DOI: 10.33896/PorJ.2023.4.2

Natalia Zawadzka-Paluektau
Instytut Podstaw Informatyki Polskiej Akademii Nauk
natalia.zawadzka-paluektau@ipipan.waw.pl
ORCID: 0000-0003-4969-2039

Aleksandra Tomaszewska
Instytut Podstaw Informatyki Polskiej Akademii Nauk
aleksandra.tomaszewska@ipipan.waw.pl
ORCID: 0000-0001-6379-3034

Joanna Wołoszyn
Instytut Podstaw Informatyki Polskiej Akademii Nauk
joanna.woloszyn@ipipan.waw.pl
ORCID: 0000-0002-8923-414X

PO NARODOWYM KORPUSIE JĘZYKA POLSKIEGO – ZMIANY W SŁOWNICTWIE POLSKIM OSTATNIEJ DEKADY NA PRZYKŁADZIE SŁÓW KLUCZOWYCH

1. WSTĘP

Podobnie jak inne języki polszczyzna ulega naturalnym zmianom, często pod wpływem języków obcych (Waszakowa 2005; Mańczak-Wohlfeld 2010: 13; Zemlanaja 2016: 354), głównie w warstwie leksykalnej (Krasowska 2018: 99). Zmiany diachroniczne słownictwa są przedmiotem badań językoznawczych już od XIX wieku (Wawrzyńczyk 2011; Siuciak 2015: 149). Stosunkowo od niedawna badane są również w korpusach – dużych zbiorach próbek naturalnie występującego języka zapisanych w formacie umożliwiającym odczyt komputerowy i spełniających kryteria reprezentatywności i skończonej wielkości (McEnery, Xiao, Tono 2006).

Celem niniejszego badania jest zdobycie wiedzy o zmianach w trendach leksykalnych w języku polskim ostatniego dziesięciolecia właśnie za pomocą metod językoznawstwa korpusowego, na podstawie dwóch wielomilionowych zbiorów tekstów. Wykorzystanie tak obszernego i zróżnicowanego materiału i zastosowanie metod statystycznych do jego analizy pozwalają na wyciągnięcie bardziej wiary-

godnych wniosków dotyczących zmian diachronicznych na przestrzeni stosunkowo długiego okresu (ponad dziesięciu lat) od tych uzyskanych z analiz jakościowych prowadzonych na znacznie mniejszych próbkach. Również pod względem zawartości – teksty publikowane w prasie i w Internecie – analizowane zbiory zostały dobrane tak, by jak najlepiej odzwierciedlać rozwój języka polskiego w ostatniej dekadzie. O ile jeszcze niedawno to prasa najżywiej podążała za zmianami i nowymi trendami językowymi (Mańczak-Wohlfeld 2006: 34), o tyle w ostatnich latach w znacznie większym stopniu na te innowacje reaguje Internet (Cierpich-Kozieł 2022). Oba te źródła obejmują szeroki zakres tematów i typów tekstów, kierowanych do osób w różnym wieku, o różnym wykształceniu, statusie socjoekonomicznym, poglądach i zainteresowaniach itp.

Wśród wielu typów korpusów warto wyróżnić korpusy ogólne, które charakteryzują się dużym rozmiarem, zwykle przekraczającym 10 milionów słów, i obejmują różne odmiany języka, dzięki czemu wnioski z analiz przeprowadzonych z ich wykorzystaniem mogą być w pewnym stopniu rozciągnięte na szerszą populację. Pełnią często funkcję korpusów referencyjnych, czyli takich, które służą za punkt odniesienia w badaniach (Paryzek 2011: 27); należą do nich między innymi tzw. korpusy narodowe. Część tych korpusów jest co jakiś czas aktualizowana, np. korpus czeski (Czech National Corpus, zob. Křen 2015) i korpus amerykańskiej odmiany angielskiego (COCA, zob. Davies, 2008–). Narodowy Korpus Języka Polskiego (Przepiórkowski i in. 2012), czyli jedyny aktualnie dostępny korpus ogólnej polszczyzny, nie był natomiast aktualizowany zasadniczo od 2010 r., a zatem na potrzeby niniejszego badania stanowi idealny korpus referencyjny: porównanie go z nowszymi tekstami pozwoli bowiem na identyfikację najistotniejszych trendów leksykalnych w języku polskim właśnie na przestrzeni ostatniej dekady – okresu, którego zbiór ten nie obejmuje.

Do wydobywania informacji z korpusów wykorzystywane są głównie kolokacje, konkordancje i listy słów (frekwencyjne). Szczególnym zastosowaniem tych ostatnich są słowa kluczowe, wprowadzone do użycia w latach 90. przez Mike'a Scotta, który zdefiniował je jako jednostki leksykalne występujące znacznie częściej w korpusie badanym niż w referencyjnym (Scott 1997: 236), szczególnie dla niego charakterystyczne (Egbert, Biber 2019: 77). Aby uzyskać listy słów kluczowych, przeprowadza się porównania statystyczne częstości wyrazów w obu korpusach. Na interpretację tego, co jest słowem kluczowym, wpływają wybrane parametry i zawartość korpusu referencyjnego. Jednostki te są jedynie pierwszymi wskaźnikami w analizie korpusu (Scott 2010: 56) – przeprowadza się zatem analizy jakościowe słów kluczowych zidentyfikowanych w ramach wstępnej analizy ilościowej. Tak zdefiniowane słowa kluczowe mogą mieć znaczenie społeczne, kulturowe lub polityczne.

W diachronicznych badaniach słów kluczowych sięga się po korpusy o odpowiedniej strukturze czasowej, które pozwalają zidentyfikować wyrazy odzwierciedlające dyskursy oraz wzorce użycia języka statystycznie istotne w jednym okresie

w porównaniu z innym (Baker, McEnery 2019: 217). Takie badania pozwalają uzyskać wgląd w „zmieniające się wartości i perspektywy społeczeństwa" (Csomay, Young 2020: 71) oraz „odkryć prawidłowości [*w języku*], których możemy nie być świadomi" (McEnery, Baker 2016: 12). Przykładem niedawno opublikowanych diachronicznych badań słów kluczowych są prace Clarke, Brookesa i McEnery'ego (2022) o zmianach w dyskursach prasowych dotyczących islamu czy Csomay i Younga (2020) o użyciu języka w popkulturze na przestrzeni trzech dekad na przykładzie dialogów z serialu *Star Trek*.

2. KORPUSY I METODOLOGIA

W celu identyfikacji słów kluczowych charakterystycznych dla języka polskiego ostatniej dekady porównałyśmy dwa korpusy tekstów opublikowanych w tym okresie z podzbiorem prasy ze zrównoważonego Narodowego Korpusu Języka Polskiego (którego podstawa materiałowa, jak wspomniano wyżej, kończy się zasadniczo na roku 2010 (Przepiórkowski i in. 2012)). Pierwszy korpus badany (nazywany dalej korpusem prasowym) składa się z artykułów publikowanych między 2011 a 2020 rokiem w czasopismach „Polityka", „Przegląd", „Gazeta Polska" i „Dziennik Gazeta Prawna", zawiera 206 milionów segmentów. Drugi (nazywany dalej korpusem webowym) jest podzbiorem korpusu danych internetowych, na który składają się statyczne strony internetowe monitorowane regularnie od listopada 2021 roku do października 2022 roku, i liczy około 990 milionów segmentów. Analiza została przeprowadzona na korpusach niezlematyzowanych, w których wszystkie słowa zostały sprowadzone do małych liter – jest to częsta praktyka w badaniach słów kluczowych, gdzie obecność wielu form jednego lematu jest traktowana jako dowód systematyczności procesu wyboru jednostek kluczowych (Kilgarriff 2012: 6). Do wygenerowania list jednostek kluczowych dla obu tych korpusów (w porównaniu z podzbiorem prasy ze zrównoważonego NKJP) posłużyłyśmy się miarą ilorazu szans, która oblicza stosunek szans wystąpienia danego słowa w jednym zbiorze do szans jego wystąpienia w innym[1]. Następnie pogrupowałyśmy 60 rzeczowników pospolitych o najwyższych wartościach miary kluczowości dla każdego z badanych korpusów w wyłonione indukcyjnie kategorie tematyczne i przeanalizowałyśmy ich funkcje dyskursywne.

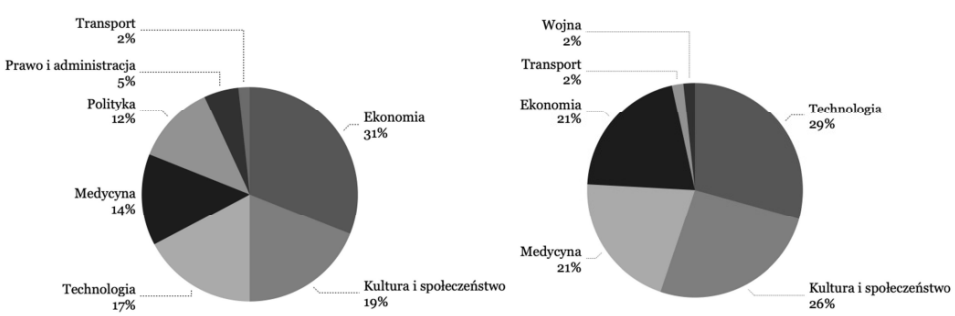
[1] Za pomoc w przygotowaniu skryptu w Pythonie służącego do wygenerowania list słów kluczowych, a także cenne uwagi do pierwszej wersji tekstu, dziękujemy dr. Witoldowi Kierasiowi.

3. WYNIKI

Opisana wyżej procedura pozwoliła na wyodrębnienie z analizowanych korpusów jednostek wskazujących na tendencje językowe charakterystyczne dla ostatniej dekady w porównaniu z poprzednią. Są to przede wszystkim słowa funkcjonujące w języku polskim od niedawna (w większości zapożyczenia z języka angielskiego), ale dobrze już w nim zakorzenione i rozprzestrzenione, o czym świadczą ich duże częstości w analizowanych korpusach (w porównaniu z – w przeważającej mierze zerowymi – frekwencjami w korpusie referencyjnym). Wartości te dla każdego z wybranych słów kluczowych (wraz z miarami kluczowości) zostały umieszczone w tabelach 6. i 7. w *Aneksie do tekstu*.

Połączenie zidentyfikowanych jednostek kluczowych w kategorie tematyczne pozwoliło uwypuklić najważniejsze trendy: o ile obecność na liście słów kluczowych pojedynczej jednostki może być wynikiem przypadku lub autorskiego stylu, o tyle grupy podobnych lub powiązanych ze sobą wyrazów są zazwyczaj uznawane za prawdziwy dowód zmian w języku (Partington 2010: 87) i, co za tym często idzie, odpowiadających im zmian pozajęzykowych. Wykres 1. pokazuje, jakie kategorie tematyczne tworzyły słowa kluczowe wyselekcjonowane w wyniku niniejszej analizy. Wyniki te sugerują, że najbardziej znaczącym zmianom w ostatnim dziesięcioleciu ulegały dyskursy ekonomiczny, technologiczny, medyczny i społeczno-kulturowy – tylko w tych czterech kategoriach tematycznych mieści się 77 i aż 96 procent słów kluczowych dla, odpowiednio, korpusu prasowego i korpusu webowego. Ich dokładniejszemu omówieniu zostały poświęcone części 3.1–3.4. Pozostałe słowa kluczowe dotyczą polityki, prawa i administracji, transportu oraz konfliktów zbrojnych (część 3.5).

Wykres 1. Kategorie tematyczne słów kluczowych w korpusie prasowym (po lewej) i webowym (po prawej)



3.1. Ekonomia

Słownictwo ekonomiczne zajmuje prawie jedną trzecią i jedną czwartą list jednostek kluczowych zidentyfikowanych, odpowiednio, w korpusach prasowym i webowym. Jak widać w tabeli 1., zdecydowanie dominują wśród nich wyrazy związane z nowymi technologiami finansowymi (używany do ich oznaczania pochodzący z angielskiego skrót *fintech* jest zresztą jednym ze słów kluczowych w korpusie prasowym), takie jak *bitcoin*, *kryptowaluty*, *blockchain* i *nft* (z angielskiego *non-fungible token*, czyli 'token niewymienialny'). Jednocześnie żadne z tych słów nie występuje w ogóle w korpusie referencyjnym: kompilacja NKJP przypadła bowiem na bardzo wczesne początki rozwoju tego nowego modelu biznesowego, który dopiero w drugiej dekadzie XXI wieku zaczął przyciągać uwagę publiczną jako coraz bardziej znaczący i kontrowersyjny.

Tabela 1. Słowa kluczowe z kategorii ekonomia[2]

Podkategoria	Korpus prasowy	Korpus webowy
Nowe technologie finansowe	*bitcoin*, *bitcoina*, *bitcoinów*, *blockchain*, *fintech*, *kryptowalut*, *kryptowaluty*	*bitcoin*, *bitcoina*, *blockchain*, *kryptowalut*, *kryptowaluty*, *nft*
Energia	*fotowoltaiki*, *mikroinstalacji*	*emisyjności*, *fotowoltaiki*, *smr*
Ogólne	*jzp*, *paymentu*	*startupów*, *esg*
Podatki	*e-deklaracji*, *pkpir*, *prewspółczynnik*, *prewspółczynnika*	
Kredyty	*frankowicze*, *frankowiczów*	
Programy społeczne	*500+*	

Drugą najliczniejszą podgrupę słów kluczowych dotyczących ekonomii stanowią nawiązania do sposobów pozyskiwania energii, w szczególności tych, które są kojarzone z mniejszą *emisyjnością* (jedno ze słów kluczowych), a co za tym idzie – mniejszymi kosztami środowiskowymi, takimi jak fotowoltaika i rozszczepianie atomu (*smr* to akronim od angielskiego terminu oznaczającego 'małe modułowe reaktory jądrowe'). Odzwierciedla to rosnącą w polskim społeczeństwie świadomość ekologiczną (Open Research 2022).

Wśród pozostałych słów kluczowych znajduje się kilka nawiązań do podatków (ale tylko w korpusie prasowym) – wśród nich uwagę zwracają *e-deklaracje*, wskazujące na zmianę w sposobie kontaktowania się obywateli z urzędami – a także do nowych modeli biznesowych, takich jak *startupy*. Obie te zmiany są też kojarzone

[2] Tu i w kolejnych tabelach słowa kluczowe zostały ułożone w kolejności alfabetycznej w celu zgrupowania wszystkich form wyrazowych danego lematu (np. *bitcoin* i *bitcoina* są dwiema różnymi formami tego samego lematu: *bitcoin*).

z rozwojem nowych technologii. W drugiej dekadzie tego tysiąclecia polska prasa poświęcała również uwagę kwestiom społeczno-gospodarczym, takim jak program wsparcia rodzin 500+ oraz problemy ze spłatą kredytów mieszkaniowych zaciągniętych we frankach szwajcarskich, które pojawiły się, gdy w 2015 roku kurs tej waluty wzrósł znacząco – w związku z czym na określenie osób zmagających się z nimi ukuto słowo *frankowicze*.

3.2. Technologia

Terminy technologiczne stanowiły 17% wszystkich zidentyfikowanych jednostek kluczowych w korpusie prasowym i dominowały na liście słów kluczowych w korpusie webowym (29%). W tym pierwszym, jak można zaobserwować w tabeli 2., są to przede wszystkim różne warianty słów *smartfon* i *dron*, a także słowa *lte, streaming* oraz *cyberbezpieczeństwo*. W tym drugim również przeważają nawiązania do smartfonów (należy do nich także słowo *flagowiec*, oznaczające telefon o najlepszych parametrach od danego producenta), ale pojawiają się tu też *smartwatche*, czyli zegarki naręczne mające wybrane funkcje smartfonu, oraz szereg bardziej specjalistycznych terminów (np. *amoled*, *thunderbolt*). Również i tu odznaczają się nawiązania do bezpieczeństwa w sieci, takie jak *cyberataki* i *ransomware* (typ tzw. złośliwego oprogramowania).

Tabela 2. Słowa kluczowe z kategorii technologia

Korpus prasowy	Korpus webowy
cyberbezpieczeństwo, *dron*, *drona*, *dronów*, *lte*, *smartfona*, *smartfonach*, *smartfonem*, *smartfonie*, *streamingu*	*amoled*, *cyberataki*, *drona*, *flagowców*, *hvec*, *lte*, *microsd*, *ransomware*, *rendery*, *smartfona*, *smartfonach*, *smartfonem*, *smartfonie*, *smartfonu*, *smartwatch*, *smartwatche*, *thunderbolt*

3.3. Kultura i społeczeństwo

Jak widać na wykresie 1., słowa dotyczące kultury i społeczeństwa stanowią około jednej czwartej obu list jednostek kluczowych. Jeszcze bardziej niż w wypadku słów z dziedziny ekonomii widać tu wpływ technologii na życie społeczne i kulturalne: zdecydowana większość jednostek kluczowych odnosi się bowiem do aktywności w przestrzeni internetowej (tabela 3.). Są to mianowicie określenia nowych form rozrywki, takie jak *tiktok* (bardzo krótki materiał wideo publikowany w aplikacji o tej samej nazwie), *selfie* (zdjęcie zrobione samemu sobie za pomocą smartfona, zwykle w celu umieszczenia go na portalu społecznościowym), *podcast* (rodzaj audycji radiowej udostępnianej przez Internet), *instastories* (w liczbie pojedynczej *instastory* – sposób publikacji zdjęć czy wideo w aplikacji Instagram), sport elektroniczny (*esports*, czyli rywalizacja w grach komputerowych) i *fanpage* (strona przeznaczona

dla fanów twórcy, produktu itd.). Wskazują one na ogromny wpływ, jaki nowe technologie wywierają nie tylko na środki komunikacji i pozyskiwania informacji, ale przede wszystkim – na sposoby spędzania wolnego czasu.

Tabela 3. Słowa kluczowe z kategorii kultura i społeczeństwo

Podkategoria	Korpus prasowy	Korpus webowy
Formy rozrywki	*fanpage*, *selfie*, *timeshare*	*esports*, *instastories*, *instastory*, *oktagonie*, *podcaście*, *selfie*, *tiktoku*
Zachowanie wyrażające postawę	*hejt*, *hejtem*, *hejtu*, *lajków*	*hejt*, *hejtu*
Osoby	*lgbt+*, *hejterów*, *milenialsów*	*influencer*, *influencerka*, *influencerów*, *youtuber*
Inne	*#metoo*	*alerty*, *metawersum*

Wiele jednostek kluczowych w tej kategorii odnosi się do zachowań w Internecie, najczęściej wyrażających skrajnie negatywną postawę wobec jakiejś osoby, rzeczy lub zjawiska. Kluczowość różnych form słowa *hejt* (z angielskiego *hate*) wskazuje na skalę problemu i trudność w wyeliminowaniu go. Jedynym słowem wyrażającym pozytywny stosunek do osób / zjawisk w tej podkategorii jest *lajk* (z angielskiego *like*), który oznacza sposób wyrażania aprobaty wobec treści publikowanych w Internecie. Warto zwrócić uwagę na spolszczenie pisowni i bezproblemową odmianę obu tych słów, wskazujące na ich istotność i przyswojenie w polszczyźnie.

Również nazwy grup osób, zidentyfikowane jako słowa kluczowe w badanych korpusach, odnoszą się często do aktywności w Internecie, szczególnie w korpusie webowym. Są to, między innymi, zapożyczenia takie jak *influencerzy* i *influencerka*, oznaczające osoby o dużym gronie odbiorców, których styl życia i wybory (często konsumpcyjne) mogą wpływać (z angielskiego *influence*) na obserwatorów. Warto tu zauważyć, że żeńska forma tego słowa występuje w korpusie webowym prawie dwukrotnie częściej i ma dużo wyższy wskaźnik kluczowości niż jej męski odpowiednik (oba te słowa mają zerową frekwencję w korpusie referencyjnym), co może sugerować, że funkcję tę pełnią przede wszystkim kobiety (forma liczby mnogiej *influencerów* natomiast, choć gramatycznie jest rodzaju męskiego, może obejmować zarówno mężczyzn, jak i kobiety). Pozostałymi słowami w tej podgrupie są: *youtuber* (twórca filmów wideo publikowanych w serwisie YouTube), *hejterzy* (czyli osoby stosujące hejt, o którym mowa wyżej), *milenialsi* (pokolenie osób urodzonych w latach 80. i 90. poprzedniego stulecia, nazywane również „pokoleniem cyfrowym" – aktywnie korzystającym z technologii i mediów cyfrowych) oraz *lgbt+* – w odpowiedzi na rosnącą obecność mniejszości seksualnych w przestrzeni publicznej oraz świadomość ich dyskryminacji.

Innym istotnym zjawiskiem o charakterze społecznym, na które wskazują zidentyfikowane słowa kluczowe, jest ruch opisywany w mediach społecznościowych za pomocą hashtagu – czyli znacznika (ang. *tag*) w formie słowa lub wyrażenia poprzedzonego kratką (ang. *hash*) opisującego skrótowo daną treść internetową – *#metoo*. Ruch #metoo (z angielskiego: *ja też*) powstał w roku 2017, by zwrócić uwagę na powszechność problemu molestowania seksualnego kobiet, a sam hashtag przynajmniej początkowo wykorzystywany był głównie przez ofiary do oznaczania ich historii.

Ostatnie dwa słowa z tej kategorii dotyczące nowych technologii to *alerty* (powiadomienia o aktywności w mediach społecznościowych, a także komunikaty o zagrożeniach wysyłane przez Rządowe Centrum Bezpieczeństwa[3]) i *metawersum*. To drugie to termin zapożyczony z angielskiego (*metaverse*), utworzony poprzez połączenie greckiego prefiksu *meta* (sugerującego przekraczanie czegoś, przechodzenie na wyższy, bardziej abstrakcyjny poziom) i słowa *universe* ('wszechświat'), oznaczający „wirtualne środowisko łączące rzeczywistość fizyczną i cyfrową, powstałe dzięki konwergencji technologii internetowych i rzeczywistości powiększonej" (*extended reality*; Lee i in. 2021: 1).

Jedyne dwa słowa kluczowe z kategorii kultura i społeczeństwo, które nie mają związku z nowymi technologiami, to *timeshare* (kolejne zapożyczenie z angielskiego) i *oktagon*: to pierwsze dotyczy zakupu nieruchomości wakacyjnej za ułamek jej prawdziwej ceny, ale pod warunkiem korzystania z niej tylko przez część roku; to drugie nawiązuje do mieszanych sztuk walki, które toczą się właśnie w ośmiokątnym ringu.

3.4. Medycyna

Ostatnia grupa słów kluczowych, która w obydwu badanych korpusach znajduje się wśród czterech dominujących kategorii, składa się z wyrazów związanych z medycyną, a ściślej mówiąc — dotyczących pandemii COVID-19. Jedyne słowo kluczowe (zidentyfikowane w korpusie prasowym), które nie odnosi się bezpośrednio do tej choroby, to *e-recepta*, która z kolei związana jest z głównym trendem zaobserwowanym w niniejszym badaniu, czyli wpływem postępu technologicznego na różne dziedziny życia (elektroniczna recepta lekarska zastępuje bowiem tradycyjny, papierowy dokument). Wśród słów dotyczących pandemii przeważają terminy nazywające chorobę, powodującego ją wirusa oraz jego warianty. Poza tym kluczowe są nawiązania do sposobów walki z pandemią, takie jak *lockdown* (czyli przymusowa izolacja i zakaz przemieszczania się) i *wyszczepienia* (słowo dotychczas kojarzone

[3] Ale wykorzystane też do przesyłania informacji dotyczącej organizacji wyborów prezydenckich w 2020 r.

raczej z językiem specjalistycznym). Pojawia się też nawiązanie do osób, które podważają sens szczepień, a nawet przekonują, że szczepionki szkodzą zdrowiu, czyli tzw. *antyszczepionkowców*. Oczywiście takie ruchy istniały już wcześniej, jednak analiza sugeruje, że zaczęły przyciągać szerszą uwagę publiczną dopiero po wybuchu pandemii COVID-19.

Tabela 4. Słowa kluczowe z kategorii medycyna

Korpus prasowy	Korpus webowy
covid, *covid-19*, *e-recepty*, *koronawirusem*, *koronawirusie*, *lockdown*, *lockdownu*, *sars-cov-2*	*antyszczepionkowców*, *covid*, *covid-19*, *lockdown*, *lockdownów*, *lockdownu*, *lockdowny*, *koronawirusem*, *omikrona*, *omikronem*, *omikronu*, *wyszczepienia*

3.5. Pozostałe kategorie

Jak widać w tabeli 5., pozostałe słowa kluczowe dotyczą transportu (obydwa korpusy), polityki i kwestii prawno-administracyjnych (tylko korpus prasowy) oraz wojny (tylko korpus webowy), przy czym w tej ostatniej grupie znajduje się tylko jedno słowo – *deeskalacji* – odnoszące się do agresji Rosji na Ukrainę. Słowa dotyczące transportu to przede wszystkim warianty rzeczownika *elektromobilność*, odzwierciedlającego zarówno postęp technologiczny w tej dziedzinie, jak i większą uwagę poświęcaną kwestii ochrony środowiska. Dodatkowo w korpusie prasowym jako kluczowe zidentyfikowano też słowo *e-myta* (przypadek analogiczny do *e-deklaracji* i *e-recept*, tylko że dotyczący opłat drogowych).

Tabela 5. Słowa kluczowe z pozostałych kategorii

Kategoria	Korpus prasowy	Korpus webowy
Transport	*elektromobilności, elektromobilność*, *e-myta*	*elektromobilności*
Polityka	*brexicie*, *brexit*, *brexitem*, *brexitu*, *miesięcznice*, *ziobryści, ziobrystów*	
Prawo i administracja	*cuw*, *sąd*, *sde*	
Wojna		deeskalacji

Słowa kluczowe dotyczące polityki wskazują na wagę wyjścia Wielkiej Brytanii z Unii Europejskiej (określanego popularnie jako *brexit*), na obecność na polskiej scenie politycznej nowej frakcji zgromadzonej wokół Zbigniewa Ziobry (*ziobryści*) oraz na debatę publiczną wokół tzw. miesięcznic smoleńskich – wydarzeń o charak-

terze polityczno-religijnym upamiętniających katastrofę samolotu TU-154 z 96 osobami na pokładzie (w tym z ówczesnym prezydentem Lechem Kaczyńskim). Wreszcie słowa z kategorii prawo i administracja dotyczą systemu dozoru elektronicznego (*sde*), centrów usług wspólnych (*cuw*) oraz systemu sądownictwa (*sąd*).

4. WNIOSKI

Analiza miała na celu identyfikację zmian w słownictwie polskim, które nastąpiły w ciągu ostatniej dekady – czyli od publikacji jedynego dostępnego korpusu ogólnej polszczyzny, Narodowego Korpusu Języka Polskiego. Wskazała ona na kilka głównych obszarów, w których zmiany te nastąpiły, a także, pośrednio, na istotne wydarzenia kreujące rzeczywistość językową, którą odzwierciedlają (przynajmniej częściowo) badane korpusy. Przede wszystkim wyraźnie zauważalna jest rosnąca rola nowych technologii: możemy ją zaobserwować na podstawie nie tylko wąsko rozumianych terminów technologicznych (które stanowią jedną z najliczniejszych grup zidentyfikowanych słów kluczowych), ale też wyrazów dotyczących ekonomii czy kultury i społeczeństwa. W obu tych grupach dominują słowa związane z nowymi technologiami finansowymi i rozrywkowymi (w tym związanymi z mediami społecznościowymi). Pozostałe słowa kluczowe dotyczą dyskursu medycznego (przede wszystkim pandemii COVID-19), a także (w mniejszym stopniu) transportu (głównie zrównoważonego), polityki, prawa i administracji oraz napaści Rosji na Ukrainę w lutym 2022 roku. Zdecydowana większość z nich funkcjonuje w języku polskim od niedawna, a jednak jest już dobrze przyswojona, o czym świadczą, przede wszystkim, ich częstości (a w wypadku wielu zapożyczeń z języka angielskiego również spolszczona pisownia i odmiana zgodna z zasadami polskiej gramatyki).

Wyniki te nie będą może zaskakujące dla uważnego obserwatora procesów diachronicznych zachodzących w języku polskim, ale wykorzystanie dużych korpusów liczonych w setkach milionów segmentów pozwoliło na zdobycie cennej wiedzy (którą trudno byłoby uzyskać za pomocą badań jakościowych) na temat skali i wagi owych zmian. Uzyskane przez nas wyniki powinny być zatem traktowane raczej jako ogólny obraz najistotniejszych trendów leksykalnych (sygnalizowanych przez ich najbardziej wyraziste wykładniki) i, być może również, jako wstęp do dalszych prac; bardziej pogłębione badania (na przykład w konkretnych kategoriach tematycznych) ujawniłyby znacznie więcej nowej leksyki i pozwoliłyby na wyciągnięcie bardziej szczegółowych i zniuansowanych wniosków.

Zastosowanie elektronicznych korpusów w językoznawstwie bywa porównywane do przełomu, jaki w astronomii przyniosło wynalezienie teleskopu (Stubbs 2004: 107), a dalszy rozwój językoznawstwa korpusowego jest w znacznej mierze zależny od stałego aktualizowania i ulepszania używanych w nim narzędzi, metod

i zasobów. Dlatego też nasza ostatnia uwaga dotyczy Narodowego Korpusu Języka Polskiego, wykorzystanego w tym badaniu jako korpus referencyjny. W świetle otrzymanych wyników NKJP wyraźnie jawi się bowiem jako narzędzie przestarzałe, niereprezentatywne dla współczesnej polszczyzny, a zatem do wielu celów niewystarczające. Słowa, które w okresie kompilowania go były neologizmami czy nieprzyswojonymi zapożyczeniami lub w ogóle nie istniały, liczą setki, a nawet tysiące wystąpień w nowszych korpusach badanych, a zjawiska, które opisują, stanowią dziś niekwestionowaną codzienność. Dlatego też ewidentna wydaje się konieczność uaktualnienia Narodowego Korpusu Języka Polskiego, tak by oddawał stan dzisiejszej polszczyzny, z myślą zarówno o badaczkach i badaczach języka i kultury, jak i o tych, którzy na co dzień korzystają z zasobów korpusowych do uczenia się, pisania i redagowania tekstów.

Bibliografia

Baker, P., McEnery, T. 2019. *The value of revisiting and extending previous studies: the case of Islam in the UK press.* W: *Quantifying approaches to discourse for social scientists*, red. R. Scholz, s. 215–249. Basingstoke: Palgrave Macmillan.

Cierpich-Kozieł, A. 2022. O patocelebrytach i patointeligencji, czyli o nowych złożeniach z (produktywnym) członem pato-. *Język Polski* 102(4), s. 5–20.

Clarke, I., Brookes, G., McEnery, T. 2022. Keywords through time. Tracking changes in press discourses of Islam. *International Journal of Corpus Linguistics* 27(4), s. 399–427.

Csomay, E., Young, R. 2020. Language use in pop culture over three decades: a diachronic keyword analysis of *Star Trek* dialogues. *International Journal of Corpus Linguistics* 26(1), s. 71–94.

Davies, M. 2008–. *The Corpus of Contemporary American English (COCA)*. Online: https://www.english-corpora.org/coca/ [dostęp: 21.11.2022].

Egbert, J., Biber, D. 2019. Incorporating text dispersion into keyword analyses. *Corpora* 14(1), s. 77–104.

Kilgarriff, A. 2012. *Getting to know your corpus*. W: *International Conference on Text, Speech and Dialogue*, s. 3–15. Berlin, Heidelberg: Springer.

Krasowska, D. 2018. Ekspresywne nazwy osób w języku licealistów i studentów białostockich. *Białostockie Archiwum Językowe* 18, s. 99–114.

Křen, M. 2015. *Recent developments in the Czech National Corpus*. W: *Proceedings of the 3rd workshop on challenges in the management of large corpora (CMLC-3),* red. P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, A. Witt, s. 1–4. Mannheim: Institut für Deutsche Sprache.

Lee, L.H., Braud, T., Zhou, P., Wang, L., Xu, D., Lin, Z., Kumar, A., Bermejo, C., Hui, P. 2021. All one needs to know about metaverse: a complete survey on technological singularity, virtual ecosystem, and research agenda. *Journal of LaTeX Class Files* 14(8), s. 1–66.

Mańczak-Wohlfeld, E. 2006. *Angielsko-polskie kontakty językowe.* Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Mańczak-Wohlfeld, E. red. 2010. *Słownik zapożyczeń angielskich w polszczyźnie.* Warszawa: Wydawnictwo Naukowe PWN.

McEnery, T., Baker, H. 2016. *Corpus linguistics and 17th-century prostitution: computational linguistics and history.* Londyn: Bloomsbury Academic.

McEnery, T., Xiao, R., Tono, Y. 2006. *Corpus-based language studies: an advanced resource book*. Londyn, Nowy Jork: Routledge.

Open Research 2022. *Badanie świadomości ekologicznej Polaków / segmentacja.*

Partington, A. 2010. Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: An overview of the project. *Corpora* 5(2), s. 83–108.

Paryzek, P. 2011. *Pozyskiwanie danych leksykalnych z tekstów elektronicznych (na materiale czasopisma naukowego)*. Poznań: UAM. Praca doktorska.

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

Scott, M. 1997. PC analysis of key words – and key key words. *System* 25(2), s. 233–245.

Scott, M. 2010. *Problems in investigating keyness, or clearing the undergrowth and marking out trails…* W: *Keyness in Texts*, red. M. Bondi, M. Scott, s. 1–12. Londyn: "Continuum".

Seetharaman, A., Saravanan, A.S., Patwa, N., Mehta, J. 2017. Impact of Bitcoin as a world currency. *Accounting and Finance Research* 6(2), s. 230–246.

Siuciak, M. 2015. Ciągłość i zmienność jako problem badań diachronicznych. *LingVaria* X 2(20), s. 149–159.

Stubbs, M. 2004. *Language corpora.* W: *Handbook of applied linguistics*, red. A. Davies, E. Catherine, s. 106–132. Oxford: Blackwell.

Waszakowa, K. 2005. *Przejawy internacjonalizacji w słowotwórstwie współczesnej polszczyzny.* Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Wawrzyńczyk, J. 2011. *Słownictwo nowopolskie. Redatacje*. Warszawa: „BEL Studio".

Zemlanaja, N. 2016. *Fighter, hejter i headliner, czyli o wyrażeniach pochodzenia angielskiego we współczesnej polszczyźnie.* W: *Globalizacja a przemiany języków słowiańskich,* red. H. Kurek, M. Święcicka, M. Peplińska, s. 354–368. Bydgoszcz: Wydawnictwo Uniwersytetu Kazimierza Wielkiego.

Evolution of Polish Lexis in the Decade Following the Introduction of the National Corpus of Polish: A Keyword Analysis

Summary

The paper reports on the results of an examination of changes in Polish lexis over the past decade. Two different, multi-million corpora spanning the years 2011–2022 were contrasted with a subset of the balanced National Corpus of Polish, which covers the period until 2010. To this end, keyword analysis was employed, and words that are particularly characteristic of the more recent set of texts, compared to the older corpus, were automatically extracted. This allowed us to identify the most salient lexical trends which differentiate the language of the last decade from the one recorded in the National Corpus of Polish, and which point to significant extralinguistic socio-cultural, economic, and political shifts across time.

Keywords: corpus linguistics – National Corpus of Polish – language change – diachronic research – keyword analysis.

Adj. Monika Czarnecka

ANEKS

Tabela 6. Słowa kluczowe w korpusie prasowym

Słowo	Iloraz szans	Częstość w korpusie badanym	Częstość w korpusie referencyjnym
covid-19	57740.26317914066	7006	0
500+	22746.18829821177	2760	0
Brexitu	17924.933720999456	2175	0
Brexit	14694.296619760018	1783	0
sars-cov-2	9683.53721212123	1175	0
Smartfona	7103.996778585464	862	0
elektromobilności	6972.135500343772	846	0
Dronów	6889.72221199129	836	0
Lte	6864.998227067833	833	0
Brexicie	4936.529653046223	599	0
Frankowiczów	4887.081799314553	593	0
Hejtu	4252.501269049178	516	0
Lockdownu	4137.123042506892	502	0
Hejt	3980.538331923385	483	0
Lockdown	3972.297032177783	482	0
Brexitem	3782.74716042425	459	0
e-myta	3733.2993748522845	453	0
Covid	3411.8888398369713	414	0
Kryptowalut	3321.2346086852676	403	0
Smartfonie	3280.0281432256425	398	0
prewspółczynnika	3214.0978027096444	390	0
Selfie	3172.8913425242717	385	0
Smartfonach	2909.170045374143	353	0
Pkpir	2810.274580364413	341	0

cd. tabeli 6.

Słowo	Iloraz szans	Częstość w korpusie badanym	Częstość w korpusie referencyjnym
Kryptowaluty	2769.068140058906	336	0
Fanpage	2587.759826813929	314	0
Payment	2563.0359689597217	311	0
Elektromobilność	2447.657975296013	297	0
e-recepty	2406.4515528418015	292	0
Frankowicze	2406.4515528418015	292	0
Bitcoin	2258.10844880308	274	0
e-deklaracji	2233.38460068594	271	0
Koronawirusie	2183.9369066424974	265	0
Prewspółczynnik	2175.6956245859215	264	0
Drone	2159.213060716196	262	0
Jzp	2126.2479339504516	258	0
Sde	2118.006652461871	257	0
#metoo	2093.282808482982	254	0
Koronawirusem	2029.8470856405359	2463	1
Lajków	2002.62872014167	243	0
cyberbezpieczeństwo	1986.1461596799159	241	0
Fintech	1961.4223195958511	238	0
Smartfonem	1887.250803725333	229	0
Mikroinstalacji	1887.250803725333	229	0
Blockchain	1879.0095245898754	228	0
Bitcoina	1821.32057291365	221	0
sąd	1788.3554594552197	217	0
Ziobrystów	1763.6316252133895	214	0
Ziobryści	1681.2188496815272	204	0
lgbt+	1664.73629554886	202	0
Cuw	1590.5648059683929	193	0

cd. tabeli 6.

Słowo	Iloraz szans	Częstość w korpusie badanym	Częstość w korpusie referencyjnym
Dron	1582.3235297540516	192	0
Bitcoinów	1557.5997015978803	189	0
Fotowoltaiki	1483.428221511039	180	0
Hejtem	1475.1869463515452	179	0
Milenialsów	1466.945671273193	178	0
Streaming	1376.2916507666982	167	0
Miesięcznice	1293.878913371621	157	0
Timeshare	1293.878913371621	157	0
Hejterów	1269.1550937353684	154	0

Tabela 7. Słowa kluczowe w korpusie webowym

Słowo	Iloraz szans	Częstość w korpusie badanym	Częstość w korpusie referencyjnym
Covid	430022.67302622425	221046	0
covid-19	82443.84320989503	42387	0
Smartfona	44481.82851600509	22870	0
Smartfonach	24561.012957256473	12628	0
Tiktoku	18774.68342939891	9653	0
Kryptowalut	18560.73581791014	9543	0
Selfie	17018.36952917112	8750	0
Nft	16954.185364627854	8717	0
Amoled	15658.833195806617	8051	0
Podcaście	15520.740205774082	7980	0
Smartfonie	14363.48287566693	7385	0
Lockdownu	14104.802027182037	7252	0
Instastories	13809.166862046452	7100	0
Drone	13760.542666475789	7075	0
Smartfonem	13715.808408851435	7052	0
Lockdown	13287.915621242517	6832	0
Omikronem	12916.427046499472	6641	0
Tiktok	12696.645919416433	6528	0
Koronawirusem	12452.73801914351	64022	1
Tiktoka	11722.218731534167	6027	0
Kryptowaluty	11045.371719721263	5679	0
Esg	10740.012744557771	5522	0
Smartfonów	10216.946779195729	52528	1
Flagowców	9808.376760917497	5043	0
Omikronu	9573.03671592468	4922	0
Hejtu	9466.063988372938	4867	0
Influencerka	9028.448416038556	4642	0

cd. tabeli 7.

Słowo	Iloraz szans	Częstość w korpusie badanym	Częstość w korpusie referencyjnym
Hejt	8886.46652012567	4569	0
Lte	8586.943141323118	4415	0
Esports	8540.264182100502	4391	0
Deeskalacji	8499.420094750005	4370	0
Smartwatch	8211.566578878732	4222	0
Alerty	8127.933479896125	4179	0
Instastory	7980.116858680696	4103	0
Emisyjności	7760.3369269058385	3990	0
Hevc	7587.235956158235	3901	0
Influencerów	7495.823097285999	3854	0
Fotowoltaiki	6924.006699643204	3560	0
Oktagonie	6762.575604971863	3477	0
elektromobilności	6706.1719762528	3448	0
Lockdownów	6612.814253661277	3400	0
Bitcoin	6408.594268970196	3295	0
Bitcoina	6375.530085294605	3278	0
Smartwatche	6365.805325619247	3273	0
Render	6216.044039776493	3196	0
Omikrona	5862.062916924489	3014	0
Blockchain	5766.760330513525	2965	0
Startupów	5727.861318527688	2945	0
Smartfonu	5494.467281622152	2825	0
Smr	5424.449082253661	2789	0
Youtuber	5420.55918244712	2787	0
Wyszczepienia	5051.018776846378	2597	0
antyszczepionkowców	5039.3490822780705	2591	0
Ransomware	5017.9546426258585	2580	0

cd. tabeli 7.

Słowo	Iloraz szans	Częstość w korpusie badanym	Częstość w korpusie referencyjnym
Metawersum	4967.3859690889785	2554	0
Cyberataki	4885.698117788699	2512	0
Microsd	4848.744092234761	2493	0
Thunderbolt	4837.074400267154	2487	0
Lockdowny	4778.725942679726	2457	0
Influencer	4763.16635462325	2449	0