2026 PRACE JĘZYKOZNAWCZE XXVIII/2
ISSN 1509-5304 DOI 10.31648/pj.12516 159–177
eISSN 2450-0801

Michał Szczyszek
Uniwersytet im. Adama Mickiewicza w Poznaniu
ORCID: https://orcid.org/0000-0002-0253-7296
e-mail: szczysze@amu.edu.pl

Empiryczna weryfikacja probabilistycznej koncepcji normy językowej na materiale polskim (badania pilotażowe)

Empirical verification of the probabilistic concept of the linguistic norm (a preliminary study based on Polish material)

Abstrakt
W artykule przedstawiona została empiryczna weryfikacja propozycji metodologicznej w zakresie normatywistyki: weryfikacja probabilistycznej koncepcji normy językowej. Zakłada ona, że struktury językowe użyte w tekstach wykazywać mogą większą lub mniejszą probabilistyczną zgodność ze znormalizowanym wzorcem języka (ZWJ). Znormalizowany wzorzec języka rozumiany jest jako centrum statystyczne wynikające ze wzorów Gaussowskich, a dokładniej „obszar" rozkładu normalnego na wykresie oscylujący wokół mediany i obejmujący jeden lub dwa centyle; natomiast obszar znormalizowanych użyć języka znajdowałby się w granicach pierwszej sigmy. Przeprowadziłem analizy – zgodnie z proponowaną w koncepcji teoretycznej – procedurą czterech konstrukcji językowych uznawanych za błędy w języku polskim: **przyszłem/ przyszedłem*; *kartka papieru*; **cofać się z powrotem*; **włanczać/ włączać*. Wynik analiz zdają się wstępnie potwierdzać operatywność proponowanej koncepcji probabilistycznej, a także – konieczność stworzenia lingwistycznego narzędzia cyfrowego: normalizatora, w którym wykorzystane zostaną wzory Gaussowskie i narzędzia statystyczne służące obliczeniu rozkładu normalnego zjawisk (tu: zjawisk językowych) zautomatyzowałyby proces kodyfikacji zjawisk językowych zgodnie z probabilistyczną koncepcją normy językowej.

Słowa kluczowe: norma językowa, probabilistyczna koncepcja normy językowej, weryfikacja probabilistycznej koncepcji normy językowej, normalizator

Abstract
The article presents empirical verification of a methodological proposal within the field of normativity studies, i.e. verification of a probabilistic concept of the linguistic norm. This framework posits that linguistic structures employed in texts may exhibit varying degrees of probabilistic alignment with a Normalized Language Pattern. This pattern

is conceptualized as a statistical centre derived from Gaussian models – specifically, as the 'area' of a normal distribution oscillating around the median and spanning one or two percentiles; the area of normalized language use would fall within the boundaries of the first standard deviation (sigma). Following the procedure outlined in the theoretical framework, analyses were conducted on four linguistic constructions traditionally classified as errors in the Polish language: **przyszłem/ przyszedłem* [*I came*]; *kartka papieru* [*a sheet of paper*]; **cofać się z powrotem* [*to go back*]; **włanczać/ włączać* [*to switch on*]. Results of these analyses appear to tentatively confirm the operability of the proposed probabilistic model. Furthermore, they underscore the necessity of developing a digital linguistic tool – a 'normaliser' – utilizing Gaussian formulas and statistic instruments to calculate the normal distribution of linguistic phenomena. Such a tool would automate the process of codifying linguistic phenomena in accordance with the probabilistic concept of the linguistic norm.

Keywords: linguistic norm, probabilistic concept of the linguistic norm, verification of the probabilistic concept of linguistic norms, normaliser

Wstęp, założenia, tezy

W artykule poddano weryfikacji koncepcję probabilistyczną koncepcję normy językowej, którą zaproponowałem w swoim artykule z 2025 r. (Szczyszek 2025b). Podstawą tej koncepcji są trzy aksjomaty, które szerzej omówiłem we wspomnianym artykule, a tu tylko je przywołam.

Pierwszym z tych aksjomatów jest istnienie wspólnoty komunikatywnej – w ujęciu Ludwika Zabrockiego (Zabrocki 1963). Drugim aksjomatem jest uznanie, że *norma językowa* istnieje i jest poziomem wewnętrznej organizacji języka, tworząc triadę z *systemem językowym* i *uzusem* (por. Bąba 1981 za: Coşeriu 1952; por. też Buttler, Kurkowska, Satkiewicz 1971). Trzecim aksjomatem jest przyjęcie założenia, że znany dotąd w lingwistyce dualny układ *kompetencja językowa* i *kompetencja komunikacyjna* jest/ powinien być uzupełniony o komponent *kompetencja normatywna*, tworząc ontologiczny ciąg trzyelementowy, tj. kolejną triadę: *kompetencja językowa* ↔ *kompetencja komunikacyjna* ↔ *kompetencja normatywna* (zob. Szczyszek 2025a).

Na tle powyższych aksjomatów sformułowana została probabilistyczna koncepcja normy językowej, którą w przywołanym tekście traktuję także jako skutek kształtowania się normy, tj. procesów normalizacji uzależnionych od zjawisk zewnętrznojęzykowych: świadomości językowej użytkowników danego języka (to pierwszy krąg wpływów na proces normalizacji języka); postawy kodyfikatorów wobec normy – od powtarzalność metanormotwórczej do innowacyjności metodologicznej (drugi krąg wpływów); komponent „polityczny", zwłaszcza w XXI w. rozumiany jako państwowotwórcza funkcja normy językowej (krąg trzeci). Normę językową na tym tle postrzegam jako jedną z norm społecznych, a więc jako fakt społeczny w ujęciu

Durkheimowkim (Durkheim 2000) konstytuujących wspólnotę komunikatywną (na poziomie państwowym).

Uwzględniając powyższe założenia, w przywołanym artykule sformułowałem probabilistyczną koncepcję normy językowej. W tym ujęciu normę językową proponuję rozumieć jako zbiór jednostek języka i reguł ich łączenia, które wykazują prawdopodobieństwo sprawczości językowej, komunikatywności językowej, zachowania porozumienia językowego i wzajemnego zrozumienia uczestników aktu komunikacji należących do danej wspólnoty komunikatywnej; zatem jest to zbiór jednostek języka i reguł ich łączenia, których zgodność ze znormalizowanym wzorcem języka (ZWJ) jest mierzalna statystycznie za pomocą prawdopodobieństwa zgodności, czyli za pomocą metod probabilistycznych znanych jako rozkład normalny (lub krzywa Gaussa). Innymi słowy, każda realizacja języka, każdy komunikat językowy, każde użycie słowa jest bliższe lub dalsze od owego znormalizowanego wzorca języka, co można wyrazić za pomocą statystycznie ujmowanego prawdopodobieństwa zgodności. Im większe prawdopodobieństwo zgodności, tym dany komunikat jest bardziej znormalizowany; im niższe prawdopodobieństwo zgodności, tym komunikat jest mniej znormalizowany. Natomiast znormalizowany wzorzec języka rozumiany jest jako centrum statystyczne wynikające ze wzorów Gaussowskich, a dokładniej „obszar" rozkładu normalnego na wykresie oscylujący wokół mediany i obejmujący jeden lub dwa centyle – jak pisałem w przywołanym artykule (Szczyszek 2025b). W toku późniejszych weryfikacji tej koncepcji doszedłem do wniosku, że trzeba by odwołać się do teorii trzech sigm (zob. Wawrzynek 2007; Zimny 2010) i przyjąć – prócz pojęcia *znormalizowany wzorzec języka* – także i pojęcie *znormalizowane użycia języka*: realizacje języka, które byłyby uznawane za znormalizowane, znajdowałyby się w granicach pierwszej sigmy, obejmując tym samym ok. 68% tekstowych realizacji języka. Znormalizowana jednostka językowa i jej użycie to taka jednostka, dla której normalizator wyliczył wyższe/ wysokie prawdopodobieństwo zgodności z ZWJ; jednostka nieznormalizowana to taka, dla której normalizator wyliczył niższe/niskie prawdopodobieństwo zgodności z ZWJ.

W przywoływanym tekście (Szczyszek 2025b) opisałem też teoretycznie *procedurę kodyfikacji probabilistycznej normy językowej*, a więc mechanizm wykorzystania funkcji. Do realizacji tej procedury potrzebne są korpusy językowe, narzędzie obliczeniowe – np. kalkulator lub specjalnie opracowane narzędzie cyfrowe (postulowany w tym przywołanym artykule normalizator, który po wprowadzeniu odpowiednich danych obliczałby poziom prawdopodobieństwa zgodności danej jednostki językowej z ZWJ), lub – obecnie – po prostu jakiś model AI. Ponadto niezbędny jest wzór Gaussowski:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

gdzie x – to zmienna losowa, m – wartość średnia, σ – wariancja tej zmiennej losowej (odchylenie standardowe).

Niniejszy artkuł jest w całości poświęcony empirycznej weryfikacji probabilistycznej koncepcji normy językowej – i jest to weryfikacja pilotażowa. Weryfikację przeprowadziłem na kilku losowo wybranych konstrukcjach językowych, uznawanych w tradycyjnym językoznawstwie normatywnym za tzw. błędy lub usterki językowe. Wykorzystując przywoływaną koncepcję, jak i procedurę analityczną, przeprowadziłem więc stosowne analizy, których wyniki przedstawiam poniżej.

Analizy wybranych tzw. błędów językowych

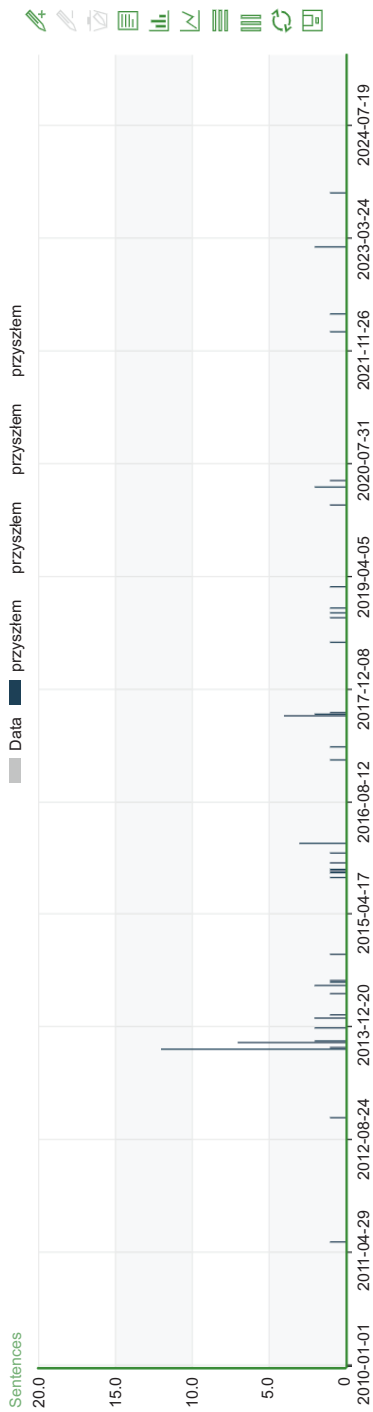
W analizach – wstępnych, pierwszych – skupiłem się na jednej zmiennej, ale bardzo istotnej dla procesu normalizacji: na frekwencji w ujęciu chronologicznym. Frekwencję ustaliłem na podstawie danych korpusowych pobranych z MoncoPL – dysponowałem zatem danymi językowymi z ostatniego piętnastolecia (a dokładniej: z okresu 1 stycznia 2010 do 19 maja 2025 – w ujęciu tygodniowym), a więc niemalże pokoleniowe; są to dane – jak podają twórcy korpusu – pochodzące z internetu, a więc z języka pisanego w internecie, a zatem mające swoje źródła zarówno w portalach informacyjnych, jak i forach dyskusyjnych itp. Jest to frekwencja skorelowana z funkcją czasu, co ten korpus umożliwia. Dzięki temu można prześledzić chronologiczne zmiany frekwencji, które z kolei rzutują na ustalenie prawdopodobieństwa zgodności z ZWJ. Do obliczeń wykorzystałem jeden z modeli AI: Gemini (https://gemini.google.com/). Analizie poddałem następujące konstrukcje językowe:

- **przyszłem / przyszedłem*;
- *kartka papieru*;
- **cofać się z powrotem*;
- **włanczać / włączać*.

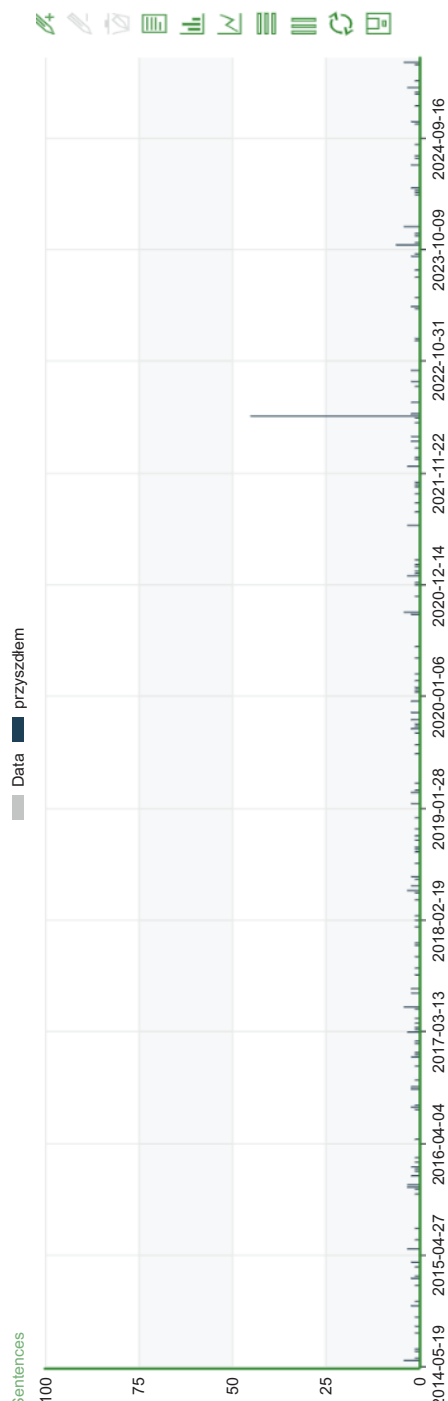
Analiza **przyszłem/ przyszedłem*

Ekscerpcja materiału z MoncoPL przy zastosowaniu odpowiedniego pytania (podyktowanego wymogami wyszukiwarki dołączonej do tego korpusu) dała wynik: **przyszłem* – 69 realizacji; *przyszedłem* – 402 realizacje.

Poniższe wykresy – 1. i 2. – wygenerowane z MoncoPL pokazują rozkład frekwencji w czasie (w ujęciu tygodniowym, tj. wykresy pokazują, ile razy



Wykres 1. Frekwencja formy tekstowej **przyszłem*
Źródło: MoncoPL.



Wykres 2. Frekwencja formy tekstowej **przyszedłem*
Źródło: MoncoPL.

w każdym tygodniu (od 1 stycznia 2010 do 19 maja 2025) w tekstach zgromadzonych w korpusie pojawiła się dana konstrukcja językowa).

Dane frekwencyjne w ujęciu chronologicznym można też pobrać z MoncoPL w postaci sekwencji liczb, które pokazują, ile razy w danym tygodniu dana forma wyrazowa się pojawiła. Dla formy **przyszłem* (wybrany losowo okres) ta sekwencja liczb (czyli rozkład frekwencji w poszczególnych tygodniach) wygląda tak:

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 12,
1, 0, 0, 7, 2, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 2, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0,
0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 4, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.

Natomiast dla formy *przyszedłem* – sekwencja liczb (czyli rozkład frekwencji w poszczególnych tygodniach) jest taka:

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 3, 0, 0, 5, 0,
2, 1, 0, 0, 1, 1, 0, 3, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 1, 2, 0, 0, 0, 0, 0,
0, 1, 0, 0, 1, 0, 0, 0, 0, 2, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,
1, 0, 0, 0, 2, 2, 0, 0, 2, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 3, 1, 0, 1, 1,
0, 1, 0, 2, 1, 0, 0, 1, 2, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 3, 6, 2, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 3, 0, 0, 2, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1,
0, 0, 0, 1, 1, 4, 0, 0, 2, 0, 3, 0, 0, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 3, 0, 0, 0, 0, 3, 1, 0, 0, 0, 0, 0,
1, 0, 1, 0, 0, 0, 0, 4, 1, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0,
0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 2, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0,
3, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 2,
4, 1, 0, 0, 0, 2, 1, 0, 3, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1, 2, 1, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0,
2, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 3, 0, 2, 0, 1, 0, 1, 0, 0, 0, 1, 3, 1, 0, 0, 0, 0,
2, 0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0,
0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 2, 1, 2, 0, 0, 0, 3, 0, 0, 0, 0,
0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1,
1, 0, 0, 0, 2, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0,
0, 0, 1, 0, 2, 1, 1, 0, 1, 1, 0, 2, 0, 0, 0, 0, 0, 2, 0, 0, 1, 0, 2, 0, 0, 0, 1, 0,
0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2,
4, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 2, 1, 2, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1,
1, 0, 0, 0, 0, 0, 0, 0, 3, 0, 1, 0, 1, 0, 0, 1, 0, 0, 2, 0, 2, 0, 0, 0, 0, 0, 1, 0,
0, 44, 2, 2, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 7, 0, 0, 1, 1, 0,
0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0, 1,
0, 0, 0, 1, 1, 0, 2, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 4, 0.

Tak pobrane dane – w postaci sekwencji liczb – można już wprowadzić do Gaussowskich wzorów matematycznych. To zadanie wykonałem przy pomocy AI Gemini, zadając pytanie: *Korzystając z tego wzoru* [przytoczonego wyżej Gaussowskiego wzoru matematycznego] *oblicz rozkład normalny dla tego zbioru danych* [powyższej sekwencji liczb]; model AI przeliczając dane, korzysta z zewnętrznych aplikacji kalkulatorowych (dla pewności przeliczyłem te dane samodzielnie – korzystając z kalkulatora)

Odpowiedzi AI przedstawiają się następująco:

Wynik analiz dla **przyszłem*:
liczba danych (N): 300
suma danych: 35
średnia (μ): ≈ 0.1167
wariancja: ok. 1.4256
odchylenie standardowe (σ): ≈ 1.194

Teraz, mając te parametry, można wstawić je do wzoru na krzywą Gaussa:

$$f(x) = 1/1.194\sqrt{2\pi}*e^{-½(x-0.1167/1.194)2}$$

Co w efekcie daje takie (przykładowe) wartości:
f(0) ≈ 0.3325
f(1) ≈ 0.2540
f(2) ≈ 0.0961
$f(7) \approx 2.29\times10^{-8}$
$f(12) \approx 6.28\times10^{-23}$

Wynik analiz dla *przyszedłem*:
liczba danych (N): 500
suma danych: 316
średnia (μ): ≈ 0.632
wariancja: ok. 1.518
odchylenie standardowe (σ): ≈ 1.232

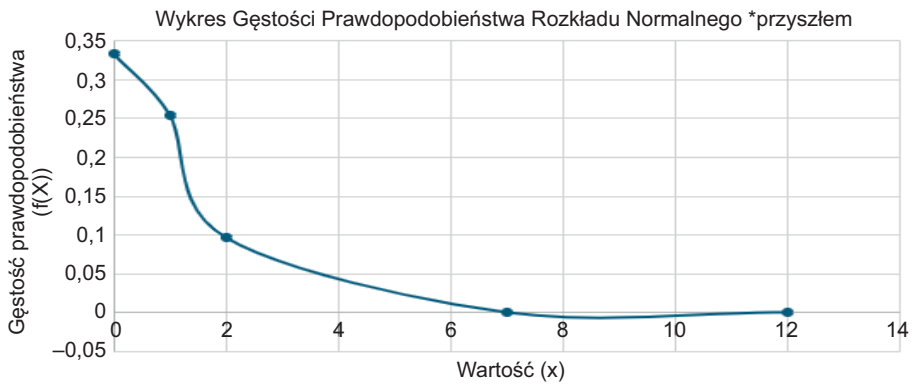
Mając te parametry, można wstawić je do wzoru na krzywą Gaussa:

$$f(x) = 1/1.232\sqrt{2\pi}*e^{-½(x-0.632/1.232)2}$$

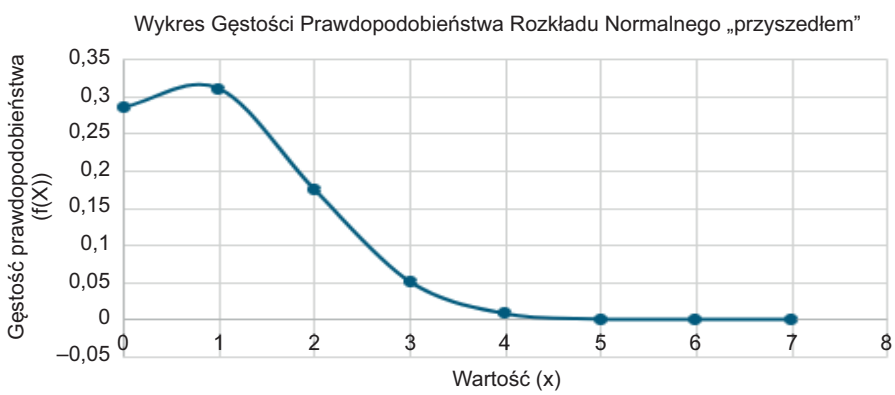
Co w efekcie daje takie przykładowe wartości:
f(0) ≈ 0.2847
f(1) ≈ 0.3106
f(2) ≈ 0.1753
f(3) ≈ 0.0511

f(4) ≈ 0.0077
f(5) ≈ 0.00061
f(6) ≈ 2.44×10−5
f(7) ≈ 5.10×10−7
f(44) ≈ bardzo bliskie 0

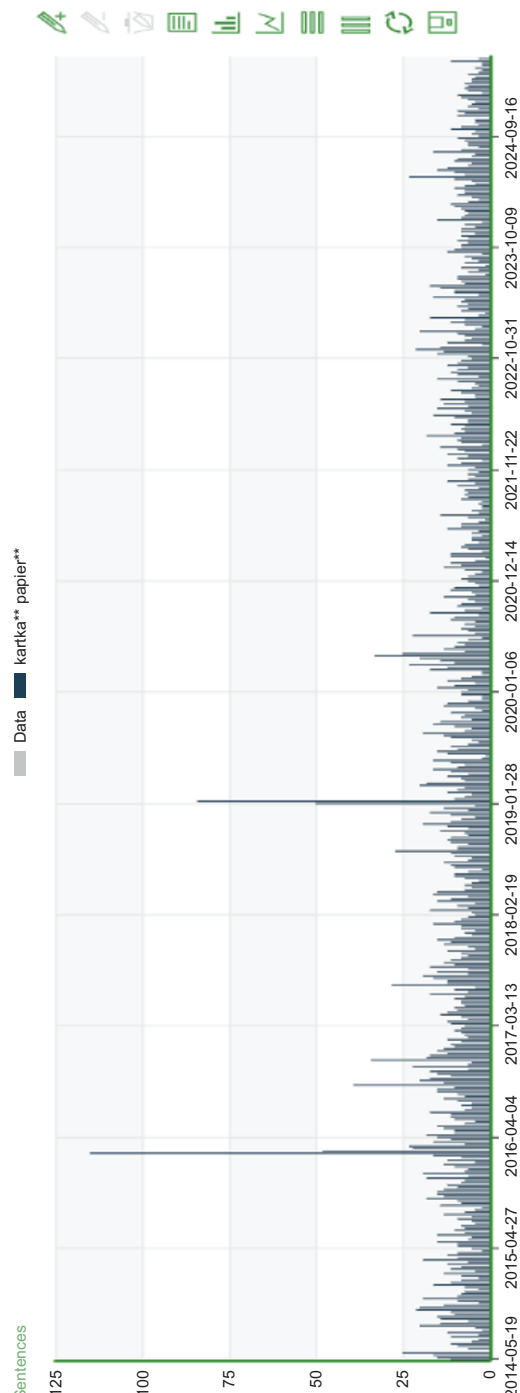
Na tej podstawie można wygenerować wykresy przedstawiające rozkład normalny (krzywą Gaussa) dla badanych form wyrazowych. Wykres 3. pokazuje krzywą Gaussa dla formy **przyszłem*, a wykres 4. – dla formy *przyszedłem*.



Wykres 3. Krzywa Gaussa dla formy tekstowej **przyszłem*
Źródło: opracowanie własne.



Wykres 4. Krzywa Gaussa dla formy tekstowej *przyszedłem*
Źródło: opracowanie własne.



Wykres 5. Frekwencja formy tekstowej *kartka papieru*
Źródło: MoncoPL.

Analiza *kartka papieru*

Identyczną analizę przeprowadziłem dla konstrukcji *kartka papieru*, która w MoncoPL została łącznie odnotowana 8134 razy. Rozkład chronologiczny tej frekwencji przedstawia wykres 5.

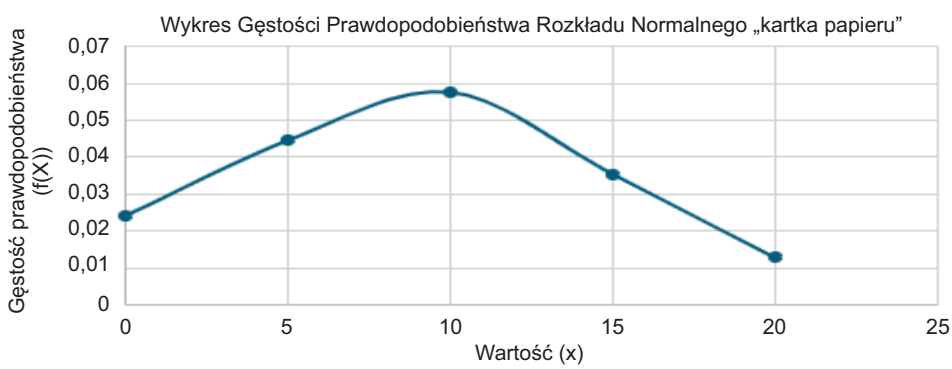
Korzystając z AI Gemini uzyskałem następujące wyniki:
liczba danych (N): 500
suma danych: 4598
średnia (μ): = 9.196
wariancja: ok. 47.78
odchylenie standardowe (σ): ≈ 6.912

Podstawiając te parametry do wzoru, uzyskałem taką jego postać:

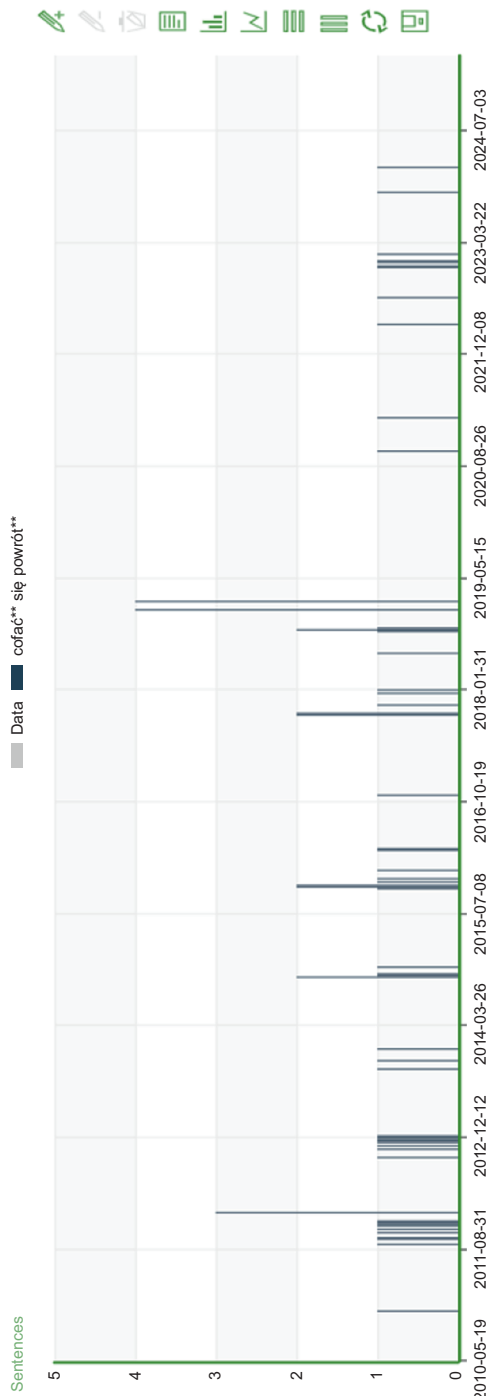
$$f(x) = 1/6.9122\sqrt{2\pi}*e^{-½(x-6.9122/9.196)2}$$

I takie wyniki (przykładowe)
f(0) ≈ 0.2847
f(5) ≈ 0.3106
f(10) ≈ 0.1753
f(15) ≈ 0.0511
f(20) ≈ 0.0077

Na ich podstawie mogłem opracować wykres 6:



Wykres 6. Krzywa Gaussa dla formy tekstowej *kartka papieru*
Źródło: opracowanie własne.



Wykres 7. Frekwencja formy tekstowej **cofać się z powrotem*
Źródło: MoncoPL.

Analiza **cofać się z powrotem*

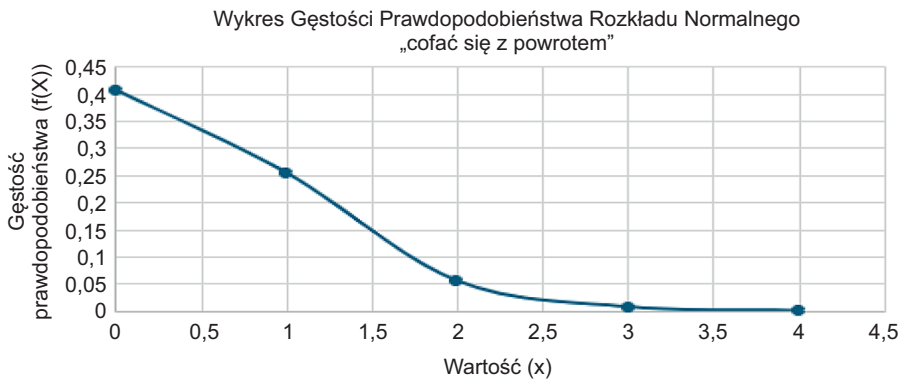
W MoncoPL ta konstrukcja miała 71 realizacji. Rozkład w czasie pokazuje wykres 7.

Na podstawie danych z MoncoPl, wzór Gaussowski ma tu postać następującą:

$$f(x) = 1/0.96162\sqrt{2\pi}*e^{-½(x-0.96162/0.145)2}$$

dzięki czemu uzyskałem takie wyniki obliczeń:
f(0)= 0.407
f(1)=0.255
f(2)=0.056
f(3)=0.008
f(4)=0.001

Na ich podstawie opracowałem wykres 8:

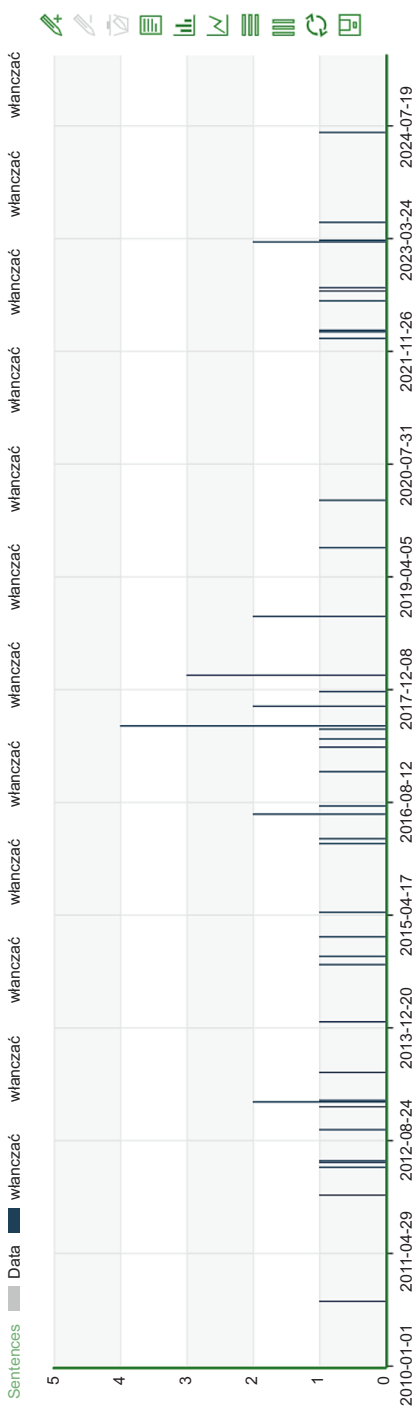


Wykres 8. Krzywa Gaussa dla formy tekstowej **cofać się z powrotem*
Źródło: opracowanie własne.

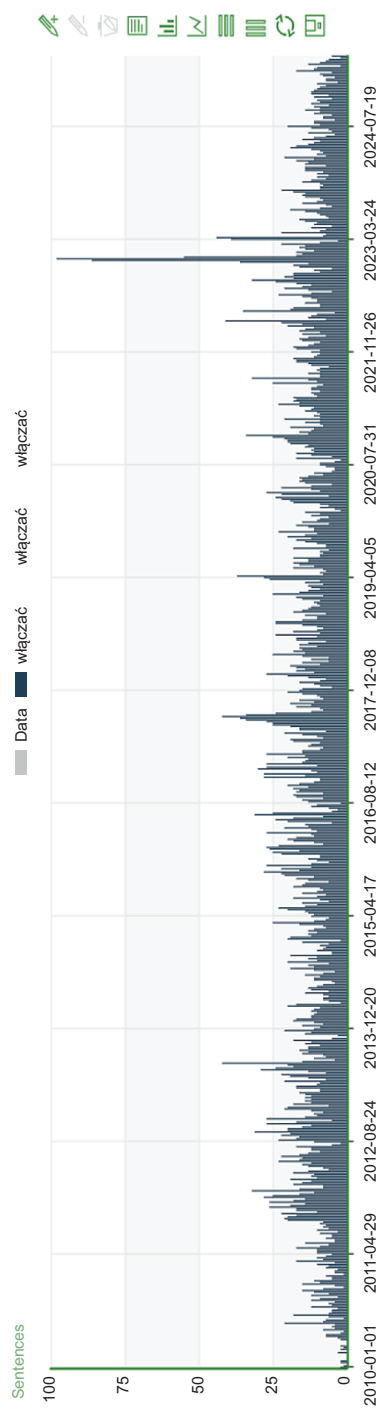
Analiza **włanczać/ włączać*

Frekwencja **włanczać* w MoncoPL wyniosła 51 realizacji, a dla formy *włączać* – 10 647 realizacji. Wykres 9. pokazuje frekwencję **włanczać*, a wykres 10. – frekwencję *włączać* w ujęciu chronologicznym.

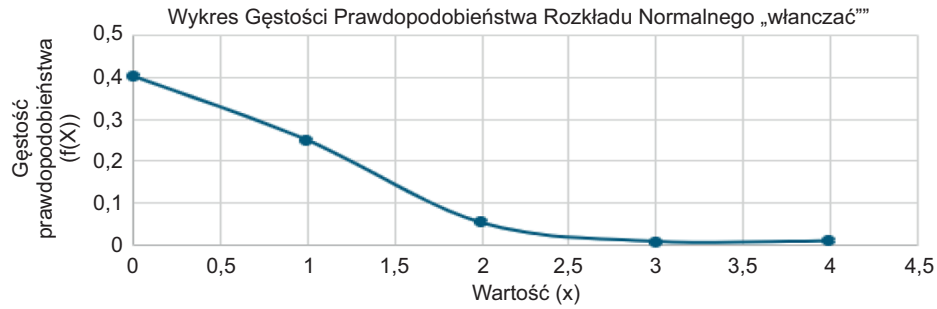
Ponownie – podobnie jak w przypadku poprzednich analiz – przeprowadziłem za pomocą AI Gemini obliczenia na podstawie sekwencji liczb reprezentujących frekwencję chronologiczną w ujęciu tygodniowym, dzięki czemu mogłem opracować wykresy Gaussowskie dla oby tych form tekstowych. Są to wykresy 11. i 12.



Wykres 9. Frekwencja formy tekstowej **włanczać*
Źródło: MoncoPL.



Wykres 10. Frekwencja formy tekstowej *włączać*
Źródło: MoncoPL.



Wykres 11. Krzywa Gaussa dla formy tekstowej **włanczać*
Źródło: opracowanie własne.

Wykres 12. Krzywa Gaussa dla formy tekstowej *włączać*
Źródło: opracowanie własne.

Wnioski

Powyższe analizy – przedstawione też i na wykresach krzywej Gaussa – pokazują prawdopodobieństwo wystąpienia danej formy językowej w danej jednostce czasu. Dane jezykowe pochodzą z korpusu MoncoPL, które pokazują, ile razy w danym tygodniu w określonej jednostce czasu pojawiła się dana jednostka, np. 2010-01-01=0, 2010-01-08=0, 2010-01-15=0, 2023-11-10=2. Zatem wykresy pokazują prawdopodobieństwo najczęstszego wystąpienia danej jednostki, ilustrując tym samym językową tendencję wyrażoną matematycznie do kształtowania się normy, tj. proces normalizacji danej konstrukcji językowej w danym okresie na podstawie samej frekwencji podawanej w funkcji czasu – podstaw liczbowych dla obliczenia prawdobodobieństw. Wykresy rozkładu normalnego – tu opartego na danych frekwencyjnych danej konstrukcji językowej z korpusu – obrazują więc prawdopodobne nasycenie tekstów daną formą językową w tygodniowej jednostce czasu.

Dla formy *włączać* na podstawie obliczeń z wykorzystaniem wzorów Gaussowskich udało się ustalić, że ma ona duże prawdopodobieństwo pojawienia np. 10 razy w ciągu tygodnia, a małe prawdopodobieństwo pojawiania się 0 razy lub 30 razy na tydzień; natomiast forma **włanczać* ma duże prawdopodobieństwo pojawienia 0 razy na tydzień, a małe prawdopodobieństwo pojawiania się 2 razy i więcej na tydzień. Zatem: forma *włączać* ma większe prawdopodobieństwo bycia/ stania się formą standardową/ znormalizowaną, a **włanczać* – nie wykazuje takiej potencji (przynajmniej nie w języku pisanym – w środowisku internetowym).

Z kolei konstrukcja **cofać się z powrotem* ma duże prawdopodobieństwo pojawienia 0 razy na tydzień, a małe prawdopodobieństwo pojawiania się 1 i więcej razy tydzień. Można zatem powiedzieć, że nie wykazuje ona potencji stania się konstrukcją ustandaryzowaną i nie ma szans na jej znormalizowanie, tj. uznanie jej za zgodną z normą współczesnej polszczyzny.

Kartka papieru ma natomiast duże prawdopodobieństwo pojawienia 10 razy na tydzień, a małe prawdopodobieństwo pojawiania się 0 razy lub 20 razy na tydzień; można zatem powiedzieć, że ta forma językowa staje się ustandaryzowana, wykazuje przesłanki probabilistyczne, aby uznać ją za znormalizowaną, zgodną z normą współczesnej polszczyzny.

Natomiast w odniesieniu do pary **przyszłem/ przyszedłem* można stwierdzić – na podstawie powyższych analiz – że: *przyszedłem* ma duże prawdopodobieństwo pojawienia 1 x na tydzień, a małe prawdopodobieństwo pojawiania się 4 i więcej razy na tydzień; forma **przyszłem* ma duże prawdopodobieństwo pojawienia 0 x na tydzień oraz 1 i 2 razy na tydzień, a małe prawdopodobieństwo pojawiania się 7 razy i więcej na tydzień. Zatem: obie formy (**przyszłem i przyszedłem*) zdają się być pod tym względem porównywalne*:* mimo istotnej różnicy w czystej frekwencji (**przyszłem* – 69 realizacji; *przyszedłem* – 402 realizacje) obie wykazują jednak porównywalne prawdopodobieństwo bycia/ stania się formą standardową (obie formy wykazują duże prawdopodobieństwo występowania 1 x na tydzień), co może być pewnym zaskoczeniem w kontekście tradycyjnych rozstrzygnięć normatywnych w tym zakresie.

Podsumowanie

Powyższe przykładowe analizy normatywne przeprowadzone zgodnie z probabilistyczną koncepcją normy językowej pokazały – moim zdaniem – zasadność i operatywność tej koncepcji i procedury analitycznej bazującej na prawdopodobieństwie i rozkładzie normalnym zjawisk – tu: językowych.

Powyższe analizy konstrukcji uznawanych w języku polskim za błędne ukazały bardzo wyraźnie procesualny charakter normalizacji: niektóre z analizowanych konstrukcji wchodzą do normy, stają się znormalizowane (np. *kartka papieru*), inne – nie wykazują potencjału normatywnego: ich wejście do normy współczesnej polszczyzny jest bardzo mało prawdopodobne (**włanczać*, **cofać się z powrotem*). Natomiast zaproponowana metoda analityczna oparta na wspomnianej koncepcji probabilistycznej ujawniła wysoce zaawansowaną rywalizację dwóch form wyrazowych **przyszłem vs. przyszedłem*. Można by zaryzykować stwierdzenie (wymagające analiz na innych formach pochodnych od czasownika *iść*, tj. **poszłem vs. poszedłem; *wyszłem vs. wyszedłem* itp.), że zaproponowana analiza ujawnia silną rywalizację obu modeli fleksyjnych, a może nawet – rozpoczęcie usuwania z systemu współczesnej polszczyzny wyjątkowego wariantu na *-szedłem*. Udowodnienie tych ostatnich stwierdzeń wymaga oczywiście dalszych badań z wykorzystaniem zaproponowanej metody analizy i koncepcji. Mogę jednakże sformułować wniosek, że probabilistyczna koncepcja normy językowej i oparta na niej metoda analizy danych językowych, która jest tu podstawą językoznawczego procesu kodyfikacji normy, wykazały się wysokim poziomem operatywności badawczej, skuteczności i precyzji analitycznej.

Dodać należy, że przeprowadzone analizy na wybranych konstrukcjach językowych współczesnej polszczyzny potwierdzają także aktualność wyrażonego przeze mnie w przywoływanym artykule (Szczyszek 2025b) postulatu o potrzebie zbudowania narzędzia cyfrowego – normalizatora – który te analizy przeprowadzałby automatycznie po wprowadzeniu do niego wymaganych danych kwantytatywnych dotyczących konstrukcji językowych, a także z (pół)automatycznym przeprowadzaniem testów normalności rozkładu (w zależności od potrzeb – np. test Shapiro-Wilka (zob. Shapiro, Wilk 1965) lub test Kołmogorowa-Smirnowa (zob. np. Bedyńska, Cypryańska 2013) czy rozkład Poissona (zob. Wawrzynek 2007). O ile przeprowadzona przeze mnie procedura – w części obliczeniowej wykorzystująca AI – nie jest nazbyt czasochłonna w odniesieniu do (tylko!) 4 przykładów, o tyle analiza językowych *big data*, czyli wszystkich konstrukcji językowych danego języka (występujących w tekstach) poddawanych językoznawczej kodyfikacji, czyli stwierdzaniu zajścia lub nie normalizacji językowej danej konstrukcji, już wymagałaby takiego narzędzia do automatycznej analizy danych na potrzeby normatywistyki. W ten sposób można by uzyskać probabilistyczny obraz normy danego języka.

Sądzę, że ta wstępna weryfikacja probabilistycznej koncepcji normy językowej, którą przeprowadziłem na potrzeby niniejszego artykułu, potwierdzić może wnioski-postulaty zawarte w przywołanym wyżej moim

tekście (Szczyszek 2025b), w którym opisałem podstawy tej koncepcji. Piszę w nim postulatywnie:

> Zaproponowana probabilistyczna koncepcja normy językowej oraz sformułowane na jej podstawie procedura kodyfikacji i narzędzie: normalizator umożliwią – w moim przekonaniu – odejście od powtarzalności metanormotwórczej (przejawiającej się m.in. niekiedy postawami kodyfikatorów oraz tendencjami kodyfikacyjnymi polegającymi na przejmowaniu wcześniejszych orzeczeń czy ustaleń normatywnych w kolejnych późniejszych opracowaniach ortoepicznych) ku innowacyjnym możliwościom orzekania w zakresie prawdopodobieństwa znormalizowania użyć tekstowych jednostek języka w danym czasie w ujęciu synchronicznym.

Po wstępnej weryfikacji empirycznej na materiale polskojęzycznym jestem wstanie z pewną ostrożnością badawczą potwierdzić – na tym etapie badań – zasadność tego wniosku-postulatu.

Potwierdzam także zasadność kolejnego postulatu, który sformułowany został tak:

> W mojej ocenie probabilistyczna koncepcja normy językowej umożliwia także odejście w procesie orzekania poprawnościowego do kryteriów zewnętrznojęzykowych (np. kryterium autorytetu kulturalnego, kryterium uzualnego) i oparciu się na danych frekwencyjno-statystycznych i odwołaniu się do żywiołu języka i jego tendencji obserwowanych w korpusach. To natomiast uniezależniłoby procesy normalizacji języka, a zwłaszcza kodyfikacji jego normy, od takich zjawisk pozajęzykowych, jak świadomość językowa i postawy wobec języka oraz – z drugiej strony – od ingerencji politycznych władz danego państwa. Uniezależnienie normy językowej i jej kodyfikacji od wpływów politycznych pozwoliłoby na traktowanie normy językowej jako istotnego faktu społecznego w ujęciu Durkheimowskim […].

Potwierdzam też – na podstawie przeprowadzonej weryfikacji empirycznej ostatni mój wniosek-postulat z przywołanego artykułu teoretycznie przedstawiającego moją koncepcję normy, który miał postać:

> […] warto zauważyć, że proponowana w tym artykule probabilistyczna koncepcja normy językowej pozwoliłaby na odejście od niekiedy problematycznego ujęcia normy znanego jako dwupoziomowość normy językowej. Po prostu – zjawiska językowe jawiłyby się jako probabilistycznie normatywne.

Ostateczne potwierdzenie wszystkich zapisanych powyżej wniosków wymaga dalszych badań – analiz empirycznych na materiale dowolnego języka wykorzystujących probabilistycznej koncepcji normy językowej i procedurę analityczną na niej opartą. Znalezienie faktu językowego falsyfikującego tę koncepcję – unieważni ją.

Literatura

Bąba S. (1981): *Z zagadnień współczesnej normy językowej*. „Studia Polonistyczne" IX, s. 27–36.
Bedyńska S., Cypryańska M. (red.) (2013): *Statystyczny drogowskaz. 1: Praktyczne wprowadzenie do wnioskowania statystycznego. 2: Praktyczne wprowadzenie do analizy wariancji*. Warszawa.
Buttler D., Kurkowska H., Satkiewicz H. (1971): *Kultura języka polskiego. Zagadnienia poprawności gramatycznej*. Warszawa.
Coşeriu E. (1952): *Sistema, norma y habla*. Montevideo.
Durkheim É. (2000): *Zasady metody socjologicznej*. Tłum. J. Szacki. Warszawa.
Shapiro S.S., Wilk M.B. (1965): *An Analysis of Variance Test for Normality (Complete Samples)*. „Biometrika". Vol. 52, nr 3/4, s. 591–611.
Szczyszek M. (2025a): *Akwizycja normy językowej. Czy zachodzi, jak i kiedy? Na przykładzie systemu słowotwórczego i – częściowo – leksykalnego*. „Język Polski" CV/2, s. 30–40.
Szczyszek M. (2025b): *Jak rozumieć trzeba normę językową? Kodyfikacja – od powtarzalności metanormotwórczej ku innowacyjnym możliwościom korpusologicznym i AI: probabilistyczna koncepcja normy językowej*. [W:] *Kultura komunikacji językowej – powtarzalność i innowacyjność. Tom dedykowany pamięci Profesora Stanisława Bąby w 10. rocznicę śmierci*. Red. A. Piotrowicz-Krenc, M. Witaszek-Samborska, K. Skibski. Poznań, s. 267–281.
Wawrzynek J. (2007): *Metody opisu i wnioskowania statystycznego*. Wrocław.
Zabrocki L. (1963): *Wspólnoty komunikatywne w genezie i rozwoju języka niemieckiego*. Cz. I: *Prehistoria języka niemieckiego*. Wrocław–Warszawa–Kraków.
Zimny A. (2010): *Statystyka opisowa. Materiały pomocnicze do ćwiczeń*. Konin.