UNIWERSYTET WARSZAWSKI
KATEDRA LINGWISTYKI FORMALNEJ

Adam Pawłowski

METODY KWANTYTATYWNE W SEKWENCYJNEJ ANALIZIE TEKSTU

WARSZAWA 2001

METODY KWANTYTATYWNE W SEKWENCYJNEJ ANALIZIE TEKSTU

UNIWERSYTET WARSZAWSKI
KATEDRA LINGWISTYKI FORMALNEJ

Adam Pawłowski

METODY KWANTYTATYWNE W SEKWENCYJNEJ ANALIZIE TEKSTU

Warszawa 2001

Recenzenci naukowi:

prof. dr hab. Jerzy Woronczak
prof. dr hab. Władysław Szczotka

Korekta redakcyjna:

Daria Demidowicz-Domanasiewicz

Redakcja techniczna:

autor

Książka wydana z funduszy Uniwersytetów Warszawskiego i Wrocławskiego

ISBN 83-910656-1-8

Spis treści

Wprowadzenie 5

I. Część teoretyczno-opisowa

1. Przedmiot i cel lingwistyki kwantytatywnej 6
   1.1 Podstawy lingwistyki modelowej 8
      1.1.1 Hipoteza 9
      1.1.2 Kwantyfikacja lub kodowanie 9
      1.1.3 Model 13
      1.1.4 Weryfikacja 14
      1.1.5 Interpretacja 14
2. Przegląd kwantytatywnych praw językowych 14
   2.1 Prawa Zipfa 15
   2.2 Prawo Menzeratha 17
   2.3 Prawo Kryłowa 18
   2.4 Prawo Beöthy 19
   2.5 Prawa Martina 21
   2.6 Prawa językowe a teoria systemów 23
   2.7 Sekwencyjna struktura tekstu a prawa językowe 24
3. Pojęcie sekwencyjnej analizy tekstu 25
4. Linearność tekstu w językoznawstwie niekwantytatywnym 26
5. Linearność tekstu w badaniach kwantytatywnych 31
   5.1 Miary spójności tekstu 34
   5.2 Teoria informacji 38
   5.3 Teoria łańcuchów Markowa 42
   5.4 Analiza widmowa i analiza szeregów czasowych 47
6. Przegląd metod sekwencyjnej analizy tekstu 55
   6.1 Test serii 56
   6.2 Podejście probabilistyczne 57
   6.3 Podejście numeryczne 63
      6.3.1 Definicja szeregu czasowego 64
      6.3.2 Stacjonarność szeregów czasowych 64
      6.3.3 Podstawowe parametry stacjonarnych szeregów czasowych 65
      6.3.4 Wybrane modele liniowe szeregów stacjonarnych 66
      6.3.5 Identyfikacja i estymacja parametrów modelu 70
7. Ograniczenia metody 72

II. Badania materiałowe

1. Porównanie struktury rytmicznej niektórych odmian stylistycznych i wersyfikacyjnych polszczyzny 75
   1.1 Badane teksty i kwantyfikacja 75
      Wiersz sylabotoniczny 77
      Wiersz sylabiczny 78

Proza artystyczna ... 78
Dyskurs oratorski ... 79
1.2 Rezultaty ... 80
Wiersz sylabotoniczny ... 80
Wiersz sylabiczny ... 83
Dyskurs oratorski ... 87
Proza artystyczna ... 89
1.3 Podsumowanie ... 91
1.4 Test efektywności modelowania sekwencyjnego ... 93
2. Analiza porównawcza prozodii języków o akcencie stałym i swobodnym ... 97
2.1 Hipoteza ... 98
2.2 Korpus tekstów, kodowanie, metoda ... 99
2.3 Rezultaty
Proza artystyczna i styl prasowo-publicystyczny – analiza wstępna ... 100
Proza artystyczna i styl prasowo-publicystyczny – podsumowanie ... 104
Wiersz – analiza wstępna ... 104
Wiersz – podsumowanie ... 110
2.4 Prozodia języków o akcencie stałym i swobodnym – podsumowanie ... 113
2.5 Rytmika tekstu a przekład ... 114
3. Struktury sekwencyjne jako kryterium taksonomii tekstów ... 116
4. Sekwencyjna analiza prozodii łacińskiej ... 118
4.1 Iloczas w łacinie – zarys problematyki ... 118
4.2 Prozodia i metryka łaciny – stan poglądów ... 119
4.3 Hipoteza badawcza ... 121
4.4 Badany korpus i kwantyfikacja tekstu ... 121
4.5 Przykład analizy szczegółowej ... 122
4.6 Wyniki sumaryczne ... 127
4.7 Dyskusja ... 128
5. Sekwencyjne modelowanie tekstu na poziomie leksemów i zdań ... 129
5.1 Sekwencje zdaniowe ... 129
5.2 Sekwencje wyrazowe ... 131
5.2.1 Metody ilościowe w typologii języków ... 131
5.2.2 Hipoteza ... 132
5.2.3 Dane i kwantyfikacja ... 133
5.2.4 Analiza szczegółowa ... 135
5.2.5 Wyniki sumaryczne ... 137
6. Zakończenie ... 139
Bibliografia ... 143
Indeks nazwisk ... 155
Indeks rzeczowy ... 159
Aneks ... 163
Quantitative Methods in Sequential Analysis of Text (summary) ... 167

WPROWADZENIE

Problem matematycznego modelowania sekwencyjnych struktur tekstu pojawiał się w historii językoznawstwa wielokrotnie. Skądinąd wartościowe obserwacje i hipotezy wysuwano jednak przy okazji innych poszukiwań – tak empirycznych, jak i teoretycznych – i nie towarzyszyła temu głębsza refleksja związana z usytuowaniem analizy sekwencyjnej w obrębie szerszej problematyki językoznawstwa kwantytatywnego i ogólnego. Mimo znaczącego postępu, jaki w tej dziedzinie nastąpił w ostatnich dekadach, stan zaawansowania ilościowych badań sekwencyjnej struktury tekstu wyraźnie odbiega od innych osiągnięć lingwistyki kwantytatywnej. Wciąż brak jest choćby przybliżonego określenia zakresu badań, a dobór stosowanych narzędzi statystycznych jest często przypadkowy. G. Altmann stwierdza wprost, iż „Theoretical research in this domain is still in its infancy." (ALTMANN 1997:17). Jednak, paradoksalnie, ten stan rzeczy otwiera przed nauką szerokie perspektywy. Wśród zagadnień oczekujących na opracowanie znajdują się podstawowe kwestie metodologiczne oraz lingwistyczne i filologiczne zastosowania analizy sekwencyjnej (prozodia tekstu, wersologia, metryka, filiacja tekstów).

Monografia niniejsza stawia sobie za cel:

- wskazanie genezy sekwencyjnej analizy tekstu;
- uporządkowanie podstawowych pojęć tej gałęzi lingwistyki;
- omówienie metod modelowania sekwencyjnego;
- przedstawienie wstępnie zweryfikowanych hipotez dotyczących sekwencyjnej struktury tekstu.

W pierwszej części pracy omówione zostały podstawy lingwistyki modelowej, najważniejsze prawa lingwistyki kwantytatywnej, dotychczasowe badania nad linearnością tekstu oraz wybrane metody analizy sekwencyjnej. W części drugiej przedstawiono wyniki badań materiałowych. Należy podkreślić, że celem naszym nie było napisanie kolejnego podręcznika lingwistyki statystycznej, uważamy bowiem, że dostępna literatura doskonale wypełnia tę niszę rynkową. Wątek dydaktyczny przy opisie technik matematycznych traktowany był więc drugoplanowo. Uwzględniając natomiast fakt, iż adresatem niniejszej pracy są językoznawcy i filologowie, położono nacisk na lingwistyczne interpretacje uzyskanych wyników.

I. CZĘŚĆ TEORETYCZNO-OPISOWA

1. PRZEDMIOT I CEL LINGWISTYKI KWANTYTATYWNEJ

Lingwistyka kwantytatywna (QL) definiuje język jako wielowarstwową i wielowymiarową strukturę, złożoną z dyskretnych jednostek połączonych ze sobą siecią relacji[1]. Celem QL jest przedstawienie sformalizowanego opisu tych relacji, uwzględniającego ich dynamiczny i ilościowy aspekt. Opis taki przyjmuje postać empirycznie weryfikowalnych i falsyfikowalnych hipotez lub praw językowych, zapisanych w formie modeli matematycznych. Empiryczny i ilościowy charakter badanych prawidłowości zakłada mierzalność i/lub kwantyfikowalność pewnych cech języka. Nie oznacza to bynajmniej rezygnacji z uwzględniania tego wszystkiego, co w języku nie może być przedmiotem pomiaru (na przykład aspektów psycholingwistycznych i semantycznych). Rozważania o takim podłożu są oczywiście istotne, ale pojawiają się na etapie formułowania hipotezy bądź też podczas interpretacji wyniku.

Lingwistyce kwantytatywnej można oczywiście postawić zarzut, iż jej podstawowym wyznacznikiem jest metodologia, a nie jasne założenia programowe. Istotnie, kwestie metodologiczne są dla QL ważne, pozwalają bowiem przenieść na grunt lingwistyki niezwykle skuteczny aparat matematyczny stosowany w naukach przyrodniczych i dzięki temu nadać jej twierdzeniom formę akceptowalną przez ogół nauk. Nie oznacza to jednak braku przesłanek ogólnolingwistycznych. Założeniem leżącym u podstaw większości badań ilościowych (sformułowanym *explicite* lub przyjętym milcząco) jest przekonanie o systemowym i samoregulującym charakterze języka. Wątek ten będzie rozwijany w dalszych rozdziałach.

Zarówno przedstawione tu ogólne założenia epistemologiczne, jak i dotychczasowa praktyka lingwistyki kwantytatywnej wskazują, że najważniejszym przedmiotem badania QL jest tekst, definiowany jako *celowa, wewnętrznie zorganizowana i spójna sekwencja znaków językowych, będąca wytworem działalności komunikacyjnej człowieka*[2]. To właśnie z tekstów wyodrębnia się jednostki fonologiczne, morfologiczne czy leksykalne, którym następnie przypisuje się wartości liczbowe odpowiadające bezpośrednio lub pośrednio ich częstości lub innej kwantyfikowalnej cesze. W oparciu o te wartości, poddane matematycznej obróbce, tworzy się następnie uogólnienia, weryfikuje lub obala hipotezy i ostatecznie formułuje prawa językowe.

---

[1] Kwestie metodologii i naukowego statusu lingwistyki kwantytatywnej obszernie omawia w swych pracach G. Altmann (1978, 1993).

[2] Pojęcie tekstu definiowane jest rozmaicie i brak w tym względzie jednomyślności. Jednak, jak zauważa L. Hřebíček, „There is no generally valid invention concerning the comprehension of this concept. Nevertheless, everybody understands it and this term is apparently used by linguists as well as by layman in the same sense." (HŘEBÍČEK 1995:5). W jednej z wcześniejszych prac ten sam autor proponuje (wraz z Altmannem) następującą, roboczą definicję: „Text is a continuous formation in a natural language that can be segmented into sequences and words." (HŘEBÍČEK&ALTMANN 1993:2).

Każda konkretna hipoteza badawcza bądź prawo językowe osadzona jest w określonej perspektywie badawczej (WĄSIK 1987:74), co wymaga przyjęcia pewnych warunków wstępnych, których skutkiem jest uznanie za relewantne niektórych tylko cech języka (tekstu), a pominięcie innych[3]. Jak dotąd, największą moc eksplanacyjną w QL miały hipotezy, u podstaw których leżały opozycje w pojmowaniu języka jako systemu lub tekstu, a także opozycje w pojmowaniu tekstu jako sekwencji lub populacji jednostek. Także przegląd dorobku QL w ostatnim półwieczu wskazuje, że większość prac można klasyfikować, stosując wspomniane człony opozycji jako deskryptory (KÖHLER 1995).

Jednak poziom zaawansowania ilościowych badań wymienionych tu aspektów języka (opozycje *system* : *tekst* oraz *linia* : *populacja*) jest zróżnicowany i w przypadku sekwencyjnej analizy tekstu wykazuje istotne luki. Przegląd literatury przedmiotu pokazuje, że badane korpusy traktowano najczęściej jak zwykłe populacje statystyczne, a więc zbiory jednorodnych, wzajemnie niezależnych elementów o określonym rozkładzie. Pomijając fakt, że podejście takie może budzić wątpliwości natury metodologicznej[4], ignoruje ono fundamentalną cechę tekstu, jaką jest powiązanie (składniowe, semantyczne, fonetyczne) następujących po sobie jednostek językowych. Efektem tak jednostronnego spojrzenia na problematykę QL jest fragmentaryczny stan wiedzy o linearnej strukturze tekstu przejawiający się w sposób najbardziej jaskrawy niewielką liczbą publikacji poświęconych temu zagadnieniu. Największa bibliografia lingwistyki kwantytatywnej (KÖHLER 1995) cytuje przeszło sześć tysięcy prac (książek, artykułów, recenzji), w której to liczbie zaledwie kilkanaście pozycji opisanych jest deskryptorami sekwencyjny, linearny, liniowy, syntagmatyczny itp. Liczbę tę uznać należy za nieznaczącą. Podobne spostrzeżenia nasuwają się po analizie znacznie starszej *Bibliographie critique de la statistique linguistique* P. Guirauda (1954). Co gorsza, lektura publikacji poświęconych sekwencyjnej strukturze tekstu pokazuje, że różni autorzy cytują różne prace i brak jest minimalnego bibliograficznego kanonu tej gałęzi lingwistyki.

Zasadę linearności przedstawił w formie opozycji binarnej F. de Saussure, pionier europejskiego strukturalizmu. I choć autor *Kursu językoznawstwa ogólnego* zapewne nie myślał o badaniu języka metodami statystycznymi, to właśnie przeciwstawienie związków syntagmatycznych paradygmatycznym i uznanie za prymarną cechę kodu językowego jego budowy linearnej stało się impulsem do podjęcia badań ilościowych sekwencyjnej struktury tekstu, a dziś pozwala lepiej uporządkować pojęcia i metody współczesnej lingwistyki kwantytatywnej.

---

[3] O zasadzie relewancji i idealizacji w naukach empirycznych piszemy na stronie 73.

[4] W przypadku danych językowych, największe wątpliwości budzą takie warunki, jak istnienie *populacji generalnej* o znanym rozkładzie, jednorodność próby oraz niezależność danych. Na przykład wnioskowanie o własnościach populacji generalnej zakłada, iż dokładność oszacowania nieznanego parametru (np. średniej) z próby będzie rosnąć wraz ze wzrostem jej liczebności. W języku warunek ten jest spełniony jedynie w odniesieniu do populacji zamkniętych o niewielkiej liczbie elementów (np. zbiór fonemów danego języka), nie jest natomiast spełniony dla populacji otwartych (np. słownictwa).

1.1 PODSTAWY LINGWISTYKI MODELOWEJ

Ze względu na problematykę niniejszej pracy szczególnie istotny jest jeden z kierunków QL, określany jako lingwistyka modelowa. Kierunek ten rozwija się od początku lat osiemdziesiątych w Niemczech i w Europie Środkowej w kręgu współpracowników G. Altmanna z Uniwersytetu w Bochum i R. Köhlera z Uniwersytetu w Trewirze. Większość programowych tekstów lingwistyki modelowej ukazała się w seriach wydawniczych Quantitative Linguistics i Glottometrika oraz w czasopiśmie „Journal of Quantitative Linguistics". Lingwistyka modelowa stawia sobie za cel „poszukiwanie ogólnych tendencji, tzw. praw statystycznych w tekście (rzadziej w systemie), i opis tych tendencji za pomocą odpowiednich funkcji matematycznych, traktowanych jako modele." (SAMBOR 1988:47). Kwantytatywny charakter praw językowych wymaga posługiwania się symbolicznym językiem matematyki. W przypadku modeli funkcyjnych, powinien być znany przebieg użytej funkcji oraz jej wartości graniczne. Występujące w modelach parametry (zarówno zmienne, jak i stałe) muszą być interpretowalne w kategoriach lingwistycznych. Punktem wyjścia do sformułowania prawa językowego jest wysunięcie tzw. uogólnionej hipotezy, spełniającej w idealnym przypadku następujące warunki[5]:

– dedukcyjność
– kwantytatywność[6]
– weryfikowalność
– falsyfikowalność
– uniwersalność[7]
– niezależność od materiału badawczego[8]
– możliwość włączenia danej hipotezy w obręb szerszego zbioru twierdzeń i praw

Jak z powyższego widać, epistemologiczne fundamenty lingwistyki modelowej osadzone są we współczesnych teoriach nauki opierających się na poglądach K. Poppera i kontynuatorów jego myśli. Fakt ten stawia współczesną lingwistykę kwantytatywną, bodaj pierwszy raz w historii, wśród dojrzałych nauk przyrodniczych[9]. Aby zilustrować tę tezę, zacytujmy *Słownik terminów i pojęć filozoficzny*ch (PODSIAD&WIĘCKOWSKI 1983:339), gdzie pod hasłem „rewolucja naukowa" czytamy: „W naukach humanistycznych uważa się za zdarzenie o charakterze rewolucyjnym wprowadzenie do językoznawstwa

[5] Por. KÖHLER 1986. Cytat na podstawie prac HAMMERL&SAMBOR 1993b:15 oraz ALTMANN 1997:18-19.
[6] Wymaga się jednak, aby każdy parametr modelu posiadał jasną interpretację lingwistyczną.
[7] „These hypotheses [...] must not concern individual languages, i.e. they must concern all languages and be testable." (ALTMANN 1997:19).
[8] „These hypotheses must not contain empirical concepts [...]" (ALTMANN 1997:18).
[9] Chociaż wpływ popperiańskiego falsyfikacjonizmu (AMSTERDAMSKI 1987:591) na metodologię lingwistyki modelowej jest niewątpliwy, teoretykiem nauki, na którego powołują się badacze tej grupy, jest M. Bunge. Na przykład G. Altmann przyjmuje bungowskie pojęcia *nauki* i *teorii naukowej* (ALTMANN 1978, 1993). Drugim epistemologicznym filarem lingwistyki modelowej jest wywodząca się z teorii systemów koncepcja języka jako *systemu synergetycznego*, wprowadzona przez R. Köhlera (HAKEN 1978, KÖHLER 1993).

pojęć i metod logiczno-matematycznych połączone z możliwością empirycznego sprawdzania teorii lingwistycznych dzięki ich powiązaniu z techniką: stosowaniu maszyn cyfrowych do przekładu, indeksowania, streszczania itp.". Liczba znanych dziś praw językowych (w rozumieniu podanej tu definicji) jest dość pokaźna, obejmuje bowiem zależności zachodzące w synchronii i w diachronii oraz w tekście i w systemie. Najlepiej zbadane i z historycznego punktu widzenia najważniejsze są bezspornie prawa Zipfa i Menzeratha. Za niezwykle interesujące należy też uznać prawidłowości zauważone w systemie leksykalnym, aspirujące do statusu praw językowych, ale nie poddane jeszcze wystarczającej liczbie testów (tzw. prawa Martina, Kryłowa i Beöthy).

Analiza sekwencyjna respektuje oczywiście wymienione wyżej ogólne zasady, jednak szczegółowa postać proponowanej procedury badawczej musi uwzględniać specyfikę linearnej struktury tekstu i z tego względu zostanie omówiona dokładniej. Dotychczasowe doświadczenia wskazują, iż w idealnym przypadku składa się ona z pięciu etapów:

1. Wysunięcie testowanej hipotezy;
2. Kwantyfikacja lub kodowanie tekstu;
3. Określenie typu procesu stochastycznego, którego realizacją jest badany tekst i wybór jego modelu;
4. Weryfikacja uzyskanego modelu na szeregach pseudolosowych;
5. Lingwistyczna interpretacja modelu i potwierdzenie lub falsyfikacja hipotezy wyjściowej.

1.1.1 Hipoteza

Hipoteza badawcza w formie pierwotnej powinna być wyrażona za pomocą terminologii i pojęć językoznawczych. Dopiero na etapie wyboru i testowania modelu terminy językoznawcze powinny zostać zastąpione wielkościami sformułowanymi w symbolicznym języku matematyki lub statystyki. I tak, w sensie lingwistycznym nie jest hipotezą przykładowe stwierdzenie, iż „funkcja $f(x)$ będzie maleć monotonicznie wraz ze wzrostem parametru x". Jest natomiast potencjalną hipotezą lub prawem stwierdzenie, iż „średnia długość składników dowolnej jednostki językowej maleje nieliniowo wraz ze wzrostem długości całej jednostki".

1.1.2 Kwantyfikacja lub kodowanie

Kwantyfikacja tekstu w QL polega na przypisaniu jednostkom tekstowym o charakterze jakościowym (kategorialnym) relewantnych wartości liczbowych. Przymiotnik „relewantny" oznacza, iż pewne przypisania uważać będziemy za istotne z punktu widzenia celu badawczego, inne zaś za nieistotne. I tak, długość jednostki tekstowej może być wyrażona w fonemach, morfemach, sylabach, literach, sekundach (jeżeli korzysta się z nagrań) itd. Wybór takiego a nie innego sposobu kwantyfikacji lub kodowania zależy oczywiście od treści testowanej hipotezy. Specyficzną cechą analizy sekwencyjnej jest możliwość

pracy na danych ilościowych i jakościowych (kategorialnych), a także dostępność szeregów prostych i kumulacyjnych odpowiadających tej samej próbie.

Praca na szeregu kategorialnym zwalnia z konieczności kwantyfikacji tekstu, wymusza jednak zastosowanie modeli probabilistycznych i/lub konekcjonistycznych (ELMAN 1990), bazujących na pojęciu *stanu* i wykorzystujących teorię łańcuchów Markowa i/lub techniki teorioinformacyjne. Odpowiednikiem *stanu* może być każda dyskretna jednostka językowa składająca się na uporządkowaną sekwencję tekstową[10]. Model buduje się obliczając proste, a następnie warunkowe prawdopodobieństwa przejścia pomiędzy poszczególnymi stanami. W oparciu o te dane można następnie obliczyć wartości entropii i redundancji systemu oraz oszacować głębokość związku kontekstowego (GUILLBAUD 1979, PETRUSZEWYCZ 1981, BAVAUD 1998, XANTOS 2000).

Poniższy przykład, zaczerpnięty z noweli A. Moravii[11], ilustruje pierwszy sposób kwantyfikacji. Słowa (a dokładniej słowoformy) tekstu „Il giorno dopo, verso le due, puntuale, con l'ombrello sul braccio perché c'era un cielo nero e minacciava di piovere, mi trovai in vie Archimede [...] ” zastąpiono liczbami reprezentującymi częstości odpowiadających im leksemów w przekrojowej próbie współczesnego języka włoskiego[12]. Kryterium reprezentatywności spełniają w tym przypadku zarówno częstości absolutne, jak i względne. Jednak znacznie lepszą miarą jest ilość informacji, liczona na podstawie wzoru C. Shannona (1948):

$$(1) \qquad I_n = -\log_2 p_n$$

Tym sposobem otrzymano sekwencję liczbową posiadającą wyrazistą interpretację lingwistyczną (I_n to przecież nic innego jak miara ilości informacji), stanowiącą zarazem potencjalny materiał dla wszechstronnych analiz kwantytatywnych, także sekwencyjnych.

Interesującym zagadnieniem jest w tym kontekście rozróżnienie szeregów prostych i kumulacyjnych (Rys. 1 i 2). Dowolny dyskretny szereg stacjonarny może zostać zastąpiony szeregiem skumulowanym dzięki sumowaniu kolejnych wyrazów. I na odwrót, dowolny szereg niestacjonarny może zostać zamieniony na szereg stacjonarny poprzez usunięcie trendu. Z metodologicznego punktu widzenia nie ma więc różnicy pomiędzy szeregiem prostym a kumulacyjnym. Różnice zaznaczają się dopiero na etapie interpretacji modelu. Klasycznym przykładem wykorzystania szeregu skumulowanego jest badanie dynamiki przyrostu słownictwa w tekście ciągłym (por. KÖHLER&GALLE 1993, PAWŁOWSKI 1994, TULDAVA 1995). Liczba różnych leksemów użytych od początku tekstu do *i*-tej słowoformy, posiada gotową interpretację lingwistyczną (mówi się o słownictwie autora, bogactwie leksykalnym, zróżnicowaniu słownictwa itd.), co wskazuje, iż szereg kumulacyjny jest w tym przypadku najbardziej efektywnym sposobem kwantyfikacji.

[10] Na przykład stopa metryczna, sylaba (BRATLEY&ROSS 1981, KOŁMOGOROW&PROCHOROW 1964, PAWŁOWSKI 1997), głoska (AZAR&KEDEM 1979) lub cecha dystynktywna (YOKOYAMA&ITASCHI 1980, KÖHLER 1983).

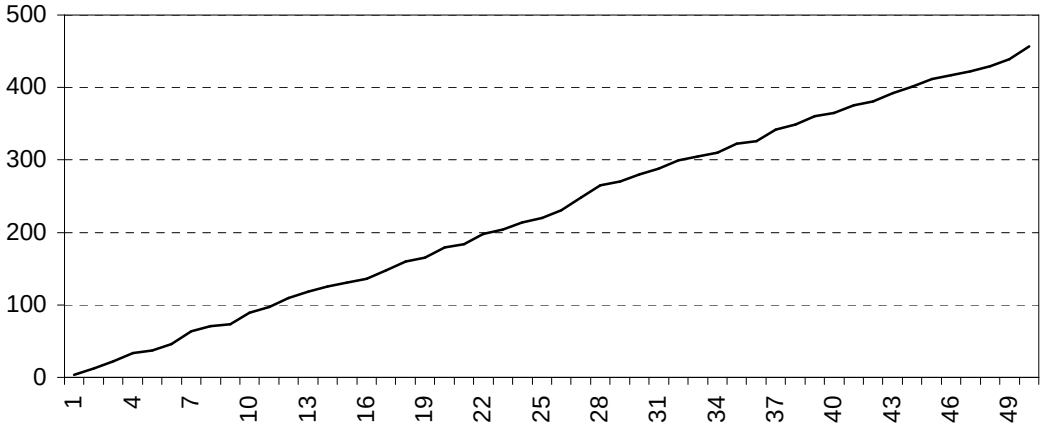
[11] A. Moravia, *Nuovi Racconti Romani di Moravia*, Bompiani 1963, 386.

[12] Punktem odniesienia był korpus 500 000 słów języka włoskiego (BORTOLINI et al. 1971).

Z kolei inne zjawiska (na przykład rytm i metryka tekstu, jego struktura informacyjna) dają się lepiej analizować na podstawie szeregów prostych.

Rys. 1 Sekwencja ilości informacji (w bitach) w kolejnych słowach tekstu włoskiego

Rys. 2 Skumulowana ilość informacji (w bitach) w kolejnych słowach tekstu włoskiego



Przykładem szeregu jakościowego może być sekwencja samogłosek i spółgłosek w tekście francuskim, otrzymana poprzez mechaniczną zamianę liter graficznych na symbole C (spółgłoska), V (samogłoska) i P (pauza lub spacja)[13]. Fragment:

L'affirmation, l'interrogation, le commandement peut-être?

po zakodowaniu miałby postać:

CVCCVCCVCVVCPCVCCVCCVCVCVVCPCVPCVCCVCCVCVCCPCVVCPVCCV

Kodowanie sekwencji literowych, a nie na przykład fonemowych w języku o tak archaicznej pisowni jak francuski może się wydać z lingwistycznego punktu widzenia problematyczne. Jednak kwestia ta nie jest przedmiotem naszych rozważań, a jedynie przykładem ilustrującym pewien postulat metodologiczny. Istotną cechą tego sposobu

[13] Przykład cytowany za pracą XANTOS 2000:362.

kodowania (nie jest to bowiem kwantyfikacja) jest łatwość skonstruowania modelu probabilistycznego, w którym tekst reprezentują stany (tu oznaczone symbolami C, V, P), oraz trudność ewentualnego ich zastąpienia liczbami. Poniżej przedstawiamy macierz prawdopodobieństw przejścia pierwszego rzędu dla sekwencji 13 438 znaków kodowanych według powyższego schematu. Macierz ta nie jest symetryczna i należy czytać ją od lewej do prawej (na przykład prawdopodobieństwo, iż po spółgłosce pojawi się samogłoska $p_{cv} = 0{,}59$ a nie 0,61).

$P =$

	P	C	V
P	0	0,78	0,22
C	0,19	0,22	0,59
V	0,21	0,61	0,18

W obu cytowanych przykładach istnieje możliwość przekształcenia szeregu liczbowego w kategorialny i na odwrót. Każdy stan dowolnego szeregu tekstowego (tu V, C, P) ma jakąś częstość w korpusie i może zostać zastąpiony liczbą. Także dowolna skala liczbowa może zostać zastąpiona szeregiem rozdzielczym, a przynależność do danego przedziału określona jako stan szeregu. W omawianym przykładzie (Rys. 1), skala liczbowa [3, 18], na której określone są wartości I_n, może zostać zredukowana do niewielkiej liczby stanów, *nazwanych* w adekwatny sposób (na przykład leksem bardzo częsty, częsty, powszechny, rzadki itd.). Sekwencja shannonowskiej ilości informacji w kolejnych słowach będzie więc opisana jako droga przejścia od stanu do stanu. Warto jednak pamiętać o tym, że zbyt duża liczba stanów (jeśli za stan uznamy na przykład jednostkę leksykalną) może znacznie utrudnić konstrukcję macierzy prawdopodobieństw przejścia bądź innego, równoważnego modelu. Z kolei liczba zbyt mała stanowić będzie przesadne uproszczenie modelu.

Jakie kryterium powinno więc ostatecznie decydować o konwersji tekstu (kwantyfikacja czy kodowanie, szereg prosty czy kumulacyjny)? Kryterium tym jest bez wątpienia lingwistyczna interpretowalność otrzymanego szeregu czasowego. Dla pewnych kategorii językowych obligatoryjna jest w zasadzie kwantyfikacja (na przykład powszechność użycia leksemu mierzona jest najlepiej, chociaż nie jedynie, jego częstością). Gdzie indziej jednak właściwe jest stosowanie szeregów kategorialnych. Badając sekwencję pojawiania się w tekście części mowy, najbardziej wyrazisty rezultat otrzyma się, kodując dane za pomocą symboli odpowiadających wyróżnionym, według przyjętej konwencji, kategoriom (N, V, Adj. itd.). Podobnie rzecz się ma z wyborem szeregów prostych (stacjonarnych) lub kumulacyjnych. Szereg liczbowy prosty z Rys. 1 można zamienić na kumulacyjny. Otrzymana tym sposobem krzywa (Rys. 2) byłaby wdzięcznym obiektem szczegółowej analizy, pozwalającym między innymi na wyodrębnienie wyraźnego trendu liniowego. Cóż jednak z tego, skoro wielkość ta nie koresponduje, przynajmniej bezpośrednio, z żadną sensowną kategorią lingwistyczną.

Dodajmy na zakończenie, że kwantyfikacja danych lingwistycznych zgodna jest z zasadami kwantyfikacji cech opisowych, przyjętymi w naukach społecznych, szczegól-

nie w badaniach ankietowych. Badacz posługuje się tam jedną ze czterech skal (GATNAR 1988:18–22). O skali nominalnej mówi się w sytuacji, gdy dane są nazywane symbolami wyrażającymi cechy niemierzalne (na przykład rodzaj męski – M, żeński – K, nijaki – N). W takim przypadku można stosować także liczby, jednak prowadzi to do nieporozumień. O skali porządkowej mówi się, jeżeli użyte symbole lub liczby dają się uszeregować według natężenia cechy (na przykład 1 – słowo niezrozumiałe, 2 – słabo zrozumiałe, 3 – zrozumiałe). O skali interwałowej mówi się wówczas, gdy obiekty dają się uporządkować według natężenia cechy, a ponadto znana jest (i najczęściej stała) szerokość interwału (na przykład klasyfikacja słownictwa na klasy częstości, oparta na frekwencji słowoform w korpusie). O skali ilorazowej mówi się w przypadku, gdy spełnione są wcześniejsze warunki, a ponadto dopuszczalne jest mnożenie i dzielenie klas.

W lingwistyce kwantytatywnej najbardziej efektywne są jednak skale oparte na cechach w pełni mierzalnych, którym można w sensowny sposób przypisać liczby rzeczywiste. Analiza sekwencyjna nie jest wyjątkiem od tej zasady. Nawet modele probabilistyczne, wykorzystujące jako dane wyjściowe cechy kategorialne z przestrzeni zdarzeń elementarnych, odwzorowują je następnie poprzez funkcję zwaną zmienną losową na zbiór liczb rzeczywistych lub poprzez rozkład prawdopodobieństwa na zbiór [0, 1][14].

1.1.3 Model

Modelowanie zjawisk językowych podlega ogólnym zasadom teorii symulacji. Opis oraz klasyfikację modeli opartą na dychotomicznych opozycjach *deterministyczny* : *stochastyczny*, *statyczny* : *dynamiczny* i *analityczny* : *numeryczny* podaje G.S. Fishman (1981:24–25). Według tej klasyfikacji modele sekwencyjne należałoby uznać za analityczne (z ich treści można wydedukować rozwiązanie problemu), dynamiczne (uwzględniają zmienną czasu lub pozycję w szeregu czasowym) i stochastyczne (część zmienności danych ma charakter losowy). Modele sekwencyjne można też podzielić stosując jako kryterium rodzaj użytej metodologii. Na podstawie dotychczasowych badań wyróżnić można: 1) modele teorioinformacyjne, wykorzystujące shannonowskie pojęcia informacji i entropii warunkowej (BAVAUD 1998:212, XANTOS 2000, HAMMERL&SAMBOR 1990:361–451); 2) probabilistyczne, ograniczające się do przedstawienia macierzy prawdopodobieństw przejścia i opartych na nich prostych wskaźników (GUILBAUD 1979, PAWŁOWSKI 1998:199–200, PETRUSZEWYCZ 1981) oraz 3) numeryczne oparte na funkcji autokorelacji, mające postać liniowych równań autoregresji lub ruchomej średniej (PAWŁOWSKI 1997, ROBERTS 1996). Przy tworzeniu modeli numerycznych wykorzystuje się także analizę widmową, stosowaną w fonetyce akustycznej (AZAR&KEDEM 1979, BRATLEY&ROSS 1981). Trzeba jednak zdawać sobie sprawę z tego, że każdy szereg tekstowy może być opisany wieloma modelami i konkretny wybór powinien wynikać z przesłanek lingwistycznych.

[14] „Zdarzenia należące do ciała zdarzeń *F* danego doświadczenia losowego mają ważną własność, mianowicie mierzalność, tzn. można przyporządkować im różne miary. W szczególności jako miarę na zdarzeniach $A \in F$ określić można funkcję P{*A*} o wartościach rzeczywistych." (GREŃ 1987:21).

1.1.4 Weryfikacja

Kryteria statystycznej oceny jakości modelu są integralnym składnikiem wszystkich prezentowanych dalej metod i z tego względu nie będą szczegółowo omawiane. Wspomnimy jedynie, że najbardziej typowe „procedury kontrolne" polegają na ocenie istotności parametrów modelu lub rozkładu statystycznego (tzw. testy parametryczne) oraz na ocenie dopasowania modelu do danych empirycznych (tzw. testy nieparametryczne). Oprócz tego istnieją jednak testy specyficzne dla sekwencyjnej analizy tekstu i im warto poświęcić więcej uwagi.

Charakterystyczną cechą „szeregów tekstowych" poddawanych analizie sekwencyjnej jest łatwość ingerencji w dane, polegająca między innymi na zmianie kolejności oraz usunięciu lub dodaniu jednostek. Można tym sposobem tworzyć mniej lub bardziej losowy „pseudotekst", stanowiący punkt odniesienia dla tekstów rzeczywistych. Manipulacje te stanowią więc pochodną stosowanych współcześnie technik symulacyjnych i w przypadku analizy sekwencyjnej mogą wzmocnić testowane hipotezy poprzez porównanie parametrów modeli zbudowanych na danych rzeczywistych, manipulowanych i sztucznie generowanych.

Test polegający na porównaniu szeregu rzeczywistego i manipulowanego zastosował na przykład A. Roberts (1996). Badał on rytmotwórczy charakter zróżnicowania długości kolejnych zdań w tekście artystycznym w języku angielskim i jako punkt odniesienia dla funkcji autokorelacji w tekstach rzeczywistych zaproponował obliczenie analogicznego parametru dla „pseudotekstów", utworzonych poprzez przypadkowe uszeregowanie zdań tekstów autorskich. Stosowano też testy wymagające ingerencji w dane. Porównano na przykład rytm prozy literackiej wyznaczony, tak jak w poprzednim przypadku, sekwencją długości zdań w próbach integralnych i takich, z których usunięto odcinki dialogowe, pozostawiając jedynie narrację opisową (PAWŁOWSKI 1998:102, 136, 155).

1.1.5 Interpretacja

Skuteczna interpretacja modelu jest procesem twórczym i z tego względu nie powinna podlegać żadnym regułom czy ograniczeniom. Należy jednak pamiętać o „przetłumaczeniu" pojęć użytego języka symbolicznego (na przykład statystyki) na pojęcia lingwistyczne bądź interpretowalne w kategoriach lingwistyki. Przy całej swej precyzji i epistemologicznej poprawności, zdania typu: „Odrzucamy / przyjmujemy hipotezę H_0 na poziomie istotności $\alpha = 0{,}05$", nawet dla lingwisty świetnie rozumiejącego treść hipotezy, a nienawykłego do statystycznego żargonu, znaczyć mogą bardzo niewiele.

2. PRZEGLĄD KWANTYTATYWNYCH PRAW JĘZYKOWYCH

Ilościowe prawa językowe były w ostatnich latach przedmiotem licznych dociekań o charakterze formalnym i empirycznym (ALTMANN 1993, KÖHLER 1986, SAMBOR 1988, HAMMERL&SAMBOR 1990, 1993a, 1993b), których podsumowaniem są odnośne roz-

działy przygotowywanego do druku poradnika *International Handbook of Quantitative Linguistics* (ALTMANN&KÖHLER 2002). Z tego względu zostaną tu przedstawione jedynie w takim zakresie, jakiego wymaga wprowadzenie do sekwencyjnej analizy tekstu. Punktem odniesienia dla niniejszego wywodu było obszerne omówienie problematyki praw językowych R. Hammerla i J. Sambor (1993b).

2.1 PRAWA ZIPFA

Prawa Zipfa uważa się za najstarsze i najlepiej zbadane ilościowe prawa językowe. Nawet z pobieżnej lektury *Bibliography of Quantitative Linguistics* R. Köhlera (1995) widać, że poświęcono im ponad 80 publikacji. Syntetyczne ujęcia tego zagadnienia w literaturze polskojęzycznej zawierają prace (HAMMERL&SAMBOR 1993b:17–19) oraz (SAMBOR 1969:30–67, 1972:59–73).

Pod pojęciem praw Zipfa rozumieć należy szereg prawidłowości językowych o charakterze ilościowym, odkrytych i opisanych przez J.K. Zipfa za pomocą modeli funkcyjnych. W szczególności wymienia się tu związki pomiędzy:

– częstością wyrazów a ich pozycją na liście rangowej
– częstością wyrazów a ich długością
– częstością wyrazów a liczbą ich znaczeń
– częstością wyrazów a ich wiekiem i pochodzeniem.

Jednak w literaturze lingwistycznej pojęcie prawa Zipfa kojarzone jest najczęściej tylko z pierwszą zależnością, opartą na powszechnie znanej prawidłowości, zaobserwowanej między innymi przez J.B. Estoupa (1916), zgodnie z którą iloczyn rang i częstości słów z listy frekwencyjnej jest wartością stałą ($fr = const.$, gdzie r – ranga, f – częstość słowa). W celu opisu tej zależności Zipf zaproponował model:

(2) $$p_r = kr^{-1}$$

gdzie p_r – prawdopodobieństwo wystąpienia wyrazu o randze r
k – stała

Pierwsze prawo Zipfa było wielokrotnie modyfikowane i znalazło zastosowania wykraczające poza obszar lingwistyki (WORONCZAK 1967, GUITER&ARAPOV 1982). Jedną z ważniejszych modyfikacji jest propozycja B. Mandelbrota (SAMBOR 1969:34), który wprowadził do równania (2) poprawki uwzględniające nieregularny kształt krzywej modelu w obszarze najniższych rang. Prawo Zipfa-Mandelbrota opisane jest modelem:

(3) $$p_r = k(r + \rho)^{-B}$$

gdzie p_r – prawdopodobieństwo wystąpienia wyrazu o randze r
k – stała
B – współczynnik modelu (stały dla konkretnego tekstu)
ρ – współczynnik spełniający warunek $\rho > 0$ dla $r < a$ i $\rho = 0$ dla $r \geq a$

Kolejna zależność odkryta przez Zipfa orzeka, iż *długość wyrazu maleje w miarę wzrostu jego częstości* (przy czym jednostką długości jest fonem lub sylaba). Zipf zaproponował dla tej zależności następujący model (SAMBOR 1972:61):

(4) $k = C \lg r$

gdzie k – długość wyrazu w fonemach
C – stała
r – ranga

Zipf sformułował też prawo mówiące, że *liczba znaczeń wyrazu jest wprost proporcjonalna do pierwiastka jego częstości* (*ibid.* 62). Zależność tę wyraził funkcyjnie jako:

(5) $m = C\sqrt{f}$

gdzie m – liczba znaczeń wyrazu (Zipf nie określił sposobu ich wyodrębniania)
C – stała
f – częstość wyrazu

Ze względu na związek pomiędzy częstością a rangą leksemów, parametr f w modelu (5) zastąpić można funkcją parametru r. Testy przeprowadzone przez Zipfa na materiale języka angielskiego pozwoliły mu na wyrażenie m następującym modelem empirycznym: $m = r^{-0,46}$.

Inna opisana przez Zipfa zależność statystyczna dotyczy relacji pomiędzy liczbą znaczeń leksemu (m), a liczbą leksemów posiadających daną liczbę znaczeń (L). Orzeka ona, iż *liczba leksemów o danej liczbie znaczeń jest odwrotnie proporcjonalna do kwadratu tej liczby znaczeń* i wyraża się modelem $L = C/m^2$, gdzie C jest stałą (*ibid.* 64).

Badając słownictwo języka angielskiego, Zipf zauważył też związek pomiędzy częstością wyrazu a jego wiekiem i pochodzeniem. Najstarsze słowa języka angielskiego (pochodzenia germańskiego) okazały się zarazem najczęstszymi. Zależność ta nie została jednak przedstawiona w postaci modelu funkcyjnego[15].

W lingwistyce kwantytatywnej prawa Zipfa odegrały ogromną rolę. Jednak z dzisiejszej perspektywy lepiej widoczne stają się ich słabe strony. Jak zauważa J. Sambor (1972:61), Zipf nie uzasadniał w swych pracach wyboru proponowanych modeli matematycznych, kierując się zapewne wizualnym podobieństwem odkrytych rozkładów empirycznych do krzywych niektórych funkcji matematycznych. Jednak te same krzywe empiryczne mogą zostać opisane wieloma funkcjami (STANISZ 1993) i kryteria wyboru modelu powinny opierać się na solidnej podstawie epistemologicznej[16]. Również materiał

[15] Badania ilościowe w diachronii (tzw. *glottochronologia*) prowadzili później m.in. M. Swadesh (1952, 1953, 1955), R.B. Lees (1953) oraz M.V. Arapow i M.M. Cherc (ARAPOV&CHERC 1974, 1983).

[16] Dedukcyjną procedurę budowy modelu prawa Menzeratha szczegółowo opisuje G. Altmann (1981).

językowy wykorzystany przez Zipfa do weryfikacji jego praw wydaje się dziś zbyt ograniczony. I chociaż odkryte przez niego tendencje wciąż pozostają w obszarze zainteresowań lingwistyki kwantytatywnej, inne jest ich ujęcie formalno-metodologiczne. Pojęcie praw Zipfa zastępuje się pojęciem *sił Zipfa*, które miałyby stanowić czynnik sprawczy w mechanizmie samoregulacji utrzymującym równowagę systemu językowego, opartym na zasadzie ekonomii wysiłku[17]. Zasada ta determinuje sposób kodowania informacji w procesie komunikacji i widoczna jest w rozkładach statystycznych większości jednostek językowych. Pewną autonomię zachowuje wciąż tylko wymienione wcześniej pierwsze prawo Zipfa, któremu w przeszłości poświęcono najwięcej studiów o charakterze zarówno lingwistycznym, jak i matematycznym. Mimo komplikacji, jakie do rozważań matematycznych i językoznawczych wprowadza pojęcie rangi, a w pewnej mierze skutkiem swoistej bezwładności, pierwsze prawo Zipfa wciąż należy, wraz z prawem Menzeratha (patrz niżej), do najczęściej cytowanych i testowanych. Zgodnie z przyjętą terminologią, w dalszej części pracy przez prawo Zipfa rozumieć będziemy właśnie wspomnianą tu zależność pomiędzy rangą i częstością.

2.2 Prawo Menzeratha

Prawo to orzeka, iż *długość konstrukcji językowej jest odwrotnie proporcjonalna do długości jej składników* („im dłuższa konstrukcja językowa, tym krótsze jej składniki") i wyraża się funkcją wykładniczą (6). Jako pierwszy zależność tę stwierdził na materiale fonetycznym niemiecki lingwista P. Menzerath, a formalną postać nadał jej G. Altmann:

(6) $y = ax^b$

gdzie
- y – średnia długość składników
- x – długość konstrukcji językowej
- a – przeciętna długość konstrukcji jednoskładnikowej
- b – nachylenie krzywej modelu wskazujące na dynamikę zmiany długości składników (prawo działa jeżeli $b<0$)

Ponieważ w trakcie weryfikacji okazało się, że pewne podsystemy języka nie spełniały powyższej zależności (długość składników nie zawsze malała monotonicznie), Altmann przedstawił wersję uogólnioną prawa, stwierdzającą, że *długość konstrukcji językowej jest funkcją wykładniczą długości jej składników*[18]. Warto zwrócić uwagę na procedurę konstruowania modelu, opartą na przesłankach dedukcyjnych, a nie indukcyjnych. Altmann nie opierał się na jakimś konkretnym zbiorze obserwacji, ale poszukiwał zależności opisującej związek między wielkościami x i y w sposób możliwie uniwersalny, uzależniając od danych empirycznych jedynie wartość parametrów a i b (Altmann 1978:19–23, 1980; Hammerl&Sambor 1993b).

[17] W podobnym duchu, choć za pomocą innej terminologii („uogólnione prawo Zipfa", „zasada ekonomii języka"), problem ten ujmuje W. Mańczak (1996:27–43).
[18] W tym przypadku mówi się raczej o prawie Menzeratha-Altmanna. Por. Altmann&Schibbe 1989.

Prawo Menzeratha-Altmanna zostało przetestowane na kilkudziesięciu językach świata, a jego funkcjonowanie potwierdzono na różnych poziomach językowych planu wyrażania (FENK&FENK-OCZLON 1993) i treści (HAMMERL&SAMBOR 1993b:32). W tym ostatnim przypadku długość całej jednostki wyrażono liczbą sylab, a długość składnika zastąpiono średnią liczbą słownikowych znaczeń leksemu. W toku dalszych poszukiwań okazało się, że prawo to w postaci wyprowadzonej dla danych językowych z powodzeniem stosuje się również w genetyce i primatologii[19], a więc poza sferą języka (*ibid.* 41–45).

2.3 PRAWO KRYŁOWA[20]

Jednym z ciekawszych zagadnień współczesnej lingwistyki kwantytatywnej jest ilościowa struktura słownictwa uwzględniająca zjawisko polisemii leksemów. Związek liczby znaczeń leksemu z jego frekwencją dostrzegł już Zipf, proponując opisanie go funkcją o postaci $L = C/m^2$, gdzie C jest stałą, a L oznaczało liczbę leksemów posiadających m znaczeń (liczba leksemów posiadających m znaczeń jest odwrotnie proporcjonalna do kwadratu m). Model Zipfa uznano jednak za niezadowalający i kontynuowano badania, szukając lepszych przybliżeń. Nieco inny i, jak pokazało doświadczenie, lepszy model zaproponował węgierski lingwista L. Papp (1967). Autor użył funkcji wykładniczej i uwzględnił wielkość słownika[21]:

(7) $$y_x = \frac{W}{2^x}$$

gdzie y_x – liczba leksemów mających x znaczeń
W – liczba leksemów w słowniku$_2$

Jednak ani Zipf, ani Papp nie przeprowadzili wystarczającej liczby testów weryfikujących proponowane modele. Także wybór proponowanych zależności funkcyjnych pozbawiony był uzasadnienia. Ju.K. Kryłow (KRYLOV 1982) oparł się początkowo na modelu Pappa, zastępując jednak liczbę leksemów (y_x we wzorze 7) prawdopodobieństwem p_x wylosowania ze *słownika*$_2$ leksemu posiadającego x znaczeń:

(8) $$p_x = \frac{1}{2^x}$$

Kryłow przeprowadził wiele testów na materiale języka rosyjskiego, wykazując, że model (8) daje bardzo efektowny (jest to przecież formuła najprostszego postępu

[19] Primatologia jest działem zoologii badającym grupowe zachowania zwierząt, m.in. ssaków naczelnych.
[20] Szczegółowy opis tej problematyki wraz z odnośnikami bibliograficznymi i przykładami znaleźć można w pracach SAMBOR 1988, SAMBOR 1989 oraz HAMMERL&SAMBOR 1993:117–125.
[21] Termin *słownik* jest w tym kontekście z konieczności dwuznaczny. Z jednej bowiem strony, realizując postulat uniwersalizmu, prawo językowe ma w założeniu opisać strukturę *słownika*$_1$ rozumianego jako inwentarz leksemów należący do systemu języka, z drugiej zaś empiryczne testy prowadzone były dotąd na konkretnych, skończonych *słownikach*$_2$ językowych. W miarę potrzeby wprowadzona tu notacja stosowana będzie w dalszej części pracy.

geometrycznego!), ale przybliżony opis rzeczywistości (negatywny wynik testu χ^2). Zaproponował więc dla tej zależności model bardziej złożony:

(9) $$p_x = \frac{(w-1)^{x-1}}{w^x}$$

gdzie p_x – prawdopodobieństwo wylosowania leksemu mającego x znaczeń
w – średnia liczba znaczeń leksemu w słowniku$_2$

Powyższa funkcja, przekształcona na funkcję liniową poprzez logarytmowanie, lepiej przybliżała dane empiryczne, choć i w tym wypadku weryfikacja testem χ^2 nie dała wyniku pozytywnego (HAMMERL&SAMBOR 1993:125). Kolejne testy prawa Kryłowa prowadzono na korpusach dobieranych bardziej selektywnie – na przykład jednorodnych pod względem gramatycznym (*passim*).

Podsumowując, bilans prac nad prawem Kryłowa uznać należy za wyjątkowo bogaty. Oprócz rozwiązania konkretnego problemu badawczego udowodniono bowiem, że mimo nieostrości pojęć semantycznych ilościowe badania lingwistyczne uwzględniające słownikowe znaczenia leksemów są możliwe. Badania takie wzbogacają wiedzę lingwistyczną i psychologiczną, nie ustępując pod względem efektywności eksplanacyjnej i dokładności pomiarowej badaniom innych, na pozór bardziej wymiernych poziomów języka. Przypuszczalnym rozszerzeniem prawa Kryłowa będzie zbadanie opisanych wyżej zależności w tekstach, a nie, jak dotychczas, tylko w słownikach$_2$ i w słowniku$_1$.

2.4 PRAWO BEÖTHY[22]

W literaturze lingwistycznej pojęciem tym określa się ilościowe związki zachodzące pomiędzy częstościami leksemów w tekście a liczbą ich znaczeń, jednak przy założeniu że każdy leksem (bądź zbiór leksemów) traktowany jest osobno. Ze względu na wymóg wieloznaczności, klasę badanych jednostek rozszerza się o morfemy gramatyczne o sprecyzowanych znaczeniach (głównie przedrostki), włączając w to funkcje składniowo-gramatyczne. Obserwacje tej zależności prowadzone na danych z języków węgierskiego, niemieckiego, polskiego i francuskiego (SAMBOR&HAMMERL 1993b:129–142) nie doprowadziły jak na razie do zadowalających uogólnień o charakterze ilościowym.

Aby wypełnić częściowo tę lukę, zilustrujemy opisaną wyżej zależność, opierając się na rozkładzie pierwszych dziesięciu słownikowych znaczeń francuskiego spójnika *et* w powieści A. Saint-Exupery'ego *Le Petit Prince*[23] (Tab. 1). Obserwowana relacja przypomina prawo Zipfa – badane są w zasadzie te same wielkości (ranga i częstość), a ich związek jest odwrotnie proporcjonalny (Rys. 3). Także ich iloczyn daje wartości w miarę stabilne (z lekkim trendem wzrostowym). Szukając modelu dla tej zależności, należałoby wyjść od prostego równania różniczkowego:

[22] Podstawą omówienia są prace BEÖTHY&ALTMANN 1984a, 1984b i 1991 oraz HAMMERL&SAMBOR 1993b:129–142.
[23] Na podstawie pracy ROTHE 1986:66, cytuję za HAMMERL&SAMBOR 1993b:132.

(10) $$\frac{dF}{dr} \approx r^{-1}$$

które po wprowadzeniu współczynnika *a* i scałkowaniu stronami daje:

(11) $$F = a \ln r + C$$

Po estymacji współczynników modelu (11) otrzymano funkcję:

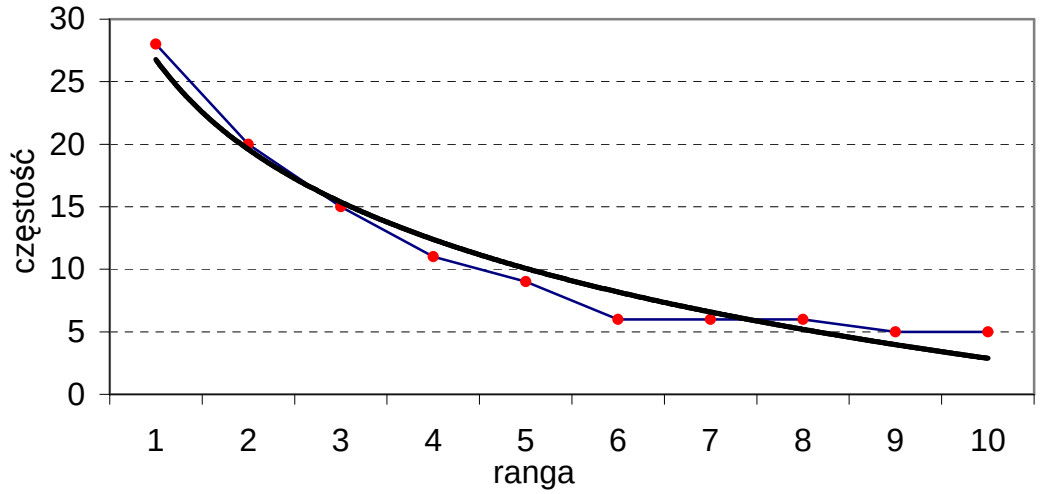
(12) $$F = -10{,}4 \ln r + 27$$

posiadającą wysoki współczynnik dopasowania[24] $R^2 = 0{,}97$.

Tab. 1 Częstości pierwszych dziesięciu znaczeń francuskiego spójnika *et* w powieści *Le Petit Prince*.

ranga (*r*)	znaczenie	częstość obserwowana	częstość teoretyczna	$r \times F$
1	*alors*$_1$	28	26,77	28
2	*puis*	20	19,58	40
3	*mais*$_1$	15	15,37	45
4	*mais*$_2$	11	12,39	44
5	konstrukcja z gerundium	9	10,07	45
6	*de même que*	6	8,18	36
7	*enfin*	6	6,58	42
8	*ci-dessus*	6	5,19	48
9	*alors*$_2$	5	3,97	45
10	*c'est pourquoi*	5	2,88	50

Rys. 3 Częstość znaczeń spójnika *et* w powieści *Le Petit Prince* z estymacją modelu



[24] R^2 wyraża procent zmienności danych empirycznych, wyjaśniony przez model: R^2=1-*SSE/SST*, gdzie *SSE* jest wariancją resztową (ang. *sum squares error*), a *SST* wariancją całkowitą (ang. *sum of squares total*). Obliczenia przeprowadzono za pomocą programu *Microsoft Excel*®.

Przedstawienie formalnego, ilościowego modelu dla jednego, konkretnego przypadku nie jest jednak równoznaczne z odkryciem prawa językowego. Otwiera raczej perspektywę wymagającą bardziej wszechstronnych badań (przekrój zjawiska w wielu językach w aspekcie porównawczym, badanie klas morfemów czy leksemów w konkretnym tekście itd.). W omawianym przypadku brakuje przede wszystkim podbudowy ogólnoteoretycznej o charakterze psycholingwistycznym, która pozwoliłaby między innymi nadać sensowną, lingwistyczną interpretację estymowanym współczynnikom modelu. Reasumując, na obecnym etapie badań tzw. prawo Beöthy należy raczej traktować jako zależność pretendującą do rangi prawa językowego i oczekującą na dalsze weryfikacje.

2.5 PRAWA MARTINA[25]

Prawa Martina opisują prawidłowości statystyczne występujące w hierarchicznie uporządkowanej strukturze leksemów danego słownika. W przeciwieństwie do przedstawionych wyżej praw Zipfa i Menzeratha, wyrażają one zależności występujące w systemie języka, a nie w tekście. Praw tych nie można też wyrazić za pomocą werbalnego schematu odpowiadającego zależnościom funkcyjnym: „zmiana wartości zmiennej niezależnej *x* powoduje takie a nie inne zachowanie zmiennej zależnej *y*". Można je natomiast wyjaśnić, wychodząc od pojęcia ciągu definicyjnego. Przykładem takiego ciągu jest sekwencja *derbista – koń – ssak – kręgowiec – zwierzę – organizm – twór*, zbudowana na podstawie definicji słownikowych, złożona z leksemów o rosnącej denotacji i malejącym stopniu szczegółowości[26]. Analizując duże próby leksemów, zauważono, że ich liczebności na kolejnych poziomach ogólności nie są przypadkowe, ale dają się zdefiniować za pomocą modeli funkcyjnych. Z powodzeniem zastosowano między innymi funkcje rekurencyjne, których wartość y_i (na przykład liczba pojęć na poziomie szczegółowości *i*) zależy od wartości y_{i-1}, a jedynie w niewielkim stopniu od jakiejś zmiennej niezależnej (na przykład od numeru poziomu *i*). Określono też rozkład statystyczny prawdopodobieństwa zdarzeń zdefiniowanych jako pojawienie się (wylosowanie) leksemu należącego do określonego poziomu szczegółowości względnie ogólności (HAMMERL&SAMBOR 1993b:75).

Ilościowe analizy struktury ciągów definicyjnych otworzyły nowe perspektywy przed badaniami z zakresu semantyki leksykalnej. Na podstawie analizy dużej liczby takich ciągów oszacowano stopień abstrakcyjności formantów słowotwórczych w języku polskim (Tab. 2). Zastosowano w tym celu wskaźnik abstrakcyjności (A), uwzględniający liczebności określonych derywatów rzeczownikowych na kolejnych poziomach ogólności ciągów definicyjnych (*ibid*. 85). Z kolei zestawienie hiperonimów końcowych

[25] Problematykę praw Martina, ciągów definicyjnych i gniazd leksykalnych szczegółowo omawiają J. Sambor i R. Hammerl (SAMBOR 1997, HAMMERL&SAMBOR 1993b). Cytowane przykłady zaczerpnięto z tych właśnie prac.

[26] Na podstawie *Słownika języka polskiego* pod redakcją M. Szymczaka, Warszawa, PWN, 1978.

dużej liczby ciągów definicyjnych pozwoliło wyodrębnić empirycznie zbiór leksemów pretendujących do statusu *indefinibiliów* semantycznych.

Tab. 2 Formanty rzeczownikowe uporządkowane według wskaźnika abstrakcyjności

ranga	formant	wskaźnik *A*	ranga	formant	wskaźnik *A*
1	*-anie*	8,6	13	*-arka*	6,8
2	*-ca*	8,5	14	*-acz*	6,7
3	*-cie*	8,4	15	*-ista*	6,7
4	*-enie*	8,2	16	*-izna*	6,6
5	*-ość*	8,0	17	*-ek*	6,5
6	*-ik*	7,4	18	*-ec*	6,4
7	*-acja*	7,3	19	*-owiec*	6,3
8	*-arz*	7,3	20	*-ówka*	5,9
9	*-izm*	7,2	21	*-ak*	5,4
10	*-nik*	7,1	22	*-nica*	5,4
11	*-stwo*	7,0	23	*-arnia*	4,0
12	*-ka*	6,9			

Z drugiej strony, warto pamiętać, że badania struktury leksykalnej i pojęciowej języka, niezależnie od użytej metodologii, napotykają wiele trudności technicznych i interpretacyjnych. Prace nad prawami Martina doprowadziły wprawdzie do sformułowania atrakcyjnych modeli funkcyjnych dla języków polskiego i niemieckiego, ale nie określono jasnych kryteriów oceny ich efektywności i nie nadano przekonującej interpretacji ich parametrom (HAMMERL&SAMBOR 1993b:56, 66). Pozornie wynikało to z różnych zasad budowania ciągów definicyjnych, ale faktycznie jest skutkiem wielkiej złożoności badanego zjawiska. O ile bowiem modele praw Menzeratha, Zipfa i Kryłowa można było wyrazić funkcjami jednej zmiennej, której lingwistyczna interpretacja nie nastręczała trudności, leksyka, czy to w słowniku$_2$, czy w psychice człowieka, jest złożoną całością, w której wzajemnie powiązane są wszystkie elementy i efektywne jej modelowanie wymaga innego podejścia. Unikając jakichkolwiek aluzji o podłożu postmodernistycznym, należy chyba zgodzić się z tezą, iż w tym akurat przypadku izolowanie zjawisk czy warstw języka będzie znacząco deformować obraz całości. Te okoliczności sprawiają, że mimo niezwykle bogatego bilansu prac nad zagadnieniami ilościowej struktury słownika$_1$, liczba nowo pojawiających się pytań wciąż góruje nad liczbą dostarczonych już odpowiedzi. Jednak i te kwestie będą stopniowo rozwiązywane. Semantyczne modele słownika są dziś jednym z kluczowych zagadnień lingwistyki formalnej funkcjonującej na odległym pograniczu akademickiej lingwistyki o podłożu humanistycznym oraz inżynierii językowej, wspomagającej prace nad systemami dialogowymi i sztuczną inteligencją (AI, AL). W nurcie tym, często niezależnie od niepodważalnych osiągnięć QL, pracuje się nad sieciowymi modelami słownika$_1$, wykorzystując do tego celu tech-

niki konekcjonistyczne, na przykład sieci neuronowe. Zadaniem tak konstruowanych modeli jest odwzorowanie w postaci sformalizowanej (ang. *machine-readable*) mechanizmów kognitywnych i struktury wiedzy człowieka[27].

2.6 PRAWA JĘZYKOWE A TEORIA SYSTEMÓW[28]

Modele funkcyjne opisujące ilościową strukturę języka występują także poza sferą języka. Potwierdziły to badania prowadzone między innymi w genetyce, socjologii, geografii kwantytatywnej i semiotyce sztuki. Prawo Zipfa, określające w ogólnej wersji relację pomiędzy rangą danej klasy jednostek a jej liczebnością (relacja *ranga* : *częstość*), z powodzeniem testowano na danych o liczbie i liczebności grup społecznych oraz na danych o liczbie miast i ich mieszkańców (HILL 1982, RAPOPORT 1982). Tym samym modelem opisywano także dzieła malarskie, traktując jako ich relewantną cechę rozłożenie jednobarwnych plam różnej wielkości na płaszczyźnie obrazu (ORLOV&VOLOŠIN 1982). Model funkcyjny *ranga* : *częstość* oraz znany w lingwistyce statystycznej rozkład Waringa-Herdana zastosowano w opisie ilościowej struktury dzieła muzycznego (KÖHLER&MARTINÁKOVÁ-RENDEKOVÁ 1998).

Z kolei prawo Menzeratha, opisujące w ogólnej wersji odwrotnie proporcjonalny związek pomiędzy wielkością całości a średnią wielkością części, zastosowano w genetyce i primatologii. W pierwszym przypadku stwierdzono odwrotnie proporcjonalną relację pomiędzy liczbą chromosomów tworzących cały łańcuch DNA (x), a ich średnią długością (y). Podobnie jak w przypadku danych leksykalnych (por. wzór 6), relacja ta opisana została funkcją potęgową $y = ax^b$ (HAMMERL&SAMBOR 1993b:43). Taki sam model dopasowano także do struktury liczbowej hierarchicznie uporządkowanych grup zwierzęcych – im liczniejsze było stado, tym mniej liczne były tworzące je grupy osobników (KAUMANNS&SCHWIBBE 1989).

Chociaż więc opisywane tu zależności sformułowano po raz pierwszy przy okazji badań lingwistycznych, tak szeroki zakres ich występowania pozwala uznać je za szczególne przypadki oddziaływania jakichś ogólniejszych prawidłowości, których opis należy do dyscypliny nadrzędnej, za jaką zwykle uważa się teorię systemów (dalej t.s.). Jak zauważa R. Köhler, jeden z reprezentantów t.s. w lingwistyce: „Systems theoretical concepts and methodology were not yet available at this time, but a reinterpretation of Zipf's notions and results in terms of modern terminology yields surprisingly up-to-date models of linguistic mechanisms." (KÖHLER&MARTINÁKOVÁ-REDENKOVÁ 1998:514). Dalej autorzy stwierdzają: „The central axiom of this approach, i.e. the assumption that language is a self-organising system, and other basic principles turned out to be at least compatible with – if not very similar to – concepts of synergetics." (*ibid.* 516). Termin *synergetyka* jest używany w literaturze anglojęzycznej jako synonim jednej ze szcze-

[27] Jednym z takich projektów jest WordNet (`http://www.cogsci.princeton.edu/~wn/`).
[28] Podstawą omówienia były prace ALTMANN&SCHWIBBE 1989, GUITER&ARAPOV 1982, HAMMERL &SAMBOR 1990 (350 i 356) oraz HAMMERL&SAMBOR 1993b.

gółowych teorii systemów. W powyższym cytacie odnosi się on do szczegółowej teorii systemów w fizyce, której twórcą jest H. Haken (HAKEN 1978).

Zważywszy, że praca niniejsza nie ma charakteru teoretyczno-filozoficznego, poprzestaniemy tu na krótkiej charakterystyce t.s. Badania systemowe zapoczątkował jeszcze w latach 30. L. von Bertalanffy, autor wielu publikacji na temat t.s., jeden z twórców i propagator koncepcji systemu otwartego. Właśnie Bertalanffy, wraz z cytowanym wcześniej A. Rapoportem oraz innymi współpracownikami, założył w 1954 r. *Society for the Advancement of General Systems Theory*, które stało się podstawą instytucjonalizacji badań nad teorią systemów i zaczątkiem rozwoju różnych szkół w ramach tego kierunku.

Traktując teorię systemów w perspektywie typologicznej, W. Gasparski wyróżnia dwa podejścia: „Teoria systemów *sensu stricto* to nazwa każdego zespołu zdań spełniającego metodologiczne kryteria poprawności, który dotyczy *systemów* w jednym z rozumień tego terminu. Teoria systemów *sensu largo* natomiast, to nazwa klasy rezultatów różnych przedsięwzięć badawczych [...] – metodologicznych, aplikacyjnych itd., dotyczących *systemów* w jednym lub wielu rozumieniach tego terminu." (GASPARSKI 1987:698). Autor stwierdza dalej, że „kombinacja dwu rozumień teorii systemów, a także wielu rozumień terminu «system» [...] składa się na współczesną, obejmującą wszystkie pozostałe, teorię systemów *sensu largo*, nazwaną też polifoniczną teorią systemów." (*ibid.*) Zgodnie z przyjętą przez Gasparskiego klasyfikacją, wymienione wyżej prawa należałyby do szczegółowej teorii systemów lingwistycznych, a jako izomorficzne z prawami szczegółowej teorii systemów biologicznych i społecznych, pretendowałyby do statusu praw uogólnionej teorii systemów.

Paradygmatu teorii systemów w lingwistyce modelowej nie należy jednak przyjmować w sposób całkowicie bezkrytyczny. Argumentem przemawiającym za ostrożnym podejściem do tej koncepcji jest, naszym zdaniem, historycznie uwarunkowana nieufność wielu współczesnych filozofów i lingwistów wobec teorii aspirujących do statusu całościowej prawdy. Jak zauważa A. Chmielewski: „Być może ma to związek z generalnym trendem w filozofii, w której Wielkie Teorie [...] o uniwersalistycznych ambicjach poddano fundamentalnej krytyce. Najogólniejsza z nauk doszła bowiem do wniosku, iż świat wymyka się jej z rąk i że nie da się już skonstruować wszechogarniającej, Jednolitej Teorii Wszystkiego, gwarantującej zbudowanie niezawodnej sieci teoretycznej, skutecznie chwytającej w jeden schemat pojęciowy najdrobniejsze cząstki i wszystkie aspekty Wszystkiego, a należy pamiętać, że ta właśnie ambicja stanowiła najsilniejszą inspirację najważniejszych systemów filozoficznych [...]" (CHMIELEWSKI 2000:41).

2.7 SEKWENCYJNA STRUKTURA TEKSTU A PRAWA JĘZYKOWE

Powyższa prezentacja lingwistyki modelowej oraz przegląd literatury przedmiotu wskazują, że w dotychczasowej praktyce badawczej QL dominowało traktowanie tekstu jako nieuporządkowanego liniowo rezerwuaru elementów (słowoform, leksemów, morfemów,

fonemów). Jednak wiele zależności obserwowanych w tekście posiada charakter wybitnie liniowy – redukowanie ich do nieuporządkowanych odcinków, na które cięta jest nieuchronnie sekwencja tekstowa, powoduje utratę relewantnej lingwistycznie informacji, obniżając efektywność hipotez, a w niektórych przypadkach uniemożliwiając wręcz ich wysunięcie. Sekwencyjnymi warstwami tekstu są, jak wcześnie wspomniano, prozodia, metryka czy struktura dyskursu traktowanego jako ciąg jednostek zdaniowych lub leksykalnych.

Podstawowym celem badań, których wyniki prezentujemy w dalszych rozdziałach, była weryfikacja pewnej liczby hipotez szczegółowych dotyczących różnych poziomów sekwencyjnej struktury tekstu i uznanych za efektywne pod względem zakresu oraz mocy eksplanacyjnej. Zależało nam też na ukazaniu wszechstronnych możliwości analizy sekwencyjnej, czemu służyć miało zróżnicowanie problematyki pod względem treści, języka i poziomu analizy. Na tym etapie badań uzyskane wyniki i ich uogólnienia nie powinny być jednak traktowane jako prawa, lecz jedynie jako zweryfikowane empirycznie i pretendujące do tego miana prawidłowości. Określanie mianem prawa każdej, nawet słabo zweryfikowanej zależności, uważamy za nadużycie – kryteria epistemologiczne są w tym względzie zupełnie jasne (KRAJEWSKI 1998:12–26).

3. POJĘCIE SEKWENCYJNEJ ANALIZY TEKSTU

Na sekwencyjną strukturę tekstu składa się całość relacji zależnych od liniowego uporządkowania w tekście jednostek fonetycznych (fonologicznych), morfologicznych, leksykalnych i składniowych. Tak definiowana struktura obejmuje więc relacje zachodzące praktycznie na wszystkich poziomach analizy lingwistycznej. Jednak z punktu widzenia lingwistyki kwantytatywnej poziomami najważniejszymi są te, które poddają się wyrazistej i sensownej segmentacji. Jako przykłady można wymienić warstwę fonetyczną i leksykalną oraz segmentacje względem nich pochodne (tekst pojmowany jako szereg akapitów, zdań, sylab, stóp metrycznych, tonów itp.). Przykładem negatywnym jest warstwa znaczeniowa, którą trudno jest zredukować do reprezentatywnej sekwencji liczb czy symboli.

Jak widać z powyższego, o sekwencyjnej analizie tekstu w lingwistyce kwantytatywnej mówić będziemy wówczas, gdy za relewantną i podlegającą matematycznemu modelowaniu cechę tekstu uznamy porządek tworzących go jednostek. Analiza sekwencyjna w lingwistyce może więc być traktowana jako dopełnienie tych klasycznych metod statystyki matematycznej, które zakładają niezależność danych. Szczegółowe hipotezy badawcze odnoszące się do sekwencyjnej budowy tekstu można sprowadzić do następującego stwierdzenia: **linearny porządek niektórych jednostek językowych w tekście stanowi realizację jakiegoś procesu stochastycznego i z tego względu nie ma charakteru losowego**. Ograniczony determinizm w sekwencji tekstowej może być skutkiem działania przesłanek lingwistycznych (na przykład reguł składni danego języka) lub psychologicznych (na przykład zasady najmniejszego wysiłku lub sił Zipfa). Jeżeli mechanizm

generujący składową deterministyczną zaobserwowanego procesu stochastycznego zostaje odkryty i wyjaśniony, a sam proces opisany modelem formalnym zweryfikowanym na obszernym materiale, można mówić o ilościowym prawie językowym.

Istotę sekwencyjnej analizy tekstu można wyjaśnić posługując się następującym przykładem. Załóżmy, że w badanym języku występują dwa rodzaje jednostek określonych jako A i B, tworzących struktury liniowe („teksty") w oparciu o pewne zasady składni. Załóżmy dalej, że dysponujemy następującym „korpusem tekstów":

(1) AAAAAABBBBBB, (2) AAABBBAAABBB, (3) AABBAABBAABB,
(4) ABABABABABAB, (5) AABBABABABAB, (6) BABBAABBBAAA

Tradycyjne miary statystyczne (parametry pozycyjne i rozkład statystyczny) nie wykażą różnic pomiędzy przedstawionymi wyżej szeregami symboli, ponieważ częstości jednostek A i B są w każdym przypadku identyczne. Jednak z językowego punktu widzenia sekwencje te różnią się w sposób zasadniczy, ponieważ różna jest kolejność tworzących je elementów. Właśnie w takich przypadkach zastosowanie znajdują metody sekwencyjnego modelowania tekstu umożliwiające rozróżnienie i efektywny opis tego rodzaju sekwencji. Opisany tu przykład jest oczywiście skrajnym uproszczeniem symulującym nader skomplikowaną strukturę języka naturalnego, gdzie zamiast dwóch, istnieje dowolna liczba jednostek, dostępne korpusy w praktyce nie są ograniczone pod względem długości, a podobieństwo miar pozycyjnych ma charakter przybliżony. Szukając analogii w języku naturalnym, wskazać można zjawiska sytuujące się na różnych poziomach analizy. Na przykład w warstwie leksykalnej spotyka się fragmenty tekstu różne pod względem treści i formy, a przy tym posiadające zbliżoną pod względem ilościowym strukturę słownictwa. Przy analizie prozodii utworów wierszowanych napotkać można różne odcinki tekstu, złożone jednak z podobnej lub tej samej liczby różnie uporządkowanych sylab nacechowanych i nienacechowanych[29]. Do podobnych wniosków prowadzi też analiza rytmu prozy generowanego przez swoiste linearne uporządkowanie jednostek językowych.

4. LINEARNOŚĆ TEKSTU W JĘZYKOZNAWSTWIE NIEKWANTYTATYWNYM

Jak już wcześniej wspomniano, w dotychczasowej praktyce lingwistyki kwantytatywnej nie uwzględniano w należytym stopniu linearnego charakteru języka. Nie znaczy to jednak, że zjawisko to nie było w ogóle zauważane. Chronologicznie, refleksja nad linearnością języka najwcześniej pojawiła się w teorii retoryki. Zasadnicza struktura dyskursu retorycznego, uformowana jeszcze w antyku, obejmuje pięć części: *inventio*, *dispositio* albo *distributio*, *elocutio*, *memoria* i *pronunciatio*. Na etapie kompozycji (*dispositio*), elementy językowe uszeregowane zostają w określonej, nieprzypadkowej kolejności. Retor może pozostać przy porządku naturalnym (*ordo naturalis*), zgodnym z ogólną normą kulturowo-cywilizacyjną, lub też zmienić kolejność elementów dyskursu (*ordo*

[29] Nacechowanie powstaje poprzez zróżnicowanie iloczasu, akcentu dynamicznego i/lub tonu sylaby.

artificialis), w celu osiągnięcia efektu perswazyjnego lub estetycznego (KOROLKO 1990:78). Linearność tekstu pojawia się także na etapie wymowy (*pronunciatio*). O sztuce wymowy Arystoteles mówi, iż „Polega ona [...] na mówieniu głośnym, cichym i pośrednim, na używaniu wysokiego, niskiego i pośredniego tonu, na używaniu rozmaitych rytmów [...]" (*Retoryka* 1403b)[30]. Przywołana powyższym cytatem kwestia rytmu prozy rozszerza zakres teoretycznych rozważań odnoszących się do linearności tekstu. Ten sam filozof stwierdza, iż: „Tekst prozy nie powinien mieć metrycznej formy wiersza, ani też nie powinien być pozbawiony rytmu. Forma metryczna nie budzi wiarygodności, ponieważ wydaje się sztuczna i rozprasza uwagę słuchacza. [...] Z drugiej strony tekst, który nie ma rytmu, nie ma też żadnych ograniczeń, a powinien je przecież posiadać, ale nie wyznaczone miarą wiersza. To bowiem, co nie ma granic, nie sprawia przyjemności i jest trudne do zrozumienia. Granicę wszystkim rzeczom wyznacza przecież liczba, a tą liczbą dla formy językowej jest właśnie rytm, podczas gdy miary wierszowe są tylko jej odcinkami. Dlatego proza musi posiadać rytm, nie może natomiast posiadać miar wierszowych, bo zamieni się w poezję. Nie wolno jednak stosować rytmu rygorystycznie, lecz w ograniczonym zakresie." (*ibid.*). Dla sekwencyjnej analizy tekstu z tego obszernego cytatu płynie wniosek, iż ze względu na poziom liniowego uporządkowania jednostek wyznaczających rytm tekstu dyskurs oratorski powinien zajmować pozycję pośrednią pomiędzy tekstem prozatorskim a wierszem. Przedstawione dalej wyniki zweryfikują słuszność tego postulatu także w odniesieniu do języków nowożytnych. Nasze wnioskowanie zasadza się wszakże na założeniu, iż najważniejsze terminy użyte w cytatach, a więc proza, wiersz, rytm, oddają sens tego, co dla Greków znaczyły przeszło dwa tysiące lat temu.

Zasada linearności była też elementem antycznej teorii wiersza przejętej przez poetyki ery nowożytnej. W strukturze wiersza wyróżnia się pewną liczbę uporządkowanych sekwencyjnie wzorców rytmicznych, takich jak stopy, człony wersowe, wersy lub strofy, złożonych z jednostek metrycznych bądź akcentowych. Ich uszeregowanie w linii tekstu ma przy lekturze dłuższych fragmentów wywołać u słuchacza bądź czytelnika wrażenie rytmu, ułatwia też zapamiętywanie dłuższych fragmentów. Fakt, iż jednostki metryczne poddają się stosunkowo łatwej segmentacji i kwantyfikacji, czyni rytmikę tekstu idealnym obszarem analiz sekwencyjnych. Nie jest zapewne przypadkiem, iż matematyczna koncepcja łańcuchów Markowa leżąca u podstaw teorii procesów stochastycznych pierwotnie inspirowana była linearną strukturą tekstu i testowana na fragmentach poematu *Eugeniusz Oniegin* A. Puszkina (MARKOW 1913, PETRUSZEWYCZ 1981). Niestety, wiele kwantytatywnych studiów wersyfikacji powtarza schemat polegający na określeniu charakterystycznych dla konkretnego utworu krótkich wzorców rytmicznych (na przykład stóp albo wersów) i badaniu częstości ich występowania. Na tej podstawie określa się następnie rozkłady empiryczne i teoretyczne, testuje hipotezy i wyciąga wnioski. Ujęcie takie ignoruje jednak liniowe uporządkowanie tekstu jako całości i z punktu

[30] *Retorykę* Arystotelesa cytujemy w przekładzie H. Podbielskiego (PODBIELSKI 1988).

widzenia analizy sekwencyjnej nie jest ani poznawczo, ani metodologicznie satysfakcjonujące[31].

Linearność języka widoczna jest także w budowie gramatycznej zdań. Od dawna gramatycy i językoznawcy zwracali uwagę na istnienie składni pozycyjnej, traktując szyk wyrazów w zdaniu jako nie wyrażoną bezpośrednio („I mean the unexpressed element in language" – BREAL 1991:169), ale relewantną z językowego punktu widzenia cechę: „This positional value exists more or less in all languages, and especially in modern languages." (*ibid*. 169).[32] Wątek ten kontynuowany jest dziś w pracach z zakresu składni, a także typologii, gdzie jako kryterium klasyfikacji przyjmuje się porządek członów składniowych w zdaniu (GREENBERG 1960, SIEWIERSKA 1988, 1997). Prace te mogą mieć dla nas jedynie charakter pomocniczy. Należy bowiem pamiętać, że reguły składniowo-gramatyczne organizują strukturę pojedynczych zdań, a nie całego tekstu. Natomiast efektywność metod analizy sekwencyjnej wynika z faktu, iż pozwalają one formułować wiarygodne uogólnienia na podstawie szeregów dowolnej długości, przekraczającej ramy zdania czy akapitu.

W sposób systematyczny i wyczerpujący kwestię linearności języka ujął dopiero strukturalizm. De Saussure opisał strukturę języka za pomocą serii prostych opozycji, zauważając przy tym, iż „Signifiant, z natury swojej słuchowy, rozwija się wyłącznie w czasie i posiada cechy, które zapożycza od czasu: a) przedstawia pewną rozciągłość i b) rozciągłość ta jest wymierna w jednym tylko kierunku: jest to linia. [...] Zasada ta jest oczywista, jak się jednak wydaje, nigdy nie starano się o jej sformułowanie, z pewnością dlatego, że wydawała się zbyt prosta; a przecież jest ona podstawowa, a jej następstwa sięgają bardzo daleko; doniosłość jej dorównuje znaczeniu pierwszego prawa [arbitralności znaku językowego – A.P.]." (SAUSSURE 1991:93–94). Genewski językoznawca wprowadził także pojęcie osi syntagmatycznej, analogicznej do osi czasu i odpowiadającej linii tekstu: „Z jednej strony, w każdej wypowiedzi wyrazy nawiązują między sobą – na mocy swego następstwa – stosunki oparte na charakterze linearnym języka, który wyklucza możliwość wymówienia dwóch elementów równocześnie. Elementy te szeregują się jeden za drugim w ciągu mówienia." (*ibid.* 147).

Przedstawiony tu związek między osią syntagmatyczną a osią czasu jest z metodologicznego punktu widzenia kluczowym elementem analizy sekwencyjnej, pozwala bowiem na zastąpienie „momentu" na osi czasu „pozycją" w linii tekstu. W szerszej perspektywie umożliwia natomiast bezpieczne przeniesienie na grunt lingwistyki kwantytatywnej technik matematycznych stosowanych przy dyskretnym modelowaniu zjawisk fizycznych i ekonomicznych o charakterze linearnym (ang. *longitudinal data*), gdzie zmienną niezależną jest właśnie czas. Jak zauważa J. Lyons: „W przypadku języków naturalnych, uporządkowanie składników rządka od lewej do prawej można traktować

[31] Literatura o tej tematyce jest obszerna (por. KÖHLER 1995). Przykładowe prace to KOŁMOGOROW 1965, KOŁMOGOROW&PROCHOROW 1963, KONDRATOW 1963 i 1965, LEVÝ 1965, SCHMIEL 1981, VASJUTOČKIN 1987, WORONCZAK 1965.

[32] Odnośniki do oryginalnych tekstów Bréala, znajdują się w cytowanej tu częściowej reedycji jego prac.

jako odzwierciedlenie porządku chronologicznego wypowiedzi w większości języków na świecie. Jednocześnie trzeba zdać sobie sprawę z tego, że ta sama abstrakcyjna zasada uporządkowania linearnego mogłaby być użyta w opisie języka także dla innych celów." (LYONS 1976:234). Współczesne metody analizy sekwencyjnej (przede wszystkim przedstawiona dalej analiza szeregów czasowych) opierają się na takiej właśnie koncepcji czasu. Pierwszym, który powiązał oś czasu występującą w modelach statystycznych z linearnym uporządkowaniem tekstu, był najprawdopodobniej N. Wiener: „The message is a discrete or continuous sequence of measurable events distributed in time – precisely what is called a time series by statisticians." (WIENER 1948:8). Definicje tekstu tworzone na potrzeby analizy sekwencyjnej w kontekście różnych metodologii badawczych nie odbiegają w istotny sposób od definicji strukturalistycznej, chociaż pojęcie szeregu czasowego może być zastąpione pojęciem procesu stochastycznego: „If the event 'W_i has property A' is abbreviated $W_i \in A$, then the text $W_1 W_2 ... W_n$ can be conceived to be a result of a sequence of choices governed by probabilistic mechanism usually called a stochastic process and is called a sample sequence of that process." (BRAINERD 1976:5).

W strukturalistycznej koncepcji języka mieściła się też analiza cech suprasegmentalnych o charakterze sekwencyjnym. Badanie zjawiska prozodii znalazło się między innymi w obszarze zainteresowań tzw. Szkoły Londyńskiej[33]. Za B. Malinowskim, jej przedstawiciele postulowali traktowanie kontekstu jako kluczowego pojęcia analizy lingwistycznej (dlatego kierunek ten określa się niekiedy jako kontekstualizm). Postulat ten odnosił się nie tylko do sytuacji komunikacyjnej (kontekst sytuacyjny) i kulturowej (kontekst kulturowy), ale także do otoczenia jednostek językowych w tekście lub ciągu mowy (kontekst wypowiedzeniowy) (FIRTH 1957:35–39). Założyciel szkoły, J.R. Firth, rozszerzył repertuar jednostek fonetycznych posiadających cechy dystynktywne o suprasegmentalne jednostki prozodyczne (tzw. prozodie), zaliczając do tej grupy sekwencje fonemów tworzących sylaby oraz grupy rytmiczne i tonalne (FIRTH 1957:121–138, PALMER 1970). Jak z powyższego widać, podobnie jak przy analizie wersyfikacji, również i w tym przypadku wystąpił schemat polegający na wyróżnieniu skończonej liczby stosunkowo krótkich, linearnych sekwencji sylabicznych, przy jednoczesnym ignorowaniu sekwencyjnego uporządkowania całości tekstu.

Kolejność jednostek językowych traktowana jako relewantna cecha tekstu pojawia się również w pracach lingwistów Szkoły Praskiej, zainteresowanych informacyjną strukturą zdania (MATHESIUS 1939, MATHESIUS 1982:128–133 i 174–178, DANEŠ 1974, EROMS 1986). Twórcy tej koncepcji, określanej jako funkcjonalna perspektywa zdania, nie dokonywali segmentacji zdań według tradycyjnych kryteriów fonologicznych, morfologicznych czy składniowych. Posługiwali się kryterium informacyjnym, wyróżniając w zdaniu (wypowiedzeniu) część znaną uczestnikom aktu komunikacji (temat) i część

[33] „Terminem *szkoła londyńska* określa się zazwyczaj językoznawstwo J.R. Firtha (1890-1960) oraz grupy jego uczniów i współpracowników ze School of Oriental and African Studies Uniwersytetu Londyńskiego." (FISIAK 1985:49).

nieznaną odbiorcy komunikatu (remat). W językach o luźnym szyku słów kolejność tych jednostek nie jest ustalona raz na zawsze i zależy od funkcji wypowiedzenia oraz od kontekstu sytuacyjnego. Problematykę informacyjnej struktury dyskursu rozwijają obecnie kierunki zaliczane do szeroko rozumianej lingwistyki tekstu i/lub pragmatyki, na przykład analiza konwersacyjna (SCHENKEIN 1978, KALLMEYER&SCHÜTZE 1976, DITTMANN 1979, ATKINSON&HERITAGE 1984) oraz analiza dyskursu (DUSZAK 1998, ADAM 1992, DIJK 1980). Choć nurty te sytuują się na obrzeżach „lingwistyki immanentnej", w pewnym sensie stanowią jedynie rozciągnięcie osi syntagmatycznej poza granice zdania pojmowanego tradycyjnie jako zamknięty składniowo, znaczeniowo i intonacyjnie układ elementów. Jak zauważa E. Benveniste, „Nous en concluons qu'avec la phrase on quitte le domaine de la langue comme système de signes, et l'on entre dans un autre univers, celui de la langue comme instrument de communication, dont l'expression est le discours." (BENVENISTE 1966:129–130). Ich znaczenie dla analizy sekwencyjnej wynika z uwzględnienia elementów semantyki tekstu, między innymi poprzez rozróżnienie tematu i rematu. Jednak to właśnie semantyka, z całym swym ładunkiem wieloznaczności i subiektywizmu, utrudnia ilościowe ujęcie struktury dyskursu. Z tego względu prace wywodzące się z tego nurtu mają dla badań kwantytatywnych znaczenie raczej drugorzędne. Warto jednak pamiętać, że przy jasno określonych kryteriach segmentacji i kwantyfikacji stać się mogą źródłem wartościowych hipotez.

Jak z powyższego wynika, szkoły strukturalistyczne akceptowały pogląd, zgodnie z którym język jest strukturą hierarchicznie uporządkowanych elementów, a jednostki kolejnych poziomów językowych mogą być definiowane jako ciągi liniowo uporządkowanych jednostek poziomu niższego – morfem jako sekwencja fonemów o określonych cechach, zdanie jako sekwencja fonemów, morfemów lub słowoform, dyskurs jako sekwencja zdań, słowoform itd. Operowanie takimi definicjami ma dla analizy sekwencyjnej istotne znaczenie, pozwala bowiem wybrać najwłaściwszą formę kwantyfikacji i segmentacji tekstu.

Wbrew pozorom, wyjątkiem od tej reguły nie są fonologiczne koncepcje przedstawicieli Szkoły Praskiej – R. Jakobsona i M. Trubeckiego – definiujących fonem jako wiązkę współbieżnie (a więc nie sekwencyjnie) występujących cech dystynktywnych (JAKOBSON 1962:231, TRUBETZKOY 1939:38). Także w tym przypadku stosuje się modelowanie sekwencyjne, z tym jednak, że zamiast pojedynczego szeregu czasowego analizuje się tzw. szereg wielokrotny (ang. *multiple time series*), złożony z kilku równoległych szeregów binarnych, z których każdy odpowiada jednej cesze dystynktywnej. Poniższy przykład (Tab. 3) przedstawia niemieckie słowo *Beispiel* zakodowane według tej zasady (KÖHLER 1983:161). Pozycjom lub momentom t_i odpowiadają tu 13-wymiarowe wektory binarne.

Podsumowując tę część naszego omówienia, należy stwierdzić, że aparat pojęciowy stosowany dziś w modelowaniu sekwencyjnej struktury tekstu przygotowało językoznawstwo strukturalistyczne. Inne działy szeroko pojmowanej nauki o języku (teoria retoryki, wersologia, analiza składni, analiza dyskursu) jawią się z dalszej perspektywy przede

wszystkim jako obszary zastosowań dostarczające ciekawego materiału badawczego, ale pod względem metodologicznym nie wnoszące niczego istotnego. O ile jednak użycie metod kwantytatywnych w tej dziedzinie może być traktowane jako rozwinięcie pewnych koncepcji de Saussure'a, zawartych w *Cours de linguistique générale*, należy pamiętać,

Tab. 3 Słowo *Beispiel* kodowane binarnie według cech dystynktywnych[34]

	son	voc	cns	ant	cor	hgh	low	bck	rnd	lat	lng	cnt	tns	oś czasu
b	[0	0	1	1	0	0	0	0	0	0	0	0	0]	t_1
a	[1	1	0	0	0	0	1	0	0	0	0	1	0]	t_2
i	[1	1	0	0	0	1	0	0	0	0	0	1	0]	t_3
ʃ	[0	0	1	0	1	0	0	0	0	0	0	1	1]	t_4
p	[0	0	1	1	0	0	0	0	0	0	0	0	1]	t_5
i:	[1	1	0	0	0	1	0	0	0	0	1	1	1]	t_6
l	[1	1	1	0	1	0	0	0	0	1	0	1	0]	t_7

że oprócz ogólnych podstaw, koniecznym warunkiem podjęcia tej tematyki było wypracowanie adekwatnej, matematyczno-statystycznej metodologii. Strukturalizm jako teoria *stricte* lingwistyczna warunku tego nie spełniał. Faktyczna możliwość rozpoczęcia kwantytatywnych badań sekwencyjnych struktur języka pojawiła się dopiero w drugiej połowie XX w., kiedy do kanonu metod statystycznych weszła teoria informacji C. Shannona, a lingwiści odkryli, po upływie dziesięcioleci, publikowane w latach 1910–1920 prace matematyka rosyjskiego A.A. Markowa, w których wykorzystywano w formie przykładów materiał językowy. Pod koniec lat 60. zaczęły się też pojawiać lingwistyczne zastosowania analizy widmowej stosowanej jednak przede wszystkim w fonetyce akustycznej. Ogniwem łączącym fundamentalne zdobycze strukturalizmu ze współczesną metodologią badań kwantytatywnych są prace brytyjskiego lingwisty G. Herdana.

5. Linearność tekstu w badaniach kwantytatywnych

Opozycje osi syntagmatycznej i paradygmatycznej oraz *langue* i *parole* próbował przenieść na grunt lingwistyki kwantytatywnej G. Herdan. Stwierdził on wprost, iż „ [...] de Saussure's *Cours de linguistique générale* was stressed, the work (*Type-Token Mathematics* – A.P.) being described as a quantification of de Saussure's fundamental concepts and distinctions." (HERDAN 1960:17). Jego główną zasługą jest jasne rozróżnienie w ana-

[34] Objaśnienie skrótów: *son* – sonorny, *voc* – wokaliczny, *cns* – spółgłoskowy, *ant* – przedni, *cor* – koronalny, *hgh* – wysoki, *low* – niski, *bck* – tylny, *rnd* – zaokrąglony, *lat* – boczny, *lng* – długi, *cnt* – trwały, *tns* – napięty. Przykład binarnego kodowania fonemów podaje także W. Jassem (1974:138–139).

lizie kwantytatywnej dwóch komplementarnych podejść do języka, opartych na wymienionych wyżej opozycjach. Podejście określane jako *analysis in the mass* zostało przeciwstawione podejściu *analysis in the line*: „In the area of language, it is the dimension of time which may have to be taken into consideration. We may deal with language in the mass, or with language in the line. In the former case, frequencies of, say phonemes, are established by phoneme counts regardless of their sequence in the morphs and chains of morphs (running texts). The statistics on the phonemic and alphabetic level are, by and large, those of the conventional type. [...] However, Shannon's Information Theory, rightly understood, represents already the introduction of the time element on the alphabetic (phonemic) level, since Entropy is calculated from the number of possible arrangements of elementary units in the line." (HERDAN 1966:423). W pierwszym przypadku (*language in the mass*), kolejność jednostek językowych w próbie nie jest brana pod uwagę, natomiast w przypadku drugim (*language in the line*) jest ona cechą relewantną.

Pod względem metodologicznym wkład Herdana w syntagmatyczną analizę tekstu jest skromniejszy i kontrastuje z jego osiągnięciami w zastosowaniach statystyki konwencjonalnej (PAWŁOWSKI 1998:50–53). Jedynym narzędziem badawczym, zaproponowanym przez niego dla analizy języka „w linii", jest shannonowska teoria informacji. Zupełnym milczeniem pominięto natomiast analizę spektralną, znaną matematykom już w początkach XIX wieku, oraz różne techniki analizy szeregów czasowych, stosowane między innymi w ekonometrii[35]. Dziwić też może fakt, że autor, który jako pierwszy zamieścił w zachodniej literaturze lingwistycznej omówienie niektórych prac Markowa (HERDAN 1960:140–153), nie powiązał ich z sekwencyjną analizą języka.

Rozróżnienie przez Herdana relacji „zbiorowościowych", nie uwzględniających porządku elementów w tekście, i sekwencyjnych, traktujących tę właściwość jako cechę relewantną, okazało się poznawczo wartościowe. Jednak analizując tę kwestię z dzisiejszej perspektywy, zauważyć można, że dominacja strukturalizmu w koncepcji Herdana spowodowała bezkrytyczne przeniesienie opozycji de Saussure'a na znacznie szerszą dziedzinę pojęciową. Koncepcja dwóch, jakoby komplementarnych perspektyw badawczych w lingwistyce kwantytatywnej, spójna z paradygmatem strukturalistycznym, obecnie wydaje się niepełna. Metody matematyczne pozwalają bowiem na modelowanie rozległych struktur sieciowych (przykładem jest leksyka), w których uporządkowania hierarchiczne i/lub sekwencyjne nie są relewantne, a mimo to wszystkie elementy układu pozostają w pośredniej lub bezpośredniej zależności.

Próbę systematycznego ujęcia analizy sekwencyjnej w całości zagadnień lingwistyki kwantytatywnej podjął też M. Dillon (1970). Opierając się na danych z listy abstraktów lingwistycznych LLBA[36], dokonał on klasyfikacji tematów prac materiałowych, wyróż-

[35] Koncepcję rozkładu dowolnej funkcji na sumę prostych funkcji trygonometrycznych J. Fourier przedstawił w pracy *Théorie analytique de la chaleur*, wydanej w 1822 r. Jedną z klasycznych pozycji opisujących obszernie teorię szeregów czasowych i analizy harmonicznej jest praca E. Morice'a i F. Chartiera (1954:423–508), dostępna w okresie naukowej działalności Herdana.

[36] Ang. *Linguistics and Language Behavior Abstracts*.

niając kilka poziomów analizy (między innymi fonetykę z fonologią, semantykę, składnię, analizę dyskursu). Przedstawił następnie listę szesnastu jednostek badawczych o coraz większym stopniu skomplikowania (fonemy, morfemy, zdania, akapity itd.). Z punktu widzenia analizy sekwencyjnej, najbardziej istotna jest lista „modeli analitycznych" (ang. *analytical models*), obejmująca nie tyle modele, ile najczęściej w lingwistyce kwantytatywnej spotykane podejścia metodologiczne (Tab. 4). Trudno nie dostrzec pewnych nieścisłości w schemacie Dillona. Nieostre jest na przykład pojmowanie pojęcia modelu – termin ten ma tu zapewne oznaczać ogólną ramę teoretyczną albo perspektywę badawczą. Trudno też orzec, czym w istocie różnią się „zależności sekwencyjne" od „miar liniowych" i „zależności strukturalnych", szczególnie, że entropia (tu wymieniana jako jedna z miar zależności strukturalnych) daje się sprowadzić do modelu Markowa, wymienionego jako metodologiczne narzędzie w drugim punkcie omawianego schematu. Jednak fakt rozróżnienia na równych zasadach miar sekwencyjnych (2, 3 i 4) i „zbiorowościowych" (1 i 5) potwierdza słuszność wybranej przez nas orientacji metodologicznej.

Tab. 4 Podejścia badawcze w lingwistyce kwantytatywnej (por. DILLON 1970:194)

Rodzaj badanej zależności	Narzędzia statystyczne
1. Miary opisowe	rozkłady częstości wybranych jednostek
2. Zależności sekwencyjne	teoria łańcuchów Markowa
3. Zależności strukturalne	teoria informacji
4. Miary liniowe	własności szeregowe tekstu (szeregi proste i kumulacyjne)
5. Klasyfikacje	grupowanie elementów jedną z metod analizy wielowymiarowej

Herdan i Dillon nie byli jedynymi badaczami wskazującymi na celowość kwantytatywnych studiów syntagmatycznej struktury języka. Postulat taki wysuwał R. Grotjahn: „Research in QL should therefore not stop with the analysis of the frequency of linguistic phenomena, but must take order into consideration." (GROTJAHN 1980:11).W programowym artykule pod tytułem *The art of quantitative linguistics* G. Altmann mówi o analizie sekwencyjnej, iż „The object of investigation in this domain is [...] everything that is positionally conditioned in text or changes in the course of text." (ALTMANN 1997:16). Podobne sugestie znaleźć można w pracach teoretyczno-metodologicznych luźno związanych z konkretnymi szkołami lingwistycznymi (SKINNER 1941, WILLIAMS 1970:105, LEVIN 1967), a także w szczegółowych studiach materiałowych, które zostaną omówione w dalszej części pracy. Przegląd literatury przedmiotu pozwala wskazać pięć filarów współczesnej analizy sekwencyjnej.

5.1 MIARY SPÓJNOŚCI TEKSTU

Spójność tekstu jest jednym z podstawowych pojęć lingwistyki tekstu, badającej na gruncie językoznawstwa teksty (dyskursy), a nie pojedyncze zdania (DRESSLER 1972, RICKHEIT 1991, STROHNER&RICKHEIT 1990, DUSZAK 1998). W sensie ogólnolingwistycznym tekst uważany jest za spójny, jeżeli stanowi semantyczno-składniową całość. Szczegółowe definicje spójności kładą nacisk na wybrane aspekty tekstu, na przykład jego składnię (mówi się więc o super- lub makroskładni – DUSZAK 1998:67), leksykę czy semantykę. Zjawisko to można definiować także za pomocą metajęzyka algebry i/lub logiki. Na przykład I. Bellert określa dyskurs (tekst spójny) jako „taki ciąg wypowiedzeń $S_1,...,S_n$, w którym interpretacja semantyczna każdego wypowiedzenia S_i (dla $2 \le i \le n$) jest zależna od interpretacji wypowiedzeń w ciągu $S_1,...,S_{i-1}$ (BELLERT 1971). Algebraiczną definicję spójności tekstu podaje także Z. Saloni (1971).

W lingwistyce kwantytatywnej wykładnikiem spójności jest rozkład w linii tekstu jednakowych bądź podobnych jednostek językowych (najczęściej leksemów). Zakłada się, że przy braku istnienia jakichkolwiek relacji syntagmatycznych (kontekstowych) badane jednostki miałyby rozkład równomierny, podczas gdy w rzeczywistości występują one w zbitkach. Układ taki ma wynikać z istnienia struktury tematycznej i/lub składniowej danego utworu oraz z uwarunkowań psychologicznych. Za miarę spójności uważa się różnicę pomiędzy spójnością tekstu obserwowanego a spójnością tekstu teoretycznego o maksymalnie równomiernym (bądź skupionym) rozkładzie elementów. Intuicyjnie (ale i matematycznie), kategoria spójności jest najbardziej przejrzysta w przypadku podziału jednostek tekstowych na dwie klasy.

Jedną z metod obliczania stopnia spójności tekstu przedstawił rosyjski lingwista Ju.I. Lewin (LEVIN 1967). Jego wskaźniki spełniają wszystkie algebraiczne wymogi stawiane tego typu wielkościom: znany jest między innymi przebieg funkcji, a jej minimum i maksimum są unormowane do jedności. Załóżmy, że wyróżnimy w tekście leksemy posiadające jakąś relewantną cechę i oznaczymy ich *i*-te wystąpienia jako α_i. Oznaczając wszystkie pozostałe leksemy przez β, dowolny tekst można zakodować jako: $\beta\beta...\alpha_1\beta\beta...\alpha_2\beta\beta...\alpha_n\beta\beta$. Parametrem pomiaru spójności tekstu jest w tym przypadku tu odległość d_i pomiędzy kolejnymi wystąpieniami wyróżnionych jednostek α_i. Tekst będzie maksymalnie spójny (symbole α_i maksymalnie skupione), jeżeli wszystkie d_i, oprócz jednego, będą równe zeru. Będzie tak w przypadku, gdy tekst składać się będzie ze zbitki wszystkich elementów α_i i drugiej zbitki wszystkich elementów β (ostatni wiersz w tabeli 5). Tekst będzie maksymalnie „rozproszony", jeżeli wszystkie d_i będą jednakowe, a więc jednostki α_i rozdzielać będzie za każdym razem ta sama liczba jednostek β (pierwszy wiersz tabeli 5). Opierając się na tym intuicyjnym pojmowaniu spójności, Lewin przedstawił kilka sposobów obliczania jej ilościowej miary, a w drugiej części pracy podał praktyczne przykłady ilustrujące zachowanie wskaźników. Za najlepsze uznał miary „zwykłą" i logarytmiczną. Wskazał także na związek tej ostatniej z pojęciem entropii.

Logarytmiczna miara spójności ma postać:

(13) $$L = 1 + \frac{\sum_{i=1}^{n} \delta_i \log \delta_i}{\log n}, \quad \text{gdzie } \delta_i = \frac{d_i}{m}$$

gdzie: m – liczba symboli β
n – liczba symboli α_i, których skupienie jest przedmiotem oceny
d_i – odległości pomiędzy kolejnymi symbolami α_i

Nielogarytmiczny wskaźnik spójności tekstu ma natomiast postać:

(14) $$Q = \frac{n\sum_{i=1}^{n} d_i^2 - m^2}{m^2(n-1)}$$ (oznaczenia jak wyżej)

Pomiar spójności stanie się pełny, jeżeli poda się dla niego punkty odniesienia w postaci wartości wskaźnika dla maksymalnego i/lub minimalnego skupienia jednostek, a także wartości uzyskane w przypadku całkowitego braku autokorelacji kolejnych wartości d_i („для дискретного случая это означает, что n экземпляров символа α случайным образом и независимо друг от друга вставляются в последовательность из m символов β” – *ibid.* 114). Wskaźniki Lewina są unormowane, tak więc rozkładom o maksymalnej i minimalnej dyspersji odpowiadają wartości 0 i 1. Wyprowadzono natomiast wzory miar dla losowego rozkładu symboli w linii tekstu. Oznaczając je odpowiednio jako $\overline{L}(n)$ i $\overline{Q}(n)$, Lewin otrzymał:

(15) $$\overline{L}(n) = 1 - \frac{\frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}}{\ln n}$$

(16) $$\overline{Q}(n) = \frac{1}{n+1}$$

We wzorach (15) i (16) nie występuje zmienna m. Wynika to z faktu, iż losowość rozkładu elementów α_i w linii tekstu nie zależy od bezwzględnej liczby rozdzielających je elementów β. Obliczenia wykonane przez Lewina na przykładowych rządkach symboli dały interesujące wyniki (Tab. 5). Na podstawie wzorów (15) i (16) za losowe należałoby uznać te sekwencje, dla których wskaźniki L i Q bliskie są wartościom $\overline{L}(12) = 0{,}15$ i $\overline{Q}(12) = 0{,}08$.

Metoda zaproponowana przez Lewina bliska jest analizie sekwencyjnej, ma jednak pewne usterki. Po pierwsze, zakłada ona wystarczalność binarnego podziału jednostek tekstowych, co w przypadku bardziej złożonych skal pomiaru (na przykład liczby bitów na fonem czy literę) nie jest rozwiązaniem praktycznym. Po drugie, sekwencyjny charakter języka jest tu uwzględniony jedynie w ograniczonym zakresie. Jest bowiem tak, że sekwencje o różnych stopniach spójności (w rozumieniu powyższej definicji) zawsze

posiadać będą różne parametry sekwencyjne, ale pewne rządki symboli, chociaż różne pod względem uporządkowania sekwencyjnego, charakteryzować się będą identycznym stopniem spójności. Jak stwierdza twórca metody: „Мы считаем существенным для оценки компактности лишь длины интервалав d_i, но не расположение этих интервалов [...]" (*ibid.* 113).

Tab. 5 Wskaźniki spójności dla wzorcowych sekwencji symboli

Sekwencja	*L*	*Q*
αβαβαβαβαβαβαβαβαβαβαβαβαβ	0,00	0,00
ααββααββααββααββααββααββ	0,28	0,09
αααβββαααβββαααβββαααβββ	0,44	0,18
αααaββββαααaββββαααaββββ	0,56	0,27
αααααaββββββαααααaββββββ	0,72	0,45
αααααααααααaββββββββββββ	1,00	1,00

Praca Lewina ma charakter wybitnie metodologiczny, ale pomimo swych walorów nie należy do często cytowanych – dlatego właśnie została tu przytoczona *in extenso*. Nie jest jednak jedyną publikacją tego rodzaju (por. STRAUSS et al. 1984, GROTJAHN 1980, ZÖRNIG 1984a i 1984b). Istnieje też pewna liczba studiów materiałowych poświęconych problemowi spójności tekstu (WORONCZAK 1976, VASJUTOČKIN 1987). Jedną z tych prac omawiamy poniżej.

J. Woronczak (1976) wykazał, że istnieje związek pomiędzy spójnością tekstu a wartościami tzw. miar Gooda (c_m), wyrażających prawdopodobieństwo wylosowania z populacji generalnej w *m* losowaniach niezależnych *m* elementów przynależnych do jednej i tej samej klasy (GOOD 1953):

$$c_m = \sum_i p_i^m \tag{17}$$

W kontekście lingwistycznym miary te mogą wyrażać leksykalne bogactwo (zróżnicowanie) tekstu, a ich zaletą jest niezależność od długości próby. Woronczak wyprowadził wzory estymatorów c_m dla $m = 2$ i $m = 3$ (WORONCZAK 1965):

$$\bar{c}_2 = \frac{\sum_i f_i^2 - N}{N^2 - N} \tag{18}$$

$$\bar{c}_3 = \frac{\sum_i f_i^3 - 3\sum_i f_1^2 + 2N}{N(N-1)(N-2)} \tag{19}$$

gdzie: f_i – frekwencja *i*-tej słowoformy
N – długość próby[37]

[37] Wzór (18) podał także G. Herdan. Obaj autorzy wskazali na podobieństwo c_2 i *charakterystyki K* Yule'a.

Uogólnieniem wzorów (18) i (19) jest formuła (20). Jej twórca nie zaleca jednak obliczania jej wartości dla $m > 3$ (WORONCZAK 1976):

$$(20) \qquad \bar{c}_m = \sum_i \frac{f_i(f_i - 1)...(f_i - m + 1)}{N(N-1)...(N-m+1)}$$

Badając dynamikę zmian średnich tych estymatorów, liczonych na coraz to dłuższych próbach tekstu ciągłego (dla $N = 2, 4, 8...$), zauważono, że przy wzroście N wartości estymatorów najpierw rosły, a następnie, mimo geometrycznego postępu N, stabilizowały się. Wartość N, przy której następuje względna stabilizacja wskaźników c_2 i c_3 (bliska ich wartości maksymalnej), wyznacza właśnie granicę spójności leksykalnej tekstu, wskazując tym samym na przeciętną długość odcinków do pewnego stopnia zamkniętych pod względem tematycznym i składniowym. Test przeprowadzony przez Woronczaka na fragmentach tekstów św. Augustyna i św. Fulgencjusza potwierdził tę hipotezę[38]. W przypadku tekstu Augustyna, adresowanego do niewykształconych warstw społeczeństwa i przez to napisanego stylem „łopatologicznym" (WORONCZAK 1976:171), wartość graniczna N przypadła na około 45 słowoform, natomiast graniczna długość N dla tekstu Fulgencjusza, trudniejszego i bardziej literackiego, wyniosła około 128 słowoform.

Warto w tym miejscu pokusić się o porównanie podejść Ju.I. Lewina i J. Woronczaka. Na pierwszy rzut oka obaj badacze zmierzyli się z tym samym problemem lingwistycznym, stosując po prostu inne modele funkcyjne. Jednak dokładniejsza analiza ich prac pokazuje, że wybór modeli matematycznych wynikał właśnie z różnych przesłanek lingwistycznych. Wymienimy najważniejsze z nich: o ile u Woronczaka uwzględnia się wszystkie klasy częstości słów, u Lewina wyróżnia się jedynie dwie klasy elementów; o ile u Lewina spójność tekstu oblicza się ze względu na rozkład elementów należących do konkretnej klasy we fragmencie dowolnej długości, metoda Woronczaka pozwala na zdefiniowanie granicznej długości rządka o maksymalnej spójności, dzięki czemu określa się wielkość swoistej „dyskursywnej porcji informacji". Ze względu na niewielką ilość przeprowadzonych testów trudno wartościować oba podejścia. Można natomiast z dużym prawdopodobieństwem twierdzić, że gdyby wskaźniki L i Q wyrażono jako funkcje długości badanego fragmentu (tak jak uczynił to ze swoim wskaźnikiem Woronczak), ich przebieg i maksima powiedziałyby więcej o spójności tekstu. Wolno też sądzić, że inne stabilne wskaźniki bogactwa leksykalnego[39], zastosowane w analogiczny sposób zamiast woronczakowych estymatorów (18) i (19), dałyby równie interesujące rezultaty.

Podsumowując powyższe rozważania, dorobek kwantytatywnych badań spójności tekstu należy uznać za istotny dla analizy sekwencyjnej. Badania te uwzględniają bowiem liniowy rozkład jednostek językowych i stosują konsekwentnie miary ilościowe. Nie jest

[38] Estymatory Woronczaka pozytywnie zweryfikowano także na tekstach polskich, m.in. na *Myślach* J. Leca.

[39] Listy i opisy takich wskaźników znaleźć można m.in. w pracach COSSETTE 1994, PAWŁOWSKI 1994, TWEEDIE&BAAYEN 1998.

też przypadkiem, że testy przeprowadzone przez Lewina pokrywają się co do zakresu i poziomu języka z hipotezami wysuwanymi w analizie sekwencyjnej (rozkład leksemów i długości zdań, rytmika tekstu). Jednak spójność tekstu nie jest tożsama z jego strukturą sekwencyjną. Oprócz statystycznego rozkładu odległości pomiędzy wyróżnionymi elementami tekstu ważna jest bowiem także ich kolejność. To sprawia, że mimo podobieństw, dokładna treść testowanych hipotez jest w obu przypadkach różna, a metody pomiaru spójności odgrywają w analizie sekwencyjnej jedynie rolę pomocniczą.

5.2 TEORIA INFORMACJI

Znaczący i niepodważalny wkład w poznanie sekwencyjnych struktur języka wniosła teoria informacji (SHANNON 1948, WEAVER&SHANNON 1949). Pojęcie entropii, stosowane od XIX wieku w termodynamice, a przez C. Shannona zdefiniowane na gruncie teorii informacji, wraz z pojęciem redundancji wyraża współzależność kolejnych jednostek językowych w linii tekstu i może w syntetyczny sposób opisać jego sekwencyjną strukturę. Jednak literatura z tego zakresu zdominowana jest problemami fonotaktyki[40], podczas gdy zastosowania teorii informacji do otwartych podsystemów języka (przede wszystkim leksyki) są mniej udane i nie tak liczne. Jest to zresztą zrozumiałe, skoro syntetyczną miarą sekwencyjnego uporządkowania tekstu nie jest entropia rzędu zerowego czy pierwszego, ale entropie rzędów wyższych, entropie *k*-gramów oraz redundancja[41]. Ich obliczanie metodą mechaniczną w systemie złożonym z kilkuset bądź kilku tysięcy różnych jednostek (a taką długość może osiągnąć lista leksemów, słowoform czy morfemów zawartych w typowym korpusie tekstów) jest, mówiąc eufemistycznie, kwestią nietrywialną. Badacze rozwiązują ten problem, posługując się metodą eksperymentalną, polegającą na tym, że pewna liczba respondentów czyta tekst poprzez ruchome „okno" o szerokości *k* i na tej podstawie odgaduje niewidoczny symbol (literę, fonem, leksem itd.) na pozycji *k*+1. Ilość informacji niesionej przez niewidoczny symbol obliczana jest dla każdej wartości *k* na podstawie średniej liczby pytań zadanych przez respondentów aż do jego odgadnięcia (HAMMERL&SAMBOR 1990:387). Natomiast posługując się metodą mechaniczną, entropie wyższych rzędów obliczyć można jedynie w przypadku podsystemów zamkniętych, złożonych z niewielkiej liczby elementów (na przykład listy fonemów bądź liter). Jednak nawet wówczas liczba obserwowanych w tekstach *k*-gramów jest bardzo wysoka i przeszkodą może być brak wystarczającej mocy obliczeniowej komputera. Dla alfabetu złożonego z N symboli, teoretyczna liczba *k*-gramów wynosi N^k. Jeżeli $N \approx 30$, nawet po eliminacji kombinacji niemożliwych (na przykład sekwencji *k* identycznych symboli), N^k będzie mieć bardzo wysoką wartość. Jak zauważa S. May, „Wyznaczanie dalszych kolejnych entropii H_n [dla liter alfabetu polskiego – A.P.] poz-

[40] „Zasadami określającymi następstwa kolejnych fonemów zajmuje się dziedzina fonematyki zwana fonotaktyką." (JASSEM 1974:201).

[41] Szczegółowy opis aparatu formalnego teorii informacji był przedmiotem wielu powszechnie dostępnych opracowań. Lingwistyczne aspekty teorii Shannona omówione zostały m.in. w pracach: MAY 1963, JASSEM 1974, HAMMERL&SAMBOR 1990:361–451, PAWŁOWSKI 1998:191–198, SZANIAWSKI 1987.

wala w zasadzie uzyskać ocenę entropii granicznej H_∞. Jednakże jest to proces niezwykle pracochłonny i dlatego znacznie korzystniej jest zastosować dla oszacowania entropii granicznej sposób Kołmogorowa, będący rozwinięciem shannonowskiej metody odgadywania kolejnych liter na podstawie znajomości liter poprzednich." (MAY 1963:370).

Fundamentem, na którym Shannon zbudował swoją teorię, wykorzystując szeroko dziś stosowane pojęcia entropii i redundancji, było sformalizowanie (kwantyfikacja) pojęcia informacji. Jeżeli empiryczne prawdopodobieństwo wystąpienia symbolu w tekście wynosi p_n, to ilość informacji I_n niesionej przez ten symbol wyniesie:

(21) $$I_n = -\log_2 p_n$$

Formalna definicja informacji opiera się więc na psychologicznej kategorii niepewności co do pojawienia się w linii tekstu kolejnego symbolu i nie może być automatycznie utożsamiana z jego treścią, chociaż, wbrew obiegowym opiniom, można tu mówić o pewnej korelacji obu pojęć. Im prawdopodobieństwo p_n jest mniejsze, tym ilość informacji niesionej przez dany symbol jest większa. Z kolei ilość informacji niesionej przez symbol całkowicie przewidywalny równa jest zeru ($\log_2 1 = 0$). Dzięki zastosowaniu funkcji logarytmicznej I_n posiada własność addytywności przy zachowaniu multiplikatywności p_n. Jest to istotne, ponieważ w rachunku prawdopodobieństwa współwystępowanie zdarzeń jest definiowane jako iloczyn, podczas gdy informacja jest intuicyjnie sumowana. Dwójkowa podstawa logarytmu sprawia, że $I_n = 1$ dla $p_n = 0{,}5$. Tę umowną jednostkę informacji określa się jako *bit*[42]. Jak wynika z powyższego, jeden bit jest ilością informacji niesioną przez pojedynczy symbol dwuelementowego alfabetu, w którym prawdopodobieństwa pojawienia się poszczególnych symboli są jednakowe i wynoszą 0,5.

W ilościowym opisie sekwencyjnej struktury tekstu szczególnie przydatne są pojęcia entropii (H) i redundancji (R). Entropię definiuje się jako miarę nieprzewidywalności (nieuporządkowania, chaotyczności) źródła informacji. Z formalnego punktu widzenia jest ona średnią ilością informacji niesionej przez symbol kodu. Dodatkowe warunki pozwalają określić różne rodzaje entropii. O entropii rzędu zerowego mówimy w przypadku, gdy wszystkie symbole kodu pojawiają się z jednakowymi prawdopodobieństwami (w dyscyplinie empirycznej przypadek taki należy uznać za czysto teoretyczny).

H_0 definiowana jest jako:

(22) $$H_0 = -\log_2 p$$

Jeżeli symbole N-elementowego kodu pojawiają się z różnymi prawdopodobieństwami p_i, mówimy o entropii rzędu pierwszego:

(23) $$H_1 = -\sum_{i=1}^{N} p_i \log_2 p_i$$

[42] Słowo *bit* utworzono poprzez kontrakcję nazwy angielskiej *binary digit*. Jego twórcą nie jest, jak niektórzy mniemają, C. Shannon, ale J.W. Tukey.

Rozważmy teraz kod o nierównomiernych prawdopodobieństwach, w którym dla każdego symbolu s_i znany jest lewostronny kontekst o długości *k*. W takim przypadku będziemy mówić o entropii warunkowej rzędu *k* (HAMMERL&SAMBOR 1990:375):

(24) $$H_k = -\sum_{i=1}^{K} p(r_i)\left(\sum_{j=1}^{N} p(s_j / r_i)\log_2 p(s_j / r_i)\right)$$

gdzie: s_j – *j*-ty symbol spośród *N* symboli kodu
r_i – *i*-ty *k*–1-gram poprzedzający symbol s_j
K – liczba *k*-1-gramów poprzedzających symbol s_j

Formułę (24) można uprościć, wykorzystując frekwencje *k*-gramów, łatwe do mechanicznego obliczenia (sposób ten zostanie omówiony dalej – por. Część I, 6.2). Można przewidzieć, że im dłuższy kontekst *k*, tym mniejsza nieprzewidywalność (a więc i entropia) nieznanego symbolu. Relację tę zapisuje się jako:

(25) $$H_0 \geq H_1 \geq H_2 \geq ... \geq H_k \geq ... \geq H_\infty$$

Pojęciem granicznym, definiowanym za pomocą wartości entropii, jest redundancja. Z nierówności (25) widać, że entropie H_k monotonicznie maleją. Jednak w rzeczywistości, dla pewnego *k* ich wartość stabilizuje się. Można odnieść ową wartość entropii rzędu *k* do teoretycznej, maksymalnej wartości H_0 i skonstruować wskaźnik wyrażający (w procentach) ich relację. Wskaźnik ten ma postać:

(26) $$R_k = 1 - \frac{H_k}{H_0}$$

R_k określane jest jako redundancja (nadmiar, rozwlekłość) rzędu *k*. Jeżeli pod H_k podstawi się wartość graniczną, po której następuje stabilizacja entropii (a więc będącą przybliżeniem nieznanej wartości H_∞), można mówić o całkowitej redundancji źródła informacji, oznaczanej przez *R*. Redundancję interpretuje się jako nadmiar informacji zawartej w sygnale (na przykład tekście) w stosunku do jej minimalnej ilości, która przy wyeliminowaniu z kanału komunikacyjnego wszystkich zakłóceń, pozwoliłaby na przekazanie tej samej wiadomości bez uszczuplenia jej treści.

Dla przeważającej liczby języków, które w przeszłości były przedmiotem kwantytatywnych badań lingwistycznych, podstawowe teorioinformacyjne parametry – wartości entropii i redundancji liter, fonemów, a niekiedy także morfemów i leksemów – zostały obliczone bądź oszacowane (HAMMERL&SAMBOR 1990:384–412). Tytułem przykładu, zacytujmy obliczone przez S. Maya entropie liter alfabetu języka polskiego: $H_0 = 1{,}52$, $H_1 = 1{,}30$, $H_2 = 0{,}98$ (MAY 1963:369). Z kolei L. Hoffmann i R.G. Piotrowski podają szacunkowe wartości H_∞ dla liter w kilku językach, a także stylach:

Tab. 6 Graniczne entropie (w bitach) liter w różnych językach i stylach[43]

	potoczny	beletrystyka	naukowy	ogólny
Niemiecki	0,74–1,24	0,83–1,36	0,56–0,97	0,71–1,36
Angielski	0,90–1,47	0,65–1,10	0,37–0,82	0,74–1,35
Rosyjski	0,83–1,10	0,70–1,19	0,49–0,83	0,82–1,37
Polski	0,69–1,18	0,83–1,29	0,53–0,83	0,76–1,28
Francuski	0,81–1,32	0,78–1,36	0,45–0,77	0,79–1,38
Rumuński	0,71–1,24	0,78–1,26	0,68–1,23	0,72–1,34

Jeżeli teraz entropię H_k potraktujemy jako funkcję zmiennej *k*, możliwe będzie określenie zakresu i dynamiki związków kontekstowych w tekście. O ile związki takie istnieją, stabilizacja wartości entropii, począwszy od pewnego *k*, wskaże maksymalny zasięg „pamięci" w tekście[44], natomiast kształt krzywej funkcji $H(k)$ określi jej dynamikę. Jak zauważają R. Hammerl i J. Sambor: „Można ją [redundancję – A.P.] ujmować jako miarę zależności strukturalnych między sygnałami, swoistą miarę stopnia zwartości systemu." (HAMMERL&SAMBOR 1990:377). Cytowani już wcześniej R.G. Piotrowski i L. Hoffmann zaproponowali nawet funkcyjny model zależności określającej tempo wzrostu ograniczeń kontekstowych w tekście. W modelu tym H_0 i H_∞ występują jako stałe (dla konkretnego przypadku), natomiast *s* jest współczynnikiem ograniczeń kontekstowych:

$$H(k) = (H_0 - H_\infty) \cdot e^{-sk} + H_\infty \tag{27}$$

Z perspektywy półwiecza widać, że teoria informacji odniosła w lingwistyce, zresztą nie tylko kwantytatywnej, wielki sukces, a jej terminy i pojęcia zadomowiły się także w estetyce, teorii literatury i filozofii (MOLES 1958, PORĘBSKI 1986, LEWICKI 1987, SZANIAWSKI 1987, ZIOMEK 1990:142–158). Swą niekwestionowaną skuteczność w analizie rozmaitych procesów komunikacyjnych zawdzięcza ona uniwersalności definicji informacji wprowadzonej przez C. Shannona. Dzięki temu ten sam aparat pojęciowy i matematyczny sprawdza się w opisie języków naturalnych i sztucznych, a także kodu genetycznego oraz wszelkich kodów semiotycznych – o ile tylko nada im się formę symboliczną i umożliwi obliczenie prawdopodobieństw *k*-gramów na wystarczająco dużym korpusie. Prymarnie symboliczny (a nie numeryczny) charakter ma właśnie tekst – przedmiot znakomitej większości analiz kwantytatywnych. Z punktu widzenia

[43] Tabela cytowana za pracą HAMMERL&SAMBOR 1990:399.

[44] Aby określić krytyczną wartość entropii, należy podać dla H_k przedział ufności. Przykład testowania hipotez statystycznych dla entropii *k*-gramów, wraz z określeniem ich rozkładu statystycznego i przedziałów ufności, podają J.M. Gottman (1990:46–49) i F. Bavaud (1998:215, por. też wzory 43 i 44 s. 59). Drugi z wymienionych autorów określa mianem entropii warunkowej (fr. *entropie conditionnelle*) entropie *k*-gramów, które nie są tożsame z entropią warunkową, obliczaną na podstawie prawdopodobieństw warunkowych.

sekwencyjnej analizy tekstu teoria informacji jest więc narzędziem komplementarnym w stosunku do teorii łańcuchów Markowa (oba podejścia posługują się pojęciem prawdopodobieństwa i nie wymagają kwantyfikacji tekstu) oraz alternatywnym w stosunku do analizy szeregów czasowych, w której konieczne jest zastąpienie badanych jednostek tekstowych liczbami.

Dodajmy na zakończenie, że w cytowanej już *Bibliography of Quantitative Linguistics* R. Köhlera (1995) teorii informacji poświęcono osobny dział, w którym cytowanych jest przeszło dwieście prac. Liczba ta jest zaniżona, gdyż nie uwzględnia publikacji książkowych, poruszających ten temat fragmentarycznie. Ponieważ lingwistyczne prace materiałowe stosujące pojęcia i metodologię teorii informacji należą do klasycznego repertuaru QL, a dla dalszego wywodu nie są szczególnie istotne, nie zostały tu omówione. Ich obszerny i reprezentatywny przegląd znaleźć można w cytowanej wyżej monografii R. Hammerla i J. Sambor (1990:361–451).

5.3 Teoria łańcuchów Markowa

Ważnym filarem wspierającym metodologicznie sekwencyjną analizę tekstu jest teoria łańcuchów Markowa, stanowiąca obecnie dział teorii procesów stochastycznych. Współcześnie znajduje ona zastosowanie w pracach z zakresu inżynierii językowej, zaliczanych do nurtu NLP (ang. *Natural Language Processing*), poświęconych między innymi takim kwestiom, jak automatyczna analiza morfologiczno-składniowa (ang. *parsing*), analiza i synteza mowy oraz budowa interfejsów komunikacyjnych typu *text-to-speech* i/lub *speech-to-text*. Zagadnienia te mają jednak charakter wybitnie inżynierski i sytuują się poza nurtem badań ogólnolingwistycznych, do którego zalicza się QL[45]. Wypada więc przypomnieć, że najczęściej przez lingwistów cytowana praca A.A. Markowa z 1913 r. poświęcona jest zastosowaniu jego teorii zdarzeń zależnych do danych językowych. Zgodnie ze współczesną terminologią należałoby ją określić jako przykład „dyskretnego modelowania sekwencyjnej struktury tekstu w analizie stylometrycznej". Streścimy tutaj lingwistyczne tezy tego artykułu, stawiając jednak na pierwszym planie jego historyczne, a nie metodologiczne znaczenie. Następnie przedstawimy wyniki uzyskane przez M. Petruszewycz, która powtórzyła testy Markowa, posługując się obszerniejszym materiałem językowym.

Markow poddał analizie losowo wybrane fragmenty poematu *Eugeniusz Oniegin* A. Puszkina. Badany przez niego korpus miał łączną długość 20 000 liter (pominięto miękkie i twarde znaki oraz odstępy międzywyrazowe). Warunki eksperymentu zdefiniowano w ten sposób, że jako zdarzenia losowe określono kolejne pojawienia się spółgłosek (C) bądź samogłosek (V). Przyjęto następnie hipotezę, iż liczba (prawdopodobieństwo)

[45] W lingwistyce teoretycznej na teorię łańcuchów Markowa wielokrotnie powoływali się badacze związani ze szkołą gramatyki generatywnej (MILLER&CHOMSKY 1963). Syntetyczne omówienie tej problematyki wraz z bibliografią znaleźć można w pracy DAMERAU 1971. Przejrzysty zarys teorii łańcuchów Markowa w ujęciu matematycznym zawiera praca FELLER 1987:338–374.

dłuższych sekwencji liter (na przykład CV, VV, CCV itd.) nie jest losową kombinacją częstości (prawdopodobieństw) zdarzeń elementarnych C i V, ale zależy od typu sekwencji. Hipoteza ta opiera się na założeniu, iż z uwagi na wyraźniejszy kontrast brzmieniowy sekwencje naprzemienne (CV, VCV itd.) powinny pojawiać się częściej od sekwencji jednorodnych (VV, CC itd.). Uogólniając, przyjęto, że prawdopodobieństwo pojawienia się określonego symbolu (ciągu symboli) w linii tekstu nie jest losowe, ale zależy od poprzedzającego go rządka liter.

Zastosowany przez Markowa sposób kodowania tekstu jest, według dzisiejszych kryteriów, niedoskonały. Można mu zarzucić między innymi pominięcie spacji (pauz?), niekonsekwentne traktowanie półsamogłosek oraz pomieszanie pojęć charakterystycznych dla grafii i fonii tekstu. Jednak na ostateczny wynik okoliczności te nie miały istotnego wpływu. M. Petruszewycz powtórzyła eksperyment Markowa z 1913 r., prowadzony pierwotnie na tekstach drukowanych według zasad starej ortografii języka rosyjskiego. Testy przeprowadzone na tych samych fragmentach kodowanych współcześnie potwierdziły rezultaty otrzymane przez Markowa: „mais nous pouvons refaire ce décompte sur un texte en orthographe moderne – ce que nous avons fait, effectivement – cela ne change pas les décomptes, du moins en ce qui concerne le nombre des doublets, triplets, mais parfois ce ne sont pas les mêmes voyelles.” (PETRUSZEWYCZ 1981:155–158).

W omawianym artykule Markow zastosował wprowadzone wcześniej (MARKOW 1907) współczynniki dyspersji (*коэффициент дисперсии*) dla rządków dwu- i trzyelementowych. Mają one wyrażać stosunek obserwowanej i teoretycznej wariancji bi- oraz trigramów (przy założeniu losowości rozkładu teoretycznego). Jeżeli także obserwowany szereg jest losowy, ich wartości powinny być równe jedności, natomiast dla szeregu „związanego w łańcuch” (*связь испытаний в цепь*) ich wartość powinna mieścić się w przedziale [0, 1]. Oba te współczynniki zostały zastosowane w zmodyfikowanej formie przez Petruszewycz (1981:27–28, *passim*), która oznaczyła je symbolami *M* (od nazwiska ich twórcy – *ibid.* 39) i C_m. Dodajmy, że wspomniany wcześniej warunek $0 < M < 1$ jest spełniony, jeżeli istotnie $p_{vv} < p_{vc}$ i $\delta < 0$. W przeciwnym wypadku, formuły (28) i (29) należy zmodyfikować. Markowa współczynnik spójności dla bigramów ma postać:

(28) $$M = \frac{1+\delta}{1-\delta}, \qquad \text{gdzie } \delta = p_{vv} - p_{cv}$$

Analogiczny współczynnik dla trigramów ma postać:

(29) $$C_m = \frac{1+\delta}{1-\delta}\left(\frac{1+\varepsilon}{2(1-\varepsilon)} + \frac{1+\eta}{2(1-\eta)}\right) + \frac{(q-p)(\eta-\varepsilon)}{(1-\varepsilon)(1-\eta)}$$

gdzie $$\delta = p_{vv} - p_{cv}, \qquad \varepsilon = \frac{p_{vvv} - p_{vv}}{p_{vc}}, \qquad \eta = \frac{p_{ccc} - p_{cc}}{p_{cv}}$$

$p_{x\ldots xy}$ – prawdopodobieństwo symbolu *y*, poprzedzonego rządkiem *x*...*x*

Za pomocą powyższych współczynników M. Petruszewycz porównała teksty A. Puszkina pisane wierszem (*Eugeniusz Oniegin*) i prozą (*Córka Kapitana*), teksty poetyckie Puszkina i W. Chlebnikowa (autor pierwszy określony jest tu jako „klasyk", drugi jako „futurysta") oraz fragmenty przemówień i pism W. Lenina (Tab. 7 i 8)[46]. Powtórzenie testów Markowa miało pokazać, że tekst jako obiekt badawczy w analizie sekwencyjnej *k*-gramów nie został przez Markowa wybrany przypadkowo, lecz ze względu na swą specyficzną cechę, jaką jest zależność kolejnych, następujących po sobie jednostek: „Ces précisions données, les considérations ci-dessus nous induisent à penser que Markov ne tenait pas pour seulement fortuit ou simplement commode son domain d'application et nous allons sur deux exemples essayer de montrer que la chaîne markovienne peut être un instrument de recherche." (PETRUSZEWYCZ 1981:40). Chociaż z lingwistycznego punktu widzenia opisywana tu metoda może budzić zastrzeżenia, hipoteza, zgodnie z którą tekst jest szeregiem zdarzeń zależnych[47], sformułowana po raz pierwszy przez Markowa, stanowi podstawę i punkt wyjścia dla późniejszych kwantytatywnych badań sekwencyjnej struktury tekstu.

Tab. 7 Porównanie wiersza i prozy A. Puszkina[48]

	poezja	proza 1	proza 2	proza 3
p_{vv}	0,117	0,130	0,128	0,126
p_{vc}	0,665	0,698	0,694	0,696
M	0,292	0,276	0,277	0,274
p_{vvv}	0,110	0,110	0,112	0,114
p_{ccc}	0,138	0,158	0,166	0,165
C_m	0,192	0,207	0,209	0,207

Analiza danych z tabeli 7 skłania do kilku refleksji. Po pierwsze, zaskoczeniem są znaczące, choć niewielkie, wartości p_{vvv} i p_{ccc}. Najpewniej są one skutkiem pominięcia pauz międzywyrazowych oraz posługiwania się sekwencjami liter, a nie głosek. Wbrew oczekiwaniom, obserwujemy też nietypowe zachowanie współczynników M i C_m. Jak ze wzorów (28) i (29) wynika, im większe M, tym bardziej niezależne od lewostronnego kontekstu będą pojawienia się kolejnych elementów badanego szeregu (podobnie C_m). Trudno oczywiście z góry przesądzać, dla jakiej odmiany stylistycznej powiązania liter powinny być silniejsze, jednak przedstawiona wyżej sytuacja, w której parametry M i C_m dają dla tych samych tekstów rozbieżne rezultaty, jest zastanawiająca.

Stosując tę samą metodę, Petruszewycz porównała kilka stylistycznych odmian języka rosyjskiego. Tabela 8 przedstawia wyniki testów przeprowadzonych na tekstach Puszkina i Chlebnikowa pisanych wierszem i prozą oraz na tekstach Lenina, reprezentujących dyskurs mówiony i pisany. Nie jest naszym celem dokonanie filologicznej inter-

[46] Zestawienie tych nazwisk jest nieco szokujące i czujemy się w obowiązku podkreślić, że figuruje tu jedynie na zasadach cytatu.
[47] Terminy *zdarzenie* i *zależność* użyte są tu w sensie statystycznym, a nie potocznym.
[48] Dane na podstawie pracy PETRUSZEWYCZ 1981:52.

pretacji tych wyników, tym bardziej, że nie są nam znane szczegółowe zasady próbkowania i kodowania tekstu, a także przebiegi i rozkłady statystyczne zmiennych M i C_m pozwalające na definicję przedziałów ufności i statystyczną ocenę różnic wartości obu tych parametrów. Należy jednak zwrócić uwagę na relację wartości M i C_m dla porównywanych grup tekstów, pamiętając o tym, że im wyższa ich wartość, tym bardziej losowe uporządkowanie elementów szeregu. Z przedstawionych danych wynika, że ze względu na występowanie po sobie samogłosek i spółgłosek, teksty Puszkina jako „klasyka" są statystycznie bardziej przewidywalne (rytmiczne) niż awangardowe teksty futurysty Chlebnikowa. Także zgodnie z oczekiwaniami, dyskurs mówiony Lenina – zręcznego oratora i populisty – okazuje się bardziej rytmiczny (przewidywalny ze względu na pojawianie się samogłosek i spółgłosek) od tekstu pisanego tegoż autora. Z kolei porównanie tekstów artystycznych i politycznych ze względu na to samo kryterium wskazuje na wyższą „spójność" tych pierwszych, co wynika z podporządkowania ich wymogom estetycznym, a w mniejszym stopniu komunikacyjnym czy perswazyjnym. Dla poszczególnych prób widoczna jest też relacja $C_m < M$ wynikająca w sposób analityczny z faktu, iż przewidywalność wystąpienia symbolu rośnie wraz z długością uwzględnionego lewostronnego kontekstu. Przy obliczaniu parametru M uwzględnia się jeden symbol (założenie, iż tekst jest szeregiem Markowa pierwszego rzędu), natomiast parametr C_m oblicza się biorąc pod uwagę dwa symbole (założenie, iż tekst jest szeregiem Markowa rzędu drugiego). Wątpliwości budzi natomiast relacja C_m i M dla fragmentów poezji i prozy tych samych autorów. Przecież właśnie w tekstach reprezentujących mowę wiązaną należy szukać wysokiego stopnia eufonii, podczas gdy cytowane parametry wskazują na tekst prozatorski jako lepiej uporządkowany.

Tab. 8 Parametry Markowa dla tekstów A. Puszkina i W. Chlebnikowa oraz W. Lenina[49]

	Puszkin (poezja)	Chlebnikow (poezja)	Puszkin (proza)	Chlebnikow (proza)	Lenin (mówiony)	Lenin (pisany)	Lenin (pisany)
p_{vv}	0,122	0,135	0,132	0,146	0,156	0,146	0,151
p_{vc}	0,659	0,663	0,714	0,686	0,682	0,660	0,652
M	0,300	0,309	0,264	0,298	0,310	0,321	0,332
p_{vvv}	0,093	0,132	0,119	0,101	0,111	0,109	0,097
p_{ccc}	0,152	0,185	0,171	0,164	0,173	0,187	0,206
C_m	0,197	0,226	0,212	0,216	0,227	0,229	0,236

Dyfuzja koncepcji Markowa w lingwistyce akademickiej miała bardzo ograniczony zasięg i nastąpiła z dużym opóźnieniem. Jedną z pierwszych prezentacji jego dorobku zawdzięczamy G. Herdanowi (1960:140–153). Cytowany wyżej artykuł Markowa (1913) omawiają także G.A. Miller i N. Chomsky (1963). Koncepcję tekstu jako swoistego procesu stochastycznego przedstawił B. Brainerd (1976). Jak dotąd, jedyne pogłębione omówienie językoznawczego dorobku Markowa sporządziła cytowana wyżej francuska

[49] Dane na podstawie pracy PETRUSZEWYCZ 1981:65 i 57.

lingwistka M. Petruszewycz. Autorka opublikowała na ten temat serię artykułów, które jako rozprawę doktorską wydała we wspólnym tomie (PETRUSZEWYCZ 1981)[50]. Teoretyczne podstawy teorii łańcuchów Markowa i jej przykładowe zastosowania w fonologii zawiera praca R. Köhlera (1983). Na uwagę zasługuje fakt, iż autor ten połączył podejście probabilistyczne, wykorzystujące jakościową kategorię stanu, z analizą danych numerycznych, w której przedmiotem analizy jest szereg liczbowy otrzymany z tekstu przez kwantyfikację, a głównym narzędziem badawczym jest funkcja autokorelacji. Ponadto, po raz pierwszy w kontekście sekwencyjnej analizy tekstu, u Köhlera pojawiła się koncepcja modelowania struktur języka w dzisiejszym rozumieniu.

Warto w tym miejscu zapytać o przyczyny wyboru przez Markowa lingwistycznego materiału badawczego dla testów jego statystycznej teorii zdarzeń zależnych. Chociaż w materii tej zdani jesteśmy na domysły[51], najbardziej prawdopodobnym powodem tego zainteresowania jest jego zetknięcie się z przedstawicielami środowiska lingwistów związanych z uniwersytetami w Petersburgu i Dorpacie (Tartu). W latach 1900–1920, podobnie jak Markow, profesorem uniwersytetu w Petersburgu był na przykład jeden z twórców strukturalizmu, J.I.N. Baudouin de Courtenay, natomiast na uniwersytecie w Dorpacie wykładał W. Lutosławski, pionier ilościowych badań nad chronologią tekstów konkretnego autora (LUTOSŁAWSKI 1897). Wiadomo też, że Markow zapoznał się z obszernym artykułem N.A. Morozowa *Лингвистические спектры* z 1915 r., traktującym o ilościowym podejściu do problemu autorstwa (PETRUSZEWYCZ 1981:139–148, MOROZOV 1915). Oprócz tych zewnętrznych uwarunkowań należy jednak podkreślić, że tekst jest konstruktem immanentnie linearnym i wprost idealnie nadaje się do testów zależności statystycznej. Wybierając materiał do badań, Markow mógł więc kierować się jedynie względami praktycznymi.

Z dzisiejszej perspektywy widać, że lingwistyczne prace Markowa należy uznać za zapowiedź (z pewnością nie jedyną) kwantytatywnych badań języka, rozwiniętych na fali strukturalizmu, a dziś kontynuowanych w obrębie teorii systemów. Prace te nie zostały jednak w porę dostrzeżone i wykorzystane przez lingwistów. Najbardziej trywialną tego przyczyną była ich ograniczona dostępność – autor publikował je w Petersburgu, w języku rosyjskim, w przededniu wybuchu pierwszej wojny światowej i rewolucji bolszewickiej. Przyczyną głębszą był ich nowatorski charakter – w początkach XX wieku badanie języka metodami ścisłymi wciąż jeszcze wykraczało poza utarte schematy myślenia. Dla lingwistów pewną barierę stanowić też mógł dość wyrafinowany aparat matematyczny stosowany w teorii zdarzeń zależnych. Przesadą byłoby twierdzić, że lingwistyczny dorobek Markowa popadł w zapomnienie, jednak rzeczywistymi i godnymi kontynuatorami jego myśli okazali się nie lingwiści, lecz matematycy i inżynierowie języka.

---

[50] Cytowana praca zawiera pełną bibliografię prac Markowa. Artykuły Petruszewycz poświęcone Markowowi ukazały się w czasopiśmie *Mathématiques et les Sciences Humaines*.

[51] „Aucune indication n'apparaissant dans les bibliographies sur les origines ou raisons de ce choix, le chercheur ne peut que faire des hypothèses." (PETRUSZEWYCZ 1981:133).

5.4 Analiza widmowa i analiza szeregów czasowych

Na rozwój współczesnych badań syntagmatycznych struktur tekstu znaczny wpływ wywarły techniki analizy szeregów czasowych oparte na analizie widmowej, a po publikacji *Time Series Analysis* G. Boxa i G. Jenkinsa w 1970 r. na metodzie ARIMA.

W 1969 r. ukazał się artykuł poświęcony syntagmatycznej analizie tekstów w języku chińskim (Dreher et al. 1969). Za pomocą analizy widmowej zbadano rytm prozy kilku chińskich autorów określony rozkładem tzw. segmentów i zdań w linii tekstu. Jako miarę długości przyjęto liczbę znaków tworzących segment lub zdanie. Uzyskane spektrogramy uznano za charakterystyczne dla każdego z pisarzy i sugerowano ich przydatność w badaniu autorstwa. Jednak bliższy ogląd przytoczonych wyników każe uznać koncepcję interpretacji spektrogramów w kategoriach stylistyki za naukowo nieuzasadnioną. Milczeniem pominięto także kontrowersje związane z problemem autorstwa, istotne dla tej problematyki (por. Muller&Brunet 1992). Omawiana praca zasługuje jednak na uwagę ze względu na przyjętą metodologię. Wnioski o charakterze materiałowym są natomiast relewantne jedynie dla języka chińskiego.

Analizę spektralną do badania rytmicznej struktury tekstów w języku angielskim zastosowali P. Bratley i D. Ross (1981). Autorzy posłużyli się sztucznie generowanymi tekstami, a także dość przypadkowymi próbkami prozy i poezji. Także ta praca ma z punktu widzenia lingwistyki wartość przede wszystkim metodologiczną.

Podobną techniką, chociaż na całkowicie odmiennym materiale językowym, posłużyli się J.B. Smith i B.A. Rosenberg (1973), których zainspirowały studia nad literaturą oralną, a w szczególności tzw. teoria formulaiczna A.B. Lorda i M. Parry'ego (Lord 1960). Autorzy próbowali wykazać, że, podobnie jak wywodzące się z przekazów ustnych teksty Homera czy pieśni bałkańskich guślarzy, kazania wygłaszane spontanicznie w kościołach południowych stanów USA mają charakter formulaiczny – składają się z gotowych, powtarzających się wzorców rytmicznych. Pomiary przeprowadzono analizując szeregi czasowe wygenerowane na podstawie liczby słów wypowiadanych przez kaznodzieję w jednostce czasu. Porównanie wyników otrzymanych dla większej liczby kazań dowodzi, zdaniem autorów, archetypicznego charakteru odkrytych wzorców rytmicznych.

Interesujące wyniki daje zastosowanie metod sekwencyjnych na poziomie fonetyczno-fonologicznym. Przykładem takiej analizy jest badanie struktury biblijnego tekstu hebrajskiego, przeprowadzone za pomocą funkcji autokorelacji oraz analizy widmowej (Azar&Kedem 1979). Stosując kwantyfikację binarną (0 – brak cechy, 1 – obecność cechy), autorzy wygenerowali szeregi czasowe złożone a) z głosek dźwięcznych i bezdźwięcznych, b) ze spółgłosek dźwięcznych i bezdźwięcznych oraz c) ze spółgłosek i samogłosek. Uzyskane spektrogramy wykazały wyraźne zróżnicowanie rytmiczne fragmentów prozatorskich i poetyckich. Z kolei sekwencyjne cechy szeregów spółgłoskowo-samogłoskowych wskazywały na silny kontrast następujących po sobie głosek. Obserwacja ta potwierdza stwierdzoną także w naszych analizach prawidłowość, iż dominującą cechą struktury syntagmatycznej na poziomie fonologicznym (ze szczególnym uwzglę-

dnieniem akcentuacji) jest kontrast bezpośrednio sąsiadujących ze sobą jednostek. Jednak mimo metodologicznej poprawności i odniesienia do omówionych wyżej koncepcji G. Herdana, w cytowanej pracy zabrakło kilku istotnych z lingwistycznego punktu widzenia informacji. Nie wiadomo na przykład, dlaczego badania przeprowadzono na tekście hebrajskim, dlaczego był to tekst biblijny i dlaczego kwantyfikowano akurat te, a nie inne cechy głosek. Zamieszczenie tych informacji z pewnością pogłębiłoby wiarygodną lingwistyczną interpretację uzyskanych wyników.

Kilkakrotnie analizie poddawano strukturę tekstu kodowanego w postaci sekwencji długości kolejnych zdań (HŘEBÍČEK 1997:124–149, ROBERTS 1996, SCHILS&DE HAAN 1993). Omówimy tutaj dwie ostatnie pozycje. A. Roberts (1996) zastosował funkcję autokorelacji do zbadania rytmu prozy wyznaczonego sekwencją długości zdań. Podobnie jak w poprzednich przypadkach, nie weryfikowano konkretnej hipotezy lingwistycznej, ograniczając się w pierwszej kolejności do odpowiedzi na pytanie o to, czy sekwencja długości zdań może być traktowana jako szereg zdarzeń zależnych. Na uwagę zasługuje zastosowana przez Robertsa procedura badawcza. Jako jeden z niewielu lingwistów porównał on wskaźniki sekwencyjne dla tekstów rzeczywistych i symulowanych, otrzymanych drogą losowego mieszania tych pierwszych. Drugie pytanie postawione przez Robertsa brzmiało więc: w jakim stopniu autorski porządek narracji porządkuje sekwencję długości zdań w prozie literackiej, jeżeli zestawić go z losowym porządkiem tego samego zbioru zdań. Porównanie pierwszych pięciu współczynników autokorelacji ze średnimi wartościami tych parametrów dla stu szeregów losowych („pseudotekstów") wykazało, że autorski sposób uporządkowania zdań pod względem długości nie jest w prozie artystycznej przypadkowy. Lewostronny kontekst pozwalał z mniejszą lub większą dokładnością przewidywać długości pojawiających się jednostek: „Since the actual and random texts differ in nothing but the order of sentences, this in itself could well convince us that the history of a sentence – i.e. the lengths of the sentences in which it is embedded – affects how long it is likely to be." (ROBERTS 1996:36).

Stylometryczną, ilościową analizę tekstu przeprowadzili E. Schils i P. de Haan (1993). Autorzy poddali empirycznym testom hipotezę, zgodnie z którą zróżnicowanie i ożywienie narracji osiąga się poprzez alternację zdań długich i krótkich[52]. Materiał badawczy złożony był z tekstów naukowych, popularnych i fikcji literackiej w języku angielskim. Autorzy zastosowali funkcję autokorelacji w klasycznej postaci oraz tzw. współczynnik von Neumana, określający stopień losowości szeregu. Przedstawimy tutaj ten współczynnik, ponieważ w literaturze przedmiotu pojawia się on stosunkowo rzadko[53]. Niech dany będzie szereg czasowy $\{x_1, x_2, ..., x_n\}$. Przy analizie autokorelacji bada się relacje par wartości (x_t, x_{t+k}), gdzie *k* określone jest jako *odstęp* (ang. *lag*) dzielący realizacje w momentach *t* i *t* + *k*. Współczynnik von Neumana konstruowany jest natomiast w oparciu o szereg utworzony przez różnice par (x_t, x_{t+1}). Przy jego definicji korzysta się

[52] Źródłem tej skądinąd zdroworozsądkowej hipotezy jest praca MARCKWORTH&BELL 1967.
[53] Zastosował go m.in. A. Salem (1988:135), analizując chronologię zmian leksyki.

z tzw. wariancji von Neumana, wyrażającej średni rozrzut różnic pomiędzy następującymi po sobie realizacjami szeregu[54]:

$$\delta^2 = \frac{1}{n-1}\sum_{t=1}^{n-1}\left(x_{t+1} - x_t\right)^2 \tag{30}$$

Jeżeli przez σ^2 oznaczy się zwykłą wariancję szeregu, współczynnik von Neumana będzie miał postać:

$$VN = \frac{\delta^2}{\sigma^2} \tag{31}$$

Zamiast wariancji zwykłej można też obliczyć wariancję wszystkich różnic pomiędzy wartościami szeregu (SALEM 1988:135):

$$\sigma^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i - x_j\right)^2 \tag{32}$$

W przypadku autokorelacji pozytywnej (długie, łagodnie rosnące bądź malejące przebiegi), δ^2 będzie mniejsze od σ^2 i zachodzić będzie zależność $VN \in (0,1)$; w przypadku korelacji negatywnej (naprzemienny porządek wartości), *VN* będzie większe od jedności, natomiast $VN \approx 1$ oznaczać będzie brak korelacji[55].

Testy przeprowadzone przez Schils i de Haan nie potwierdziły postawionej hipotezy. O pewnym poziomie alternacji w tekście literackim można mówić jedynie w przypadku stosunkowo krótkich odcinków tekstu, natomiast przy odcinkach dłuższych pojawiają się zakłócenia rytmu, wynikające z przemieszania fragmentów dialogowych i opisowych. Wykazano też, że w tekstach naukowych i popularnonaukowych korelacja długości następujących po sobie zdań w ogóle nie zachodzi. Jednak nawet w tekstach artystycznych, w założeniu spójnych pod względem treściowym i formalnym, stwierdzono słabsze od oczekiwanego powiązanie długości kolejnych zdań.

Bardzo dobre rezultaty w badaniu syntagmatycznej struktury tekstu daje metoda ARIMA[56], opracowana przez amerykańskich statystyków G. Boxa i G. Jenkinsa. Przeznaczona początkowo do zastosowań inżynierskich i ekonomicznych (prognozowania i sterowania procesami technologicznymi), w latach 80-tych została przeniesiona na grunt nauk społecznych, przede wszystkim socjologii i psychologii.

---

[54] Stosujemy notację A. Salema (1988). W cytowanym artykule E. Schils i P. de Haan zamiast δ^2 stosuje się skrót MSJ (ang. *mean square jump*). Autorzy podają też kryteria oceny wartości współczynnika *VN*.

[55] A. Salem nie podaje dokładniejszych kryteriów oceny wyniku (np. przedziałów ufności). Znaleźć je można w cytowanej pracy E. Schils i P. de Haan (1993) oraz w źródłowym tekście J. von Neumana (1941).

[56] ARIMA jest akronimem angielskiej nazwy pełnego modelu szeregu czasowego (***A**uto**R**egressive **I**ntegrated **M**oving **A**verage*). W literaturze przedmiotu nazwa ta stosowana jest na określenie metody Boxa i Jenkinsa nawet wówczas, gdy, tak jak przy analizie danych tekstowych, wykorzystuje się jedynie modele stacjonarne ARMA i sezonowe SARMA.

Pierwsze zastosowania metody Boxa i Jenkinsa w lingwistyce pojawiły się dopiero w latach 90. M. Corduas (1995) poddała analizie teksty trzech współczesnych pisarzy włoskich (A. Manzoni, C. Pavese i D. Buzzati). Przyjmując jako podstawę generowania szeregów czasowych graficzne długości kolejnych słów, autorka estymowała modele procesów typu autoregresji (AR), autoregresji i ruchomej średniej (ARMA) oraz model mieszany ARCH[57]. Corduas wykazała, że relacje długości następujących po sobie słów można traktować jak zdarzenia zależne i opisać modelem autoregresji AR(1) (Tab. 9). Nie porównano jednak parametrów tego modelu, obliczonych dla poszczególnych autorów. Rezultaty analizy sekwencyjnej porównano z parametrami pozycyjnymi badanych populacji (średnią, modą, medianą itd.). W celu uproszczenia statystycznego opisu danych autorka zastosowała rozwiązanie, polegające na dekompozycji modelu mieszanego typu ARMA(4,2) na dwa modele proste typu AR. Okazało się, że na jeden z modeli prostych przypadało aż 85% całkowitej wariancji wyjaśnionej przez model mieszany.

Tab. 9 Modele autoregresyjne opisujące sekwencję długości słów (CORDUAS 1995)[58]

Autor	Typ modelu	Model
A. Manzoni[59]	AR(1)	$x_t = m_x - 0{,}28x_{t-1} + e_t$
C. Pavese	AR(1)	$x_t = 5{,}60 - 0{,}22x_{t-1} + e_t$
D. Buzzati	AR(1)	$x_t = 5{,}86 - 0{,}19x_{t-1} + e_t$

Niestety, podobnie jak w uprzednio cytowanych artykułach, także w pracy Corduas odczuwa się brak mocnych podstaw lingwistycznych. Wykazano co prawda, że sekwencja długości słów w języku włoskim może być traktowana jako realizacja procesu Markowa rzędu pierwszego, jednak fakt ten nie został należycie zinterpretowany. Poważniejszym mankamentem jest brak przesłanek o charakterze dedukcyjnym, sprawiający, że nie wiadomo, jakich parametrów statystycznych należy oczekiwać: czy wskazujących na podobieństwo tekstów (ten sam język, gatunek, epoka), czy też na ich różnicę (różni autorzy).

Jak już wspomniano, jednym z preferowanych przez lingwistów parametrów ilościowych tekstu jest długość zdania. R. Oppenheim zastosowała metodę ARIMA do badania sekwencji długości zdań w tekście literackim (OPPENHEIM 1988). Wzorem innych badaczy[60], autorka przyjęła założenie, iż długości kolejnych zdań nie są od siebie niezależne: „we hypothesize that the lengths of successive sentences are mathematically related, or correlated. When such correlation exists, the length of a sentence can be described, quantitatively, in terms of the lengths of previous sentences." (*ibid.* 242) Na plus omawianej pracy warto odnotować, że autorka przedstawiła argumenty nie tylko zwolenników, ale także przeciwników stosowania metod ilościowych w badaniu autor-

[57] Ang. *AutoRegressive Conditional Heterodastic process*.
[58] Oznaczenia: e_t – szum o rozkładzie N(0,1), x_t – wartość szeregu w chwili lub na pozycji t.
[59] Dla tego autora nie podano wartości średniej szeregu m_x.
[60] Cytowana jest m.in. praca G.U. Yule'a na temat rozkładu statystycznego długości zdań (YULE 1939).

stwa tekstu. W pewnym uproszczeniu, argumenty te mają odpowiedzieć na pytanie o istnienie w statystycznej strukturze tekstu literackiego jakiegoś indywidualnego piętna (w literaturze anglosaskiej używa się terminu *fingerprint* – odcisk linii papilarnych) pozwalającego na odróżnienie stylu autora od ogólnej normy językowej oraz od stylu innych autorów. Jak dotąd, w debacie tej przeważają przeciwnicy stosowania narzędzi statystycznych w badaniu stylu osobniczego. Jednak przewaga ta nie jest wystarczająca, by wykluczyć możliwość przynajmniej poszlakowego ustalenia autorstwa, o ile tylko spełnione zostaną pewne warunki wstępne.

Tab. 10 Modele opisujące sekwencję długości zdań w prozie artystycznej[61]

Autor, tytuł	Typ modelu	Model
J. Joyce – *Ulysses* (1)	MA(1)	$x_t = 8{,}22 + 0{,}142e_{t-1} + e_t$
J. Joyce – *Ulysses* (2)	AR(2)	$x_t = 8{,}07 + 0{,}195x_{t-1} + 0{,}365x_{t-2} + e_t$
J. Joyce – *The Dead* (1)	AR(2)	$x_t = 15{,}4 - 0{,}077x_{t-1} + 0{,}198x_{t-2} + e_t$
J. Joyce – *The Dead* (2)	ARMA(1,1)	$x_t = 9{,}58 + 0{,}413x_{t-1} - 0{,}283e_{t-1} + e_t$
J. Joyce – *The Dead* (3)	ARMA(1,1)	$x_t = 2{,}56 + 0{,}814x_{t-1} - 0{,}695e_{t-1} + e_t$
E. Hemingway – *Big Two-Hearted River* (1)	AR(2)	$x_t = 23{,}32 - 0{,}28x_{t-1} - 0{,}216x_{t-2} + e_t$
E. Hemingway – *Big Two-Hearted River* (2)	ARMA(1,1)	$x_t = 3{,}43 + 0{,}737x_{t-1} - 0{,}620e_{t-1} + e_t$
E. Hemingway – *Big Two-Hearted River* (3)	ARMA(1,1)	$x_t = 2{,}31 - 0{,}770x_{t-1} + 0{,}861e_{t-1} + e_t$

Wysunięta przez Oppenheim hipoteza zweryfikowana została jedynie częściowo. Wykazano co prawda, że sekwencje zdań są realizacją jakichś procesów stochastycznych, jednak rozbieżności w wartościach odpowiednich parametrów dla tych samych autorów i dzieł były zbyt duże, by określić je jako charakterystyczne dla stylu osobniczego czy konkretnego dzieła (Tab. 10). Rozumowania tego nie należy jednak uogólniać na całość problematyki stylometrycznej, ponieważ autorka zbadała stosunkowo niewiele próbek – trzy stuzdaniowe fragmenty powieści E. Hemingway'a oraz łącznie pięć fragmentów powieści J. Joyce'a. Uzyskanego rezultatu, z całą pewnością wiarygodnego, nie próbowano też wyjaśnić w kategoriach językoznawczych czy teoretycznoliterackich.

Monografię poświęconą zastosowaniom metody ARIMA do problemu autorstwa opublikował też A. Pawłowski (1998). Ze względu na objętość tej pracy, streścimy tu jej najważniejsze tezy. Autorem badanych tekstów był francuski pisarz R. Gary, publikujący pod koniec życia także jako E. Ajar. Genialne oszustwo literackie Gary'ego, jedynego w historii pisarza dwukrotnie uhonorowanego nagrodą Goncourtów[62], ośmieszyło zastępy paryskich krytyków, przy okazji dostarczając lingwistom zainteresowanym problemem autorstwa wyjątkowego materiału badawczego. Nie ma bowiem wątpliwości co do faktu,

[61] Na podstawie pracy OPPENHEIM 1988. Oznaczenia: e_t – szum o rozkładzie N(0,1), x_t – wartość szeregu w chwili lub na pozycji t.

[62] Akademia Goncourtów nie nagradza dwukrotnie tego samego autora. R. Gary otrzymał to prestiżowe wyróżnienie w 1956 r. za powieść *Les racines du ciel*, a w 1975 r., jako E. Ajar, za powieść *La vie devant soi*.

iż Gary i Ajar to jedna i ta sama osoba. Ale równie bezsporną kwestią pozostaje przekonanie czytelników powieści sygnowanych rzeczonymi nazwiskami, iż obcują z autorami różnymi pod względem tożsamości fizycznej i artystycznej. Warto też dodać, że dopóki za Ajara nie podstawiono figuranta, ówczesna prasa prześcigała się w domysłach co do jego prawdziwej tożsamości, proponując jako kandydatów innych pisarzy – L. Aragona, R. Quenneau i M. Tourniera. Jak wynika z powyższej prezentacji, za skuteczną więc uznać należy taką metodę badania autorstwa tekstu, która wykaże istotne podobieństwo tekstów Gary'ego i Ajara oraz różnicę pomiędzy tekstami Ajara i innych autorów.

Analizie sekwencyjnej poddano szeregi czasowe wygenerowane z tekstów Gary'ego, Ajara, Aragona, Queneau i Tourniera (łącznie prawie tysiąc próbek) oraz z anglojęzycznych wersji kilkudziesięciu fragmentów powieści Gary'ego[63]. Zastosowano trzy rodzaje kwantyfikacji: 1) ilość bitów informacji w kolejnych wyrazach tekstu; 2) sekwencję odstępów (ang. *gaps*) pomiędzy kolejnymi wystąpieniami najczęstszych morfemów gramatycznych oraz 3) sekwencję długości zdań.

W wyniku testów przeprowadzonych na kilkuset próbach ustalono, że sekwencje ilości informacji niesionej przez kolejne wyrazy tekstu w językach francuskim i angielskim można opisać modelem średniej ruchomej MA(1). Mimo, że współczynniki tego modelu miały niewielką wartość, powtarzały się w sposób regularny w większości badanych szeregów (*ibid*. 96–112). Stwierdzono też ujemną autokorelację bezpośrednio sąsiadujących wartości szeregu, wskazującą na statystycznie znaczącą alternację wyrazów o dużej i małej zawartości informacyjnej. W językach o tendencji analitycznej zjawisko to można wyjaśnić względnie równomiernym przemieszaniem wyrazów bardzo częstych, niosących niewiele informacji (przede wszystkim zaimków, rodzajników i przyimków) oraz wyrazów o niskich frekwencjach i dużej zawartości informacyjnej (por. Część II, 5.2). Zastosowanie otrzymanych tu sekwencyjnych parametrów do przedstawionej wyżej kwestii autorstwa dało pozytywny wynik: spośród analizowanych tekstów najbliżej Ajara sytuowały się dzieła Gary'ego.

W przypadku szeregów czasowych generowanych przez wystąpienia kolejnych morfemów gramatycznych o najwyższych częstościach[64] otrzymano zerowe autokorelacje (*ibid*. 113–123). Jako realizacje procesu losowego, szeregi takie są więc nieprzydatne w stylometrii. Jest to o tyle interesujące, że, jak się uważa, ze względów psychologicznych i/lub kompozycyjnych pewne wyrazy występują w zbitkach, a ich nierównomierny rozkład powinien wpływać na zmianę sekwencyjnych parametrów odpowiedniego szeregu czasowego. Z przeprowadzonych testów należy wnioskować, że pewna nierównomierność rozkładu w linii tekstu wystąpić może jedynie w przypadku wyrazów tematycznych, specyficznych dla każdego tekstu.

---

[63] W początkach swej kariery literackiej R. Gary napisał kilka powieści w języku angielskim. Nie bez wpływu na to pozostał zapewne fakt, iż jako dyplomata spędził on wiele lat w USA, a jego pierwsza żona, L. Blanche, była angielską pisarką.

[64] W przypadku leksemów o niskich częstościach nie udałoby się wygenerować dostatecznie długich szeregów czasowych.

Tab. 11 Procesy stochastyczne opisujące sekwencję długości zdań (PAWŁOWSKI 1998:130)

Model	AR(1)	AR(2)	AR(3)	ARMA	brak korelacji
Liczba fragmentów	64	39	9	103	132
Procentowo	18,4%	11,2%	2,6%	29,7%	38,1%

Testy prowadzone na sekwencjach długości zdań pokazały, iż mogą one stanowić realizację procesu Markowa, co w kategoriach lingwistycznych oznacza, że lewostronny kontekst każdego zdania może w pewnych przypadkach determinować statystycznie jego długość (*ibid.* 124–153). Odkryte procesy stochastyczne były jednak bardzo zróżnicowane: oprócz sekwencji losowych (gdzie długości zdań są od siebie statystycznie niezależne), stwierdzono obecność procesów prostych typu AR oraz mieszanych typu ARMA (Tab. 11).

Wykorzystując uśrednione wartości współczynników procesu prostego AR(1) oraz długości zdań, raz jeszcze pokazano, że topologiczna odległość analizowanych tekstów wskazuje na Gary'ego jako najbardziej prawdopodobnego autora powieści podpisanych pseudonimem E. Ajara (*ibid*. 137).

Powyższy przegląd dorobku naukowego w zakresie ilościowej, sekwencyjnej analizy tekstu skłania do kilku refleksji. Zastanawiająca jest z pewnością względnie skromna liczba publikacji poświęconych tej tematyce. Argument o niehumanistycznym charakterze lingwistyki kwantytatywnej jest niepoważny nie tylko ze względu na liczbę prac wykorzystujących statystykę konwencjonalną w badaniu języka, ale przede wszystkim ich merytoryczny zasięg i efektywność wysuwanych hipotez[65]. Także argument o braku bądź małym znaczeniu sekwencyjnych struktur w języku jest nie do utrzymania. Podstawową manifestacją języka jest bowiem tekst – struktura *par excellence* liniowa. Otóż, jak się wydaje, najważniejszą przyczyną opóźnienia w badaniach sekwencyjnych struktur języka jest nowatorski charakter stosowanej metodologii. O ile bowiem pierwsze znaczące prace z zakresu rachunku prawdopodobieństwa markiz P.S. Laplace publikował już u schyłku XVIII wieku, o tyle koncepcja statystycznej zależności zdarzeń i jej formalizacja narodziły się dzięki pracom A.A. Markowa dopiero w początkach wieku XX. W momencie, w którym lingwistyka dojrzała do stosowania metod ilościowych, między innymi dzięki pojawieniu się na uniwersytetach pierwszych maszyn liczących, naukowcy (w tej liczbie wielu filologów bez doświadczenia matematycznego) sięgnęli po dostępne i sprawdzone techniki statystyczne, wśród których metod sekwencyjnych najpewniej zabrakło.

Innym charakterystycznym rysem wielu omówionych wyżej prac jest rażący niekiedy kontrast pomiędzy precyzją wykorzystanego aparatu matematycznego a miałkością poruszanej problematyki językoznawczej. Można odnieść wrażenie, że zagadnienia językoznawcze są dla wielu autorów jedynie pretekstem służącym zastosowaniu w humanistyce

[65] Jak już we wstępnych rozdziałach pracy wspomniano, bibliografia lingwistyki kwantytatywnej Köhlera cytuje ponad sześć tysięcy pozycji ujętych w różne działy tematyczne (KÖHLER 1995).

metodologii typowej dla nauk przyrodniczych. Ale jeśli tak, to brak pogłębionej analizy zjawisk i prób ich wyjaśnienia podważa w ogóle celowość prowadzenia podobnych badań. Jak wiele pułapek czyha na tych, którzy silnie akcentują kwestie metodologiczne, lekceważąc jednocześnie aspekty językoznawcze i filologiczne, pokazują wypowiedzi doświadczonych francuskich lingwistów-statystyków poświęcone kwestii autorstwa (MULLER&BRUNET 1988). Na wstępie autorzy stawiają tyleż proste, co ważkie pytanie: „A quoi sert-il de soumettre les textes littéraires aux traitements informatiques et statistiques?". Następnie rozważają argumenty przemawiające za i przeciw stosowaniu metod statystycznych w stylometrii. Obserwacje Ch. Mullera i E. Bruneta wskazują, iż zmienna określana jako „styl" jest bardzo nieostra, zaś pojęcie „autora" reprezentowanego przez spójny statystycznie zbiór tekstów nie znajduje potwierdzenia w analizach stylometrycznych. „Toutes [nos observations – A.P.] convergent vers une constatation décevante, mais qu'il serait vain de taire: la variable *style*, dans une oeuvre littéraire assez étendue, et à plus forte raison dans un corpus comprenant des textes d'un même auteur, mais de genres divers, d'époques différentes, crée dans les données quantitatives autant et plus d'écarts que la variable *auteur*." (*ibid*. 371) Jest to o tyle znaczące, że cytowani autorzy są czołowymi przedstawicielami lingwistyki statystycznej we Francji, a kwestia atrybucji zajmuje w ich obszernym dorobku naukowym poczesne miejsce. Powołując się na liczne przykłady z literatury francuskiej, Muller i Brunet pokazali, że standardowa technika analizy wielowymiarowej, zastosowana na poziomie leksykalnym, nie wskazuje konkretnych autorów, ale gatunki bądź rodzaje literackie. Oznacza to, że niezależnie od rzeczywistej liczby uwzględnionych autorów należałoby się spodziewać wyraźnych skupień punktów wskazujących na powieść, dramat i poezję. Testy przeprowadzone na tekstach A. Lamartina, V. Hugo i A. Musseta skłoniły cytowanych badaczy do sformułowania wniosku: „Si donc les textes qui nous servent de témoins se trouvaient anonymes et que les méthodes quantitatives fussent appelées seules à trancher, elles inviteraient à conclure que ces textes sont imputables à trois auteurs différents dont l'un serait poète, l'autre dramaturge et le troisième romancier. [...] La vérité ne trouverait évidemment pas son compte dans ces histoires à dormir debout." (*ibid*. 378). Wypada jedynie uśmiechnąć się na myśl o atrybucjach, jakich można by dokonać, gdyby tytułem eksperymentu jako anonimowe potraktować teksty autorów polskich.

Krytyczne opinie Mullera i Bruneta nie podważają sensowności stosowania metod ścisłych w badaniu języka – przeczy temu zresztą ich własny, ogromny dorobek badawczy w tej dziedzinie. Przykłady udanych atrybucji potwierdzają, iż przy zachowaniu pewnych warunków wstępnych także kontrowersyjny problem autorstwa może zostać rozstrzygnięty. Wywód powyższy miał natomiast pokazać, że każdy, nawet pozornie trywialny problem językoznawczy, wymaga gruntownego opracowania filologicznego, a tok myślowy musi prowadzić od problemu do metody numerycznej, a nie odwrotnie. Porównanie spostrzeżeń Mullera i Bruneta z pewną liczbą cytowanych wyżej prac pokazuje, jak wiele można przeoczyć, odwracając ten porządek, a wraz z nim właściwą hierarchię ważności.

6. PRZEGLĄD METOD SEKWENCYJNEJ ANALIZY TEKSTU

Klasyfikacja dostępnych prac z zakresu sekwencyjnej analizy tekstu pozwala rozróżnić dwa podejścia metodologiczne: probabilistyczne i numeryczne. W pierwszym przypadku, tekst podlega jakiejś formie konwersji, nie traci jednak przy tym swego symbolicznego charakteru. Przyjmując na przykład, że jako jego relewantną cechę potraktuje się opozycję samogłosek (V), spółgłosek (C) i sonantów (S), a pominie pauzy międzywyrazowe, zdanie „Ogary poszły w las" można będzie zastąpić rządkiem VCVSVCVCSVCSVC. Metoda probabilistyczna pozwoli na obliczenie prawdopodobieństw przejścia pomiędzy stanami (tu określonymi jako V, C i S lub ich kombinacje), a także innych parametrów charakteryzujących między innymi siłę i głębokość związków kontekstowych w tekście. Przyjmując natomiast, że relewantną cechą tekstu są, na przykład, długości kolejnych zestrojów akcentowych wyrażone liczbą głosek, powyższe zdanie zostanie zakodowane jako sekwencja {5,5,4}. Przy podejściu numerycznym narzędzia badawcze pozwolą określić, czy kolejne wartości szeregu (oczywiście dłuższego) są statystycznie zależne, a jeżeli tak, to jaki model najlepiej opisuje ich powiązania kontekstowe. Zarówno przy podejściu probabilistycznym, jak i numerycznym możliwe jest też porównywanie i klasyfikowanie różnych fragmentów tekstu ze względu na ich własności sekwencyjne.

Przedstawiony tu podział na podejście numeryczne i probabilistyczne nie jest oczywiście wyczerpujący. W wielu przypadkach badacze tworzyli dla potrzeb jednostkowych analiz metody oryginalne, luźno powiązane z paradygmatem teorii prawdopodobieństwa czy procesów stochastycznych, bądź też mieszające różne podejścia. Z lingwistycznego punktu widzenia prace tego rodzaju bywają inspirujące jako źródło nowych hipotez, jednak o własnościach syntagmatycznych tekstu można dzięki nim wnioskować jedynie w sposób fragmentaryczny. Ponadto, jako nietypowe pod względem metodologicznym, nie zawsze dają się zestawić z innymi pracami w celu weryfikacji wyniku. Metodologia uznana przez przedstawicieli kilku dyscyplin naukowych staje się natomiast ogólnym standardem, ułatwiającym badania dzięki dostępności opracowań i oprogramowania oraz szerokiej bazie weryfikacyjnej. Jednorodność metodologiczna umożliwia też realizację postulatu redukcjonizmu (URBANEK 1987), poszerzającego zakres teorii i pozwalającego na daleko idące porównania. Biorąc pod uwagę te argumenty, nie uznano za celowe szczegółowego omawiania prac, których autorzy stosowali techniki niestandardowe[66], czyniąc wszakże wyjątek dla wybranych metod pomiaru spójności tekstu (por. 4.1).

Dotychczasowe doświadczenia wskazują, że najbardziej wszechstronnym i skutecznym narzędziem analizy sekwencyjnej tekstu jest metoda ARIMA. Pozwala ona na opis i porównywanie dowolnych szeregów czasowych, niezależnie od stylu tekstu, długości próby i rodzaju zastosowanej kwantyfikacji. Daje też łatwe w interpretacji wyniki, uwzględniające silnie stochastyczny charakter sekwencji tekstowych. Bardzo skuteczna

[66] Por. BAEVSKIJ&OSIPOVA 1987, BORODA 1994, FUCKS 1952, GROTJAHN 1979, KRASNOPEROVA 1987, HUG 1979, PÄÄKKÖNEN 1993, SKINNER 1941, WIOLAND 1985. Niektóre z tych prac zostały omówione w monografii PAWŁOWSKI 1998:61–65.

jest także metoda probabilistyczna, wykorzystująca aparat pojęciowy teorii informacji. W kolejnych rozdziałach omówione zostaną oba podejścia, jednak większość prezentowanych w części empirycznej zastosowań wykorzystuje metodę ARIMA. Na wstępie krótko omówimy także nieparametryczny test serii, służący między innymi do oceny losowości szeregu binarnego[67].

6.1 TEST SERII

Jedną z prostszych technik statystycznych pozwalających na wstępną ocenę losowości szeregu binarnego jest tzw. test serii (MOOD 1940, GROTJAHN 1980, BAVAUD 1998:206). Załóżmy, że w eksperymencie statystycznym otrzymano następujący wynik[68]:

(33) {1100101000111010100101010001110100101010010000000100000100111}

W szeregu takim ciąg symboli jednakowego typu określać będziemy jako *serię*. Liczba zer w szeregu (33) wynosi $n_0 = 36$, liczba jedynek $n_1 = 25$, a liczba serii $r = 35$. Zakłada się, że przy statystycznie losowym pojawianiu się symboli 0 i 1 (czyli ich równomiernym rozkładzie) liczba serii będzie oscylować wokół pewnej średniej, zależnej od wartości n_0 i n_1. Jako nieparametryczną hipotezę zerową H_0 przyjmuje się, że przy losowym uporządkowaniu symboli 0 i 1, liczbę serii *r* będzie najlepiej przybliżać rozkład normalny $N(m, s^2)$ o parametrach:

(34) $$m = \frac{2n_0 n_1}{n_0 + n_1} + 1$$

(35) $$s^2 = \frac{2n_0 n_1 (2n_0 n_1 - n_0 - n_1)}{(n_0 + n_1)^2 (n_0 + n_1 - 1)^2}$$

Zmienna *r* w postaci standaryzowanej ma postać:

(36) $$d = \frac{r - m}{s}$$

Jeżeli *d* znajdzie się poza przedziałem ufności na zadanym poziomie istotności α, hipotezę H_0 o losowym rozkładzie symboli w tekście można będzie odrzucić.

Dla $n_0 = 36$ i $n_1 = 25$, oczekiwana liczba serii przy losowym rozkładzie elementów w szeregu wynosi $m = 30{,}5$, a więc mniej niż w pseudolosowym szeregu (33). Korzystając ze wzorów (34–36) można obliczyć zmienną decyzyjną $d = 9{,}4$. Jej wartość znacznie wykracza poza przedział ufności [–1,96, 1,96] na poziomie istotności $\alpha = 0{,}05$ i wskazuje, że w szeregu (33), generowanym metodą „na chybił trafił" przez niżej podpisanego, liczba alternacji jest zbyt wysoka. Można obliczyć, że przy podanych wyżej wartościach

[67] Test serii stosuje się także do porównywania dwóch populacji – por. GREŃ 1987:492.

[68] Użyte symbole mogłyby oznaczać zarówno liczby, jak i cechy jakościowe (na przykład wynik rzutu monetą, wystąpienie określonej cechy w linii tekstu itd.).

n_0 i n_1, liczba serii w statystycznie losowym szeregu powinna wynosić 30 lub 31. Psychologicznie, rezultat ten jest jednak zgodny z oczekiwaniami. Jak zauważa F. Bavaud: „Il s'agit là d'une observation psychologique classique, maintes fois vérifiée dans les études de génération de hasard par les sujets humains, que l'on peut rapprocher de la croyance répandue qu'une série de plusieurs « 0 » consécutifs produite par une pièce de monnaie équilibrée tendra préférentiellement à être suivie d'un « 1 » plutôt que d'un autre « 0 » (*gambler's fallacy*)[69]." (BAVAUD 1998:207).

Jednak w przeciwieństwie do omawianych dalej metod analizy sekwencyjnej test serii nie uwzględnia kolejności pojawiających się jednostek, a jedynie rozkład długości jednakowych odcinków w szeregu. Oznacza to, że zmiana kolejności serii w szeregu, przy zachowaniu pozostałych parametrów, nie wpłynęłaby na wynik testu, mimo iż mogłaby być lingwistycznie istotna. Ponadto, stosowanie tego testu do szeregów bardziej zróżnicowanych (na przykład alfabetu) jest zdecydowanie trudniejsze i wymaga bardziej złożonego formalizmu matematycznego. W lingwistyce test serii może więc służyć jedynie jako miara spójności tekstu bądź jako wstęp do analizy sekwencyjnej, prowadzonej inną, skuteczniejszą metodą.

6.2 PODEJŚCIE PROBABILISTYCZNE

Analiza sekwencyjna bazująca na pojęciach teorii prawdopodobieństwa jest dziś obszernym działem matematyki, znajdującym zastosowania między innymi w akustyce, ekonometrii i telekomunikacji. Badania lingwistyczne stanowiące przedmiot niniejszej monografii nie wymagają jednak stosowania tak zróżnicowanej i rozbudowanej metodologii jak wymienione wyżej dyscypliny. Kryterium wyboru metody była dla nas jedynie jej efektywność i przydatność w badaniach lingwistycznych oraz, po spełnieniu tych warunków, dostępność źródeł i oprogramowania[70]. Omówiona niżej technika wykorzystuje shannonowską definicję informacji, opartą oczywiście na paradygmacie probabilistycznym, zgodnie z którym analizowany tekst nie musi podlegać kwantyfikacji (bada się rządki symboli, a nie liczb), rozszerza jednak jej stosowalność na sekwencje znaków dowolnej długości (tzw. *k*-gramy), czyniąc podstawowym narzędziem badawczym pojęcie *entropii*.

Elementy teorii informacji, z uwzględnieniem jej historycznego dorobku, zostały omówione we wcześniejszych rozdziałach (por. Część I, 5.2). Przypomnijmy, że w rozumieniu shannonowskim, informacja definiowana jest jako ilościowa miara niepewności

---

[69] W piśmiennictwie polskim na określenie tego psychologicznego mechanizmu używa się wyrażenia *sofizmat gracza*. C.R. Rao ilustruje go dość makabryczną anegdotą przypisywaną statystykowi G. Pólya: „Lekarz zwraca się do pacjenta takimi słowy: – Choruje Pan na bardzo poważną chorobę. Z dziesięciu chorych na nią tylko jeden przeżywa. Ale proszę się nie martwić. To szczęście, że przyszedł Pan do mnie, ponieważ ostatnio miałem dziewięciu pacjentów cierpiących na tę chorobę i wszyscy umarli." (RAO 1998:33).

[70] W omawianym przypadku użyto programu ENTROPIZER, opracowanego przez A. Xantosa i umieszczonego pod adresem URL: `http://www.unil.ch/ling` (stan na rok 2001). Jego opis dostępny jest także w formie drukowanej (XANTOS 2000).

związana z pojawieniem się danego symbolu:

(37) $$I_n = -\log_2 p_n$$

natomiast entropia źródła jest średnią ilością informacji niesioną przez symbol i dla kodu nierównomiernego, złożonego z *N* symboli, wynosi:

(38) $$H_1 = -\sum_{i=1}^{N} p_i \log_2 p_i$$

Wzór (38) wyraża średnią informację niesioną przez pojedynczy symbol kodu. Jednak badanie sekwencyjnej struktury tekstu ma w założeniu służyć opisowi i wyjaśnianiu związków symboli w linii tekstu. W celu przedstawienia siły i zasięgu takich związków można posłużyć się entropią warunkową, opartą na prawdopodobieństwie warunkowym pojawienia się danego symbolu (por. wzór 24). Jednak intuicyjnie, a przede wszystkim technicznie prostszym sposobem jest obliczenie entropii *k*-gramów (diad, triad itd.), a na tej podstawie entropii warunkowej i resztowej (BAVAUD 1998:212).

Niech we wzorze (38) p_i wyraża prawdopodobieństwo wystąpienia *i*-tego *k*-gramu w tekście utworzonym z symboli *N*-elementowego kodu, a *M* liczbę różnych *k*-gramów, które wystąpiły w tekście (zachodzi oczywiście relacja $M \leq N^k$). Przez entropię *k*-gramów rozumieć będziemy funkcję:

(39) $$H_k = -\sum_{i-1}^{M} p_i \log_2 p_i$$

Można teraz zdefiniować entropię warunkową rzędu *k*, wyrażającą nieprzewidywalność wystąpienia symbolu w sytuacji, gdy znane jest poprzedzających go $k-1$ symboli:

(40) $$h_k = H_k - H_{k-1}$$

oraz entropię resztową rzędu *k*, interpretowaną jako spadek niepewności związanej z pojawieniem się danego symbolu w sytuacji, gdy zamiast $k-1$ poprzedzających go symboli znanych jest *k* takich symboli:

(41) $$d_k = h_k - h_{k+1}$$

Im dłuższy lewostronny kontekst $k-1$, tym mniejsza nieprzewidywalność (a więc i entropia) nieznanego symbolu. Można wykazać, że entropia warunkowa, obliczona według wzoru (40), spełnia relację[71]:

(42) $$h_0 \geq h_1 \geq h_2 \geq \ldots \geq h_k \geq \ldots \geq h_\infty \quad \text{(przy czym } h_0 = H_0 \text{ i } h_1 = H_1\text{)}$$

Nierówność (42) wskazuje, iż spadek entropii dla rosnącego *k* jest monotoniczny i można opisać jego dynamikę, analizując kształt krzywej, którą utworzyłyby kolejne

[71] Por. HAMMERL&SAMBOR 1990:373–375, PAWŁOWSKI 1998:195 oraz wzór 25 s.40.

wartości h_k. Można też domniemywać, że dla pewnego *k*, wartości h_k ustabilizują się, a d_k spadną do zera. Wartość *k*, przy której to nastąpi, wskaże rząd procesu stochastycznego (a więc głębokość związku kontekstowego), którego realizacją jest badany tekst. Określenie rzędu procesu, a więc minimalnej liczby realizacji, których znajomość pozwoli efektywnie obniżyć niepewność związaną z pojawieniem się kolejnego symbolu, może zostać oszacowana na zadanym poziomie istotności (BAVAUD 1998:214). Dla szeregu długości *n*, zawierającego *N* różnych symboli, efektywna estymacja rzędu procesu dopuszcza wartości *k* nie większe niż:

$$(43) \qquad k = \mathrm{int}\left[\frac{\log_2 n}{\log_2 N}\right] \quad \text{dla} \quad N > 2 \quad \text{i} \quad k = \mathrm{int}[\log_2 n] \quad \text{dla} \quad N = 2$$

W celu określenia *k*, należy wysunąć dwie alternatywne hipotezy:

H_0: proces jest rzędu *k*
H_1: proces jest rzędu *k*+1

Odrzucamy H_0 na korzyść H_1 na poziomie istotności α jeżeli zachodzi[72]:

$$(44) \qquad \chi^2_{emp} = 2\,(n-k)\,d_{k+1} \ln 2 \geq \chi^2_\alpha\left[N^k\,(N-1)^2\right]$$

gdzie $\chi^2_\alpha[i]$ – wartość rozkładu *chi* ^{2} na poziomie istotności α przy *i* stopniach swobody
n – długość badanego szeregu
d_i – entropia resztowa rzędu *i*

Przykładowej analizie poddano fragment łacińskiego heksametru[73]. Założono, iż jego struktura rytmiczna mogła opierać się zarówno na iloczasie, jak i dynamicznym akcencie metrycznym określanym jako *ikt*. Sylaby długie zakodowano jako „D", krótkie jako „K", natomiast sylaby akcentowane i nie akcentowane oznaczono odpowiednio symbolami „A" i „N". Tym sposobem uzyskano dwa szeregi symboli odpowiadające jednemu fragmentowi tekstu. Symbole „N" i „A" bądź „D" i „K", a więc *de facto* odpowiadające im typy sylab, można określić jako *stany* układu i oznaczyć przez E_k. Jeżeli na przykład *n*-ta sylaba szeregu jest akcentowana, symbolicznie notujemy $E_n = A$. Prawdopodobieństwo sekwencji $E_j \to E_k$ określa się mianem prawdopodobieństwa przejścia i oznacza przez p_{ik}. Prawdopodobieństwa przejścia obliczone dla danego układu tworzą macierz prawdopodobieństw przejścia (FELLER 1987:340). W omawianym przypadku wstępnym etapem

[72] Zasady testowania hipotez statystycznych, w szczególności użytego tu testu *chi* ^{2}, omówione są m.in. w pracy HAMMERL&SAMBOR 1990:291, 305.
[73] Hor. *Ars* 147–156. Ponieważ omawiany tu przykład ma jedynie ilustrować zastosowanie określonej metody badawczej, komentarze filologiczne ograniczono do minimum. Obszerny opis tej problematyki, z podaniem zasad kodowania tekstu łacińskiego, znajduje się w części materiałowej niniejszej pracy (Część II, 4).

analizy było sporządzenie macierzy prawdopodobieństw przejścia dla obu szeregów z podaniem w nawiasach częstości tych *k*-gramów, które wystąpiły w tekście (Tab. 12 i 13). Na wejściu macierzy uwzględniono także dłuższe sekwencje znaków (diady i triady).

Pobieżna analiza tabeli 12 pozwala na wyciągnięcie następujących wniosków:

1. Akcentuacja sylab heksametru łacińskiego jest silnie determinowana lewostronnym kontekstem (sylaba akcentowana zawsze wymusza pojawienie się sylaby nie akcentowanej, a sekwencja dwóch sylab nie akcentowanych wymusza w sposób konieczny pojawienie się sylaby akcentowanej). Oznacza to, że metrum to jest realizacją pewnego procesu stochastycznego;
2. Prawdopodobny rząd procesu jest niski, ponieważ już dwie poprzedzające sylaby skutecznie determinują trzecią.

Tab. 12 Macierz prawdopodobieństw przejścia dla heksametru jako sekwencji akcentowej[74]

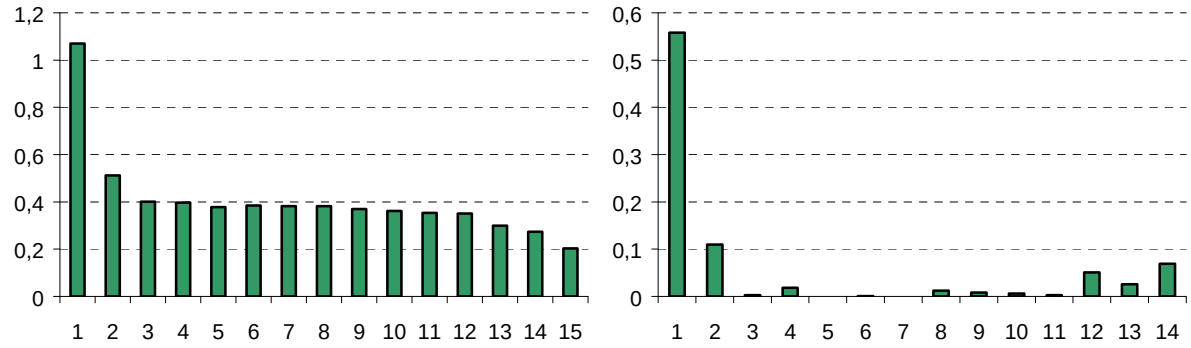
	N	A		N	A
N (84)	0,29	0,71	NNA (59)	1,00	0,00
A (60)	1,00	0,00	NAN (35)	0,40	0,60
NN (60)	0,00	1,00	ANN (24)	0,00	1,00
NA (59)	1,00	0,00	ANA (24)	1,00	0,00
AN (24)	0,41	0,59			

Spostrzeżenia te znajdują potwierdzenie w przebiegach funkcji entropii warunkowej h_k i resztowej d_k (Rys. 4). Stabilizację entropii warunkowej h_k zaobserwować można począwszy od odstępu $k=3$, co oznacza, że dowolny symbol jest statystycznie determinowany już przez dwa poprzedzające go symbole. Uwzględnianie szerszego lewostronnego kontekstu jest oczywiście możliwe, ale nie przyniesie istotnego obniżenia niepewności co do rodzaju kolejnego symbolu.

Zgodnie ze wzorem (43), przy długości szeregu $n=144$ efektywna estymacja rzędu procesu możliwa jest o ile $k \leq 7$. W oparciu o wzór (44), hipotezę H_0: *proces jest rzędu k na poziomie istotności* $\alpha = 0{,}01$ odrzucamy przy $k=0$, gdyż $\chi^2_{emp.} = 111{,}4 \geq \chi^2_{0{,}01}[1] = 6{,}6$, a także przy $k=1$, gdyż $\chi^2_{emp.} = 21{,}8 \geq \chi^2_{0{,}01}[2] = 9{,}2$, a przyjmujemy dopiero przy $k=2$, gdyż $\chi^2_{emp.} = 5{,}9 < \chi^2_{0{,}01}[4] = 13{,}3$. Oznacza to, że zgodnie z tym, co w oczywisty sposób sugerują wykresy na Rys. 4, proces stochastyczny, którego realizacją jest obserwowany szereg, będzie najprawdopodobniej rzędu drugiego.

[74] Lektura macierzy prawdopodobieństw przejścia przebiega zawsze od lewej do prawej. Na przykład liczba 0,41 w dolnym wierszu jest prawdopodobieństwem przejścia AN → ANN, czyli pojawienia się sylaby nie akcentowanej po sekwencji sylab akcentowanej i nie akcentowanej. Należy podkreślić, że w idealnym przypadku macierz prawdopodobieństw przejścia powinna być macierzą kwadratową, zawierającą prawdopodobieństwa przejścia z, i do każdego stanu. W tym przypadku, ze względu na specyfikę kodu językowego, zasady tej nie przestrzegano (punktem wyjścia są także diady i triady symboli).

Rys. 4 Entropia warunkowa (wykres lewy) i resztowa (wykres prawy) dla heksametru łacińskiego kodowanego jako sekwencja akcentów[75]



Bardziej złożony obraz otrzymamy analizując heksametr kodowany w postaci sekwencji iloczasowej (Tab. 13). Macierz prawdopodobieństw przejścia sugeruje w takim przypadku, iż:

1. Długość sylab heksametru łacińskiego determinowana jest lewostronnym kontekstem (na przykład sekwencja dwóch sylab krótkich wymusza pojawienie się sylaby długiej). Oznacza to, że badany tekst jest realizacją pewnego procesu stochastycznego;
2. Prawdopodobny rząd procesu (czyli głębokość związku kontekstowego) jest wysoki, ponieważ dla pewnych kombinacji nawet trzy sylaby nie determinują w sposób konieczny kolejnej, czwartej sylaby.

Tab. 13 Macierz prawdopodobieństw przejścia dla heksametru jako sekwencji iloczasowej

	K	D		K	D
K (91)	0,45	0,55	KDK (24)	0,62	0,38
D (53)	0,33	0,67	KDD (24)	0,00	1,00
KK (60)	0,00	1,00	DKK (15)	0,00	1,00
KD (29)	0,47	0,53	DKD (15)	0,40	0,60
DK (29)	0,83	0,17	DDK (13)	1,00	0,00
DD (24)	0,25	0,75	DDD (5)	0,33	0,67
KKD (45)	0,48	0,52			

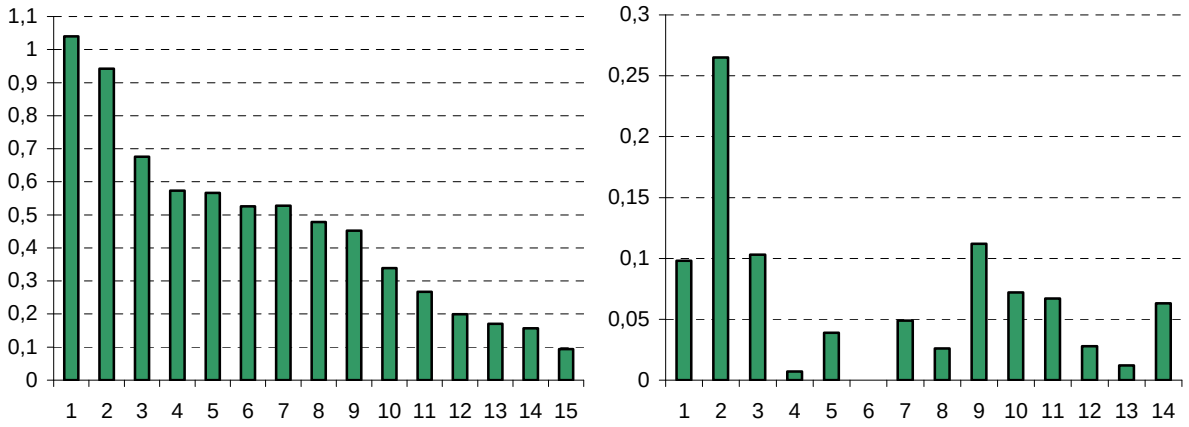
Jak z powyższego wynika, dokładniejsze określenie struktury i zasięgu związków kontekstowych w heksametrze kodowanym iloczasowo wymaga narzędzia badawczego lepiej syntetyzującego informację. Kryterium takie spełniają opisane wyżej funkcje entropii.

Silny spadek entropii warunkowej następuje dla $k = 4$, a jej stabilizacja dopiero począwszy od wartości $k = 12$ (Rys. 5). Można więc przyjąć, że do względnie dobrego określenia rodzaju kolejnej sylaby w tekście potrzeba co najmniej trzech poprzedzających

[75] Na osiach rzędnych umieszczono wartości h_k i d_k, na osiach odciętych wartości k (por. wzory 40 i 41).

sylab, a efektywna predykcja może ich wymagać nawet kilkanaście (dopiero począwszy od $k=15$ entropia warunkowa ma rzeczywiście niską wartość).

Rys. 5 Entropia warunkowa (wykres lewy) i resztowa (wykres prawy) dla heksametru łacińskiego kodowanego jako sekwencja iloczasowa



Podobnie jak w poprzednim przypadku, efektywna estymacja rzędu procesu możliwa jest dla $k \leq 7$ (wzór 43). W oparciu o wzór (44), hipotezę H_0: *proces jest rzędu k* na poziomie istotności $\alpha = 0{,}01$ odrzucamy przy $k=0$, gdyż $\chi^2_{emp.} = 19{,}5 \geq \chi^2_{0,01}[1] = 6{,}6$, przy $k=1$, gdyż $\chi^2_{emp.} = 52{,}5 \geq \chi^2_{0,01}[2] = 9{,}2$, oraz przy $k=2$, gdyż $\chi^2_{emp.} = 20{,}3 \geq \chi^2_{0,01}[4] = 13{,}3$. Hipotezę H_0 przyjmujemy dopiero przy $k=3$, gdyż $\chi^2_{emp.} = 1{,}4 < \chi^2_{0,01}[8] = 20{,}1$. Oznacza to, że proces stochastyczny, którego realizacją jest obserwowany szereg iloczasowy, może być efektywnie opisany modelem rzędu trzeciego. W tym kontekście wysokie wartości d_k dla $k \approx 10$ nie mogą być uznane za statystycznie znaczące. Trudno jednak na podstawie jednego testu osądzić, czy ich pojawienie się było kwestią przypadku. Naszym zdaniem mogło ono wynikać z powtarzalności jakiegoś dłuższego wzorca rytmicznego w wersach heksametru. Ich długość sylabiczna i uporządkowanie są wprawdzie zmienne, ale tylko w pewnych granicach: badany tekst złożony jest jedynie ze spondejów i daktyli (a więc stóp dwu- i trzysylabowych), i to w takich proporcjach, że przeciętna długość wersu wynosi około piętnastu sylab (por. Część II, 4). Hipoteza ta, przedstawiona w tym miejscu jedynie w charakterze przykładu, wymagałaby jednak szczegółowej analizy metrycznej i weryfikacji na większej liczbie fragmentów.

Porównanie wykresów 4 i 5 wskazuje, iż w heksametrze łacińskim zależności kontekstowe iloczasów są bardziej złożone niż analogiczne zależności sylab akcentowanych i nie akcentowanych dynamicznie (o ile oczywiście zaakceptuje się hipotezę o istnieniu *iktu*). Przypuszczalnie porządek akcentowy jest bardziej przewidywalny (a więc rytmiczny) od porządku iloczasowego. Z metodologicznego punktu widzenia widać natomiast, że entropia jest miarą opisującą w sposób syntetyczny i efektywny sekwencyjną strukturę tekstu kodowanego jako ciąg symboli i w przypadku testowania nowych hipotez może oddać duże usługi. Należy jednak pamiętać, że zbyt duża liczba stanów może

utrudnić obliczenia i prezentację wyniku. W takiej sytuacji warto rozważyć rezygnację z metody probabilistycznej na korzyść metody numerycznej, w której ograniczenie to nie występuje.

6.3 PODEJŚCIE NUMERYCZNE

Omówiona w poprzednim rozdziale metoda stosowana jest przy analizie szeregów kategorialnych złożonych z elementów określanych jako zdarzenia bądź stany[76]. Przy podejściu numerycznym zakłada się, że dostępne dane mają postać dyskretnych szeregów czasowych, a więc ciągów liczb stanowiących realizację pewnej zmiennej losowej. Warto przypomnieć, że te same teksty mogą być podstawą generowania zarówno szeregów kategorialnych, jak i numerycznych: w pierwszym przypadku, jednostki językowe kodowane są jako stany – zachowują więc charakter symboliczny; w przypadku drugim, zastąpione zostają wartościami mierzalnej cechy, uważanej za relewantną ze względu na testowaną hipotezę (por. Część I, 1.1.2).

Podstawowym narzędziem badawczym używanym przez nas przy podejściu numerycznym jest metoda ARIMA G. Boxa i G. Jenkinsa. Jej pierwsze lingwistyczne zastosowania, ujęte w perspektywie przeglądowo-historycznej, omówiono wcześniej (Część I, 5.4). W tym miejscu przedstawione zostaną jej formalne podstawy, oczywiście w zakresie niezbędnym do zrozumienia wyników prac materiałowych omawianych w następnych rozdziałach. Jednym z argumentów przemawiających za taką formą prezentacji jest fakt, iż prowadzone dotąd badania tekstu wykorzystują jedynie skromny fragment rozbudowanego aparatu metody ARIMA. Z przyczyn oczywistych nie korzysta się na przykład z technik prognozowania i sterowania zaprojektowanych na potrzeby ekonomii i inżynierii, nie było też konieczne użycie analizy spektralnej (od zasady tej uczyniono jeden wyjątek – por. Cześć II, 1). Należy podkreślić, że analiza szeregów czasowych była już przedmiotem kilku opracowań o charakterze podręcznikowym adresowanych między innymi do praktyków nauk społecznych, głównie socjologii i psychologii[77]. Właśnie w naukach społecznych metoda Boxa i Jenkinsa cieszy się jak dotąd największym powodzeniem. P.S. Nurius porównała spotykane w literaturze przedmiotu techniki analizy danych sekwencyjnych uporządkowanych na osi czasu. Najbardziej rozpowszechnioną okazała się właśnie ARIMA: „ARIMA modelling is perhaps the most commonly encountered and widely used of several stochastic process models adapted for use with time-series data." (NURIUS 1983:222). Warto w tym kontekście nadmienić, że propozycje Boxa i Jenkinsa nie spotkały się w środowisku statystyków z jednoznacznie pozytywnym

[76] „Zamiast mówić: «wynikiem *n*-tej próby jest zdarzenie E_k », będziemy mówić, że w momencie *n* układ znajduje się w stanie E_k ." (FELLER 1987:340)

[77] Praca Boxa i Jenkinsa z 1970 r. (przekład polski BOX&JENKINS 1983) jest tekstem „kanonicznym" metody ARIMA. Godne polecenia są opracowania o charakterze podręcznikowym, zawierające niekiedy oprogramowanie: BROCKWELL&DAVIES 1991 i 1996, CHAGHAGHI 1985, COURTROT&DROESBEKE 1984: 67–76, GLASS et al. 1975, GOTTMAN 1981, MCCLEARY&HAY 1980, MONTGOMERY&JOHNSON 1976: 188–240, NURIUS 1983, PAWŁOWSKI 1998, STIER 1989, WHITELEY 1980.

przyjęciem[78]. Jak jednak wskazuje powyższy cytat, poparty solidnym przeglądem bibliograficznym, kontrowersje te nie wpłynęły na efektywność metody w badaniach z zakresu nauk społecznych. Pozytywne rezultaty testów prowadzonych na danych tekstowych pozwalają sądzić, że w badaniach lingwistycznych metoda ARIMA okaże się narzędziem równie skutecznym.

6.3.1 Definicja szeregu czasowego

Szeregiem czasowym nazywać będziemy każdy ciąg liczb będący realizacją pewnej zmiennej losowej X_t. Zmienna niezależna *t* reprezentująca tradycyjnie czas rzeczywisty, zastąpiona jest w badaniach tekstu tzw. czasem syntagmatycznym odpowiadającym sekwencyjnemu uporządkowaniu jednostek językowych (PAWŁOWSKI 1998:4). Pojęciu chwili na osi czasu rzeczywistego odpowiada więc pojęcie pozycji w linii tekstu. Interwał oddzielający realizacje szeregu w chwilach lub na pozycjach t_i i t_j określać będziemy jako odstęp (ang. *lag*) i oznaczać symbolem $k = t_j - t_i$. W tej konwencji, realizacje x_{ti} i x_{tj} notowane będą jako x_t i x_{t+k}.

6.3.2 Stacjonarność szeregów czasowych

Ważną cechą szeregów czasowych generowanych z tekstu jest stacjonarność. Przez stacjonarność rozumieć należy swoistą stabilność szeregu wyrażającą się brakiem trendu i stałością parametrów niezależnie od tego, które odcinki szeregu zostały uwzględnione w obliczeniach. Rozróżniamy trzy rodzaje stacjonarności (PRISTLEY 1981:112):

– szereg czasowy jest stacjonarny w sensie ścisłym, jeżeli jego autokowariancja oraz wszystkie momenty statystyczne[79] zależą jedynie od odstępu *k*
– szereg czasowy jest stacjonarny rzędu *s*, jeżeli jego autokowariancja oraz momenty statystyczne rzędu $m < s$ zależą jedynie od odstępu *k*
– szereg czasowy jest stacjonarny w sensie szerokim, jeżeli jego średnia jest stała, natomiast autokowariancja zależy jedynie od odstępu *k*

Metoda ARIMA pozwala na analizę dowolnych szeregów czasowych, także niestacjonarnych. Wiadomo jednak, że długości jednostek językowych (na przykład słowoform) oraz wartości innych mierzalnych parametrów tekstu nie mogą być dowolne – mieszczą się w pewnych granicach. Owa stabilność ilościowych cech badanych jednostek językowych sprawia, że szeregi uzyskiwane poprzez kwantyfikację tekstów mogą być *a priori*

[78] „Box i Jenkins przyjęli zasadę, że książka [*Analiza szeregów czasowych* – A.P.] ma być dostępna dla czytelnika niemal całkowicie surowego w statystyce i matematyce, w wyniku czego zdecydowali się na daleko idące odstępstwo od ścisłości definicji i rozważań. [...] Takie postępowanie może czasami dezorientować czytelnika, tym bardziej, że autorzy odstąpili od formułowania i dowodzenia jakichkolwiek twierdzeń. Jesteśmy jednak przekonani, że książka może oddać duże usługi przy analizie szeregów czasowych, jeśli tylko dostanie się w ręce myślącego i krytycznego czytelnika." (ze wstępu do pracy BOX&JENKINS 1983:8).

[79] Definicję momentu statystycznego znaleźć można m.in. w pracy GREŃ 1987:46.

traktowane jako stacjonarne w sensie szerokim. Takich właśnie szeregów dotyczyć będzie dalszy opis metody.

6.3.3 Podstawowe parametry stacjonarnych szeregów czasowych

Średnią μ_x szeregu czasowego X_t jest wyrażenie:

(45) $$\mu_x = E(X_t)$$

którego estymatorem jest:

(46) $$m_x = \frac{1}{N}\sum_{t=1}^{N} x_t$$

gdzie N – długość szeregu
x_t – wartość szeregu w momencie lub na pozycji t

Wariancję szeregu czasowego X_t definiuje się jako:

(47) $$\sigma_x^2 = E(X_t - \mu_x)^2$$

Estymatorem wariancji (przy tych samych oznaczeniach) jest wielkość:

(48) $$s_x^2 = \frac{1}{N}\sum_{t=1}^{N}(x_t - m_x)^2$$

Specyficzną cechą szeregów czasowych jest możliwość definiowania ich autokowariancji i autokorelacji. Funkcje te mają duże znaczenie przy identyfikacji i estymacji modeli procesów stochastycznych, których realizacją są badane szeregi czasowe. Ponadto, wyrażają one statystyczne powiązanie realizacji szeregu oddalonych o ustalony odstęp k i nawet bez estymacji jakiegokolwiek modelu mogą posłużyć jako miary ich sekwencyjnego uporządkowania.

Autokowariancję szeregu czasowego X_t przy odstępie k definiujemy jako:

(49) $$\gamma_k = E\{(X_t - \mu_x)(X_{t+k} - \mu_x)\}$$

Estymatorem funkcji autokowariancji jest funkcja:

(50) $$c_k = \frac{1}{N-k}\sum_{t=1}^{N-k}(x_t - m_x)(x_{t+k} - m_x)$$

Autokorelację szeregu (ACF – ang. *autocorrelation function*) można określić jako unormowaną funkcję autokowariancji:

(51) $$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_k}{\sigma_x^2}$$

a jej estymatorem jest funkcja:

(52) $$r_k = \frac{c_k}{c_0} = \frac{c_k}{s_x^2}$$

Nietrudno zauważyć, że dla $k = 0$ zachodzi $\gamma_0 = \sigma_x^2$, a więc autokowariancja staje się wariancją i korelacja szeregu „z samym sobą" równa jest jedności ($\rho_0 = 1$). W praktyce analizy stacjonarnych szeregów czasowych najczęściej wykorzystuje się właśnie funkcje autokorelacji i, dodatkowo, autokorelacji cząstkowej (PACF – ang. *partial autocorrelation function*). Kształt obu funkcji pozwala wnioskować o typie procesu stochastycznego, którego realizacją jest obserwowany szereg. PACF wyprowadzana jest z równań Yule'a-Walkera (BOX&JENKINS 1983:71), stosowanych do estymacji współczynników modelu autoregresji, i zostanie przedstawiona w rozdziale 6.3.5.

6.3.4 Wybrane modele liniowe szeregów stacjonarnych

Omawiane niżej modele szeregów czasowych są szczególnymi przypadkami tzw. ogólnego procesu liniowego (BOX&JENKINS 1983:54–59). Określa się je jako modele (bądź filtry) liniowe, ponieważ każda wartość szeregu x_t jest liniową kombinacją wartości szeregu (model AR) bądź realizacji procesu losowego (model MA) w chwilach (na pozycjach) poprzedzających *t*. Szczególnym typem szeregu jest proces czysto losowy.

Szeregiem losowym nazywać będziemy ciąg statystycznie niezależnych realizacji zmiennej losowej $X_t = \{e_1, e_2, ...\}$. Z uwagi na niezależność e_i, autokowariancja i autokorelacja szeregu losowego przyjmą wartości:

(53) $$\gamma_k = \begin{cases} \sigma_e^2 & \text{dla } k = 0 \\ 0 & \text{dla } k \neq 0 \end{cases} \qquad \rho_k = \begin{cases} 1 & \text{dla } k = 0 \\ 0 & \text{dla } k \neq 0 \end{cases}$$

Poprzez analogię do spektrum świetlnego, szereg wartości e_i posiadających rozkład normalny $N(0,1)$ określany jest jako biały szum (ang. *white noise*), podczas gdy szeregi losowe o innych rozkładach bywają nazywane szumami kolorowymi.

Szeregiem autoregresji rzędu *p*, oznaczanym AR(*p*), nazywać będziemy ciąg wartości x_t opisywanych modelem o postaci[80]:

(54) $$x_t = a_1 x_{t-1} + a_2 x_{t-2} + ... + a_p x_{t-p} + e_t$$

gdzie a_i – współczynniki modelu
e_i – wartości losowe o rozkładzie $N(0,1)$

W kategoriach lingwistycznych rząd procesu autoregresji odpowiada głębokości związku kontekstowego, określanej również jako „pamięć tekstu". Autokorelacja procesu

[80] Modele szeregów czasowych zapisuje się w formie „jawnej" bądź za pomocą operatorów. W podanych niżej definicjach zastosowano zapis jawny, czyniąc wyjątek jedynie dla modeli sezonowych. W części materiałowej stosowane będą oba systemy notacji.

typu AR(p) może być zdefiniowana jako funkcja rekurencyjna o postaci (BOX&JENKINS 1983:62):

(55) $$\rho_k = a_1\rho_{k-1} + a_2\rho_{k-2} + ... + a_p\rho_{k-p}$$

Z własności tej wynika, iż ρ_k będzie funkcją gasnącą dla rosnącego k. Na przykład autokorelację procesu AR(1) można przedstawić, drogą kolejnych podstawień w równaniu (55), jako funkcję wykładniczą $\rho_k = a_1^k$. Fakt ten ma istotne znaczenie przy identyfikacji typu procesu i wyborze modelu.

Szeregiem średniej ruchomej rzędu q, oznaczanym MA(q), nazywać będziemy ciąg wartości x_t opisywanych modelem o postaci:

(56) $$x_t = e_t - b_1e_{t-1} - b_2e_{t-2} - ... - b_qe_{t-q}$$

gdzie b_i – współczynniki modelu
e_i – wartości szeregu losowego o rozkładzie $N(0,1)$

Interpretacja modelu MA(q) w kategoriach lingwistycznych jest trudniejsza. Mamy tu co prawda do czynienia z pewną formą zależności wartości szeregu od kontekstu lewostronnego, ponieważ dla każdego x powtarza się ta sama kombinacja współczynników b_i. Jednak poprzedzające realizacje szeregu nie są w modelu bezpośrednio uwidocznione, a co ciekawsze, równanie (56) jest filtrem liniowym przekształcającym w uporządkowany i do pewnego stopnia deterministyczny szereg $\{x_1, x_2, ...x_t...\}$ wartości losowe e_t. Otóż, związek realizacji procesu ruchomej średniej z „przeszłością" szeregu istnieje dzięki *odwracalności* modeli AR i MA. Na mocy tej własności każdemu procesowi typu AR(p) odpowiada proces MA(∞), natomiast każdemu procesowi MA(q) odpowiada proces AR(∞). Ponadto, patrząc na to zjawisko z szerszej perspektywy, warto zauważyć, że modelowanie zjawisk deterministycznych za pomocą liczb losowych (używając nieco pretensjonalnej metafory, proces ten można by określić jako swoistą transformację *chaosu* w *kosmos*...) jest we współczesnej matematyce i w naukach przyrodniczych częstą i, co ważniejsze, niezwykle skuteczną procedurą[81].

Można wykazać, że funkcja autokorelacji (ACF) szeregu ruchomej średniej MA(q) wyraża się wzorem (BOX&JENKINS 1983:75):

(57) $$\rho_k = \begin{cases} \dfrac{-b_k + b_1b_{k+1} + b_2b_{k+2} + ... + b_{q-k}b_q}{1 + b_1^2 + b_2^2 + ... + b_q^2} & \text{dla } k \le q \\ 0 & \text{dla } k > q \end{cases}$$

Jak widać, wartości ACF są różne od zera jedynie dla $k \le q$. Podczas analizy rzeczywistych szeregów czasowych fakt ten wykorzystuje się przy identyfikacji procesu i wyborze optymalnego modelu.

[81] Więcej informacji na ten temat znaleźć można w pracach GLEICK 1987, STEWART 1994 i PEITGEN et al. 1997.

Szeregiem mieszanym typu ARMA(*p*, *q*) nazywać będziemy ciąg wartości opisywanych następującym modelem:

(58) $$x_t = a_1 x_{t-1} + a_2 x_{t-2} + ... + a_p x_{t-p} - b_1 e_{t-1} - b_2 e_{t-2} - ... - b_q e_{t-q} + e_t$$

gdzie a_i – współczynniki modelu AR
b_i – współczynniki modelu MA
e_i – wartości szeregu losowego o rozkładzie $N(0,1)$

Jak widać, modele mieszane są addytywną kombinacją modeli prostych. Praktyka pokazuje, że stosuje się je stosunkowo rzadko, najczęściej ze względu na kryterium oszczędności modelu (ang. *parsimony*), a więc w sytuacji, gdy modele proste zawierają zbyt dużą liczbę parametrów. Jak zauważają R. McCleary i R. Hay: „If our experiences are typical, only few social science time series in a thousand will have both p and $q \neq 0$." (MCCLEARY&HAY 1980:64). Dodatkowe warunki muszą też spełniać współczynniki tych modeli. Obaj autorzy zauważają, że szereg generowany modelem ARMA(1,1) będzie losowy, jeżeli jego współczynniki będą mieć jednakową wartość $a_1 = b_1$, a zredukuje się do modeli prostych przy niewielkiej różnicy współczynników $a_1 \approx b_1$ (*ibid*. 1980:65). Funkcja autokorelacji modelu mieszanego jest gasnąca, podobnie jak w przypadku modelu prostego AR(*p*), ale nie ma charakteru jednorodnego[82]. Pierwsze *q* wartości ACF modelu mieszanego zależne jest od składowej MA, pozostałe wartości ACF zależą od składowej AR. Względy te są dodatkowym utrudnieniem w stosowaniu i interpretacji modeli mieszanych (por. BOX&JENKINS 1983:80–86, MCCLEARY&HAY 1980:64–66).

Szeregiem sezonowym określa się szereg, którego realizacje wykazują regularności okresowe ze stałym odstępem *s*. Modele sezonowe stosuje się w ekonometrii, gdzie wiele zjawisk rynkowych (na przykład ceny) wykazuje regularność w cyklach miesięcznych lub rocznych. Zapewne fakt ten zaskoczy zarówno ekonomistę, jak i filologa, ale właśnie modele sezonowe okazały się w lingwistyce kwantytatywnej narzędziem wyjątkowo skutecznym. Zaskoczenie to będzie jednak mniejsze, jeżeli zestawimy dwie następujące definicje: „Ogólnie mówimy, że szereg zachowuje się w sposób okresowy z okresem *s*, jeżeli po *s* jednostkach czasu szereg wykazuje podobne właściwości." (BOX&JENKINS 1983:303) i „Wiersz jest to wypowiedź swoiście zorganizowana. Tak mianowicie, że jej rozczłonkowanie wytwarza odcinki ekwiwalentne pod względem wyboru i układu określonych elementów językowych." (MAYENOWA 1963:8). Pomijając różnice stylistyczne, zauważalne jest wyraźne podobieństwo obu definicji, pozwalające na zastosowanie matematycznych modeli szeregów sezonowych w analizie wersyfikacji.

Szereg określa się jako sezonowy SARMA$(P, Q)_s$, jeżeli opisujący go model ma postać:

(59) $$x_t = a_s x_{t-s} + a_{2s} x_{t-2s} + ... + a_{ps} x_{t-ps} - b_s e_{t-s} - b_{2s} e_{t-2s} - ... - b_{qs} e_{t-qs} + e_t$$

[82] Własności funkcji ACF dla szeregu mieszanego opisują m.in. G. Box i G. Jenkins (1983:81–82).

gdzie a_i – współczynniki modelu AR
b_i – współczynniki modelu MA
s – odstęp sezonowy
e_i – wartości szeregu losowego o rozkładzie $N(0,1)$

Jeżeli przyjmiemy, iż sezonowość szeregu najlepiej opisują modele proste pierwszego rzędu SAR(1)$_s$ lub SMA(1)$_s$, wzór (59) znacznie się uprości. Na przykład model SARMA(1,1)$_s$ miałby postać:

(60) $$x_t = a_s x_{t-s} - b_s e_{t-s} + e_t$$

Pewnym ułatwieniem notacji jest też zapis operatorowy, tym bardziej, że obserwowane procesy sezonowe na ogół zawierają w sobie także procesy proste, a całość wyrażana jest w postaci jednego modelu. I tak, operatorem cofającym B rzędu s określa się przekształcenie:

(61) $$B^s x_t = x_{t-s}$$

W tej konwencji szeregi SAR(1)$_s$ i SMA(1)$_s$ można zapisać jako:

(62) $$e_t = (1 - a_s B^s) x_t$$

(63) $$x_t = (1 - b_s B^s) e_t$$

Natomiast szereg złożony SARMA(1,1)$_s$ będzie mieć postać:

(64) $$x_t (1 - a_s B^s) = e_t (1 - b_s B^s)$$

Model ogólny stacjonarnego procesu sezonowego SARMA(P, Q)$_s$ zapiszemy jako:

(65) $$x_t (1 - a_s B^s - a_{2s} B^{2s} ... - a_{Ps} B^{Ps}) = e_t (1 - b_s B^s - b_{2s} B^{2s} - ... - b_{Qs} B^{Qs})$$

Jednak w dotychczasowych badaniach wersyfikacji obserwowano zazwyczaj kombinacje procesów prostych i sezonowych, w których odstęp sezonowy s równy był długości wersu. W przypadku ogólnym, modele opisujące takie procesy zapisywane są jako SARMA(p, q)(P, Q)$_s$, gdzie p i q oznaczają rzędy normalnych składowych autoregresji i ruchomej średniej, natomiast P i Q rzędy autoregresji i ruchomej średniej składowych sezonowych. Jak już wspomniano, w praktyce składowa sezonowa jest najczęściej zredukowana do modelu prostego rzędu pierwszego. W przeciwieństwie do modeli mieszanych, złożone modele sezonowe nie są jednak addytywne, ale multiplikatywne. Właśnie z tego względu zapis operatorowy, nie wymagający mnożenia czynników, jest w ich przypadku bardziej dogodny. Typowy model mieszany zawierający składową sezonową SARMA(1,0)(1,0)$_s$ miałby w tej konwencji postać:

(66) $$(1 - a_1 B)(1 - a_s B^s) x_t = e_t$$

Jak widać, po wykonaniu mnożenia, w modelu tym pojawi się składnik $a_1 a_s B^{s+1}$, który nie występowałby w modelu addytywnym. Jednak z uwagi na fakt, iż współczynniki modeli stacjonarnych przyjmują wartości z przedziału [–1, 1], składniki takie nie osiągają znaczących wartości.

Sezonowość szeregu zakłada, że realizacje w chwilach t będą silnie powiązane z realizacjami w chwilach $t \pm s$. Fakt ten ma wpływ na kształt funkcji autokorelacji. ACF jako funkcja odstępu *k* posiadać będzie znaczące wartości dla *k* równych wielokrotnościom odstępu sezonowego. Na przykład autokorelacja odpowiednio zakodowanego trzynastozgłoskowca przyjmować będzie znaczące wartości dla $k = 13$, $k = 26$ itd. Cecha ta ułatwia identyfikację szeregu sezonowego. Należy dodać, że Box i Jenkins oraz autorzy innych konsultowanych prac nie podają analitycznych wyprowadzeń funkcji ACF dla szeregów sezonowych.

6.3.5 Identyfikacja i estymacja parametrów modelu

Identyfikacji modelu dokonuje się na podstawie kształtu funkcji autokorelacji (ACF) i autokorelacji cząstkowej (PACF). ACF została zdefiniowana już wcześniej (wzory 51 i 52), natomiast osobnego wyjaśnienia wymaga PACF. W statystyce funkcja ta jest bardzo użyteczna, ponieważ w układzie wielu zmiennych losowych pozwala na badanie bezpośredniej korelacji dwóch wybranych zmiennych bez uwzględniania niepożądanego wpływu pozostałych zmiennych (BOURSIN 1981:37). W analizie szeregów czasowych PACF wyraża związek nie sąsiadujących ze sobą wartości szeregu (x_t i x_{t+k} gdzie $k > 1$) bez uwzględniania wpływu wartości pośrednich ($x_{t+1}, x_{t+2}, ..., x_{t+k-1}$).

Omawiając procesy autoregresji, funkcję ACF przedstawiono w postaci równania rekurencyjnego o postaci:

(67) $$\rho_k = a_1 \rho_{k-1} + a_2 \rho_{k-2} + ... + a_p \rho_{k-p}$$

Wykorzystując symetryczność funkcji ACF (a więc własność, dzięki której $\rho_{-k} = \rho_k$), równanie (67) można rozwinąć w układ *k* równań, zwanych równaniami Yule'a-Walkera (BOX&JENKINS 1983:62), zapisanych w formie matrycowej jako:

(68) $$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ ... \\ \rho_k \end{bmatrix} = \begin{bmatrix} \rho_0 & \rho_1 & \rho_2 & ... & \rho_{k-1} \\ \rho_1 & \rho_0 & \rho_1 & ... & \rho_{k-2} \\ \rho_2 & \rho_1 & \rho_0 & ... & \rho_{k-3} \\ ... & ... & ... & ... & ... \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & ... & \rho_0 \end{bmatrix} * \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ ... \\ a_k \end{bmatrix}$$

Po zastąpieniu ρ_k wartościami r_k obliczanymi z próby, układ równań (68) można rozwiązać ze względu na a_k i tym sposobem otrzymać estymowane wartości nieznanych współczynników modelu AR[83]. Równanie (68) pozwala jednak także na obliczenie auto-

[83] Przy znajdowaniu współczynników innych typów procesów schemat postępowania jest podobny,

korelacji cząstkowej szeregu, niezależnie od tego, czy potraktujemy go jako realizację procesu MA, czy AR. Ponieważ układ równań 68 rozwiązuje się dla kolejnych wartości *k*, otrzymanym współczynnikom a_i można dopisać drugi indeks, notując je jako a_{ik}. Na przykład rozwiązanie dla $k = 3$ dałoby a_{13}, a_{23} i a_{33}. Przez funkcję autokorelacji cząstkowej szeregu czasowego rozumieć będziemy tak zdefiniowane wartości współczynników a_{kk}, traktowane jako funkcja *k*. PACF posiada bardzo przydatną własność, a mianowicie urywa się na odstępie $p+1$, jeżeli szereg jest typu AR(*p*), a wygasa jak funkcja wykładnicza lub gasnąca sinusoida, jeżeli szereg jest typu MA(*q*). W połączeniu z dokładnie odwrotnym zachowaniem funkcji ACF, nasz wywód zmierzający do przedstawienia metody identyfikacji szeregu można podsumować prostym schematem:

Tab. 14 Identyfikacja prostych modeli liniowych[84]

	ACF	PACF
AR(*p*)	wygasa	urywa się na odstępie *p*+1
MA(*q*)	urywa się na odstępie *q*+1	wygasa
ARMA(*p*, *q*)	wygasa	wygasa

Problemem pojawiającym się przy identyfikacji procesu na podstawie powyższej tabeli jest oszacowanie istotności funkcji ACF i PACF. Ich wartości nie zawsze bowiem urywają się w sposób wyraźny i nie budzący wątpliwości. M.S. Bartlett (1946) pokazał, że błąd standardowy współczynników autokorelacji liczonych z próby długości *N* wynosi:

$$S(r_k) = N^{-0,5}\left(1 + 2\sum_{i=1}^{k-1} r_i^2\right)^{0,5} \tag{69}$$

natomiast M.H. Quenouille (1947) wykazał, że błąd standardowy dla współczynnika PACF liczonego na podstawie szeregu długości *N* wynosi $N^{-0,5}$.

W praktyce autokorelacje uznaje się za statystycznie istotne, o ile ich wartości leżą poza przedziałem dwóch błędów standardowych określanych jako tzw. wstęga Bartletta (ang. *Bartlett band*). W następnych rozdziałach przedział ten będzie rysowany bezpośrednio na wykresach funkcji autokorelacji.

Zgodnie z przedstawionymi wcześniej założeniami ostatnim etapem procedury badawczej jest ocena stopnia dopasowania modelu do obserwowanych danych. Zmienną decyzyjną jest w tym przypadku procent wariancji szeregu obserwowanego wyjaśnionej przez model. W celu jego obliczenia generuje się tzw. szereg resztowy, stanowiący ciąg różnic pomiędzy wartościami obserwowanymi a generowanymi na podstawie modelu, a następnie porównuje jego wariancję z wariancją szeregu obserwowanego. Należy się spodziewać, że model dobrze dopasowany będzie pozostawiał szereg resztowy o niewiel-

---

tzn. za pomocą obserwowanych wartości autokorelacji (bądź autokorelacji cząstkowej) estymuje się nieznane współczynniki modelu. Algorytmów tych jednak nie omawiamy.

[84] Por. MONGOMERY&JOHNSON 1976:208 oraz GOTTMAN 1981:142.

kiej wariancji. Oznaczając wariancję szeregu obserwowanego przez s^2_{obs}, a wariancję szeregu resztowego przez s^2_r, poszukiwany procent V_e obliczymy na podstawie wzoru:

(70) $$V_e = 100\%\left(1 - \frac{s_r^2}{s_{obs}^2}\right)$$

Przyjmuje się, że im wyższa jest wartość V_e, tym lepiej dobrany jest model. Wariancja resztowa (s^2_r) interpretowana jest bowiem jako ta część całkowitej zmienności (a więc informacji) zawartej w szeregu, której proponowany model nie wyjaśnia. W idealnym przypadku model powinien odfiltrować z szeregu obserwowanego tyle informacji, by pozostający szereg resztowy był całkowicie losowy. Jednak z wartości V_e wnioskować też można o bardzo istotnej własności tekstu, który był podstawą wygenerowania szeregu czasowego. Jest bowiem tak, że szeregi rytmiczne są łatwiejsze w modelowaniu równaniami autoregresji czy ruchomej średniej, podczas gdy szeregi arytmiczne modelować jest znacznie trudniej. *Ipso facto*, im większy rzeczony procent wariancji wyjaśnionej przez model, tym regularniejszy albo bardziej rytmiczny obserwowany szereg (w domyśle tekst). Parametr V_e jest więc niezwykle istotny z lingwistycznego punktu widzenia – uważać go można za **syntetyczną miarę sekwencyjnego uporządkowania tekstu**. Zupełnie inną i, naszym zdaniem, nierozstrzygalną na gruncie naukowym kwestią pozostaje natomiast interpretacja tego wskaźnika w kategoriach estetyki.

7. OGRANICZENIA METODY

Metody ilościowe nie są magiczną różdżką, której dotknięciem można w sposób ścisły i bezbłędny rozwiązać wszystkie problemy lingwistyki. Wymagają one wielkiego nakładu pracy, dyscypliny badawczej i, jak wszystkie metody, mają swoje ograniczenia, o których nie wolno zapominać podczas interpretacji wyników.

W trakcie prowadzonych analiz (por. Część II) pojawiły się problemy praktyczne, które nie pozostają bez wpływu na uzyskany wynik. Badane teksty reprezentowane były przez szeregi czasowe uzyskane drogą kwantyfikacji relewantnych cech jednostek językowych. Ponieważ wiele uwagi poświęcono zagadnieniom prozodii, pojawił się znany w lingwistyce problem kwantyfikacji pauz międzywyrazowych. Po długich deliberacjach zrezygnowano z ich kwantyfikacji, uznając, że 1) ich rozkład i długość są tylko w niewielkim stopniu przewidywalne, a przez to charakteryzują raczej konkretne odczytanie niż tekst stanowiący jego podstawę; 2) ich potencjalna wartość nie jest porównywalna ze skalą stosowaną przy kwantyfikacji akcentu dynamicznego lub długości sylab.

Wypada też zwrócić uwagę na problem ogólniejszy. Nie od dziś wiadomo, że jedną z podstawowych barier lingwistyki kwantytatywnej, a więc także analizy sekwencyjnej, jest nieunikniona redukcja wielowarstwowej struktury języka do jednego, najczęściej formalnego poziomu i wynikająca z tego znaczna utrata informacji. Intuicja podpowiada nam, że najważniejszymi, chociaż trudno uchwytnymi atrybutami jednostek językowych

– morfemów, leksemów czy zdań – nie są wcale ich cechy formalne (na przykład długość), ale cechy semantyczne rozpatrywane w odniesieniu do pojedynczych jednostek oraz większych całości. Problem polega jednak na tym, że mimo intensywnych prac nad formalną reprezentacją wiedzy, mimo badań lingwistycznych uwzględniających słownikowe znaczenia pojedynczych leksemów (SAMBOR 1997) efektywna kwantyfikacja znaczenia nie jest możliwa. Z ograniczeniem tym można się uporać, wychodząc z założenia, że niektóre warstwy języka – fonematyczna, prozodyczna – nie posiadają znaczenia referencjalnego, natomiast innych warstw nie da się efektywnie opisać, nie izolując ich, przynajmniej we wstępnej fazie badań. Dopiero dysponując szczegółowymi modelami, można poszukiwać uogólnień, powołując się na epistemologiczno-metodologiczny postulat redukcjonizmu (URBANEK 1987). Jak bowiem zauważa K. Popper, bądź co bądź krytyk redukcjonizmu skrajnego: „Metoda nauki polega na naszych próbach opisania świata za pomocą prostych teorii: teorie złożone mogą okazać się niesprawdzalne, nawet jeżeli są prawdziwe. Nauka to sztuka systematycznych uproszczeń – sztuka określania tego, co możemy z korzyścią dla siebie pominąć.” (POPPER 1996:70).

Z kolei redukcja indywidualnych odczytań tekstu do jego uproszczonej reprezentacji (na przykład zastąpienie go binarną sekwencją sylab akcentowanych i nie akcentowanych) realizuje metodologiczny postulat *idealizacji* (KRAJEWSKI 1998:104). Jest więc formą abstrahowania służącego konstrukcji ogólnego wzorca lub modelu, wyrażonego w języku matematyki. Wspomniana utrata informacji, która w tym procesie następuje, rekompensowana jest szerszym zakresem modelu skonstruowanego na bazie takiej idealizacji. Warto w tym miejscu nadmienić, że w lingwistyce *idealizację* można określić za K. Bühlerem jako zasadę abstrahującej relewancji (BÜHLER 1934). Faktycznie, zasada ta jest odpowiednikiem (i uszczegółowieniem) koncepcji typów idealnych M. Webera (WEBER 1973). Jednak kres uogólnień leży tak naprawdę jeszcze dalej: „Jako ciekawostkę można przytoczyć fakt, że Weber i niektórzy jego następcy sądzili, iż tworzenie typów idealnych jest osobliwością nauk humanistycznych, nie zdając sobie sprawy z tego, że idealizacja jest podstawową metodą fizyki od czasów Galileusza.” (KRAJEWSKI 1988:113)

Inną problematyczną kwestią jest sposób pojmowania losowości i determinizmu. Można mianowicie podać przykłady funkcji deterministycznych generujących liczby losowe, a także funkcji lub algorytmów prowadzących od wartości losowych do wyników całkowicie przewidywalnych[85]. Sytuacja ta stawia w dość kłopotliwym położeniu tych wszystkich badaczy (nie wyłączając niżej podpisanego), którzy w wysuwanych przez siebie hipotezach posługują się pojęciem losowości, przeciwstawionym jakiejś formie determinizmu. Cóż bowiem znaczyć ma stwierdzenie, iż dany szereg jest losowy, skoro, przynajmniej teoretycznie, możliwe jest wygenerowanie go przy pomocy jakiejś funkcji?

[85] O funkcjach pierwszego rodzaju wspomina C.R. Rao (1994:41–43), powołując się na prace M. Kaca. Przykładami drugiego rodzaju są tzw. *gry w chaos*. Intrygujący przykład algorytmu generującego z liczb losowych strukturę w pełni deterministyczną (chodzi o fraktal określany jako *trójkąt Sierpińskiego*) opisany jest w pracy PEITGEN et al. 1997:378–383.

I na odwrót, jak rozumieć stwierdzenie, że dany szereg zawiera składową deterministyczną, skoro, potencjalnie, istnieć może algorytm generujący ów deterministyczny szereg z liczb losowych?

Naszym zdaniem, odpowiedzi na powyższe wątpliwości należy szukać wychodząc od kuhnowskiej koncepcji paradygmatu naukowego rozumianego jako jednolita rama pojęciowa i metodologiczna, akceptowana w danej epoce przez środowisko naukowe i stanowiąca swoisty wzorzec dla wszystkich teorii naukowych (KUHN 1968, por. też WĄSIK 1987:26). Obserwując rozwój nauk przyrodniczych w ostatnich dziesięcioleciach, zauważyć można narodziny nowego paradygmatu opartego na teorii tzw. nieliniowych układów dynamicznych. Przyjęcie postulatów tej teorii prowadzi do daleko idącej redefinicji pojęć losowości i determinizmu, tak istotnych w sekwencyjnej analizie tekstu. Zachodzi więc pytanie, czy lingwistyka kwantytatywna nie powinna podążyć tym samym torem[86].

Aby uniknąć zbyt ogólnych ocen, ograniczmy się do sekwencyjnej analizy tekstu, w której kluczowymi pojęciami są właśnie losowość i determinizm. Naszym zdaniem, kwestionowane ostatnio tradycyjne znaczenia tych terminów powinny pozostać integralnym składnikiem obowiązującego paradygmatu lingwistyki kwantytatywnej, ponieważ dobrze odpowiadają intuicyjnemu rozumieniu elementarnych, przednaukowych kategorii poznawczych, takich jak rytm, porządek, symetria, chaos. Ten swoisty antropocentryzm lingwistyki kwantytatywnej znajdować będzie uzasadnienie tak długo, jak długo podmiotem, a zarazem adresatem wiedzy o języku pozostanie człowiek. Rysujący się powoli nowy formalizm lingwistyki kwantytatywnej skłania natomiast do postawienia niełatwego pytania o granicę pomiędzy konwencją a „naturą rzeczy” w pojmowaniu losowości i determinizmu w języku.

[86] Wysoce pouczająca w kwestii łatwych zapożyczeń metodologicznych jest prześmiewcza praca A. Sokala i J. Bricmonta (1998), wskazująca na opłakane skutki mechanicznego i bezmyślnego mieszania aparatu pojęciowego nauk ścisłych i humanistyki.

II. BADANIA MATERIAŁOWE

1. PORÓWNANIE STRUKTURY RYTMICZNEJ NIEKTÓRYCH ODMIAN STYLISTYCZNYCH I WERSYFIKACYJNYCH POLSZCZYZNY[87]

Jak już wspomniano, struktury linearne występują w tekście na wszystkich poziomach analizy, jednak szczególnie dogodnym obszarem badań sekwencyjnych jest prozodia i rytmika. Kwantyfikacja lub kodowanie warstwy rytmicznej najczęściej redukuje akcentuację (iloczas, poziom tonalny) do dyskretnego szeregu złożonego z niewielkiej liczby prostych elementów lub wartości. Metody nie uwzględniające ich porządku mają w takiej sytuacji bardzo ograniczone możliwości – można obliczyć i porównać jedynie proste parametry pozycyjne (średnią, rozrzut), które zresztą zachowują stabilne wartości w dostatecznie długich próbach. Znacznie bogatsza informacja płynie z analizy następstwa akcentów lub tonów w linii tekstu. Teoretycznie możliwe są bowiem wszystkie uporządkowania, można też mówić o stopniowalności zrytmizowania tekstu.

Przedmiotem prezentowanych poniżej analiz jest struktura rytmiczna kilku odmian stylistycznych i wersyfikacyjnych polszczyzny. Warunkiem rozpoczęcia badań empirycznych było przyjęcie następujących ogólnych założeń:

1. Możliwa jest segmentacja i kodowanie tekstu jako sekwencji sylab akcentowanych i nie akcentowanych dynamicznie (a więc metodą binarną);
2. Kodowanie to (lub kwantyfikacja) jest reprezentatywne dla relewantnych językowo zjawisk, na przykład prozodii lub rytmiki tekstu;
3. Szeregi czasowe wygenerowane tym sposobem nie są sekwencjami losowymi;
4. Parametry ich składowych deterministycznych, takie jak zasięg i siła związku kontekstowego, pozwalają odróżnić od siebie style lub odmiany wersyfikacyjne;
5. Rzeczone parametry mogą w pewnych przypadkach charakteryzować style osobnicze;
6. Te same zasady obowiązują w innych językach o podobnych systemach prozodii;

Celem przeprowadzonych analiz było znalezienie formalnych modeli rytmicznej struktury tekstu i wnioskowanie na ich podstawie o istotnych z lingwistycznego punktu widzenia własnościach badanych stylów. Użyto opisanej wcześniej metody ARIMA, rozszerzając w tym jednym przypadku jej podstawowy zakres o funkcję periodogramu, stanowiącą podstawowe narzędzie analizy widmowej dyskretnych szeregów czasowych.

1.1 BADANE TEKSTY I KWANTYFIKACJA

Testom poddano fragmenty wiersza sylabotonicznego i sylabicznego oraz prozy artystycznej i dyskursu oratorskiego[88]. Jako hipotezę przyjęto, iż struktura rytmiczna wiersza sylabotonicznego będzie najbardziej regularna, mniej rytmiczny będzie wiersz sylabiczny,

[87] Wykorzystano dane empiryczne opublikowane w pracy PAWŁOWSKI 1997.
[88] Szczegółowe dane bibliograficzne na temat wykorzystanych tekstów podano w ANEKSIE.

a jeszcze mniej regularne uporządkowanie znajdziemy w obu odmianach tekstów pozbawionych wersyfikacji. O ile jednak kolejność silnie kontrastujących odmian wersyfikacyjnych i stylistycznych wynika w sposób analityczny z ich cech strukturalnych, o tyle trudno osądzić *a priori*, czy proza artystyczna (tu powieść czytana przez zawodowego lektora) będzie posiadać budowę regularniejszą od dyskursu oratorskiego. Istnieją przesłanki przemawiające zarówno za taką tezą, jak i przeciw niej. Jeżeli zachowanie sylabotonika i wiersza sylabicznego okaże się zgodne z przewidywaniami, uwiarygodni tym samym wnioski wyciągnięte na podstawie analizy pozostałych stylów.

Każda odmiana stylistyczna i wersyfikacyjna reprezentowana była przez trzy próby o przeciętnej długości siedmiuset sylab. Wyjątkiem był wiersz sylabotoniczny, w przypadku którego całkowicie wystarczyły szeregi o średniej długości dwustu pięćdziesięciu sylab. Kwantyfikację przeprowadzono w oparciu o skalę porządkową, złożoną z wartości 0 (sylaba nie akcentowana), 0,5 (akcent poboczny) i 1 (sylaba akcentowana). Na równych zasadach traktowano akcent wyrazowy, logiczny i emocjonalny. Ze względu na obecność pośredniej wartości przypisanej akcentowi pobocznemu, skala ta formalnie nie może być określona jako binarna, jednak faktycznie taką właśnie jest, ponieważ akcent poboczny występował niezmiernie rzadko. Pauzy międzywyrazowe nie były uwzględniane. Podstawą kwantyfikacji były nagrania tekstów w profesjonalnych, aktorskich interpretacjach (patrz ANEKS). Wszystkie zapisy były wyrywkowo weryfikowane przez inne osoby, co czyni je bardziej wiarygodnymi.

Analiza rytmiki tekstu w proponowanej tu wersji wyrasta w prostej linii z nurtu fonologii. Przy opisie materiału językowego nie stosuje się kategorii fizycznych, lecz skale oparte na pewnym abstrakcyjnym wzorcu, który realizuje się w konkretnych aktach komunikacji. Foniczne kontinuum zostaje więc zredukowane do dyskretnej skali kontrastujących ze sobą elementów binarnych, a złożony charakter niektórych cech prozodycznych jest ignorowany[89]. Fakt ten nie wywołuje jednak utraty relewantnej językowo informacji. Binarność jako podstawa kodowania rytmiki tekstu wynika bowiem z odporności na zakłócenia i prostoty kodów dwuwartościowych, spotykanych powszechnie zarówno w technice, jak i w przyrodzie. Argument o fonologicznej genezie proponowanego tu podejścia przemawia też za metodą percepcyjną, a przeciwko stosowaniu urządzeń elektro-akustycznych, które rozpoznają bardzo dokładnie parametry fizyczne dźwięku, w tym również zupełnie dla nas nieistotne cechy indywidualne podmiotu mówiącego, pomijają za to segmentację tekstu na jednostki językowe, takie jak sylaby, morfemy, wyrazy, stopy metryczne itd. Mimo pewnej redukcji informacji, przyjęta metoda kwantyfikacji respektuje więc powszechnie akceptowaną normę komunikacyjną i kulturową[90].

Innym uzasadnieniem takiego sposobu kwantyfikacji jest stosowanie podobnej skali w metryce i wersologii: „A binary «either/or» choice [...] is a reasonable approximation

[89] Akcentowi dynamicznemu może towarzyszyć podniesienie tonu i wydłużenie samogłoski.
[90] Warto zwrócić uwagę, że rozwój metryki i wersologii trwa nieprzerwanie od starożytności, a brak urządzeń technicznych nigdy nie stanął temu na przeszkodzie.

of what happens in speech, and has been used in several major studies of metrics, both traditional and generative." (BRATLEY&ROSS 1981:42–43). Podobnie postąpili M. Azar i B. Kedem (AZAR&KEDEM 1979) badając sekwencyjną strukturę języka hebrajskiego. Autorzy przypisali zera i jedynki pewnym klasom fonemów (bezdźwięcznym i dźwięcznym, samogłoskom i spółgłoskom itd.), a następnie poddali analizie sekwencyjnej uzyskane szeregi czasowe. Uzyskane przez nich rezultaty dowodzą, iż stosowanie binarnej skali porządkowej do konwersji tekstu na szereg czasowy jest całkowicie dopuszczalne i może posłużyć jako podstawa wnioskowania o własnościach tekstu[91].

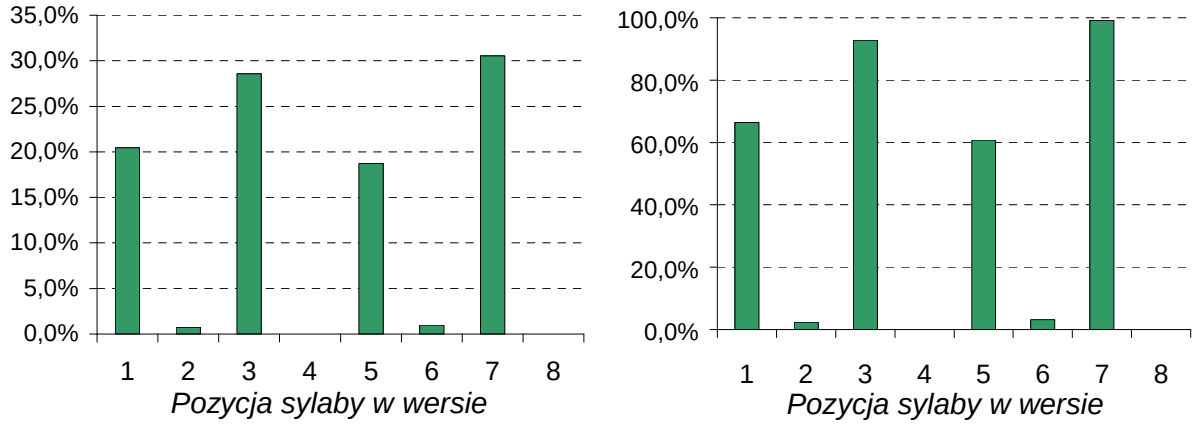
WIERSZ SYLABOTONICZNY

Za przykład wiersza sylabotonicznego posłużyły fragmenty bajki *Opowiedział dzięcioł sowie* J. Brzechwy (patrz ANEKS). Struktura tego utworu jest bardzo regularna (ośmiozgłoskowiec z średniówką po czwartej sylabie), wręcz monotonna, i zmusza do bardzo rytmicznej deklamacji. W założeniu, wszystkie wersy składają się z czterech trochejów, chociaż, jak pokazuje poniższy fragment, w konkretnych interpretacjach reguła ta nie musi być przestrzegana w sposób ortodoksyjny (ANEKS – BRZECHWA 1965:42):

Każdy ptak i każde zwierzę	⊥	–	⊥	–	⊥	–	⊥	–
W swej komorze coś wybierze	–	–	⊥	–	⊥	–	⊥	–
I natychmiast to przyniesie	–	–	⊥	–	–	–	⊥	–
Jako udział w interesie	–	–	⊥	–	–	–	⊥	–

Aby w bardziej syntetyczny sposób ukazać regularność badanego sylabotonika, obliczono procentowy rozkład przycisków na kolejnych sylabach wersu (Rys. 6).

Rys. 6 Rozkład przycisków akcentowych w ośmiozgłoskowcu sylabotonicznym



Po stronie lewej podano procentowy rozkład akcentów na poszczególne sylaby, po stronie prawej, procent sylab akcentowanych na danej pozycji (średnia 125 losowo

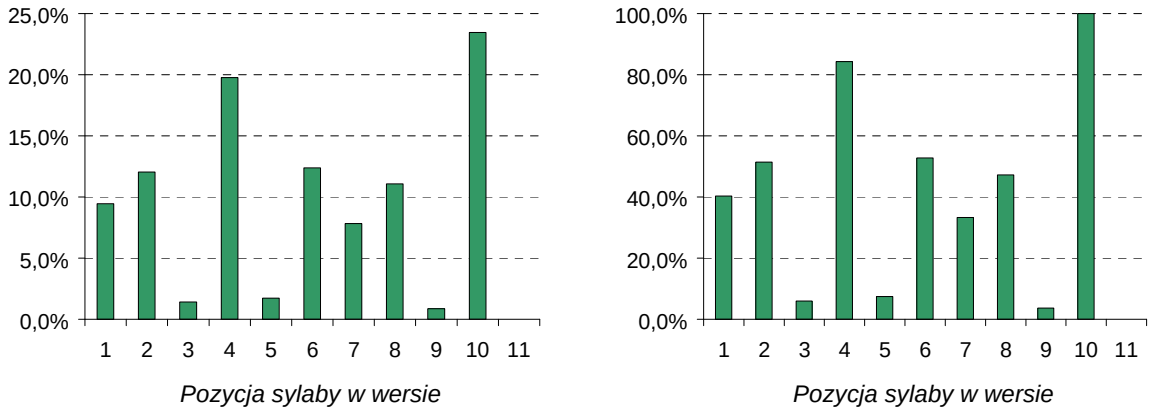
[91] Więcej informacji na temat kodowania tekstu wierszowanego znaleźć można w pracach CAIRNS et al. 1981, CHRISHOLM 1981 oraz LOGAN 1982.

wybranych wersów). Jak widać, na czterech pozycjach w wersie znajduje się w sumie ponad 98% sylab akcentowanych (wykres lewy), przy czym sylaby trzecia i siódma są akcentowane częściej.

WIERSZ SYLABICZNY

Kolejnym analizowanym utworem jest *Beniowski* J. Słowackiego. Na wybór tego poematu dygresyjnego napisanego jedenastozgłoskowcem miały wpływ dwa argumenty. Pierwszym była jego struktura wersyfikacyjna oparta na sylabizmie, a więc systemie mniej regularnym od sylabotonicznego, chociaż w znacznym stopniu przewidywalnym. Argumentem drugim była stylistyczno-formalna jednorodność utworu, pozwalająca na wygenerowanie praktycznie dowolnej liczby szeregów czasowych o odpowiedniej długości (w naszym przypadku każda z prób liczy około 700 sylab). Średni rozkład przycisków w wersie (na podstawie 216 losowo wybranych wersów) przedstawiony jest na wykresie 7. Podobnie jak w przypadku wiersza sylabotonicznego, wykres lewy zawiera procentowy rozkład akcentów padających na poszczególne sylaby w wersie, a wykres prawy procent sylab akcentowanych na danej pozycji.

Rys. 7 Rozkład przycisków akcentowych w jedenastozgłoskowcu sylabicznym



Oba wykresy potwierdzają nasze przypuszczenia co do struktury wersu w wierszu sylabicznym. Za konstanty stabilizujące rytm uznać należy pozycje 3, 4, 5 oraz 9, 10 i 11, czyli sylaby poprzedzające klauzulę i średniówkę. Sylaby na pozycjach 1, 2, 6, 7, 8, akcentowane średnio w co drugim wersie, są natomiast źródłem zmienności rytmicznej wiersza, którą dokładnie zbadać można tylko metodą analizy sekwencyjnej[92].

PROZA ARTYSTYCZNA

Analizę rytmu prozy przeprowadzono na fragmentach powieści I. Newerlego *Wzgórze Błękitnego Snu* (patrz ANEKS). Wybór tego akurat tekstu nie był podyktowany żadnymi

[92] Podobne wyniki uzyskały Z. Kopczyńska i L. Pszczołowska (1986).

specjalnymi względami o podłożu merytorycznym. Chodziło nam jedynie o zbadanie reprezentatywnych fragmentów współczesnej prozy artystycznej, respektującej konwencję gatunku i unikającej eksperymentów stylistycznych (szczególnie w warstwie prozodii). Podobnie jak w poprzednim przypadku, tekst kodowano na podstawie interpretacji aktorskiej.

DYSKURS ORATORSKI

W przeciwieństwie do pojęć statystyki, style językowe nie dają się zdefiniować w sposób ścisły i jednoznaczny. W naszym przypadku, przez dyskurs oratorski rozumiemy zbiór tekstów obejmujący przede wszystkim (choć nie jedynie) mowy polityczne, sądowe, kaznodziejskie oraz, w mniejszym stopniu, język reklamy. Oprócz informowania, teksty tego rodzaju mają za zadanie przekonywać, a w odróżnieniu od tekstów prozatorskich czy potocznych, silnie nasycone są elementami retorycznymi. Aby uniknąć niepotrzebnych konotacji pozajęzykowych, a zarazem pozostać w obszarze żywego języka i dysponować dostatecznie długimi fragmentami tekstu, jako reprezentatywną dla tej odmiany potraktowano jedną z homilii papieskich (patrz ANEKS). Argumentem, który zaważył na naszej decyzji, był specyficzny kontekst sytuacyjny, w którym ów tekst zaistniał. Wszak osoba zwracająca się na żywo do wielotysięcznej rzeszy słuchaczy kładzie na ogół szczególny nacisk na komunikatywność i siłę perswazji swego wywodu. Dysponuje przy tym rozmaitymi środkami retorycznymi i stylistycznymi, wśród których jest właśnie nadawanie mowie pożądanego rytmu poprzez odpowiednią konstrukcję zdań i akcentowanie. Nie bez znaczenia był również fakt, że zgodnie z powszechną opinią, polszczyzna Jana Pawła II uchodzi za wzorcową.

O wpływie tych czynników na strukturę prozodyczną tekstu świadczy poniższy fragment homilii, stanowiący pointę poprzedzającego wywodu. Uwagę zwraca parokrotne akcentowanie pierwszej sylaby słów trzy- i czterosylabowych, odbiegające od typowego dla polszczyzny systemu paroksytonicznego. Może to wynikać z podporządkowania zasad prozodii polskiej wymogom retoryki, może też być spowodowane czynnikiem emocjonalnym, mogło wreszcie pojawić się jako interferencja języka włoskiego, w którym dość często akcent pada na pierwszą sylabę słowa, a którym szef Państwa Watykańskiego posługuje się na co dzień:

W ewangelicznym przykazaniu miłości	– ⊥ – ⊥ – – – ⊥ – – ⊥ –
tkwi bowiem najgłębsze źródło	⊥ ⊥ – – ⊥ – ⊥ –
duchowego rozwoju każdego człowieka	⊥ – – – ⊥ – – ⊥ – – ⊥ ⊥ –

Rozczłonkowanie powyższej kwestii na trzy zestroje intonacyjne, mimo że w jakimś stopniu arbitralne, respektuje miejsca wystąpienia pauz w tekście. Warto zwrócić uwagę na liczbę i rozkład akcentów w każdym wersie, przypominające swym układem wersyfikację wiersza tonicznego. Ta jednostkowa obserwacja zostanie zweryfikowana w trakcie obliczeń przeprowadzonych na dłuższych fragmentach.

1.2 REZULTATY

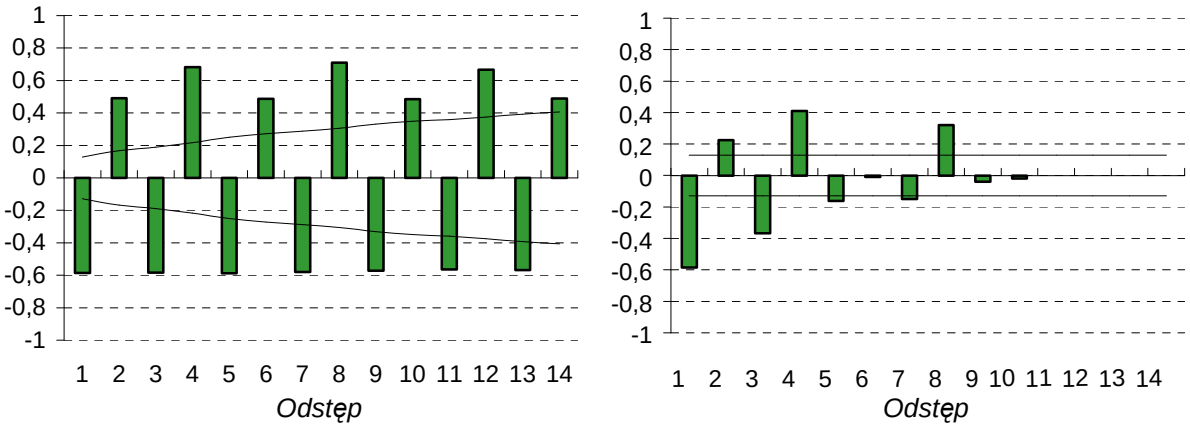
Celem naszej analizy było w pierwszej kolejności sprawdzenie, czy badane szeregi czasowe (a więc odcinki tekstu) zawierają jakieś składowe deterministyczne. Posłużono się w tym celu funkcją autokorelacji. Dla szeregów zawierających składowe deterministyczne określono typ i rząd procesów (na podstawie kształtu funkcji ACF i PACF), a następnie estymowano ich modele. Dodatkowo przedstawiono także wykres funkcji periodogramu. Parametrem liczbowym, który pozwolił na porównanie wyników i ich językoznawczą interpretację był procent wariancji szeregu obserwowanego wyjaśniony przez model (V_e).

WIERSZ SYLABOTONICZNY

Zgodnie z oczekiwaniami, najbardziej wyraziste procesy stochastyczne znaleziono w wierszu sylabotonicznym J. Brzechwy. Prezentujemy poniżej przykład analizy sekwencyjnej jednego z fragmentów *Opowiedział dzięcioł sowie* (ANEKS – BRZECHWA 1965:34–35). Struktura rytmiczna pozostałych próbek pochodzących z tego utworu była bardzo podobna – obserwacje i wnioski odnoszą się więc do całego wiersza.

Funkcje autokorelacji i autokorelacji cząstkowej wskazują na silnie deterministyczny charakter szeregu. ACF właściwie nie gaśnie, natomiast PACF gaśnie powoli (Rys. 8):

Rys. 8 Autokorelacja (wykres lewy) i autokorelacja cząstkowa (wykres prawy) sekwencji przycisków akcentowych w wierszu sylabotonicznym[93]



Znając postać funkcji ACF i PACF, można przystąpić do wyboru prawdopodobnego modelu procesu. Wbrew pozorom, nie jest to łatwe, ponieważ uwzględnić należy zarówno modele proste, jak i sezonowe. Co najmniej dwa argumenty przemawiają za zastosowaniem w tym przypadku modelu sezonowego: pierwszym jest sylabotonizm wiersza wymuszający regularne powtarzanie akcentów na stałych pozycjach w wersie, drugim jest charakterystyczny dla procesów sezonowych kształt funkcji ACF i PACF. Z drugiej strony, korelacja szeregu (pozytywna dla odstępów parzystych i negatywna dla nieparzy-

[93] Linie ciągłe na wykresach funkcji ACF i PACF wyznaczają przedziały ufności (tzw. wstęga Bartletta) (por. Część I, 6.3.5).

stych) jest tak silna, że nie należy wykluczać sytuacji, w której wystarczającym okazałby się jakiś model prosty. Z tych względów serię testów rozpoczęto od modeli prostych, przechodząc stopniowo do modeli sezonowych (Tab. 15).

Tab. 15 Identyfikacja modelu rytmiki wiersza sylabotonicznego

Typ modelu ($s_0^2 = 0{,}234$)	s_r^2	V_e	N
AR(1)	0,154	34%	1
AR(2)	0,147	37%	2
MA(1)	0,178	24%	1
ARMA(1,1)	0,104	55%	2
SARMA(1,1)(1,0)$_4$	0,098	58%	3
SARMA(1,1)(1,1)$_4$	0,088	62%	4
SARMA(1,1)(1,1)$_8$	0,082	65%	4

Oznaczenia:

N – liczba parametrów modelu
s_0^2 – wariancja szeregu obserwowanego
s_r^2 – wariancja szeregu resztowego
V_e – procent wariancji wyjaśniony przez model (por. wzór 70)

Uzyskane wyniki są pewnym zaskoczeniem. Istotną poprawę jakości dopasowania obserwuje się nie przy przejściu od modelu prostego do sezonowego, ale przy przejściu od modelu prostego do złożonego, czyli od AR(1) do ARMA(1,1). Mimo to uwzględnienie składowej sezonowej równej długości wersu poprawia jeszcze stopień dopasowania modelu. Przyjmując jako kryterium wyboru procent wyjaśnionej wariancji, za najlepszy uznano czteroparametryczny model sezonowy SARMA(1,1)(1,1)$_8$ o postaci:

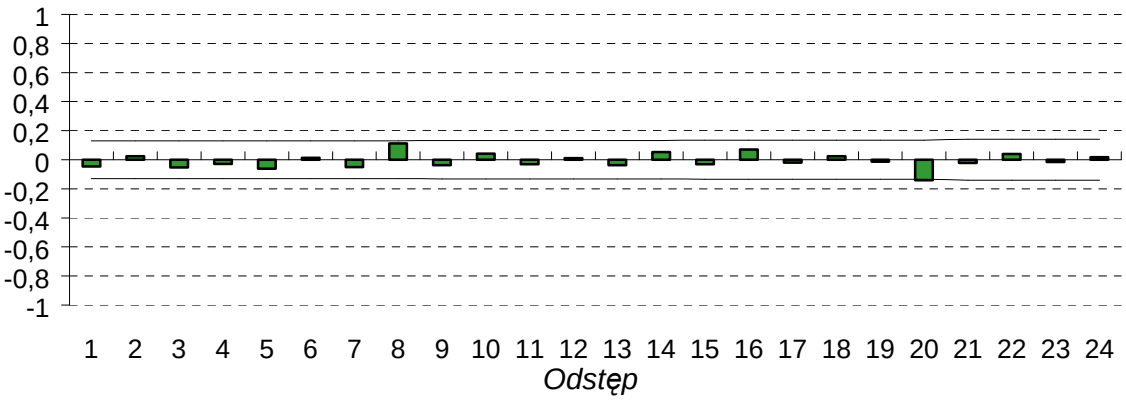
$$(71) \qquad (1+0{,}99B)(1-1{,}02B^8)x_t = (1+0{,}91B)(1-0{,}92B^8)e_t$$

Model SARMA(1,1)(1,1)$_8$ okazał się najlepszy także w przypadku pozostałych prób wiersza sylabotonicznego, gdzie procent wyjaśnionej przez model zmienności szeregu obserwowanego wynosił *ca* 61% i 69% (Tab. 17). Miarą dopasowania modelu do danych jest funkcja autokorelacji szeregu resztowego, czyli szeregu różnic pomiędzy wartościami obserwowanymi a generowanymi przez model. Jeżeli jakiś model dobrze opisuje dane obserwowane, szereg pozostający po odfiltrowaniu składowych deterministycznych ma bardzo niską autokorelację. Jak widać (Rys. 9), ACF szeregu resztowego dla omawianego tu modelu sezonowego (wzór 71) nie zawiera żadnych znaczących wartości.

Warto podkreślić, że bardzo dobre rezultaty dał zarówno model sezonowy z odstępem $s = 8$, odpowiadającym długości wersu, jak i model oparty na odstępie sezonowym

równym połowie tej długości ($s = 4$). Interpretując to spostrzeżenie w kategoriach lingwistycznych, należałoby powiedzieć, że mimo formalnej ośmiosylabowej struktury wersyfikacyjnej, wyraźna ekwiwalencja jednostek rytmicznych pojawia się już na poziomie hemistychu, czyli odcinka o długości czterech sylab.

Rys. 9 Autokorelacja szeregu resztowego dla modelu SARMA(1,1)(1,1)$_8$



W przypadku danych silnie deterministycznych interesujące rezultaty daje też model częstotliwościowy. Czyniąc wyjątek od przyjętej wcześniej zasady korzystania jedynie z metod modelowania w dziedzinie czasu, przedstawiamy poniżej wykres funkcji periodogramu stanowiącego jedno z podstawowych narzędzi badawczych analizy sekwencyjnej w dziedzinie częstotliwości[94]. Stosowanie periodogramu w analizie dyskretnych szeregów czasowych opiera się na twierdzeniach analizy matematycznej o rozkładach funkcji na szeregi trygonometryczne (tzw. rozkład Fouriera). Przy podejściu takim oblicza się ilości energii przypadające na częstotliwości harmoniczne szeregu. Spośród różnych postaci funkcji periodogramu przedstawiamy tę, która pozwala na estymację energii widmowej $I(f)$ bezpośrednio na podstawie obserwowanych wartości szeregu (GOTTMAN 1981:205):

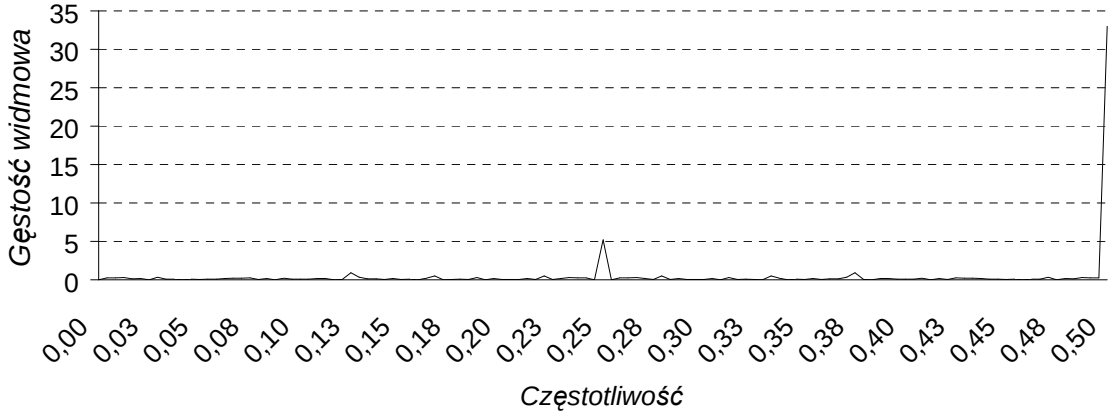
$$(72) \qquad I(f) = \frac{1}{2\pi N}\left[\sum_{t=1}^{N}(x_t \cos 2\pi f t)^2 + \sum_{t=1}^{N}(x_t \sin 2\pi f t)^2\right]$$

gdzie f – częstotliwość oscylacji szeregu
x_t – wartość szeregu czasowego w chwili lub na pozycji t
N – całkowita długość szeregu

[94] Opis analizy w dziedzinie częstotliwości był przedmiotem wielu obszernych publikacji (BLOOMFIELD 1976, PRIESTLEY 1981, GOTTMAN 1981:197–214) i został w tym miejscu pominięty. Argumentem, który w jakimś stopniu zaważył na tej decyzji, było domniemanie, iż w pracy o tematyce lingwistycznej cały rozdział poświęcony praktycznie wyłącznie analizie szeregów trygonometrycznych byłby niestosowny. Przede wszystkim jednak, uwzględniono rezultaty prowadzonych dotąd testów, z których wynika, iż w przypadku szeregów silnie stochastycznych (a takie właśnie szeregi spotyka się najczęściej przy kwantyfikacji tekstów) lepsze wyniki daje modelowanie w dziedzinie czasu.

Periodogram szeregu akcentowego wiersza sylabotonicznego (Rys. 10) wskazuje na możliwość wygenerowania takiej sekwencji jedynie dwiema częstotliwościami podstawowymi $f = 0{,}25$ i $f = 0{,}5$, przy czym energia przypadająca na częstotliwość $f = 0{,}5$ (okres $T = 2$, powtarzalność akcentu na co drugiej sylabie) jest wyraźnie większa od energii przypadającej na częstotliwość $f = 0{,}25$ (okres $T = 4$, powtarzalność akcentu na co czwartej sylabie). Dodajmy, że dokładnie te same częstotliwości dominowały w pozostałych próbkach. W kategoriach lingwistycznych układ taki odpowiada tyleż regularnej, co rzadkiej alternacji sylab akcentowanych i nie akcentowanych.

Rys. 10 Periodogram sekwencji przycisków akcentowych w wierszu sylabotonicznym



Nie da się ukryć, że na tym etapie badań ciężar gatunkowy użytego aparatu formalnego pozostaje w pewnej dysproporcji w stosunku do wagi problemu lingwistycznego. Cząstkowość poczynionych dotąd obserwacji skłania więc do postawienia pytania o celowość prowadzenia tak żmudnych obliczeń. Otóż prostota rytmiczna sylabotonizmu, szczególnie w przypadku poezji dziecięcej, jest zjawiskiem wyjątkowym. Sytuacja zaczyna się jednak komplikować przy innych systemach wersyfikacji oraz w prozie, gdzie nie istnieje formalny i odczuwalny system ekwiwalencji powtarzających się odcinków tekstu. Metoda sekwencyjna w analizie rytmiki i wersyfikacji ujawni więc pełnię swych możliwości dopiero podczas analizy szeregów reprezentujących bardziej złożone systemy rytmiczne.

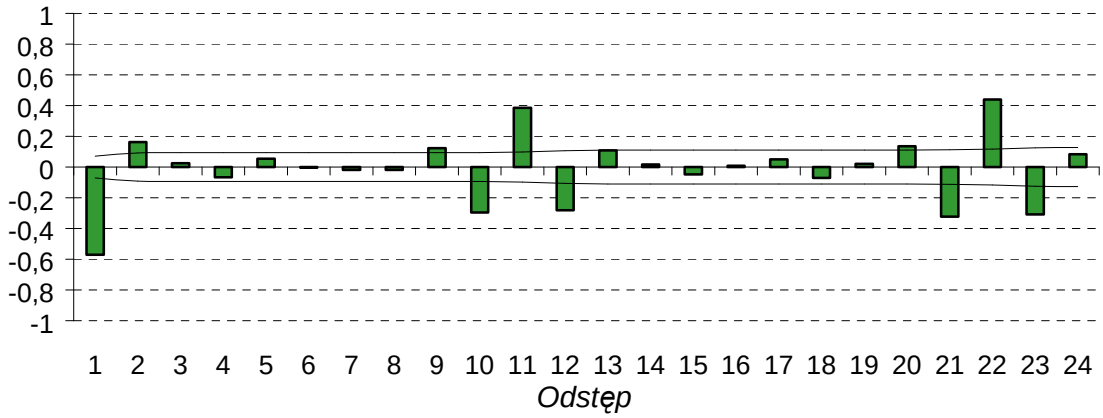
WIERSZ SYLABICZNY

Wiersz sylabiczny, reprezentowany tu przez fragment *Beniowskiego* J. Słowackiego (ANEKS – SŁOWACKI 1974:97–98), okazał się także bardzo rytmiczny, chociaż nie tak monotonny jak sylabotonik[95]. Funkcja autokorelacji zawiera znaczące wartości dla odstępów 1 i 2 oraz dla wielokrotności odstępu 11 (Rys. 11). Funkcja autokorelacji cząstkowej (Rys. 12) wygląda na gasnącą, choć także zawiera znaczące prążki pojawiające się w 11-sylabowych interwałach. Wynik ten dowodzi, iż sekwencja sylab akcentowa-

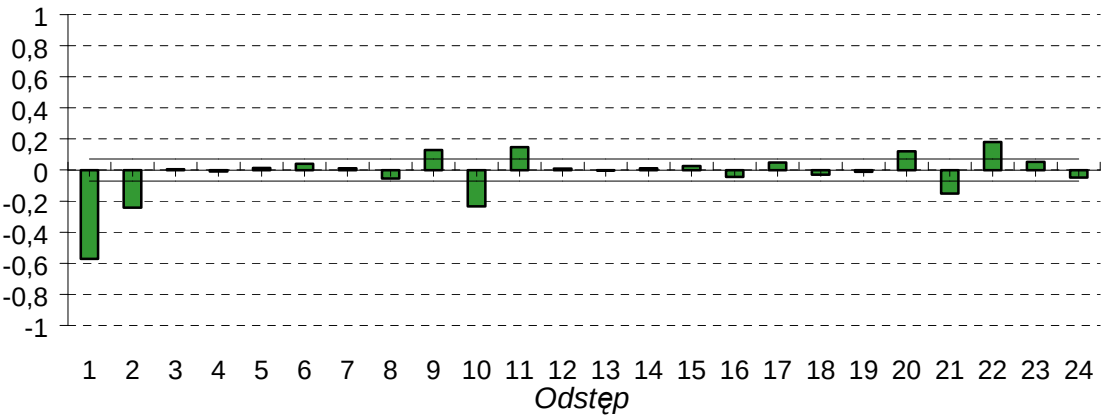
[95] Przedstawione obserwacje i wnioski odnoszą się do wszystkich badanych fragmentów *Beniowskiego*.

nych i nie akcentowanych w wierszu sylabicznym jest niezaprzeczalnie realizacją jakiegoś procesu stochastycznego, najprawdopodobniej sezonowego.

Rys. 11 Autokorelacja sekwencji przycisków akcentowych w wierszu sylabicznym



Rys. 12 Autokorelacja cząstkowa sekwencji akcentów w wierszu sylabicznym



Kształt funkcji ACF i PACF sugeruje model złożony ze składowej sezonowej (prążki na odstępach 11 i 22) i prostej typu MA(1) (pierwsze dwie autokorelacje cząstkowe wyglądają na gasnące, autokorelacja zwykła urywa się po odstępie pierwszym). Modele proste i sezonowe, estymowane osobno, dawały raczej przeciętne wyniki – każdy z nich pozostawiał szereg reszt z dużą ilością nie odfiltrowanej wariancji (Tab. 16). Dopiero połączenie ich w jednym modelu typu SARMA dało rezultat w pełni satysfakcjonujący. Estymowany model SARMA(0,1)(1,1)$_{11}$ wyjaśnia aż 48% wariancji szeregu obserwowanego i ma następującą postać:

(73) $$(1-0{,}99B^{11})x_t = (1-0{,}48B)(1-0{,}91B^{11})e_t$$

Wykres ACF szeregu reszt otrzymanego z szeregu obserwowanego po odfiltrowaniu powyższych składowych deterministycznych nie zawiera żadnych znaczących wartości i potwierdza dobrą jakość powyższego modelu (Rys. 13).

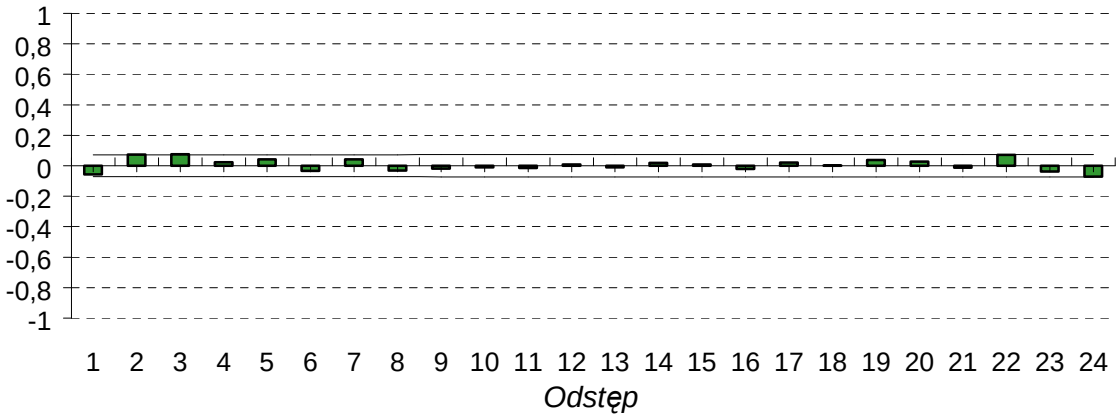
Tab. 16 Identyfikacja modelu rytmiki wiersza sylabicznego

Typ modelu ($s_0^2 = 0{,}239$)	s_r^2	V_e	N
MA(1)	0,162	32%	1
SAR(1)	0,204	15%	1
SMA(1)	0,218	9%	1
SARMA(0,0)(1,1)$_{11}$	0,157	34%	2
SARMA(0,1)(1,1)$_{11}$	0,123	48%	3

Oznaczenia:

N – liczba parametrów modelu
s_0^2 – wariancja szeregu obserwowanego
s_r^2 – wariancja szeregu resztowego
V_e – procent wariancji wyjaśniony przez model (por. wzór 70)

Rys. 13 Autokorelacja szeregu resztowego dla modelu SARMA(0,1)(1,1)$_{11}$

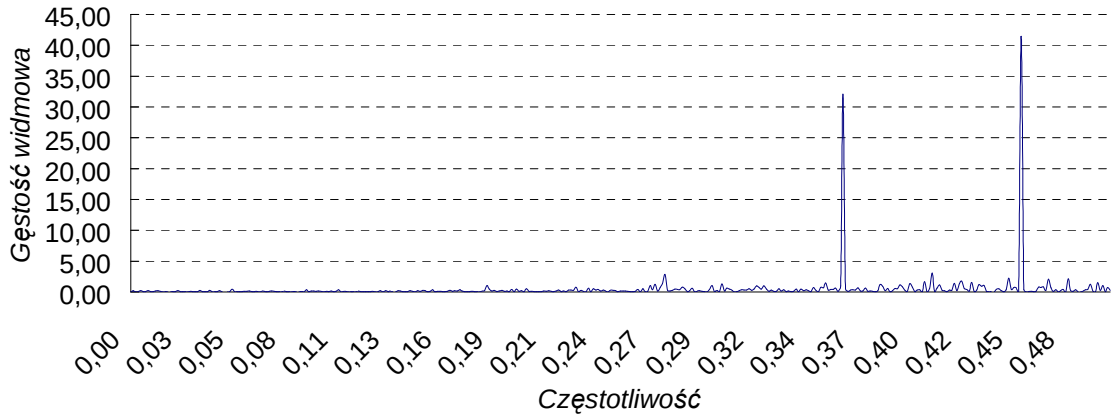


Podobnie jak w przypadku sylabotonika, dla porównania przedstawiono także wykres periodogramu (Rys. 14). Okazuje się, że rytm poematu Słowackiego pisanego jedenastozgłoskowcem daje się wygenerować za pomocą dwóch częstotliwości składowych. Ich wartości ($f_1 = 0{,}364$ i $f_2 = 0{,}454$) odpowiadają długościom okresu $T_1 = 2{,}75$ i $T_2 = 2{,}20$, czyli, po zaokrągleniu, dwu- i trzysylabowym interwałom oddzielającym sylaby najczęściej akcentowane. Jednak widoczna tu postać periodogramu nie powstała na skutek takiego czy innego rozkładu statystycznego tych interwałów, ale z uwagi na ich kolejność w tekście.

Porównanie otrzymanych dotychczas wyników dla wiersza sylabicznego i sylabotonicznego pokazuje, że w przypadku danych tekstowych zawierających stosunkowo słabe składowe deterministyczne (w prozie będą one jeszcze słabsze niż w wierszu) modelowanie w dziedzinie częstotliwości (periodogram) jest mniej efektywne niż modelowanie w dziedzinie czasu. Rząd modelu sezonowego, widoczny już na wykresach autokorelacji szeregu akcentowego, posiada przejrzystą lingwistyczną interpretację – odpowiada rze-

czywistej długości minimalnego odcinka tekstu, ekwiwalentnego pod względem rytmicznym i/lub metrycznym. W przypadku sylabotonika, stwierdzono, że odcinkiem takim może być sekwencja nie tylko ośmiu, ale i czterech sylab. Z kolei rząd modelu prostego daje się w odniesieniu do tekstu interpretować jako głębokość związku kontekstowego. Informacji takich nie dostarcza periodogram, wyodrębniając jedynie częstotliwości składowe szeregu, trudniejsze do zinterpretowania w kategoriach lingwistycznych[96].

Rys. 14 Periodogram sekwencji akcentów w jedenastozgłoskowcu sylabicznym[97]



Aby lepiej przyjrzeć się strukturze rytmicznej wiersza sylabotonicznego, analizie sekwencyjnej poddano szereg reszt otrzymany po odfiltrowaniu z szeregu obserwowanego składowych sezonowych, powstałych jedynie na skutek stosowania w tekście wersyfikacji. Matematycznie operacja jest wykonalna, ponieważ w metodzie ARIMA poszczególne składowe procesu są przez algorytm estymacji modelu rozdzielane i przedstawiane w postaci addytywnej lub multiplikatywnej. W kategoriach lingwistycznych szereg taki należałoby interpretować jako zapis struktury rytmicznej jedenastozgłoskowca sylabicznego z pominięciem „efektu wersyfikacji".

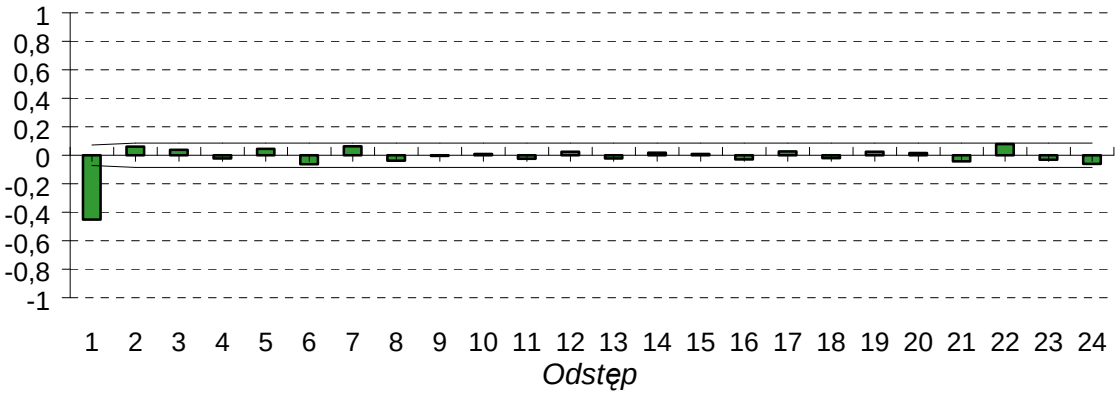
W omawianym przypadku przedmiotem zainteresowania jest źródło różnic pomiędzy modelem czysto sezonowym SARMA(0,0)$(1,1)_{11}$ i mieszanym SARMA(0,1)$(1,1)_{11}$. Jak się okazuje, funkcje ACF i PACF otrzymane po odfiltrowaniu „efektu wersyfikacji" nie są losowe (Rys. 15 i 16). Zgodnie z oczekiwaniami zawierają składową deterministyczną, którą można opisać modelem typu MA(1) o postaci:

(74) $$x_t = (1 - 0{,}47B)e_t$$

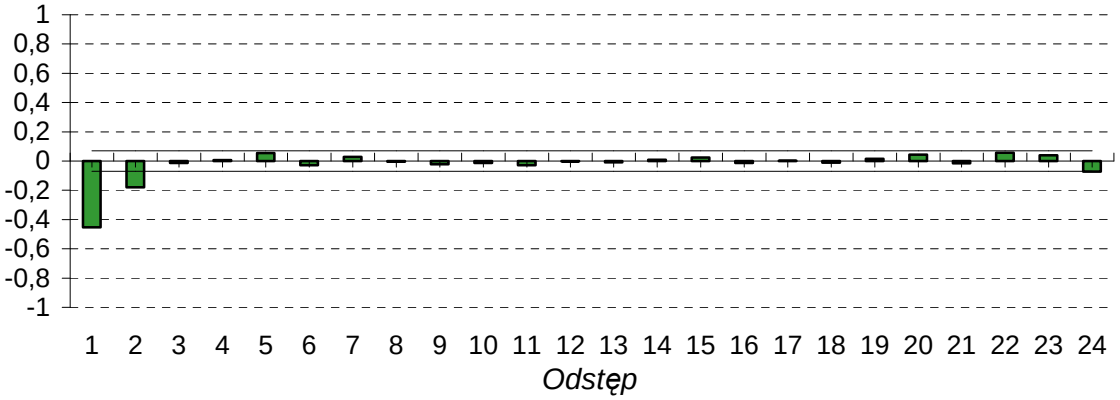
Mimo niewielkiej liczby parametrów, model (74) wyjaśnia 21% wariancji szeregu powstałego poprzez odfiltrowanie „efektu wersyfikacji" z jedenastozgłoskowca sylabicznego.

[96] Brak na przykład znaczącej wartości periodogramu wskazującej na długość wersu T=11.
[97] Dokładnie te same częstotliwości dominowały w pozostałych fragmentach.

Rys. 15 Autokorelacja sekwencji przycisków akcentowych w wierszu sylabicznym po usunięciu „efektu wersyfikacji”



Rys. 16 Autokorelacja cząstkowa sekwencji przycisków akcentowych w wierszu sylabicznym po usunięciu „efektu wersyfikacji”



Nasuwa się w tym miejscu pytanie o to, czym właściwie jest tekst wiersza bez rytmicznego „efektu wersyfikacji” – pytanie o tyle istotne, że otrzymany szereg nie jest losowy. Otóż pod względem struktury rytmicznej, tekst taki powinien najbardziej przypominać prozę. Wyniki kolejnych testów prezentowane w dalszych rozdziałach pozwolą zweryfikować tę hipotezę.

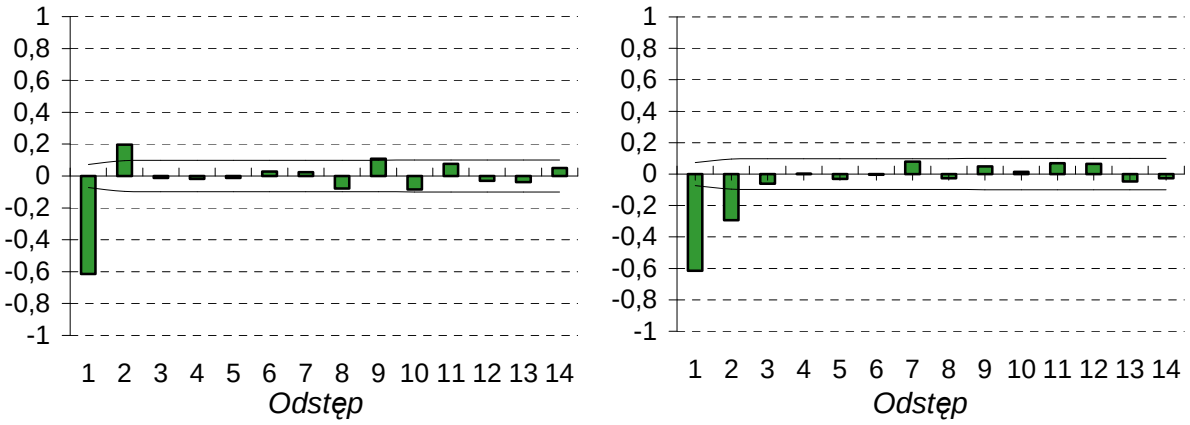
DYSKURS ORATORSKI

Wszystkie dotychczas badane teksty posiadały formalną strukturę wersyfikacyjną, generującą w warstwie rytmicznej tekstu silne procesy stochastyczne typu sezonowego. Można więc sądzić, że fragmenty pozbawione stałej wersyfikacji będą w najlepszym wypadku realizacjami procesów prostych.

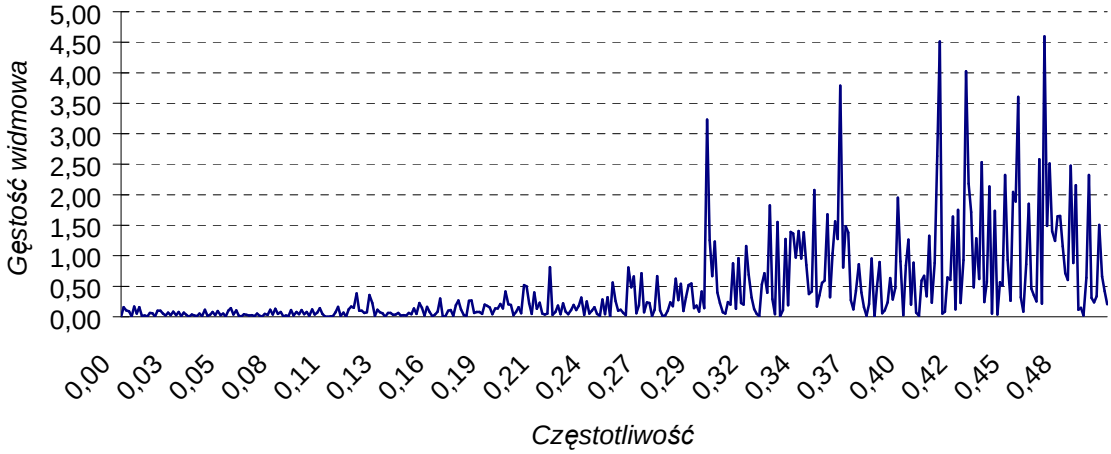
Kształty funkcji ACF i PACF dla próby pochodzącej z wygłoszonej we Wrocławiu homilii Jana Pawła II (Rys. 17) wskazują na obecność w strukturze rytmicznej szeregu wyraźnej składowej deterministycznej typu prostego[98].

[98] Dane nagrania zawiera ANEKS. Analizowany tutaj fragment miał długość 754 sylab. Prezentowane obserwacje i wnioski odnoszą się do wszystkich analizowanych fragmentów homilii.

Rys. 17 Autokorelacja (wykres lewy) i autokorelacja cząstkowa (wykres prawy) sekwencji przycisków akcentowych w dyskursie oratorskim



Rys. 18 Periodogram sekwencji akcentów w dyskursie oratorskim



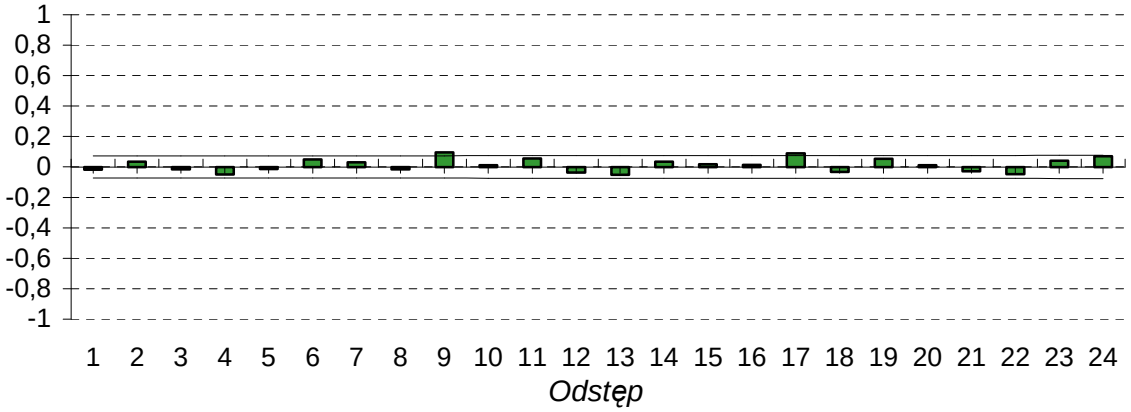
Z kształtu funkcji ACF urywającej się po drugim odstępie i łagodnie gasnącej autokorelacji cząstkowej (Rys. 17) należy wnioskować, że obserwowany szereg jest realizacją procesu ruchomej średniej typu MA(2). W połączeniu z brakiem wersyfikacji w tekście sugeruje to estymację modelu w dziedzinie czasu. Wniosek taki potwierdza forma periodogramu (Rys. 18). Wartość energii przypadającej na poszczególne częstotliwości, począwszy od $f = 0{,}3$ rozkłada się równomiernie.

Jak wynika z lektury przedstawionych wyżej wykresów, najlepszym modelem badanego procesu stochastycznego jest MA(2), wyjaśniający 43% wariancji szeregu początkowego. Model MA(1) wyjaśnia jej co prawda aż 38%, jednakże pozostający szereg reszt nie jest zupełnie losowy. Natomiast modele wyższych rzędów nie poprawiają jakości estymacji, zwiększając jedynie liczbę parametrów. Estymowany model MA(2) ma postać:

(75) $$x_t = (1 - 0{,}79B + 0{,}26B^2)e_t$$

Autokorelacja szeregu resztowego pokazuje, że wszystkie znaczące składowe obserwowanego procesu zostały przez model MA(2) usunięte (Rys. 19).

Rys. 19 Autokorelacja szeregu resztowego dla modelu MA(2)



Otrzymany wynik można uznać za zgodny z oczekiwaniami. Potwierdzono hipotezę mówiącą, że sekwencja rytmiczna dyskursu oratorskiego nie jest losowa, ale zawiera silną składową deterministyczną. Procent wariancji szeregu obserwowanego, wyjaśniony przez model MA(2), jest w tym przypadku niższy niż w wierszu ($V_e = 43\%$), jednak jego wartość i tak uznać należy za wysoką (por. Tab. 17). Rząd estymowanego procesu wskazuje na głębokość związku kontekstowego. Okazuje się, że akcent na dowolnej sylabie (lub jego brak) determinowany jest przez jedną lub co najwyżej dwie sylaby poprzedzające.

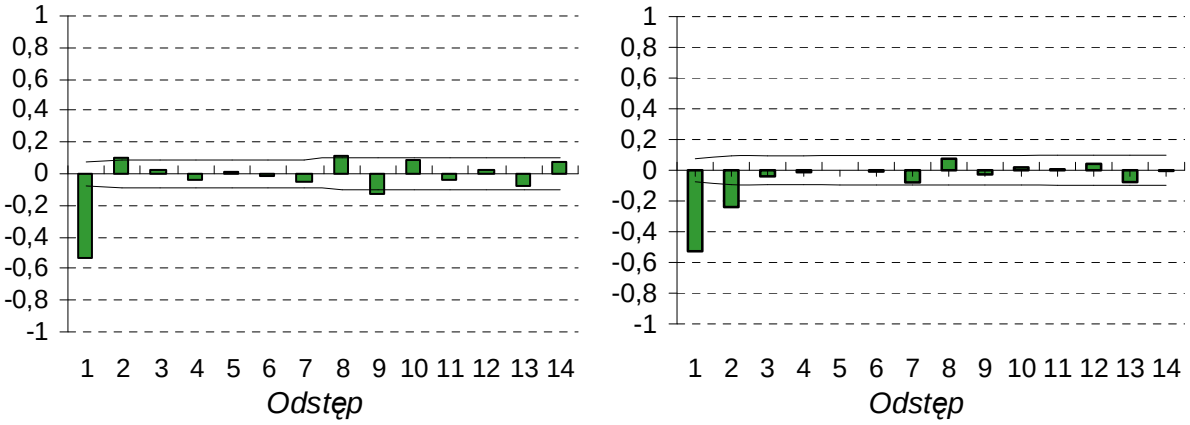
Warto też porównać otrzymane tu wykresy funkcji ACF i PACF oraz typ estymowanego modelu z analogicznymi danymi dla wiersza sylabicznego, z którego usunięto „efekt wersyfikacji". Ich podobieństwo pod względem kształtu i ten sam typ modelu świadczą o tym, że być może zawierają podobną składową deterministyczną. Płynie z tego wniosek, iż różne odmiany prozy w polszczyźnie mogą zawierać ten sam model rytmiczny. Natomiast różnica wartości współczynników oraz parametru V_e dla modelu ruchomej średniej estymowanego w obu przypadkach pokazuje, że stopień zrytmizowania tekstów jest zróżnicowany (przypomnijmy, że dla wiersza sylabicznego bez „efektu wersyfikacji" otrzymano $V_e = 21\%$). Przypuszczenia te zweryfikuje przedstawiona poniżej analiza fragmentów prozy artystycznej.

PROZA ARTYSTYCZNA

Podobnie jak w poprzednim przypadku, analizę obserwowanego szeregu reprezentującego prozę artystyczną[99] rozpoczynamy od przedstawienia funkcji autokorelacji i autokorelacji cząstkowej (Rys. 20). Jak widać, ACF zawiera tylko jeden znaczący prążek, natomiast PACF dość szybko wygasa. Układ taki charakterystyczny jest dla procesu typu MA(1).

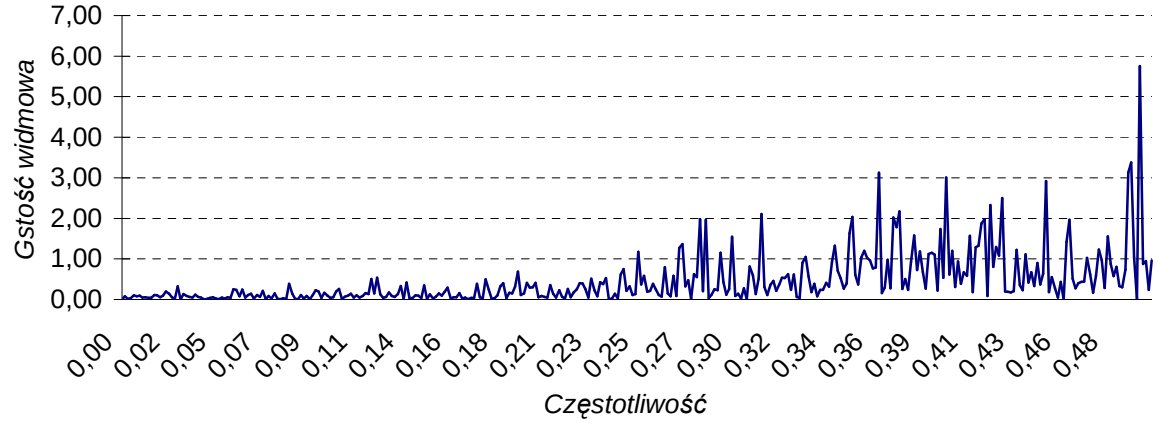
[99] Analizowano fragment powieści I. Newerlego *Wzgórze Błękitnego Snu* (ANEKS – NEWERLY 1986:18–19). Długość próby wynosiła 703 sylaby.

Rys. 20 Autokorelacja i autokorelacja cząstkowa sekwencji przycisków akcentowych w prozie artystycznej



Silnie stochastyczny charakter szeregu potwierdza wykres periodogramu (Rys. 21). Tak jak w przypadku dyskursu oratorskiego, energia rozkłada się na dużą liczbę częstotliwości i w efekcie żadna z nich nie może być uznana za dominującą.

Rys. 21 Periodogram sekwencji akcentowej w prozie artystycznej

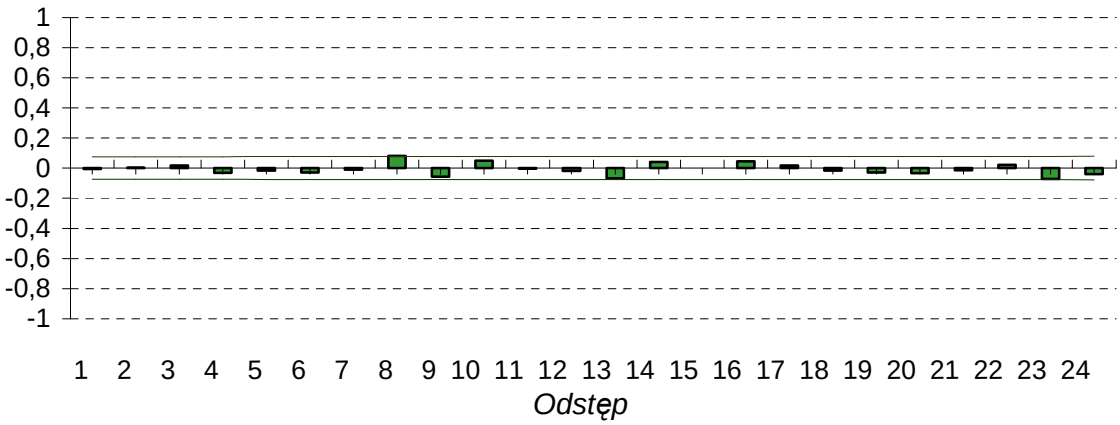


Wybór właściwego modelu nie nastręcza większych trudności. Jak już wspomniano, funkcje ACF i PACF wyraźnie wskazują na MA(1), który wyjaśnia 30% wariancji szeregu wyjściowego. Zważywszy jednak, że autokorelacja dla odstępu 2 jest na granicy przedziału ufności (Rys. 20), estymowano także model MA(2). W zamian za jeden dodatkowy parametr, wyjaśnia on nieco więcej, bo 31,7% wariancji początkowej. Ostatecznie przyjęto więc model MA(2) mający postać:

(76) $$x_t = (1 - 0{,}66B + 0{,}16B^2)e_t$$

Wykres ACF szeregu resztowego (Rys. 22) nie zawiera żadnych znaczących wartości, co potwierdza, że model MA(2) dobrze opisuje szereg obserwowany.

Rys. 22 Autokorelacja szeregu resztowego dla modelu MA(2)



Uzyskany wynik wskazuje na istnienie wyraźnego rytmu prozy. Głębokość związku kontekstowego jest wprawdzie niewielka (praktycznie jedna sylaba determinuje typ akcentu na sylabie następnej), jednak jego obecność w tekście jest faktem bezdyskusyjnym. Potwierdzono też wcześniejsze przypuszczenia o istnieniu jednakowego wzorca rytmicznego prozy artystycznej, dyskursu oratorskiego i wiersza, z którego usunięto „efekt wersyfikacji". W rzeczonych przypadkach odkryto w tekście proces stochastyczny opisany modelem MA(2) lub MA(1) (drugi współczynnik modelu miał stosunkowo niewielką wartość i może zostać pominięty). Odnotowano także wyraźną tendencja spadkową wartości współczynnika V_e, wyrażającego stopień rytmicznego uporządkowania tekstu: spośród porównywanych tu czterech odmian wersyfikacyjno-stylistycznych polszczyzny najmniej rytmiczną okazała się właśnie proza artystyczna.

1.3 PODSUMOWANIE

Przeprowadzone testy wykazały, iż matematyczne modelowanie sekwencyjnej struktury tekstu jest możliwe i poznawczo efektywne. Stosując liniowe modele procesów stochastycznych, wykazano, że sylaby akcentowane i nie akcentowane rozłożone są w tekście w sposób mniej lub bardziej uporządkowany, a stopień tego uporządkowania jest cechą wymierną i charakteryzuje różne stylistyczno-wersyfikacyjne odmiany polszczyzny. Poddane analizie próby opisane zostały specyficznymi dla każdego stylu modelami. Różnice między próbami należącymi do tej samej odmiany stylistycznej były natomiast niewielkie. Wykazano, że dla wiersza najefektywniejsze były modele sezonowe, natomiast dla prozy najlepszy okazał się model ruchomej średniej MA(2). Potwierdzono skuteczność wskaźnika V_e (por. wzór 70), wyrażającego procent wariancji (zmienności) szeregu obserwowanego wyjaśnionej przez model. Empirycznie wykazano, że wskaźnik ten stanowi syntetyczną miarę stopnia zrytmizowania tekstu i umożliwia porównywanie dowolnych stylów i odmian językowych (Tab. 17).

W omawianym przypadku wskaźnik V_e wyraźnie odróżnił wiersz sylabotoniczny, sylabiczny oraz prozę. Ponadto stwierdzono zauważalną różnicę pomiędzy dyskursem

oratorskim (reprezentowanym przez homilię papieską) a prozą artystyczną. O ile jednak różnice pewnych stylów czy odmian (tu dobranych na zasadzie kontrastu) wynikają analitycznie z ich cech strukturalnych i funkcjonalnych, o tyle szczegółowa klasyfikacja stylów prozatorskich nie jest możliwa bez adekwatnych narzędzi matematycznych. Można więc spierać się o przydatność modeli opisujących najprostsze systemy wersyfikacji, jednak trudno nie zgodzić się z twierdzeniem, że dokładny pomiar stopnia uporządkowania rytmicznego dowolnego tekstu, w szczególności prozy i stylów „pośrednich" (dyskursu oratorskiego, swobodnych układów metrycznych), jest tradycyjnymi metodami nieosiągalny.

Tab. 17 Uporządkowanie rytmiczne stylistyczno-wersyfikacyjnych odmian polszczyzny

	Typ modelu	Próba			Średnia
		1	2	3	
Wiersz sylabotoniczny	SARMA(1,1)(1,1)$_8$	61%	65%	69%	65%
Wiersz sylabiczny	SARMA(0,1)(1,1)$_{11}$	45%	48%	51%	48%
Dyskurs oratorski	MA(2)	39%	43%	39%	40%
Proza artystyczna	MA(2)	36%	32%	33%	34%

Za szczególnie ważne dla wersologii uznać należy modele sezonowe, pozwalające wykryć rzeczywistą długość ekwiwalentnych pod względem rytmicznym odcinków tekstu. Na ogół, długość ta pokrywa się z długością wersu, ale zdarza się, że jest inna. Pokazał to przykład regularnego sylabotonika J. Brzechwy, pokażą też testy przeprowadzone na tekście *Eugeniusza Oniegina* A. Puszkina, które przedstawiamy w następnym rozdziale.

Uogólniając otrzymane rezultaty w kontekście założeń i celów lingwistyki modelowej (por. Część I, 1.1 i 1.2), odkrycie różnych procesów stochastycznych w sekwencyjnych strukturach tekstu należy uznać za wiarygodną podstawę do sformułowania praw językowych opisujących tego rodzaju struktury. I tak, jako hipotezę roboczą dla dalszych poszukiwań można przyjąć, że model ruchomej średniej MA(2) jest najlepszym liniowym modelem opisującym sekwencję sylab akcentowanych i nie akcentowanych w polszczyźnie. To ostatnie spostrzeżenie potwierdza rozkład jedenastozgłoskowca sylabicznego na składową sezonową, wyrażającą „efekt wersyfikacji", i pozostający szereg resztowy, który swą strukturą liniową przypominał właśnie tekst prozy (Rys. 15 i 16).

W podsumowaniu tego rozdziału zacytujmy raz jeszcze słowa Arystotelesa z III księgi *Retoryki*, mówiące, iż „Tekst prozy nie powinien mieć metrycznej formy wiersza, ani też nie powinien być pozbawiony rytmu. [...] Dlatego proza musi posiadać rytm, nie może natomiast posiadać miar wierszowych, bo zamieni się w poezję." (*Retoryka* 1408b)[100] Wyniki naszych dociekań opartych na aparacie matematycznym metody ARIMA i wszechstronnych testach empirycznych potwierdzają słuszność tego postulatu.

[100] Cytat na podstawie przekładu H. Podbielskiego (PODBIELSKI 1988).

1.4 TEST EFEKTYWNOŚCI MODELOWANIA SEKWENCYJNEGO[101]

Stosowane przez nas dotąd modele matematyczne opisują jedynie sekwencyjną strukturę języka. Aby lepiej ocenić ich efektywność, uzyskane dotychczas wyniki porównano z rezultatami, jakie przy tych samych założeniach ogólnych dałaby metoda statystyki konwencjonalnej, oparta nie na kolejności elementów w próbie, ale ich częstościach. Uwzględniając wielkie zróżnicowanie dostępnych danych językowych oraz dwa dominujące (choć nie jedyne) podejścia badawcze, można rozróżnić trzy możliwe sytuacje:

1. **Stosowalne jest jedynie podejście nie uwzględniające porządku sekwencyjnego**
   Przykłady danych językowych nie podlegających modelowaniu sekwencyjnemu podać można wychodząc od elementarnej dla QL opozycji systemu i tekstu[102]. O ile każdy tekst daje się przedstawiać zarówno w formie zbioru, jak i sekwencji jednostek, o tyle dane systemowe (na przykład słownictwo konkretnego autora, tekstu czy epoki) mogą być analizowane jedynie jako zbiory elementów (w sensie matematycznym).
2. **Stosowalne i efektywne są oba podejścia**
   Prowadzone dotąd badania pokazały, że wartościowe wyniki uzyskuje się między innymi badając teksty jako zbiory i/lub sekwencje leksemów. W pierwszym przypadku otrzymuje się rozkłady statystyczne leksemów, w drugim modele sekwencyjne (por. Część II, 5.2).
3. **Stosowalne są oba podejścia, ale ich efektywność jest różna**
   Tekst jest strukturą linearną i należy się spodziewać, że mimo niekwestionowanej skuteczności metody konwencjonalnej, pewne jego warstwy będą lepiej opisywane metodą sekwencyjną. Prawdopodobnie warstwą taką jest rytmika (metryka) tekstu.

Przypadek trzeci potraktowano jako hipotezę i poddano weryfikacji. Na danych, które były już przedmiotem analizy sekwencyjnej (wyniki przedstawione w poprzednim rozdziale), przeprowadzono konwencjonalne testy statystyczne i porównano otrzymane wyniki. Testy te opierały się na założeniu, iż badane fragmenty nie są sekwencjami, ale zbiorami sylab akcentowanych i nie akcentowanych. Porządek występowania sylab jest w takim przypadku pomijany, a jedyną relewantną informacją jest częstość występowania sylab różnych typów. Przypomnijmy, że testy sekwencyjne pozwoliły wyraźnie odróżnić nie tylko odmiany ekstremalnie różne pod względem wersyfikacji i rytmiki, ale także style na pozór bardzo do siebie podobne (dyskurs oratorski i prozę artystyczną). Choć trudno z góry cokolwiek przesądzać, z uwagi na to, że tekst jest strukturą linearną, rezygnacja z uwzględniania porządku sylab powinna raczej pogorszyć lub wręcz uniemożliwić rozróżnienie poszczególnych stylów i odmian wersyfikacyjnych. W tabeli 18 przedstawiono liczby sylab akcentowanych i nie akcentowanych w badanych próbach[103]:

---

[101] Wykorzystano tu wyniki empiryczne opublikowane w pracy PAWŁOWSKI 1999.
[102] Rozróżnienie to szczegółowo omawiane jest w pracy HAMMERL&SAMBOR 1990:15–16.
[103] Tym razem pominięto występujący niezwykle rzadko akcent poboczny.

Tab. 18 Liczby sylab akcentowanych i nie akcentowanych w badanych próbach

Próba	n	n_0	n_1	s^2	s^2/n	m_1
Wiersz sylabotoniczny 1	312	188	124	0,240	0,0008	0,3974
Wiersz sylabotoniczny 2	240	151	89	0,234	0,0010	0,3708
Wiersz sylabotoniczny 3	264	141	123	0,250	0,0009	0,4659
Wiersz sylabotoniczny	816	480	336	0,242	0,0003	0,4118
Wiersz sylabiczny 1	792	506	286	0,231	0,0003	0,3611
Wiersz sylabiczny 2	792	478	314	0,240	0,0003	0,3965
Wiersz sylabiczny 3	792	471	321	0,241	0,0003	0,4053
Wiersz sylabiczny	2376	1455	921	0,237	0,0001	0,3876
Dyskurs oratorski 1	731	465	266	0,221	0,0003	0,3639
Dyskurs oratorski 2	753	464	289	0,225	0,0003	0,3838
Dyskurs oratorski 3	668	405	263	0,224	0,0003	0,3937
Dyskurs oratorski	2152	1334	818	0,223	0,0001	0,3801
Proza artystyczna 1	727	461	266	0,227	0,0003	0,3659
Proza artystyczna 2	703	449	254	0,230	0,0003	0,3613
Proza artystyczna 3	531	339	192	0,228	0,0004	0,3616
Proza artystyczna	1961	1249	712	0,228	0,0001	0,3631

Oznaczenia:

n – wielkość próby
n_0 – liczba sylab nie akcentowanych
n_1 – liczba sylab akcentowanych
s^2 – wariancja z próby
m_1 – średnia liczba sylab akcentowanych

Celem analizy było wskazanie statystycznie istotnych różnic pomiędzy badanymi próbami. Gdyby efektywność testu była optymalna, fragmenty reprezentujące tę samą odmianę stylistyczną byłyby statystycznie podobne, natomiast próby reprezentujące odmiany różne różniłyby się także pod względem statystycznym. Pewnym problemem jest wybór zastosowanego testu. Binarny zapis rytmicznej struktury tekstu może być bowiem traktowany zarówno jako zbiór liczb (dostępnym parametrem jest wówczas średnia z próby), jak i zbiór symboli (dostępnym parametrem jest wtedy częstość elementów 0 i 1). Aby uwzględnić obie możliwości, posłużono się testem istotności na różnicę średnich oraz testem na wskaźnik struktury (tzw. test frakcji)[104].

[104] Dokładny opis testu średnich znaleźć można m.in. w pracach HAMMERL&SAMBOR 1990:253–254, SOBCZYK 1996:155–158, GREŃ 1987:419–424. Test frakcji omawiany jest m.in. w pracach SOBCZYK 1996:166–167, GREŃ 1987:170–173, HAMMERL&SAMBOR 1990:269–271.

W celu porównania średnich z prób buduje się statystykę *U* o postaci:

$$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{77}$$

gdzie:

n_1, n_2 – wielkości porównywanych prób
$\bar{x}_1$, $\bar{x}_2$ – średnie z próby
s_1^2, s_2^2 – wariancje z próby

Hipoteza zerowa H_0 ma postać równości $\bar{x}_1 = \bar{x}_2$. Ponieważ statystyka *U* ma rozkład asymptotycznie normalny *N*(0,1), hipotezę H_0 przyjmujemy na poziomie istotności $\alpha = 0{,}05$, jeżeli jej wartość zawarta jest w przedziale krytycznym [–1,96, 1,96]. Pozytywny wynik testu oznacza, że pomiędzy próbami nie ma znaczącej różnicy ze względu na średnią.

Wskaźnik struktury dla dwóch populacji oblicza się na podstawie statystyki *U* o postaci:

$$U = \frac{w_1 - w_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}} \tag{78}$$

gdzie:

$$w_1 = \frac{m_1}{n_1}; \qquad w_2 = \frac{m_2}{n_2}; \qquad n = \frac{n_1 n_2}{n_1 + n_2}; \qquad \bar{p} = \frac{m_1 + m_2}{n_1 + n_2};$$

n_1 – wielkość pierwszej próby;
n_2 – wielkość drugiej próby;
m_1 – liczba elementów pierwszej próby posiadających daną cechę;
m_2 – liczba elementów drugiej próby posiadających daną cechę;

Formułujemy hipotezę zerową H_0 o postaci $w_1 = w_2$. Tak jak poprzednio, statystyka *U* ma rozkład asymptotycznie normalny *N*(0,1) i hipotezę H_0 przyjmuje się na poziomie istotności $\alpha = 0{,}05$ jeżeli wartości *U* należą do przedziału krytycznego [–1,96, 1,96]. Pozytywny wynik testu (przyjęcie hipotezy H_0) zakłada, iż obie próby pochodzą z tej samej populacji ogólnej i nie różnią się w sposób statystycznie istotny.

W tabeli 18 przedstawiono wyniki porównania poszczególnych prób za pomocą obu testów. Liczby powyżej przekątnej reprezentują wartość statystyki *U* obliczanej jako wskaźnik struktury (wzór 78), liczby poniżej przekątnej reprezentują natomiast wartości statystyki *U* obliczonej dla porównania średnich (wzór 77). Szarym tłem wyróżniono komórki zawierające wartości spoza przedziału krytycznego, przy których hipoteza H_0 była odrzucana. Tylko tam mówić można o statystycznie istotnych różnicach między próbami i stylami.

Tab. 19 Porównanie struktury rytmicznej wybranych odmian polszczyzny[105]

	B1	B2	B3	S1	S2	S3	W1	W2	W3	N1	N2	N3
B1		0,636	-1,654	1,125	0,030	-0,240	1,025	0,416	0,111	0,963	1,099	1,038
B2	0,637		**-2,159**	0,274	-0,713	-0,956	0,194	-0,360	-0,624	0,138	0,265	0,247
B3	-1,653	**-2,169**		**3,027**	**1,984**	1,728	**2,912**	**2,338**	**2,016**	**2,851**	**2,971**	**2,832**
S1	1,115	0,273	**2,978**		-1,450	-1,809	-0,113	-0,922	-1,281	-0,193	-0,008	-0,017
S2	0,030	-0,717	**1,964**	-1,450		-0,359	1,308	0,510	0,107	1,225	1,398	1,280
S3	-0,240	-0,964	1,713	-1,810	-0,359		1,659	0,864	0,450	1,576	1,745	1,600
W1	1,025	0,194	**2,886**	-0,114	1,324	1,682		-0,793	-1,149	-0,079	0,101	0,084
W2	0,417	-0,363	**2,326**	-0,934	0,516	0,876	-0,812		-0,383	0,712	0,887	0,810
W3	0,112	-0,632	**2,016**	-1,302	0,109	0,458	-1,181	-0,394		1,070	1,237	1,139
N1	0,959	0,138	**2,819**	-0,194	1,233	1,587	-0,081	0,725	1,093		0,180	0,157
N2	1,091	0,264	**2,930**	-0,008	1,400	1,751	0,103	0,899	1,259	0,181		-0,010
N3	1,036	0,247	**2,812**	-0,018	1,289	1,614	0,085	0,823	1,162	0,158	-0,010	

Wynik ten potwierdza się, jeżeli zamiast pojedynczych fragmentów porówna się zsumowane próby każdego z czterech badanych stylów (oznaczenia jak wyżej):

Tab. 20 Porównanie struktury rytmicznej wybranych odmian polszczyzny

	B	S	W	N
B		1,218	1,579	**2,411**
S	1,213		0,519	1,661
W	1,582	0,527		1,129
N	**2,396**	1,670	1,149	

Porównanie wyników uzyskanych metodą konwencjonalną (Tab. 19, 20) i sekwencyjną (Tab. 17) wskazuje na większą efektywność modeli sekwencyjnych. Jeżeli trzymać się ściśle zasad interpretacji testów statystycznych, jedynie najbardziej skrajne odmiany stylistyczne i wersyfikacyjne (sylabotonik i proza artystyczna) mogą być uznane za statystycznie różne pod względem częstości pojawiania się sylab akcentowanych i nie akcentowanych. Co więcej, przy analizie pojedynczych fragmentów jedyną wyróżniającą się próbą jest B3, reprezentująca wiersz sylabotoniczny. Próba ta jest tak nietypowa, że różni się nawet od próby B2, pochodzącej z tego samego utworu. Nie należy oczywiście całkowicie rezygnować z obu zastosowanych tu testów. Tabela sumaryczna (Tab. 20) wskazuje, że wartość statystyki *U* jest większa przy stylach różniących się wyraźnie (N–B, S–B), a mniejsza przy stylach podobnych. Jednak przy analizie wszystkich fragmentów obraz ten nie jest ani tak przejrzysty, ani przekonujący. Biorąc pod uwagę całkowite pominięcie wersyfikacji, którą bezbłędnie rozszyfrowuje funkcja autokorelacji i modele sezonowe, a także możliwość izolowania i osobnej analizy składowych szeregu, powyższy rezultat udowadnia, że przy badaniu rytmiki tekstu zdecydowanie większą moc eksplananacyjną posiadają modele sekwencyjne budowane za pomocą metody ARIMA.

[105] Oznaczenia badanych prób: B1, B2, B3 – wiersz sylabotoniczny; S1, S2, S3 – wiersz sylabiczny; W1, W2, W3 – dyskurs oratorski; N1, N2, N3 – proza artystyczna.

2. ANALIZA PORÓWNAWCZA PROZODII JĘZYKÓW O AKCENCIE STAŁYM I SWOBODNYM[106]

W klasyfikacji języków stosuje się trzy podstawowe kryteria: genetyczne, geograficzne i typologiczne (MAJEWICZ 1989:10). Klasyfikacja typologiczna opiera się na kryteriach fonologicznych, morfologicznych, składniowych i semantycznych. Zaprezentowane w poprzednim rozdziale wyniki badań wskazują na możliwy związek metod sekwencyjnych z klasyfikacją fonologiczną opartą na prozodycznych cechach języka. W poniższym rozdziale wykażemy, iż związek taki istotnie występuje i umożliwia testowanie nowych hipotez badawczych. Należy w tym miejscu wyraźnie zastrzec, że w przeciwieństwie do języków sztucznych, języki naturalne w zasadzie nigdy nie są typologicznie jednorodne. Mówiąc o określonym typie, ma się więc zawsze na myśli jedynie dominującą statystycznie tendencję.

Do efektywnych kryteriów fonologicznych zalicza się cechy prozodyczne języka (akcentuację, intonację). Pozwalają one wyróżnić języki, w których akcent wpływa na znaczenie wyrazu (tzw. języki prozodyczne) i takie, w których związek pomiędzy akcentem a znaczeniem nie zachodzi (MILEWSKI 1965:220–221, MAJEWICZ 1989:187). Z punktu widzenia analizy sekwencyjnej, przydatna jest klasyfikacja języków ze względu na rodzaj akcentu. Na poziomie ogólnym kryterium to pozwala wyróżnić:

- języki o akcencie dynamicznym
- języki iloczasowe
- języki o akcencie tonicznym
- języki tonalne

Wyjaśnienia wymagają podobnie brzmiące pojęcia akcentu tonicznego i tonalności. Jak stwierdza A. Majewicz (*ibid.* 188), „Akcent toniczny polega na wyróżnianiu pewnych ściśle określonych sylab lub ich części (mor) poprzez zrelatywizowaną do kontekstu poprzedzającego zmianę wysokości tonu.” Wśród akcentów tonicznych wyróżnia się rosnący, malejący i rosnąco-opadający[107]. Języki tonalne charakteryzują się „stałym, leksykalnie istotnym relewantnym tonem na każdej (lub prawie każdej sylabie), będącym w opozycji do innego bądź innych relatywnych tonów, które mogą wystąpić na takiej samej sylabie.” (*ibid.* 190). Tonalność określa się jako rejestrową, jeżeli wysokość tonów w obrębie sylaby nie ulega zmianie, lub konturową, jeżeli wysokość tonu w obrębie sylaby jest zmienna.

W przypadku języków o akcencie dynamicznym bardziej szczegółowa klasyfikacja opiera się natomiast na określeniu pozycji akcentu w wyrazie bądź grupie rytmicznej. Wyróżnia się tu języki o akcencie stałym, padającym zawsze na tę samą sylabę wyrazu

[106] Część prezentowanych tu wyników empirycznych opublikowana została wcześniej (por. PAWŁOWSKI 2000a i 2000b). Problem klasyfikacji języków opartej na kryteriach ilościowych (chodzi o tzw. wskaźniki Greenberga) omawiany jest także w dalszych rozdziałach (por. Część II, 5.2.1)

[107] W odniesieniu do języków klasycznych używa się też terminów akut, grawis i cyrkumfleks, przy czym cyrkumfleks może oznaczać również akcent opadająco-rosnący.

lub zestroju akcentowego (na przykład w węgierskim jest to sylaba pierwsza, w macedońskim trzecia od końca, w polskim przedostatnia, we francuskim ostatnia) oraz nie mniej zróżnicowaną klasę języków o akcencie zmiennym. W klasie tej na uwagę zasługują języki o akcencie *swobodnym*, czyli takim, którego pozycja w różnych wyrazach jest zmienna, ale dla konkretnego leksemu względnie ustabilizowana w obrębie jego paradygmatu (włoski, niemiecki), oraz języki o akcencie swobodnym i ruchomym, czyli takim, którego pozycja nie jest określona regułami, a ponadto może się zmieniać się w obrębie paradygmatu konkretnego leksemu (rosyjski). Podstawą wysuniętych i testowanych dalej hipotez jest właśnie rozróżnienie języków o akcencie stałym oraz języków o akcencie swobodnym i ruchomym.

2.1 HIPOTEZA

W poprzednim rozdziale wykazano, że sekwencję sylab akcentowanych i nie akcentowanych w języku polskim można traktować jako realizację pewnego procesu stochastycznego, który daje się opisać modelem formalnym. Model ten był na tyle czuły, że wykazywał różnice pomiędzy niektórymi odmianami stylistycznymi i wersyfikacyjnymi polszczyzny. Otóż jest bardzo prawdopodobne, że porządek sylab akcentowanych i nie akcentowanych w językach o odmiennych systemach akcentuacji będzie także generować różne szeregi czasowe. I tak, teksty w językach o akcencie stałym powinny być przeciętnie bardziej rytmiczne od tekstów w językach o akcencie swobodnym i ruchomym, co wynika z faktu, iż średnia długość wyrazów i zestrojów pojawiających się w linii tekstu jest stabilna i przy stałej pozycji sylaby akcentowanej musi generować silny, niemalże monotonny rytm. Jakiś poziom zrytmizowania pojawi się także w przypadku języków o akcencie swobodnym, ale różne rozłożenie przycisków w wyrazach będzie tę regularność zakłócać.

Pewne przesłanki mogą jednak podważyć poprawność tego rozumowania. W tekstach artystycznych, szczególnie takich, gdzie stosowana jest wersyfikacja, systemowa regularność przycisków może utrudnić autorowi tworzenie niektórych wzorców rytmicznych[108]. W takich przypadkach języki o akcencie swobodnym okazują się znacznie lepszym tworzywem pisarskim. Posiadają one szeroki repertuar synonimicznych środków reprezentujących różne schematy akcentowe i pozwalają dowolnie kształtować tkankę rytmiczną tekstu bez nadmiernych ustępstw na poziomie treści[109]. Jest to widoczne przy porównaniu języków polskiego (akcent stały, paroksytoniczny) i rosyjskiego (akcent swobodny i ruchomy). Jak zauważa J. Tuwim: „Przyczyniły się do tych krzywd, poezji Puszkina wyrządzonych, dobrze tłumaczom z rosyjskiego znane utrapienia: niedostatek rymów męskich w polszczyźnie, przy bogactwie ich w mowie rosyjskiej, i na mur ustabilizowany akcent słów polskich, gdy w rosyjskich jest zmienny i ruchliwy.” (TUWIM 1937).

[108] Na przykład każdy dwusylabowiec w języku o akcencie oksytonicznym jest potencjalnym jambem, trochej może być więc stworzony jedynie jako zestrój akcentowy.
[109] Na przykładzie języków czeskiego i angielskiego zagadnienie to omawia J. Levý (1965).

Także R. Łużny we wstępie do polskiego wydania *Eugeniusza Oniegina* w przekładzie A. Ważyka zwraca uwagę na specyfikę prozodii rosyjskiej: „Tak więc mimo formalnie jednolitego metrum wierszowego oraz stałości stroficznego układu Puszkin osiąga zdumiewającą różnorodność rytmiczną dzięki użyciu całego systemu środków organizacji wierszowej. [...] Jednym z owych sposobów było pełne wykorzystanie możliwości tej ważnej cechy języka rosyjskiego, którą stanowi niestały i ruchomy akcent wyrazowy." (ŁUŻNY 1993:LXXVI).

Konkludując, testowaną hipotezę należy sformułować następująco: teksty, których prymarną funkcją jest komunikowanie (teksty popularne, prasa, w jakimś stopniu proza artystyczna) będą bardziej rytmiczne w językach o akcencie stałym, a mniej rytmiczne w językach o akcencie swobodnym. W przypadku tekstów, w których nad komunikacyjną wyraźnie dominuje funkcja estetyczna, w szczególności artystycznych tekstów wierszowanych, poziom rytmu będzie zawsze wysoki, choć tak naprawdę sytuacja jest tutaj nieprzewidywalna i w najwyższym stopniu zależna od kunsztu autora i zastosowanej konwencji. W przypadku tekstów prymarnie komunikacyjnych, rytm będzie miał charakter systemowy, wynikający z cech języka, a tylko w niewielkim stopniu ze świadomej decyzji piszącego. W przypadku tekstów wierszowanych o silnym nacechowaniu estetycznym, na rytmikę tekstu przemożny wpływ będzie miał wybrany przez autora system wersyfikacyjny. Treść testowanej hipotezy w syntetycznej formie przedstawiono w tabeli 21.

Tab. 21 Struktura rytmiczna tekstów w językach o akcencie stałym i swobodnym

Prymarna funkcja tekstu	Porządek sylab akcentowanych i nie akcentowanych	
	Języki o akcencie stałym	Języki o akcencie swobodnym
Komunikatywna	wysoki	niski
Estetyczna	wysoki	wysoki

2.2 KORPUS TEKSTÓW, KODOWANIE, METODA

Wysunięta tu hipoteza została poddana testom na materiale języków polskiego i rosyjskiego. Oba języki są genetycznie blisko spokrewnione, wykazują też wiele podobieństw składniowych i morfologicznych (kapitalne znaczenie ma tu na przykład średnia długość wyrazu). Jedną z cech przeciwstawiających je sobie w sposób bezdyskusyjny jest miejsce akcentu wyrazowego. W dalszej części przedstawiono wyniki porównania tekstów równoległych, czyli tych samych fragmenty w oryginale i w przekładzie. Aby uniknąć niejasności związanych z kierunkiem przekładu, dla każdej grupy tekstów badano zawsze dwóch autorów – polskiego i rosyjskiego wraz z przekładami.

Teksty o funkcji prymarnie komunikacyjnej pochodzą z powieści J. Iwaszkiewicza *Sława i chwała* (2*15 prób), M. Bułhakowa *Mistrz i Małgorzata* (2*15 prób) oraz z polskiej i rosyjskiej prasy (2*20 prób). Długość każdej próby wynosi około 150 sylab.

Teksty o funkcji prymarnie estetycznej zaczerpnięto z *Pana Tadeusza* A. Mickiewicza (2*60 prób, każda długości 12 wersów) oraz z poematu *Eugeniusz Oniegin* A. Puszkina (2*64 14-wersowe strofy)[110]. Wybór tekstów powszechnie znanych autorów podyktowany był wymogami o charakterze czysto pragmatycznym – zależało nam nie tylko na próbach spełniających określone kryteria typologiczne, ale także na dostępności i jakości ich przekładów. Warunek ten, raczej nieistotny w przypadku stylu prasowego, ma wielkie znaczenie w odniesieniu do tekstów artystycznych i wiersza.

Przyjęty system kodowania był taki sam jak w przypadku analizy stylistyczno - wersyfikacyjnych odmian polszczyzny i został już szczegółowo omówiony w poprzednich rozdziałach (por. Część I, 1.1.2). Wyróżniono trzy poziomy akcentu: akcent główny (kodowany jako 1), akcent poboczny (kodowany jako ½) i brak akcentu (kodowany jako 0). Przykładowe zdanie w języku polskim „Krzyki i gromkie śmiechy dobiegały także z innego miejsca" po zakodowaniu dałoby sekwencję {1001010½010½001010}. To samo zdanie w języku rosyjskim „Крики и ревущий хохот донеслись и из другово места" byłoby reprezentowane sekwencją {100010100010001010}. Z kolei fragment *Eugeniusza Oniegina* „Когда же юности мятежной пришла Евгению пора" należałoby zakodować jako {01010001001010001}, a jego odpowiednik w wersji polskiej „A kiedy już podrastał uczeń i przyszła lat burzliwych pora" byłby reprezentowany sekwencją {0100010100101010 10}.

Szeregi czasowe wygenerowane z badanych fragmentów poddane zostały analizie metodą ARIMA w dziedzinie czasu. Zastosowano modele proste (proza) i sezonowe (teksty wierszowane). Porównanie typów i parametrów modeli umożliwiło klasyfikację prób i wnioskowanie o własnościach badanych stylów i odmian wersyfikacyjnych.

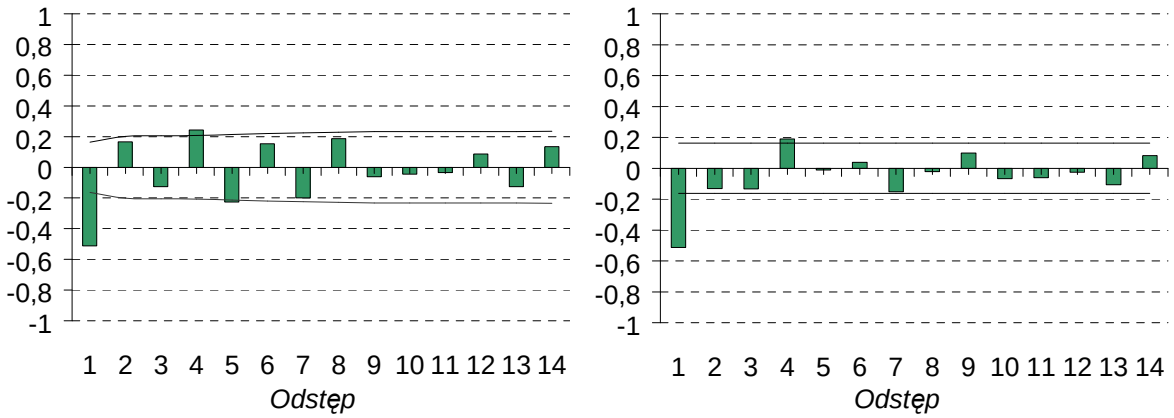
2.3 REZULTATY

PROZA ARTYSTYCZNA I STYL PRASOWO-PUBLICYSTYCZNY – ANALIZA WSTĘPNA

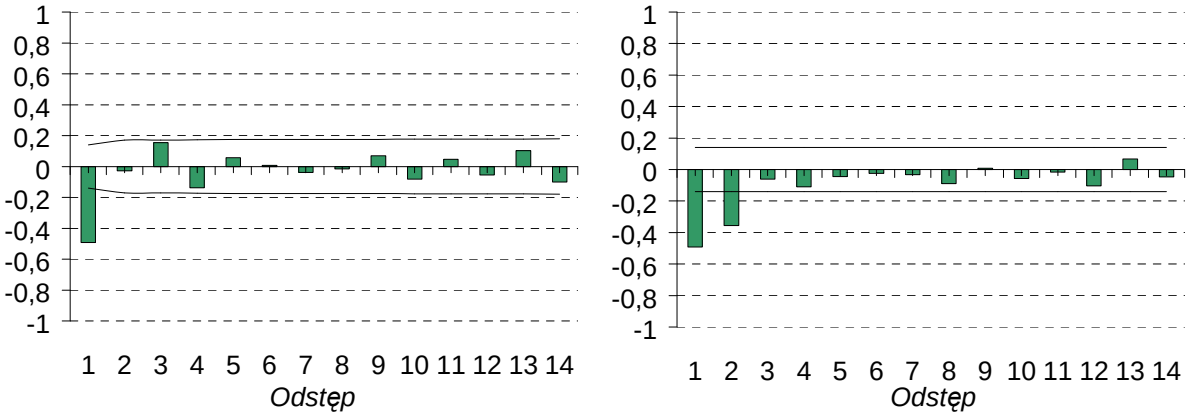
W pierwszej kolejności porównano prozę artystyczną i styl prasowo-publicystyczny. Na podstawie przeprowadzonych analiz stwierdzono, że sekwencja sylab akcentowanych i nie akcentowanych w prozie artystycznej i w tekstach prasowych w językach polskim i rosyjskim jest realizacją pewnego procesu stochastycznego i w żadnym razie nie może być uznana za losową. Kształt funkcji autokorelacji i autokorelacji cząstkowej dla większości badanych fragmentów wskazuje, iż modelem tego procesu jest MA(1) lub MA(2). Wykres 23 przedstawia funkcje ACF i PACF jednego z fragmentów prozy J. Iwaszkiewicza w języku polskim. Podobną strukturę rytmiczną posiadają teksty reprezentujące styl prasowo-publicystyczny w polszczyźnie. Funkcje ACF i PACF (Rys. 24) obliczone dla przykładowego fragmentu prasy nie różnią się w istotny sposób od analogicznych funkcji obliczonych dla prozy artystycznej.

[110] Szczegółowe dane bibliograficzne podano w ANEKSIE.

Rys. 23 Funkcje ACF (wykres lewy) i PACF (wykres prawy) dla prozy polskiej[111]



Rys. 24 Funkcje ACF (wykres lewy) i PACF (wykres prawy) dla stylu prasowo-publicystycznego w polszczyźnie[112]



Estymowane dla obu porównywanych prób modele procesów stochastycznych są jednakowe pod względem typu (ruchoma średnia), różnią się natomiast co do rzędu – dla prozy artystycznej najlepszy okazał się model rzędu drugiego (wzór 79), natomiast dla tekstu prasowego model rzędu pierwszego (wzór 80):

(79) $x_t = (1 - 0{,}61B + 0{,}21B^2)e_t$

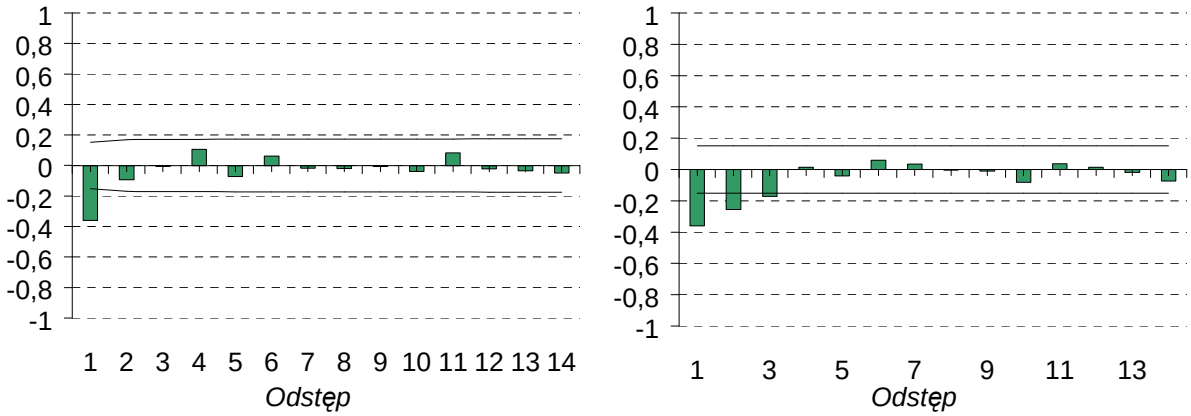
(80) $x_t = (1 - 0{,}68B)e_t$

Procent wariancji szeregu obserwowanego wyjaśniony przez oba modele jest jednak różny. Dla modelu wyrażonego wzorem (79) otrzymujemy $V_e = 27\%$, natomiast dla modelu wyrażonego wzorem (80) $V_e = 33\%$. Fakt ten jest o tyle zaskakujący, że jako zorientowany na obiektywny przekaz informacji i pozbawiony walorów estetycznych, styl dziennikarski powinien być raczej mniej rytmiczny od prozy artystycznej. Dalsze próby pokazały jednak, że wynik ten jest odosobniony i wyraźnie odbiega od średniej (por. Tab. 22).

[111] ANEKS – IWASZKIEWICZ 1973:380 (t.1).
[112] Analizowany fragment pochodzi z dziennika „Rzeczpospolita” (patrz ANEKS).

Podobieństwo struktury rytmicznej stylu prasowo-publicystycznego i prozy artystycznej widoczne jest także w języku rosyjskim. Wykresy 25 i 26 przedstawiają funkcje ACF i PACF dla fragmentu prozy J. Iwaszkiewicza w przekładzie rosyjskim i dla tekstu prasowego.

Rys. 25 Funkcje ACF (wykres lewy) i PACF (wykres prawy) dla prozy rosyjskiej[113]

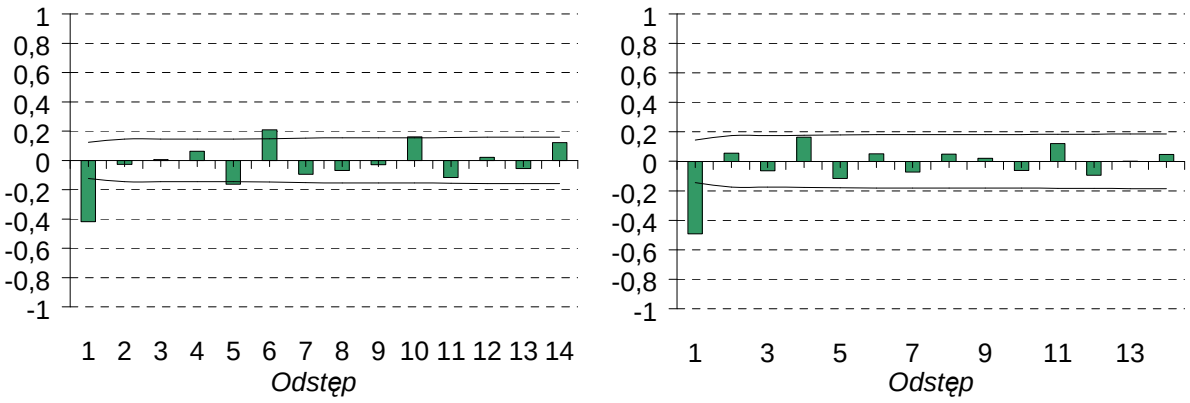


Rys. 26 Funkcje ACF (wykres lewy) i PACF (wykres prawy) dla stylu prasowo-publicystycznego w języku rosyjskim[114]

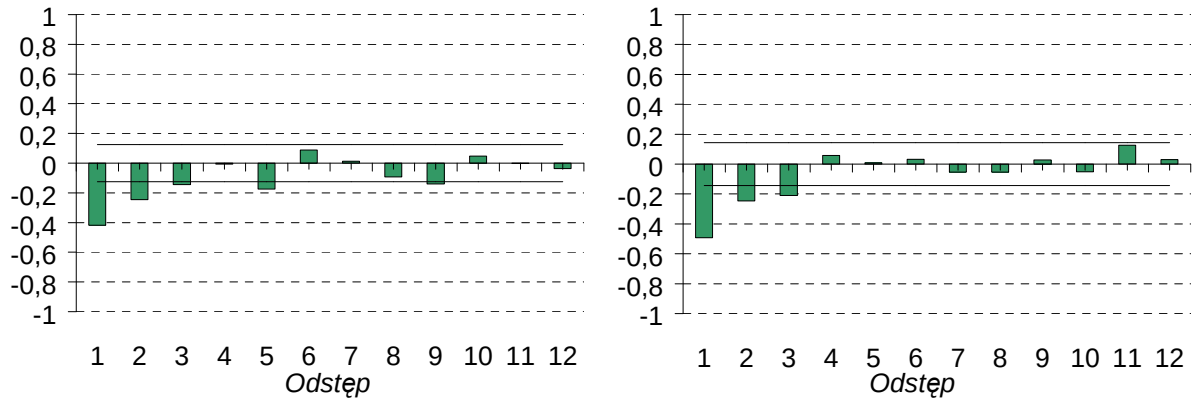
Porównano następnie charakterystykę rytmiczną prozy artystycznej w językach polskim i rosyjskim. Wykresy 27 i 28 przedstawiają zestawienie funkcji ACF i PACF dla równoległych fragmentów powieści *Mistrz i Małgorzata* w wersji rosyjskiej i polskiej. Ich analiza wskazuje na podobieństwo szeregów ze względu na ich porządek sekwencyjny: w obu przypadkach obserwujemy negatywną korelację dla odstępu pierwszego, po którym ACF się urywa, także w obu przypadkach PACF jest funkcją gasnącą. Taki kształt funkcji ACF i PACF jednoznacznie wskazuje na model MA(1).

[113] ANEKS – ИВАШКЕВИЧ 1975:368.
[114] Analizowany fragment pochodzi z dziennika „Известия” (patrz ANEKS).

Rys. 27 Funkcje ACF dla równoległych fragmentów prozy rosyjskiej (wykres lewy) i polskiej (wykres prawy)[115]



Rys. 28 Funkcje PACF dla równoległych fragmentów prozy rosyjskiej (wykres lewy) i polskiej (wykres prawy)[116]



Bardzo istotna jest widoczna na wykresach 27 i 28 różnica wartości funkcji ACF i PACF dla kolejnych odstępów. Zgodnie z oczekiwaniami, dla języka rosyjskiego są one niższe niż dla polskiego. Można z tego wywnioskować, że także wartość współczynnika V_e, czyli procent zmienności szeregu obserwowanego wyjaśniony przez model, będzie dla języka rosyjskiego niższy niż dla polskiego. W istocie, dla wersji oryginalnej badanej próby (język rosyjski) otrzymujemy $V_e = 24\%$, podczas gdy w wersji polskiej stwierdzono $V_e = 30\%$. Podobne spostrzeżenia nasuwają się przy porównaniu funkcji ACF i PACF obliczonej dla tekstów publicystyczno-prasowych w językach polskim i rosyjskim (por. Rys. 24 i 26), gdzie stwierdzono odpowiednio $V_e = 33\%$ i $V_e = 17\%$. Gdyby prawidłowość ta wystąpiła w innych badanych fragmentach, otrzymalibyśmy wyraźne potwierdzenie pierwszej części testowanej hipotezy (Tab. 21).

[115] ANEKS – БУЛІГАКОВ 1998:182 oraz BUŁHAKOW 1988:236.
[116] ANEKS – БУЛІГАКОВ 1998:182 oraz BUŁHAKOW 1988:236.

PROZA ARTYSTYCZNA I STYL PRASOWO-PUBLICYSTYCZNY – PODSUMOWANIE[117]

W tabeli 22 zamieszczono uśrednione wyniki otrzymane dla badanych fragmentów prozy polskiej i rosyjskiej oraz dla stylu prasowego. Wartości V_e dla różnych stylów tego samego języka są do siebie zbliżone, co dowodzi, iż rytm prozy oparty na akcencie wyrazowym nie jest cechą stylu osobniczego. Uwaga ta istotna jest przy interpretacji niewątpliwych różnic, jakie ujawniły się pomiędzy badanymi językami. Jak widać, wszystkie próby w języku polskim są rytmiczniejsze od równoległych (dla prozy) bądź analogicznych (w przypadku prasy) prób w języku rosyjskim. Skoro różnicy tej nie da się wyjaśnić cechami stylu osobniczego, tak wysoki (względnie niski) stopień uporządkowania rytmicznego musi wynikać z systemowych cech języka o stałym (względnie swobodnym) akcencie wyrazowym. Wnioskowanie to potwierdza przedstawioną na wstępie hipotezę, iż proza i styl prasowo-publicystyczny w językach o stałym akcencie wyrazowym są rytmiczniejsze od analogicznych tekstów w językach o akcencie swobodnym i ruchomym. Systemowy i mierzalny charakter tej różnicy pozwala traktować ją jako jedno z potencjalnych ilościowych kryteriów w typologicznej klasyfikacji języków.

Tab. 22 Struktura rytmiczna tekstów w językach o stałym i swobodnym akcencie wyrazowym

	Język polski (akcent stały)	Język rosyjski (akcent swobodny)
Proza artystyczna (Bułhakow)	**35,5%**	**22,3%**
Proza artystyczna (Iwaszkiewicz)	**33,7%**	**20,7%**
Styl prasowo-publicystyczny	**31,7%**	**17,8%**

Otrzymany rezultat nie potwierdza natomiast hipotezy, zgodnie z którą teksty prasowe jako mniej staranne i nastawione przede wszystkim na informowanie powinny być mniej rytmiczne od prozy artystycznej, której ambicją jest wywołanie przeżycia estetycznego. Wynik ten można jednak racjonalnie wytłumaczyć, przywołując pewne aspekty stylu prasowego, na które nie zwrócono dotąd uwagi, a mianowicie jego perswazyjność i stereotypowość (formułkowość, korzystanie z gotowych zwrotów). Cechy te mogą wywoływać zjawisko rytmu bez jakiegokolwiek związku z estetyką tekstu.

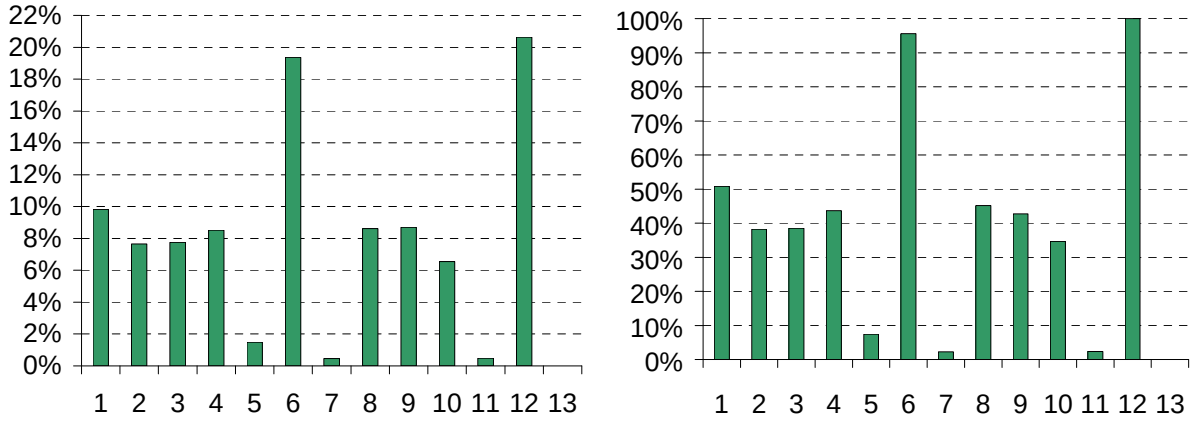
WIERSZ – ANALIZA WSTĘPNA

Pan Tadeusz (dalej PT) napisany jest trzynastozgłoskowym wierszem sylabicznym z paroksytoniczną klauzulą i cezurą po siódmej sylabie. Układ ten jest wyraźnie widoczny na wykresie przedstawiającym średni rozkład przycisków na kolejne sylaby wersu (Rys. 29). Przyciski padające na sylaby wyznaczające główne działy wersu (5, 6, 7 oraz 11, 12, 13)

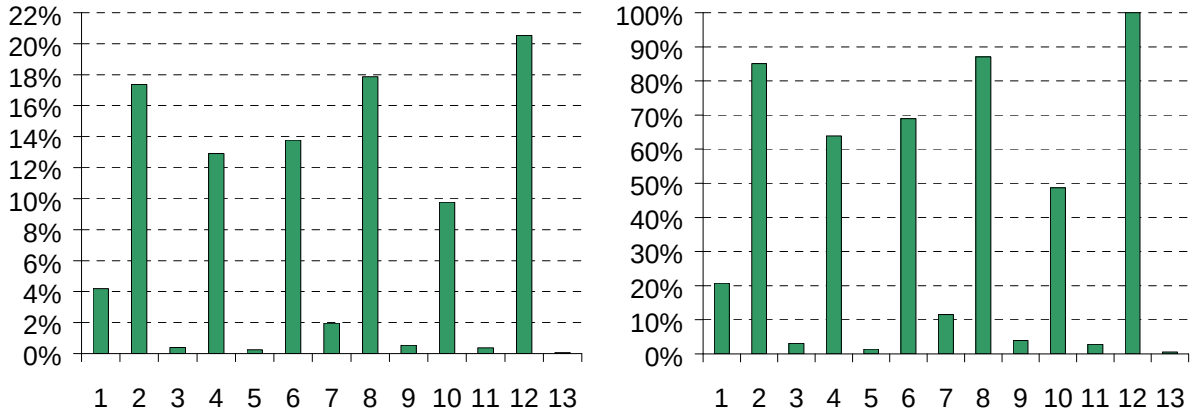
[117] Szczegółowe wyniki analiz zamieszczono w ANEKSIE.

są niemal całkowicie zdeterminowane formalną strukturą wiersza. Natomiast rozkład przycisków na pozycjach 1, 2, 3, 4 oraz 8, 9, 10 jest mniej przewidywalny i stanowi podstawowe źródło zmienności szeregów czasowych generowanych przez tekst PT. Wersja rosyjska w ogólnych zarysach zachowuje ten układ, jednak rozkład przycisków jest bardziej monotonny, brak jest też wyraźnie wyodrębnionej klauzuli i średniówki (Rys. 30).

Rys. 29 Rozkład przycisków akcentowych w trzynastozgłoskowcu sylabicznym (na przykładzie *Pana Tadeusza*)[118]



Rys. 30 Rozkład przycisków akcentowych w trzynastozgłoskowcu sylabicznym (na przykładzie rosyjskiego przekładu *Pana Tadeusza*)



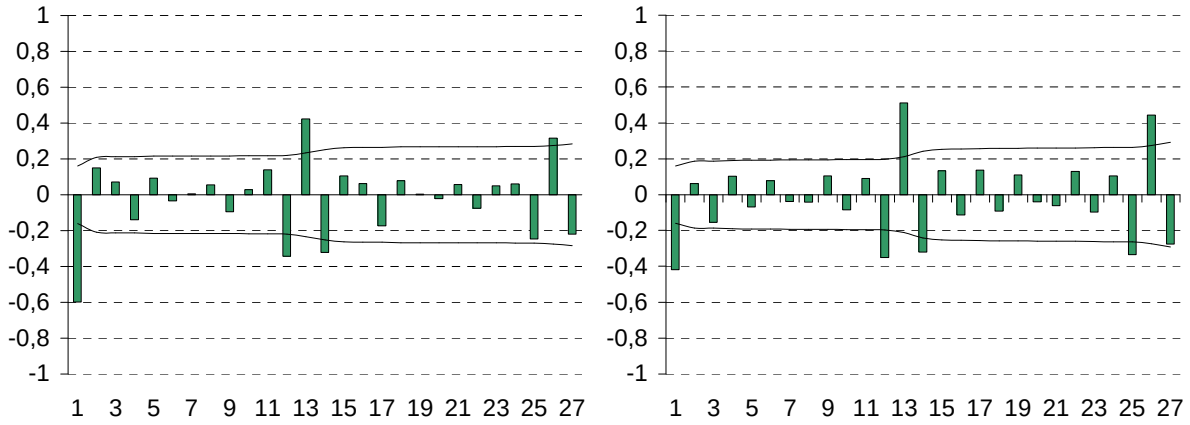
Eugeniusz Oniegin w cytowanym przekładzie (dalej EO) jest utworem stroficznym złożonym z czternastowersowych strof o naprzemiennym układzie wersów ośmio- i dziewięciosylabowych. W wersach ośmiozgłoskowych występuje klauzula oksytoniczna, a w wersach dziewięciozgłoskowych paroksytoniczna. W polskim przekładzie A. Ważyka (patrz ANEKS) strofy składają się z dwunastu wersów dziewięciosylabowych z klauzulą paroksytoniczną i dwóch wersów ośmiosylabowych z klauzulą oksytoniczną. Układ taki

[118] Wykresy 29 i 30 sporządzono na podstawie wszystkich badanych prób (por. ANEKS). Wykres lewy przedstawia średni rozkład przycisków w wersie, natomiast wykres prawy procent sylab akcentowanych na danej pozycji. Podobne wyniki uzyskały Z. Kopczyńska i L. Pszczołowska (1968).

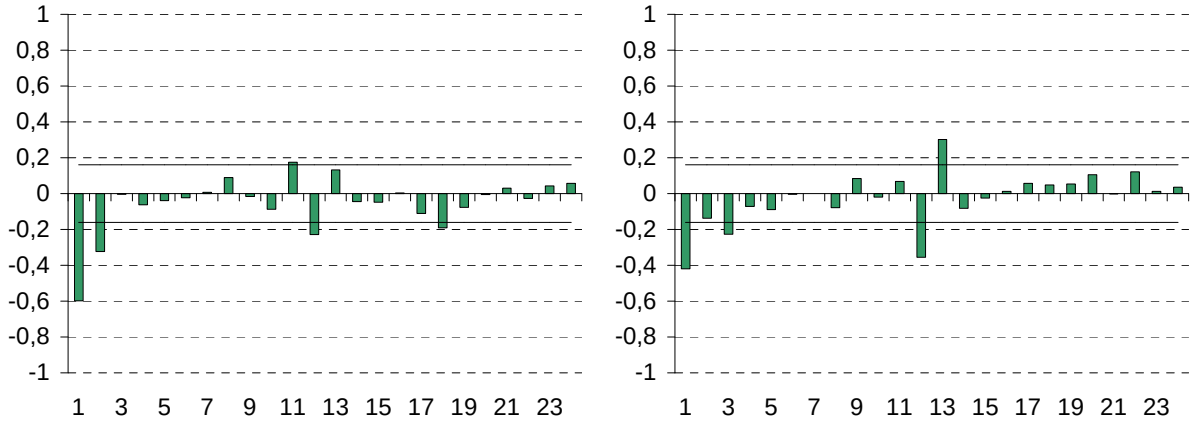
utrudnia przedstawienie średniego rozkładu przycisków w wersie, ponieważ faktycznie ekwiwalentną, powtarzalną jednostką formalną jest tu cała strofa.

Wykresy 31 i 32 przedstawiają funkcje autokorelacji zwykłej (ACF) i cząstkowej (PACF) dla wybranych fragmentów obu analizowanych utworów w oryginale i w przekładach[119]. Nawet pobieżny ogląd funkcji ACF pozwala zauważyć bardzo wyraziste składowe deterministyczne wszystkich badanych szeregów, niezależnie od tego, jaki język i utwór reprezentują. Jednak, jak pokazały dalsze testy, zróżnicowanie kolejnych fragmentów jest duże i nie pozwala wnioskować, na podstawie dwóch prób, o cechach całych utworów lub o relacjach zachodzących pomiędzy oryginałami i przekładami. Przykłady poniższe mają natomiast za zadanie zilustrować tok postępowania i rozumowanie towarzyszące analizie wszystkich pozostałych fragmentów i tym samym uwiarygodnić przedstawione dalej uogólnienia.

Rys. 31 Funkcja ACF dla sekwencji akcentowej trzynastozgłoskowca w języku polskim (lewa) i rosyjskim (prawa)[120]



Rys. 32 Funkcja PACF dla sekwencji akcentowej trzynastozgłoskowca w języku polskim (lewa) i rosyjskim (prawa)



[119] PT – V/199, EO – II/2.
[120] Wykresy 31 i 32 sporządzono dla fragmentu PT V/199 (patrz ANEKS).

W przypadku PT w języku polskim (Rys. 31), wyraźny prążek funkcji autokorelacji pojawia się przy odstępie pierwszym, a następnie przy odstępach 13 i 26. Prążek pierwszy wskazuje na silną negatywną korelację kolejnych akcentów, co oznacza, że sylaba akcentowana będzie najczęściej wymuszać pojawianie się na następnej pozycji sylaby nie akcentowanej i *vice versa*, podczas gdy dłuższe sekwencje sylab tego samego typu będą stosunkowo rzadkie. Znaczące, dodatnie wartości ACF na wielokrotnościach liczby 13 potwierdzają wyższą niż przeciętna powtarzalność tego samego akcentu na co trzynastej sylabie. Sugeruje to estymację nieznanego procesu stochastycznego modelem sezonowym.

Autokorelacja obliczona dla PT w przekładzie rosyjskim (Rys. 31) ma podobny wygląd, chociaż układ prążków jest nieco bardziej regularny. Spostrzeżenie to potwierdza się, jeżeli porównamy uśrednione rozkłady nacisków w wersie oryginału i przekładu (Rys. 29, 30). W wersji polskiej stabilizująco działa jedynie sześć sylab (pozycje 5, 6, 7 oraz 11, 12, 13) podczas gdy pozostałe są źródłem zmienności, natomiast w wersji rosyjskiej przyciski są w zasadzie ustabilizowane na wszystkich pozycjach (jak widać, sylabizm wiersza oryginału oddano w przekładzie sylabotonizmem). Znacznie trudniej jest określić jednoznacznie kształt funkcji autokorelacji cząstkowej (Rys. 32). Dla tekstu polskiego PACF raczej wygasa, natomiast w wersji rosyjskiej możliwe są obie interpretacje.

Kształt funkcji ACF i PACF widocznych na wykresach 31 i 32 sugeruje estymację modelu złożonego ze składowej prostej typu MA(1) lub AR(1) oraz jakiejś składowej sezonowej z odstępem równym długości wersu, czyli $k = 13$. Jeśli chodzi o składowe proste, dla wersji polskiej PT najlepszy okazał się model prosty MA(1), a dla przekładu rosyjskiego model AR(1):

(81) $x_t = (1 - 0{,}65B)e_t$ (oryginał polski)

(82) $x_t = -0{,}42x_{t-1} + e_t$ (przekład rosyjski)

Podobnie jak w poprzednich przypadkach, efektywność modelu mierzy się procentem wyjaśnionej wariancji (zmienności) szeregu obserwowanego (por. wzór 70). Efektywność modeli prostych stosowanych do tekstu wierszowanego jest na ogół niska – w tym przypadku dla oryginału polskiego mamy $V_e = 38\%$, a dla przekładu rosyjskiego $V_e = 22\%$. Jest to skutkiem pominięcia „efektu wersyfikacji", czyli regularności wynikającej z powtarzania tych samych wartości (cech sylab) na stałych pozycjach wersu. Mankamentu tego nie posiadają modele sezonowe. Badany fragment *Pana Tadeusza* w wersji polskiej opisano modelem SARMA(0,1)(1,1)$_{13}$ natomiast dla przekładu rosyjskiego użyto modelu SARMA(1,0)(1,1)$_{13}$:

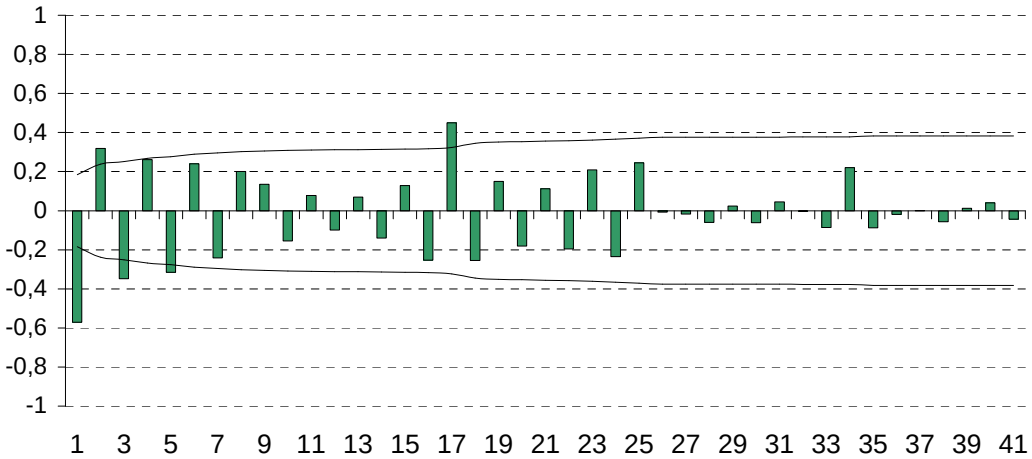
(83) $(1 - B^{13})x_t = (1 - 0{,}55B)(1 - 0{,}84B^{13})e_t$ (oryginał polski)

(84) $(1 - 0{,}21B)(1 - 0{,}98B^{13})x_t = (1 - 0{,}72B^{13})e_t$ (przekład rosyjski)

Jak widać, do obu wersji językowych zastosowano podobny typ modelu i ten sam odstęp sezonowy pokrywający się z długością wersu. Po uwzględnieniu wersyfikacji efektywność obu modeli dla badanego fragmentu znacząco wzrasta. W wersji polskiej PT $V_e = 50\%$, natomiast w wersji rosyjskiej $V_e = 38\%$. Mimo to nadal zauważalna jest różnica pomiędzy wartościami parametru V_e dla oryginału i przekładu. Błędem byłoby jednak interpretować ten jednostkowy wynik już teraz i wnioskować na jego podstawie o własnościach całych utworów czy odmian wersyfikacyjnych. Przy interpretacji należałoby bowiem uwzględnić co najmniej dwie grupy czynników, których wpływ nie jest do końca jasny. Po pierwsze, może (ale wcale nie musi) istnieć jakaś formalna relacja pomiędzy rytmiczną strukturą oryginału i przekładu. W tym przypadku jest to jedynie ta sama długość wersu, co skutkuje takim samym odstępem sezonowym modelu, natomiast inny jest rozkład wewnętrznych działów wersu. Po drugie, oba języki posiadają różne systemy akcentuacji, co w przypadku stylu prasowo-publicystycznego i prozy znacząco wpłynęło na strukturę rytmiczną tekstu i, *ipso facto*, wartość współczynników modelu. Nie jest jednak wcale pewne, czy tekst artystyczny, stanowiący skrajne przeciwieństwo wypowiedzi spontanicznej, zachowa się podobnie. Wyjątkowość wypowiedzi tego rodzaju skazuje więc badacza na indukcjonizm – dopiero uśrednienie uzyskanych dla dużej liczby prób pozwala na bardziej miarodajne uogólnienia.

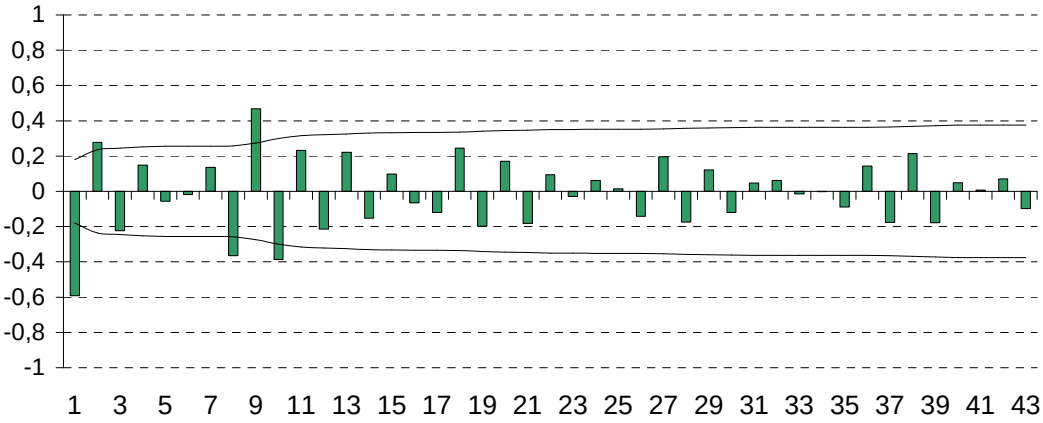
Zupełnie inaczej przedstawia się struktura rytmiczna tekstu *Eugeniusza Oniegina*. Układ prążków funkcji ACF jest w obu wersjach językowych naprzemienny (Rys. 33 i 34) i wskazuje na bardzo regularny rytm tekstu wyznaczony w sekwencją następujących bezpośrednio po sobie sylab akcentowanych i nie akcentowanych.

Rys. 33 Funkcja ACF dla sekwencji akcentowej *Eugeniusza Oniegina* w wersji oryginalnej[121]



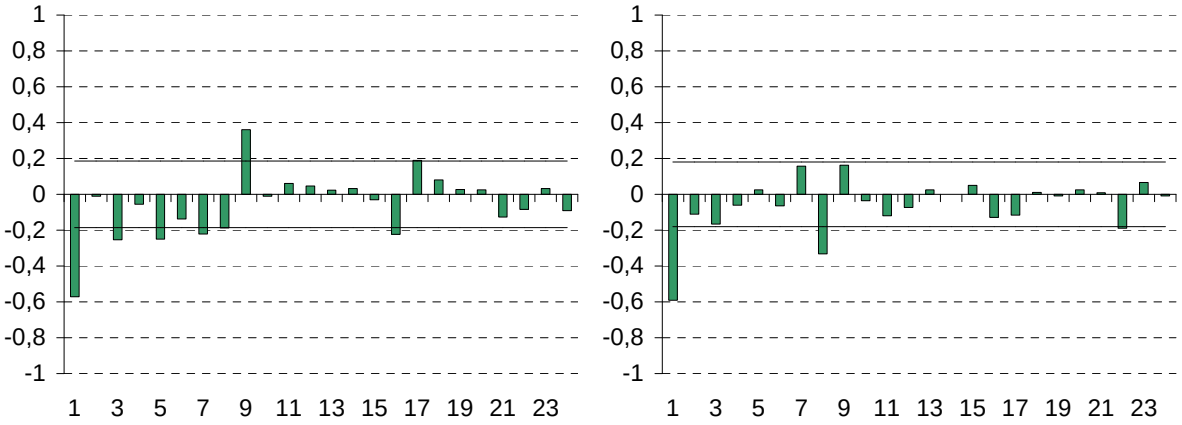
[121] Wykresy 33–35 sporządzono dla fragmentu EO II/2 (patrz ANEKS).

Rys. 34 Funkcja ACF dla sekwencji akcentowej *Eugeniusza Oniegina* w przekładzie polskim



W wersji rosyjskiej uwagę zwraca sezonowy prążek ACF dla odstępu 17 i jego wielokrotności. Jest to o tyle interesujące, że formalnie tzw. strofę onieginowską tworzy naprzemienny układ wersów dziewięcio- i ośmiosylabowych. Brak stałego rytmu klauzulowego w przypadku rosyjskiej wersji EO (dla odstępów 8 i 9 nie zauważamy nic istotnego) jest więc rekompensowany regularnym akcentem występującym średnio na każdej 17 sylabie. Oznaczałoby to, że formalny układ krótkich wersów *Eugeniusza Oniegina* jest pozorny i w rzeczywistości ukrywa strukturę rytmiczną wyznaczoną sekwencją ekwiwalentnych odcinków siedemnastosylabowych. W cytowanym polskim przekładzie EO także dominuje rytm naprzemienny (Rys. 34), ale przy odstępie 9 pojawia się sezonowość, co wskazuje na istnienie wyraźnego działu klauzulowego w wersie.

Rys. 35 Funkcja PACF dla sekwencji akcentowej *Eugeniusza Oniegina* w wersji oryginalnej (lewa) i w przekładzie (prawa)



Kolejnym stadium analizy jest estymacja modeli procesów stochastycznych zawartych w strukturze rytmicznej badanych tekstów. Wybór właściwego modelu wymaga jednak znajomości funkcji autokorelacji cząstkowej. Jak widać (Rys. 35), w obu wersjach językowych PACF w zasadzie urywa się po pierwszym odstępie. Uwzględniając gasnący

charakter oraz sezonowość funkcji ACF, poszukiwany model będzie więc kombinacją składowej prostej typu AR(1) i jakiejś składowej sezonowej.

Dla obu wersji językowych estymowano najpierw modele proste typu AR(1), a następnie modele złożone typu sezonowego. W wersji rosyjskiej najlepszym modelem złożonym okazał się SARMA(1,0)(1,1)$_{17}$, zaś dla przekładu polskiego estymowano model SARMA(1,0)(1,1)$_9$. Typ modelu jest więc taki sam, natomiast odstęp sezonowy wskazujący na rzeczywistą długość ekwiwalentnej jednostki rytmicznej w tekście jest różny. Modele proste typu autoregresji mają postać:

(85) $x_t = -0{,}58x_{t-1} + e_t$ (oryginał rosyjski)

(86) $x_t = -0{,}6x_{t-1} + e_t$ (przekład polski)

Ich efektywność jest zaskakująco wysoka: dla tekstu rosyjskiego $V_e = 33\%$, a dla polskiego $V_e = 35\%$. Wynika to zapewne z regularnego, naprzemiennego rytmu, który widoczny był dzięki wykresom funkcji ACF dla obu wersji (Rys. 33 i 34). Modele sezonowe mają następującą postać:

(87) $(1-0{,}58B)(1-0{,}64B^{17})x_t = (1-0{,}16B^{17})e_t$ (oryginał rosyjski)

(88) $(1-0{,}53B)(1-0{,}64B^{9})x_t = (1-0{,}33B^{9})e_t$ (przekład polski)

Efektywność modeli sezonowych jest wyższa od efektywności modeli prostych: dla oryginału rosyjskiego $V_e = 48\%$, natomiast dla przekładu polskiego $V_e = 42\%$. Tak jak w przypadku analizowanego wcześniej fragmentu PT, nie należy tego jednostkowego rezultatu uogólniać. Z dużym prawdopodobieństwem można jedynie oczekiwać, że w innych próbach z danego utworu odstęp sezonowy będzie taki sam, a modele sezonowe będą przeciętnie efektywniejsze od prostych. Warto też odnotować już teraz potwierdzenie wcześniejszego spostrzeżenia, iż w wersji rosyjskiej EO rzeczywista długość ekwiwalentnego odcinka rytmicznego wynosi siedemnaście sylab, a więc jest długością złożenia dwóch wersów. Zjawiska tego nie obserwujemy w cytowanym przekładzie polskim.

WIERSZ – PODSUMOWANIE

We wszystkich badanych próbach stwierdzono, że sekwencje sylab akcentowanych i nie akcentowanych są realizacjami procesów stochastycznych, a więc nie mają charakteru losowego. Wyniki uzyskane wcześniej dla prozy i wiersza polskiego uzasadniały estymację, także w tym przypadku, modeli prostych – opisujących porządek rytmiczny tekstu z pominięciem „efektu wersyfikacji” – oraz złożonych modeli sezonowych uwzględniających wersyfikację. Dla każdej próby i modelu obliczano też procent wyjaśnionej wariancji szeregu obserwowanego (V_e). Pozwoliło to ustalić, jak bardzo dzięki wersyfikacji wzrasta poziom zrytmizowania tekstu. Przypomnijmy, że jako syntetyczna miara rytmicznego uporządkowania tekstu, V_e jest parametrem stanowiącym podstawę przyjęcia bądź odrzucenia wysuniętej na wstępie hipotezy.

Jeśli chodzi o modele proste, ich największe zróżnicowanie stwierdzono w tekście *Pana Tadeusza* w języku polskim (Tab. 23). Dominuje co prawda model ruchomej średniej MA(1), ale pojawiają się także procesy autoregresji. Fakt ten świadczy o wielkim bogactwie wersyfikacji mickiewiczowskiej. Pozostałe utwory, w szczególności przekłady, charakteryzują się bardziej unormowaną i przewidywalną strukturą rytmiczną – wszędzie tam wykryto obecność procesów typu AR(1), dobrze reprezentujących proste szeregi naprzemienne (Tab. 24). Różnicę wartości V_e dla procesów prostych w przypadku epickiego tekstu *Pana Tadeusza* można jednak łatwo wyjaśnić, przywołując wyniki podobnych badań dla polskiej i rosyjskiej prozy (Tab. 22). Okazuje się, że poziom rytmu wyjaśniony modelem prostym w polskim i rosyjskim tekście PT jest porównywalny ze średnim poziomem rytmu prozy artystycznej w tych językach. Wzrost rytmicznego uporządkowania tekst zawdzięcza dopiero wersyfikacji. Dla polskiego tekstu PT wynosi on +10%, a dla rosyjskiego aż +22%. Jak wykazano, fakt ten jest skutkiem stałego, a w rosyjskim zmiennego akcentu wyrazowego (proza polska jest systemowo bardziej rytmiczna). Jednak ostatecznie, wartości V_e reprezentujące całość struktury rytmicznej są bardzo podobne (Tab. 24).

Tab. 23 Typy modeli estymowanych w polskiej wersji *Pana Tadeusza*

model	AR(1)	AR(2)	MA(1)	MA(2)
liczba fragmentów	15	2	37	6

W tym przypadku najważniejsze są złożone modele sezonowe syntetyzujące cechy systemowe języka (zwykle jest to naprzemienny porządek sylab akcentowanych i nie akcentowanych reprezentowany w modelu przez składową prostą) oraz specyficzne cechy wersyfikacji (długość wersu i rozkład przycisków w wersie reprezentowane w modelu przez składową sezonową). Ostateczne wyniki testów (Tab. 24) potwierdzają hipotezę mówiącą o braku zależności pomiędzy pozycją akcentu wyrazowego (stały lub ruchomy) a poziomem zrytmizowania wierszowanych tekstów artystycznych. Co prawda zróżnicowanie wartości V_e dla modeli prostych jest w jednym przypadku znaczne (dla *Pana Tadeusza* w oryginale obserwujemy $V_e = 35\%$, a w przekładzie $V_e = 25\%$), jednak, jak już wspomniano, modele te pomijają wersyfikację i nie mogą być traktowane jako reprezentatywne dla utworów wierszowanych. Modele sezonowe (uwzględniające wersyfikację) wykazują po uśrednieniu zbliżony poziom rytmicznego uporządkowania tekstów w porównywanych tu wersjach językowych. Różnica w wartości parametru V_e obserwowana w obu przypadkach wynosi zaledwie 2%. Wynika z tego, że obserwowana różnica wartości parametru V_e dla różnych utworów (dla PT mamy średnio $V_e = 46\%$ a dla EO $V_e = 38\%$) nie ma związku z systemem akcentuacji danego języka, a jedynie z rodzajem użytego metrum. Oryginalność i niekonwencjonalność strofy onieginowskiej urozmaica rytm wiersza obniżając jednak wartość V_e. W przypadku PT w wersji oryginalnej wartość V_e jest natomiast zaskakująco wysoka, jeżeli uwzględnić brak regularnego powtarzania przycisków na parzystych sylabach wersu (por. Rys. 29).

Tab. 24 Typy estymowanych modeli i średnie wartości V_e

Badany utwór	Język	Typ modelu	V_e
Pan Tadeusz	polski	modele proste (Tab. 23)	35%
Pan Tadeusz	rosyjski	prosty AR(1)	25%
Pan Tadeusz	polski	modele sezonowe[122]	45%
Pan Tadeusz	rosyjski	sezonowy SARMA(1,0)(1,1)$_{13}$	47%
Eugeniusz Oniegin	rosyjski	prosty AR(1)	32%
Eugeniusz Oniegin	polski	prosty AR(1)	33%
Eugeniusz Oniegin	rosyjski	sezonowy SARMA(1,0)(1,1)$_{17}$	37%
Eugeniusz Oniegin	polski	sezonowy SARMA(1,0)(1,1)$_{9}$	39%

Z przeprowadzonych analiz wynika więc, że brak w leksyce języka o akcencie stałym pewnych układów rytmicznych w postaci gotowych leksemów nie jest dla dobrego autora żadną przeszkodą. Konkluzja ta przeczy obiegowym poglądom o rzekomym istnieniu języków uprzywilejowanych – szczególnie melodyjnych, śpiewnych, poetyckich, lepiej poddających się zewnętrznym schematom wersyfikacyjnym. Co prawda wyjątkowy artyzm użytych tekstów skłania do ostrożnej interpretacji, jednak trudno zaprzeczyć, iż w przeciwieństwie do prozy, rytmiczna struktura wierszowanego tekstu artystycznego o dużych walorach estetycznych jest przede wszystkim wyrazem talentu autora, a jedynie w minimalnym stopniu odzwierciedla charakter prozodii użytego tworzywa językowego.

Badanie potwierdziło też, że stwierdzony empirycznie odstęp sezonowy pokrywa się z rzeczywistą długością minimalnej ekwiwalentnej jednostki rytmicznej tekstu (lub jego wielokrotnością). W przeważającej liczbie przypadków zachodzi zgodność pomiędzy długością takiej jednostki a długością wersu. Jednak w przypadku strofy onieginowskiej okazało się, że faktyczna, przeciętna długość powtarzalnego odcinka odpowiada złożeniu dwóch kolejnych wersów (9+8=17 sylab). Dodajmy, że już wcześniej zauważono, iż w ośmiosylabowym sylabotoniku polskim minimalnym odcinkiem ekwiwalentnym może być czterosylabowy hemistych (Tab. 15). Zastosowany przez J. Brzechwę układ ośmiosylabowy miał oczywiście przekonujące uzasadnienie w tym, że lepiej odpowiadał przeciętnej długości typowej jednostki składniowej (por. PSZCZOŁOWSKA 1965).

Zastanawiający jest też fakt, iż w obu przypadkach minimalnie bardziej rytmiczne od oryginałów są przekłady. Zbadany materiał nie jest naszym zdaniem wystarczająco zróżnicowany, by traktować tę relację jako empirycznie potwierdzoną prawidłowość. Można jednak na tej podstawie pokusić się o wysunięcie dobrze ugruntowanej hipotezy mówiącej, iż w przypadku utworów tej klasy oryginały będą literacko lepsze od przekładów, co objawiać się powinno między innymi występowaniem niekonwencjonalnych

[122] W zależności od typu modelu prostego (Tab. 23) estymowano modele sezonowe SARMA(1,0)(1,1)$_{13}$, SARMA(0,1)(1,1)$_{13}$, SARMA(2,0)(1,1)$_{13}$ oraz SARMA(0,2)(1,1)$_{13}$.

i przez to trudniejszych w tłumaczeniu rozwiązań wersyfikacyjnych. Widać to w przedstawionych tu utworach: polski tłumacz *Eugeniusza Oniegina* uprościł strukturę wersyfikacyjną oryginału, zastępując większość ośmiosylabowych wersów z klauzulą oksytoniczną i rymem męskim wersami z paroksytoniczną klauzulą i typowym dla polszczyzny rymem żeńskim. Z kolei tłumacz *Pana Tadeusza* wprowadził regularny, typowy dla sylabotonizmu akcent padający na parzystych sylabach wersu, rezygnując ze specjalnego wyróżnienia średniówki i klauzuli (Rys. 5).

2.4 PROZODIA JĘZYKÓW O AKCENCIE STAŁYM I SWOBODNYM – PODSUMOWANIE

Sumaryczne wyniki przeprowadzonych testów (Tab. 25) pozwalają utrzymać wysuniętą na wstępie hipotezę (Tab. 21), zgodnie z którą teksty pozbawione wersyfikacji, tworzone w sposób mniej lub bardziej spontaniczny w językach o akcencie stałym, będą rytmiczniejsze od analogicznych tekstów w językach o akcencie swobodnym. Oznacza to, że procent zmienności obserwowanego szeregu jednostek akcentowych wyjaśniony przez model procesu stochastycznego (V_e) może być uwzględniany w klasyfikacji języków opartej na kryteriach fonologicznych. Warto podkreślić, że V_e ma charakter ilościowy i dzięki temu może zostać użyty jako jeden z parametrów w algorytmach analizy wielowymiarowej, stosowanych z powodzeniem w klasyfikacji typologicznej (por. BATAGELJ et al. 1992).

Tab. 25 Rytm tekstu w językach o akcencie stałym i swobodnym – wyniki

Prymarna funkcja tekstu (styl)	**Typ akcentuacji**	
	a. stały (polski)	**a. swobodny (rosyjski)**
Komunikacyjna		
Proza artystyczna (M. Bułhakow)	35,5%	22,3%
Proza artystyczna (J. Iwaszkiewicz)	33,7%	20,7%
Styl prasowo-publicystyczny	31,7%	17,8%
Estetyczna		
Wiersz (*Pan Tadeusz*)	45%	47%
Wiersz (*Eugeniusz Oniegin*)	39%	37%

Wbrew pozorom, materiałem językowym, na którym można oprzeć rzeczoną klasyfikację fonologiczną, mogą być dowolne teksty, także artystyczne utwory wierszem. W takim przypadku należy po prostu usunąć („odfiltrować") „efekt wersyfikacji", pozostawiając jedynie rytm lingwistyczny (operację taką przeprowadzono, analizując fragment *Beniowskiego* – por. Rys. 15, 16). Jest to jednak droga okrężna i w praktyce z pewnością prościej jest posłużyć się od razu tekstami prozatorskimi. O możliwości tej wspominamy w tym miejscu jedynie po to, by oddalić ewentualny zarzut arbitralnego ograniczenia zakresu występowania badanej prawidłowości. Bądź co bądź każdy język jest całością

funkcjonującą na jednakowych zasadach i hipoteza pretendująca do statusu prawa językowego nie może obowiązywać tylko dla pewnej grupy tekstów.

W podsumowaniu warto też zwrócić uwagę na zróżnicowanie modeli charakteryzujących badane teksty (Tab. 23 i 24). Ich rząd jest na ogół niski (przeciętnie $k = 1$), co oznacza, że do określenia charakteru każdej sylaby (akcentowana *vs* nie akcentowana) wystarcza jedna, względnie dwie sylaby poprzedzające. Szczególną uwagę warto poświęcić strukturze rytmicznej utworów wierszowanych. Jak już wspomniano, ich zastosowanie do klasyfikacji języków jest możliwe, ale ze względu na nakład pracy zupełnie nieopłacalne. Jednak składowe sezonowe charakteryzujące wersyfikację zawierają istotne informacje o charakterze genologicznym. Zauważono mianowicie, iż odstęp sezonowy pokrywał się z rzeczywistą, średnią długością powtarzalnej jednostki rytmicznej, natomiast wartość parametru V_e, mimo iż zawsze wyższa niż w prozie, zmieniała się w zależności od systemu wersyfikacyjnego, a nie języka. Chociaż ilość i zróżnicowanie badanych tekstów wierszowanych nie pozwalają na zbyt daleko idące uogólnienia, wszystko wskazuje na to, że analiza sekwencyjna pozwoli na stworzenie typologii tekstów wierszowanych opartej na kryterium ilościowym. Gdyby ograniczyć się do porównywania tekstów różniących się od siebie w sposób jaskrawy, tworzenie takiej typologii byłoby zbyteczne. Jednak oprócz przypadków oczywistych istnieje szeroki obszar twórczości sytuującej się „na pograniczach" typowych klasyfikacji genologicznych (proza poetycka, wiersz biały, literatura ustna okresu przedpiśmiennego etc.). Właśnie tam metoda ARIMA lub inne matematyczne narzędzie analizy sekwencyjnej może oddać wielkie usługi. Do tego wątku powrócimy w następnym rozdziale.

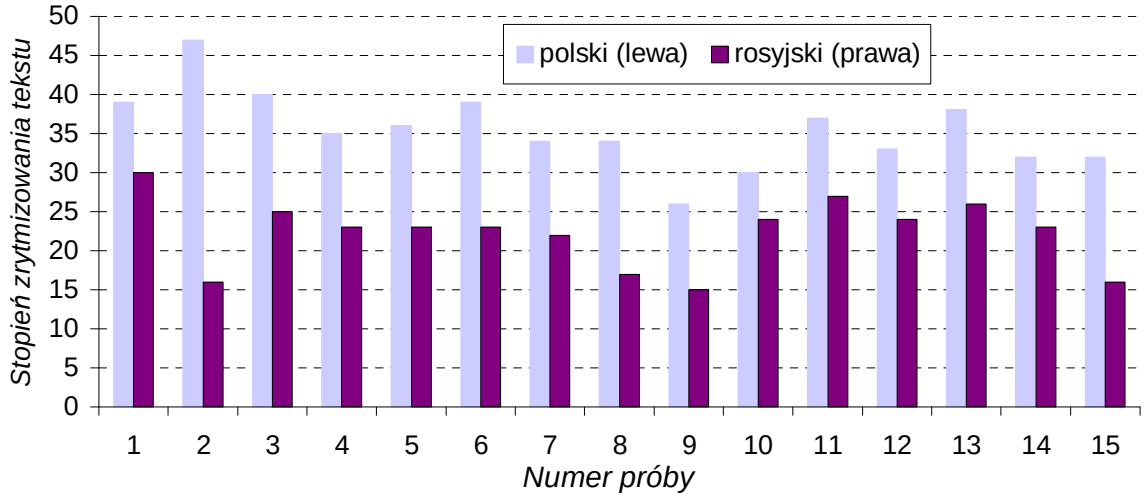
2.5 RYTMIKA TEKSTU A PRZEKŁAD

Fakt, iż analizowano równoległe fragmenty prozy artystycznej w dwóch językach, zachęcił nas do dokładniejszego porównania struktury rytmicznej oryginałów i przekładów. Wykorzystano w tym celu szczegółowe wyniki uzyskane dla tekstów równoległych, czyli prozy M. Bułhakowa i J. Iwaszkiewicza (ANEKS, Tab. 1 i 2), oraz wszystkich prób *Pana Tadeusza* i *Eugeniusza Oniegina*. Na wykresach 36 i 37 przedstawiono jako przykład wartości parametru V_e obliczone dla odpowiadających sobie fragmentów prozy w wersji oryginalnej i w przekładzie.

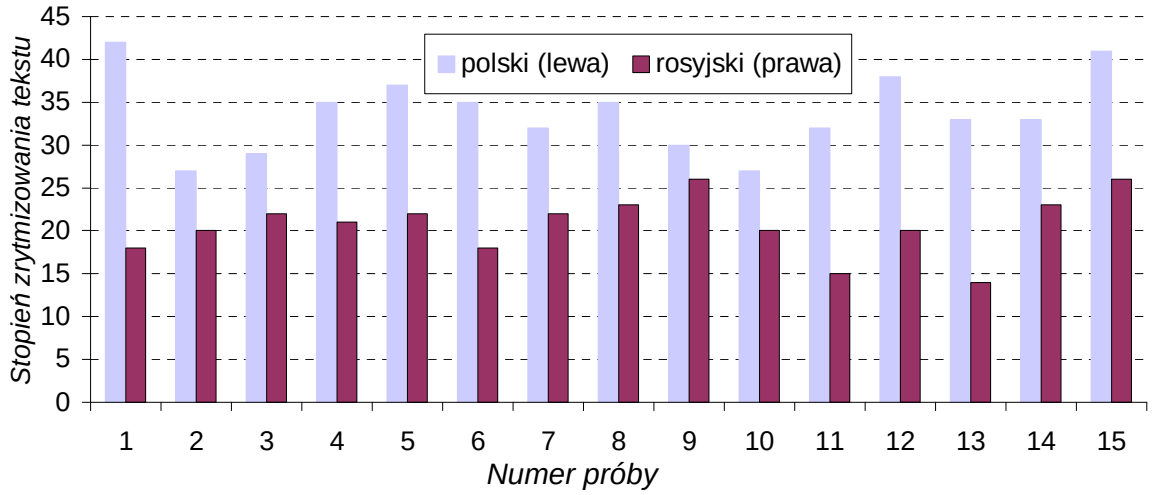
Przyglądając się obu wykresom, nie można wykluczyć słabej korelacji pomiędzy odpowiadającymi sobie fragmentami. Jako hipotezę roboczą przyjęto więc założenie, zgodnie z którym rytm prozy może być w pewien sposób powiązany z treścią każdego fragmentu i dobry tłumacz może oddać go wiernie w przekładzie. Wynik porównania prób za pomocą współczynnika korelacji liniowej Pearsona (por. GREŃ 1987:123, HAMMERL&SAMBOR 1990:93–102) nie pozostawia jednak żadnych złudzeń: dla tekstów M. Bułhakowa otrzymujemy $r_{xy} = 0{,}23$, dla tekstów J. Iwaszkiewicza $r_{xy} = 0{,}06$, dla A. Mickiewicza $r_{xy} = 0{,}19$, a w przypadku poematu A. Puszkina $r_{xy} = -0{,}01$. Nawet ma-

nipulacja danymi polegająca na pominięciu najbardziej nietypowych prób nie wpłynęła na znaczącą zmianę tego rezultatu. O statystycznej korelacji struktur rytmicznych oryginałów i przekładów nie ma więc mowy.

Rys. 36 Procent wariancji wyjaśnionej przez model dla równoległych fragmentów prozy M. Bułhakowa w językach rosyjskim i polskim[123]



Rys. 37 Procent wariancji wyjaśnionej przez model dla równoległych fragmentów prozy J. Iwaszkiewicza w językach polskim i rosyjskim[124]

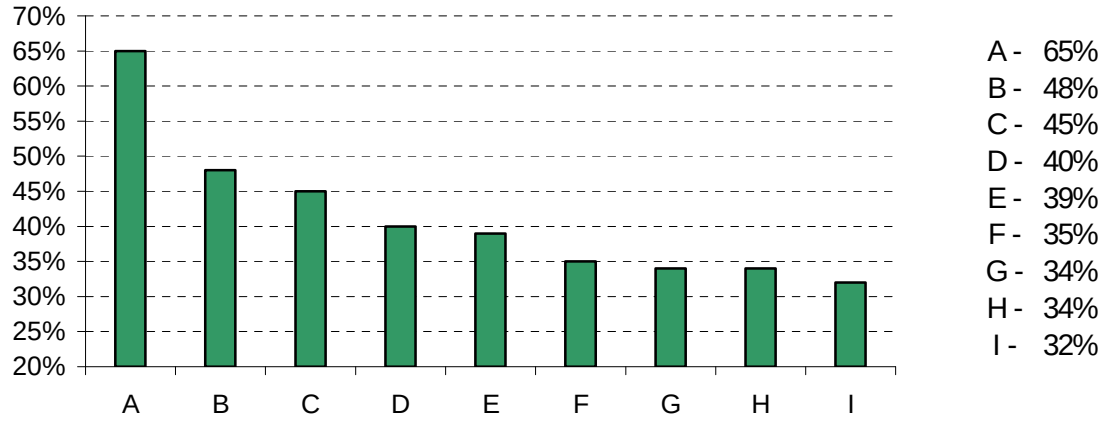


[123] ANEKS – БУЛГАКОВ 1998 oraz BUŁHAKOW 1988.
[124] ANEKS – IWASZKIEWICZ 1973 oraz ИВАШКЕВИЧ 1975.

3. STRUKTURY SEKWENCYJNE JAKO KRYTERIUM TAKSONOMII TEKSTÓW

Przegląd wszystkich wyników uzyskanych dotąd dla języka polskiego pozwala dostrzec pewną prawidłowość. Stwierdzono już, że ogólnie rozumiany styl tekstu, a także systemy metryczne charakteryzują się różnymi wartościami parametru V_e. Otóż kolejność analizowanych wcześniej tekstów polskojęzycznych uporządkowanych według parametru V_e, a więc według stopnia ich zrytmizowania, w zasadzie pokrywa się z formalnym (wynikającym z reguł) poziomem zdeterminowania ich struktury wersyfikacyjnej (Rys. 38). Najwyższe wartości uzyskano dla ośmiosylabowego wiersza sylabotonicznego dla dzieci (A). W miarę rozluźniania reguł metrycznych następuje spadek wartości V_e. Zgodnie z oczekiwaniami mniej rytmiczny jest wiersz sylabotoniczny o stałej długości wersu (B i C). W tym miejscu dość niespodziewanie pojawia się dyskurs oratorski (D). Jak widać, brak formalnej struktury metrycznej może być przez mówcę rekompensowany właściwym doborem innych środków stylistycznych. Następny w kolejności jest stroficzny wiersz sylabiczny (E), na którego niższą pozycję wpłynął zapewne brak stałej długości wszystkich wersów (strofa składa się z dwunastu wersów dziewięciozgłoskowych i dwóch ośmiozgłoskowych), a także fakt, iż mamy do czynienia ze współczesnym przekładem. Dalej są teksty prozatorskie, a najniższą wartość współczynnika V_e stwierdzono w stylu publicystyczno-prasowym.

Rys. 38 Zrytmizowanie wybranych odmian stylistycznych i wersyfikacyjnych polszczyzny



Oznaczenia:

A – wiersz sylabotoniczny 8-zgłoskowy (J. Brzechwa)
B – wiersz sylabiczny 11-zgłoskowy (J. Słowacki)
C – wiersz sylabiczny 13-zgłoskowy (A. Mickiewicz)
D – dyskurs oratorski (K. Wojtyła)
E – wiersz sylabiczny stroficzny (A. Puszkin)
F – proza artystyczna (M. Bułhakow)
G – proza artystyczna (I. Newerly)
H – proza artystyczna (J. Iwaszkiewicz)
I – styl prasowo-publicystyczny

Na diagramie 38 zwraca uwagę brak wyraźnej cezury oddzielającej pod względem rytmiki wiersz i prozę. Fakt ten można interpretować różnie. Z jednej strony można przyjąć za pewnik, iż taka cezura istnieje i jeżeli w jednoznaczny sposób nie stwierdzono jej obecności, tym gorzej dla metody. Podejście takie nie jest jednak poparte przekonującymi argumentami. Warto uzmysłowić sobie, że metoda ARIMA (oraz każda inna metoda analizy sekwencyjnej) narzuca pewne ograniczenia polegające na redukcji wielowarstwowości i wieloznaczności każdej indywidualnej interpretacji do zobiektywizowanego ciągu rytmicznego zapisanego kodem binarnym. Redukcja ta pozwala jednak na uogólnienia dające się wyrazić formalnymi i falsyfikowalnymi modelami, które wykrywają rzeczywiste (nie związane z długością wersu) rozczłonkowanie strumienia sylab o praktycznie dowolnej długości. Pozytywnym tego efektem jest możliwość porównania struktury sekwencyjnej tekstów traktowanych dotychczas jako nieporównywalne, a więc prozy, wiersza, dyskursu oratorskiego czy literatury ustnej. I chociaż przedstawienie pełnej taksonomii systemów wersyfikacyjnych konkretnego języka etnicznego opartej na metodach analizy sekwencyjnej jest tematem obszernym, wymagającym napisania osobnej dysertacji, przeprowadzone tu testy stanowią dla takiego przedsięwzięcia doskonały punkt wyjścia.

Wyniki przedstawione na diagramie 38, mimo że oparte na niepełnym materiale językowym, pozwalają już na sformułowanie pewnych uogólnień. Cezurę oddzielającą w polszczyźnie prozę i teksty wierszowane wyznaczyć można na poziomie $V_e = 35\%$. Teksty o wyrazistym układzie metrycznym sytuowałyby się natomiast powyżej wartości $V_e = 40\%$. Przejście rytmiczne pomiędzy prozą a wierszem miałoby charakter ciągły, a nie skokowy – jego płynność zapewniałyby teksty o luźnym układzie wersyfikacyjnym (system toniczny, proza rytmiczna). Wzorcowy rozkład wartości parametru V_e, oparty na zróżnicowanej i obszernej próbie, mógłby też stać się probierzem pozwalającym na ocenę jednostkowych, empirycznych wyników, które nie mieszczą się w tradycyjnych klasyfikacjach.

Warto w tym miejscu powtórzyć, że przełożenie ilościowej miary rytmiki tekstu na kategorie estetyczne jest kwestią złożoną i nie do końca rozstrzygalną. Współczynnik V_e nie może więc być uważany za wyrazistą miarę literackości czy poetyckości utworu. M.R. Mayenowa stwierdza, iż „Podział na wiersz i prozę nie jest równoznaczny z podziałem na sztukę i niesztukę.” (MAYENOWA 1979:369). Jakościową cechą stanowiącą przedmiot pomiaru jest raczej to, co autorka określa jako „wierszowość” (*passim*). Cecha ta współtworzy estetykę utworu („...ograniczenia są źródłem istotnych znalezisk, bez których sztuka jest niemożliwa. [...] poezja to właśnie wiersz, swoista organizacja harmonii brzmieniowej. [...] Gdzie nie ma tej organizacji, nie ma poezji. [...] w naturze człowieka tkwi potrzeba takiej właśnie harmonii.” – *ibid.* 372), choć w skrajnych przypadkach może się sprowadzać do „kompromitującej akrobatyki” służącej jedynie celom mnemotechnicznym (*ibid.*).

4. SEKWENCYJNA ANALIZA PROZODII ŁACIŃSKIEJ[125]

Dotychczasowe analizy empiryczne prozodii przeprowadzono na materiale języków nowożytnych mających akcent dynamiczny. Linearność jest jednak uniwersalną cechą każdego tekstu, niezależnie od tego, kiedy powstał i jaki typ prozodii realizuje. Potwierdziły to już udane analizy języków tonalnych (DREHER et al. 1969). Kolejną egzemplifikacją powyższej tezy, wynikającej z przesłanek ogólnolingwistycznych, jest przedstawiony poniżej przykład, w którym analizie poddano łacinę epoki klasycznej, kodując jako relewantne cechy długości samogłosek oraz tzw. akcent metryczny. Testy wykazały, że dzięki przejrzystej segmentacji tekstu na dyskretne jednostki, takie jak stopy metryczne czy sylaby, łacina doskonale poddaje się sekwencyjnej analizie ilościowej.

4.1 ILOCZAS W ŁACINIE – ZARYS PROBLEMATYKI

Przegląd źródeł encyklopedycznych oraz akademickich podręczników lingwistyki wskazuje, iż najczęściej cytowanym przykładem języka stosującego iloczas jest łacina. Iloczas uważa się też za podstawowy, jeśli nie jedyny środek rytmotwórczy w poezji łacińskiej. Na przykład w wielokrotnie wznawianej *The Cambridge Encyclopædia of Language* czytamy: „By contrast, the length of a syllable (whether long or short) was a crucial feature of rhythm in Latin." (CRYSTAL 1997:171). Z kolei J. Wikarjak stwierdza: „Istnieje zasadnicza różnica pomiędzy wierszem antycznym a wierszem nowożytnym. Rytm wiersza nowożytnego wiąże się ściśle z akcentem wyrazowym, rytm wiersza antycznego jest niezależny od akcentu poszczególnych wyrazów, wiąże się natomiast z iloczasem, z określonym następstwem sylab długich i krótkich." (WIKARJAK 1978:153). Prozodię iloczasową przeciwstawia się systemowi tonalnemu, występującemu głównie w językach Azji, oraz akcentowi dynamicznemu, spotykanemu w większości języków indoeuropejskich. Znaczenie iloczasu w łacinie jest najbardziej widoczne w tekstach stosujących specyficzne dla tego języka normy wersyfikacji. Ich znajomość pozwala przedstawić tekst (dziś z oczywistych względów pisany) w postaci sekwencji stóp metrycznych, zbudowanych z sylab długich i krótkich.

Ten klarowny obraz wydaje się jednak zawierać pewne rysy. Wypada przecież zapytać, dlaczego w zdecydowanej większości języków indoeuropejskich, a w szczególności w językach romańskich, wywodzących się przecież z łaciny, powszechnie występuje akcent dynamiczny, natomiast iloczasu nie stosuje się w ogóle[126]. Trudno też zgodzić się z tezą, iż niekwestionowana rola akcentu dynamicznego w łacinie ludowej była jedynie skutkiem oddziaływania innych języków i przejawem degeneracji języka klasyków, a nie miała żadnego oparcia w immanentnych cechach prozodii łacińskiej. Do zastano-

[125] Wykorzystane tu wyniki opublikowane zostały w pracy PAWŁOWSKI&EDER 2000.

[126] Opozycja głoska długa / krótka występuje w niektórych językach indoeuropejskich i może mieć nawet charakter fonologiczny (na przykład w czeskim i słowackim). Jednak powszechność występowania akcentu dynamicznego w językach nowożytnych oraz jego funkcjonalne obciążenie sprawiają, że iloczas w rodzinie indoeuropejskiej uważać należy za zjawisko peryferyjne.

wienia skłania wreszcie fakt, iż przez długie wieki łacina, tak w mowie, jak i w piśmie, była ponadnarodowym językiem elit, funkcjonującym w swoistej symbiozie z językami etnicznymi nowożytnej Europy, nie stosującymi prozodii iloczasowej. Patrząc na to zjawisko w perspektywie historycznej, warto przypomnieć, że już w pierwszych wiekach n.e. pisarze prowincjonalni nie stosowali iloczasu w sposób poprawny i konsekwentny, a od III w. n.e. jego poczucie zanikło zupełnie. Czynnikiem dominującym w wymowie łacińskiej stał się wyłącznie akcent dynamiczny (MYŚLIWIEC 1959:141). W ciągu kilkunastu następnych wieków powstawały co prawda utwory konstruowane według zasad metryki antycznej (i niejednokrotnie były to teksty wybitne), ale iloczas – podstawa wersyfikacji epoki klasycznej – pozostał dla większości Europejczyków swobodnie posługujących się językiem Wergiliusza zjawiskiem obcym, a w najlepszym wypadku wyuczonym i nienaturalnym. Argumentacja ta znajduje pewne potwierdzenie w badaniach filologicznych. Badając rozwój fonetyczny łaciny przedklasycznej zauważono, że samogłoski w pewnych pozycjach nie podlegały modyfikacji, z czego można wnioskować, iż były akcentowane dynamicznie. Jak zauważa J. Safarewicz: „Ten znany nam akcent łaciński uwzględniał, jak widać, dwa czynniki: miejsce w wyrazie (padał na sylabę przedostatnią albo trzecią od końca) i iloczas sylab (bo wybór sylaby drugiej czy trzeciej od końca zależał od iloczasu sylaby przedostatniej).” (SAFAREWICZ 1988:521). Fakt, że klasyczna wersyfikacja łacińska nie brała pod uwagę akcentuacji wyrazów, a jedynie długość sylab wyjaśnia się tym, że w jakimś momencie w łacinie nastąpił powrót do akcentu muzycznego (*ibid.*).

4.2 PROZODIA I METRYKA ŁACINY – STAN POGLĄDÓW

Iloczasowa wersyfikacja łacińska wraz ze wszystkimi odmianami metrów została w całości przejęta z poezji greckiej. Tymczasem przed rokiem 240 p.n.e., uważanym za narodziny literatury rzymskiej, istniały prawdopodobnie dość liczne teksty pisane w rodzimym metrum italskim, tzw. wierszem saturnijskim (*versus Saturnius*). Wśród badaczy panują sprzeczne poglądy na ten temat – jedni uważają go za poezję akcentuacyjną, inni za iloczasową (LEO 1905). Po 240 r. p.n.e. panuje już wyłącznie wersyfikacja iloczasowa, której podstawą jest swoiste uszeregowanie sylab długich i krótkich tworzących stopy metryczne, przy czym stopy nie muszą pokrywać się z granicami jednostek leksykalnych.

Niektóre wiersze (*versus Alcaius enneasyllabus, versus Asclepiadeus minor, metrum Hipponecteum*) mają stałą ilość sylab w wersie, inne (na przykład trymetr jambiczny czy heksametr daktyliczny) dopuszczają pewną swobodę w zamianach stóp i tym samym nie determinują ilości sylab w wersie, pod warunkiem zachowania takiej samej całkowitej długości stopy[127]: daktyl (– ∪∪) można zastąpić spondejem (– –), jamb (∪ –) trybrachem (∪∪∪) itp. W heksametrze, stanowiącym materiał poniższej analizy, klasyczny układ sześciu daktyli (ostatnia sylaba w wersie może mieć dowolną długość):

[127] Sylabę długą oznaczamy symbolem –, a krótką symbolem ∪.

sive vetebat, an hoc inhonest(um) et inutile factu (Hor. Serm. I 4,124)
– ∪∪| – ∪∪| – ∪∪| – ∪∪| – ∪∪| – ∪

może zostać zastąpiony nawet przez sześć spondejów, choć jest to dość rzadki przypadek:

olli respondit rex Albai Longai (Enn. Ann. 169)
– –| – –| – –| – –| – –| – –

Daje to 32 możliwości praktycznej realizacji heksametru, na przykład:

dissuadere licet: non est tua tota voluntas! (Ov. *Met.* II 53)
– –| – ∪ ∪ | – –| – ∪ ∪ | – ∪ ∪ | – –

Osobnym zagadnieniem jest kwestia akcentu dynamicznego. Wśród badaczy panuje dość zgodna opinia, że rytmu wiersza nie można wyczerpująco opisać za pomocą jednego tylko elementu, na przykład opozycji sylab długich i krótkich. Jak zauważa G.L. Hendrickson: „Without the moulding power of rhythmic movement a purely quantitative rhythm cannot be sustained in language" (HENDRICKSON 1899:209). Podstawę do tych sądów dali już starożytni metrycy greccy, którzy wyróżniali w stopie część mocną (*thesis*) i słabą (*arsis*). Terminy te odnosiły się do ruchów tanecznych: *teza* oznaczała opuszczenie stopy tancerza (część mocniejsza), *arsa* jej podniesienie (część słabsza). Rzymscy teoretycy odwrócili znaczenie obu terminów, dziś więc *arsa* to część mocna, a *teza* – słaba (oba terminy stosowane są dalej w tym właśnie znaczeniu). Niektórzy badacze zaprzeczają jednak, jakoby część mocna miała oznaczać akcent dynamiczny, argumentując, że po pierwsze, iloczasowa poezja indyjska nie posiadała przycisku akcentowego, a jedynie akcent słowny – a jak się przypuszcza, heksametr grecki, będący pierwowzorem dla łacińskiego, jest również adaptacją obcego, być może indyjskiego, wzoru (NAGY 1974). Po drugie, rytm w muzyce można uzyskać również na instrumentach, które nie mają możliwości stosowania akcentu dynamicznego (przykładem są organy). By wyjaśnić istotę rytmu wiersza antycznego, badacze ci, rezygnując z binarnej opozycji sylab długich i krótkich, różnicują długość sylaby na kilka wartości czasowych, na przykład sześć (POSTAL 1968) lub siedem (WEST 1970).

Z kolei zwolennicy akcentu dynamicznego wprowadzają pojęcie *iktu* (*ictus*). Przez ikt rozumieją oni akcent metryczny, różny od akcentu słownego, padający zawsze na mocną część stopy (arsę). Na poparcie tezy o istnieniu iktu przedstawiają kilka argumentów. Po pierwsze, dłuższy szereg metryczny złożony z samych krótkich lub długich sylab – na przykład wspomniany przykładowy typ heksametru – – | – – | – – | – – | – – | – – , staje się bez iktu zupełnie arytmiczny, w przeciwieństwie do tego samego, ale iktowanego szeregu ⊥ – | ⊥ – | ⊥ – | ⊥ – | ⊥ – | ⊥ – . Po drugie, pewne rozwiązania i ściągnięcia stóp byłyby bez iktu nierozróżnialne, na przykład anapest (∪∪⊥) rozwiązany w daktyl (– ∪/∪) byłby identyczny z daktylem właściwym (⊥∪∪). Po trzecie, wzdłużenia u Homera

są o wiele częstsze w arsie niż w tezie, co pośrednio dowodzi istnienia iktu. Po czwarte wreszcie – trudno wyobrazić sobie pozbawione iktu pieśni, towarzyszące na przykład tańcowi (SĄDEJOWA 1959:13–14).

Trzeba jednak wyraźnie zaznaczyć, że akcent metryczny nie jest odpowiednikiem paroksytonicznego lub proparoksytonicznego akcentu wyrazowego (muzycznego lub dynamicznego). W wielu metrach wręcz celowo podkreśla się tę rozbieżność: na przykład w heksametrze dąży się do rozsunięcia obu rodzajów akcentów w początkowej części wersu (mniej więcej do głównej cezury), podczas gdy w części końcowej, oba akcenty zazwyczaj się pokrywają. Jest to widoczne w poniższym wersie, gdzie wytłuszczonym drukiem zaznaczono akcent metryczny (ikt), a podkreśleniem akcent słowny:

abs<u>e</u>nt**e**s pro s**<u>e</u>** || m<u>e</u>mor**i** r<u>o</u>gat **<u>o</u>**re sal**<u>u</u>**tent (Ov. *Met.* VI 508)

Oczywistych i jednoznacznych dowodów na istnienie iktu nie ma ani u teoretyków antycznych, ani u późniejszych badaczy, jednak argumenty jego zwolenników są dość przekonujące, tym bardziej, że jakiś rodzaj akcentu dynamicznego istnieje we wszystkich formach wersyfikacyjnych języków europejskich, z pewnością również w łacinie ludowej. W dalszej części pracy termin „akcent” używany będzie w znaczeniu „akcentu metrycznego” (iktu).

4.3 HIPOTEZA BADAWCZA

Opierając się na powyższych argumentach, przyjęto, iż rytm tekstu w łacinie klasycznej mogły wyznaczać zarówno iloczas, jak i akcent metryczny o charakterze dynamicznym. Postawiono następnie pytanie o wzajemną relację tych dwóch sekwencyjnych porządków. Z uwagi na wyjątkową nośność kulturową wersyfikacji iloczasowej i panujący w tej kwestii stan poglądów należałoby się spodziewać, że sekwencje sylab kodowanych według długości sylab będą przedstawiać wysoki stopień uporządkowania, szczególnie w tekstach artystycznych epoki klasycznej. Z drugiej jednak strony, wyłożone wcześniej argumenty przemawiałyby za niebagatelną rolą akcentuacji w procesie generowania rytmu tekstu. Wychodząc od tych przesłanek, najbardziej prawdopodobną hipotezą jest stwierdzenie, iż oba porządki – zarówno ten, który tworzą sylaby długie i krótkie, jak i ten, który wyznaczają sylaby akcentowane i nie akcentowane – będą generować rytm tekstu, jednak bardziej wyrazisty będzie rytm iloczasowy.

4.4 BADANY KORPUS I KWANTYFIKACJA TEKSTU

W celu weryfikacji wysuniętej hipotezy zakodowano teksty Horacego (20 próbek), Owidiusza (10 próbek) i Wergiliusza (20 próbek), stosujące strukturę metryczną heksametru (szczegółowe dane bibliograficzne podano w ANEKSIE). Przeciętna długość próby wynosiła 150 sylab. Taki dobór materiału uzasadniony jest treścią weryfikowanej hipotezy, przy wszystkich swych ograniczeniach heksametr pozostawia bowiem autorowi pewien stopień swobody w kształtowaniu rytmicznej struktury wiersza. Struktura ta nie

jest więc całkowicie przewidywalna, tak jak na przykład w wierszu sylabotonicznym, nie jest też, wzorem prozy, pozbawiona formalnych wyznaczników rytmu.

Kwantyfikacja tekstu miała na celu wygenerowanie z każdej próby dwóch odpowiadających sobie sekwencji: iloczasowej i akcentowej. Podobnie jak w poprzednich analizach, zastosowano skalę liczbową typu porządkowego, przypisując liczbę 1 sylabom wyróżnionym (długim bądź akcentowanym), oraz liczbę 0 sylabom nie wyróżnionym (krótkim bądź nie akcentowanym).

Metodę kwantyfikacji ilustruje przykład dwuwiersza (Verg. *Aen*. III 233–234):

haud secus ac iussi faciunt tectosque per herbam
disponunt ensis et scuta latentia condunt.

który można przedstawić jako sekwencję stóp metrycznych (oznaczenia: d – daktyl, s – spondej, c – daktyl katalektyczny):

d s d s d c
s s s d d s

lub sylab długich i krótkich:

– ⏑ ⏑ – – – ⏑ ⏑ – – – ⏑ ⏑ – ⏑
– – – – – – – ⏑ ⏑ – ⏑ ⏑ – –

lub sylab akcentowanych i nie akcentowanych:

⊥ ⏑ ⏑ ⊥ – ⊥ ⏑ ⏑ ⊥ – ⊥ ⏑ ⏑ ⊥ ⏑
⊥ – ⊥ – ⊥ – ⊥ ⏑ ⏑ ⊥ ⏑ ⏑ ⊥ –

4.5 Przykład analizy szczegółowej

Zastosowaną na całym korpusie procedurę badawczą ilustruje szczegółowa analiza wybranego losowo fragmentu *Eneidy* (Verg. *Aen.* III 229–238). Jej pierwszym etapem jest segmentacja tekstu na stopy metryczne oraz przypisanie sylabom (samogłoskom) długości i/lub akcentów według wcześniej opisanych zasad.

Rurs(um) in secessu longo sub rupe cavata,
⊥ – / ⊥ – / ⊥ || – / ⊥ – / ⊥ ⏑ ⏑ / ⊥ –
arboribus clausi circ(um) atqu(e) horrentibus umbris,
⊥ ⏑ ⏑ / ⊥ – / ⊥ || – / ⊥ – / ⊥ ⏑ ⏑ / ⊥ –
instruimus mensas arisque reponimus ignem:
⊥ ⏑ ⏑ / ⊥ – / ⊥ || – / ⊥ ⏑ ⏑ / ⊥ ⏑ ⏑ / ⊥ ⏑
rurs(um) ex diverso caeli caecisque latebris
⊥ – / ⊥ – / ⊥ || – / ⊥ – / ⊥ ⏑ ⏑ / ⊥ –
turba sonans praedam pedibus circumvolat uncis,
⊥ ⏑ ⏑ / ⊥ – / ⊥ || ⏑ ⏑ / ⊥ – / ⊥ ⏑ ⏑ / ⊥ –

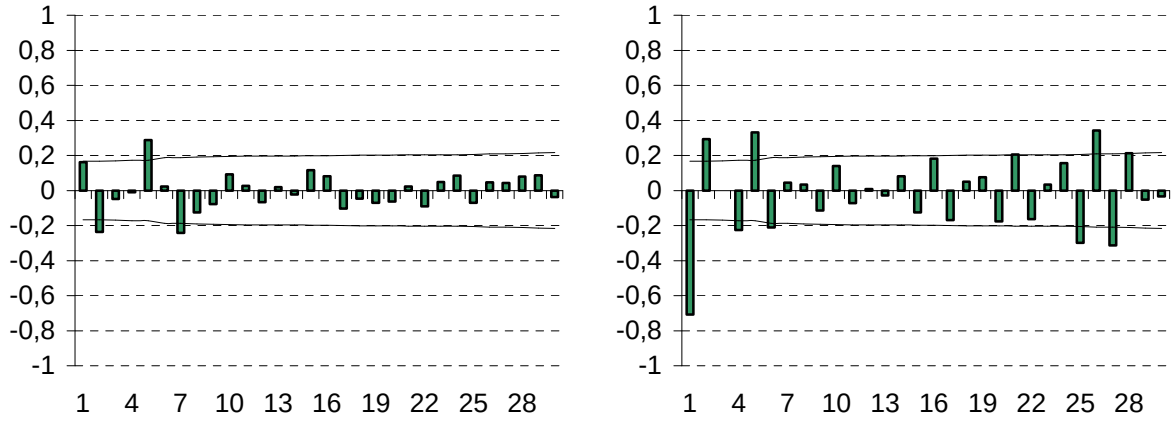
polluit ore dapes. Sociis tunc, arma capessant,
⊥ ∪ ∪ / ⊥ ∪ ∪ / ⊥ || ∪ ∪ / ⊥ – / ⊥ ∪ ∪ / ⊥ –
edic(o), et dira bellum cum gente gerendum.
⊥ – / ⊥ – / ⊥ || – / ⊥ – / ⊥ ∪ ∪ / ⊥ ∪
haud secus ac iussi faciunt tectosque per herbam
⊥ ∪ ∪ / ⊥ – / ⊥ || ∪ ∪ / ⊥ – / ⊥ ∪ ∪ / ⊥ ∪
disponunt ensis et scuta latentia condunt.
⊥ – / ⊥ – / ⊥ || – / ⊥ ∪ ∪ / ⊥ ∪ ∪ / ⊥ –
erg(o) ubi delapsae sonitum per curva dedere
⊥ ∪∪ / ⊥ – / ⊥ || ∪∪ / ⊥ – / ⊥ ∪∪ / ⊥ ∪

Sekwencja powyższa, zapisana w postaci iloczasów i akcentów, miałaby postać:

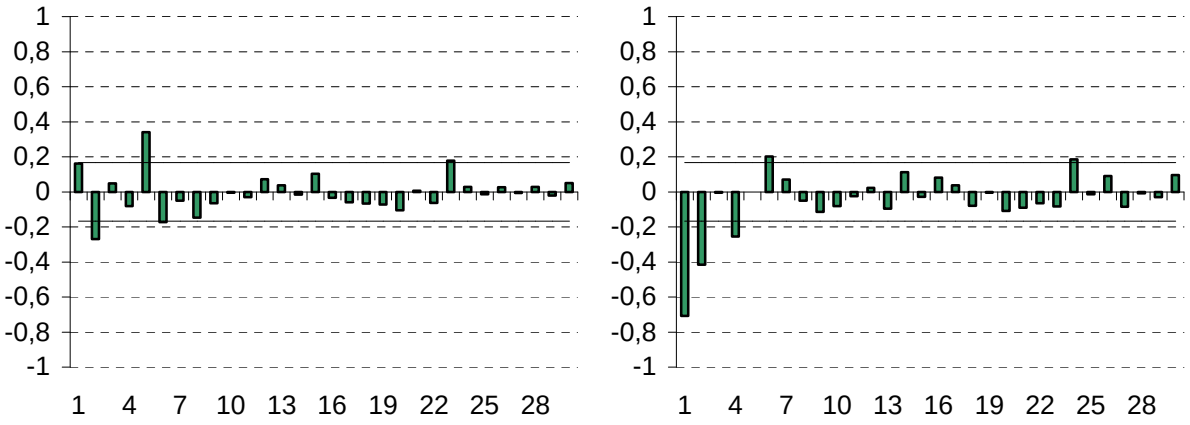
Iloczas	Akcent metryczny
1111111110011	1010101010010
10011111110011	10010101010010
100111110010010	100101010010010
1111111110011	1010101010010
100111001110011	100101001010010
1001001001110011	1001001001010010
1111111110010	1010101010010
100111001110010	100101001010010
11111110010011	10101010010010
100111001110010	100101001010010

Uzyskane tym sposobem dwa binarne szeregi czasowe poddano następnie analizie metodą ARIMA. Wykres 39 przedstawia funkcję autokorelacji (ACF), natomiast wykres 40 funkcję autokorelacji cząstkowej (PACF), obliczone dla szeregów iloczasowego i akcentowego.

Rys. 39 Funkcja autokorelacji dla łacińskiego heksametru kodowanego jako sekwencja iloczasowa (wykres lewy) i akcentowa (wykres prawy)



Rys. 40 Funkcja autokorelacji cząstkowej dla łacińskiego heksametru kodowanego jako sekwencja iloczasowa (wykres lewy) i akcentowa (wykres prawy)



Wbrew oczekiwaniom, postać funkcji ACF i PACF dla obu szeregów sugeruje w oczywisty sposób, iż to sekwencja akcentowa, a nie iloczasowa generuje najsilniejszy rytm tekstu. Wartość ACF dla kroku 1 przy sekwencji akcentowej (Rys. 39) jest bardzo wysoka: $r_1 = -0{,}71$ przy tym, że za statystycznie znaczące uważa się wartości spoza przedziału [–0,17, 0,17]. Kolejne wartości r_i układają się w dwie gasnące sinusoidy. W przypadku szeregu iloczasowego układ prążków funkcji ACF i PACF jest bardziej chaotyczny. Co prawda niektóre wartości obu funkcji są statystycznie znaczące (dla odstępów 2 i 5), jednak wyraźnie niższe od analogicznych wartości obliczonych dla szeregu akcentowego. Co więcej, trudno doszukać się tu wyrazistego kształtu funkcji ACF lub PACF, sugerującego konkretny model procesu stochastycznego. Także w przypadku sekwencji akcentowej, gdzie wartości odnośnych funkcji są wyższe, wybór modelu nie jest sprawą łatwą (por. Tab. 14). Tak naprawdę trudno bowiem orzec, która funkcja wygasa, a która raptownie się urywa. Sytuację komplikuje dodatkowo podejrzenie, iż sekwencja akcentowa może zawierać składową sezonową – przy odstępach 25–27 pojawiają się bowiem znaczące prążki funkcji ACF (Rys. 39).

Po przeprowadzeniu serii testów jako składowe proste dla sekwencji iloczasowej przyjęto AR(5) lub MA(2), a dla sekwencji akcentowej AR(2) lub AR(4). Uwzględniono też znaczące wartości funkcji ACF sugerujące sezonowość sekwencji akcentowej (w kategoriach lingwistycznych oznaczałoby to istnienie wewnętrznego rozczłonkowania tekstu na ekwiwalentne pod względem rytmicznym odcinki o długości około 26 sylab). Oprócz procesów prostych, dla sekwencji akcentowej estymowano więc także modele sezonowe (Tab. 26). Podobnie jak w poprzednich przypadkach, miarą jakości dopasowania modelu do danych jest procent wyjaśnionej zmienności szeregu obserwowanego (V_e).

Wyniki analizy pokazały, że sezonowość zaobserwowana w sekwencji akcentowej jest bardzo słaba i nie polepsza stopnia dopasowania modelu (Tab. 26). Testy przeprowadzone na pozostałych próbach potwierdziły te spostrzeżenia: 1) korelację o charakterze sezonowym stwierdzono jedynie w około 60% próbek; 2) nigdzie nie pojawiły się stałe

odstępy sezonowe, a jedynie przedziały wartości (odstęp od 20 do 30); 3) wartości sezonowe sytuowały się na pograniczu przyjętego przedziału ufności (tzw. wstęgi Bartletta). Uwzględniając te argumenty oraz treść testowanej hipotezy, zrezygnowano w tym przypadku z estymacji modeli sezonowych. Poczynione obserwacje sugerują natomiast celowość prowadzenia bardziej szczegółowych analiz o podłożu stylometrycznym i filologicznym, które pozwoliłyby wyjaśnić zauważoną regularność. Zachodzi bowiem zbieżność (być może przypadkowa) pomiędzy długością odstępu sezonowego (od 20 do 30 sylab), a długością wersu heksametru wyrażoną w morach (24 mory).

Tab. 26 Identyfikacja modelu rytmiki heksametru łacińskiego kodowanego jako sekwencja akcentowa

Typ modelu ($s_0^2 = 0{,}244$)	s_r^2	V_e	N
AR(1)	0,122	50%	1
AR(2)	0,102	58%	2
AR(4)	0,096	61%	4
ARMA(1,1)	0,103	58%	2
SARMA(2,0)(1,0)$_{25}$	0,102	58%	3
SARMA(2,0)(1,0)$_{26}$	0,102	58%	3
SARMA(2,0)(1,0)$_{27}$	0,101	59%	3

Oznaczenia:

N – liczba parametrów modelu
s_0^2 – wariancja szeregu obserwowanego
s_r^2 – wariancja szeregu resztowego
V_e – procent wariancji wyjaśniony przez model (por. wzór 70)

Jak już wspomniano, kryterium jakości dopasowania modelu do danych empirycznych jest procent wyjaśnionej wariancji szeregu obserwowanego (V_e). Wartości V_e dla estymowanych modeli prostych zawiera tabela 27. Potwierdzają one domysły oparte na obserwacji wykresów 39–40. Szereg oparty na iloczasie zawiera co prawda składową deterministyczną, ale wyjaśnia ona zaledwie kilkanaście procent całkowitej zmienności obserwowanej sekwencji. Zupełnie inny wynik otrzymujemy przy analizie szeregu akcentowego: proponowane modele procesów stochastycznych wyjaśniają do 61% wariancji szeregu obserwowanego, co dowodzi, iż ze statystycznego punktu widzenia sekwencja taka jest zdecydowanie bardziej rytmiczna i przewidywalna. Jednak na ostateczny wybór modelu, oprócz procentu wyjaśnionej wariancji, wpływ ma także liczba jego parametrów. Przyjmuje się, że model prostszy jest „oszczędniejszy” (ang. *parsimonious*) i przez to lepszy. Z tego względu, w omawianym przypadku jako optymalne wybrano modele MA(2) dla sekwencji iloczasowej i AR(2) dla sekwencji akcentowej.

Tab. 27 Rytm heksametru łacińskiego kodowanego jako sekwencja iloczasowa i akcentowa

	sekwencja iloczasowa		sekwencja akcentowa	
typ modelu	AR(5)	MA(2)	AR(2)	AR(4)
wartość V_e	19%	15%	58%	61%

Model ruchomej średniej MA(2) dla analizowanej sekwencji iloczasowej miałby postać:

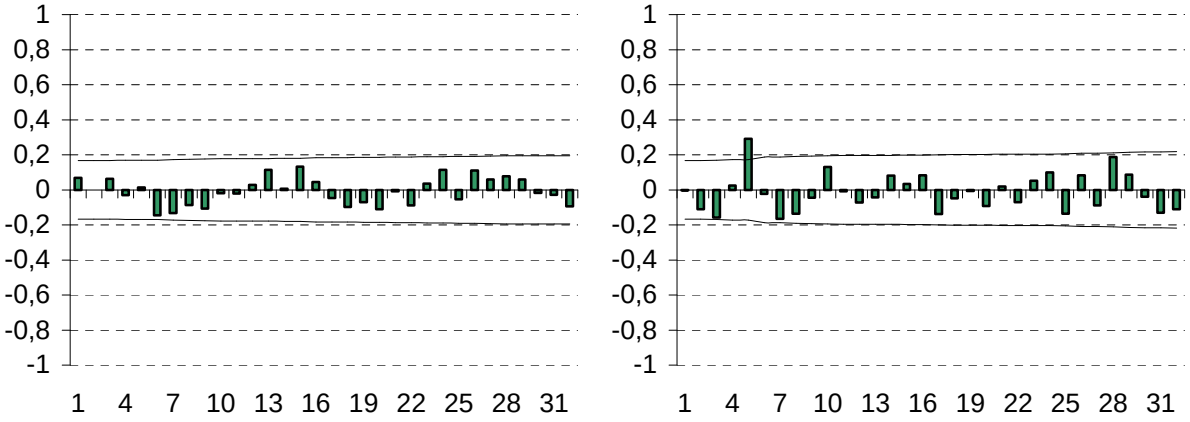
(89) $$x_t = e_t + 0{,}34e_{t-1} - 0{,}37e_{t-2}$$

natomiast model autoregresji AR(2) dla sekwencji akcentowej miałby postać:

(90) $$x_t = -x_{t-1} + 0{,}47x_{t-2} + e_t$$

W obu przypadkach x_t oznacza wartość szeregu odpowiadającą chwili lub pozycji *t*, natomiast e_t jest wartością szumu o rozkładzie N(0,1), także odpowiadającą chwili lub pozycji *t*. Jakość dopasowania obu modeli do danych potwierdza autokorelacja szeregów resztowych (Rys. 41). Co prawda z sekwencji akcentowej nie odfiltrowano znaczącej wartości dla odstępu piątego, jednak porównanie wielu prób wskazuje, że cecha ta nie występuje w sposób systematyczny i nie można uważać jej za istotną z punktu widzenia celu analizy.

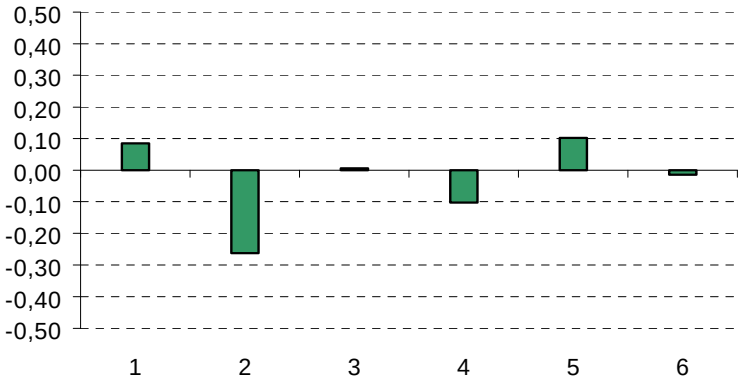
Rys. 41 Autokorelacja szeregów resztowych dla modeli MA(2) (szereg iloczasowy, wykres lewy) i AR(2) (szereg akcentowy, wykres prawy)



Analizując próbki sekwencji iloczasowych, stwierdzono też, że trudno jest wskazać jeden, powtarzający się i wyrazisty wzorzec funkcji autokorelacji. Z tego względu uznano za celowe obliczyć średnie wartości ACF dla pierwszych sześciu odstępów (Rys. 42). Za statystycznie znaczącą można uznać jedynie autokorelację przy odstępie drugim, natomiast przy odstępach trzecim i szóstym charakterystyczna jest zerowa wartość współ-

czynników, co dowodzi całkowitego zerwania związku kontekstowego pomiędzy sylabami na pozycjach t i $t \pm 3$ oraz t i $t \pm 6$. Wskazuje to na stosunkowo słaby poziom liniowego uporządkowania sylab długich i krótkich w heksametrze klasycznym i potwierdza raz jeszcze, że rytmotwórcza funkcja iloczasu jest, przynajmniej w tym przypadku, wątpliwa. Dodajmy, że stworzenie podobnego uśrednionego wykresu dla autokorelacji szeregu akcentowego nie było konieczne, ponieważ kształt ACF dla początkowych odstępów, zaprezentowany wcześniej (Rys. 39), powtarzał się dość regularnie we wszystkich próbach.

Rys. 42 Uśrednione wartości autokorelacji dla sekwencji iloczasowej



Ważnym momentem każdej analizy ilościowej danych tekstowych jest nadanie matematycznym modelom przekonującej lingwistycznej interpretacji. Lingwistycznym odpowiednikiem parametru x_t jest oczywiście cecha sylaby (0 – krótka bądź nie akcentowana, 1 – długa bądź akcentowana). Współczynniki modeli (88 i 89) wskazują na siłę związku pomiędzy cechą danego elementu, a cechami elementów poprzedzających. Podobnie można interpretować wartości funkcji autokorelacji. Na przykład niski współczynnik stojący przy x_{t-1} (w modelu 89 $a_1 = -1$) oznacza bardzo silną negatywną korelację i wskazuje, iż najczęściej sylaba akcentowana będzie wymuszać następstwo sylaby nie akcentowanej i na odwrót. Wyraźniejszy prążek funkcji ACF dla szeregu iloczasowego przy odstępie piątym ($r_5 = 0{,}288$) oznacza istnienie słabej pozytywnej korelacji pomiędzy cechami sylab t i $t \pm 5$ w linii tekstu. Istotnie, szereg iloczasowy zawiera dłuższe sekwencje sylab długich rozdzielane dwu- lub jednosylabowymi sekwencjami sylab krótkich.

4.6 Wyniki sumaryczne

Szczegółowa analiza poszczególnych próbek uwzględniająca treść testowanej hipotezy doprowadziła do zaskakujących uogólnień. Podobnie jak w dotychczas prowadzonych analizach, jako zmienną decyzyjną potraktowano współczynnik V_e syntetyzujący informację na temat sekwencyjnego porządku badanych szeregów. Dla każdej próby obliczono jego wartość, wybierając uprzednio najodpowiedniejszy model. Przy sekwen-

cjach iloczasowych był to model MA(2), natomiast przy sekwencjach akcentowych 32% prób opisano modelem AR(2), a pozostałe 68% modelem AR(4). Otrzymany po uśrednieniu rezultat nie pozwolił na utrzymanie testowanej hipotezy, zgodnie z którą rytmotwórcza funkcja iloczasu i akcentu metrycznego powinna być wysoka, a przy tym porównywalna: dla szeregów opartych na iloczasie obliczono $V_e = 15\%$, natomiast dla szeregów opartych na akcencie dynamicznym $V_e = 61\%$. Jak widać, różnica jest bardzo wyraźna. Okazało się więc, że wbrew przyjętym poglądom to nie długości sylab, ale powracające regularnie akcentowane części kolejnych stóp metrycznych tworzą rytmiczną tkankę łacińskiego heksametru. Jak z powyższego wynika, teksty łacińskie czytane z pominięciem iloczasu, a z uwzględnieniem akcentu metrycznego, brzmieć będą bardziej rytmicznie niż te same teksty czytane z uwzględnieniem tylko iloczasu lub iloczasu i akcentu równocześnie. Na tym etapie badań kończą się jednak możliwości analizy formalnej, opartej w mniejszym lub większym stopniu na empirii. Kwestia, czy „rytmiczniej" znaczy „ładniej", „przyjemniej" bądź „lepiej", jest bowiem na gruncie naukowym nierozstrzygalna.

4.7 DYSKUSJA

Dzięki przeprowadzonym testom wykazano, że funkcję rytmotwórczą w łacińskim heksametrze pełni akcent metryczny (ikt), a nie iloczas. Podczas gdy przeciętna wartość współczynnika V_e dla sekwencji iloczasowych jest bardzo niska, nieporównywalna nawet z analogiczną wartością obliczoną dla prozy w polszczyźnie, jego wartość dla sekwencji akcentowych jest wysoka i odpowiada rytmice prostego wiersza sylabotonicznego (por. Rys. 38).

Jednak fakt, iż uzyskany rezultat pozostaje w pewnej sprzeczności wobec głoszonej i akceptowanej od stuleci teorii, zmusza do refleksji i bardzo ostrożnej interpretacji. Wyrażenie „pewna sprzeczność" nie jest tu zresztą czczym eufemizmem. Nie sposób bowiem zaprzeczyć, że klasyczna wersyfikacja łacińska oparta jest właśnie na stosunkach *stricte* metrycznych, a więc na iloczasowej ekwiwalencji wersów. Z drugiej jednak strony, także ukazana tu dysproporcja wartości V_e jest empirycznie stwierdzonym faktem i nie może zostać zwyczajnie zignorowana.

Otóż, jak się wydaje, iloczas łaciński należy traktować jako naddaną kulturowo formę, ponad wszelką wątpliwość organizującą wersyfikację i mającą silny związek z muzycznym kontekstem odtwarzania tekstów w okresie antyku. Jednak struktura sekwencyjna oparta na metrum iloczasowym nie pokrywa się z naturalnymi rytmicznymi wzorcami („naturalność" oznacza tu pierwotny związek rytmiki języka z regularnym rytmem oddechów), wspomagającymi generowanie, rozumienie i zapamiętywanie tekstu w dowolnym kontekście komunikacyjnym[128]. Jak zauważa T. Milewski: „Mowa realizuje się na wydechu, który rytmicznie wzmaga się i słabnie, dla ekonomicznego więc jego wykorzystania w toku mowy dźwięki mające większą doniosłość akustyczną

[128] Właśnie rytm jest jednym z mechanizmów wykorzystywanych często w technikach mnemotechnicznych.

przeplatają się z dźwiękami mającymi doniosłość mniejszą, wskutek czego ciąg mowy przedstawia się pod względem doniosłości akustycznej jako linia falista, odpowiadająca w pewnej mierze falistej linii natężenia wydechu." (MILEWSKI 1965:217). We wcześniejszych rozdziałach wykazano obecność takich wzorców w prozodii języków nowożytnych i wiele wskazuje na to, iż uważać je można za uniwersalną cechę języka. Przy całej swej przewrotności, otrzymany wynik dowodzi więc, iż łacina nie była wyjątkiem od tej reguły.

5. SEKWENCYJNE MODELOWANIE TEKSTU NA POZIOMIE LEKSEMÓW I ZDAŃ

Badane dotychczas struktury sekwencyjne charakteryzowały organizację rytmiczną tekstu na poziomie prozodii, gdzie jednostką podlegającą kwantyfikacji jest sylaba. Struktury takie występują jednak także na innych poziomach języka. Przedmiotem zainteresowania lingwistów były między innymi teksty analizowane jako sekwencje zdań lub leksemów. W sekwencjach tych stwierdzono obecność procesów stochastycznych wskazujących na związek pewnych cech następujących po sobie elementów. Poniżej przedstawiamy omówienie dotychczasowych badań nad modelowaniem sekwencji zdaniowych oraz testujemy pewną hipotezę wskazującą na znaczenie analogicznych struktur na poziomie leksykalnym.

5.1 SEKWENCJE ZDANIOWE

Jednym z zagadnień cieszących się od lat niezmiennym zainteresowaniem lingwistyki kwantytatywnej jest rozkład długości zdań w tekście. Rozkład ten najczęściej bada się traktując tekst jak typową populację statystyczną, a więc bez uwzględniania porządku, w jakim zdania wystąpiły (por. KÖHLER 1995). Prace dotyczące rozkładu w znaczeniu *kolejności* zdań w linii tekstu nie są może tak liczne, ale wystarczająco pogłębione, by wskazać, iż sekwencje tego rodzaju spotykane w niektórych tekstach nie są losowe w sensie statystycznym i dają się opisać modelami matematycznymi. Hipoteza ta wydaje się bardzo prawdopodobna, zważywszy, że teksty w języku naturalnym tworzą struktury kompozycyjne podporządkowane w większym lub mniejszym stopniu wymogom retoryki i zorientowane na realizację funkcji komunikacyjnej, perswazyjnej, estetycznej itd. W pierwszej kolejności wymóg ten odnosi się oczywiście do planu treści, jednak, jak pokazują testy, widoczny jest także w planie powierzchniowym w postaci zależności zachodzących pomiędzy długościami kolejnych zdań.

Badania tekstów w językach francuskim i angielskim dowiodły, iż szeregi długości zdań otrzymane z prozy artystycznej są realizacjami procesów stochastycznych opisywanych modelami autoregresji AR(p) oraz mieszanymi ARMA(p, q) (PAWŁOWSKI 1998:124–153). Analiza przeszło 250 próbek z języka francuskiego (por. Część I, 5.4) pozwoliła ustalić, że zróżnicowanie występujących w tekstach modeli jest duże, tym bardziej, że jak na to wskazują wcześniejsze analizy, określenie typu modelu nie zawsze

jest jednoznaczne. Stwierdzono, że 18,4% prób opisywał model AR(1), 11,2% model AR(2), 2,5% model AR(3), 29,7% różne modele mieszane typu ARMA, natomiast aż 38,1% fragmentów stanowiły sekwencje zupełnie nieskorelowane (por. Tab. 11). Ten zróżnicowany obraz potwierdziła analiza około czterdziestu próbek języka angielskiego (*ibid.*), a także cytowane rezultaty A. Robertsa (ROBERTS 1996) i innych badaczy (OPPENHEIM 1988:41–42, SCHILS&DE HAAN 1993:40), którym nie udało się w pełni potwierdzić proponowanych hipotez (por. też Cześć I, 5.4).

Wyniki te pod wieloma względami odbiegają od dotychczas uzyskiwanych. Niewątpliwie wartościowe jest pokazanie, że długości zdań w niektórych tekstach można w przybliżeniu wyrazić funkcjami długości pewnej liczby zdań poprzedzających. Dowodzi to skuteczności analizy sekwencyjnej w badaniach dyskursu i otwiera nowe perspektywy badawcze. Oto przykład: analizując cytowane wyżej wyniki, przeprowadzono test pozwalający na odkrycie źródła obserwowanych regularności. Niezły wynik dało usunięcie z tekstów dialogów, w których dominują zdania krótkie i pozostawienie jedynie sekwencji opisowych. Autokorelacja otrzymanych tym sposobem szeregów była we wszystkich przypadkach wyraźnie niższa, co dowodzi, iż rozkład partii dialogowych w tekście nie był statystycznie losowy (PAWŁOWSKI 1998:136). Można oczywiście przeprowadzić więcej podobnych testów polegających na manipulacjach dostępnym materiałem tekstowym. Jednak z punktu widzenia lingwistyki modelowej, poszukującej prawidłowości ogólnych, ich celowość wydaje się wątpliwa. Warto bowiem zastanowić się, czy w kontekście tak zróżnicowanych wyników jakiekolwiek hipotezy mogą prowadzić do praw opisujących i wyjaśniających regularności rozkładu długości zdań w tekstach. Warunkiem zaproponowania takiej hipotezy powinno być stwierdzone empirycznie występowanie określonego typu modelu, pozwalające przewidzieć, oczywiście w pewnych granicach, cechy dowolnie wybranej próbki tekstu. Otóż w badanym i, dodajmy, wyjątkowo obszernym materiale, nic podobnego nie zachodzi. Co więcej, okazało się, że wartości współczynników otrzymanych modeli tego samego typu miały dużą dyspersję. Na przykład współczynniki a_1 modelu AR(1) (18,4% prób) sytuowały się w przedziale od 0,2 do 0,5.

Fakt, iż nie stwierdzono w tekstach obecności jednego, typowego modelu opisującego sekwencje długości zdań, jasno sugeruje, że nie można w tym przypadku wysuwać hipotez pretendujących do statusu praw językowych, lecz należy koncentrować się na opisie zjawisk jednostkowych, na przykład stylu konkretnego utworu czy autora. W każdym indywidualnym przypadku uzyskane modele powiedzą coś na temat kompozycji utworu czy upodobań stylistycznych autora. Jednak zawsze, nawet podczas analizy różnych fragmentów tego samego utworu, owo „coś" będzie nieco inne. Uwzględniając te argumenty oraz pierwotne założenia niniejszej monografii, nie przedstawiano szczegółowych analiz sekwencji zdaniowych, ograniczając się do powyższego omówienia wyników uzyskanych we wcześniejszych badaniach.

5.2 SEKWENCJE WYRAZOWE

W teorii lingwistycznej mówi się często o informacyjnej strukturze tekstu narzucającej pewien porządek elementów składniowych i tematycznych w zdaniu i w dłuższych partiach tekstu. Problematyka ta pozostaje jednak poza obszarem lingwistyki kwantytatywnej, ograniczającej się do tych poziomów tekstu, które poddają się wyrazistej segmentacji i kwantyfikacji. Trudno przeliczyć informację semantyczną zawartą w słowach, wyrażeniach lub zdaniach, aktualizowaną w sposób jedyny i niepowtarzalny w każdym akcie komunikacji. Naszym zdaniem, pewnym przybliżeniem informacji w potocznym rozumieniu może być koncepcja C. Shannona definiująca to pojęcie na gruncie teorii prawdopodobieństwa. Dowodzą tego badania prowadzone na szeregach czasowych utworzonych przez ilości informacji (w bitach), niesione w kolejnych słowach tekstu (PAWŁOWSKI 1998:96–111). Dotychczas przedmiotem takich badań były teksty w językach francuskim i angielskim, gdzie wykryto słabe ($V_e \approx 8\%$), ale powtarzające się we wszystkich próbach regularności, które opisano modelem MA(1) (*ibid.*). Uznano, że ich źródłem może być mniej lub bardziej regularne, naprzemienne występowanie w linii tekstu morfemów leksykalnych i gramatycznych, a więc jednostek o niskiej i wysokiej frekwencji. Zjawisko to w największym nasileniu występuje w językach o tendencji analitycznej stosujących zasady składni pozycyjnej. Można oczekiwać, że szeregi czasowe wygenerowane tą samą metodą z tekstu w języku o tendencji syntetycznej dadzą zdecydowanie niższą wartość V_e. Pozytywna weryfikacja tej hipotezy pozwoliłaby uznać V_e za jedno z kryteriów w klasyfikacji typologicznej języków.

5.2.1 Metody ilościowe w typologii języków

Miary ilościowe w klasyfikacji języków najczęściej kojarzone są z osobą J.H. Greenberga, amerykańskiego lingwisty i antropologa kultury. Przedstawił on listę wskaźników liczbowych (tzw. wskaźniki Greenberga) pozwalających na typologiczną klasyfikację języków opartą na kryteriach ilościowych (GREENBERG 1960). Od chwili publikacji metoda Greenberga, podobnie zresztą jak kwestia taksonomii językowej w ogóle, traktowana była przez część lingwistów nieufnie: “Some of the most obvious and frequently mentioned syntactical differences do not easily lend themselves to this technique [...] all these [language characteristics – A.P.], and many more like them, are difficult to reduce to a meaningful number.” (HOUSEHOLDER 1960:195). Charakterystycznym rysem tej krytyki jest podkreślenie niewystarczalności listy Greenberga i, co logiczne, jakiejkolwiek listy wskaźników, ponieważ zawsze będzie można znaleźć kolejne, nie uwzględnione cechy języka, które wpłyną na zmianę klasyfikacji, o ile tylko nada im się formę liczbową. Zdaniem krytyków, niepożądany efekt tego stanu rzeczy jest taki, że klasyfikacje tego samego zbioru obiektów mogłyby prowadzić do różnych wyników. W innym duchu, chociaż także krytycznie, na temat taksonomii wypowiada się G. Altmann: „At early stage of explorative research, one usually *classifies* texts, languages or particular phenomena in order to obtain a map of the scope of taxa. [...] One can gain useful impulses

but one observes that empirical taxonomies lead quickly to a dead end." (ALTMANN 1997:15, por. także ALTMANN&LEHFELDT 1973).

Na przekór temu, dla wielu lingwistów lista wskaźników Greenberga stała się źródłem inspiracji badawczej (SILNITSKY 1993), a klasyfikacje języków oparte na kryteriach ilościowych i zaawansowanych technikach matematycznych pojawiają się w dalszym ciągu (BATAGELJ et al. 1992). Dzieje się tak, ponieważ oba stanowiska można pogodzić, wychodząc z założenia, że nie istnieją klasyfikacje ostateczne i definitywne, a jedynie takie, które podporządkowane są pewnym wyselekcjonowanym kryteriom, a więc *de facto* oparte na jakiejś teorii. Konkretne obiekty (tu języki) mogą więc występować w klasyfikacji w różnych kombinacjach i zależy to jedynie od kryteriów uznanych przez badacza za istotne.

Prowadzone przez nas testy i weryfikacje hipotez nie są głosem w dyskusji o epistemologicznym statusie taksonomii lingwistycznych. Zależy nam raczej na tym, aby pozytywnie zweryfikować zarysowaną już i sformułowaną dalej hipotezę i znaleźć tym samym ilościowy „wskaźnik analityczności" języka, który mógłby być zastosowany w klasyfikacji języków. Wskaźniki Greenberga, wywodzące się jeszcze z koncepcji E. Sapira („The method of classification is fundamentally that of Sapir" – GREENBERG 1960:185), są bowiem zdefiniowane jako proste relacje liczbowe (liczba jednostek posiadających daną cechę dzielona przez całkowitą liczbę jednostek) i nie uwzględniają w żaden sposób kolejności elementów językowych (*ibid.* 181–184).

5.2.2 Hipoteza

Nawet powierzchowna analiza list frekwencyjnych języków o tendencji analitycznej pozwala zauważyć, że bardzo niewielka liczba słów o bardzo wysokich częstościach pokrywa dużą część tekstu. Na przykład dziesięć najczęściej występujących leksemów włoskich pokrywa około 32,5% tekstu, analogiczna wartość dla hiszpańskiego wynosi 33,5%, a dla francuskiego 30,5% (dane pochodzą z reprezentatywnych korpusów o długości 500000 słowoform)[129]. W przypadku języków o tendencji syntetycznej odnośne wartości są wyraźnie niższe. Widać to na przykładzie języków słowiańskich: dziesięć najczęściej występujących słów rosyjskich pokrywa jedynie 18% tekstu (dane z reprezentatywnej próby 1000000 słowoform), dla polskiego wartość ta jest identyczna, dla ukraińskiego wynosi 17% (obie próby złożone z 500000 słowoform), a dla czeskiego wynosi 18,5% (próba złożona z 1623527 słowoform)[130].

Mimo pewnych różnic w długości korpusów różnica jest uderzająca. Pojawia się więc pytanie, czy dysproporcja ta przekłada się w jakiś sposób na sekwencyjną strukturę tekstu. Nie można wykluczyć, że alternacje leksemów rzadkich i bardzo częstych będą pojawiać się częściej w językach o tendencji analitycznej, a rzadziej w językach o tendencji syntetycznej. Sytuację wyjaśnia porównanie zasad składni i morfologii języków

[129] Por. JUILLAND et al. 1971, JUILLAND et al. 1964, BORTOLINI et al. 1971.
[130] ZASORINA 1977, KURCZ et al. 1990, ORLOVA et al. 1981, JELINEK et al. 1961.

obu typów. W językach słowiańskich szyk wyrazów w zdaniu jest raczej swobodny, a liczba morfemów gramatycznych o wysokich frekwencjach stosunkowo niewielka. Podstawowym nośnikiem informacji o charakterze gramatycznym jest bowiem fleksja. W językach o tendencji analitycznej obciążenie funkcjonalne fleksji jest niższe. W zamian za to na określonych regułami pozycjach występuje więcej morfemów gramatycznych o bardzo wysokich częstościach. Patrząc na to zjawisko w perspektywie analizy sekwencyjnej, można domniemywać, że w językach analitycznych zaobserwuje się w miarę regularne przemieszanie wyrazów o częstościach niskich i bardzo wysokich (podział ten pokrywa się z podziałem na morfemy gramatyczne i leksykalne), natomiast w językach o tendencji syntetycznej występowanie w linii tekstu wyrazów o wysokich częstościach będzie mniej regularne.

Tak sformułowana hipoteza poddana zostanie testom na szeregach czasowych reprezentujących ilości informacji (w rozumieniu shannonowskim) niesione przez poszczególne słowa tekstu. Przyjmuje się zarazem, że rozróżnienie dwóch idealnych typów języków alfabetycznych (analityczne i syntetyczne) jest koncepcją poprawną, choć oczywiście podział taki nie jest jedyny. Założenie powyższe można kwestionować, opierając się na przedstawionej wcześniej krytycznej argumentacji. Jednak za jego poprawnością przemawia wiele faktów językowych, co więcej, bez minimalnych choćby ograniczeń i założeń trudno sobie wyobrazić sensowne prowadzenie badań i formułowanie uogólnień.

5.2.3 Dane i kwantyfikacja

Testy przeprowadzono na tekstach prozatorskich w językach włoskim (tendencja analityczna) i polskim (tendencja syntetyczna). Język polski reprezentowany był przez dwadzieścia prób pochodzących z powieści T. Konwickiego (10) i A. Szczypiorskiego (10). Język włoski reprezentowało dziesięć fragmentów powieści A. Moravii (szczegółowe informacje zawiera ANEKS). Przeciętna długość próby wynosiła około stu wyrazów. Wyniki badań zostały też porównane z uzyskanymi wcześniej rezultatami dla języka francuskiego (PAWŁOWSKI 1998:96–111).

Jak już wspomniano, kwantyfikacja danych polegała na zastąpieniu kolejnych słowoform tekstu shannonowską ilością informacji obliczoną według podanego wcześniej wzoru (por. Część I, 5.2 i 6.2):

(91) $$I_n = -\log_2 p_n$$

gdzie I_n – ilość informacji (w bitach) niesionej przez symbol n
p_n – prawdopodobieństwo wystąpienia symbolu n

Warto zastanowić się nad techniką obliczania wartości p_n. Prawdopodobieństwo pojawienia się konkretnej słowoformy w tekście nie jest ustalone raz na zawsze i waha się w pewnych granicach. Przyjęto, że dobrą metodą jego oszacowania będzie wykorzy-

stanie danych ze słowników frekwencyjnych badanych języków[131], gdzie podane są częstości słowoform i odnośnych form hasłowych (lematyzowanych), znana też jest całkowita długość korpusu. Pewna ilość słowoform znalezionych w tekstach nie wystąpiła w słownikach (przeciętnie około 6%). Dotyczyło to przede wszystkim nazw własnych. W takich przypadkach nadawano im częstość równą maksymalnej częstości wyrazów nie uwzględnianych przez słownik ($f = 3$). Ponieważ zdarzało się, iż w słownikach nie występowały pewne formy fleksyjne odnalezione w tekstach, natomiast pojawiały się odnośne formy hasłowe (na przykład słownik notował jedynie liczbę pojedynczą pewnego rzeczownika, a w tekście występował on w liczbie mnogiej), kwantyfikację przeprowadzono na podstawie częstości form lematyzowanych. Metoda taka nie jest oczywiście jedynym rozwiązaniem. Informację I_n można przybliżać przeprowadzając eksperyment z odgadywaniem kolejnych słów w tekście (HAMMERL&SAMBOR 1990:438–443, por. też Część I, 5.2). Można też stworzyć własny korpus tekstów, w którym obliczy się empiryczne prawdopodobieństwa poszczególnych słowoform i/lub haseł. Dobrym rozwiązaniem jest na przykład wykorzystanie kompletu dzieł danego autora (można wtedy zrezygnować z estymacji i mówić o indukcji zupełnej). Wstępne testy pokazały, że wyjąwszy sposób pierwszy (odgadywanie), metoda pozyskania empirycznych prawdopodobieństw p_n nie wpływa w znaczący sposób na ostateczny rezultat.

Zgodnie z zastosowaną tu metodą kodowania zdanie w języku włoskim „Io alzai le spalle ed uscii in punta di piedi" byłoby reprezentowane sekwencją liczbową {6,00 11,95 3,19 12,36 5,30 11,19 5,82 11,05 4,30 11,49} (por także Rys. 1). Jako że częstości wyrazów (a w konsekwencji odpowiadające im ilości bitów) pozostają w pewnych granicach stabilne, badane szeregi czasowe można uważać za stacjonarne w sensie szerokim (por. Część I, 6.3.2).

Inną kwestią jest uzasadnienie rezygnacji z posługiwania się częstościami absolutnymi lub względnymi na rzecz ilości informacji. Z technicznego punktu widzenia obliczanie p_n i zamiana go na I_n jest bądź co bądź obciążeniem, które spowalnia procedurę badawczą. Jednak wartości I_n, w przeciwieństwie do liczb bezwzględnych, posiadają przekonującą lingwistyczną interpretację. Z jednej strony są uniwersalne i mogą reprezentować dowolny kod lub język, z drugiej strony stanowią jakieś przybliżenie pojęcia informacji w rozumieniu potocznym (por. Część I, 5.2).

Aby zilustrować zastosowaną procedurę, przedstawiono przykładowy przelicznik niektórych częstości wyrazów na ilości informacji przy założeniu, że długość korpusu wynosi 500000 słowoform:

Częstość	3	5	10	25	50	250	900	1500	2500	5000	9000
Liczba bitów	17,35	16,61	15,61	14,29	13,29	10,97	9,12	8,38	7,64	6,64	5,80

Podobnie jak w dotychczas prowadzonych testach, otrzymane szeregi czasowe poddano analizie metodą ARIMA, która daje bardzo dobre rezultaty przy estymacji

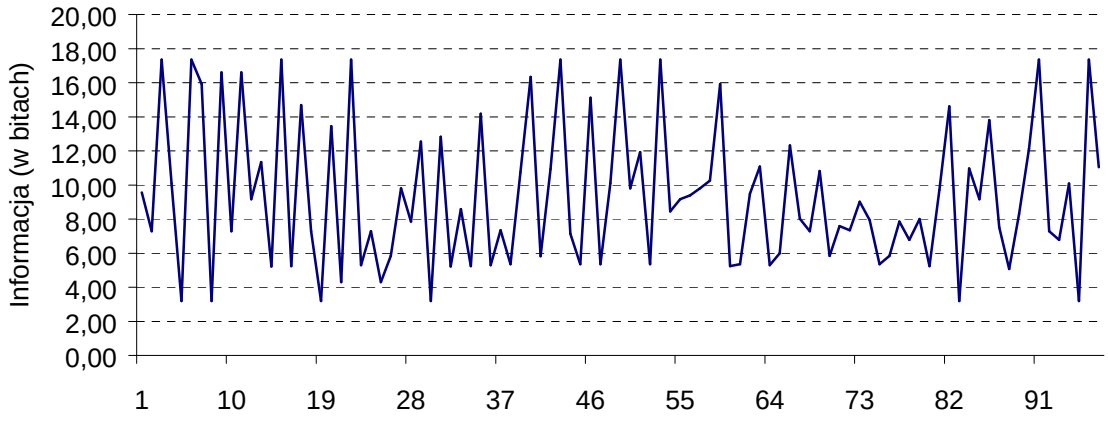
[131] KURCZ et al. 1990 oraz BORTOLINI et al. 1971.

danych zawierających słabe składowe deterministyczne, pozwala też na stosunkowo łatwe obliczenie procentu wariancji (zmienności) szeregu obserwowanego wyjaśnionej przez model (współczynnik V_e).

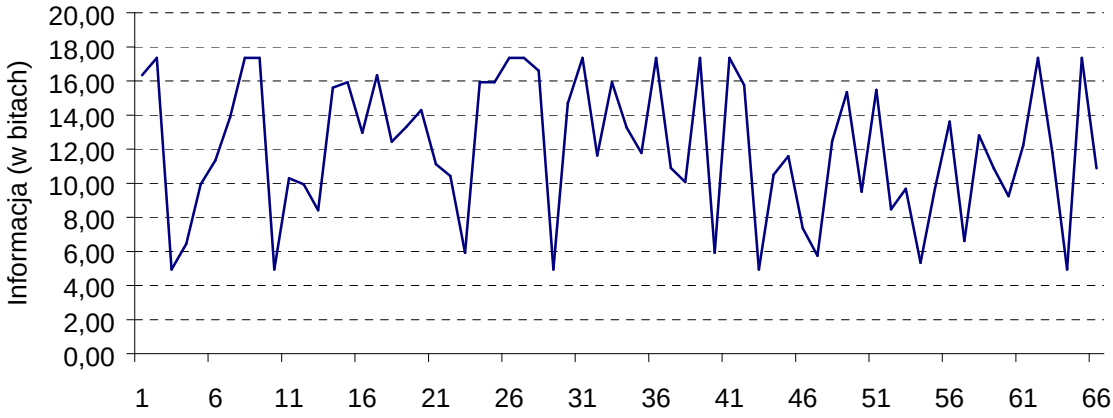
5.2.4 Analiza szczegółowa

Analizie szczegółowej poddano dwie próby reprezentujące oba badane języki[132]. Na wstępie prezentujemy histogramy obu szeregów (Rys. 43 i 44). Kształty krzywych wyglądają na nieuporządkowane, chociaż w tekście włoskim widoczna jest dość regularna alternacja. Próby rysowania na wykresach linii trendu bądź ruchomych średnich także nie wnoszą do analizy nic nowego. Trudno więc na tym etapie powiedzieć cokolwiek o ich ewentualnych składowych deterministycznych.

Rys. 43 Histogram sekwencji ilości informacji w słowach tekstu włoskiego



Rys. 44 Histogram sekwencji ilości informacji w słowach tekstu polskiego

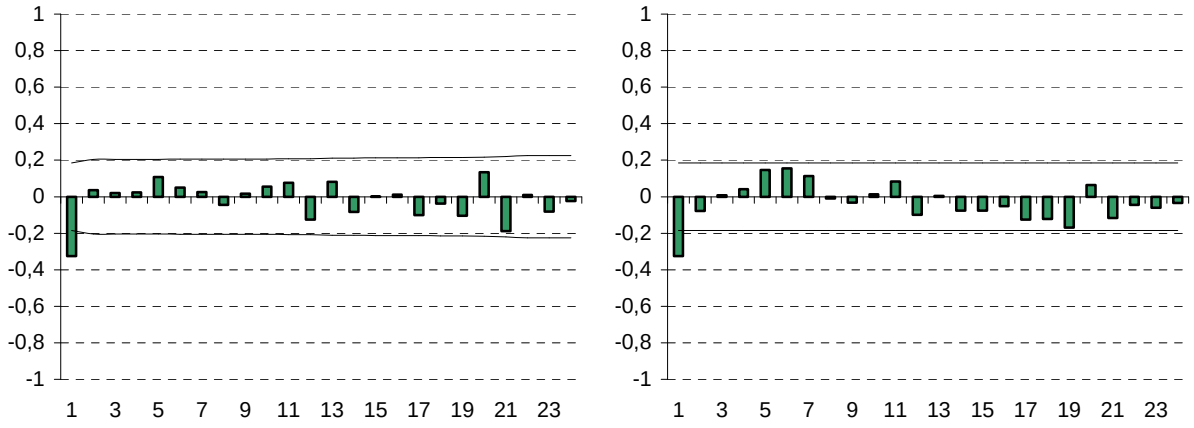


W następnej kolejności przedstawiamy wykresy funkcji autokorelacji i autokorelacji cząstkowej dla obu analizowanych prób. W przypadku języka włoskiego (Rys. 45) za-

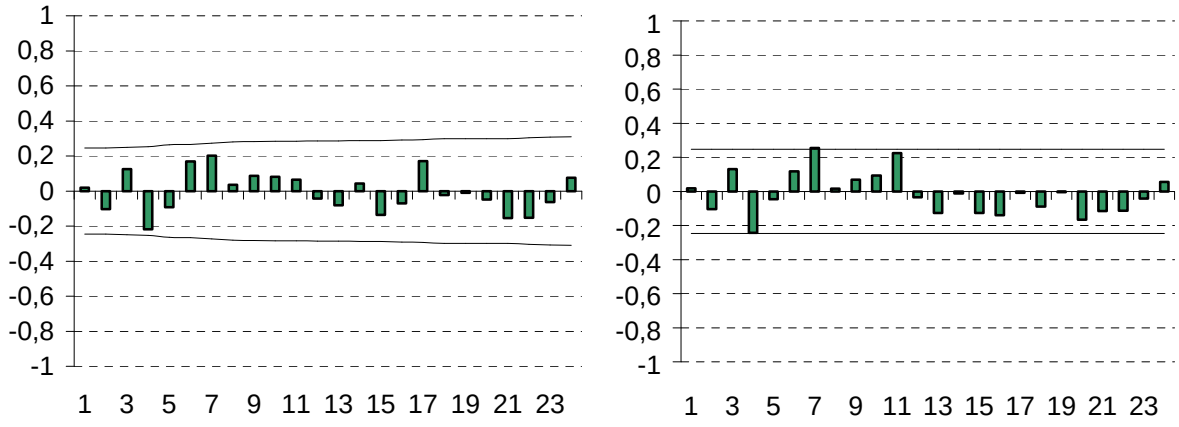
[132] ANEKS – MORAVIA 1963:182 oraz KONWICKI 1982:68.

uważalna jest korelacja negatywna dla pierwszego odstępu ACF. Układ prążków PACF nie jest jednoznaczny, ale raczej uznać go należy za gasnący. Takie kształty obu funkcji sugerują estymację nieznanego procesu stochastycznego modelem typu MA(1).

Rys. 45 Autokorelacja i autokorelacja cząstkowa dla sekwencji ilości informacji w tekście włoskim (język analityczny)



Rys. 46 Autokorelacja i autokorelacja cząstkowa dla sekwencji ilości informacji w tekście polskim (język syntetyczny)



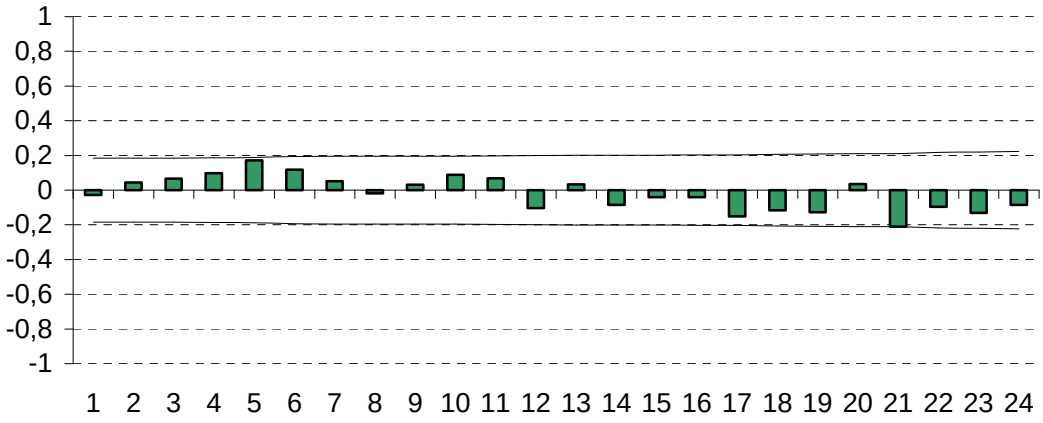
Zupełnie inaczej prezentują się analogiczne funkcje obliczone dla tekstu polskiego (Rys. 46). Brak jest statystycznie znaczącej autokorelacji, co wskazuje na losowy charakter szeregu. Wnioski płynące z porównania obu wykresów (45 i 46) mają na tym etapie charakter jednostkowy, jednak sugerują wyraźnie, iż wysunięta hipoteza może znaleźć oparcie w faktach.

Model procesu estymowano jedynie dla sekwencji włoskiej. Brak autokorelacji danych w szeregu reprezentującym tekst polski oznacza bowiem nieobecność składowej deterministycznej, którą należałoby wyjaśnić. Dla fragmentu włoskiego otrzymano model MA(1) o postaci:

(92) $$x_t = (1 - 0{,}28B)e_t$$

Wyjaśnia on 10% wariancji szeregu obserwowanego i może być uznany za satysfakcjonujący. Funkcja autokorelacji szeregu resztowego (Rys. 47) nie jest co prawda całkowicie płaska, jednak po pierwsze, nie odfiltrowane wartości są niewielkie i statystycznie nieistotne, po wtóre, jak pokazały dalsze testy, każdy analizowany fragment pozostawiał po odfiltrowaniu szeregu MA(1) nieco inny wzór, co wyklucza możliwość uogólnienia tej obserwacji.

Rys. 47 Autokorelacja i autokorelacja cząstkowa dla sekwencji ilości informacji w tekście polskim (język syntetyczny)



Warto porównać funkcje ACF i PACF obu szeregów z ich histogramami: z kształtu krzywych (Rys. 43 i 44) raczej trudno jest wywnioskować, czy w przedstawionych danych są jakieś składowe deterministyczne. Trudno też, na podstawie kształtu funkcji ACF i PACF dla tekstu włoskiego (Rys. 45), określić opisywany przez nie szereg jako rytmiczny (Rys. 43). Zauważona regularność jest bowiem zbyt słaba ($V_e = 7\%$) i zapewne dla podmiotu mówiącego niewyczuwalna. Jest jednak na tyle wyraźna, by jej ewentualne potwierdzenie w innych próbach pozwoliło uznać ten rodzaj miary za potencjalne narzędzie klasyfikacji języków.

5.2.5 Wyniki sumaryczne

Aby porównać wszystkie badane próby, obliczono i uśredniono wartości współczynników ACF przy odstępach $k = 1$ i $k = 2$ (r_1 i r_2 w tabeli 28). Wynik ten nie pozostawia żadnych wątpliwości. Sekwencje ilości informacji w słowach tekstu w języku o tendencji syntetycznej tworzą szereg pozbawiony autokorelacji. Natomiast sekwencje wygenerowane z języka włoskiego (tendencja analityczna) zawierają słabą składową deterministyczną wskazującą na obecność procesu rzędu pierwszego. Istotnie, wszystkie próby języka włoskiego najlepiej opisywał model MA(1).

W celu oszacowania jakości dopasowania otrzymanych modeli obliczono wartości współczynnika V_e stanowiącego tu zmienną decyzyjną, na podstawie której odrzuca się bądź pozytywnie weryfikuje przyjętą hipotezę. Tabela 29 zawiera wyniki obliczeń. Mimo

braku składowych deterministycznych w tekstach polskojęzycznych, obliczono dla nich wartości współczynnika V_e. Wyjaśnienie tego faktu znajduje się w samej tabeli: nieliczne, pojedyncze próby w języku polskim zawierały składowe deterministyczne ($V_e > 0$), co zwiększyło uśrednione wartości V_e.

Tab. 28 Współczynniki autokorelacji obliczone dla sekwencji ilości informacji w kolejnych słowach tekstów w językach włoskim i polskim[133]

numer próby	→	1	2	3	4	5	6	7	8	9	10	*Średnia*
j. włoski	r_1	-,11	-,33	-,34	-,31	-,28	-,27	-,21	-,26	-,23	-,31	**–,27**
(A. Moravia)	r_2	,10	,02	,07	-,03	,15	,06	-,07	-,01	,01	,05	**,04**
j. polski	r_1	-,03	-,09	-,27	-,17	-,07	,04	-,02	-,09	-,13	-,06	**–,09**
(A. Szczypiorski)	r_2	,02	-,07	-,04	,18	,15	-,08	-,02	,14	,06	-,03	**,03**
j. polski	r_1	-,03	,04	,08	,10	,02	,15	,15	-,26	-,13	-,22	**–,01**
(T. Konwicki)	r_2	-,06	,05	-,07	,01	,12	,02	,04	-,12	,14	-,09	**,00**

Wynik ten ostatecznie potwierdza słuszność przeprowadzonego na wstępie rozumowania i pozwala utrzymać wysuniętą hipotezę. Wartość V_e dla języka włoskiego (6,9%) trudno wprawdzie uznać za imponującą – V_e obliczane dla szeregów akcentowych były znacznie wyższe. Jednak fakt, iż pojawia się ona regularnie, pozwala uznać współczynnik V_e za miarę analityczności języka przydatną w klasyfikacji języków. Zaobserwowane anomalie wskazują natomiast, że miara ta powinna być stosowana jako średnia z większej liczby prób.

Tab. 29 Procent wariancji szeregu obserwowanego wyjaśniony przez model MA(1)

numer próby	→	1	2	3	4	5	6	7	8	9	10	*Średnia*
j. włoski	s_0^2	20,0	19,0	19,2	18,7	17,7	16,3	19,5	15,3	17,1	19,2	
(A. Moravia)	s_r^2	19,8	17,1	16,7	17,3	16,4	15,0	18,8	14,4	16,4	17,7	
	V_e	1%	10%	13%	8%	7%	8%	4%	6%	4%	8%	**6,9%**
j. polski	s_0^2	11,3	13,9	14,7	17,8	15,7	14,2	9,6	17,0	15,3	18,4	
(A. Szczypiorski)	s_r^2	11,3	13,9	13,9	17,4	15,7	14,2	9,6	17,0	15,2	18,4	
	V_e	0%	0%	5%	2%	0%	0%	0%	0%	1%	0%	**0,08%**
j. polski	s_0^2	15,5	15,1	15,2	13,5	16,6	10,7	13,3	18,0	14,1	16,0	
(T. Konwicki)	s_r^2	15,5	15,1	15,2	13,5	16,6	10,7	13,3	17,1	14,1	15,4	
	V_e	0%	0%	0%	0%	0%	0%	0%	5%	0%	4%	**0,09%**

Oznaczenia:

s_0^2 – wariancja szeregu obserwowanego
s_r^2 – wariancja szeregu resztowego
V_e – procent wariancji wyjaśnionej przez model MA(1)

[133] Za znaczące uznaje się wartości leżące poza przedziałem [–0,2, 0,2].

Ważnym argumentem wspierającym te spostrzeżenia są wyniki wcześniejszych badań prowadzonych na językach francuskim i angielskim (PAWŁOWSKI 1998:96–111). Na przykład modele estymowane dla języka francuskiego (szeregi generowano w podobny sposób) wyjaśniały przeciętnie około 8% zmienności szeregu wyjściowego, czyli więcej niż modele estymowane dla włoskiego (*ibid.* 103–104). Ta różnica wartości może oczywiście wynikać z różnych przyczyn, jest jednak prawdopodobne, że dzięki współczynnikowi V_e języki, podobnie jak style, można uszeregować według stopnia ich analityczności.

6. ZAKOŃCZENIE

Celem przeprowadzonych badań była weryfikacja hipotezy, zgodnie z którą „linearny porządek niektórych jednostek językowych w tekście stanowi realizację jakiegoś procesu stochastycznego i z tego względu nie ma charakteru losowego" (por. s. 25). Tak ogólne sformułowanie nie mogło być przedmiotem testów empirycznych, jednak pozwoliło na wysunięcie kilku hipotez szczegółowych, dotyczących wyraźnie określonych poziomów i/lub zakresów analizy językoznawczej.

Najbardziej efektywne okazały się hipotezy testowane na poziomie struktury prozodycznej. Modelowanie szeregów czasowych otrzymanych poprzez kwantyfikację symbolicznych sekwencji tekstowych, reprezentujących ciągi złożone z sylab akcentowanych i nie akcentowych, pozwoliło zastąpić mgliste pojęcie „rytmu tekstu" ilościowym, syntetycznym wskaźnikiem zrytmizowania tekstu (oznaczanym przez V_e). Wskaźnik ten można obliczyć dla każdego szeregu czasowego, co umożliwia porównywanie dowolnych tekstów, niezależnie od ich języka, stylu, systemu wersyfikacji i długości. Zestawiając wartości wskaźnika V_e obliczone dla różnych próbek pokazano, że na rytmikę tekstu wpływają pozycja akcentu wyrazowego oraz styl i system wersyfikacji.

Testując hipotezy szczegółowe wykazano, że wskaźnik V_e jest efektywnym ilościowym kryterium klasyfikacji języków ze względu na typ akcentuacji. Co prawda, baza weryfikacyjna ograniczona była do dwóch języków słowiańskich (polskiego i rosyjskiego), jednak żadne merytoryczne względy nie stoją na przeszkodzie, by ująć w klasyfikacji inne języki alfabetyczne stosujące akcent dynamiczny. Ponadto, zacytowano przykład analizy sekwencyjnej języka chińskiego (por. DREHER et al 1969), co sugeruje możliwość dokonania taksonomii języków tonalnych, opartej na tej samej metodologii. W takim przypadku cechą poddaną kwantyfikacji, byłaby swoista „melodia" tekstu wyznaczona sekwencją tonów.

Przeprowadzono także analizę rytmiki języka polskiego, wykazując statystycznie znaczące różnice pomiędzy różnymi stylami oraz systemami wersyfikacyjnymi. Baza weryfikacyjna była w tym przypadku stosunkowo szeroka (por. Rys. 38, s. 116), choć nie objęła niektórych, trudnych do sklasyfikowania, odmian wersyfikacyjnych (wiersz biały, proza poetycka, systemy wersyfikacji oparte na tonizmie). Na tym etapie badań brak ten nie jest jednak szczególnie istotny. Każda metoda badawcza, zanim zostanie zastosowana

do rozwiązania konkretnego problemu, powinna przejść wiele testów. Temu właśnie celowi służył dobór tekstów poddanych analizie (jako kryterium selekcji przyjęto silne zróżnicowanie stylistyczne), nawet jeżeli nie można go uznać za kompletny. Uzyskane wyniki potwierdziły jednak możliwość przeprowadzenia znacznie obszerniejszej taksonomii stylów opartej na modelowaniu rytmicznej struktury tekstu (w dowolnym języku alfabetycznym). Porównując równoległe fragmenty prozy w językach polskim i rosyjskim wykazano także możliwość zastosowania parametrów sekwencyjnych tekstu, takich jak funkcja autokorelacji i współczynnik V_e, w przekładoznawstwie.

Metoda ARIMA okazała się bardzo skutecznym narzędziem analizy wersyfikacji. Funkcja autokorelacji szeregu czasowego uzyskanego poprzez kwantyfikację sekwencji sylab akcentowanych i nie akcentowanych wykazywała nie tylko głębokość i siłę związku kontekstowego bezpośrednio sąsiadujących ze sobą jednostek językowych, ale także minimalną długość ekwiwalentnego pod względem rytmicznym, powtarzalnego odcinka tekstu. W większości przypadków długość takiego odcinka pokrywała się z długością wersu, jednak testy udowodniły, że wcale tak być nie musi. Przy analizie rosyjskiego tekstu *Eugeniusza Oniegina* okazało się, że statystycznie najczęściej powtarzany jest odcinek długości siedemnastu sylab, powstały poprzez złożenie dwóch następujących po sobie wersów (9+8). Natomiast minimalnym, powtarzalnym odcinkiem rytmicznym w formalnie ośmiozgłoskowym sylabotoniku J. Brzechwy, był czterosylabowy hemistych.

Bazy weryfikacyjnej nie ograniczono do języków stosujących akcent dynamiczny. Interesujące wyniki dało porównanie struktury rytmicznej heksametru łacińskiego kodowanego w postaci sekwencji iloczasowej, a więc złożonej z sylab długich i krótkich, oraz „iktowanej", opartej na sekwencji sylab akcentowanych i nie akcentowanych dynamicznie. W drugim przypadku przyjęto hipotezę zakładającą istnienie tzw. iktu (łac. *ictus*) – akcentu metrycznego (albo wierszowego) pojawiającego się na niektórych pozycjach w wersie podczas deklamacji poezji łacińskiej. Wynik ten dlatego uznano za interesujący, że wbrew obiegowym opiniom nośnikiem rytmu w łacińskim heksametrze okazała się sekwencja akcentowa, a nie iloczasowa. O ile więc wersyfikacja łacińska strukturalnie oparta jest na iloczasie, twierdzenie, przynajmniej w odniesieniu do heksametru, iż „rytm poezji łacińskiej wyznacza sekwencja sylab długich i krótkich", nie odpowiada wynikom przeprowadzonych testów.

Prawidłowości ujawnione podczas badań prozodycznej struktury tekstu można wyjaśnić, odwołując się do jednej z podstawowych zasad językoznawstwa (cytowanej tu za T. Milewskim, por. s. 128–129), mówiącej, że „mowa realizuje się na wydechu", a jej periodyczność jest skutkiem regularnego rytmu oddechów. Uzyskane wyniki pozwalają nadać tej zasadzie ścisłą formę ilościowego prawa językowego. W tym ujęciu, stwierdzona we wszystkich badanych tekstach regularność akcentuacji byłaby uwarunkowana fizjologicznie, natomiast różnice w stopniu zrytmizowania tekstów (wyrażone współczynnikiem V_e) oraz odmienne wzorce rytmiczne (wyrażone funkcją autokorelacji i/lub modelami procesów stochastycznych) byłyby efektem oddziaływania ograniczeń formalnych o podłożu kulturowym, takich jak miejsce akcentu wyrazowego, styl tekstu lub system wersyfikacji.

Dobre wyniki dało także modelowanie sekwencyjnych struktur tekstu na poziomie leksykalnym. Wykazano, że przedmiotem efektywnej analizy metodą ARIMA mogą być sekwencje ilości informacji (w rozumieniu definicji C. Shannona) niesionej przez kolejne słowoformy tekstu. Testowano hipotezę, zgodnie z którą na poziom tego „informacyjnego rytmu" (mierzony współczynnikiem V_e) wpływ powinien mieć typ składni danego języka. Wynik weryfikacji był satysfakcjonujący: szeregi czasowe oparte na tekstach polskojęzycznych (składnia typu syntetycznego, swobodny szyk wyrazów w zdaniu) były praktycznie pozbawione rytmu, natomiast sekwencje oparte na języku włoskim (składnia typu analitycznego, względnie uporządkowany szyk wyrazów w zdaniu) wykazywały słabe, ale statystycznie znaczące regularności. Biorąc pod uwagę publikowane wcześniej wyniki analogicznych testów prowadzonych na innych językach, uznano, że współczynnik V_e, wyrażający swoisty stopień „informacyjnego uporządkowania tekstu", można uważać za efektywne kryterium w typologicznej klasyfikacji języków.

Bilans przeprowadzonych badań należy uznać za pozytywny. Odkryto wiele prawidłowości występujących w sekwencyjnych strukturach języka i przedstawiono je w formie sformalizowanych modeli matematycznych, nadając im jasną interpretację lingwistyczną. Wszystkie testowane hipotezy oparte były na przesłankach dedukcyjnych, niezależnych od konkretnego materiału językowego. Wyraźnie określono jednak sposób i warunki, w jakich można poddać je falsyfikacji. Dla uogólnień o charakterze teoretycznym przedstawiono także przykładowe zastosowania praktyczne, głównie w zakresie klasyfikacji języków, taksonomii tekstów i wersologii.

O tym, czy analiza sekwencyjna znajdzie dalsze zastosowania w językoznawstwie, trudno dziś jednak przesądzać. Jak zauważył Saint-Evremond, XVII-wieczny francuski moralista i myśliciel: „Les mathématiques, à la vérité, ont beaucoup plus de certitude; mais, quand je songe aux profondes méditations qu'elles exigent, comme elles vous tirent de l'action et des plaisirs, pour vous occuper tout entier, ces démonstrations me semblent bien chères, et il faut être fort amoureux d'une vérité, pour la chercher à ce prix-là."[134] Mimo upływu lat, refleksja ta zachowała zadziwiającą aktualność.

[134] SAINT-EVREMOND, *Jugement sur les sciences où peut s'appliquer un honnête homme*, 1662. Cytat na podstawie wydania SAINT-EVREMOND, *Œuvres mêlées*, Edizioni dell'Ateneo, Roma 1966, 88.

BIBLIOGRAFIA

ADAM J.-M. (1992), *Les textes: types et prototypes: récit, description, argumentation, explication et dialogue*. Paris: Nathan Université.

ALTMANN G. (1978), Towards a Theory of Language. In: *Glottometrika* 1, 1–25.

— (1980), Prolegomena to Menzerath's Law. In: *Glottometrika* 2, 1–10.

— (1993), *Science and Linguistics*. In: R. Köhler, B.B. Rieger (1993) (red.), *Contributions to Quantitative Linguistics*. Dodrecht: Kluwer, 3–10.

— (1997), The Art of Quantitative Linguistics. In: *Journal of Quantitative Linguistics* 4/1–3, 13–22.

ALTMANN G., KOCH W.A. (1998) (red.), *Systems*. Berlin, New York: Walter de Gruyter.

ALTMANN G., KÖHLER R. (2002) (red.), *Handbook of Quantitative Linguistics*. Berlin itd.: Walter de Gruyter (w przygotowaniu).

ALTMANN G., LEHFELDT W. (1973), *Allgemeine Sprachtypologie*. München: Fink.

ALTMANN G., SCHWIBBE M. (1989), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim itd.: Olms.

AMSTERDAMSKI S. (1987), Rozwój nauki. In: *Filozofia a nauka. Zarys encyklopedyczny*. Wrocław itd.: Ossolineum, 589–598.

ARAPOV M.V., CHERC M.M. (1974), *Matematičeskie metody v istoričeskoj lingvistike*. Moskva: Nauka.

— (1983), *Matematische Methoden in der historischen Linguistik*. Bochum: Brockmeyer.

ARYSTOTELES, *Retoryka, Poetyka*. Warszawa, 1988, PWN. Przełożył i wstępem opatrzył H. Podbielski.

ATKINSON J.M., HERITAGE J. (1984) (red.), *Structure of Social Action. Studies in Conversational Analysis*. London, Paris: Oxford University Press, Maison des Sciences de l'homme.

AZAR M., KEDEM B. (1979), Some Time Series in the Phonetics of Biblical Hebrew. In: *Bulletin of the ALLC* 7(2) 111–129.

BAEVSKIJ V.S., OSIPOVA L.JA. (1987), The Algorithm and some Results of a Statistical Investigation of Alternating Rhythm on the Minsk-32 Computer. In: *Glottometrika* 8, 157–177.

BARTLETT M.S. (1946), On the Theoretical Specification of Sampling Properties of Autocorrelated Time-Series. In: *Journal of the Royal Statistical Society*, B8, 27–41.

BATAGELJ V., PISANSKI T., KERŽIČ D. (1992), Automatic Clustering of Languages. In: *Computational Linguistics* 1992/18 (3), 339–352.

BAVAUD F. (1998), *Modèles et données*. Paris: L'Harmattan.

BELLERT I. (1971), *O pewnym warunku spójności tekstu*. In: M.R. Mayenowa (1971) (red.), *O spójności tekstu*. Wrocław itd.: Ossolineum, 47–76.

BENVENISTE E. (1966), *Problèmes de linguistique générale*. Paris: Gallimard.

BEÖTHY E., ALTMANN G. (1984a), The diversification of meaning of Hungarian verbal prefixes. II. *ki-*. In: *Finisch-Ugrische Mitteilungen* 8, 29–37.

— (1984b), Semantic diversification of Hungarian verbal prefixes. III. *föl-*, *el-*, *be-*. In: U. Rothe (1984) (red.), *Glottometrika* 7, 45–56.

— (1991), The diversification of meaning of Hungarian verbal prefixes. I. *meg-*. In: U. Rothe (1991) (red.), *Diversification Processes in Language: Grammar*. Hagen: Rottman., 60–66.

BLOOMFIELD P. (1976), *Fourier analysis of time series: an introduction.* New York itd.: John Wiley.

BORODA M.G. (1994), Complexity Oscillations in a Coherent Text: Towards the Rhythmic Foundations of Text Organization. In: *Journal of Quantitative Linguistics* 1(1), 87–97.

BORTOLINI U., TAGLIAVINI C., ZAMPOLLI A. (1971), *Lessico di frequenza della lingua italiana contemporanea*. Milano: Garzanti.

BOURSIN J.-L. (1981), *Méthodes statistiques de la gestion*. Paris: Vuibert.

BOX G., JENKINS G. (1983), *Analiza szeregów czasowych*. Warszawa: PWN (z angielskiego przełożył W. Herer).

BRAINERD B. (1976), On the Markov Nature of Text. In: *Linguistics* 176, 5–30.

BRATLEY P., ROSS D. (1981), Syllabic Spectra. In: *ALLC Journal* 2(2), 41–50.

BRÉAL M. (1991), *The Beginnings of Semantics*. Stanford: Stanford University Press.

BROCKWELL P., DAVIES R. (1991), *ITSM: an interactive time series modelling package for the PC*. Berlin itd.: Springer Verlag.

— (1996), *Introduction to time series and forecasting*. New York itd.: Springer.

BÜHLER K. (1934), *Sprachteorie: die Darstellungsfunktion der Sprache*. Jena: G. Fischer.

CAIRNS F., CRAVEN P.G., HOWIE J.G.H. (1981), Textcode: Grammatical, Syntactical, metrical and accentual information in machine-readable form. In: *The ALLC Journal* 9/3, 1981, 13–18.

CHAGHAGHI F. (1985), *Time series package (TSPACK)*. Berlin itd.: Springer.

CHMIELEWSKI A. (2000), *Haute couture* czy *prêt-à-porter?* czyli czekając na nową polską szkołę filozoficzną. In: *Odra* 6/2000, 40–46.

CHRISHOLM D. (1981), Prosodic approaches to twentieth-century verse. In: *The ALLC Journal* 2/1981, 34–39.

CORDUAS M. (1995), *La struttura dinamica dei dati testuali*. In: S. Bolasco et al. (1995) (red.), *Analisi Statistica dei Dati Testuali, III Journées Internationales d'Analyse Statistique des Données Textuelles*, Roma 11–13 XII 1995, 345–352.

COSSETTE A. (1994), *La richesse lexicale et sa mesure*. Paris: Champion.

COUTROT B., DROESBEKE J.J. (1984), *Les méthodes de prévision*. Paris: PUF.

CRYER J. (1986), *Time series analysis*. Boston: Duxbury Press.

CRYSTAL D. (1997), *The Cambridge Encyclopædia of Language*. Cambridge: Cambridge University Press.

DAMERAU F.J. (1971), *Markov models and linguistic theory*. The Hague, Paris: Mouton.

DANEŠ F. (1974) (red.), *Papers on Functional Sentence Perspective*. Praha: Academia.

DIJK T.A. VAN (1980), *Macrostructures. An Interdisciplinary Study of Global Structures in Discourse, Interaction and Cognition*. Hillsdale itd.: Erlbaum.

DILLON M. (1970), The Quantitative Analysis of language: Preliminary Considerations. In: *Computer Studies in the Humanities and Verbal Behahior* 3, 191–207.

DITTMANN J. (1979) (red.), *Arbeiten zur Konversationsanalyse*. Tübingen: Niemeyer Verlag.

DREHER J., YOUNG E., NORTON R., MA J. (1969), Power Spectral Densities of Literary Speech Rhythms. In: *Computer Studies in the Humanities and Verbal Behahior* 2, 170–191.

DRESSLER W.U. (1972), *Einführung in die Textlinguisitk*. Tübingen: Niemeyer.

DUSZAK A. (1998), *Tekst, dyskurs, komunikacja międzykulturowa*. Warszawa: PWN.

ELMAN J. (1990), Finding structure in time. In: *Cognitive Science* 14, 179–211.

EROMS H.-W. (1986), *Funktionale Satzperspektive*. Tübingen: Niemeyer.

ESTOUP J.B. (1916), *Gammes sténographiques. Méthode et exercices pour l'acquisition de la vitesse*. Paris: Institut Sténographique.

FELLER W. (1987), *Wstęp do rachunku prawdopodobieństwa*. Warszawa: PWN (z angielskiego przełożyli R. Bartoszyński, B. Bielecki).

FENK A., FENK-OCZLON G. (1993), Menzerath's Law and the Consonant Flow of Linguistic Information. In: R. Köhler and B.B. Rieger (1993) (red.), *Contributions to Quantitative Linguistics*. Dodrecht: Kluwer Academic Publishers, 11–31.

FIRTH J.R. (1957), *Papers in Linguistics, 1934–1951*. London: Oxford University Press.

FISHMAN G.S. (1981) *Symulacja komputerowa. Pojęcia i metody*. Warszawa: Państwowe Wydawnictwo Ekonomiczne (z angielskiego przełożyli S. Bartosiewicz i J. Jakubczyc).

FISIAK J. (1985), *Wstęp do współczesnych teorii lingwistycznych*. Warszawa: Wydawnictwa Szkolne i Pedagogiczne.

FUCKS W. (1952), On mathematical analysis of style. In: *Biometrika* 39, 122–129.

GASPARSKI W. (1987), Systemów teoria. In: *Filozofia a nauka. Zarys encyklopedyczny*. Wrocław itd.: Ossolineum, 1978, 696–703.

GATNAR E. (1998), *Symboliczne metody klasyfikacji danych*. Warszawa: PWN.

GLASS G., WILSON V., GOTTMAN J. (1975), *Design and Analysis of Time-Series Experiments*. Colorado: Colorado Associated University Press.

GLEICK J. (1987), *Chaos: Making a New Science*. New York: The Viking Press.

GOOD I.J. (1953), On the Population frequencies of Species and estimation of population parameters. In: *Biometrika* 40/1953, 237–264.

GOTTMAN J.M. (1981), *Time-series analysis: a comprehensive introduction for social scientists*. Cambridge itd.: Cambridge University Press.

— (1990), *Sequential analysis*. Cambridge itd.: Cambridge University Press.

GREENBERG J.H. (1960), A Quantitative Approach to the Morphological Typology of Language. In: *International Journal of American Linguistics* 26(3), 178–194 (pierwodruk: R.F. Spencer (1954), *Methods and Perspectives in Anthropology: Papers in Honor of Wilson D. Wallis*. University of Minnesota Press.

GREŃ J. (1987), *Statystyka matematyczna*. Warszawa: PWN.

GROTJAHN R. (1979), *Linguistische und statistische Methoden in Metrik und Textwissenschaft*. Bochum: Brockmeyer.

— (1980), The Theory of Runs as an Instrument for Research in Quantitative Linguistics. In: *Glottometrika* 2, 11–43.

— (1981) (red.), *Hexameter Studies*. Bochum: Brockmeyer.

GUILBAUD G.TH. (1979), Note sur les comptabilités markoviennes. In: *Mathématiques et Sciences Humaines* 66, 99–112.

GUIRAUD P. (1954), *Bibliographie critique de la statistique linguistique*. Utrecht: Spectrum.

GUITER H., ARAPOV M.V. (1982) (red.), *Studies on Zipf's Law*. Bochum: Brockmayer.

HAKEN H. (1978), *Synergetics*. Berlin itd.: Springer.

HAMMERL R. (1991), *Untersuchungen zur Struktur der Lexik: Aufbau eines Lexikalischen Basismodells*. Trier: RAM.

HAMMERL R., SAMBOR J. (1990), *Statystyka dla językoznawców*. Warszawa: PWN.

— (1993a), *Synergetic Studies in Polish*. In: R. Köhler, B.B. Rieger (1993) (red.), *Contributions to Quantitative Linguistics*. Dodrecht: Kluwer, 331–359.

— (1993b), *O statystycznych prawach językowych*. Warszawa: Zakład Semiotyki Logicznej Uniwersytetu Warszawskiego.

HENDRICKSON G.L. (1899), Book review of: C.E. Bennet (1898). What was ictus in Latin Prosody? *American Journal of Philology*, 19/4, 361–383. In: *American Journal of Philology*, 20/2, 198–210.

HERDAN G. (1960), *Type-token Mathematics*. The Hague: Mouton.

— (1966), *The Advanced Theory of Language as Choice and Chance*. Berlin itd.: Spriger Verlag (pierwsze wydanie *Language as Choice and Chance*, Groningen: Mouton, 1956).

HILL B.M. (1982), A theoretical derivation of the Zipf (Pareto) law. In: H. Guiter, M.V. Arapov (1982) (red.), *Studies on Zipf's Law*. Bochum: Brockmayer, 53–64.

HOUSEHOLDER F.W. (1960), First Thoughts on Syntactic Indices. In: *International Journal of American Linguistics* 26/3, 195–203.

HŘEBÍČEK L. (1994), Fractals in Language. In: *Journal of Quantitative Linguistics* 1, 82–86.

— (1995), *Text Levels*. Trier: WVT.

— (1997), *Lectures on Text Theory*. Prague: Oriental Institute.

HŘEBÍČEK L., ALTMANN G. (1993), *Prospects of Text Linguistics*. In: L. Hřebíček, G. Altmann (1993) (red.), *Quantitative Text Analysis*. Trier: WVT, 1–28.

HUG M. (1979), *La distribution des phonèmes en français. Die Phonemverteilung im Deutschen. Essais statistiques*. Genève: Slatkine.

ISAAC L.W., GRIFFIN L.J. (1989), Ahistoricism in Time-Series Analyses of Historical Process: Critique, Redirection and illustrations from U.S. Labor History. In: *American Sociological Review* 1989/54, 873–890.

JAKOBSON R. (1962), *Selected Writings I: Phonological Studies*. The Hague: Mouton and Co.

JASSEM W. (1974), *Mowa a nauka o łączności*. Warszawa: PWN.

JELÍNEK J., BEČKA J.V., TĚŠITELOVÁ M. (1961), *Frekvence slov, slovních druhů a tvarů v českém jazyce*. Praha: Státní Pedagogické Nakladitelství.

JOB U. (1981), *Annotated Bibliography of the Statistical study Hexameter Verse*. In: R. Grotjahn (1981) (red.), *Hexameter Studies*. Bochum: Brockmeyer, 226–262.

JUILLAND A., BRODIN D., DAVIDOVITCH C. (1971), *Frequency Dictionary of French Words*. The Hague: Mouton.

JUILLAND A., CHANG-RODRIGUEZ E. (1964), *Frequency Dictionary of Spanish Words*. The Hague: Mouton.

KALLMEYER W., SCHÜTZE F. (1976), Konversationsanalyse. In: *Studium Linguistik* 1, 1–28.

KAUMANNS W., SCHWIBBE M.H. (1989), *Strukturmerkamale von Primatensozietäten unter dem Gesichtspunkt der Menzerathschen Regel*. In: G. Altmann, M.H. Schwibbe (1989), *Das Menzeratsche Gesetz in informationsverarbeitenden Systemen*. Hildersheim itd.: Olms, 99–107.

KÖHLER R. (1983), Markov-Ketten und Autokorrelation in der Sprach- und Textanalyse. In: *Glottometrika* 5, 134–167.

— (1986), *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

— (1993), *Synergetic Linguistics*. In: R. Köhler, B.B. Rieger (1993) (red.), *Contributions to Quantitative Linguistics*. Dodrecht: Kluwer Academic Publishers, 41–51.

— (1995), *Bibliography of Quantitative Linguistics*. Amsterdam: John Benjamins.

KÖHLER R., GALLE M. (1993), *Dynamic Aspects of Text Characteristics*. In: L. Hřebíček, G. Altmann (1993) (red.), *Quantitative Text Analysis*. Trier: WVT, 46–53.

KÖHLER R., MARTINÁKOVA-RENDEKOVÁ Z. (1998), *A systems theoretical approach to language and music*. In: G. Altmann, W.A. Koch (1998) (red.), *Systems*. Berlin, New York: Walter de Gruyter, 514–546.

KOŁMOGOROW A.N. (1965), Zamecanija po povodu analiza rifma „Stixov o sovetskom pasporte" Majakovskogo. In: *Voprosy jazykoznanija* 1965/3, 70–75.

KOŁMOGOROW A.N., PROCHOROW A.V. (1963), O dolnike sovremennoj russkoj poezii (Obščaja xarakteristika). In: *Voprosy jazykoznanija*, 1963/6, 84–95.

— (1964), O dolnike sovremennoj russkoj poezii. In: *Voprosy jazykoznanija* 1964/1, 75–94.

KONDRATOW A.M. (1963), Statistika tipov russkoj rifmy. In: *Voprosy Jazykoznanija* 1963/6, 96–106.

— (1965), *Czterostopowy jamb N. Zabłockiego i niektóre zagadnienia statystyki wiersza*. In: M.R. Mayenowa (1965) (red.), *Poetyka i matematyka*. Warszawa: Państwowy Instytut Wydawniczy, 97–111.

KOPCZYŃSKA Z., PSZCZOŁOWSKA L. (1968), Z zagadnień struktury językowej polskiego sylabowca. In: *Pamiętnik Literacki* 59/2, 183–193.

KOROLKO M. (1990), *Sztuka retoryki*. Warszawa: Wiedza Powszechna.

KRAJEWSKI W. (1998), *Prawa nauki*. Warszawa: Książka i Wiedza (wydanie drugie).

KRASNOPEROVA M.A. (1987), The Relationship between Degrees of Contrast in Rhythmic Structures. In: *Glottometrika* 8, 140–156.

KRYLOV JU.K. (1982), *Eine Untersuchung statistischer Gesetzmäßigkeiten auf der paradigmatischen Ebene der Lexik natürlicher Sprachen*. In: H. Guiter, M.V. Arapov (1982) (red.), *Studies on Zipf's law*. Bochum: Brockmayer, 234–262.

KUHN T.S. (1968), *Struktura rewolucji naukowych*. Warszawa: PWN (z angielskiego przełożyła H. Ostromęcka).

KURCZ I., LEWICKI A., SAMBOR J., SZAFRAN K., WORONCZAK J. (1990), *Słownik frekwencyjny polszczyzny współczesnej*. Kraków: Polska Akademia Nauk, Instytut Języka Polskiego.

LEBART L., SALEM A. (1994), *Statistique textuelle*. Paris: Dunod.

LEES R.B. (1953), The basis of glottochronology. In: *Language* 1953/2, b.d.

LEO F. (1905), Der Saturnische Vers. In: *Abhandlungen der Gesellschaft der Wissenschaften zu Göttingen*, Neue Folge 8/5.

LEVIN JU.I. (1967), O količestvennyx xarakteristikax raspredelenija simvolov v tekste. In: *Voprosy Jazykoznanjia* 6, 112–121.

LEVÝ J. (1965), *W sprawie ścisłych metod analizy wiersza*. In: M.R. Mayenowa (1965), *Poetyka i matematyka*. Warszawa: Państwowy Instytut Wydawniczy, 23–71.

— (1971), *Bude literární věda exaktní vědou?* Praha: Československý spisovatel.

LEWICKI Z. (1987), *Rudolf Clausius i Herman Melville – entropia jako pojęcie literackie*. In: *Od kodu do kodu*. Prace ofiarowane profesorowi Olgierdowi Adrianowi Wojtasiewiczowi na 70-lecie jego urodzin. Warszawa: Wydawnictwo Uniwersytetu Warszawskiego, 279–286.

LOGAN H.M. (1982), The computer and metrical scansion. In: *The ALLC Journal* 3/1982, 9–14.

LORD A.B. (1964), *The Singer of Tales*. Cambridge Mass.: Harvard Univ. Press.

LUTOSŁAWSKI W. (1897), *The origin and growth of Plato's logic*. London: b.d. Reprint Hildesheim itd.: Olms, 1983.

LYONS J. (1976), *Wstęp do językoznawstwa*. Warszawa: PWN (z angielskiego przełożył K. Bogacki).

ŁUŻNY R. (1993), *Wstęp*. In: A. Puszkin (1993), *Eugeniusz Oniegin*. Wrocław itd.: Ossolineum, V–CXXII.

MAJEWICZ A.F. (1989), *Języki świata i ich klasyfikacja*. Warszawa: PWN.

MAŃCZAK W. (1996), *Problemy językoznawstwa ogólnego*, Wrocław itd.: Ossolineum.

MARCKWORTH M., BELL L. (1967), *Sentence-length distribution in the Brown Corpus*. In: H. Kučera, W.N. Francis (1967) (red.), *Computational analysis of present-day American English*. Providence: Brown University Press, 368–405.

MARKOV A.A. (1910), Recherches sur un cas remarquable d'épreuves dépendantes. In: *Acta Mathematica* 33, 87–104.

— (1907), Issledovanie zamečateľnogo slučaja zavisimyh ispytanij. In: *Bulletin de l'Académie Impériale des Sciences* 6/1, 61–80.

— (1911), Ob odnom slučae ispytanij, svjazannych v složnuju cep'. In: *Bulletin de l'Académie Impériale des Sciences* 6/5, 171–186.

— (1913), Primer statističeskogo issledovanija nad tekstom „Evgenija Onegina", illjustrirujuščij svjaz' ispytanij v cep'. In: *Bulletin de l'Académie Impériale des Sciences* 6/7, 153–162.

— (1961), *Izbrannye Trudy. Teorija čisel. Teorija verojatnostej*. Moskva: Izdateľstvo Akademii Nauk SSSR.

MATHESIUS V. (1939), O tak zvaném aktuálním členění věty. In: *Slovo a Slovesnost* 5, 171–174.

— (1982), *Jazyk, kultura a slovesnost*. Praga: Odeon.

MAY S. (1963), *0 pewnych właściwościach statystycznych języka polskiego*. In: A.M. Jagłom, J.M. Jagłom (1963), *Prawdopodobieństwo i informacja*. Warszawa: Książka i Wiedza, 368–372.

MAYENOWA M.R. (1963), *Wiersz*. In: M.R. Mayenowa (1963) (red.), *Poetyka. Zarys encyklopedyczny*. Tom II, część I – *Rytmika* (red. J. Woronczak). Wrocław itd.: Ossolineum.

— (1965) (red.), *Poetyka i matematyka*. Warszawa: Państwowy Instytut Wydawniczy.

— (1971) (red.), *O spójności tekstu*. Wrocław itd.: Ossolineum.

— (1979), *Poetyka teoretyczna*. Wrocław itd.: Ossolineum.

MCCLEARY R., HAY R. (1980), *Applied Time Series Analysis for the Social Sciences*. Beverly Hills itd.: Sage Publications.

MILEWSKI T. (1965), *Językoznawstwo*. Warszawa: PWN.

MILLER G.A., CHOMSKY N. (1963), *Finitary models of language users*. In: R.D. Luce, R.R. Bush, E. Galanter (1963) (red.), *Handbook of Mathematical Psychology* vol.II. New York: Wiley, 419–491.

MOLES A. (1958), *Théorie de l'information et perception esthétique*. Paris: Flammarion.

MONTGOMERY D., JOHNSON L. (1976), *Forecasting and time series analysis*. New York itd.: McGraw-Hill.

MOOD A.M. (1940), The Distribution Theory of Runs. In: *Annals of Mathematical Statistics* 11/1940, 367–392.

MORICE E., CHARTIER F. (1954), *Méthode statistique. Deuxième partie: analyse statistique*. Paris: Imprimerie Nationale.

MOROZOW N.A. (1915), Lingwističeskije spektry. In: Izvestja Otdelenija Russkogo Jazyka i Slovesnosti, *Bulletin de l'Académie Impériale des Sciences* 20/3, 93–134.

MULLER CH., BRUNET E. (1988), La statistique résout–elle les problèmes d'attribution? In: *Strumenti Critici*, n.s., a.III, n.3, settembre 1988, 367–387.

MYŚLIWIEC H. (1959), *Zarys wersyfikacji łacińskiej średniowiecza*. In: M. Dłuska, W. Strzelecki (1959) (red.), *Metryka grecka i łacińska*. Wrocław itd.: Ossolineum.

NAGY C. (1974), *Comparative Studies in Greek and Indic Meter.* Cambridge (Mass.): Harvard University Press.

NEUMAN J. VON (1941), Distribution of the ratio of the mean square difference to the variance. In: *Annals of Mathematical Statistics* 1941/12, 367–395.

NURIUS P.S. (1983), Methodological Observations on Applied Behavioral Science. In: *The Journal of Applied Behavioral Science* 19(3), 215–228.

OPPENHEIM R. (1988), The mathematical analysis of style: a correlation-based approach. In: *Computers and the Humanities* 22 (1988), 241–252.

ORLOV JU.K., VOLOŠIN B.A. (1982), Das verallgemeinerte Zipf-Mandelbrotsche Gesetz und die Verteilung der Anteile von Farbflächen in der Malerei. In: Ju.K. Orlov, M.G. Boroda, I.S. Naderejšvili (1982) (red.), *Sprache, Text, Kunst*. Bochum: Brockmeyer, 263–270.

ORLOVA L.V., PEREBIJNIS V.S. (1981) (red.), *Častotnij slovnik sučasnoï ukaïnskoï xudožnoï prozy*. Kiev: Vidavnictvo Naukova Dumka.

PÄÄKKÖNEN M. (1993), Graphems in Context: Statistical Data on the Graphology of Standard Finish. *Glottometrika* 14, 1–53.

PALMER F.R. (1970) (red.), *Prosodic analysis*. London: Oxford University Press. Coll. Language and Language Learning vol. 25.

PAPP F. (1967), Eine Bearbeitung des ungarischen Wortschatzes auf Lochenkartenmachinen. In: *Acta Linguistica Academiae, Scientarum Hungaricae* 17, b.s.

PAWŁOWSKI A. (1994), Ein problem der Klassischen Stilforschung: Vergleich einiger Indikatoren des Lexikonumfangs. In: *ZET – Journal of Empirical Text Research* 1, 67–74.

— (1997), Time-Series Analysis in Linguistics. Application of the ARIMA Method to Some Cases of Spoken Polish. In: *Journal of Quantitative Linguistics* 4(1–3), 203–221.

— (1998), *Séries temporelles en linguistique. Avec application à l'attribution de textes: Romain Gary et Emile Ajar*. Paris: Champion.

— (1999), Language in the line vs. language in the mass: On the efficiency of sequential modelling in the analysis of rhythm. In: *Journal of Quantitative Linguistics* 6(1), 70–77.

— (2000a), Analyse quantitative comparée de la prosodie des langues à accent fixe et à accent libre. In: M. Rajman, J.C. Chappelier (2000) (red.), *JADT 2000, Actes des 5es journées internationales d'analyse statistique des données textuelles*. Lausanne: EPFL, 531–534.

— (2000b), Sequential approach to prosody: contrastive analysis of fixed- and free-accent languages: examples of Polish and Russian. In: *35 Linguistisches Kolloquium, Universität Innsbruck, 20–22 September 2000*, 41.

PAWŁOWSKI A., EDER M. (2000), Quantity or stress? Sequential analysis of Latin prosody. In: Proceedings of the fourth conference on the International Quantitative Linguistics Association *QUALICO 2000*. Prague: b.w.

PEITGEN H.-O., JÜRGENS H., SAUPE D. (1997), *Granice chaosu. Fraktale*. Warszawa: PWN (z angielskiego przełożyły K. Pietruska-Pałuba, K. Winkowska-Nowak).

PETRUSZEWYCZ M. (1981), *Les chaînes de Markov dans le domaine linguistique*. Genève: Slatkine.

PODBIELSKI H. (1988), patrz *Arystoteles*.

PODSIAD A., WIĘCKOWSKI Z. (1983) (red.), *Mały słownik terminów i pojęć filozoficznych*. Warszawa: Instytut Wydawniczy PAX.

POPPER K.R. (1996), *Wszechświat otwarty*. Kraków: Znak (z języka angielskiego przełożył A. Chmielewski).

PORĘBSKI M. (1986), *Sztuka a informacja*. Kraków: Wydawnictwo Literackie.

POSTAL P.M. (1968), *Aspects of Phonological Theory*. New York: Harper and Row.

PRIESTLEY M.B. (1981), *Spectral Analysis and Time Series*, London itd.: Academic Press.

PSZCZOŁOWSKA L. (1965), *Długość wiersza a budowa zdania*. In: M.R. Mayenowa (1965) (red.), *Poetyka i matematyka*. Warszawa: Państwowy Instytut Wydawniczy, 79–96.

QUENOUILLE M.H. (1947), A large-sample of autoregressive schemes. In: *Journal of the Royal Statistical Society* 110, 123–129.

RAO C.R. (1994), *Statystyka i prawda*. Warszawa: PWN (z angielskiego przełożyli M. Abrahamowicz i M. Męczarski).

RAPOPORT A. (1982), *Zipf's Law Revisited*. In: H. Guiter, M.V. Arapov (1982) (red.), *Studies on Zipf's law*. Bochum: Brockmayer, 1–28.

RICKHEIT G. (1991) (red.), Kohärenzprozesse. Modellierung von Sprachverarbeitung in Texten und Diskursen. Opladen: Westdeutsche Verlag.

ROBERTS A. (1996), Rhythm in Prose and the Serial Correlation of Sentence Lengths: a Joyce Cary Case Study. In: *Literary and Linguistic Computing* (ALLC) 11(1), 33–39.

ROTHE U. (1986), *Die Semantik des kontextuellen „et"*. Frankfurt a.M. itd.: Lang.

SĄDEJOWA H. (1959), *Zarys metryki greckiej*. In: M. Dłuska, W. Strzelecki (1959) (red.), *Metryka grecka i łacińska*. Wrocław itd.: Ossolineum.

SAFAREWICZ J. (1988), *Języki italskie*. In: L. Bednarczuk (1988) (red.), *Języki indoeuropejskie* t.2. Warszawa: PWN, 515–570.

SALEM A. (1988), Approches du temps lexical, statistique textuelle et séries chronologiques. In: *Mots* 1988/17, 105–143.

SALONI Z. (1971), *Definicja spójności tekstu*. In: M.R. Mayenowa (1971) (red.), *O spójności tekstu*. Wrocław itd.: Ossolineum, 89–94.

SAMBOR J. (1969), *Badania statystyczne nad słownictwem. Na materiale „Pana Tadeusza"*. Wrocław, Warszawa itd.: Ossolineum.

— (1972), *Słowa i liczby. Zagadnienia językoznawstwa statystycznego*. Wrocław itd.: Ossolineum.

— (1988), Lingwistyka kwantytatywna – stan badań i perspektywy rozwoju. In: *Biuletyn Polskiego Towarzystwa Językoznawczego* 41, 47–67.

— (1989), Struktura kwantytatywna wyrazów polisemicznych w słowniku, czyli o tzw. prawie Kryłowa. In: *Polonica* XIV, 13–31.

— (1997) (red.), *Z zagadnień kwantytatywnej semantyki kognitywnej*. Warszawa: Polskie Towarzystwo Semiotyczne.

SAUSSURE F. DE (1991), *Kurs językoznawstwa ogólnego*. Warszawa: PWN (z francuskiego przełożyła K. Kasprzyk).

SCHENKEIN J. (1978) (red.), *Studies in the Organisation of Conversational Interaction*. New York: Academic Press.

SCHILS E., DE HAAN P. (1993), Characteristics of sentence length in running text. In: *Literary and Linguistic Computing* 8/1, 1993, 20–26.

SCHMIEL R. (1981), *Rhythm and accent: texture in Greek epic poetry*. In: R. Grotjahn (1981) (red.), *Hexameter Studies*. Bochum: Brockmeyer.

SHANNON C. (1948), The Mathematical Theory of Communication. In: *Bell System Technical Journal* 27, 379–423.

SIEWIERSKA A. (1988), *Word Order Rules*. London itd.: Croom Helm.

— (1997) (red.), *Constituent Order in the Languages of Europe*. Berlin itd.: Mouton de Gruyter.

SILNITSKY G. (1993), Typological Indices and Language Classes: a Quantitative Study. In: *Glottometrika* 14, 139–160.

SKINNER B.F. (1941), A quantitative estimate of certain types of sound-patterning in poetry. In: *American Journal of Psychology* 54, 64–79.

SMITH J.B., ROSENBERG B.A. (1973), Rhythms in Speech: the Formulaic Structure of Four Fundamentalist Sermons. In: *Computer Studies in the Humanities and Verbal Behahior* 4, 166–173.

SOBCZYK M. (1996), *Statystyka*, Warszawa: PWN.

SOKAL A., BRICMONT J. (1988), *Intellectual impostures*. London: Profile Books.

STANISZ T. (1993), *Funkcje jednej zmiennej w badaniach ekonometrycznych*. Warszawa: PWN.

STEWART I.N. (1996), *Czy Bóg gra w kości?* Warszawa: PWN (z angielskiego przełożyli M. Tempczyk, W. Komar).

STIER W. (1989), Basic Concepts and New Methods of Time Series Analysis in Historical Research. In: *Historical Social Research/Historische Sozialforschung* 14(1), 3–24.

STRAUSS U., SAPPOK CH., DILLER H.J., ALTMANN G. (1984), Zur Theorie der Klumpung von Textentitaeten. In: *Glottometrika* 7, 73–100.

STROHNER H., RICKHEIT G. (1990), Kognitive, kommunikative und sprachliche Zusammenhänge. Eine systemtheoretische Konzeption linguistischer Kohärenz. In: *Linguistische Berichte* 125, 3–23.

SWADESH M. (1952), Lexico-Statistic Dating of Prehistoric Ethnic Contacts, with Special Reference to North American Indians and Eskimos. In: *Proceedings of the American Philosophical Society* XCVI, 452–463.

SWADESH M. (1953), Archeological and Linguistic Chronology of Indoeuropean Groups. In: *American Anthropologist* LV, 349–352.

— (1955), Towards Greater Accuracy in Lexicostatistic Dating. In: *Internationa Journal of American Linguistics* 21, 121–137.

SZANIAWSKI K. (1987), *Informacja*. In: *Filozofia a nauka. Zarys encyklopedyczny*. Wrocław itd.: Ossolineum, 244–251.

TRUBECKOJ N.S. (1939), *Grundzüge der Phonologie*. Prague: Travaux du Cercle Linguistique de Prague 7.

TULDAVA J. (1995), On the Relation between Text Length and Vocabulary Size. In: J. Tuldava (1995) (red.), *Methods in Quantitative linguistics*. Trier: WVT, 131–150.

TUWIM J. (1937), *Lutnia Puszkina*. Warszawa: J. Przeworski. Cytaty z antologii *Julian Tuwim. Dom mój: cztery ściany wiersza*. Red. Z. Fedecki, Warszawa 2000, Świat Książki, 377–396.

TWEEDIE F.J., BAAYEN R.H. (1998), How Variable May a Constant be? Measures of Lexical Richness in Perspective. In: *Computers and the Humanities* 32/1998, 323–352.

URBANEK A. (1987), *Redukcjonizm*. In: *Filozofia a nauka. Zarys encyklopedyczny*. Wrocław itd.: Ossolineum, 564–576.

VASJUTOČKIN G.S. (1987), Das rhytmische System der „Aleksandrinischen Gesänge“ In: *Glottometrika* 8, 178–191.

WĄSIK Z. (1987), *Semiotyczny paradygmat językoznawstwa*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego.

WEAVER W., SHANNON C. (1949), *The Mathematical Theory of Communication*. Illinois: Urbana.

WEBER M. (1973), *Gesammelte Aufsatze zur Wissenschaftslehre*. Tübingen: J.C.B. Mohr.

WEST M.L. (1970), A new approach to Greek prosody. In: *Glotta* 48, 185.

WHITELEY P. (1980), Time Series Analysis. In: *Quality and Quantity* 14, 225–247.

WIENER N. (1948), *Cybernetics or control and communication in the animal and the machine*. Cambridge (Mas.): MIT Press.

WIKARJAK J. (1978), *Gramatyka opisowa języka łacińskiego*. Warszawa: PWN.

WILLIAMS C.B. (1970), *Style and vocabulary: numerical studies*. London: Griffin.

WIOLAND F. (1985), *Les structures syllabiques du français*. Genève: Slatkine-Champion.

WORONCZAK J. (1965), *Rytmika akcentowa sylabowca* In: M.R. Mayenowa (1965) (red.), *Poetyka i matematyka*. Warszawa: Państwowy Instytut Wydawniczy, 72–78.

— (1967), *On an attempt to generalize Mandelbrot's distribution*. In: *To Honor Roman Jakobson*, vol.2. The Hague: Mouton, 2254–2268.

— (1976), O statystycznym określeniu spójności tekstu. In: M.R. Mayenowa (1976) (red.), *Semantyka tekstu i języka*. Wrocław itd.: Ossolineum, 165–173.

XANTOS A. (2000), Entropizer 1.1: un outil informatique pour analyse séquentielle. In: M. Rajman, J.C. Chappelier (2000) (red.), *JADT 2000, Actes des 5es journées internationales d'analyse statistique des données textuelles*. Lausanne: EPFL, 357–364.

YOKOYAMA S., ITAHASHI S. (1980), On the Distance of Japanese Words Based on Distinctive Features and a Second-Order Model. In: *Glottometrika* 3, 62–81.

ZASORINA L.N. (1971) (red.), *Častotnyj slovar' russkogo jazyka*. Moskva: Izdatelstvo Russkij Iazyk.

ZIOMEK J. (1990), *Retoryka opisowa*. Wrocław itd.: Ossolineum.

ZÖRNIG P. (1984a), The Distribution of the Distance Between Like Elements in a Sequence (I). In: *Glottometrika* 6, 1–13.

— (1984b), The Distribution of the Distance Between Like Elements in a Sequence (II). In: *Glottometrika* 7, 1–14.

INDEKS NAZWISK

Czcionką zwykłą zaznaczono wystąpienia w tekście głównym, czcionką pochyłą w przypisie. Jeżeli nazwisko pojawiło się na tej samej stronie w tekście głównym i w przypisie, podajemy tylko odnośnik do tekstu głównego. Indeks nie obejmuje bibliografii i aneksu.

ADAM Jean-Michel 30
AJAR Emile 51–53
ALTMANN Gabriel 5, *6*, 8, 14–17, *19*, 23, 33, 131–132
AMSTERDAMSKI Stefan *8*
ARAGON Louis 52
ARAPOW Michał W. 15, *16*, *23*
ARYSTOTELES 27, 29
ATKINSON John M. 30
AUGUSTYN św. 37
AZAR M. *10*, 13, 47, 77
BAAYEN Harald R. *37*
BAEVSKIJ Vadim S. *55*
BARTLETT Maurice S. 71
BATAGELJ Vladimir 113, 132
BAUDOUIN DE COURTENAY Jan N. 46
BAVAUD François 10, 13, *41*, 56–59
BELL L. *8*
BELLERT Irena 34
BENVENISTE Emile 30
BEÖTHY Erszébet *19*
BERTALANFFY Ludwig von 24
BLANCHE Lesley *52*
BLOOMFIELD Peter *82*
BORODA Mojsej G. *55*
BORTOLINI Umberta *10*, *132*, *134*
BOURSIN Jean-Louis 70
BOX George 47, 49, 63, *64*, 66–68, 70
BRAINERD Barron 29, 45
BRATLEY Paul *10*, 13, 47, 77
BRÉAL Michel 28
BRICMONT Jean *74*
BROCKWELL Peter *63*
BRUNET Etienne 47, 54
BRZECHWA Jan 77, 80, 92, 112, 116
BÜHLER Karl 73
BUŁHAKOW Michał 99, *103*, 104, 113–116
BUNGE Mario *8*
BUZZATI Dino 50
CAIRNS Francis *77*
CHAGHAGHI Francois *63*
CHARTIER Fernand *32*
CHERC Maja M. *16*
CHLEBNIKOW Wiktor 44–45
CHMIELEWSKI Adam 24
CHOMSKY Noam *42*, 45
CHRISHOLM David *77*
CORDUAS Marcella 49–50
COSSETTE André *37*
COUTROT Bernard *63*
CRYSTAL David 118
DAMERAU Frederic J. *42*
DANEŠ František 29
DAVIES Richard *63*
DIJK Teun A. van 30
DILLON M. 32–33
DITTMANN Jürgen 30
DREHER John 47, 118
DRESSLER Wolfgang U. 33
DROESBEKE Jean-Jacques *63*
DUSZAK Anna 30, 34
ELMAN Jeffrey 10
ENNIUSZ 120
EROMS Hans-Werner 29
ESTOUP J.B. 15
FELLER William *42*, 59, *63*

FENK August 18
FENK-OCZLON Gertraud 18
FIRTH John R. 29
FISHMAN George S. 13
FISIAK Jacek *29*
FOURIER Joseph *32*
FUCKS Wilhelm *55*
FULGENCJUSZ św. 37
GALLE Matthias 10
GARY Romain 51–52
GASPARSKI Wojciech 24
GATNAR Eugeniusz 13
GLASS Gene V. *63*
GLEICK James *67*
GOOD Irving J. 36
GOTTMAN John M. *41*, *63*, 82
GREENBERG Joseph H. 28, *97*, 131
GREŃ Jerzy *13*, *56*, *64*, *94*, 114
GROTJAHN Rüdiger 33, 36, *55*, 56
GUILBAUD Georges Th. 10, 13
GUIRAUD Pierre 7, *23*
GUITER Henri 15
HAAN Petr de 48–49, 130
HAKEN Hermann *8*, 24
HAMMERL Rolf *8*, 13, 15, 17–19
HAY Richard A. *63*, 68
HEMINGWAY Ernest 51
HENDRICKSON G.L. 120
HERDAN Gustaw 31–33, *36*, 45, 48
HERITAGE John 30
HILL B.M. 23
HOFFMANN Lothar 40–41
HOMER 47
HORACY *59*, 120–121
HOUSEHOLDER Fred W. 131
HŘEBÍČEK Luděk *6*, 48
HUG Marc *55*
HUGO Victor 54
ITAHASHI Shuichi *10*
IWASZKIEWICZ Jarosław 99, *101*, 102, 104, 113–116
JAKOBSON Roman 30
JASSEM Wiktor *31*, *38*
JELÍNEK Jaroslav *132*
JENKINS Gwilym 47, 49, 63, *64*, 66–68, 70
JOHNSON Lynwood *63*, *71*
JOYCE James 51
JUILLAND Alphonse *132*
KAC Marek *73*
KALLMEYER Werner 30
KAUMANNS W. 23
KEDEM Benjamin *10*, 13, 47, 77
KÖHLER Reinhardt 7–8, 10, 14–15, 23, *28*, 30, 42, 46, *53*, 129
KOŁMOGOROW Andrej N. *10*, *28*, 39
KONDRATOW Aleksander *28*
KONWICKI Tadeusz 133
KOPCZYŃSKA Zdzisława *78*, *105*
KOROLKO Mirosław 27
KRAJEWSKI Władysław 25, 73
KRASNOPEROVA M.A. *55*
KRYŁOW Jurij K. 18
KUHN Thomas S. 74
KURCZ Ida *132*, *134*
LAMARTINE Auguste 54
LAPLACE Pierre S. 53
LEC Jerzy *37*
LEES Robert B. 16
LEHFELDT Werner 132
LEO Friedrich 119
LEVIN Jurij I. 33, 34–37
LEVÝ Jiří *28*, *98*
LEWICKI Zbigniew 41
LOGAN H.M. *77*
LORD Albert B. *47*
LUTOSŁAWSKI Wincenty *46*

LYONS John 28–29
ŁUŻNY Ryszard 99
MAJEWICZ Alfred F. 97
MALINOWSKI Bronisław 29
MANDELBROT Benoit 15
MANZONI Alessandro 50
MAŃCZAK Witold *17*
MARCKWORTH M. *48*
MARKOW Andrej A. 27, 31, 33, 42–46, 53
MARTINÁKOVÁ-Rendeková Zuzana 23
MATHESIUS Vilém 29
MAY Stanisław 38–40
MAYENOWA Maria R. 68, 117
MCCLEARY Richard *63*, 68
MICKIEWICZ Adam 110–111, 113–114, 116
MILEWSKI Tadeusz 97, 128–129
MILLER George *42*, 45
MOLES Abraham 41
MONTGOMERY Douglas *63*, *71*
MOOD Alexander M. 56
MORAVIA Alberto 10, 133, *135*
MORICE Eugène *32*
MOROZOW N.A. 46
MULLER Charles 47, 54
MUSSET Alfred de 54
MYŚLIWIEC Herbert 119
NAGY C. *120*
NEUMAN John von 49
NEWERLY Igor 78, *89*, 116
NURIUS Paula S. 63
OPPENHEIM Rosa 50–51, 130
ORLOV Jurij K. 23
ORLOVA L.V. *132*
OSIPOVA L.Ja. *55*
OWIDIUSZ 120–121
PÄÄKKÖNEN Matti *55*
PALMER Frank R. 29
PAPP Laszlo 18
PARRY Milman 47
PAVESE Cesare 50
PAWŁOWSKI Adam 10, 13–14, 32, *37*, *38*, 51, 53, *55*, *58*, *63*, 64, *75*, *93*, *97*, *118*, 130–133, 139
PEITGEN Heinz-Otto *67*, *73*
PETRUSZEWYCZ Micheline 10, 13, 27, 42–46
PIOTROWSKI Rajmund G. 40–41
PODBIELSKI Henryk *27*, *92*
PODSIAD Antoni 8
POPPER Karl 73
PORĘBSKI Mieczysław 41
POSTAL Paul M. 120
PÓLYA George *57*
PRIESTLEY Maurice B. 64, *82*
PROCHOROW A. *10*, *28*
PSZCZOŁOWSKA Lucylla *78*, *105*, 112
PUSZKIN Aleksander 27, 42, 44–45, 92, 100, 113–114, 116
QUENEAU Raymond 52
QUENOUILLE Maurice H. 71
RAO C. Radhakrishna *57*, *73*
RAPOPORT Anatol 23–24
RICKHEIT Gert 34
ROBERTS Alan 14, 48, 130
ROSENBERG Bruce A. 47
ROSS Donald *10*, 13, 47, 77
ROTHE Ursula *19*
SAFAREWICZ Jan 119
SAINT-EXUPÉRY Antoine de 19
SALEM André *48*, 49
SALONI Zygmunt 34
SAMBOR Jadwiga 8, 13–19, 21–23, 38, 40–42, *58*, *59*, 73, *93*, *94*, 114, 134
SAPIR Eduard 132
SAUSSURE Ferdinand de 7, 28, 30, 32
SĄDEJOWA Helena 121
SCHENKEIN Jim 30

SCHILS Erik 48–49, 130
SCHMIEL Robert *28*
SCHÜTZE Fritz 30
SCHWIBBE Michael *17*, 23
SHANNON Claude E. 10, 31, 38–39, 41, 131
SIEWIERSKA Anna 28
SILNITSKY George 132
SKINNER Burrhus F. 33, *55*
SŁOWACKI Juliusz 78, 83, 113, 116
SMITH John B. 47
SOBCZYK Mieczysław *94*
SOKAL Alan *74*
STANISZ Tadeusz *16*
STEWART Ian N. *67*
STIER Winfried *63*
STRAUSS Udo 36
STROHNER Hans 34
SWADESH Maurice *16*
SZANIAWSKI Klemens *38*, 41
SZCZYPIORSKI Andrzej 133
SZYMCZAK Mieczysław *21*
TOURNIER Michel *52*
TRUBECKI Mikołaj 30
TUKEY John W. *39*
TULDAVA Juhan 10
TUWIM Julian 98
TWEEDIE Fiona J. *37*
URBANEK Adam 55, 73
VASJUTOČKIN G.S. *28*, 36
VOLOŠIN B.A. 23
WAŻYK Adam 99, 105
WĄSIK Zdzisław 7, 74
WEAVER Warren 38
WEBER Max 73
WERGILIUSZ 121–122
WEST Martin L. 120
WHITELEY Paul *63*
WIENER Norbert 29
WIĘCKOWSKI Zbigniew 8
WIKARJAK Jan 118
WILLIAMS Carrington B. 33
WIOLAND François *55*
WOJTYŁA Karol 79, 87, 116
WORONCZAK Jerzy 15, *28*, 36–37
XANTOS Aris 10, *11*, 13, *57*
YOKOYAMA Shochi *10*
YULE George U. *50*
ZASORINA L.N. *132*
ZIOMEK Jerzy 41
ZIPF George K. 15–16, 18
ZÖRNIG Peter 36

INDEKS RZECZOWY

Czcionką zwykłą zaznaczono wystąpienia w tekście głównym, czcionką pochyłą w przypisie. Indeks nie obejmuje bibliografii i aneksu.

ACF, patrz *autokorelacji funkcja*

akcent
- – dynamiczny *26*, 27, 48, 60–62, 72, 80, 83, 85–114, 118–128
- – emocjonalny 76, 79
- – logiczny 76
- – metryczny, patrz *ikt*
- – paroksytoniczny 79, 98, 121
- – poboczny 6, *93*
- – stały 97–99, 104, 109, 113
- – swobodny i ruchomy 97–99, 104, 113
- – toniczny 97

akut *97*

analiza
- – sekwencyjna 5, 7, 9, 10, 13, 14, 25–33, 35–39, 42, 46–48, 53, 55–63
- – szeregów czasowych 29, 32, 42, 47–56, 63–72, 78, 82, 105, 120, 133
- – w dziedzinie czasu 82, 85, 88, 100
- – w dziedzinie częstotliwości 82, 85
- – widmowa 13, 31, 47, 75, 82
- – wielowymiarowa 33, 54, 113

AR, patrz *proces*

ARIMA metoda 49–56, 63–64, 75, 86, 92, 96, 100, 114, 117, 123, 134

ARMA, patrz *proces mieszany*

arsa 120–121

autokorelacji cząstkowej funkcja (PACF) 66, 70–71, 80, 84, 86–87, 89–90, 100–103, 106–107, 109, 123–124, 136–137

autokorelacji funkcja (ACF) 65–68, 70–71, 80–81, 84, 86–90, 100–103, 106–110, 123–124, 126–127, 136–137

autokowariancji funkcja 65

autorstwa problem 46–47, 50–54

binarne skale 30–31, 35, 47, 56, 73, 75–77, 94, 117, 120, 123

bit 11, 35, 39

błąd standardowy 71

bogactwo leksykalne 10, 36–37

cecha dystynktywna *10*, 29–31

chaos 67, 74

charakterystyka K *36*

ciąg definicyjny 21, 22

cyrkumfleks *97*

czasu koncepcja 28–29

dedukcyjność 8, *16*, 17, 50

determinizm 13, 25–26, 73–74

dopasowanie modelu 14, 20, 71, 81, 124–126, 137

dyskurs oratorski 75–79, 87–94, 116

dyspersja 35, 43

efekt wersyfikacji 89, 91–92, 113

efektywność modelowania 22, 25, 28, 53, 57, 64, 93–94, 96, 107–110

ekonomii wysiłku zasada 17

entropia 10, 13, 33–34, 38–41, 57–62

estetyczne kategorie 117

falsyfikowalność 6, 8–9, 117

fonem *7*, 9, 11, 16, 24, 29, 30, *31*, 33, 35, 38, 40, 77

fonologia 30, 33, 46–47, 76, 97, 113, *118*

fonotaktyka 38

formant słowotwórczy 21–22

formulaiczna teoria 47

funkcje tekstu
- – estetyczna 27, 99–101, 104, 112, 113, 117, 129
- – komunikacyjna 45, 79, 99, 113, 129
- – perswazyjna 27, 44–45, 79, 104, 129

funkcjonalna perspektywa zdania 29

genetyka 18, 23, 41

geografia kwantytatywna 23
glottochronologia *16*
gniazdo leksykalne *21*
grawis *97*
gry w chaos *73*
heksametr łaciński 59–62, 119–128
hiperonim końcowy 21–22
hipoteza 6–9, 14, 25, 27, 30, 37–38, 43–44, 48–49, 53, 56, 59–62, 73, 87, 92–98, 104, 111, 114, 121, 128–133
idealizacja *7*, 73
ikt 59, 62, 120–121, 128
iloczas *26*, 53, 61–62, 75, 97, 118–128
indefinibilia semantyczne 22
indukcja 17, 108, 134
informacji ilość 10–13, 37–39, 52, 57–58, 133–138
inżynieria językowa 22, 42
jedenastozgłoskowiec 78, 85–86, 92
język
– definicja 6
– angielski 14, 16, 41, 47–48, 52, *98*, 129–131, 139
– chiński 47
– francuski 11, 19–21, 41, 52, 54, 98, 129, 131–133, 139
– łaciński 59, 62, 118–128
– niemiecki 19, 22, 31, 41, 98
– polski 19, 21–22, 38, 40–41, 75–96, 98–117, 133–138
– rosyjski 18, 41, 43–44, 46, 98–115, 132
– włoski 10–11, 50, 79, 98, 132–139
języki
– analityczne 52, 131–133, 136–138
– syntetyczne 131–133, 136–137
klasyfikacja języków 28, 97, 100, 104, 113–114, 131–132, 137–138
kodowanie 9, 11–12, 17, 30–31, 43, 45, 48, 55, 59–63, 75–76, 79, 100, 121, 134
konekcjonizm 10, 23
kontekstowy związek 10, 29, 30, 34, 40–41, 44–45, 48, 53, 55, 58–62, 66–67, 75, 86, 87, 91, 97, 126
kontekstualizm 29
kwantyfikacja 6, 9–10, 12, 27, 30, 39, 42, 46–47, 52, 55, 57, 64, 72–73, 75–76, 121–122, 129, 131–134
langue 31
linearności zasada 7, 26–28, 31, 118
linearność tekstu, patrz *sekwencyjna struktura tekstu*
lingwistyka
– kwantytatywna 5–8
– modelowa 5, 8, 13, 24–26, 28, 32, 42, 46, 91–93, 130
– tekstu 30, 34
lista frekwencyjna 15, 132, 134
losowość 56, 73–74
łańcuch Markowa 10, 27, 31, 33, 42–46
MA, patrz *proces średniej ruchomej*
macierz prawdopodobieństw przejścia 12–13, 59–61
metryka 5, 11, 25, 27, 62, 86, 92–93, 116–123, 128
miary Gooda 36
modelowanie
– numeryczne 13, 46, 54–55, 63–72
– probabilistyczne 10, 12–13, 46, 55–63
– teorioinformacyjne 10, 13, 40
modelu
– addytywność 68, 69, 70, 86
– identyfikacja 65, 67, 70–71, 81, 85, 125
– interpretacja *8*, 9–10, 12, 14, 21–22, 48, 67–68, 72, 76, 80, 82, 85–86, 104, 107–108, 127–128, 134
– multiplikatywność 69, 86
– odwracalność 67
– oszczędność 68
– pojęcie 13
– weryfikacja 8, 9, 14

moment statystyczny 64
mora 125
morfem 6, 9, 19, 21, 25, 29–30, 33, 38, 40, 42, 52, 73, 76, 97, 99, 131–133
odstęp (ang. *gap*) 42, 52
odstęp (ang. *lag*) 48, 60, 64–65, 68–71, 80–84, 88, 90, 102–103, 107–114, 124–127, 136–137
operatorowy zapis *66*, 69
opozycje językowe 7, 28, 31–32, 55, 93, 97, *118*, 120
pamięć w tekście 41, 66
oś paradygmatyczna 7, 31
oś syntagmatyczna 7, 30–31
ośmiozgłoskowiec 77, 82, 86, 105, 116
PACF, patrz *autokorelacja*
paradygmat naukowy 24, 32, 55, 74
parole 31
pauzy międzywyrazowe 11, 43–45, 72, 76, 79
periodogram 75, 80, 82–83, 85–86, 88, 90
perspektywa badawcza 7, 32–33
pokrycie tekstu słownictwem 132
polisemia 18
populacja generalna *7*, 36, 95, 129
prasowo-publicystyczny styl 100–104, 108, 113, 116
prawdopodobieństwo przejścia, patrz *macierz prawdopodobieństw przejścia*
prawo
– językowe 6–9, 14–26, *18*
– Beöthy 19–21
– Kryłowa 18–19
– Martina 21–23
– Menzeratha *16*, 17–18, 23
– Zipfa 15–17, 19, 23
primatologia 18, 23
proces stochastyczny
– autoregresji (AR) 50–54, 66–71, 81, 107, 110–112, 124–126, 128–130
– losowy 43, 49, 53, 56, 66–67
– mieszany (ARMA) 50–51, 53, 68, 71, 81, 125, 129
– sezonowy (SARMA) 68–69, 81–82, 84–86, 92, 107, 110, 112, 125
– średniej ruchomej (MA) 51–51, 66–67, 71, 81, 84–86, 88–92, 100, 102, 107, 111, 124–127, 131, 136–138
prognozowanie 49, 63
proza artystyczna 14, 44–47, 76–79, 89–92, 94, 96, 99, 104, 111, 113–114, 116–117, 133
proza poetycka 114
prozodia tekstu 5, 25–26, 29, 72–73, 75–76, 79, 97, 99, 112–113, 118–119, 129
przekładu analiza 114–115
przestrzeń zdarzeń 13
pseudolosowy szereg czasowy 9, 14, 48, 56
pseudotekst 14, 48
redukcjonizm 55, 73
redundancja 10, 38–42
relewancji zasada 73
remat 30
retoryka 26–27, 79, 92, 129
rewolucja naukowa 8–9
rozkład Waringa 23
równania Yule'a-Walkera 66, 70
rytm prozy 14, 26–27, 47–48, 78, 91, 104, 111, 114,
rząd procesu 45, 50, 59–62, 64, 66–67, 69, 80, 85–86, 89, 101, 114, 137
samoregulacji zasada 6, 17
SARMA, patrz *proces sezonowy*
segmentacja *6*, 25, 27, 29–30, 75–76, 118, 122, 131
sekwencyjna struktura tekstu 25–26
semantyka 6–7, 18–19, 21–22, 30, 33–34, 73, 97, 131
semiotyczne kody 23, 41
sieciowe struktury 22, 32
siły Zipfa 17, 25

skala
– binarna 30–31, 35, 47, 56, 73, 75–77
– ilorazowa 13
– interwałowa 13
– nominalna 13
– porządkowa 13, 76
składnia pozycyjna 28, 131
socjologia 23, 49, 63
sofizmat gracza *57*
spondej 62, 119, 120–122
spójność tekstu 34–38 43, 45, 49, 54–55
stacjonarność 10, 12, 54–66, 69–70, 134
stan procesu 10, 12, 46, 55, 59, 63
stopa metryczna *10*, 25, 27, 62, 118–120, 122, 128
strofa onieginowska 109, 111, 112
strukturalizm 7, 28–32, 46
styl osobniczy 51, 75, 104
stylometria 42, 48, 51–5, 54, 125
symetria 74
symulacja danych, patrz *pseudolosowy szereg*
synergetyka *8*, 23
szereg czasowy
– definicja 29
– kategorialny (jakościowy) 9–11
– kumulacyjny 10–11
– numeryczny 9–10
– prosty 10–11
– resztowy 71–72
– stacjonarny 10, 12, 54–66, 69–70, 134
– wielokrotny 30
Szkoła Londyńska *29*
Szkoła Praska 29
szyk wyrazów 28, 30, 130
średnia szeregu 65
tekstu definicja 6
temat 30
teoria
– informacji 32–33, 38–42
– systemów 23–24
test frakcji 93–94
test serii 56–57
teza 120–121
tonalność 97
trójkąt Sierpińskiego *73*
trzynastozgłoskowiec 70, 104–107
typy idealne 73
wariancja szeregu 65
wariancja wyjaśniona przez model 72
wersologia 5, 30, 76, 92
weryfikacja modelu 14
weryfikowalność 6, 8
wiersz
– definicja 68
– biały 114
– sylabiczny 78, 83–92, 104–105, 116
– sylabotoniczny 75–83, 85–89, 91–94, 107, 112–113, 116, 122, 128
– toniczny 79, 117
Wordnet *23*
wskaźnik struktury, patrz *test frakcji*
wskaźniki Greenberga 131
współczynnik von Neumana 48–49
wstęga Bartletta 71, *80*
wyrazowe sekwencje 131–138
zdaniowe sekwencje 14, 24, 48–51, 53, 129–131
zestrój intonacyjny 79
zmienna losowa 13, 63–64, 66

ANEKS

A. LISTA WYKORZYSTANYCH TEKSTÓW I NAGRAŃ

BUŁHAKOW Michał, *Mistrz i Małgorzata*. Warszawa 1988, Czytelnik.
Strony: 12, 58, 64, 78, 90, 94, 98, 193, 197, 236, 268, 333, 336, 396, 452;

БУЛГАКОВ Михаил, *Мастер и Маргарита*. Москва 1998, Художественная Литература.
Strony: 14, 50, 55, 65, 75, 78, 81, 151, 154, 182, 207, 254, 255, 302, 344;

BRZECHWA Jan, *Opowiedział dzięcioł sowie* z tomu *Brzechwa dzieciom*. Warszawa 1965, Nasza Księgarnia, 32–46.
Wykorzystano fragmenty: 32–33, 34–35, 35–36.
Nagranie: *Bajki dla dzieci*, cz. 1. Warszawa 1995, Fann Music.
Wykonanie: G. Barszczewska, J. Bończak, W. Drzewicz, M. Kondrat, J. Zelnik i inni.

IWASZKIEWICZ Jarosław, *Sława i chwała*, tom 1 i 2. Warszawa 1973, PIW.
Strony: tom 1 – 28, 59, 97, 146, 198, 251, 261, 305, 316, 380, 428, 439, 481; tom 2 – 22, 96;

ИВАШКЕВИЧ Ярослав, *Собрание сочинений т. 6, Хвала и слава*. Москва 1975, Художественная Литература.
Strony: 28, 55, 93, 139, 188, 240, 251, 293, 304, 368, 414, 426, 468, 504, 575;

KONWICKI Tadeusz, *Wniebowstąpienie*. Warszawa 1982, Iskry.
Strony: 19 (1), 33 (2), 42 (3), 54 (4), 68 (5), 72 (6), 80 (7), 134 (8), 152 (9), 176 (10).

MICKIEWICZ Adam, *Dzieła*, tom 4. Warszawa 1995, Czytelnik.

МИЦКЕВИЧ Адам, *Собрание сочинений*, том 2, перевод Светлана Мар-Аксенова. Москва 1949, Государственное издательство художественной литературы.

W obu wersjach językowych wykorzystano następujące wersy: I/26-37 216-227 386-397 503-514 572-583 810-821; II/49-60 225-236 350-361 624-635 815-826; III/79-90 172-183 395-406 469-480 633-644; IV/57-68 217-228 409-420 518-529 756-767 910-921; V/85-96 199-210 453-464 648-659 836-847; VI/70-81 266-277 497-508; VII/61-72 167-178 288-299 405-416; VIII/35-46 141-152 336-347 460-471 645-656 763-774; IX/72-83 165-176 317-328 439-450 537-548 671 682; X/11-22 209-220 314-325 474-485 584-595 687-698; XI/37-48 162-173 383-394 656-667; XII/184-195 349-360 453-464 769-780.

MORAVIA Alberto, *Nuovi Racconti Romani di Moravia*. Roma 1963, Bompiani.
Strony: 158 (1), 188 (2), 212 (3), 230 (4), 236 (5), 316 (6), 348 (7), 386 (8), 476 (9), 530 (10).

NEWERLY Igor, *Wzgórze Błękitnego Snu*. Warszawa 1986, Czytelnik.
Strony: 5–6, 18–19, 23–24.
Nagranie: Zakład Wydawnictw i Nagrań Polskiego Związku Niewidomych, Warszawa 1988.
Wykonanie: H. Machalica.

ПУШКИН Александр С., *Евгений Онегин*. Москва 1977, Детская Литература.

PUSZKIN Aleksander, *Eugeniusz Oniegin*, przekład Adam Ważyk. Wrocław itd., 1993, Ossolineum.

W obu wersjach językowych wykorzystano następujące strofy: I/4 11 20 22 27 30 47 57; II/2 4 14 17 23 29 36 39; III/1 6 11 16 20 27 35 40; IV/8 11 16 20 27 38 44 48; V/2 7 9 18 25 28 37 42; VI/6 9 22 24 28 32 41 45; VII/2 5 8 19 24 28 35 53; VIII/5 9 11 16 22 40 46 51.

SŁOWACKI Juliusz, *Beniowski*. Wrocław itd., 1989, Ossolineum.
Wykorzystano fragmenty: pieśń pierwsza 1–72, pieśń druga 1–72, pieśń trzecia 1–72;
Nagranie: Zakład Wydawnictw i Nagrań Polskiego Związku Niewidomych, Warszawa 1981.
Wykonanie: S. Zaczyk.

SZCZYPIORSKI Andrzej, *I ominęli Emaus*. Poznań 1992, Kantor Wydawniczy SAWW.
Strony: 16 (1), 35 (2), 66 (3), 72 (4), 88 (5), 112 (6), 130 (7), 146 (8), 156 (9), 172 (10).

WOJTYŁA Karol (JAN PAWEŁ II), Homilia wygłoszona 21 czerwca 1983 we Wrocławiu podczas drugiej pielgrzymki papieskiej do ojczyzny (nagranie prywatne, 3 fragmenty).

Teksty prasowe:

POLITYKA nr 41/1999 (2214) – 10 fragmentów;

RZECZPOSPOLITA nr 237/1999 (5402) – 10 fragmentów;

ИЗВЕСТИЯ nr 25381–25384/1999 – 20 fragmentów;

B. LISTA BADANYCH TEKSTÓW ŁACIŃSKICH

Hor. *Serm.* I 3, 66–75; I 4, 115–124; I 9, 63–72; I 10, 41–50; II 3, 176–185; II 4, 4–13; II 5, 28–37; II 7, 108–117; II 8, 71–80; Hor. *Epist.* I 1, 53–62; I 7, 1–10; I 11, 19–28; I 13, 10–19; I 16, 24–33; I 20, 4–13; II 1, 199–208; Hor. *Ars* 60–69; 147–156; 220–229; 351–360; Verg. *Aen.* I 419–428; III 229–238; IV 634–643; V 72–81; VI 585–594; VII 770–779; IX 88–97; X 668–677; XI 764–773; XII 233–242; Verg. *Georg.* I 84–93; I 328–3347; II 184–193; III 16–25; III 414–423; IV 206–215; Verg. *Ecl.* I 59–68; IV 11–20; VII 57–66; IX 32–41; Ov. *Met.* I 454–463; II 52–61; III 583–592; IV 742–751; V 228–237; VI 503–512; VII 147–156; VIII 713–722; IX 370–379; X 702–711.

C. Wyniki badań prozy artystycznej i stylu prasowo-publicystycznego w językach polskim i rosyjskim

Tab. 1 Struktura rytmiczna równoległych fragmentów prób prozy polskiej i rosyjskiej (na przykładzie powieści *Mistrz i Małgorzata* M. Bułhakowa)

	Tekst polski (akcent stały)				Tekst rosyjski (akcent swobodny)			
	Model	V_{org}	V_{res}	V_e	Model	V_{org}	V_{res}	V_e
1	MA(2)	0,211	0,129	39%	MA(1)	0,203	0,143	30%
2	MA(2)	0,212	0,113	47%	MA(1)	0,211	0,177	16%
3	MA(2)	0,216	0,129	40%	MA(1)	0,211	0,158	25%
4	MA(1)	0,210	0,136	35%	MA(1)	0,204	0,157	23%
5	MA(2)	0,213	0,137	36%	MA(1)	0,233	0,179	23%
6	MA(1)	0,209	0,127	39%	MA(1)	0,215	0,165	23%
7	MA(1)	0,214	0,142	34%	MA(1)	0,213	0,167	22%
8	MA(1)	0,209	0,137	34%	MA(1)	0,214	0,178	17%
9	MA(1)	0,197	0,146	26%	MA(1)	0,207	0,176	15%
10	MA(1)	0,202	0,142	30%	MA(1)	0,218	0,165	24%
11	MA(2)	0,220	0,139	37%	MA(1)	0,226	0,164	27%
12	MA(1)	0,205	0,137	33%	MA(1)	0,217	0,165	24%
13	MA(2)	0,210	0,130	38%	MA(1)	0,221	0,164	26%
14	MA(1)	0,216	0,147	32%	MA(1)	0,225	0,172	23%
15	MA(1)	0,208	0,142	32%	MA(1)	0,216	0,182	16%
			Średnia*:*	**35,5%**			Średnia*:*	**22,3%**

Tab. 2 Struktura rytmiczna równoległych fragmentów prób prozy polskiej i rosyjskiej (na przykładzie powieści *Sława i chwała* J. Iwaszkiewicza)

	Tekst polski (akcent stały)				Tekst rosyjski (akcent swobodny)			
	Model	V_{org}	V_{res}	V_e	Model	V_{org}	V_{res}	V_e
1	MA(2)	0,211	0,123	42%	MA(1)	0,216	0,176	18%
2	MA(1)	0,205	0,149	27%	MA(1)	0,220	0,176	20%
3	MA(1)	0,189	0,134	29%	MA(1)	0,232	0,181	22%
4	MA(2)	0,227	0,148	35%	MA(1)	0,233	0,184	21%
5	MA(1)	0,213	0,134	37%	MA(1)	0,220	0,172	22%
6	MA(1)	0,217	0,142	35%	MA(1)	0,229	0,188	18%
7	MA(2)	0,212	0,144	32%	MA(1)	0,215	0,167	22%
8	MA(1)	0,205	0,133	35%	MA(1)	0,214	0,164	23%
9	MA(1)	0,213	0,149	30%	MA(1)	0,208	0,153	26%
10	MA(2)	0,223	0,162	27%	MA(1)	0,224	0,180	20%
11	MA(1)	0,209	0,143	32%	MA(1)	0,194	0,164	15%
12	MA(1)	0,205	0,127	38%	MA(1)	0,205	0,165	20%
13	MA(1)	0,210	0,140	33%	MA(1)	0,224	0,192	14%
14	MA(1)	0,204	0,137	33%	MA(1)	0,212	0,164	23%
15	MA(2)	0,217	0,128	41%	MA(1)	0,219	0,163	26%
			Średnia*:*	**33,7%**			Średnia*:*	**20,7%**

Tab. 3 Struktura rytmiczna stylu prasowo-publicystycznego w językach polskim i rosyjskim

	Tekst polski (akcent stały)				Tekst rosyjski (akcent swobodny)			
	Model	V_{org}	V_{res}	V_e	Model	V_{org}	V_{res}	V_e
1	MA(2)	0,213	0,127	40%	MA(1)	0,222	0,191	14%
2	MA(2)	0,208	0,134	36%	MA(1)	0,203	0,163	20%
3	MA(2)	0,204	0,124	39%	MA(1)	0,215	0,179	17%
4	MA(2)	0,222	0,148	33%	MA(1)	0,217	0,178	18%
5	MA(1)	0,203	0,140	31%	MA(1)	0,205	0,154	25%
6	MA(1)	0,211	0,143	32%	MA(1)	0,208	0,192	8%
7	MA(1)	0,205	0,138	33%	MA(1)	0,199	0,159	20%
8	MA(2)	0,215	0,156	27%	MA(1)	0,208	0,158	24%
9	MA(1)	0,210	0,147	33%	MA(1)	0,187	0,156	17%
10	MA(1)	0,218	0,147	33%	MA(1)	0,202	0,168	17%
11	MA(1)	0,200	0,153	23%	MA(1)	0,206	0,181	12%
12	MA(1)	0,203	0,126	38%	MA(1)	0,223	0,198	15%
13	MA(1)	0,208	0,151	27%	MA(1)	0,201	0,169	16%
14	MA(1)	0,207	0,146	29%	MA(1)	0,198	0,160	19%
15	MA(1)	0,224	0,149	33%	MA(1)	0,199	0,166	17%
16	MA(1)	0,192	0,135	30%	MA(1)	0,209	0,166	21%
17	MA(1)	0,214	0,143	33%	MA(1)	0,212	0,183	14%
18	MA(1)	0,225	0,173	23%	MA(1)	0,210	0,171	19%
19	MA(1)	0,203	0,159	22%	MA(1)	0,212	0,166	22%
20	MA(2)	0,216	0,134	38%	MA(1)	0,214	0,172	20%
			Średnia:	**31,7%**			Średnia:	**17,8%**

Oznaczenia stosowane w tabelach 1, 2 i 3:

V_{org} – wariancja szeregu obserwowanego
V_{res} – wariancja resztowa
V_e – procent wariancji szeregu obserwowanego wyjaśniony przez model

QUANTITATIVE METHODS IN SEQUENTIAL ANALYSIS OF TEXT (summary)

The goal of the study was to verify the hypothesis: "the linear order of some linguistic units in text is a realisation of a stochastic process and for this reason cannot be regarded as random" (cf. p. 25). This general statement could not be the object of empirical verification, but enabled us to advance and test a series of detailed hypotheses concerning some well-determined levels and scopes of linguistic analysis.

Of all the hypotheses advanced, the most effective were those tested on the prosodic level. Modelling the time-series representing the sequences of stressed and unstressed syllables in a continuous text lets us replace the vague notion of "text rhythm" with the quantitative, synthetic measure of rhythmical orderedness in text based upon the percentage of the original variance (in the observed series), explained by the estimated linear model of the stochastic process underlying the series (V_e). The V_e coefficient can be calculated for any series of numbers and, consequently, for any text, irrespective of its language, style, versification system and length.

It was demonstrated that the V_e coefficient (calculated for the stress-based series) is a good typological criterion in language taxonomy. Although the basis of verification was limited to two Slavic languages, there is no substantial obstacle to analysing any dynamically accentuated alphabetical language in the same way. Successful studies of Chinese (DREHER et al 1969) prove that this approach to linguistic data can also yield worthwhile results in the case of tonal languages, where the text would be represented as a series of tones.

The analysis of samples in Polish (quantification of stressed and unstressed syllables) has shown that sequential characteristics, such as V_e or the autocorrelation function, are an efficient, quantitative criterion in the taxonomy of styles and/or versification systems (Fig. 38, p. 116). Sequential modelling at the prosodic level can also be applied in the theory of translation. As an example, parallel fragments of Polish and Russian prose were compared with regard to their rhythmical structure. The results obtained were satisfactory.

The ARIMA method turned out to be a powerful tool in the analysis of versification. When analysing the time-series based on the sequences of stressed and unstressed syllables in versified texts, it revealed not only the strength and the depth of the contextual connection between the directly neighbouring syllables, but also statistically significant connections between the non-adjacent units – i.e. every 10^{th} or 11^{th} syllable etc. (Fig. 8, 11, 31, 33–34). This "seasonal lag" usually coincides with the verse length. However, it was shown that the real length of the minimal, equivalent rhythmical pattern, repeatedly occurring in the line of text, can be different (e.g. two verses or a half-verse).

The linguistic material analysed in the study was not limited to the stress-based languages. We compared the degree of linear order in Latin hexameter, coded respectively as a sequence of long vs. short and stressed vs. unstressed syllables. In the latter case, we assumed the existence of the dynamic, metrical stress called "ictus", placed on the strong parts of the consecutive metrical feet and appearing, supposedly, during

the declamation of Latin poetry. The average percentage of the original variance explained by the estimated models was much higher for the stress series than for the series based on quantity coding. Contrary to the received opinion, the underlying basis of rhythm in the Latin hexameter was not quantity but dynamic stress.

Good results were obtained in the modelling of text on the lexical level. We advanced the hypothesis that the linear arrangement of words represented by some quantitative measure related to their frequency (e.g. Shannon's quantity of information) depends on the morphosyntactical structure of a language and, in some cases, is not random. The tests carried out showed that in languages of inflectional syntax without a rigid word order in a sentence (Polish) no sequential regularity was detected, whereas in languages having an analytic and positional syntax (Italian) weak but significant stochastic processes were found. This fact was due to the alternate appearance of grammatical (very frequent) and lexical (rare) morphemes in Italian. Although the research was based on a limited number of samples, it clearly indicates a general tendency which further tests on multilingual corpora should confirm. V_e and the coefficients of sequential models could be then applied in linguistic typology.

ISBN 83-910656-1-8