości, które dany człowiek ze sobą niósł. Ów film nie jest jednak wyłącznie melancholijnym spojrzeniem w przeszłość – zadumą nad rzeczywistością, z którą należy się pożegnać. Jest on przede wszystkim rozrachunkiem z samym sobą, spojrzeniem w głąb swych myśli, swoistym odkrywaniem własnych

uczuć bez ukrywania czegokolwiek – zarówno przed światem, jak i samym sobą.

Kiedy rozważamy kwestie związane z emigracją, zaobserwować możemy, że i w tym przypadku opowieść o świecie i innych staje się jednocześnie tematem autobiograficznym. Emigracja stanowi bowiem swoistą przestrzeń między przeszłością a przyszłością, wypełnioną rozterkami i wątpliwościami, nostalgią – tęsknotą za krajem i bliskimi. Badania dotyczące stanu psychicznego osób przebywających na emigracji pokazują, jak trudnych emocji (stresu, osamotnienia) i problemów adaptacyjnych doświadczą wiele z nich.

Zdjęcia wykonane zostały podczas podróży do Irlandii (kwiecień 2016 roku), w miejscowości Bray (Ir. *Bré*) położonej około 20 km od Dublinia. Przedstawiają one polską rodzinę przebywającą na emigracji zarobkowej od ponad dekady, a mimo to wciąż niemogącą odnaleźć swojego miejsca w świecie.

Przemysław Wrochna





PSJ

Dostępny online
www.przegladsocjologiijakosciowej.org

„PSJ” stworzyliśmy, aby umożliwić swobodny przepływ informacji w społecznym świecie socjologii jakościowej. Adresujemy go do wszystkich socjologów, dla których paradygmat interpretacyjny i badania jakościowe stanowią podstawową perspektywę studiowania rzeczywistości społecznej.

Big Data i CAQDAS w badaniach jakościowych

Tom XIII ~ Numer 2
31 maja 2017

REDAKTORZY NUMERU TEMATYCZNEGO: Maciej Brosz,
Grzegorz Bryda oraz Piotr Siuda

REDAKTOR NACZELNY: Krzysztof T. Konecki
REDAKCJA: Waldemar Dymarczyk, Marek Gorzko,
Anna Kacperczyk, Łukasz T. Marciniak,
Jakub Niedbalski, Izabela Ślęzak

KOREKTA: Magdalena Chudzik-Duczmańska,
Magdalena Wojciechowska

EDYCJA: Magdalena Chudzik-Duczmańska
PROJEKT OKŁADKI: Anna Kacperczyk

ISSN: 1733-8069

