FOLIA 397

Annales Universitatis Paedagogicae Cracoviensis

Studia Linguistica 19 (2024)

ISSN 2083-1765

DOI 10.24917/20831765.19.18

Michał Szczyszek
Uniwersytet Adama Mickiewicza, Poznań, Polska
ORCID 0000-0002-0253-7296

Czy można przewidzieć *WOJNĘ* i *KRYZYS* na podstawie danych językowych? Wstępna propozycja procedury lingwometrycznej

Słowa kluczowe: cyfrowa analiza języka, korpus językowy, frekwencja leksykalna, metoda lingwometryczna

Keywords: digital language analysis, language corpus, lexical frequency, linguometric method

Wstęp

W artykule zastanawiam się nad frekwencją leksykalną jako znacznikiem nadchodzących zmian w rzeczywistości pozajęzykowej. W tym celu zbadałem występujące w polskich korpusach językowych wyrazy związane z dwoma przykładowymi polami tematycznymi (zob. Batko-Tokarz 2019): KRYZYS oraz WOJNA. Następnie skupiłem się na korelacji między zmianami we frekwencji wyrazów tekstowych z tych dwóch pól w korpusach a zaistniałym w przeszłości wydarzeniami kryzysowo-wojennymi. Punktami historycznymi, które poddałem szczególnej obserwacji, są: globalny kryzys ekonomiczny lat 2008–2009, wybuch II wojny światowej w 1939 r., agresja Rosji na Ukrainę w 2022 r. Przyjrzałem się zwłaszcza danym frekwencyjno-leksykalnym pochodzącym z okresów poprzedzających te wydarzenia historyczne.

Na podstawie przeprowadzonych obserwacji i dokonanych analiz statycznych – wzrostu/spadku frekwencji wyrazów tekstowych z tych dwóch pól tematycznych – podjąłem próbę wypracowania metody/procedury oszacowywania/prognozowania na bazie danych frekwencyjno-leksykalnych zwartych w korpusach językowych tego, czy jest możliwe przewidzenie zmian w rzeczywistości pozajęzykowej zachodzących w nadchodzącej/przyszłej perspektywie czasowej; innymi słowy: sprawdziłem, czy w języku, w danych leksykalno-frekwencyjnych związanych z realizacją języka w tekstach, można dostrzec nadchodzącą zmianę pozajęzykową, czyli: czy język (w tym i frekwencja leksykalna) zapowiada taką zmianę.

Wobec tego celem, jaki sobie tu stawiam, jest próba wypracowania metody analizy danych frekwencyjno-leksykalnych, dzięki której będzie można oszacować nadchodzącą zmianę w rzeczywistości pozajęzykowej (kolejny kryzys, kolejna wojna). Przyświecają mi tu – jako swoisty wzór i punkt odniesienia – narzędzia ekonometryczne (por. np. Bartosiewicz 1978; Krzysztofiak 1984) służące ekonomistom do prognozowania rozwoju gospodarczego w skali świata czy danego regionu/państwa. Chodzi mi o wypracowanie **narzędzia lingwometrycznego**. Podstawą lingwistyczną – swoistym aksjomatem językoznawczym (aksjomat 1) – są dla mnie teoria kognitywistyczna (Lakoff, Johnson 1980) i jej polska wersja: językowy obraz świata (Bartmiński 2006), zakładające wszak, że w języku odbija się wizja świata danej społeczności. Na potrzeby artykułu doprecyzowuję/zawężam tę teorię, przyjmując aksjomatycznie (aksjomat 2), że zmiany we frekwencji, w częstości występowania struktur leksykalnych są wskaźnikami/markerami nadchodzących zmian w rzeczywistości pozajęzykowej; opisuję zatem nie tyle językowy obraz świata, ile tekstowy obraz świata (wyłaniający się z analizowanych tekstów, a więc jedynie z użyć języka). Opieram się na danych pochodzących z języka polskiego (z korpusów polszczyzny) – i traktuję je jako przykład, a także na polskiej wspólnocie komunikatywnej (dalej: PWK; w ujęciu L. Zabrockiego (Zabrocki 1963), z doprecyzowaniem S. Borawskiego (Borawski 2005)). Założeniem tego tekstu jest również przekonanie, że wspólnota komunikatywna (PWK) jako wspólnota trwała i funkcjonująca na określonym terytorium w sensie geograficznym, historycznym, geopolitycznym oraz mentalno-kognitywnym nie tylko wytwarza swoją wizję świata, lecz także „odnotowuje podkorowo" zmiany zachodzące w świecie, które odbijają się w języku, a dokładniej – we frekwencji leksykalnej i w tekstach.

Materiał leksykalny wyekscerpowałem z NKJP (http://nkjp.pl/, zob. Przepiórkowski, Bańko, Górski, Lewandowska-Tomaszczyk 2012; korpus ten notuje polskie wyrazy z okresu od ok. 1950 r. do ok. 2010 r., zawiera ok. 3 mld segmentów), z Monco (http://monco.frazeo.pl/, zob. Pęzik 2020; korpus ten notuje polskie wyrazy z okresu od ok. 2010 r. do dziś i jest nieustannie pomnażany i uaktualniany o nowe jednostki funkcjonujące w polskojęzycznym internecie, zawiera ok. 7 mld jednostek), a także z narzędzia korpusowego Odkrywka (nieudostępnionego publicznie przez jej autora, Filipa Gralińskiego, który opisuje je w swojej książce, zob. Graliński 2019; korpus ten notuje polskie wyrazy z okresu od ok. 1800 r. do dziś, zawiera ok. 23 mld jednostek leksykalnych).

Będę starał się osiągnąć założony cel, udowadniając przydatność języka, lingwistyki, korpusologii, narzędzi frekwencyjno-leksykalnych do oszacowywania i prognozowania nadchodzących zmian w rzeczywistości pozajęzykowej. Od razu trzeba dodać, że ten cel jest (potencjalnie) możliwy do osiągnięcia tylko i wyłącznie dzięki rozwojowi korpusologii (por. Lewandowska-Tomaszczyk 2009). Bez nieustannie rozwijanych językowych korpusów ogólnych (m.in. polszczyzny ogólnej, np. NKJP) czy specjalistycznych (np. Korpusu Dyskursu Parlamentarnego (KDP) – http://sejm.nlp.ipipan.waw.pl/query_corpus/, jak również Polskiego Korpusu Sejmowego – http://clip.ipipan.waw.pl/PSC) prognozowanie zmian pozajęzykowych na podstawie danych językowych (frekwencyjnych) nie byłoby i nie będzie możliwe.

W artykule przedstawię wybrane wykresy frekwencji chronologicznej kilku przykładowych leksemów (tj. ich słowoform) wygenerowanych z ekscerpowanych korpusów językowych, a także tabele z danymi frekwencyjnymi (z tych samych korpusów) wybranych leksemów należących do tych dwóch pól tematycznych. Wyniki poddam analizie statystycznej i dyskusji, całość zakończę wnioskami, w których przedstawię moją ocenę na temat możliwości wypracowania narzędzia lingwometrycznego oraz potencjału badawczo-poznawczego zaproponowanej metody/procedury prognozowania zmian w rzeczywistości pozajęzykowej na podstawie danych językowych.

Podobne prace, ale przeprowadzane wyłącznie na materiale literackim, czyli w odniesieniu do zjawisk literackich, podejmuje – niezależnie od mojej lingwistycznej propozycji – zespół pod kierunkiem Jürgena Wertheimera i Moniki Wolting w ramach projektu „Cassandra" (https://www.ifg.uni.wroc.pl/projekt-cassandra/).

Pytania badawcze

Sformułowany powyżej cel przekuwam na potrzeby artykułu w cztery operatywne pytania, na które będę starał się odpowiedzieć:

- Czy w korpusach widać wzrost frekwencji wyrazów tekstowych (realizacji leksemów) należących do pól tematycznych KRYZYS i WOJNA, a więc wyrazów *kryzys* i *wojna* oraz ich synonimów?
- Czy to oznacza, że można na poziomie języka przewidzieć kryzys, wojnę – analogicznie do narzędzi i danych ekonomicznych, politologicznych, socjologicznych itp.?
- Czy analizy języka (języków) mogą być komplementarne/wyprzedzające wobec analiz ekonomicznych, politologicznych, socjologicznych i innych?
- Czy można wypracować metodologię tego rodzaju badań?

Ekscerpcja materiału z korpusów

Aby przeanalizować ogromny materiał zgromadzony w korpusach, należy wykorzystać narzędzia cyfrowe wraz z zastosowaniem określonego (informatycznego) sposobu zapytań. Korzystałem zatem z konkondarserów tych korpusów i/lub ich wyszukiwarek, stosując zapytanie typu: [base="(*kryzys*)"]; bazowałem przy tym na lematach, a nie na formach tekstowych. W ten sposób uzyskiwałem dane frekwencyjne danego lematu i na tej podstawie mogłem wygenerować wykresy frekwencji chronologicznej.

Analizy materiału

W pierwszej kolejności analizie poddałem dane frekwencyjno-leksykalne związane z pojęciem KRYZYS, a w drugiej kolejności – z pojęciem WOJNA. Dobór synonimów do analiz został zaczerpnięty ze *Słownika synonimów* (https://www.synonimy.pl/).

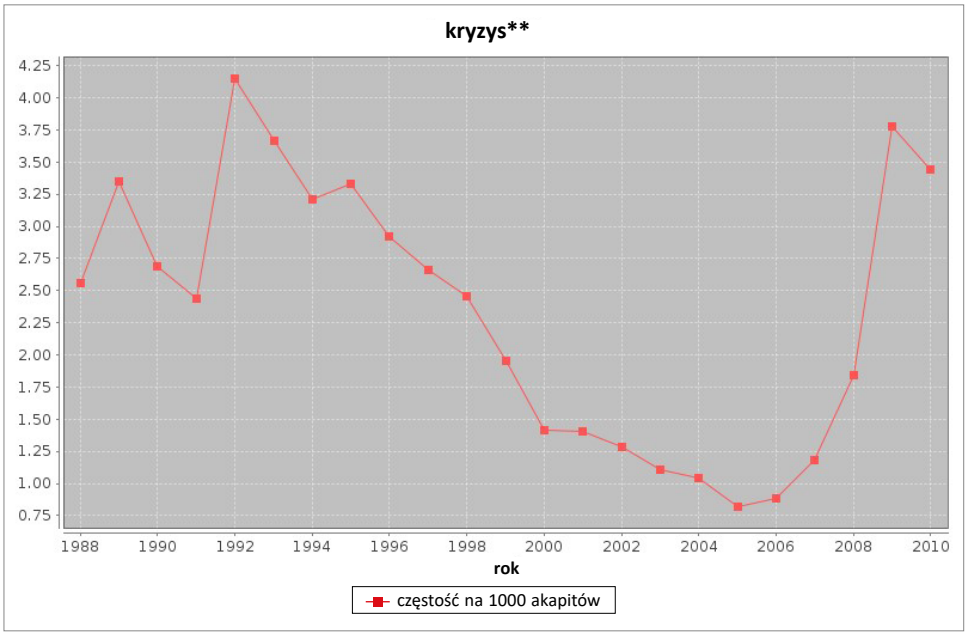
Poniżej przedstawię wyniki obserwacji materiału korpusowego, na podstawie których będę pokazywał zależności frekwencyjno-leksykalne zachodzące przed danym wydarzeniem historycznym. Materiał korpusowy to realizacje wyrazów centralnych obu pól tematycznych oraz synonimów wyrazów centralnych.

Analiza materiału leksykalnego związanego z globalnym *KRYZYSEM* lat 2008–2009

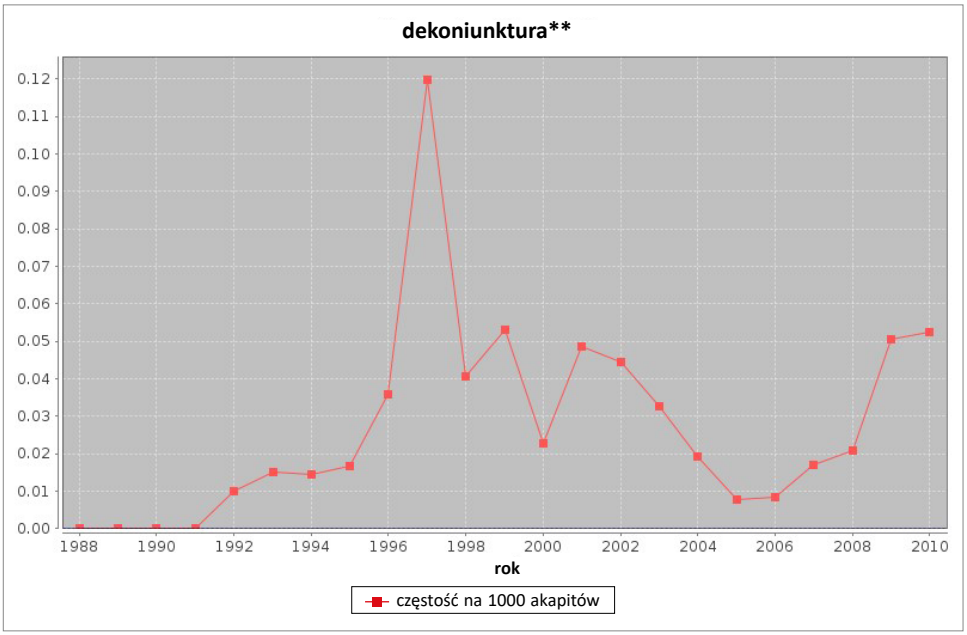
Obserwacje i analizy oparłem na grupach synonimów związanych z centralnym leksemem pola KRYZYS. Szeregi synonimiczne przestawiają się następująco:

- Synonimy słowa *KRYZYS* poddawane ekscerpcji korpusowej i analizie frekwencyjnej, np.: 1. przesilenie, załamanie, komplikacja, konflikt, impas, pat, klincz, trudności, przełom, dołek, moment zwrotny, martwy punkt, sytuacja bez wyjścia; 2. regres, regres gospodarczy, dekoniunktura, recesja, zastój, stagnacja, bessa, zapaść, załamanie, zła koniunktura.
- Synonimy słowa *KŁOPOT*: problem, przeszkoda, trudność, troska, zmartwienie, przykrość, bolączka, nieprzyjemność, szkopuł, zagwozdka, ambaras, ból, pierepałka, pasztet, posp. kanał, posp. poruta, perturbacja.
- Synonimy słowa *PŁACZ*: szloch, szlochanie, lament, łkanie (brak w NKJP), spazmy, kwilenie (brak w NKJP), chlipanie (brak w NKJP), zawodzenie (brak w NKJP), łzy, ryk, bek, wycie (brak w NKJP).

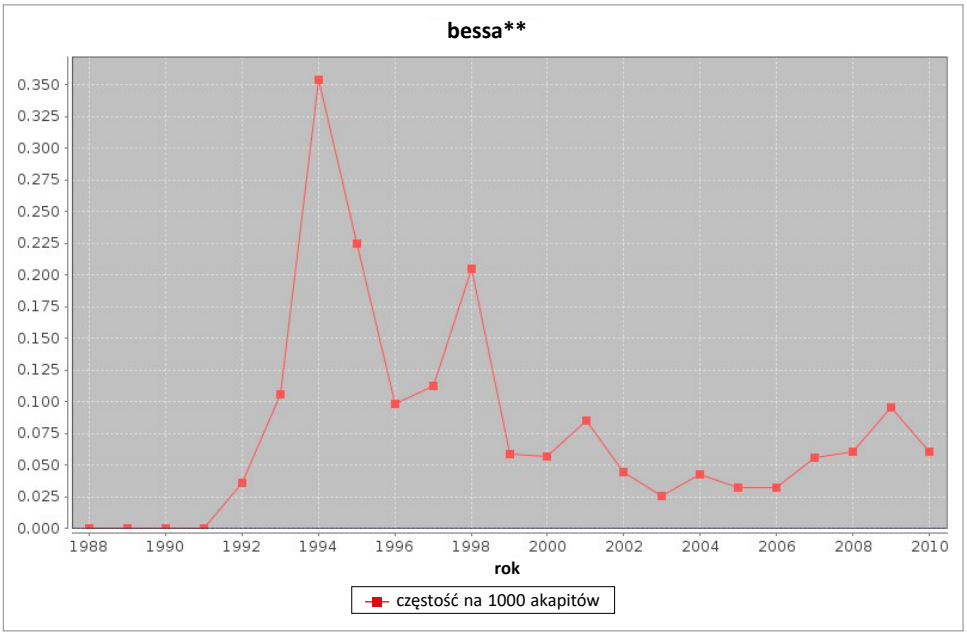
Poniżej prezentuję kilka przykładowych wykresów frekwencyjnych dla wybranych słowoform. Wykresy zostały wygenerowane w NKJP. Na wykresach 1–4 widoczna jest wyraźna regularność w zakresie frekwencji.



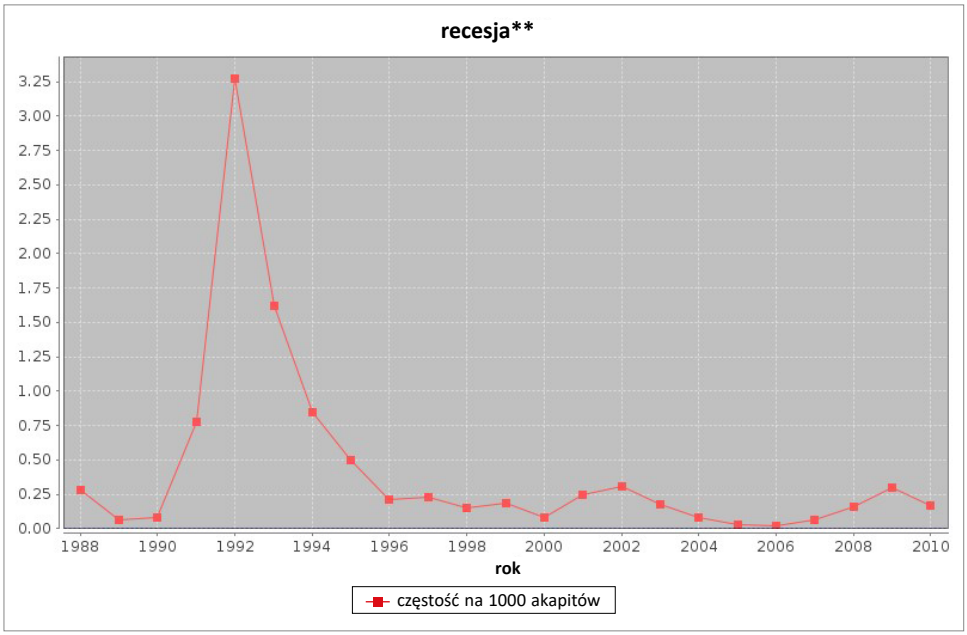
Wykres 1. Frekwencja wyrazu *kryzys* w NKJP w latach 1988–2010.
Źródło: NKJP.



Wykres 2. Frekwencja wyrazu *dekoniunktura* w NKJP w latach 1988–2010.
Źródło: NKJP.



Wykres 3. Frekwencja wyrazu *bessa* w NKJP w latach 1988–2010.
Źródło: NKJP.



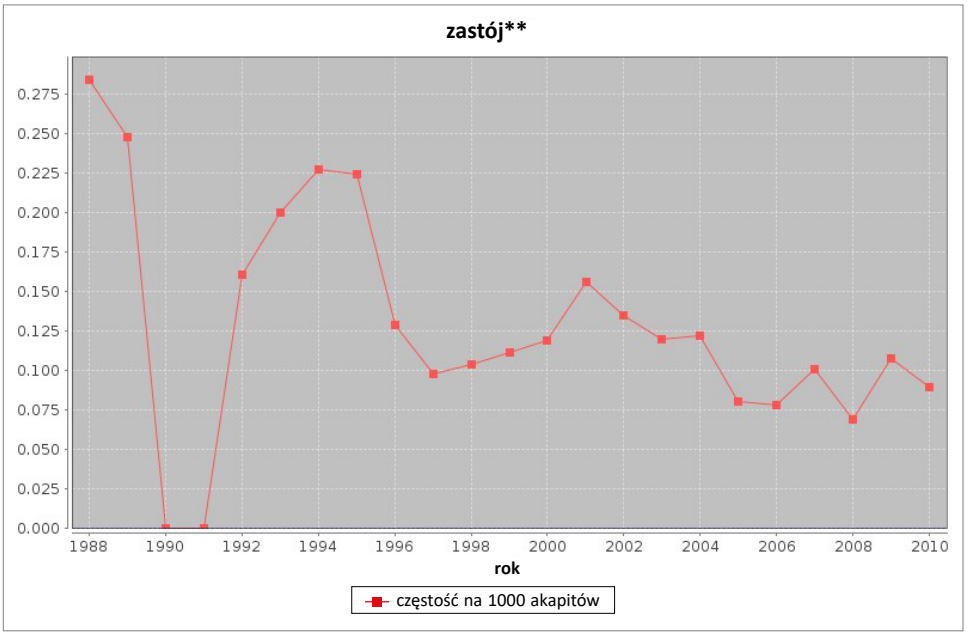
Wykres 4. Frekwencja wyrazu *recesja* w NKJP w latach 1988–2010.
Źródło: NKJP.

Na przykładowych wykresach z NKJP (a także na innych wykresach, których ze względów objętościowych nie zamieściłem w artykule) widać wyraźnie, że w okresie poprzedzającym kryzys, a więc w latach 2005/2006–2008, frekwencja realizacji tekstowych tych leksemów rośnie, a w trakcie kryzysu lub tuż po jego zakończeniu spada. Taki obraz jest prawie regularny i powtarzalny, jeśli chodzi o terminy ekonomiczne (wiązane powszechnie z ekonomią, uznawane za terminy ekonomiczne). Istnieje zaledwie kilka wyjątków od tej reguły – zob. wykresy 5 i 6, na których nieco inaczej rozkłada się przebieg frekwencji.

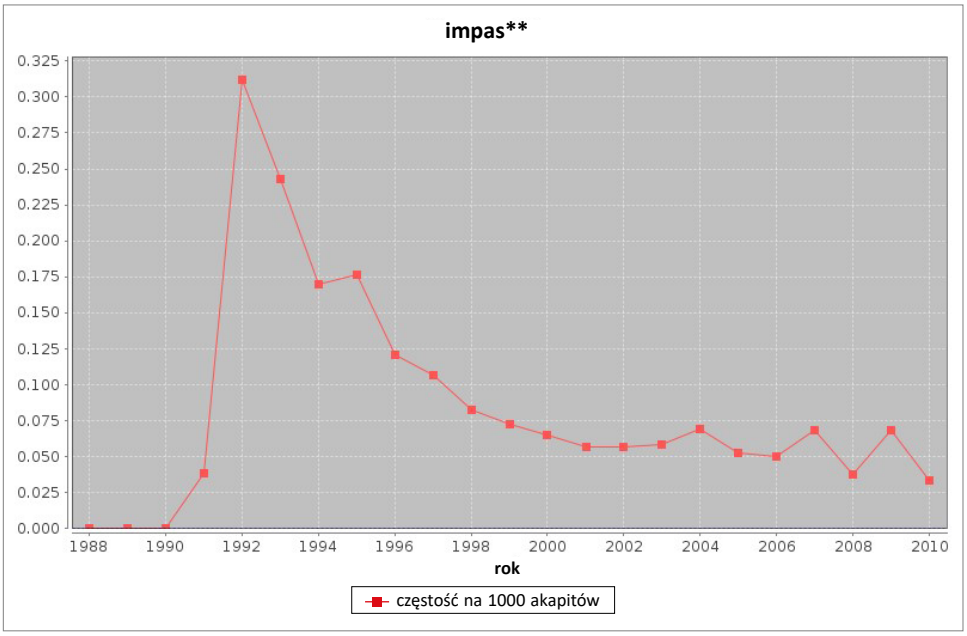
Na wykresach 5 i 6 widać, że frekwencja wpierw rośnie, następnie rok przed kryzysem spada, po czym zaczyna fluktuować: raz rośnie, raz spada.

Z kolei w grupie wyrazów niebędących terminami ekonomicznymi, czyli leksemami nieekonomicznymi, ale należącymi do pola KRYZYS i będącymi (bliższymi i dalszymi) synonimami centralnego wyrazu tego pola, wykresy chronologiczne frekwencji przedstawiają się różnie i nie są tak jednoznaczne. Wzrost frekwencji albo zaczyna się później niż w wypadku terminów ekonomicznych, albo początkowy wzrost nagle się stabilizuje lub nawet – niekiedy gwałtownie – spada.

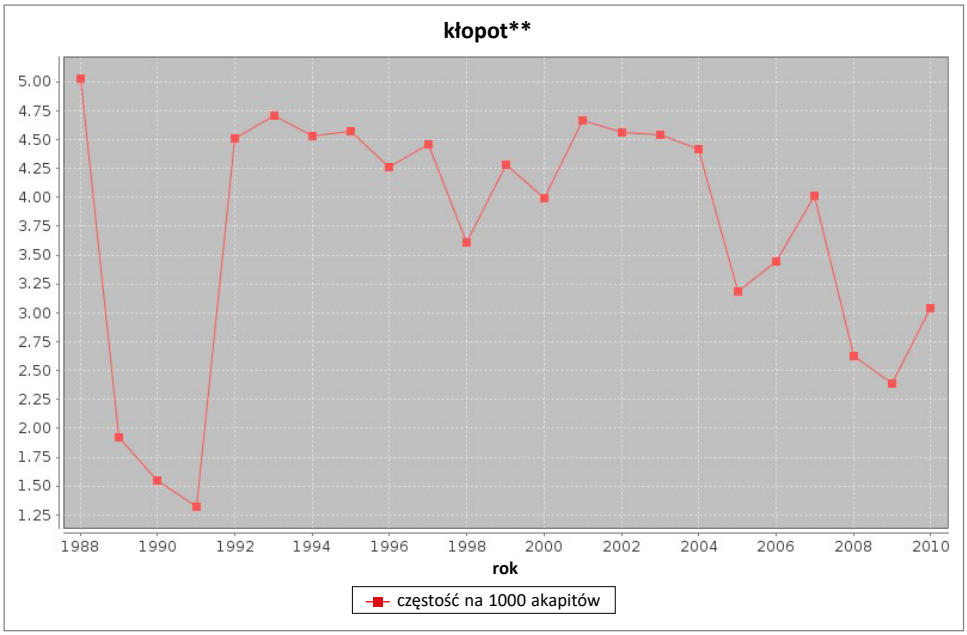
Dla porównania podaję tu dane z narzędzia korpusowego Odkrywka. Dane te – zestawione w tabelach – jeszcze wyraźniej prezentują zauważone na powyższych wykresach zależności. Tabele 1–3 przedstawiają zbiorczo wyniki dla szeregów synonimicznych z podziałem na terminy ekonomiczne oraz leksemy nieekonomiczne. Dane chronologicznie dotyczą okresu tuż przed kryzysem ekonomicznym, czyli z lat 2005–2008.



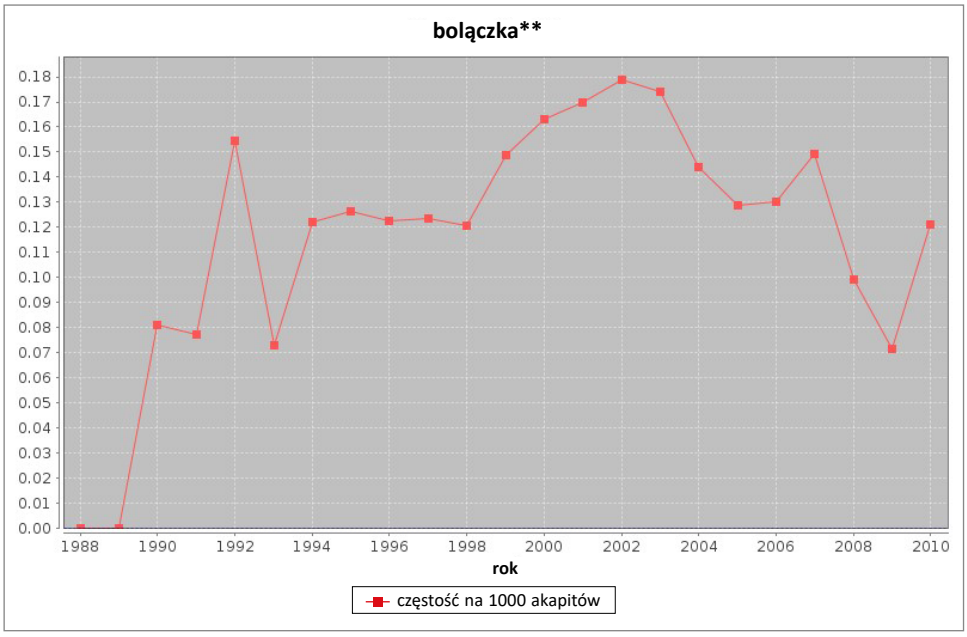
Wykres 5. Frekwencja wyrazu *zastój* w NKJP w latach 1988–2010.
Źródło: NKJP.



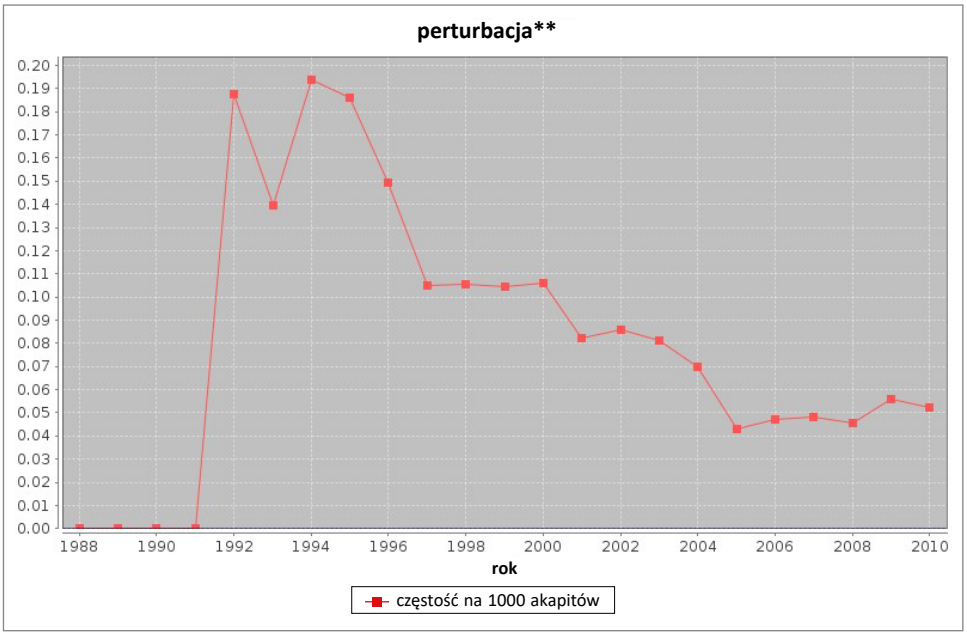
Wykres 6. Frekwencja wyrazu *impas* w NKJP w latach 1988–2010.
Źródło: NKJP.



Wykres 7. Frekwencja wyrazu *kłopot* w NKJP w latach 1988–2010.
Źródło: NKJP.



Wykres 8. Frekwencja wyrazu *bolączka* w NKJP w latach 1988–2010.
Źródło: NKJP.



Wykres 9. Frekwencja wyrazu *perturbacja* w NKJP w latach 1988–2010.
Źródło: NKJP.

Parametr korpusowy wygładzanie (np. wygładzanie na poziomie 30) jest operacją matematyczną oferowaną przez Odkrywkę, statystycznym wyrównaniem frekwencji chronologicznej. Jest to narzędzie statystyczno-korpusowe automatycznie wyliczające średnią frekwencję danego wyrazu dla danego roku, obejmującą podaną liczbę lat, np. wygładzanie na poziomie 30 podaje średnią frekwencją dla roku 2005, liczoną z 30 lat (15 lat przed i 15 lat po roku 2005). Na potrzeby przeprowadzonych analiz wyrazów z pola KRYZYS parametr wygładzanie ustawiłem na poziomie 30 oraz na poziomie 1 (czyli rok do roku).

Tabela 1. Terminy ekonomiczne Odkrywka (lata 1970–2021); częstość na 1 mln wyrazów, wygładzanie na poziomie 30.

KRYZYS i synonimy				
Wyraz	**2005**	**2006**	**2007**	**2008**
bessa	7,22	7,36	7,52	7,68
dekoniunktura	5,32	5,43	5,55	5,66
dołek	19,92	20,31	20,73	21,16
impas	17,24	17,52	17,79	18,09
klincz	6,73	6,87	7,02	7,14
komplikacja	58,04	59,17	60,35	61,59
konflikt	657	671	685	699
kryzys	605	617	633	644

martwy punkt	89,16	90,81	92,53	94,38
moment zwrotny	27,77	28,29	28,86	29,46
pat	32,97	35,5	34	34,55
przełom	863	882	901	920
przesilenie	20,44	20,87	21,3	21,76
recesja	14,63	14,85	15,08	15,33
regres	41,53	42,37	43,24	44,16
regres gospodarczy	4,58	4,66	4,73	4,81
stagnacja	174	177	181	185
sytuacja bez wyjścia	190	193	197	201
trudność	1129	1152	1175	1199
załamanie	179	182	186	190
zapaść	297	304	310	317
zastój	271	276	282	288
zła koniunktura	4,97	5,02	5,08	5,15

Źródło: Odkrywka.

Tabela 2. Leksemy nieekonomiczne, Odkrywka (lata 1970–2021); częstość na 1 mln wyrazów, wygładzanie na poziomie 30.

KŁOPOT i synonimy				
Wyraz	**2005**	**2006**	**2007**	**2008**
bolączka	51,42	52,44	53,51	54,62
ból	1008	1029	1051	1074
kanał	747	762	778	795
kłopot	534	543	554	564
nieprzyjemność	13,73	14,01	14,3	14,61
pasztet	15,2	15,5	15,83	16,16
perturbacja	35,28	36	36,74	37,53
pierepałka	0,0098	0,01	0,0093	0,0095
poruta	0,0741	0,0744	0,0746	0,0745
problem	6208	6338	6473	6615
przeszkoda	500	510	521	532
przykrość	61,04	62,25	63,51	64,86
szkopuł	12,02	12,26	12,5	12,76
troska	611	624	636	649
trudność	1129	1152	1175	1199
zagwozdka	3,29	3,36	3,44	3,52
zmartwienie	78,13	79,66	81,27	82,96

Źródło: Odkrywka.

Tabela 3. Leksemy nieekonomiczne, Odkrywka (lata 1970–2021); częstość na 1 mln wyrazów, wygładzanie na poziomie 30.

PŁACZ i synonimy				
Wyraz	**2005**	**2006**	**2007**	**2008**
chlipanie	0,247	0,249	0,249	0,252
kwilenie	5,34	5,44	5,54	5,66
lament	24,54	25,02	25,51	26,04
łkanie	15	15,26	15,53	15,82
łza	342	349	357	365
płacz	248	253	259	264
ryk	28,24	28,7	29,18	29,66
spazm	0,479	0,475	0,473	0,469
szloch	8,5	8,67	8,85	9,03
szlochanie	2,35	2,39	2,42	2,46
wycie	68,74	69,76	70,77	71,79
zawodzenie	73,89	75,37	76,92	78,54

Źródło: Odkrywka.

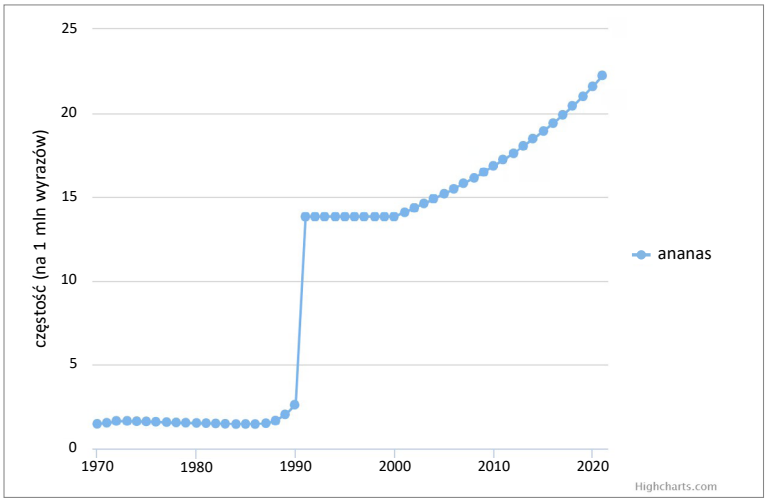
We wszystkich powyższych tabelach widać wyraźny wzrost częstotliwości analizowanych wyrazów w badanym okresie. Dla porównania przyjrzałem się także innym wyrazom, które – z perspektywy badanych pól tematycznych – nazwałem *wyrazami placebo*, gdyż nie należą one do badanych pól tematycznych i są ogólnopolskimi leksemami, np. niespecjalistycznymi. Do *wyrazów placebo* na potrzeby tego badania zaliczyłem słowa: *dom*, *człowiek*, *ananas*, *chomik*.

Tabela 4. *Wyrazy placebo*, Odkrywka (lata 1970–2021); częstość na 1 mln wyrazów, wygładzanie na poziomie 30.

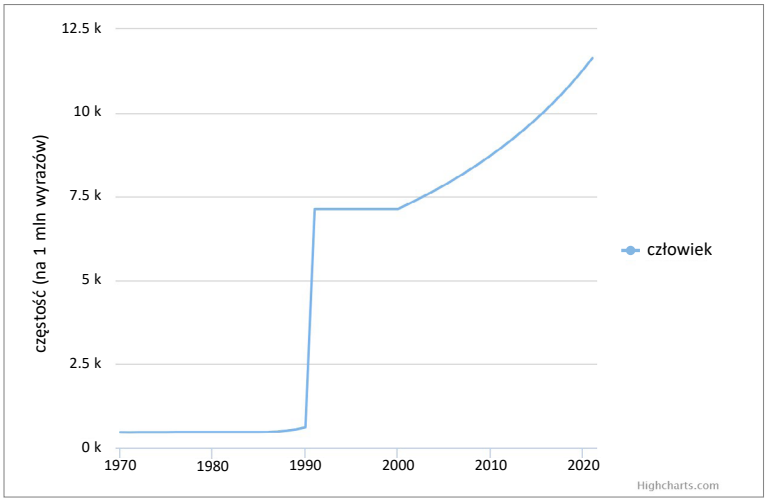
Wyrazy placebo				
Wyraz	**2005**	**2006**	**2007**	**2008**
dom	5119	5244	5334	5450
człowiek	7840	8001	8170	8347
ananas	15,18	15,49	15,8	16,14
chomik	6,29	6,42	6,54	6,68

Źródło: Odkrywka.

Dane zawarte w tabeli 4 pokazują, że częstotliwość *wyrazów placebo* w badanym czteroleciu rośnie. Potwierdzają to poniższe wykresy przestawiające chronologię frekwencji wyrazów *ananas* i *człowiek*.



Wykres 10. Frekwencja wyrazu *ananas* w Odkrywce (lata 1970–2021).
Źródło: Odkrywka.



Wykres 11. Frekwencja wyrazu *człowiek* w Odkrywce (lata 1970–2021).
Źródło: Odkrywka.

W tym miejscu warto pokazać wzrost frekwencji wyrażony procentowo – działanie to ma na celu uniezależnienie obserwacji i statystyk dla różnych grup wyrazów (tj. *wyrazów placebo* oraz wyrazów nieekonomicznych i terminów ekonomicznych) od zmiennych liczbowych wyrażanych w liczbach rzeczywistych, a więc o odmiennych wartościach rzeczywistych dla każdej z grup wyrazowych. Procentowy wzrost częstości wyrazów z poszczególnych grup (na podstawie danych z Odkrywki przy wygładzaniu na poziomie 30) przedstawiają tabele 5–7.

Tabela 5. *Wyrazy placebo* – wzrost częstotliwości wyrażony procentowo, wygładzanie na poziomie 30.

***Wyrazy placebo* – wzrost częstotliwości [%]**	
dom	6,46
człowiek	6,46
ananas	6,32
chomik	6,20
średnia	**6,36**

Źródło: Odkrywka.

Tabela 6. Terminy ekonomiczne – wzrost częstotliwości wyrażony procentowo, wygładzanie na poziomie 30.

Terminy ekonomiczne – wzrost częstotliwości [%]	
KRYZYS i synonimy (wybrane przykłady)	
dekoniunktura	6,39
kryzys	6,44
recesja	4,78
stagnacja	6,32
średnia	**5,98**

Źródło: Odkrywka.

Tabela 7. Leksemy nieekonomiczne – wzrost częstotliwości wyrażony procentowo, wygładzanie na poziomie 30.

Leksemy nieekonomiczne – wzrost częstotliwości [%]	
KŁOPOT i PŁACZ i synonimy (wybrane przykłady)	
kanał	6,42
kłopot	5,61
problem	6,55
trudność	6,20
lament	6,11
łza	6,72
płacz	6,45
średnia	**6,29**

Źródło: Odkrywka.

Widać wyraźnie, że w grupie badanych wyrazów (terminy ekonomiczne oraz leksemy nieekonomiczne) średni wzrost częstotliwości występowania jest niższy niż w grupie *wyrazów placebo*. Większą różnicę miedzy średnim wzrostem częstości użycia widać w parze *wyrazy placebo* – terminy ekonomiczne, mniejsza różnica jest zaś zauważalna w parze *wyrazy placebo* – leksemy nieekonomiczne. Niemniej zaznacza

się tu wyraźna i istotna różnica statystyczna między *wyrazami placebo* a badanymi wyrazami. W celu porównania i weryfikacji tego wyniku podobną operację (przeliczenie danych liczbowych na procenty) przeprowadziłem na danych z Odkrywki dotyczących obu grup wyrazów (*wyrazy placebo* oraz wyrazy badane), przy czym parametr wygładzanie ustawiłem na poziomie 1 (tabele 8–10).

Tabela 8. *Wyrazy placebo* – wzrost częstotliwości wyrażony procentowo, wygładzanie na poziomie 1.

***Wyrazy placebo* – wzrost częstotliwości [%]**	
dom	–14,44
człowiek	–2,23
ananas	–15,44
chomik	17,80
średnia	**–3,58**

Źródło: Odkrywka.

Warto zauważyć, że przy wygładzaniu na poziomie 1 w grupie *wyrazów placebo* widoczne jest duże rozchwianie: od dużych spadków (nawet do ujemnych wartości) po duże wzrosty. Średni wzrost jest jednakże ujemny, a zatem w obserwowanym czteroleciu *wyrazy placebo* odnotowały – w ujęciu uśrednionym – spadek częstotliwości występowania w korpusie.

Tabela 9. Terminy ekonomiczne – wzrost częstotliwości wyrażony procentowo, wygładzanie na poziomie 1.

Terminy ekonomiczne – wzrost częstotliwości [%]	
KRYZYS i synonimy (wybrane przykłady)	
dekoniunktura	47,38
kryzys	53,76
recesja	106,34
stagnacja	–10,62
średnia	**49,22**

Źródło: Odkrywka.

Tabela 10. Leksemy nieekonomiczne – wzrost częstotliwości wyrażony procentowo, wygładzanie na poziomie 1.

Leksemy nieekonomiczne – wzrost częstotliwości [%]	
KŁOPOT i PŁACZ i synonimy (wybrane przykłady)	
kanał	21,24
kłopot	–26,48
problem	7,33

trudność	3,97
lament	–12,50
łza	–18,87
płacz	–17,74
średnia	**–6,11**

Źródło: Odkrywka.

Widać więc, że przy wygładzaniu na poziomie 1 różnice są bardziej dynamiczne, większe jest też rozchwianie wyników liczbowych. *Wyrazy placebo* – w ujęciu uśrednionym – zanotowały spadek częstotliwości; podobnie leksemy nieekonomiczne, jednak tu średni spadek jest jeszcze większy. Z kolei w przypadku terminów ekonomicznych trzeba odnotować bardzo duży uśredniony wzrost częstotliwości występowania w danych korpusowych

Podsumowując tę część analiz i rozważań, należy stwierdzić, że przy wygładzaniu na poziomie 30 uśredniony wzrost procentowy badanych wyrazów w stosunku do *wyrazów placebo* jest niewielki, natomiast przy wygładzaniu na poziomie 1 widoczny jest ogromny wzrost wyrazów badanych, zwłaszcza terminów ekonomicznych, w stosunku do *wyrazów placebo*. Z kolei leksemy nieekonomiczne zanotowały większy spadek niż *wyrazy placebo*.

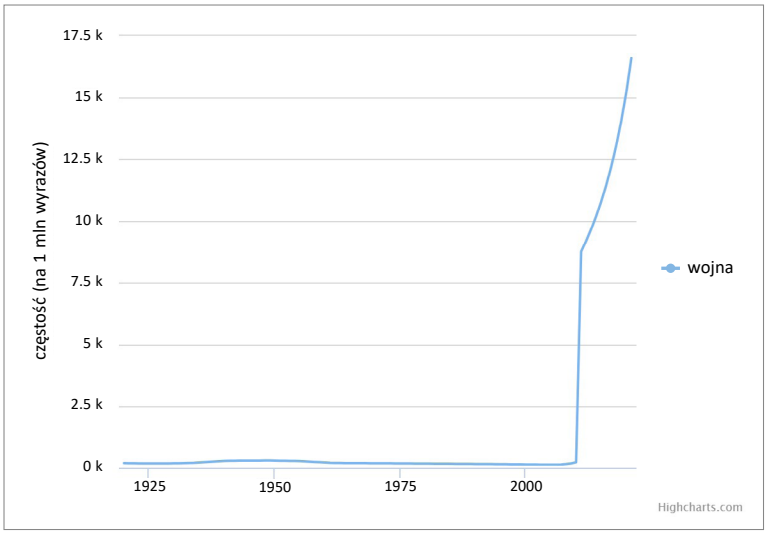
Zauważalna jest więc coraz wyraźniej rysująca się istotna statystycznie różnica. Na tym etapie należałoby postulować wykorzystywanie danych z dwóch poziomów wygładzania – na poziomie 1 i na poziomie 30. Komplementarne ich wykorzystanie daje bowiem obraz różnic statystycznych między grupą wyrazów badanych a grupą *wyrazów placebo*. Dopiero na tym tle widać, że w badanym czteroleciu przed wielkim globalnym kryzysem ekonomicznym lat 2008–2009 następował powolny, ledwo zauważalny gołym okiem, choć istotny statystycznie wzrost frekwencji i częstotliwości użycia badanych wyrazów: terminów ekonomicznych (przede wszystkim) i leksemów nieekonomicznych (obserwowany średni wzrost przy wygładzaniu na poziomie 30 i wyraźniejszy spadek przy wygładzaniu na poziomie 1) w stosunku do *wyrazów placebo*.

Analiza materiału leksykalnego związanego z WOJNĄ

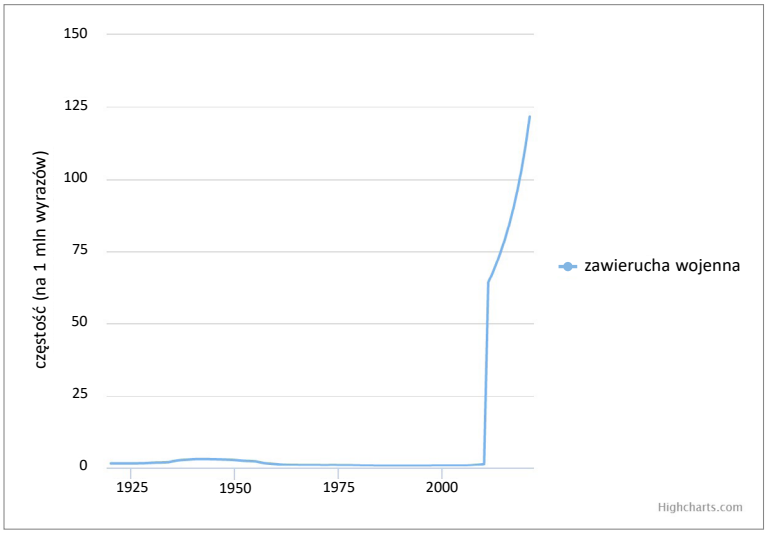
Podobną procedurę zastosowałem w odniesieniu do danych związanych z wybuchem II wojny światowej w 1939 r. oraz z agresją Rosji na Ukrainę w 2022 r. Obserwacje i analizy oparłem na grupie synonimów związanych z centralnym leksemem pola WOJNA. Szereg synonimiczny słowa *KRYZYS* poddawany ekscerpcji korpusowej i analizie frekwencyjnej przestawia się zatem następująco: 1. wojenka, kampania, batalia, bitwa, bój, blitzkrieg, konflikt, konflikt zbrojny, awantura, zawierucha wojenna, konfrontacja, konfrontacja zbrojna, operacja specjalna, działania wojenne, szczęk oręża; 2. przen. walka, kłótnia, zatarg, spór, scysja, starcie, zwada, utarczka, szarpanina, zmagania, polemika, kontrowersja, rywalizacja, nieporozumienie, antagonizm, waśń.

II wojna światowa – 1939 r.

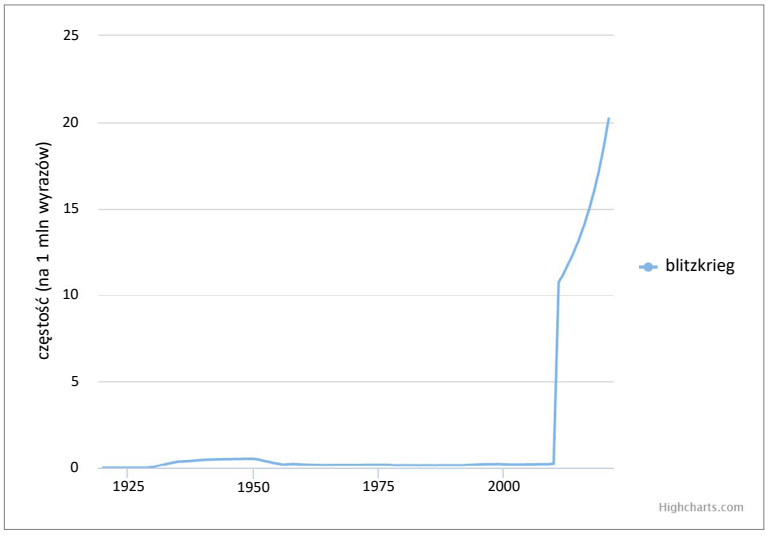
Poniżej przedstawiam wykresy frekwencji w ujęciu chronologicznym wyrazu *wojna* i wybranych synonimów obejmujące lata 30. XX w.; wykresy wygenerowałem za pomocą narzędzia Odkrywka.



Wykres 12. Frekwencja wyrazu *wojna* w Odkrywce (lata 1920–2021).
Źródło: Odkrywka.

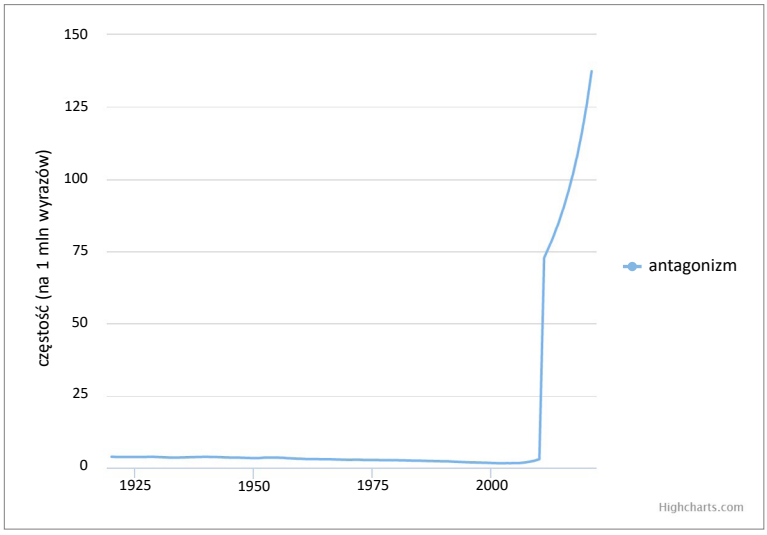


Wykres 13. Frekwencja bigramu *zawierucha wojenna* w Odkrywce (lata 1920–2021).
Źródło: Odkrywka.

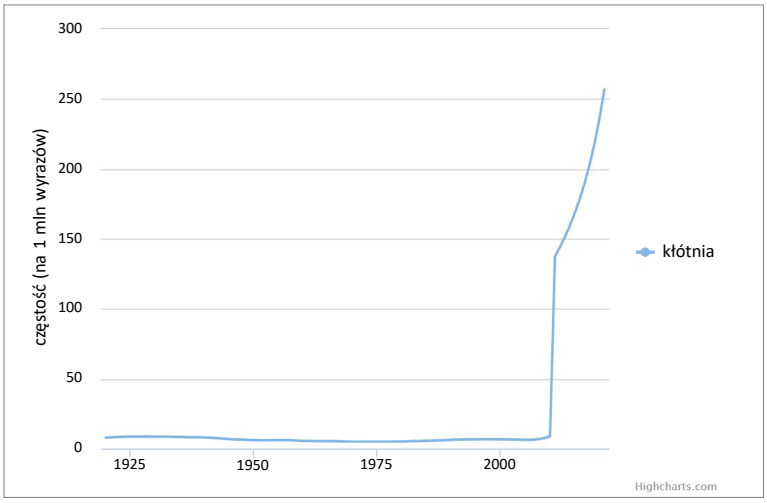


Wykres 14. Frekwencja wyrazu *blitzkrieg* w Odkrywce (lata 1920–2021).
Źródło: Odkrywka.

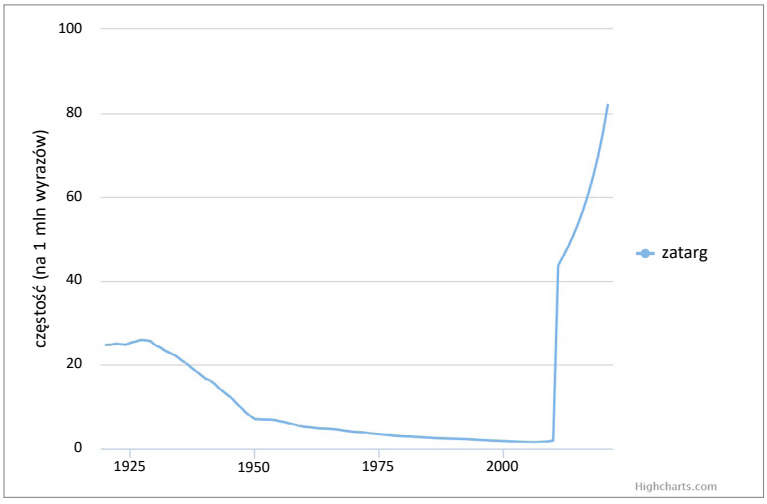
Poniżej zaś znajdują się wykresy frekwencyjne synonimów wyrazu *wojna*, które nie mają charakteru terminologicznego.



Wykres 15. Frekwencja wyrazu *antagonizm* w Odkrywce (lata 1920–2021).
Źródło: Odkrywka.



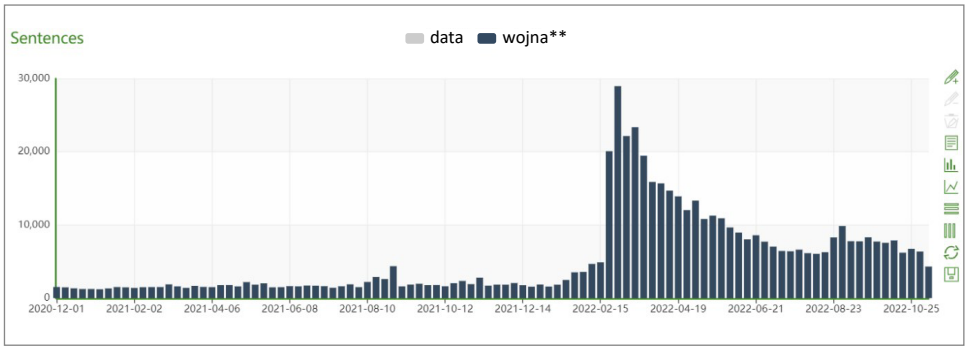
Wykres 16. Frekwencja wyrazu *kłótnia* w Odkrywce (lata 1920–2021).
Źródło: Odkrywka.



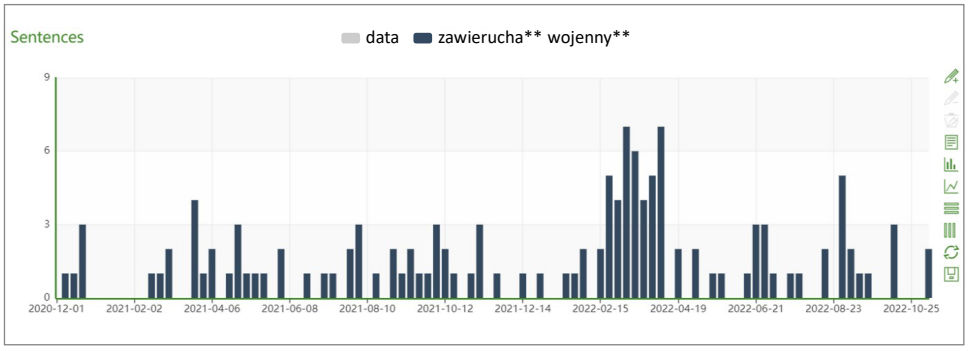
Wykres 17. Frekwencja wyrazu *zatarg* w Odkrywce (latach 1920–2021).
Źródło Odkrywka.

Agresja Rosji na Ukrainę – 2022 r.

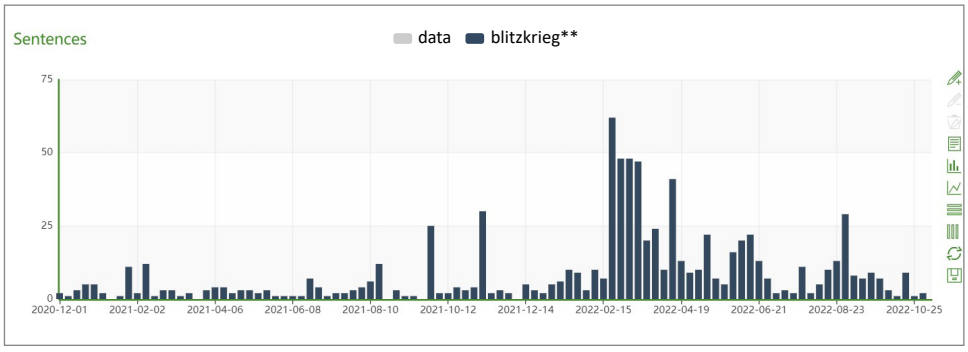
Dane dotyczące tej wojny pochodzą z korpusu Monco (NKJP nie obejmuje już tego okresu, Odkrywka co prawda go obejmuje, jednak warto pokazać dane z innego korpusu). Wygenerowane wykresy przedstawiają w ujęciu chronologicznym frekwencję wyrazu *wojna* i jego wybranych synonimów.



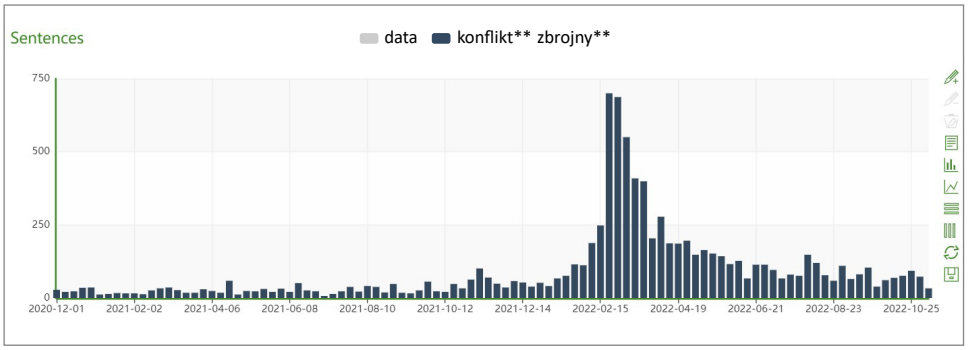
Wykres 18. Frekwencja wyrazu *wojna* w Monco (lata 2020–2022).
Źródło: Monco.



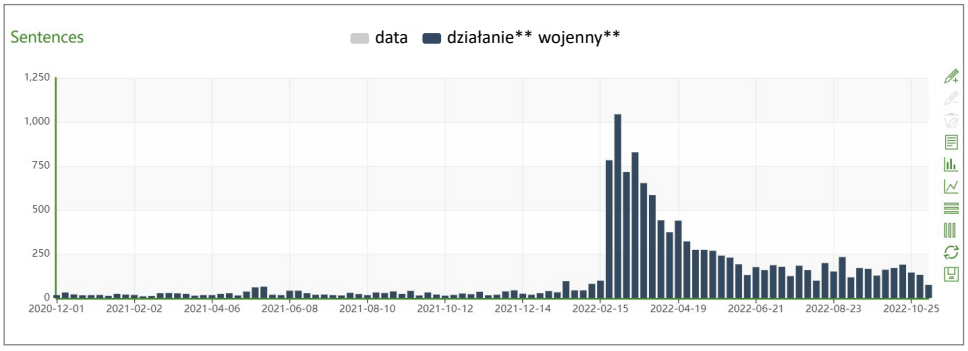
Wykres 19. Frekwencja bigramu *zawierucha wojenna* w Monco (latach 2020–2022).
Źródło: Monco.



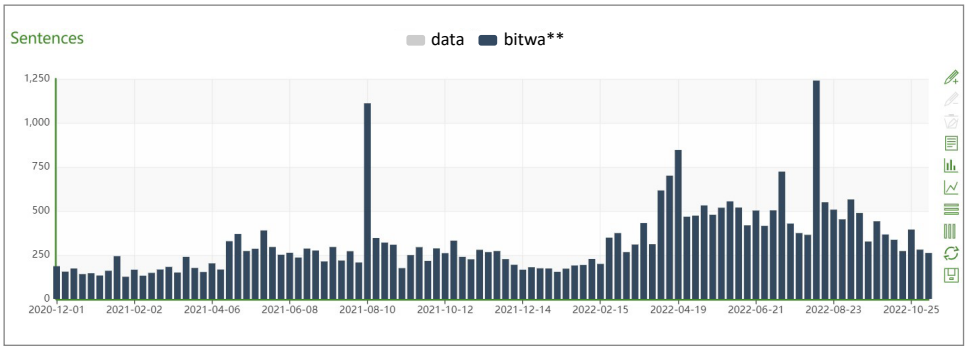
Wykres 20. Frekwencja wyrazu *blitzkrieg* w Monco (lata 2020–2022).
Źródło: Monco.



Wykres 21. Frekwencja bigramu *konflikt zbrojny* w Monco (lata 2020–2022).
Źródło: Monco.

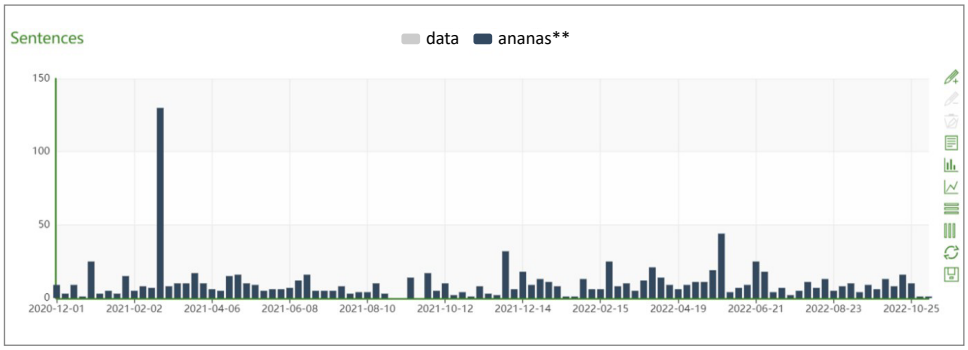


Wykres 22. Frekwencja bigramu *działania wojenne* w Monco (lata 2020–2022).
Źródło: Monco.

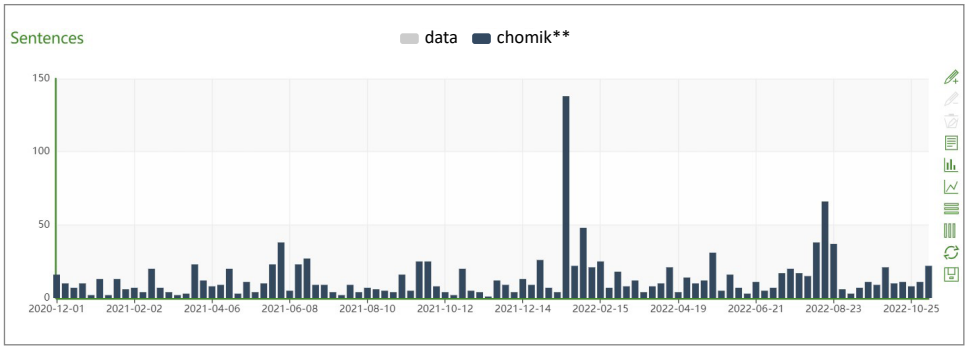


Wykres 23. Frekwencja wyrazu *bitwa* w Monco (lata 2020–2022).
Źródło: Monco.

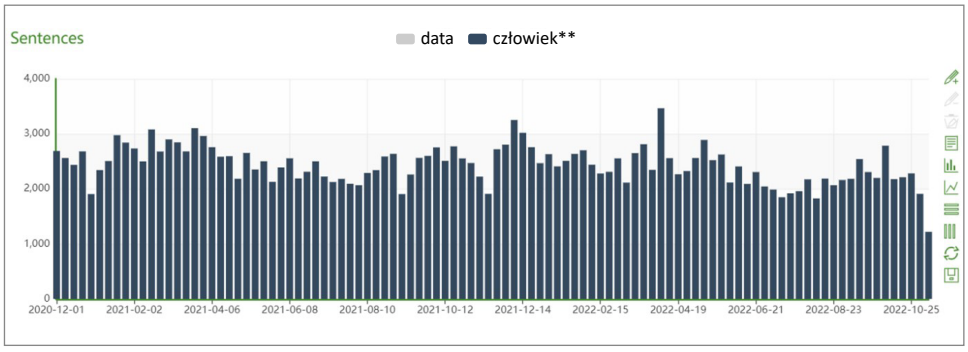
Dla porównania poniżej przedstawiam wykresy frekwencyjne *wyrazów placebo* w korpusie Monco.



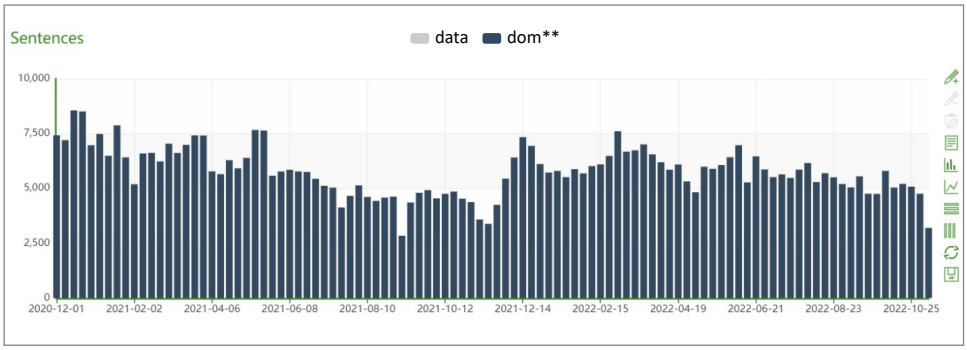
Wykres 24. Frekwencja wyrazu *ananas* w Monco (lata 2020–2022).
Źródło: Monco.



Wykres 25. Frekwencja wyrazu *chomik* w Monco (lata 2020–2022).
Źródło: Monco.



Wykres 26. Frekwencja wyrazu *człowiek* w Monco (lata 2020–2022).
Źródło: Monco.



Wykres 27. Frekwencja wyrazu *dom* w Monco (lata 2020–2022).
Źródło: Monco.

Z kolei w zaprezentowanych poniżej tabelach przedstawiam zbiorcze wyniki pochodzące ponownie z Odkrywki – dotyczą one synonimów wyrazu *wojna*. Dla analiz porównawczych w zakresie metody parametr wygładzanie ustawiłem na poziomie 10 oraz na poziomie 1. Podobny poziom wygładzania ustawiłem dla *wyrazów placebo* (por. tabele 11–12).

Tabela 11. WOJNA i synonimy, Odkrywka (lata 1920–2021); częstość na 1 mln wyrazów, wygładzanie na poziomie 10.

WOJNA i synonimy								
Wyraz	**1936**	**1937**	**1938**	**1939**	**2018**	**2019**	**2020**	**2021**
antagonizm	3,65	3,68	3,73	3,77	108	116	126	137
awantura	16,95	16,71	16,98	17,24	198	213	230	250
batalia	1,78	1,92	2,1	2,31	267	287	311	339
bitwa	46,53	47,19	48,91	50,62	2938	3162	3424	3733
blitzkrieg	0,38	0,39	0,42	0,44	15,91	17,12	18,54	20,2
bój	71,89	72,5	75,08	77,58	2816	3028	3276	3568
działania wojenne	55,84	60,46	65,12	69,45	1649	1775	1921	2094
kampania	29,62	30,92	32,97	35,5	2019	2269	2455	2675
kłótnia!	8,04	7,92	7,96	7,94	202	218	235	256
konflikt	27,14	27,55	27,87	28,56	2094	2252	2436	2654
konflikt zbrojny	6,15	6,31	6,44	6,66	223	240	260	283
konfrontacja	1,552	1,637	1,705	1,971	577	620	671	731
konfrontacja zbrojna	0,1214	0,1371	0,1411	0,1564	5,94	6,37	6,87	7,47
walka	232	239	252	266	9733	10473	11336	12356
wojenka!	0,67	0,67	0,66	0,64	22,11	23,76	25,69	27,97

wojna	239	250	262	275	13 063	14 057	15 215	16 586
zatarg!	20,45	19,47	18,55	17,62	64,71	69,59	75,27	82
zawierucha wojenna	2,56	2,71	2,8	2,9	95	103	111	121
zmagania	13,63	14,05	14,73	15,21	1646	1770	1915	2087
zwada!	1,0169	1,0058	0,9888	0,9809	11,02	11,84	12,79	13,92

Źródło: Odkrywka.

Tabela 12. WOJNA i synonimy – dane w ujęciu procentowym, Odkrywka (lata 1920–2021).

WOJNA i synonimy [%]				
	wygładzanie na poziomie 10		**wygładzanie na poziomie 1**	
Wyraz	**1936–1939**	**2018–2021**	**1936–1939**	**2018–2021**
bitwa	8,79	27,05	28,25	37 588
blitzkrieg	15,78	26,96	35 000	45 858
działania wojenne	24,73	26,98	98,78	16 368
konflikt zbrojny	8,29	26,90	48,97	9 431
walka	14,65	26,94	0,94	16 289
wojna	15,06	26,96	43,78	18 677
zawierucha wojenna	13,28	27,36	54,09	33 000
średnia	**14,37**	**23,17**	**5039,26**	**25 315,86**

Źródło: Odkrywka.

W dwóch poniższych tabelach (13–14) przedstawiam dane dotyczące *wyrazów placebo* – także w ujęciu procentowym.

Tabela 13. *Wyrazy placebo*, Odkrywka (lata 1920–2021); częstość na 1 mln wyrazów, wygładzanie na poziomie 10.

Wyrazy placebo										
Wyraz	**1936**	**1937**	**1938**	**1939**	**2018**	**2019**	**2020**	**2021**	**Wzrost w ciągu 4 lat [%]**	
									1936–1939	**2018–2021**
dom	320	321	327	331	16 459	17 702	19 152	20 866	3,43	26,77
człowiek	366	370	379	391	25 234	27 138	29 360	31 983	6,83	26,74
ananas	1,738	1,734	1,707	1,699	47,13	50,69	54,85	59,73	–2,24	26,73
chomik	0,543	0,554	0,558	0,568	18,29	19,58	21,05	22,73	4,60	24,27
średnia									**3,16**	**26,13**

Źródło: Odkrywka.

Tabela 14. *Wyrazy placebo* – dane w ujęciu procentowym, Odkrywka (lata 1920–2021).

***Wyrazy placebo* [%]**		
Wyraz	**1936–1939**	**2018–2021**
dom	–2,51	9653
człowiek	–5,66	9713
ananas	–6,69	2700
chomik	13,15	2465
średnia	**–0,285**	**6132,75**

Źródło: Odkrywka.

Na podstawie powyższych danych dotyczących pola tematycznego WOJNA widać wyraźnie istotną statystycznie różnicę między grupą wyrazów badanych (synonimów wyrazu *wojna*) a *wyrazami placebo*. Widać też, że ustawienie parametru wygładzanie na poziomie 10 i korelowanie danych z tego poziomu z danymi z wygładzania na poziomie 1 daje precyzyjniejsze wyniki niż analiza wygładzania na poziomie 30 w kontekście wygładzania na poziomie 1.

W odniesieniu do roku 1939 (wybuch II wojny światowej) w czteroleciu poprzedzającym to wydarzenie historyczne widać wyrazisty wzrost średniej częstotliwości wyrazów badanych w stosunku do wzrostu *wyrazów placebo* (wygładzanie na poziomie 10; przy wygładzaniu na poziomie 1 widać jeszcze większy średni wzrost częstotliwości wyrazów badanych w stosunku do wzrostu *wyrazów placebo*, który jest nawet ujemny). Natomiast w odniesieniu do agresji Rosji na Ukrainę w trzyleciu poprzedzającym to wydarzenie widać – przy wygładzaniu na poziomie 10 – mniejszy średni wzrost wyrazów badanych w stosunku do *wyrazów placebo*, z kolei przy wygładzaniu na poziomie 1 zauważalny jest drastycznie wyższy średni wzrost częstości badanych wyrazów w stosunku do *wyrazów placebo*. Są to dane bardzo istotne statystycznie, pokazujące ogromną różnicę w dynamice frekwencji między wyrazami badanymi a *wyrazami placebo*. Na tej podstawie można zatem prognozować nadchodzące wydarzenia – w tym wypadku wojenne.

Dodatkowo należy zauważyć, że im dawniej wydarzyła się dana sytuacja, tym dłuższy okres jest potrzebny do wychwycenia nadchodzącej zmiany: w wypadku wojny z 1939 r. były to aż 4 lata, w wypadku wojny w Ukrainie był to okres 3 lat (a ostatecznie nawet 2). Innymi słowy – dane leksykalno-frekwencyjne poprzedzające wybuch II wojny światowej pozwalały z czteroletnim wyprzedzeniem prognozować to wydarzenie. Współcześnie możemy liczyć na 2 lub maksymalnie 3 lata przygotowań do ewentualnej wojny (jest to zatem okres o 50–75% krótszy niż w latach 30. XX w.). Ponadto dane współczesne są bardziej dynamiczne i rozchwiane: przy wygładzaniu na poziomie 10 zauważalny jest wolniejszy wzrost częstości badanych wyrazów niż *wyrazów placebo*, a przy wygładzaniu na poziomie 1 pojawia się ogromna różnica we wzrostach częstotliwości między obiema grupami na korzyść wyrazów badanych (ich częstotliwość jest wykładnikowo wyższa). To wiąże się oczywiście z faktem, że dane wynikające z wygładzania na poziomie 1 obejmują pojedyncze lata, więc ujawniają

dynamikę wzrostu rok do roku. Z kolei przy wygładzaniu na poziomie 10 dane obejmują okresy dziesięcioletnie, przez co mogą nie uwzględniać danych z lat późniejszych niż moment badania. Niemniej jestem przekonany, że – tak jak w wypadku KRYZYSU – udowodniłem przydatność tej metody lingwometrycznej do prognozowania zmian w rzeczywistości pozajęzykowej i oszacowywania nadchodzących wydarzeń. Wydaje się też, że skorelowanie poziomu wygładzania na poziomie 10 z wygładzaniem na poziomie 1 daje precyzyjniejsze rezultaty prognostyczne niż użycie pary wygładzanie na poziomie 30 i wygładzanie na poziomie 1.

Podsumowanie, wnioski ogólne, końcowa dyskusja

Na podstawie powyższych analiz można sformułować wnioski dotyczące proponowanej procedury lingwometrycznej – jej operatywności i użyteczności, a także wnioski na temat opisywanych wydarzeń historycznych i ich odbicia w danych leksykalno-frekwencyjnych pochodzących z korpusów polszczyzny. Tym samym można zbliżyć się do odpowiedzi na pytania wyjściowe:

- Czy w korpusach widać wzrost frekwencji wyrazów tekstowych (realizacji leksemów) należących do pól tematycznych KRYZYS i WOJNA, a więc wyrazów *kryzys* i *wojna* oraz ich synonimów?
- Czy to oznacza, że można na poziomie języka przewidzieć kryzys, wojnę – analogicznie do narzędzi i danych ekonomicznych, politologicznych, socjologicznych itp.?
- Czy analizy języka (języków) mogą być komplementarne/wyprzedzające wobec analiz ekonomicznych, politologicznych, socjologicznych i innych?
- Czy można wypracować metodologię tego rodzaju badań?

Przypomnę też: celem artykułu była próba wypracowania metody analizy danych frekwencyjno-leksykalnych, tj. procedury lingwometrycznej, która umożliwi – na bazie danych językowych, tj. użyć języka, zwłaszcza frekwencji leksykalnej – oszacowanie, czy nadchodzi jakaś zmiana w rzeczywistości pozajęzykowej (kolejny kryzys, kolejna wojna).

Odpowiadając na pytanie pierwsze, należy stwierdzić: tak, w korpusach widać wzrost frekwencji wyrazów tekstowych należących do pól tematycznych KRYZYS i WOJNA, a więc wyrazów *kryzys* i *wojna* oraz ich synonimów. To bezsprzeczny fakt. Widać także zmianę częstotliwości (zazwyczaj jej wzrost) wyrazów badanych, co nie zostało ujęte w pytaniu z początku artykułu, lecz okazało się niezwykle istotnym znacznikiem dla proponowanej procedury lingwometrycznej.

W konsekwencji – odpowiadając na pytanie drugie – trzeba stwierdzić: tak, wydaje się, że na podstawie języka (jego użycia w tekstach) można przewidzieć nadchodzące kryzys i wojnę nawet z kilkuletnim wyprzedzeniem[1]. Można zatem prognozować

[1] Warto dodać, że propozycja metody, którą wstępnie opracowałem, jest być może nieczuła na bardzo losowe zmienne, takie jak np. zamach terrorystyczny, w wyniku którego może dojść do wybuchu wojny. Widzę tu dwa rozwiązania: 1. podniesienie czułości narzędzia, tak aby mogło ono wychwytywać randomowe zmiany w rzeczywistości pozajęzykowej,

zbliżającą się zmianę, bazując na danych korpusowych w zakresie frekwencji w funkcji czasu badanych wyrazów. Widać tu analogię między narzędziami lingwometrycznymi a narzędziami ekonometrycznymi, politologicznymi bądź socjologicznymi.

Odnosząc się do pytania trzeciego, należy stwierdzić: analizy lingwometryczne mogą być komplementarne wobec analiz ekonomicznych, politologicznych, socjologicznych i innych. Czy lingwometria wcześniej pokazuje symptomy nadchodzących zmian pozajęzykowych niż narzędzia ekonomiczne, politologiczne itp.? Na to pytanie nie udało się odpowiedzieć, ale uzyskany rezultat łatwo można porównać z danymi z innych dziedzin wiedzy. Wstępne analizy przedstawione powyżej pokazują, że lingwometria ujawnia symptomy nadchodzących zmian na 2–4 lata przed zmianą (przed danym wydarzeniem historycznym).

Wreszcie odpowiedź na ostanie pytanie, zapowiadające główny cel artykułu, powinna brzmieć: tak, w moim przekonaniu jest to możliwe. W odniesieniu do badanych fenomenów – WOJNY i KRYZYSU – wydaje się, że można na podstawie języka (a dokładniej: na podstawie danych leksykalno-frekwencyjnych związanych z użyciem języka w tekstach) przewidzieć nadchodzący kryzys i wojnę – z maksymalnie czteroletnim wyprzedzeniem. Dość precyzyjnie udało się przewidzieć nadchodzący kryzys ekonomiczny (z lat 2008–2009) na podstawie analiz terminów ekonomicznych (co jest wtórne względem analiz ekonomistów). Nieco mniej precyzyjne rezultaty uzyskałem na podstawie analizy leksemów nieekonomicznych. Dobrze udało się zaś przewidzieć wojnę, analizując synonimy tego leksemu.

Postulaty

Dalsze prace w zakresie analiz leksyki i jej frekwencji powinny zmierzać w kierunku ustalenia zakresów wzrostów procentowych wyrazów badanych względem neutralnych *wyrazów placebo*, tj. przedziałów procentowych, dzięki którym można by powiedzieć, że w języku odbijają się nadchodzące kryzys, wojna bądź inne wydarzenie pozajęzykowe. Ponadto wychwycenia wymagałoby ustalenie odległości przedziału procentowego między *wyrazami placebo* a wyrazami badanymi. Dopiero wtedy będzie można uznać wyniki za zadowalające i w pełni miarodajne, obecnie szacuję skuteczność opracowywanej metody lingwometrycznej na poziomie ok. 70–90%.

Aby podwyższyć skuteczność proponowanej metody lingwometrycznej, należałoby rozwijać korpusy językowe – im większe, tym lepiej – oraz narzędzia do automatycznej analizy obszernej bazy tekstów, dające możliwość przeprowadzenia tego

---

a w konsekwencji – w tekstach, w korpusach językowych. Obecnie wydaje się to trudne, ale nie niemożliwe; 2. zaufanie statystyce i ujawnianym dzięki niej trendom/tendencjom. Przywołam tu jako przykład zamach na arcyksięcia Ferdynanda w Sarajewie 28 czerwca 1914 r., który był bezpośrednią przyczyną wybuchu I wojny światowej (*nota bene* – nie analizowałem danych frekwencyjno-leksykalnych sprzed jej wybuchu), aczkolwiek sytuacja polityczna w Europie na długo przez lipcem 1914 r. była bardzo trudna, nabrzmiała i dynamiczna – zatem zakładam, że analiza lingwometryczna wykryłaby to napięcie przedwojenne. Jeśli będzie możliwe powołanie zespołu informatyczno-lingwistycznego do opracowania tak zaprojektowanego narzędzia lingwometrycznego, to warto byłoby, aby zespół przetestował oba zarysowane tu podejścia.

rodzaju badań frekwencyjno-leksykalnych. Warto też rozwijać systematycznie korpusy zarówno ogólne (podobne do NKJP), jak i specjalistyczne (podobne do KDP). Dopiero do pracy na takich korpusach można by zaimplementować narzędzie informatyczne, tj. narzędzie lingwometryczne powstałe na podstawie zaproponowanej tu metody/procedury lingwometrycznej. To narzędzie cyfrowe powinno automatycznie zbierać dane korpusowe (frekwencyjno-leksykalne) związane z badaniem danego zjawiska, czyli dane dotyczące jakiejkolwiek zmiany w pozajęzykowej rzeczywistości, np. wydarzenia. Proponowane urządzenie lingwometryczne powinno bazować na lematach – jako reprezentantach pojęcia – i wyszukiwać wszystkie synonimy lematu centralnego dla danego pola tematycznego (można by dołączyć komponent słownika synonimów). Należałoby także włączyć do takiego cyfrowego narzędzia lingwometrycznego dwa kolejne analizatory: parametr wzorowany na wygładzaniu w Odkrywce (jako istotny składnik procedury lingwometrycznej), a także automatyczny przelicznik danych liczbowych na dane procentowe, dotyczący frekwencji i częstotliwości jednostek językowych w danej jednostce czasu.

Można by też odwoływać się do korpusów obcojęzycznych, aby za pomocą danych z różnych języków precyzyjniej i na szerszą skalę obserwować nadchodzące zmiany oraz prognozować rozwój wydarzeń.

Wydaje się więc, że ową analizę lingwometryczną można sobie wyobrazić tylko jako procedurę maszynową (cyfrową), ponieważ analizy ręczne, liczenie ręczne jest niezwykle czasochłonne. Przy obecnym stanie techniki i informatyki można takie narzędzie przygotować w zespole informatyczno-językoznawczym. Mogłoby się ono opierać na założeniach i opracowanej metodzie/procedurze lingwometrycznej, jakie przyjąłem i zaproponowałem w tym artykule.

Ostatnia uwaga

Wydaje się, że można przewidzieć przyszłość na podstawie danych językowych – tekstowych, frekwencyjno-leksykalnych. Wydaje się też, że da się zauważyć pewne zmiany w rzeczywistości pozajęzykowej na mniej więcej 2–4 lata przed danym wydarzeniem. Dane frekwencyjno-leksykalne wcześniejsze niż czteroletnie (licząc wstecz od analizowanego wydarzenia) nie wykazują statystycznie istotnych właściwości. Podejrzewam, że łatwiej – stosując metodę lingwometryczną – przewidzieć takie wydarzenie jak WOJNA niż KRYZYS (choć wpływ na wynik badawczy mogło mieć także inne ustawienie parametru wygładzanie). W planie kognitywnym może się to wiązać z językowymi obrazami *wojny* oraz *kryzysu* w polskiej (i chyba każdej innej) wspólnocie komunikatywnej. *Wojna* jawi się jako namacalne fizycznie i psychicznie zagrożenie, potencjalnie śmiercionośne oraz niosące druzgocące skutki społeczne, polityczne, psychiczne, ekonomiczno-gospodarcze; w PWK mamy prototyp wojny w postaci II wojny światowej. Z *kryzysem* jest natomiast nieco inaczej: w PWK odczuwamy go nie jako fizyczno-psychiczne zagrożenie, ale jako problem do obejścia, przeżycia (tzw. lata chude), które trzeba przetrwać. PWK pamięta kryzys w Polsce lat 80. XX w. Wszak trzeba mieć świadomość, że ostatni wielki kryzys globalny miał miejsce w 1929 r. i w początkach lat 30. XX w. (co *nota bene* było jednym z czynników

wybuchu II wojny światowej). W zbiorowej – globalnej i polskiej – świadomości tak naprawdę od 1945 r. do 2008 r. nie było globalnego kryzysu gospodarczego. Były kryzysy lokalne, np. w Polsce lat 80. XX w., w Rosji na początku XXI w., w Brazylii. Jednak te lokalne kryzysy można było ominąć – np. wyjeżdżając z regionu/kraju objętego lokalnym kryzysem. Zatem także to – czyli językowy obraz *kryzysu* ujawniający się w PWK – może wpływać na wynik przeprowadzonego badania lingwometrycznego: PWK nie konceptualizuje tak wyraziście *kryzysu* jak *wojny*. Jednakże sądzę, że zaproponowana metoda, na tym etapie jeszcze w wersji analogowej, pozwala prognozować przyszłość, oszacowywać nadchodzące zmiany w rzeczywistości pozajęzykowej. Jest to więc metoda/procedura gotowa do zaaplikowania do cyfrowego narzędzia lingwometrycznego.

Rozwiązanie skrótów

KDP – Korpus Dyskursu Parlamentarnego, http://clip.ipipan.waw.pl/PSC, http://sejm.nlp.ipipan.waw.pl/query_corpus/ (dostęp: 1.03.2024).

Monco – Wyszukiwarka korpusowa Monco, http://monco.frazeo.pl (dostęp: 1.03.2024).

NKJP – Narodowy Korpus Języka polskiego, http://www.nkjp.pl/ (dostęp: 1.03.2024).

Słownik synonimów – Słownik synonimów, https://www.synonimy.pl/ (dostęp: 1.03.2024).

Bibliografia

Bartmiński J., 2006, *Językowe podstawy obrazu świata*, Lublin.

Batko-Tokarz B., 2019, *Tematyczny podział słownictwa współczesnego języka polskiego – teoria, praktyka, leksykografia*, Kraków.

Borawski S., 2005, *Podstawy idei poznawczej studiów nad dziejami używania języka. Esej o diachronii*, [w:] *Rozprawy o historii języka polskiego*, red. S. Borawski, Zielona Góra, s. 13–61.

Graliński F., 2019, *Against the arrow of time. Theory and practice of mining massive corpora of Polish historical texts for linguistic and historical research*, Poznań.

Lakoff G., Johnson M., 1980, *Metaphors we live by*, Chicago.

Lewandowska-Tomaszczyk B., 2009, *Corpus linguistics. Computer tools, and applications – state of the art*, Frankfurt am Main.

Ogrodniczuk M., 2018, *Polish Parliamentary Corpus*, [w:] *Proceedings of the LREC 2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, red. D. Fišer, M. Eskevich, and F. de Jong, Paris, s. 15–19.

Pęzik P., 2020, *Budowa i zastosowania korpusu monitorującego MoncoPL*, „Forum Lingwistyczne", 7, s. 133–150.

Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B. (red.), 2012, *Narodowy Korpus Języka Polskiego*, Warszawa.

Zabrocki L., 1963, *Wspólnoty komunikatywne w genezie i rozwoju języka niemieckiego*, cz. I: *Prehistoria języka niemieckiego*, Wrocław–Warszawa–Kraków.

Can WAR and CRISIS be predicted from linguistic data? Preliminary proposal for a linguometric procedure

Abstract

In this article, I discuss a linguometric method/procedure that I am developing myself, with which it will be possible to forecast upcoming changes in the extra-linguistic reality. Econometric analyses are a model for this procedure. Underlying the proposed linguometric method, I made a cognitive assumption related to the linguistic picture of the world (or more precisely here: textual worldview) of a some kind of communicative community (here: Polish communicative community) and conducted corpus-based frequency-lexical analyses with reference to three exemplary historical events: a) the global economic crisis of 2008–2009; b) the outbreak of World War II in 1939; c) the Russian aggression against Ukraine in 2022. In the corpus-based linguistic data, I tracked the frequency-lexical symptoms of the coming of the mentioned events. It was possible to establish by means of statistical data (frequency and textual frequency of synonymic words from the thematic fields CRISIS and WAR) that language 'foretells' an imminent change in reality 2 to 4 years before it occurs (it is therefore possible to predict future events on the basis of linguistic data). It seems, therefore, that the proposed linguometric procedure can be complementary to econometric, political science, sociological, and many other analyses. I also think that the method developed can be implemented into a digital linguometric tool operating on language corpora, which – for the purposes of linguometrics – should be continuously developed. Such a linguometric tool could be developed in the information and linguistics team.