Anna Bączkowska
Uniwersytet Gdański
ORCID 0000-0002-0147-2718

doi.org/10.34765/sp.0420.a11

# TEMATYKA KORONAWIRUSA W POLSKIEJ PRASIE *ONLINE* – *BADANIE KORPUSOWE*

## Streszczenie

Celem badań przedstawionych w niniejszym artykule była analiza tekstów związanych z tematyką koronawirusa opublikowanych przez „Gazetę Wyborczą", tj. polską gazetę opiniotwórczą. Zastosowano metodologię językoznawstwa korpusowego, która pozwala przeprowadzić analizę ilościową dużych zbiorów danych (korpusów językowych). Metodologia ta jest używana w językoznawstwie, jednak – jak pokazuje niniejsza analiza – można ją z powodzeniem zastosować do badań ilościowych dyskursu prasowego i politycznego. Dane analizowane były automatycznie przez programy komputerowe opracowane dla języka polskiego (Korpusomat) oraz przez narzędzia dostępne w systemie Sketch Engine, które pozwalają analizować dane w języku angielskim. Przeprowadzona analiza wykazała, że w publikacjach prasowych w marcu i kwietniu koncentrowano się głównie na opisie wirusa i konsekwencjach zakażenia (hospitalizacja), natomiast artykuły publikowane między majem a lipcem zawierają informacje o potencjalnych szczepionkach i – w przeciwieństwie do faktów – podkreślają malejący trend zachorowań.

**Słowa kluczowe**: korpusy językowe, dyskurs prasowy, dyskurs polityczny, koronawirus, polska prasa opiniotwórcza.
**Kody JEL**: C80, I10

## Wprowadzenie

Analiza języka polityków według metodologii językoznawstwa korpusowego jest dość powszechną metodą badawczą używaną przez językoznawców w krajach anglojęzycznych, a w Polsce popularna jest głównie wśród anglistów (Charteris-Black 2004; Cap 2006; Koteyko 2007; Ädel 2010; Bączkowska 2017; Dalman 2017; Liu, Lei 2018; Chen, Yan, Hu 2019). W badaniach medioznawczych również z powodzeniem stosuje się metodologię językoznawstwa korpusowego (Schneider 1999; Bednarek 2006; Baker i in. 2008; Khosravinik 2010; Bednarek 2015; Bączkowska 2016; Moon 2016; Bączkowska 2019a; 2019b; Bączkowska, Khohlacheva 2019; Sofyaningrat, Suhardijanto, Yuwono 2019; Bączkowska 2020a; Bączkowska 2020b; Bączkowska, Gabdrakhmanova, Akhmetova 2020). Nie jest to jednak metodologia powszechnie stosowana wśród polskich badaczy medioznawców i politologów. Metodologia językoznawstwa korpusowego reprezentuje paradygmat badań ilościowych, mogłaby zatem stanowić cenne uzupełnienie dla przeważających w badaniach medioznawczych i politologicznych badań jakościowych.

## Cel badania

Celem ogólnym artykułu jest przybliżenie metod i narzędzi wykorzystywanych w językoznawstwie korpusowym badaczom reprezentującym środowisko politologów i medioznawców, w szczególności prasoznawców. Celem szczegółowym jest natomiast sprawdzenie czy sposób przedstawiania informacji dotyczących koronawirusa i wpływu epidemii na życie społeczne i polityczne w Polsce różnił się w tekstach publikowanych w prasie przed wyborami planowanymi na czerwiec 2020 r. oraz po tym terminie, tj. przed faktycznymi wyborami przeprowadzonymi 12.07.2020 r.

Motywacją podjęcia badania przez autorkę tego artykułu było nagłaśniane przez media prorządowe (telewizję i radio) w okresie przed niedoszłymi wyborami korespondencyjnymi w czerwcu (głównie od maja do końca czerwca), że koronawirus nie jest aż tak niebezpieczny, aby uniemożliwić przeprowadzenie wyborów prezydenckich. W okresie od marca do końca kwietnia głoszono jednak tezy dokładnie odwrotne, tj. że koronawirus stanowi ogromne zagrożenie dla życia i zdrowia obywateli. Z tego powodu zastosowano tzw. *lockdown*, uniemożliwiając obywatelom swobodne poruszanie się w przestrzeni

publicznej (praktycznie zamykając ich w domach), szkoły przeszły na system *e-learningowy*, a administracja i zakłady pracy – na *home office*, o ile było to możliwe. Wiele miejsc pracy zostało zlikwidowanych, co spowodowało zamrożenie niektórych gałęzi gospodarki na dwa-trzy miesiące, a w konsekwencji naraziło gospodarkę na wielomilionowe straty i ostatecznie powstanie dziury budżetowej, wzrost inflacji i bezrobocia oraz załamanie się niektórych dziedzin gospodarki. Analiza przedstawiona poniżej miała na celu zaobserwowanie, czy i jak ewoluował dyskurs prasowy w obliczu zmieniających się wytycznych dotyczących życia publicznego. Założono, że w tekstach prasowych będą przytaczane, komentowane i krytykowane toczące się w życiu politycznym rozmowy, w szczególności zmieniane oceny powagi sytuacji i skalę potencjalnego zagrożenia głoszone przez telewizję państwową w początkowym okresie pandemii (marzec, kwiecień), w okresie przed niedoszłymi wyborami korespondencyjnymi (maj, czerwiec) oraz po tym okresie, a przed faktycznymi wyborami 12.07.2020 r.

## Metodologia językoznawstwa korpusowego

Językoznawstwo korpusowe ma swoje początki w latach 60. XIX w. w Stanach Zjednoczonych. Pierwsze korpusy były zbiorami tekstów niewielkich rozmiarów, zgodnie z dzisiejszymi standardami liczyły około 1 miliona segmentów (w przybliżeniu – wyrazów). Następnie powstały pierwsze korpusy elektroniczne, nadal jednak stosunkowo niewielkich rozmiarów, choć znacznie przekraczające zakres ich poprzedników (np. British National Corpus), liczące już ok. 100 mln segmentów. Dane pozyskiwano głównie przez skanowanie książek, gazet i innych tekstów oraz transkrybowanie zapisów rozmów w formacie audio lub wideo. Ostatnia faza rozwoju korpusów to ogromne zbiory danych, liczące miliardy segmentów (np. enTenTen, plTenTen), które pozyskują dane z internetu za pomocą robotów indeksujących. W krajach anglojęzycznych i wśród badaczy anglosaskich językoznawstwo korpusowe rozwija się dynamicznie od dawna. Grono zwolenników badań korpusowych rozszerza się też od jakiegoś czasu w Polsce, głównie jednak wśród anglistów (od końca lat 90.), choć stopniowo zdobywa też pewne zainteresowanie w kręgach innych filologów.

Korpusy językowe to ustrukturyzowane zbiory językowe tekstów autentycznych, zapisanych w formie elektronicznej (Lewandowska-Tomaszczyk 2005). Należy podkreślić, że nie każdy zbiór tekstów to korpus. Za korpusy

w rozumieniu językoznawstwa korpusowego można uznać jedynie takie repozytoria, które są ustrukturyzowane, sformatowane i zapisane elektronicznie. Wynika to z tego, że aby dokonać analizy (ilościowej) danych tekstowych, trzeba je oznaczyć (tagować), kategoryzując je według części mowy (znakowanie morfoskładniowe) i zwykle też według funkcji, które pełnią w zdaniu (parsowanie). Dzięki temu możliwe jest dokonanie obliczeń ilościowych, takich jak np. liczba słów (lub segmentów) w tekstach, liczba rzeczowników, czasowników, przymiotników itd. oraz ich stosunek względem siebie (np. liczba przymiotników przypadająca na rzeczownik). Te i inne, bardziej zaawansowane metody badawcze tekstów, pozwalają określić m.in. cechy stylu tekstów, wartościowanie opisywanych zjawisk przez autorów (lub dany tytuł prasowy) czy orientację polityczną wyrażaną przez badane zbiory tekstów (np. konkretną gazetę). Duża liczba danych pozwala zwykle w przekonujący i obiektywny sposób zweryfikować hipotezy badawcze, bowiem badania nie opierają się na analizie pojedynczych przypadków (zdań, słów) i spekulacyjnym oraz subiektywnym charakterze wnioskowania badacza odwołującego się do własnej intuicji językowej, lecz na identyfikacji powtarzających się wzorców zachowań jednostek leksykalno-składniowych w dużych zbiorach tekstów. Analiza rekurencji pozwala zauważyć pewne trendy istniejące w całym korpusie danych poddanych analizie.

Ogólnie rzecz biorąc korpus można utworzyć na dwa sposoby: wgrywając własne pliki z tekstami lub zlecając systemowi wyszukiwanie tekstów na dany temat (tj. zawierających zestawy, tzw. krotki, konkretnych słów lub tekstów z określonego przez badacza portalu internetowego) za pomocą tzw. robotów indeksujących. Ta druga metoda bardzo przyśpiesza agregację tekstów, jednak ma swoje ograniczenia, np. w Sketch Engine (SK) korpusy teoretycznie nie mogą być większe niż 10 mln słów, w praktyce jednak osiąga się zwykle znacznie mniejsze wielkości korpusu (Bączkowska 2020b). Ponadto SK stosuje do wyszukiwania dokumentów w internecie wyszukiwarkę Bing, która stanowi niewielki procent rynku wyszukiwarek internetowych (więcej w sekcji *Metody i materiał*). Dane można zbierać z internetu również według różnych metod (popularne metody to *web crawling* i *web scraping*). W niniejszym badaniu, z przyczyn technicznych, zastosowano ten pierwszy sposób tworzenia korpusu, tj. korpus utworzono na podstawie plików tekstowych wgranych do systemu SK, a pozyskanych z internetu w wyniku ręcznego ich kopiowania i czyszczenia.

# Metody i materiał

## Metoda i źródło pozyskiwania danych

Dane zebrano z internetowej wersji Gazety Wyborczej (GW) po wpisaniu w wyszukiwarkę Google „koronawirus site:wyborcza.pl" oraz określeniu konkretnych dat artykułów za pomocą narzędzi dostępnych w wyszukiwarce Google[1]. Wszystkie dane zebrane były w ten sam dzień, z jednego adresu IP w trybie „incognito" i zdeponowane na twardym dysku PC. Wybrano Google z uwagi na to, że stanowi ona 81% rynku wyszukiwarek internetowych, w związku z tym uznano ją za wiarygodne źródło pozyskiwania danych. Dla porównania wyszukiwarka Bing stanowi 10% rynku, a Yahoo tylko 3,9% (www1).

Gazeta Wyborcza reprezentuje prasę społeczno-polityczną o orientacji liberalno-demokratycznej (www2). Gazetę wybrano jako źródło danych, bowiem jest to dziennik o zasięgu ogólnopolskim, o stosunkowo wysokiej sprzedaży oraz powszechnie uważany za opiniotwórczy. Nakład i sprzedaż gazety w 2019 r. określa się na ok. 100 tys. egzemplarzy płatnych (www2). Dla porównania inny dziennik ogólnopolski – Rzeczpospolita – ma nakład i sprzedaż na poziomie ok. 40 tys. egzemplarzy (www3), Gazeta Polska Codziennie z kolei przy średnim nakładzie ok. 50 tys. egzemplarzy może pochwalić się sprzedażą ok. 13 tys. egzemplarzy (www4), a ogólnopolski dziennik Gazeta Polska jedynie ok. 13 tys. egzemplarzy (www5). Teksty poddane analizie pochodzą zarówno z głównego wydania GW, jak i z jej wersji regionalnych. W skład korpusu wchodzą tylko artykuły udostępnione bezpłatnie *online* w wersji pełnotekstowej. Artykuły włączone do korpusu wybierano według kolejności wyświetlenia się w wyszukiwarce.

## Kompilacja korpusu i struktura danych

Zanim skompilowano korpus artykuły pobrane z internetu wyczyszczono z metadanych, np. nazwiska autorów, informacje o wolnym dostępie do

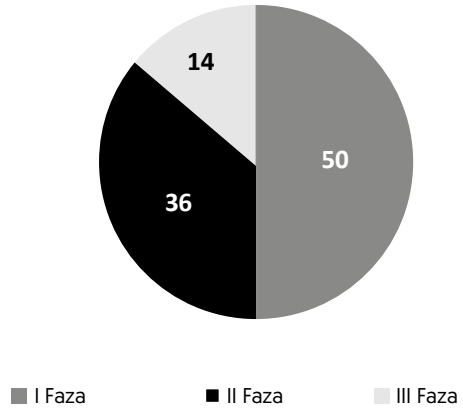
[1] Trzeba przyznać, że jest to dość czasochłonny sposób pozyskiwania danych, jednak jednocześnie najprostszy, niewymagający znajomości języka programowania ani działania robota automatycznie zbierającego dane z internetu. Ponieważ artykuł ma na celu przybliżenie analizy opartej na metodologii zapożyczonej z językoznawstwa korpusowego badaczom medioznawcom i politologom, bardziej zaawansowane metody ekstrakcji danych (np. *webscraping* czy *webcrawling*) zastąpiono ręcznym wyszukiwaniem tekstów.

tekstu, *hiperlinki* do innych artykułów (np. poprzedzonych frazą „Przeczytaj także") itp. Teksty bez metadanych zapisano w kodzie UTF-8 i załadowano do systemu Sketch Engine (SK, sketchengine.eu) oraz do Korpusomatu (KM, korpusomat.pl), a następnie skompilowano, tj. poddano oznakowaniu morfosyntaktycznemu (tagowaniu) oraz parsowaniu. Te dwie czynności odbywają się automatycznie po załadowaniu danych do systemu SK lub KM. W SK istnieje też możliwość wyboru narzędzi do kompilacji korpusu. System SKjest dostępny zasadniczo w wersji komercyjnej, jednak można skorzystać z darmowej aplikacji KM, która ma najważniejsze narzędzia wykorzystywane w językoznawstwie korpusowym (analiza frekwencji, słów kluczowych, kolokacji itp.). Przy wyborze artykułów minimalną długość tekstu określono na 200 segmentów (słów). Teksty krótsze od 200 słów wydawały się niemiarodajne i zawierające zbyt mało danych dla obliczeń statystycznych. Cały korpus liczy niecałe 37 tys. segmentów (tj. ok. 30 tys. słów). W tym miejscu trzeba zaznaczyć, że im większy korpus, tym bardziej wiarygodne wyniki badań, trudno jest jednak określić, ile słów powinien liczyć korpus, aby był wystarczająco duży do wyciągania obiektywnych wniosków. Nie istnieją wytyczne dotyczące wymagań w kwestii wielkości korpusu, bowiem ta wielkość adekwatna do weryfikacji danych zależy od wielu czynników. Utworzony na potrzeby niniejszego badania korpus nie jest rzecz jasna wyczerpujący, stanowi pewien wycinek wszystkich artykułów opublikowanych przez GW na temat koronawirusa w określonych ramach czasowych (i tylko tych dostępnych bezpłatnie), zatem badanie można uznać za wstępne (pilotażowe), a wyniki analizy traktować niekategorycznie, jako informacje dotyczące pewnych tendencji obserwowalnych w przypadku tego konkretnego korpusu tekstów.

W skład korpusu, nazwanego korpusem „Korona" (KK), wchodzą artykuły z kilku miesięcy: – z marca, kwietnia, maja, czerwca i lipca 2020 r. Z każdego miesiąca wybrano po 10 tekstów, które pojawiły się w wyszukiwarce jako pierwsze (dostępnych bezpłatnie, dłuższych niż 200 słów po oczyszczeniu danych). Utworzono również podkorpusy, które reprezentowały artykuły z każdego miesiąca oddzielnie. Struktura korpusu jest następująca: podkorpus marcowy stanowi 11,9% całego korpusu, podkorpus kwietniowy to 6,5%, majowy 7,7% danych, czerwcowy 5,7%, a podkorpus lipcowy tworzy 5% całego KK. Ponieważ celem analizy było sprawdzenie tematów tekstów do wyborów prezydenckich (12.07), ostatni artykuł dodany do korpusu opublikowano 11.07. W sumie KK składa się zatem z 50 tekstów, które podzielono na trzy części, odpowiadające etapom rozwoju sytuacji.

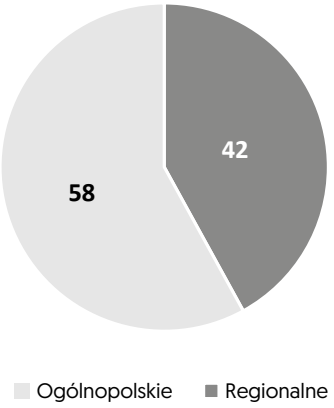
Początkowy etap obejmuje dyskurs prasowy dotyczący koronawirusa z marca i kwietnia (faza I) czyli z okresu poprzedzającego niedoszłe wybory korespondencyjne i obfitującego w dramatyczne (a wręcz paniczne) komentarze dotyczące błyskawicznie wzrastającej liczby zakażeń i nieuchronnie nadchodzącej sytuacji epidemicznej w kraju. Następny okres obejmuje maj i czerwiec (faza II), tj. czas, kiedy mimo lawinowo wzrastającej liczby zakażeń i zgonów, politycy prorządowi zapewniali, że koronawirus jest w odwrocie, co miało uspokoić obywateli i przekonać ich, że uczestniczenie w czerwcowych wyborach jest bezpieczne. Ostania część dotyczy okresu przed wyborami lipcowymi, na którą składają się teksty opublikowane w dniach 1–11.07 (faza III). 42% artykułów w Korpusie Korona stanowią teksty opublikowane w ogólnopolskim wydaniu GW, 58% to teksty pochodzące z wydań regionalnych, korpus zatem jest pod tym względem stosunkowo dobrze zrównoważony. Z krakowskiego wydania pochodzi 9 tekstów, z poznańskiego – 1, z warszawskiego – 6, z radomskiego – 2, z olsztyńskiego –1, z toruńskiego –1, z rzeszowskiego – 4, z katowickiego – 4 i z trójmiejskiego – 1. Strukturę korpusu przedstawiono na wykresie 1a i 1b.

**Wykres 1a.** Procent liczby segmentów w korpusie ze względu na datę publikacji tekstu




Źródło: opracowanie własne.

**Wykres 1b.** Procent liczby tekstów pochodzących z wydania ogólnopolskiego i z wydań regionalnych




Źródło: opracowanie własne.

## Metoda analizy danych

Do analizy danych wykorzystano narzędzia dostępne w systemie Sketch Engine (Kilgarriff i in. 2004) oraz powstałą kilka lat temu darmową aplikację *webową* dla danych w języku polskim Korpusomat (Kieraś, Kobyliński, Ogrodniczuk 2018). Większość analiz parametrów można wykonać zarówno w SK, jak i w KM, a wybór ten w niniejszej analizie podyktowany był preferencjami autorki i przyczynami technicznymi (KM powstał stosunkowo niedawno, a ściąganie danych w czytelnym formacie wymaga instalacji wyszukiwarki Poliqarp, która jest mało stabilna). Przedstawiona poniżej analiza zawiera omówienie wyników badania dotyczącego kilku parametrów: frekwencji, kolokacji, n-gramów, dyspersji i słów kluczowych.

## Technikalia

Niniejsza sekcja przeznaczona jest dla osób dociekliwych, chcących poznać szczegóły techniczne badań korpusowych. Do analizy przedstawionej poniżej wykorzystano różne aplikacje i narzędzia informatyczne: dostępne w systemie Sketch Engine (Sketch Word, listy frekwencyjne, ekstrakcja słów kluczowych

i terminów) oraz na stronie Korpusomatu (C-value, ekstrakcja słów kluczowych, ich dystrybucja oraz listy frekwencyjne). Trzeba pamiętać, że wyniki uzyskane z SK i KK dla tych samych parametrów mogą się różnić, a to za sprawą użytych odmiennych tagerów morfosyntaktycznych, innych korpusów referencyjnych i wzorów matematycznych do identyfikacji i obliczania wartości słów kluczowych, a także innych wzorów do określania współzależności wyrazów w kolokacji. W SK stosuje się ten sam analizator morfologiczny używany do oznaczania części mowy, który stosowany jest również w NKJP. Jest to analizator morfologiczny Morfeusz, który „etykietuje" wyrazy bez uwzględnienia kontekstu, natomiast do dezambiguacji (ujednoznacznienia) morfosyntaktycznej zaimplementowano tager Pantera. W KK do niniejszej analizy użyto Morfeusza 2, który jest nowszą wersją Morfeusza, i który do dezambiguacji stosuje tager Concraft. Morfeusz jest już nieco przestarzały i nieaktualizowany, przeciwnie do Morfeusza 2. Bardziej precyzyjne wyniki osiąga się więc w przypadku korzystania z nowszej wersji. Błąd w anotowaniu dla NKJP, który korzysta z tagera morfosyntaktycznego Pantera, wynosi ok. 8%, osiągając tym samym gorsze wyniki niż tager Concraft (Waszczuk 2012; Kobyliński, Kieraś 2016). Warto wspomnieć, że w Korpusomacie istnieje możliwość wyboru analizatora morfologicznego, dzięki czemu możliwe jest dokonywanie porównań z NKJP, trzeba jednak pamiętać o zmianie ustawień sposobu anotacji morfologicznej.

Druga kwestia dotyczy słów kluczowych, które w KK obliczane są na podstawie NKJP traktowanego jako korpus referencyjny, natomiast w SK na podstawie korpusu plTenTen12. Ten ostatni jest korpusem tzw. trzeciej generacji, tworzony na podstawie danych z internetu, więc różnice między nimi mogą wpływać na wyniki identyfikacji słów kluczowych z uwagi na ich odmienną strukturę oraz inny sposób i różne źródła pozyskiwania danych. NKJP jest korpusem zrównoważonym, natomiast plTenTen12 nieporównywalnie większym (plTenTen12 liczy aktualnie około 10 mld segmentów). Wybór korpusu referencyjnego ma zasadnicze znaczenie przy określaniu słów kluczowych, zatem w zależności od tego, który wybierzemy, możliwe jest uzyskanie nieco odmiennych słów kluczowych.

Warto wspomnieć, że sposób obliczania wartości dla słów kluczowych też jest istotny. Do tego celu różne programy używają odmiennych wzorów matematycznych, np. test zgodności chi-kwadrat, logarytm wskaźnika prawdopodobieństwa, procent różnicy (*%Diff*), BIC, czy metodę prostych obliczeń matematycznych (*simple math method*). Każdy z tych wskaźników opiera się na innych kalkulacjach, dlatego trzeba to uwzględnić przy wyborze testu

statystycznego; dobrze jest też zastosować więcej niż jeden wzór, bowiem każdy z nich może podkreślać nieco inne cechy badanych kolokacji. Chi-kwadrat, popularny dwadzieścia lat temu i wcześniej, koncentruje się na obliczaniu poziomu istotności statystycznej różnicy frekwencji. Wraz z pojawieniem się aplikacji *WordSmith* statystyka ta została wyparta przez logarytm prawdopodobieństwa. Testy mierzące kluczowość na podstawie poziomu istotności mają swoje uzasadnienie w przypadku małych korpusów. Jeśli istnieje potrzeba uwzględnienia różnic w wielkości korpusów, wówczas można obliczyć tzw. parametr BIC (Wilson 2013). Zamiast sprawdzania istotności statystycznej można weryfikować efekt wielkości używając indeksu *%Diff* (Gabrielatos, Marchi 2011). Inny parametr statystyczny – metoda prostych obliczeń matematycznych – uwzględnia różnice w kluczowości wynikające z przedziału wyrazów o bardzo wysokiej frekwencji oraz wyrazów rzadkich (Kilgarriff 2009). Warto dodać, że w ostatnich latach w językoznawstwie korpusowym można zauważyć zwrot metod statystycznych dla obliczania kluczowości z tych opartych na obliczeniach poziomu istotności ku metodom badającym efekt wielkości oraz stosującym kombinację różnych wzorów matematycznych. Metoda prostych obliczeń matematycznych podawana jest przez SK, algorytm wskaźnika prawdopodobieństwa – przez KM.

Kolejna kwestia to kryteria identyfikacji kolokacji. Wyrazy, które mają tendencję do współwystępowania, takie jak wspomniany wyżej *rzęsisty deszcz*, muszą wykazywać się jakąś konwergencją semantyczną, czyli wewnętrznym zespoleniem. Kryterium semantyczne w definiowaniu kolokacji, typowe dla paradygmatu analitycznego we frazeologii (szczególnie w nurcie winogradowskim), nie jest jedynym możliwym rozwiązaniem. Inne przypadki, takie jak syntaktycznie niekompletne związki *no tak, ale* czy *uważam, że* nie są kolokacjami (*sensu stricto*) nawet jeśli występują często ze sobą w danym korpusie i można je traktować w analizie syntaktycznej jako całości, jak proponował Lewicki (1976), na podstawie kryterium frekwencji. Stopień łączliwości elementów konstytutywnych kolokacji oraz prawdopodobieństwo ich współwystępowania można obliczyć statystycznie i określić liczbowo na podstawie wzorów matematycznych. KM używa wielu różnych wskaźników prawdopodobieństwa, w tym dość powszechnie używanego „Dice", który jest również zaimplementowany do narzędzia Sketch Word w SK. Sketch Engine podaje również wartości stopnia łączliwości kolokatów, co może stanowić cenne uzupełnienie danych. Trzeba pamiętać, że różne wzory matematyczne są używane do mierzenia stopnia spójności czy

prawdopodobieństwa współwystępowania słów w zależności od tego, czy związki wyrazowe są stałe czy luźne oraz jak są częste.

## Wyniki analizy danych

### Frekwencja

Okurencje najczęściej pojawiających się wyrazów w danym podkorpusie można podać w postaci frekwencji liczby wystąpień. Średnia frekwencja zredukowana (ŚFZ), uwzględniająca rozkład słów w całym korpusie, pozwala stosunkowo precyzyjnie (bardziej precyzyjnie niż frekwencja absolutna) wyciągnąć trafne wnioski dotyczące faktycznej okurencji analizowanych wyrazów i ich znaczenia dla tekstów (tabela 1).

**Tabela 1.** Lista frekwencyjna czasowników w podkorpusach mierzonych wartością średniej frekwencji zredukowanej (ŚFZ)

| Marzec | | Kwiecień | | Maj | | Czerwiec | | Lipiec | |
|---|---|---|---|---|---|---|---|---|---|
| lemat | ŚFZ | lemat | ŚFZ | lemat | ŚFZ | lemat | ŚFZ | lemat | ŚFZ |
| być | 150,9 | Być | 78,2 | być | 72,6 | być | 64,3 | być | 59,4 |
| mieć | 67,4 | Mieć | 20,3 | mieć | 34,5 | mieć | 21,5 | zakazić | 16,0 |
| móc | 49,6 | Móc | 15,8 | móc | 16,0 | zakazić | 17,2 | mieć | 15,9 |
| zakazić | 20,1 | zakazić | 11,0 | zakazić | 15,9 | móc | 15,3 | potwierdzić | 10,9 |
| potwierdzić | 13,2 | mówić | 9,8 | zemrzeć | 5,9 | wykryć | 9,1 | zostać | 9,3 |
| zostać | 11,1 | pracować | 7,2 | wiedzieć | 5,8 | zemrzeć | 8,3 | podawać | 8,5 |
| mówić | 10,4 | przyjmować | 6,5 | musieć | 5,6 | potwierdzić | 5,9 | zemrzeć | 7,8 |
| musieć | 8,5 | potwierdzić | 5,9 | zostać | 5,1 | czekać | 5,0 | mówić | 5,9 |
| pojawić | 8,3 | będzie | 4,6 | wynikać | 4,5 | zostać | 4,9 | przebywać | 5,4 |
| przenosić | 7,1 | wykryć | 4,5 | mówić | 4,4 | wynosić | 3,8 | informować | 4,8 |
| objąć | 6,4 | prowadzić | 4,4 | zrobić | 4,3 | odbyć | 3,7 | dotyczyć | 4,8 |
| odwołać | 6,1 | wstrzymać | 4,1 | przeprowadzić | 4,0 | trwać | 3,3 | wynikać | 4,4 |
| zachorować | 6,0 | zostać | 3,8 | przebadać | 3,7 | będzie | 3,2 | przeprowadzić | 3,9 |
| stosować | 5,6 | informować | 3,6 | będzie | 3,5 | chorować | 3,1 | badać | 3,8 |
| chodzić | 5,6 | wynikać | 3,5 | pochodzić | 3,3 | przygotować | 3,0 | wykryć | 3,8 |
| działać | 5,5 | działać | 3,5 | stwierdzić | 3,3 | podawać | 3,0 | móc | 3,8 |
| unikać | 5,3 | zamknąć | 3,4 | chodzić | 3,3 | przebywać | 3,0 | odnotować | 3,7 |
| pamiętać | 5,3 | przygotowywać | 3,4 | zamknąć | 3,2 | przypominać | 2,8 | pobrać | 3,7 |

Źródło: opracowanie własne.

W tabeli 1 przedstawiono wartości ŚFZ dla okurencji czasowników. Podobne analizy przeprowadzono dla rzeczowników i przymiotników. Dane te pozwalają zauważyć, że w I fazie dyskursu prasowego na temat koronawirusa są czasowniki *być*, *mieć* i *móc*, które w II fazie (w czerwcu) zmieniają się na *być*, *mieć* i *zakazić*, a w III fazie – na *być*, *zakazić* i *mieć*. Potencjalność wyrażoną przez *móc* występującą z dużą częstością w marcu i kwietniu zastąpiono zatem w II i III fazie raportem o konkretnych przypadkach zachorowań. Ponadto w maju, czerwcu i lipcu pojawia się czasownik *zemrzeć*, który sygnalizuje pogorszenie się sytuacji epidemicznej.

Jeśli chodzi o przymiotniki to na uwagę zasługuje częste pojawianie się w tekstach czerwcowych słowa *schyłkowy* (w odniesieniu do rozwoju epidemii), a także przymiotników*wyborczy*, *zagraniczny* i *międzynarodowy*. Punkt ciężkości wydaje się przesuwać w czerwcu na kończącą się według polityków epidemię koronawirusa w Polsce oraz na arenę międzynarodową w nawiązaniu do przeprowadzonych z sukcesem wyborów korespondencyjnych w innych krajach, co nagłaśniała państwowa Telewizja Polska.

Wśród rzeczowników w marcu najczęściej pojawiają się wyrazy oznajmiające zaistnienie problemu, np. *osoba*, *wirus*, *zakażenie*; w kwietniu na pierwsze miejsce wysuwają się określenia związane z hospitalizacją (*szpital*, *oddział*, *pacjent*); maj to miesiąc kontynuacji problemu hospitalizacji (*szpital*, *pacjent*) osób *zakażonych* oraz dyskusji o zbyt małej liczbie przeprowadzanych *testów* na obecność koronawirusa, a także zamiany kwalifikacji problemu z *epidemii* na *pandemię*. Interesujące jest również to, że w czerwcu wysoka frekwencja dotyczy wyrazów niezwiązanych bezpośrednio z koronawirusem (*powiat*, *kraj*, *mieszkaniec*), co ma związek z przygotowywanymi wyborami, a także wzmożoną dyskusją na temat badań nad *szczepionkami* lub ich zbytecznością (ponieważ, jak twierdzą niektórzy internauci, koronawirus nie istnieje i całe zamieszanie wokół niego to spisek), co jest informacją sugerującą szybką poprawę sytuacji lub w ogóle brak problemu. Jest to informacja głoszona (i komentowana przez GW) tuż przed planowanymi wyborami czerwcowymi.

Z powyższej analizy wynika, że wyrazy o wysokiej frekwencji zmieniały się na przestrzeni kilku miesięcy i ilustrowały dyskusje polityków sugerujące „trend pozytywny" w II fazie, zwłaszcza w tekstach z czerwca, ignorujących faktyczne zagrożenie wypływające z szybkiego rozprzestrzeniania się wirusa.

## N-gramowe jednostki charakterystyczne

Z korpusu można pozyskać w aplikacji Korpusomat ciągi słów tworzące semantycznie wspólną całość (w większym lub mniejszym stopniu), składającą się z dwóch lub więcej wyrazów (tzw. n-gramy), które stanowią słownictwo charakterystyczne dla analizowanego korpusu. Ekstrakcja ta opiera się na obliczeniu statystycznym (obliczaniu tzw. wartości C), co pozwala na identyfikację fraz mających status jednostek wielowyrazowych (lub dwuwyrazowych), w tym terminów zagnieżdżonych w innej, dłuższej jednostce wielowyrazowej. Poniżej podano słownictwo charakterystyczne dla poszczególnych podkorpusów, w nawiasie podano ich wartość C. Za dolną granicę wartości C przyjęto liczbę 4.

Marzec: *osoba chora* (6), *ministerstwo zdrowia* (6), *kraj objęty* (5), *izba przyjęć* (5), *szpital zakaźny* (5), *inspektor sanitarny* (5), *pierwszy przypadek zakażenia* (4), *światowa organizacja zdrowia* (4), *ciepła woda* (4), *północne Włochy* (4).

Kwiecień: *chirurgia ogólna* (10), *oddział wewnętrzny* (8), *przypadek zakażenia* (7), *dyrektor szpitala* (6), *cały oddział* (6).

W korpusie marcowym są wzmianki o problemie koronawirusa poza Polską (*północne Włochy*, *kraj objęty*) oraz o pojedynczych przypadkach zachorowań (*osoba chora*, *pierwszy przypadek zakażenia*). Wyraz *kraj* dotyczy w każdym przypadku zagranicy, a konkretnie osób, które przyjeżdżają do naszego kraju, np. *osoby przyjeżdżające z krajów objętych (…) monitoringiem epidemiologicznym*, *pracownicy i studenci (…), którzy wrócili z krajów objętych zagrożeniem epidemiologicznym* itp. W porównaniu z marcem teksty z kwietnia wyraźnie koncentrują się na hospitalizacji, co wskazuje na rozwój epidemii.

Maj: *szpital wolski* (6), *Stany Zjednoczone* (6), *ognisko koronawirusa* (6), *pandemia koronawirusa* (6), *większa liczba* (5), *test genetyczny* (4).

Czerwiec: *ministerstwo zdrowia* (8), *nowe zakażenie* (7), *nowy przypadek* (5), *przypadek zakażenia* (5), *ognisko koronawirusa* (5), *lokal wyborczy* (4), *faza schyłkowa* (4).

Lipiec: *Ministerstwo Zdrowia* (17), *nowy przypadek* (14), *nowe zakażenie* (10), *przypadek zakażeń* (8), *powiat jarosławski* (7), *zakład mięsny* (7), *mieszkaniec województwa podkarpackiego* (5), *wynik dodatni* (4), *osoba zakażona* (4), *powiatowy sanepid* (4), *województwo śląskie* (4).

W maju sytuacja rozwija się, mowa jest już o *pandemii*, o *większej liczbie* zachorowań, nowych *ogniskach zachorowań* i rozwoju epidemii w *Stanach Zjednoczonych*. W czerwcu z jednej strony czytamy o *nowych przypadkach* i *zakażeniach*, z drugiej jednak przebija się wątek związany z przygotowaniami do wyborów (*lokal wyborczy*), które możliwe są dlatego, jak zapewniają

politycy, że następuje *faza schyłkowa* epidemii. Teksty prasowe w korpusie majowym wydają się zatem być podobne w treści do kwietniowych, a zmianę retoryki można zauważyć w czerwcu. W korpusie lipcowym, tuż po niedoszłych wyborach korespondencyjnych, a przed wyborami przeprowadzonymi 12.07, ponownie mowa jest o wzroście zachorowań i rozwoju epidemii (co jednak nie przeszkodziło w przeprowadzeniu wyborów).

## Kolokacje

Kolokacje to wyrazy mające tendencję do współwystępowania. Typowymi przykładami kolokacji przymiotnika z rzeczownikiem są np. *zjełczałe masło* czy *rzęsisty deszcz*. Przymiotnik *zjełczałe* występuje bardzo rzadko z innymi rzeczownikami (może jeszcze ewentualnie wystąpić np. z rzeczownikiem *szminka*), a *rzęsisty* spotyka się praktycznie tylko w kontekście z rzeczownikiem *deszcz*. Przymiotniki te zatem niejako automatycznie przywołują z pamięci rzeczowniki, które zwykle po nich następują i na tym polega ich tendencja do współwystępowania. Kolokacje to jednak nie tylko współwystępowanie przymiotników z rzeczownikami (z przymiotnikami w prepozycji lub postpozycji w stosunku do rzeczowników), ale też innych części mowy, np. rzeczowników z czasownikami, rzeczowników z przyimkami itd.

Poniżej podano kilka rodzajów kolokacji: przymiotników z rzeczownikami w pre- lub postpozycji (Prz-Rz; Rz-Prz) oraz rzeczowników z czasownikami (Rz-Cz, Cz-Rz), a także czasownika *być* z przymiotnikiem w postpozycji (B-Prz). Dane te pozyskano z aplikacji dostępnej na stronie KM.

Marzec:

Prz-Rz: *niebezpieczny mikroorganizm*, *północne Włochy*, *światowa organizacja*,

Rz-Prz: *spektakl teatralny*, *presja selekcyjna*, *obróbka termiczna*, *ciepła ręka*.

Kwiecień:

Prz-Rz: *intensywna terapia*, *niebezpieczny mikroorganizm*, *otwarta interna*, *pełna obsada*, *planowe przyjęcie*, *prawdopodobne źródło*, *wolne łóżko*; *regularna dezynfekcja*,

Rz-Prz: *chirurgia ogólna*, *oddział wewnętrzny*, *system ostrodyżurowy*, *obróbka chemiczna*, *kondycja psychiczna*, *kwarantanna własna*, *onkologia pulmonologiczna*, *niewydolność oddechowa*; *test dodatni*, *choroba zakaźna*.

Artykuły prasowe publikowane w marcu odwołują się do sytuacji epidemicznej w północnych Włoszech i do komunikatów Światowej Organizacji Zdrowia.

Informują również o tym, czym jest i jak powstał koronawirus (*niebezpieczny mikroorganizm*; *mógł się namnażać przy stosunkowo słabszej presji selekcyjnej*), co należy robić, aby się przed nim uchronić (*myć ręce w ciepłej wodzie*; *obróbka termiczna, np. gotowanie zabija wirusy*) oraz że wiele instytucji publicznych będzie zamkniętych (*odwołane spektakle teatralne*). W kwietniu głównym tematem poruszanym przez GW jest hospitalizacja zakażonych pacjentów.

Wątek hospitalizacji zaczęty w kwietniu jest kontynuowany w II fazie (skonkretyzowany geograficznie i instytucjonalnie). W szczególności w czerwcu widać zainteresowanie wyborami prezydenckimi w tle dyskusji o koronawirusie oraz *obiecującą szczepionkę* opracowywaną w *zagranicznych laboratoriach*, a także zarysowującą się *fazę schyłkową* epidemii.

Maj:

Prz-Rz: *duża liczba*, *negatywny wynik*, *cicha epidemia*, *bakteryjne jelito*, *dwukrotna dezynfekcja*;

Rz-Prz: *szpital wolski*, *test genetyczny*, *flora bakteryjna*, *ofiara śmiertelna*, *badanie przesiewowe*, *układ odpornościowy*, *szpital specjalistyczny*.

Czerwiec:

Prz-Rz: *cały świat*, *Wielka Brytania*, *obiecująca szczepionka*, *niedzielny wybory*, *szwedzki król*, *szybki dostęp*, *zaawansowany praca*, *zagraniczne laboratorium*, *światowa gospodarka*, *nowy zakażenie*, *nowy przypadek*, *poszczególny rząd*;

Rz-Prz: *województwo małopolskie*, *faza schyłkowa*, *ciąg ostatni*, *lokal wyborczy*, *wybory prezydenckie*, *powiat tarnowski*.

Zaskakujące jest to, że po *fazie schyłkowej*, którą odnotowano pod koniec czerwca, nagle w pierwszej połowie lipca epidemia ogarnęła *całą Polskę* (wymienia się szereg województw i powiatów), pojawiają się *nowe przypadki i zakażenia* oraz *kolejne ofiary śmiertelne*, Sanepid i stacje epidemiologiczne pracują z pełna mocą, działa *całodobowa infolinia* dla osób z objawami koronawirusa, na którą można dzwonić ze *wszelkimi wątpliwościami* itd. Cała opisywana w prasie sytuacja pogrąża w kryzysie *branżę turystyczną*.

Lipiec:

Prz-Rz: *nowy przypadek*, *nowe zakażenie*, *kolejna ofiara*, *powiatowa stacja*, *wojewódzka stacja*, *powiatowy sanepid*, *cała Polska*, *wszelka wątpliwość*, *całodobowa infolinia*, *60-procentowy alkohol*;

Rz-Prz: *zakład mięsny*, *ofiara śmiertelna*, *wynik dodatki*, *województwo podkarpackie*, *urządzenie pomiarowe*, *powiat jarosławski*, *województwo śląskie*, *powiat przemyski/poznański/przeworski*, *branża turystyczna*, *służba sanitarna*, *opieka medyczna*.

Kolokacje kilku wybranych słów pozyskano korzystając z aplikacji Word Sketch (system SK). Poniżej przykładowy wynik takiego poszukiwania dla wyrazu *koronawirus* w podkorpusie marcowym (rysunek 1).

**Rysunek 1.** Kolokacje rzeczownika *koronawirus* w podkorpusie marcowym

| a_modifier |
|---|
| stwierdzić |
| pierwszy |
| nowy |

| is_subj |
|---|
| przenosić |
| wyewoluować |
| odpuścić |
| znosić |
| wywoływać |
| przebiegać |
| rozprzestrzeniać |
| zmieniać |

| post_verb |
|---|
| przenosić |
| wyewoluować |
| odpuścić |
| znosić |
| wywoływać |
| zmieniać |
| rozprzestrzeniać |

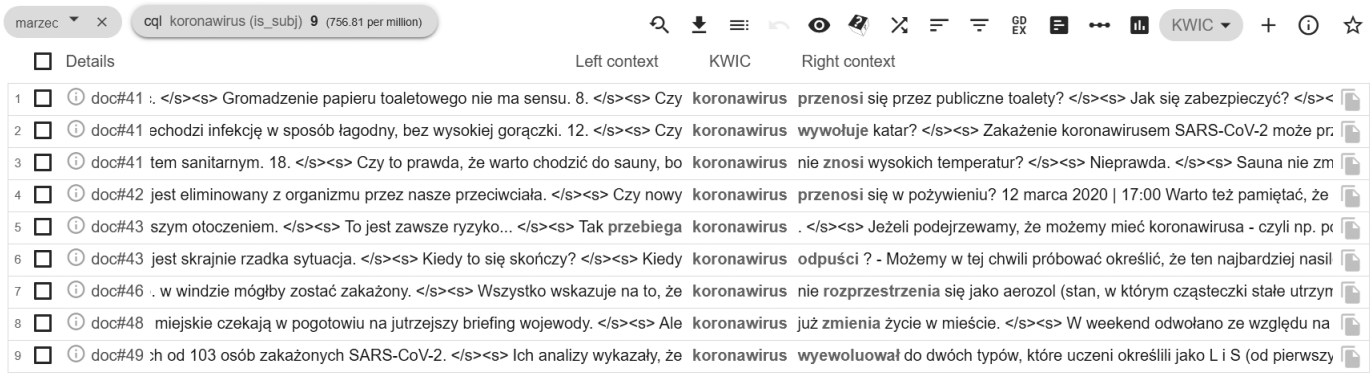
| być_adj |
|---|
| zaraźliwy |
| niebezpieczny |

| prec_verb |
|---|
| zachować |
| występować |
| przebiegać |

| post_w |
|---|
| powietrze |

Źródło: opracowanie własne.

**Rysunek 2.** Konkordancje dla wyrazu *koronawirus* w podkorpusie marcowym




Źródło: opracowanie własne.

Z danych na rysunku 1 wynika, że przymiotnik w prepozycji do wyrazu *koronawirus* to *pierwszy* i *nowy*. Jest to podkorpus z marca, zatem określenia te nie dziwią – pojawiały się wówczas pierwsze zachorowania (u tzw. pacjenta „zero" wykryto koronawirusa 4.03), a koronawirus to nowy typ wirusa. Czasowniki współwystępujące z lematem *koronawirus* to: *przenosić*, *wyewoluować*, *rozprzestrzeniać*, *występować* itd. (rubryka druga [is_subj], np. *koronawirus jest przenoszony*); rubryka trzecia [post_verb] to np. *wywoływać koronawirusa*; rubryka piąta [prec_verb] to np. *koronawirus występuje*. Przymiotniki współwystępujące z lematem *koronawirus* to *zaraźliwy* i *niebezpieczny* ([być_adj]), a z przyimkiem *w* to np. *koronawirus w powietrzu*. Przykłady z korpusu w szerszym kontekście ilustrują poniższe konkordancje (krótkie konteksty automatycznie wyekstrahowane i „przycięte" przez konkordancer, który jest wbudowany w SK i KM). Korzystając z aplikacji SK wygenerowano konkordancje dla każdego podkorpusu; poniżej przytoczono dla przykładu jeden zestaw konkordancji (z korpusu marcowego).

Z powyższej analizy wynika, że teksty opublikowane w marcu koncentrują się na opisie cech charakteryzujących koronawirusa, np. w jaki sposób przenosi się na innych ludzi, w jakich warunkach się (nie) rozwija, od czego ewoluował, jakie wywołuje potencjalne skutki itd. W marcu koronawirus określany był mianem *niebezpieczny* i *zaraźliwy*. Warto wspomnieć przy okazji, że w maju, kiedy zarażeń i zgonów było kilkakrotnie więcej niż w marcu, zmienił swoją kwalifikację na *niegroźny*. W podkorpusie kwietniowym występuje dodatkowo wyraz *pojawia się*, zatem teoretyczny opis cech wirusa konkretyzuje się. Analiza danych w pierwszej fazie rozwoju dyskursu na temat koronawirusa (marzec-kwiecień) wykazuje, że prasa zajmowała się przede wszystkim charakterystyką wirusa oraz określaniem miejsc, w których się pojawił. Na początku II fazy (maj-czerwiec) z kolei czytamy, że koronawirus *potężnie uderzył w polską gospodarkę* (21.05). Z charakterystyki wirusa typowego dla I fazy punkt ciężkości przesunął się więc na skutki jego działania, które przedstawione są jako trudne do opanowania i wyrządzające ogromne szkody. Jednakże już trzy tygodnie później, 5.05 (i jednocześnie na trzy tygodnie przed planowanymi wyborami), GW pisała optymistycznie o tym, że *wciąż są miejsca, gdzie nie dotarł koronawirus*. 10.05 GW krytykowała upowszechniane stwierdzenia prorządowych polityków o małej szkodliwości koronawirusa, które propagowano w telewizji państwowej w związku z wyborami planowanymi na 28.06: *Ale argumentowanie, że pandemii nie ma, a koronawirus nie jest groźniejszy niż ciężkie przeziębienie, to absurd*. W lipcu koronawirus ponownie *pojawił*

*się* w wielu powiatach w Polsce, spowodował zgony wielu pacjentów i pogrąża branżę turystyczną w głębokim kryzysie.

Najważniejsze kolokacje dla lematu *koronawirus* w całej bazie KK ilustruje tabela 2. Dane te mierzone statystyką *logDice* przy minimalnym progu równym wartości 10 i przy założeniu, że minimalna okurencja danego wyrazu jest równa 5. Warto zauważyć, że kolokaty o najwyższej wartości to przymiotniki zagrożenia, kodujące negatywne emocje (*zaraźliwy*, *niebezpieczny*, *groźny*).

**Tabela 2.** Kolokaty wyrazu *koronawirus* według wartości statystycznej logDice

| Kolokaty | LogDice |
|---|---|
| zaraźliwy | 13,0 |
| niebezpieczny | 13,0 |
| groźny | 12,41 |
| przenosić | 12,41 |
| powietrze (~ w powietrzu) | 12,19 |
| czynny | 11,99 |
| weselny | 11,99 |
| stwierdzony | 11,83 |
| Podkarpacie | 11,83 |
| pandemia | 11,09 |
| Sanepid | 10,57 |
| Śląsk | 10,54 |
| żywność | 10,35 |

Źródło: opracowanie własne.

## Kluczowość i dyspersja

Kluczowość to zbiór wyrazów, które występują w analizowanym korpusie z częstością wyższą niż to wynika z normy. Za normę uznaje się dane pochodzące z innego korpusu, który przyjmuje się jako punkt odniesienia dla analizy (jest to tzw. korpus referencyjny); zwykle jest to korpus języka ogólnego. Słowa, które są kluczowe, mają okurencję niewspółmiernie wyższą od oczekiwanej. Nie należy mylić kluczowości z frekwencją, te pierwsze bowiem mogą, ale nie muszą, mieć wysoką frekwencję w danym korpusie, wystarczy aby miały frekwencję wyższą niż w przypadku przyjętego przez badacza korpusu referencyjnego. Ponadto kluczowość mierzy się nie tyle różnicą w frekwencji okurencji ile poziomem

istotności tej różnicy (zwykle podawanej w wartości statystycznej $G^2$). Punktem odniesienia dla wyrazów kluczowych nie są zatem inne wyrazy w analizowanym korpusie tylko wyrazy w innym korpusie. Korpusami referencyjnymi w niniejszym badaniu są Narodowy Korpus Języka Polskiego (NKJP) oraz korpus plTenTen12. Słowa kluczowe są dobrym wskaźnikiem tematyki analizowanych tekstów, bowiem ich nadreprezentacja w stosunku do korpusu referencyjnego zdradza wątki, wokół których toczy się w tekście dyskusja.

Dyspersja to rozkład danych słów, w wypadku poniższej analizy słów kluczowych, w całym korpusie. Pozwala ona zaobserwować, czy interesujące nas słowo używane jest w nielicznych tekstach, czy też systematycznie powtarza się w różnych tekstach. Im więcej tekstów zawiera analizowany wyraz, tym bardziej uwiarygadnia się obserwacja o jego wysokiej charakterystyce frekwencyjnej; a im niższa dyspersja, tym mniej istotne są obserwacje dotyczące frekwencji. Słowo o wysokiej częstości występujące tylko w pojedynczym dokumencie nie jest miarodajnym wykładnikiem najczęściej używanych słów w danym korpusie i w zasadzie, nie należałoby go uwzględniać w listach frekwencyjnych z racji jego niskiej wiarygodności.

Słowa kluczowe w podkorpusach KK, przy zastosowaniu NKJP jako referencyjnego, są następujące (w nawiasach podano sumę wystąpień dla dokumentu):

Korpus marcowy: *osoba* (98), *wirus* (91), *móc* (88), *koronawirusa* (49), *przypadek* (44), *zakazić* (42), *szpital* (39), *żywność* (38), *choroba* (27), *epidemia* (26), *potwierdzić* (28), *ryzyko* (25).

Korpus kwietniowy: *szpital* (81), *koronawirusa* (39), *żywność* (37), *osoba* (35), *pacjent* (34), *zakazić* (31), *przypadek* (29), *liczba* (18), *wirus* (18), *epidemia* (17), *test* (16).

Korpus majowy: *osoba* (70), *zakazić* (41), *zakażenie* (38), *koronawirusa* (37), *test* (30), *liczba* (28), *szpital* (27), *grupa* (21), *COVID-19* (20), *maj* (20), *wirus* (18), *pacjent* (18), *pandemia* (16), *górnik* (14).

Korpus czerwcowy: *osoba* (78), *zakażenie* (46), *Polska* (37), *powiat* (37), *przypadek* (33), *zakazić* (33), *epidemia* (29), *Koronawirus* (16), *liczba* (29), *szczepionka* (19), *test* (18), *ognisko* (17),

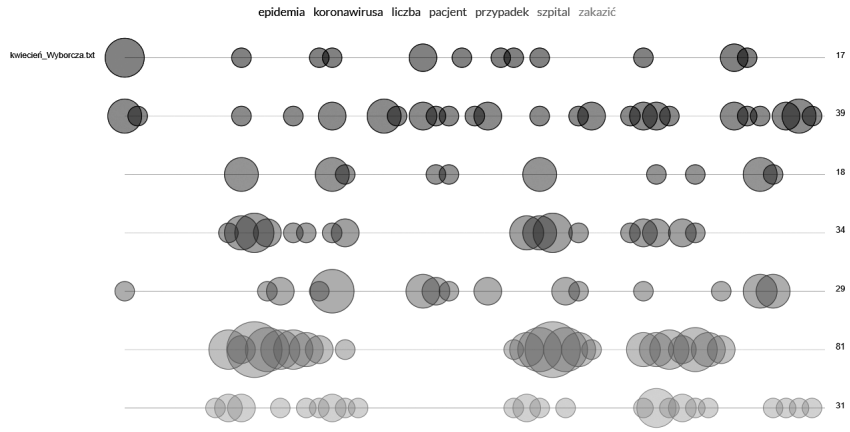
Korpus lipcowy: *osoba* (86), *zakażenie* (46), *przypadek* (45), *nowy* (32), *powiat* (29), *Sanepid* (29), *zakazić* (27), *Ministerstwo* (23), *zdrowie* (23), *Koronawirus* (21), *potwierdzić* (21), *województwo* (20), *Podkarpacie* (18),

W marcu tematyka artykułów oscylowała głównie wokół pojedynczych potwierdzonych przypadków osób zakażonych koronawirusem. Dużo uwagi poświęcono też *żywności*, konkretnie braku zalecenia gromadzenia jej dużych ilości oraz zapewnianiu, że teoretycznie nie jest ona źródłem wirusa, choć

zalecane jest gotowanie, bowiem w temperaturze pokojowej *mikroorganizmy mogą namnażać się bardzo szybko*. Temat żywności pojawiał się w marcu i w kwietniu i nie występował już w późniejszych artykułach w KK. W kwietniu artykuły były zdominowane przez tematykę funkcjonowania szpitali w czasie epidemii. W maju z kolei główne wątki dotyczyły zakażeń koronawirusem w czasie *pan*demii. Czerwiec to kontynuacja tematu zakażeń, ale pojawiały się też kwestie dotyczące szczepień i testów. W lipcu teksty koncentrowały się wokół instytucji – ogłoszeń Ministerstwa Zdrowia oraz Sanepidu, a także pojawiania się *nowych* ognisk zakażeń.

Niżej przedstawiono działanie aplikacji w KK, graficzną reprezentację niektórych słów kluczowych i ich dyspersji (KK) dla trzech wybranych podkorpusów:

**Rysunek 3.** Dystrybucja słów kluczowych w podkorpusie kwietniowym

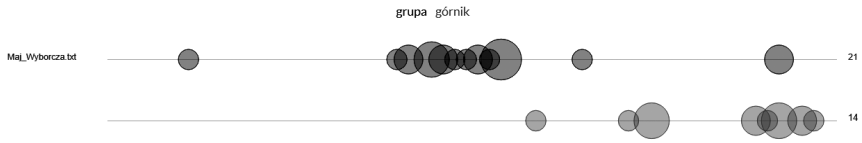



Źródło: opracowanie własne.

W kwietniu wyraz *epidemia* nie pojawiał się we wszystkich tekstach, ale jego okurencja jest stosunkowo równomiernie rozłożona w całym korpusie. Wyraz *pacjent* występował w kontekście *szpitala*, co sugeruje, że mowa jest głównie o pacjentach hospitalizowanych (a nie np. ambulatoryjnych), czyli o tych w ciężkim stanie. W sąsiedztwie tych słów pojawiało się również *zakazić*, co potwierdza zwiększenie się potwierdzonych przypadków zachorowań.

W połowie maja pojawiały się teksty w GW dotyczące *grupy* internetowej, która nie wierzy w koronawirusa, natomiast pod koniec miesiąca – o epidemii wśród górników i ich rodzin.

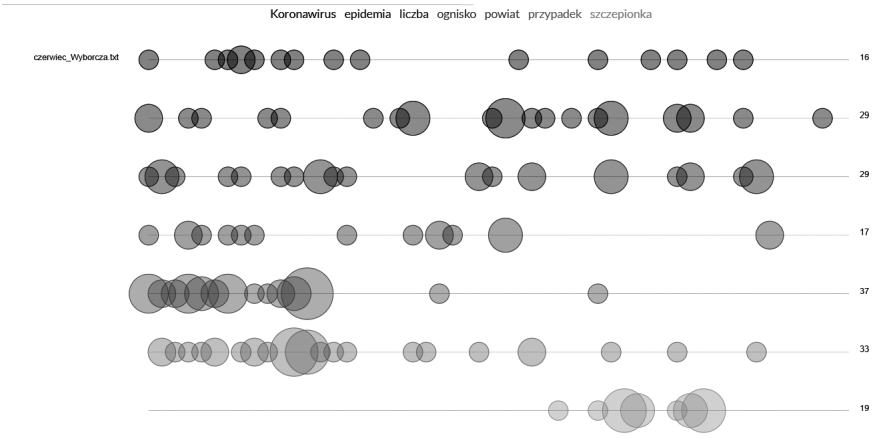
**Rysunek 4.** Dystrybucja słów kluczowych w podkorpusie majowym




Źródło: opracowanie własne.

W czerwcu, kiedy prasa nie pisała o *ogniskach* zachorowań, pojawiały się dyskusje na temat *szczepionki*, co kierowało uwagę czytelników z pogrążania się w rozważaniach nad trudną sytuacją epidemiczną na badania naukowe niosące nadzieję na jej zakończenie lub na spekulacje o braku konieczności jej poszukiwania, bowiem koronawirusa nie ma. Znamienne jest to, że opinie głoszące brak istnienia koronawirusa pojawiały się w drugiej połowie czerwca, tj. przed planowanymi wyborami korespondencyjnymi. Istnieje duża korelacja słów *przypadek* i *powiat*, które najczęściej pojawiały się w pierwszej połowie miesiąca, by później ustąpić miejsca rozważaniom na temat *szczepionki*.

**Rysunek 5.** Dystrybucja słów kluczowych w podkorpusie czerwcowym




Źródło: opracowanie własne.

Uwzględniając frekwencję słów w paśmie o najwyższej i najniższej częstości występowania, z wykorzystaniem korpusu referencyjnego plTenTen12, słowa kluczowe w korpusie czerwcowym są następujące (wartości kluczowości podano w nawiasie, za minimalną wartość przyjęto 400): *pandemia* (1264), *ozdrowieniec* (1126), *wyzdrowiały* (476), *schyłkowy* (410) itd. Dla porównania w danych z marca są następujące słowa: *kwarantanna* (753), *zakazić* (510), *aerozol* (334); w korpusie kwietniowym: *zakazić* (735), *pandemia* (716), *kwarantanna* (653), *sanepid* (465); w korpusie majowym: *zakazić* (735), *pandemia* (716), *kwarantanna* (653), *wyzdrowiały* (412); w korpusie lipcowym: *zarazić* (735), *pandemia* (716), *wyzdrowiały* (412).

Z powyższych danych wynika, że w maju i w czerwcu, oprócz informacji o pandemii, czyli poważnym stanie epidemicznym, wyjątkowo wysokie frekwencje odnotowano również dla wyzdrowień (*wyzdrowiały*, *ozdrowieniec*) i opinii o kończącej się pandemii (*schyłkowy*) lub krytyce takiego stanowiska przez GW: *Epidemia w Polsce nie jest w fazie schyłkowej, bo 100 chorych zaraża 111 osób* (25.06). Teksty publikowane w II fazie odnoszą się (krytycznie) zatem do bardziej optymistycznych tez głoszonych przez polityków niż te w I i III fazie. Warto wspomnieć, że pod koniec kwietnia rozpoczęło się odmrażanie gospodarki: 20.04 otwarto sklepy; galerie handlowe i hotele ponownie zaczęły działać od 4.05, od 18.05 – salony kosmetyczne i fryzjerskie oraz obiekty gastronomiczne, a od 25.05 planowano nawet częściowe wznowienie zajęć dydaktycznych dla studentów. Ponadto w II fazie toczyła się dyskusja na temat ewentualnych szczepionek, które mają zaradzić tej trudnej sytuacji. Co ciekawe, w III fazie, tj. przed kolejną datą wyborów, pojawiało się słowo kluczowe *wyzdrowiały*, czego nie wykazały inne statystyki wcześniej prezentowane.

Kończąc analizę danych trzeba wspomnieć o pewnych ograniczeniach niniejszego badania. Kolokacje, jak i słowa kluczowe oraz frekwencje prezentowane w niniejszym artykule, ilustrują dane pochodzące z tekstów zgromadzonych na potrzeby omawianych badań. Nadreprezentacja słowa *Podkarpacie* w kolokacjach, jest pochodną wszystkich artykułów prasowych składających się na utworzony korpus. Wynika ona zapewne częściowo z faktu, że korpus zawierał kilka artykułów publikowanych w rzeszowskim wydaniu GW (które stanowiły 8% KK). Możliwe jest, że w innym korpusie opartym na artykułach z GW, na przykład takich, które nie stanowiłyby pierwszej dziesiątki wyświetlonych *hiperlinków* w wyszukiwarce, ale reprezentowałyby drugą dziesiątkę, statystyka kolokacji (słów kluczowych i frekwencji) wyglądałaby odmiennie. Aby mieć całkowicie obiektywne informacje należałoby zgromadzić wszystkie artykuły z GW, również te, które są dostępne tylko po subskrypcji gazety.

Drugim ograniczeniem powtarzalności przedstawionych powyżej badań jest kwestia doboru wzorów matematycznych, za pomocą których oblicza się poszczególne parametry. Ponieważ teoretycznie niektóre parametry można określać odwołując się do różnych metod ich obliczania, wyniki dotyczące w szczególności kluczowości i kolokacji mogą się różnić w zależności od przyjętej metody statystycznej (zob. szczegóły w części *Technikalia*). Z tych powodów prezentowane dane należy traktować jako probierz, bowiem kluczową kwestią w badaniach korpusowych jest kompozycja korpusu oraz przyjęte metody statystyczne.

## Podsumowanie

Badanie korpusowe pozwoliło zauważyć, że w tekstach prasowych publikowanych online przez Gazetę Wyborczą w okresie 1.03–11.07.2020 obserwowalne są zmiany w sposobie przedstawiania epidemii koronawirusa, które współgrają z wypowiedziami polityków i dziennikarzy w mediach. W pierwszej fazie narracji o koronawirusie artykuły koncentrują się na opisie wirusa i cechach choroby, którą wywołuje. Widać narastające napięcie i panikę spowodowaną gwałtownym rozprzestrzenianiem się wirusa i rosnącą liczbą przypadków zachorowań oraz hospitalizacji pacjentów w Polsce. W drugiej fazie zauważalna jest zmiana retoryki, w szczególności w czerwcu, kiedy ton wypowiedzi na temat epidemii zaczęto zmieniać z pesymistycznego na bardziej optymistyczny, co ewidentnie było związane ze zbliżającymi się wyborami korespondencyjnymi, planowanymi początkowo na koniec czerwca. Od lipca ponownie pojawiły się w prasie teksty nagłaśniające ogromną skalę problemu, które w bardziej dramatyczny sposób opisywały rozwój epidemii, mimo iż w połowie lipca temat wyborów znowu był aktualny.

W analizie ilościowej tekstów prasowych dotyczących tematyki koronawirusa przedstawionej w niniejszym artykule omówiono kilka narzędzi i metod używanych w językoznawstwie korpusowym, które z powodzeniem mogą być stosowane do analizy tekstów prasowych i dyskursu politycznego. Przeprowadzając ilościowe analizy korpusowe tekstów należy wziąć pod uwagę fakt, że metody obliczeń wartości słów kluczowych, częstości słów w korpusie czy stopnia współwystępowania kolokacji, a także wybór zastosowanych narzędzi informatycznych do identyfikacji, segmentacji i znakowania słów, mają istotne znaczenie w korpusowej analizie tekstów i w związku z tym mogą wpływać

na wyniki badań. Niepodważalną zaletą analizy korpusowej z kolei jest to, że pozwala ona obserwować tendencje i wzorce leksykalno-gramatyczne na podstawie kryterium rekurencji w dużych zbiorach tekstów oraz że taka analiza przebiega niezwykle sprawnie; przetwarzanie danych określa się bowiem w sekundach, co nie byłoby możliwe w przypadku analizy wykonywanej ręcznie w badaniach jakościowych.

## Bibliografia


Ädel A. (2010), *How to use corpus linguistics in the study of political discourse*, (w:) O'Keeffe A., McCarthy M. (red.), *The Routledge Handbook of Corpus Linguistics*, Routledge, London.

Baker P. (2004), *Unnatural acts: discourse of homosexuality within the House of Lords debates on gay male law reform*, „Journal of Sociolinguistics", No. 8(1).

Baker P., Gabrielatos C., Khosravinik M., Krzyżanowski M., McEnery T., Wodak R. (2008), *A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press*, „Discourse and Society", No. 19(3).

Bączkowska A. (2016), *Korpusowa analiza dyskursu związanego z tematyką imigracji w brytyjskiej prasie opiniotwórczej*, „Conversatoria Linguistica", No. 10.

Bączkowska A. (2017), *Krytyczna analiza dyskursu prawicowo-populistycznego: analiza korpusowa przemówień wyborczych Donalda Trumpa*, (w:) Pierzchalski F., Rydliński B. (red.), *Populizmu w XXI wieku. Krytyczna analiza sukcesów prawicowego populizmu w Europie i USA*, Elipsa, Warszawa.

Bączkowska A. (2019a), *Obraz Polaka imigranta w brytyjskiej prasie opiniotwórczej: analiza korpusowa i krytyczna analiza dyskursu*, (w:) Benenowska I., Bączkowska A., Czechowski W. (red.), *Komunikowanie wartości – wartość komunikowania*, Wydawnictwo Uniwersytetu Kazimierza Wielkiego, Bydgoszcz.

Bączkowska A. (2019b), *A Corpus-Assisted Critical Discourse Analysis of "Migrants" and "Migration" in the British Tabloids and Quality Press*, (w:) Lewandowska-Tomaszczyk B. (red.), *Contacts and Contrasts in Cultures and Languages*, Springer, Cham.

Bączkowska A. (2020a), *Framing the conceptualization of obesity in online Chinese and British Quality Newspapers: A corpus-assisted study*, (w:) Lewandowska-Tomaszczyk B. (red.), *Cultural Conceptualiztions in Language and Communication*, Springer, Cham.

Bączkowska A. (2020b), *Healthy lifestyle, dieting, fitness and bodybuilding: compliments in the context of Polish online discussion forums and message boards*, (w:) Placencia M.E., Eslami Z.R. (red.), *Complimenting Behavior and (Self-)Praise across Social Media*, John Benjamins, Amsterdam.

Bączkowska A., Khokhlacheva Y. (2019), *The world's coldest, newest and most remote capital" – the perception of Astana by the British and American quality press: a corpus-assisted analysis*, „Scripta Neophilologica Posnaniensia", No. 19.

Bączkowska A., Gabdrakhmanova S., Akhmetova G. (2020), *The representation of the capital of Kazakhstan in Central Asia online news coverage: a corpus-assisted analysis*, „Studia Linguistica".

Bednarek M. (2006), *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*, Bloomsbury, London.

Bednarek M. (2015), *Corpus-assisted multimodal discourse analysis of television and film narratives*, (w:) Baker P., McEnery T. (red.), *Corpora and Discourse Studies. Integrating Discourse and Corpora*, Palgrave, London.

Cap P. (2006), *Legitimization in Political Discourse: A Cross-Disciplinary Perspective on the Modern US War Rhetoric*, Cambridge Scholars Press, Newcastle.

Charteris-Black J. (2004), *Corpus Approaches to Critical Metaphor Analysis*, Palgrave, London.

Chen X., Yan Y., Hu J. (2019), *A corpus-based study of Hilary Clinton's and Donald Trump's linguistic styles*, „International Journal of English Linguistics", No. 9(3).

Dalman A. (2017), *A corpus analysis of Donald Trump's political communications*, „International Journal of Current Research", No. 9(11).

Gabrielatos K., Marchi A. (2011), *Keyness: matching metrics to definitions*, (w:) *Corpus Linguistics in the South 1*, University of Portsmouth.

Kieraś W., Kobyliński Ł., Ogrodniczuk M. (2018), *Korpusomat — a tool for creating searchable morphosyntactically tagged corpora*, „Computational Methods in Science and Technology", No. 24(1).

Kilgarriff A. (2009), *Simple maths for keywords*, (w:) Mahlberg M., González-Díaz V., Smith C. (red.) *Proceedings of the Corpus Linguistics Conference,* CL2009, Liverpool.

Kilgarriff A., Baisa V., Bušta J., Smrž P., Tugwell D. (2004), *The Sketch Engine*, „Information Technology".

Khosravinik M. (2010), *The representation of refugees, asylum seekers and immigrants in British Newspapers: A critical discourse analysis*, „Journal of Language and Politics", No. 9(1).

Kobyliński Ł., Kieraś W. (2016), *Part of speech tagging for Polish. State of the art and future perspectives*, (w:) *Proceedings of the 17th International Conference*

*on Intelligent Task Processing and Computational Linguistics*, http://nlp.ipipan.waw.pl/Bib/kob:kie:16.pdf [dostęp: 16.04.2020].
Koteyko N. (2007), *A diachronic approach to meaning: English loanwords in Russian opposition discourse*, „Corpora", No. 2(1).
Lewandowska-Tomaszczyk B. (2005), *Podstawy językoznawstwa korpusowego*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.
Lewicki A.M. (1976), *Wprowadzenie do frazeologii syntaktycznej. Teoria zwrotu frazeologicznego*, Uniwersytet Śląski, Katowice.
Liu D., Lei L. (2018), *The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election*, „Discourse, Context and Media", No. 25.
Moon R. (2016), *A corpus-linguistic analysis of news coverage in* Kenya's Daily Nation *and* The Times of London, „International Journal of Communication", No. 10.
O'Halloran K. (2010), *How to use corpus linguistics in the study of media discourse*, (w:) O'Keeffe A., McCarthy M. (red.), *The Routledge Handbook of Corpus Linguistics*, Routledge, London.
Schneider K. (1999), *Exploring the roots of popular English news-writing – a preliminary report on a corpus-based project*, (w:) Diller H-J., Otto E., Stratmann G. (red.), *English via Various Media*, Universitätsverlag Winter, Heidelberg.
Sofyaningrat S., Suhardijanto S., Yuwono U. (2019), *Representation of Ulama collocations in online Kompas media: corpus-based critical discourse analysis*, (w:) Handoko S.S., Gusdi S., *Proceedings of the 4th International Seminar on Linguistics (ISOL-4)*, Padang.
Waszczuk J. (2012) *Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language*, (w:) *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India, http://zil.ipipan.waw.pl/Concraft?action=AttachFile&do=view&target=coling2012.pdf [dostęp: 12.07.2020].
Wilson A. (2013), *Embracing Bayes factors for key item analysis in corpus linguistics*, (w:) Bieswanger M., Koll-Stobbe A. (red.), *New Approaches to the Study of Linguistic Variability. Language Competence and Language Awareness in Europe*, Peter Lang, Frankfurt.

**Publikacje internetowe**

(www1) EbizMBA, *Top 15 most popular search engines*, http://www.ebizmba.com/articles/search-engines [dostęp: 12.04.2019].
(www2) https://pl.wikipedia.org/wiki/Gazeta_Wyborcza#Nak%C5%82ad_i_sprzeda%C5%BC [dostęp: 12.04.2019].
(www3) https://pl.wikipedia.org/wiki/Rzeczpospolita_(gazeta)#Nak%C5%82ad_i_sprzeda%C5%BC [dostęp: 12.04.2019].
(www4) https://pl.wikipedia.org/wiki/Gazeta_Polska_Codziennie [dostęp: 12.04.2019].
(www5) https://pl.wikipedia.org/wiki/Gazeta_Polska_Codziennie [dostęp: 12.04.2019].

# THE THEME OF CORONAVIRUS IN POLISH ONLINE PRESS – A CORPUS-ASSISTED STUDY

## Abstract

The aim of the study presented in this paper was to analyse how the topic of coronavirus was described by the Polish quality newspaper "Gazeta Wyborcza" in the period between the 1st of March and the 12th of July 2020, i.e. the day of presidential elections in Poland. The methodology used in the investigation follows a corpus-assisted analysis, which allows one to conduct a quantitative study of large collections of data (language corpora). The methodology is typically used in linguistics, yet the study shows that it can be successfully employed in a quantitative analysis of press and political discourse. The data were analysed automatically by computer software dedicated to Polish language (Korpusomat) and by tools available in the Sketch Engine system that allow one to examine English language data. The study demonstrates that the theme of coronavirus presented in March and April focused largely on the description of the virus and the consequences of being infected (hospitalization), whilst the articles published between May and July contain more information about possible vaccines and, contrary to facts, they emphasise a declining trend in morbidity rates.

**Keywords**: language corpora, press discourse, political discourse, coronavirus, Polish quality press.
**JEL Codes**: C80, I10

**Afiliacja:** **dr hab. Anna Bączkowska**
Uniwersytet Gdański
Instytut Anglistyki i Amerykanistyki
ul. Wita Stwosza 51
80-308 Gdańsk
e-mail: anna.k.baczkowska@gmail.com