Korpus Współczesnego Języka Polskiego

Dekada 2011–2020 w tekstach polskich

Marek Łaziński

Dariah.lab

Zespół Inżynierii Lingwistycznej
Instytut Podstaw Informatyki
Polskiej Akademii Nauk

IPI PAN

Fundusze Europejskie Inteligentny Rozwój

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

KORPUS REFERENCYJNY

Korpus Współczesnego Języka Polskiego (KWJP) obejmujący teksty drugiej dekady XXI wieku to korpus referencyjny. Zawiera teksty pisane redagowane (bez mediów społecznościowych). Ma niezmienny skład i korzystamy z niego z takim samym zaufaniem jak ze słowników i encyklopedii z założeniem, że częstość, towarzyszące słowa i skojarzenia (kolokacje) będą takie, jak u przeciętnych użytkowników języka.

Podobne cechy ma **Narodowy Korpus Języka Polskiego** (nkjp.pl), gromadzący teksty z lat 1918–2010.

STRUKTURA

Zrównoważony podkorpus KWJP ma 100 milionów segmentów i jest dostępny obecnie do przeszukiwania. Składają się na niego:

- teksty prasowe 35% — głównie dzienniki i tygodniki,
- literatura piękna (*fiction*) 30%,
- non-fiction (książki i periodyki) 35%.

Wielki niezrównoważony korpus docelowo będzie zawierał ponad miliard słów, głównie z prasy.

W korpusie mamy 693 książki 677 autorów. Fikcja to 295 książek, non-fiction — 398. Są tu książki nagradzane w konkursach literackich (także książki noblistki) i czytane masowo powieści obyczajowe oraz sensacyjne (te kategorie nie są rozłączne). Wśród tekstów non-fiction jest literatura popularnonaukowa i naukowa sensu stricte, a także dużo historii, publicystyki historycznej, politycznej i społecznej kojarzonej zarówno z prawicą, jak i z lewicą, a także reportaży.

AUTORKI I AUTORZY (WYBÓR)

Tadeusz Bartoś, Agata Bielik-Robson, Marek Bieńczyk, Włodzimierz Bolecki, Martyna Bunda, Max Cegielski, Sławomir Cenckiewicz, Wojciech Chmielarz, Sylwia Chutnik, Jacek Dehnel, Eugeniusz Dębski, Ludwik Dorn, Agnieszka Frączek, Andrzej Friszke, Maciej Gdula, Piotr Gontarczyk, Maciej Hen, Jacek Hugo-Bader, Tadeusz Isakowicz-Zaleski, Inga Iwasiów, Jerzy Jarniewicz, Agnieszka Jeż, Piotr Ibrahim Kalwas, Jaś Kapela, Ignacy Karpowicz, Wojciech Kuczok, Adam Leszczyński, Paweł Lisicki, Cezary Łazarewicz, Tomasz Łubieński, Dorota Masłowska, Marcin Meller, Zygmunt Miłoszewski, Remigiusz Mróz, Katarzyna Nosowska, Artur Nowak, Stanisław Obirek, Tomasz Organek, Wojciech Orliński, Agata Passent, Aleksandra Pawlicka, Jacek Pawlicki, Jarema Piekutowski, Krzysztof Pomian, Zbigniew Rokita, Zyta Rudzka, Eustachy Rylski, Tomasz Sakiewicz, Piotr Semka, Maciej Siembieda, Justyna Sobolewska, Natasza Socha, Paweł Sołtys, Jerzy Sosnowski, Ewa Stankiewicz, Maja Staśko, Maciej Stuhr, Klementyna Suchanow, Justyna Suchecka, Katarzyna Surmiak-Domańska, Ziemowit Szczerek, Mariusz Szczygieł, Monika Sznajderman, Wojciech Szot, Monika Szwaja, Wiesław Paweł Szymański, Robert Tekieli, Olga Tokarczuk, Wojciech Tomczyk, Teresa Torańska, Agata Tuszyńska, Szczepan Twardoch, Tymon Tymański, Mariusz Urbanek, Krzysztof Varga, Andrzej Walicki, Ewa Wanat, Henryk Waniek, Marcin Wicha, Bronisław Wildstein, Mariusz Wilk, Janusz L. Wiśniewski, Marcin Wolski, Rafał Woś, Krzysztof Ziemiec, Jakub Żulczyk

ZRÓWNOWAŻENIE

- Oprócz reprezentatywności tematycznej, zadbaliśmy o regionalną i genderową. Prasa pochodzi ze wszystkich regionów Polski. Mamy 130 różnych tytułów z 44 miast (bez Polska Press).
- Kobiety napisały 291 książek o łącznej wielkości 23 mln słów, mężczyźni zaś – 379 książek, 29 mln słów (pomijamy prace zbiorowe).
- Choć nie jest to podział równy, udział kobiet jest większy niż na przeciętnej liście lektur szkolnych.
- Korpus ma odzwierciedlać przeciętne poczucie językowe, ale też podąża za zmianami społecznymi.

TYTUŁY PRASOWE (WYBÓR)

7 Dni, Akant, Astronomia, Autoportret, Biały Kruk, CD-Action, Chorzowianin, Co Tydzień, Czas Chojnic, Czas Ostrzeszowski, Cztery Kąty, Czuwaj, Do Rzeczy, Duży Format, Dziennik Gazeta Prawna, Dzikie Życie, Esensja, Fakt, Ferment, Filozofuj, Gazeta Lokalna — Białystok, Gazeta Wyborcza z dodatkami tematycznymi i wydaniami regionalnymi: Bydgoszcz, Częstochowa, Gdańsk, Katowice, Kielce, Kraków, Lublin, Łódź, Olsztyn, Opole, Płock, Poznań, Radom, Rzeszów, Szczecin, Toruń, Wrocław, Zielona Góra, Gazeta Nieruchomości, Gazeta Polska, Gazeta Polska Codziennie, Gazeta Powiatowa — Wiadomości Oławskie, Glissando, Głos Adwentu, Głos Milicza, Głos Powiatu Średzkiego, Gość Niedzielny, Gwiazdy Mówią, JazzPRESS, Kocie Sprawy, Kronika Beskidzka, Kultura Współczesna, Kurier Szczeciński, Le monde diplomatique, Liberté!, Logo, Ładny Dom, Magnolia, Mazowieckie To i Owo, Monitor Polski, Mój Biznes, Mówią Wieki, Namiary na Morze i Handel, Newsweek Polska, Niedziela, Nieznany Świat, Nowa Europa Wschodnia, Nowa Konfederacja, Nowe Państwo, Nowości Badawcze, Nowy Obywatel, Pałuki, Parkiet, Pasieka, Pismo Folkowe, Polityka, Polski Przegląd Dyplomatyczny, Poznaj Świat, Pressje, Programista, Projektor, Przegląd, Przekrój, Przełom, Przewodnik Katolicki, Rebel Times, Replika, Rynek Kolejowy, Rzeczpospolita, Słowo Podlasia, Super Express, Szum, Świat Elit, Tygodnik Nowodworski, Tygodnik Powszechny, Tygodnik Pułtuski, Tygodnik Siedlecki, Tygodnik Zamojski, Tylko Zdrowie, Vege, W Drodze, Weranda Country, Wiadomości Krajeńskie, Wiadomości Wrzesińskie, Wiedza i Życie, Więź, Wprost, Wróżka, wSieci, wSieci Historii, Wysokie Obcasy, Wyspa, Z Biegiem Szyn, Znak, Zwierciadło

SŁOWNICTWO LAT 2011-2020

Spróbujmy teraz scharakteryzować słownictwo drugiej dekady XXI wieku na podstawie list frekwencyjnych korpusu z podziałem na typ tekstu. Mamy listę leksemów i form, bigramów, czyli najczęstszych połączeń dwuwyrazowych, odpowiednio także trigramów i tetragramów złożonych z form lub leksemów.

Porządek częstości słów w korpusie jest taki jak w przeciętnym tekście pisanym: **w**, **i**, **się**, **z**, **na** itd. Najczętsze trigramy to

- **na to że**,
- **być w stanie**,
- **w ten sposób**,

ale w tekstach prasowych na pierwszym miejscu jest **w tym roku**, a w tekstach fikcjonalnych — **nie móc być**.

SŁOWA KLUCZOWE KWJP NA TLE NKJP

Ciekawsze będzie porównanie częstości słów w KWJP i NKJP, słowa kluczowe ostatniej dekady w porównaniu z okresem 1918–2010. Listę otwierają wyrazy pełnoznaczne:

- KORONAWIRUS,
- PIS,
- KOBIETA

(pomijamy słowa gramatyczne jak SIĘ).

Liczne słowa dopiero pojawiły się w dyskursie publicznym, np. TWEET, SELFIE, LAJK, KRYPTOWALUTA, DRON, INFLUENCER, MIESIĘCZNICA, SYMETRYSTA, PRAWAK (LEWAK był od 100 lat). Pojawiły się nowe lub wróciły stare feminatywy, jak GOŚCINI, nowe słownictwo slangowe i covidowe.

SŁOWA KLUCZOWE I

1. koronawirus
2. głowa
3. **kobieta**
4. dane
5. ziemia
6. wzrok
7. **dłoń**
8. **oko**
9. wciąż
10. pandemia
11. przestrzeń
12. **mężczyzna**
13. czuć
14. zacząć
15. imię
16. **ciało**
17. poczuć
18. **narracja**
19. wieczór
20. wyglądać
21. twarz
22. najwyraźniej
23. **historia**
24. słowo
25. drzwi
26. zrobić
27. negocjacje
28. głębia
29. próbować
30. kolejny
31. **relacja**
32. **opowieść**
33. wiedzieć
34. nawet
35. marihuana
36. smartfon

SŁOWA KLUCZOWE II

Na liście zaznaczyliśmy inne słowa kluczowe, które wynikają

- albo z dużego udziału literatury pięknej, jak CIAŁO i jego części,
- albo z „kultury narracji” (RELACJA, OPOWIEŚĆ).

Pominięto wyrazy gramatyczne i nazwy własne.

KOBIETY I MĘŻCZYŹNI

Wśród słów kluczowych są KOBIETA (3) oraz MĘŻCZYZNA (12). Oba słowa mają częstość na milion słów ponad półtora raza większą niż w NKJP. Oba występują często w liczbie mnogiej.

	KOBIETA			MĘŻCZYZNA		
	l.p.	l.m.	na mln	l.p.	l.m.	na mln
KWJP	36 083	33 006	680	29 206	17 093	460
NKJP	49 659	76 965	420	47 363	37 353	280

DYSTANS MIĘDZY RODZAJEM A PŁCIĄ

Współczynnik dystansu między rodzajem a płcią jest to stosunek form męskich do żeńskich l.p. czasowników oznaczających sytuacje, w których uczestniczą tak samo często mężczyźni i kobiety. W tekstach XX wieku wynosi między 2,5 a 3.

	PWN	NKJP	KWJP
„wiedział” : „wiedziała”	3	2,7	1,6
„myślał” : „myślała”	3,2	3,2	2,2
„mówił” : „mówiła”	2,6	2,2	1,6
wszystkie czasowniki		1,45	1,43

W XXI w. współczynnik spada, ale wciąż jest daleki od 1.

STARE SŁOWA TEŻ COŚ MÓWIĄ

Porównajmy częstość w NKJP i w KWJP słownictwa codziennego.

- Trzy najczęstsze dania obiadowe w KWJP to **pizza**, **kebab**, **kotlet**, w NKJP były to: **pierogi**, **kotlet**, **pizza**.
- Najczęściej wspominane w tekstach urządzenia komunikacji (media) to w NKJP **telewizja/telewizor**, **telefon** (w tym komórkowy od lat 90.), **komputer**, **radio**, w KWJP: **telefon**, **telewizja/telewizor**, **komputer**, **radio**.

WNIOSKI (Z KORPUSU I Z PREZENTACJI)

- Korpus mówi nam coś o wyobrażeniach i skojarzeniach odzwierciedlonych w języku, nie o faktach. To nie jest rocznik statystyczny ani krytycznie opracowane źródło historyczne.
- Nie istnieją korpusy odzwierciedlające poczucie językowe i obraz świata każdego użytkownika, bo te odczucia mamy różne.
- Zespół KWJP dołożył wszelkich starań, by przygotować bezstronne źródło do badań słownictwa i systemu językowego ostatniej dekady. Mamy nadzieję, że nam się udało – także dzięki Państwu.

Dziękujemy!