Instytut Fizyki Jądrowej
im. Henryka Niewodniczańskiego
PAN
Zakład Teorii Systemów Złożonych

Ilościowe charakterystyki złożoności języka naturalnego

Andrzej Kulig

Rozprawa doktorska przygotowana pod kierunkiem
dra hab. Jarosława Kwapienia

Kraków 2014

Spis treści

PODZIĘKOWANIA

Pragnę podziękować mojemu promotorowi, dr hab. Jarosławowi Kwapieniowi, za wszelką pomoc i wykazaną życzliwość podczas pisania tej pracy.
Dziękuje mojemu promotorowi pomocniczemu, dr Pawłowi Oświęcimce, za cenne uwagi i wskazówki przy wykonywaniu obliczeń.
Podziękowania składam kierownikowi Zakładu Teorii Systemów Złożonych, prof. dr hab. Stanisławowi Drożdżowi, za stworzenie wspaniałej, przyjaznej atmosfery podczas całych moich studiów w IFJ PAN.

ABSTRACT

This doctoral dissertation includes the following main theses:

- As samples of natural language, literary texts show several properties of complex systems: they have internal organization, including a hierarchical structure, and the interactions between their components such as words are of complicated nature, which among others can be a consequence of imposed rules of grammar and an author's style of writing. One also observes formation of large-scale effects that are inexplicable on a basis of the sole knowledge of the individual words. Such effect can include content, emotional charge, and artistic value of the text.

- Interactions between words defined by their mutual adjacency, after expressing them in the network representation, show certain features of networks with accelerated growth and, approximately, scale-free degree distribution of nodes. Such networks are also characterized by unique tendency to condensation, which leads to shortening of the path lengths between nodes if the number of nodes increases.

- Despite strong differences in grammar, different European languages do not show comparable differences in network topology. Substantially larger differences can be seen within one language, when one compares texts that represent different literary genres.

- Modelling of the empirical word adjacency networks is possible either directly, via the appropriate network models (e.g., by various kinds of the networks with accelerated growth), or indirectly, via network representation of the relevant stochastic processes. Comparing topology of the model networks with the empirical ones shows, however, that language has some subtleties, which cannot fully be expressed by relatively simple, generic models.

- Literary texts, if parameterized by sentence lengths and expressed in a form of time series, show clear fractal structure, and in some cases even the multifractal structure. On the literary science ground, the latter group of texts can be linked with a narrative technique called the stream of consciousness.

This dissertation is divided into 5 chapters. Chapter 1 contains a short introduction with listed the main objectives and theses of the work. Chapter 2 is devoted to description of the phenomenon of natural language - its origins, evolution, and morphology. The main theories of the language origin and formal classification of languages is also discussed in this part of the work. Chapter 3 contains an introduction to complex systems science. It begins with the explanation, why physics is a branch of science the best equipped to examine such systems and the natural language in

particular. Later on, the term of complexity is introduced and the most important properties of complex systems are discussed together with the methodology allowing for their study.

Chapter 4 is a container that includes description of all the analyses and discussion of the obtained results. It is composed of several sections devoted to specific issues. Section 4.1 presents a statistical analysis of empirical data consisting of vocabulary of six European languages with particular emphasis put on the Zipf approach. In Section 4.2 literary texts expressed by word adjacencies are a subject to network analysis. Of interest are the topological properties of these networks, especially the node connectivity distributions and the average shortest path lengths. Empirical results are confronted with the results of simulations according to different network models. Last Section 4.3 presents the results of the fractal analysis applied to time series of sentence lengths with the main stress put on identification of multifractal properties.

Finally, Chapter 5 contains a summary with critical discussion of the results presented throughout this work, as well as an indication of possible directions of future research.

Rozdział 1

Wprowadzenie i cel pracy

1.1 Wstęp

Umiejętność posługiwania się językiem jest jedną z kluczowych cech, która jakościowo odróżnia ludzi od zwierząt, pozwalając na wzajemną wymianę informacji w sposób nieobserwowalny nigdzie indziej. Wprawdzie nauki zoosemiotyczne[1] wyróżniają pewne formy komunikacji między zwierzętami, takie jak: przekazy środowiskowe, wytyczanie granic terytorium czy informacje o stanie emocjonalnym, jedynie człowiek rozwinął ją w sposób nieporównywalnie złożony, tworząc tym samym niezwykle skomplikowany, ale i również efektywny system symboli i reguł, pozwalający na wzajemne komunikowanie się czy wyrażanie myśli. Zdolność posługiwania się językiem stanowi kluczowy element naszej ewolucji, jest on czynnikiem determinującym istnienie oraz rozwój społeczny niezaobserwowany wśród innych istnień żywych na całym świecie.

Właściwy opis języka naturalnego oraz jego wielopłaszczyznowa analiza jest konieczna w kontekście prawidłowego zrozumienia jego pochodzenia, struktury i dynamiki. Samo jego pojęcie jest interdyscyplinarne, w najszerszym możliwym rozumieniu, począwszy od nauk ścisłych, poprzez przyrodnicze skończywszy na naukach humanistycznych. Istnienie szeregu analiz językoznawczych, opisujących jego statystyczne właściwości, morfologię czy własności strukturalne pozwalają wysunąć tezę, że jest on jednym z przykładów *układów złożonych*[2]. Podobnie jak w przypadku ilościowej analizy innych układów tego typu, opis języka naturalnego jest zadaniem skomplikowanym, niedającym się łatwo i bezstratnie zredukować i sformalizować w ramach kilku praw dotyczących relacji pomiędzy jego elementami składowymi (np. słowami). Do najważniejszych metod badawczych można zaliczyć m.in. mechanikę statystyczną procesów nierównowagowych, modelowanie stochastyczne, teorię sieci złożonych i analizę fraktalną.

Temat ten będzie dokładniej przedstawiony w rozdziale 3, w tym miejscu warto wspomnieć, że użyteczność opisu układów złożonych w języku mechaniki statystycznej wynika z ich budowy, ponieważ składają się z wielkiej liczby składników i ich

[1] Zoosemiotyka – nauka zajmująca się komunikacją zachodzącą pomiędzy zwierzętami.

[2] Układy złożone – termin interdyscyplinarny, określający pewną klasę układów występujących w przyrodzie, których opis w oparciu o właściwości elementów składowych jest niewystarczający; układy takie zostaną szerzej omówione w dalszych rozdziałach pracy.

opis w języku tradycyjnej dynamiki jest praktycznie niemożliwy, biorąc pod uwagę dostępną współcześnie moc obliczeniową. Ponadto układy złożone, w tym i język naturalny, przebywają w stanach oddalonych od równowagi termodynamicznej, stąd konieczność wyjścia z analizą poza granice klasycznej, XIX-wiecznej mechaniki statystycznej. W modelowaniu ewolucji układów złożonych wykorzystuje się najczęściej procesy stochastyczne o leptokurtycznych, niegaussowskich rozkładach fluktuacji i nieliniowych, długozasięgowych korelacjach. Z kolei teoria sieci złożonych pozwala na określenie struktury danego układu poprzez wyrażenie oddziaływań pomiędzy elementami składowymi za pomocą grafów i analizę ich topologii oraz, jeśli to możliwe, także dynamiki. Analiza fraktalna, wraz z jej wersją multifraktalną, pozwala na ilościowe wyrażenie złożoności, zarówno w przypadku struktury badanego układu, jak i dynamiki związanych z nim obserwabli. Ma to związek z często spotykaną w przyrodzie fraktalnością, a także z intuicyjnym postrzeganiem obiektów fraktalnych jako złożonych.

Cechy języka naturalnego, takie jak statystyki słów, wzajemne relacje pomiędzy ich występowaniem czy ilościowe charakterystyki jego struktury pozwalają na swobodne zastosowanie powyższych metodologii badawczych. Dzięki temu możliwe jest uzyskanie szeregu wyników znamionujących istnienie w języku nietrywialnych właściwości, charakterystycznych dla układów złożonych. Wnioski płynące z tak przeprowadzonej analizy mogą być cenną informacją w kontekście klasycznych analiz językoznawczych czy literaturoznawczych, ale również tych przeprowadzanych w ramach pokrewnych dziedzin – komunikacji i przetwarzania języka naturalnego[3].

1.2 Tezy i zakres pracy

Praca zawiera następujące zasadnicze tezy:

- Teksty literackie jako próbki języków naturalnych wykazują niektóre własności układów złożonych: posiadają wewnętrzną organizację, w tym hierarchiczną budowę, a interakcje pomiędzy elementami składowymi, takimi jak słowa, mają skomplikowany charakter, narzucony przez reguły gramatyki i styl pisarski autora i prowadzą do powstawania efektów wielkoskalowych, niewytłumaczalnych na gruncie znajomości poszczególnych słów. Do takich efektów można zaliczyć treść, ładunek emocjonalny i wartość artystyczną tekstu.

- Interakcje pomiędzy słowami, określone poprzez ich wzajemne sąsiedztwo, po wyrażeniu w reprezentacji sieciowej wykazują cechy sieci złożonych o przyspieszonym wzroście i (w przybliżeniu) bezskalowym rozkładzie krotności wierzchołków. Sieci konstruowane w oparciu o teksty charakteryzują się ponadto wyjątkowo silną tendencją do kondensacji, co prowadzi do zmniejszania się długości ścieżek pomiędzy wierzchołkami wraz ze wzrostem ich liczby.

- Różne języki europejskie, pomimo istotnych różnic w gramatyce, nie wykazują porównywalnie dużych różnic w topologii reprezentujących je sieci sąsiedztwa

[3]Ang. *natural language processing* (NLP) – interdyscyplinarna dziedzina, zajmująca się zautomatyzowaną analizą języka naturalnego przez komputer i wykorzystującą zagadnienia z zakresu językoznawstwa i sztucznej inteligencji.

słów. Większe różnice widoczne są w ramach jednego języka, gdy porównuje się teksty przynależne rożnym formom wypowiedzi, np. powieści i teksty naukowe.

- Modelowanie sieci sąsiedztwa słów jest możliwe zarówno bezpośrednio, poprzez odpowiednie modele sieciowe (np. modyfikacje modeli sieci o przyspieszonym wzroście), jak i pośrednio, przez sieciowe reprezentacje procesów stochastycznych. Porównanie topologii takich sztucznych sieci z sieciami empirycznymi pokazuje jednak, że język zawiera pewne subtelności, których nie da się w pełni wyrazić przez stosowanie stosunkowo prostych, generycznych modeli.

- Teksty literackie sparametryzowane przez długości zdań je tworzących i wyrażone w formie szeregów czasowych wykazują budowę fraktalną, a wśród nich są także teksty o budowie multifraktalnej. Te drugą grupę tekstów można powiązać na gruncie literaturoznawczym z techniką narracyjną strumienia świadomości.

Tekst pracy jest podzielony na 5 rozdziałów, poniżej znajduje się krótkie omówienie zawartości każdego z nich.

W rozdziale 2 zostanie przedstawiona teoria dotycząca genezy powstania języka naturalnego, jak również opis jego struktury w kontekście teorii informacji oraz teorii gramatyk formalnych. Wiedza ta jest istotna ze względu na właściwą interpretację wyników przedstawionych w dalszych rozdziałach pracy. Omówione zostaną trudności w sformułowaniu genezy pojawienia się języka naturalnego wśród ludzi i ścisłego określenia jego morfologii.

Rozdział 3 rozpoczyna się wyjaśnieniem, dlaczego fizyka jest dziedziną nauki w pełni uprawnioną do badania języka naturalnego. W dalszej części rozdział ten zawiera wprowadzenie do tematyki złożoności, omówienie najważniejszych fizycznych własności układów złożonych, a także przedstawienie głównych koncepcji i metodologii wykorzystywanych w badaniach będących przedmiotem pracy.

Opis wykonanych analiz oraz wszystkie uzyskane wyniki zostały zamieszczone w rozdziale 4. Składa się on z kilku podrozdziałów, w których przedstawione zostały wyniki pokrewnych analiz. W podrozdziale 4.1 przedstawiono statystyczną analizę słownictwa sześciu języków europejskich, ze szczególnym uwzględnieniem zależności opisanych prawem Zipfa. W podrozdziale 4.2 przedstawiono język naturalny w reprezentacji sieci sąsiedztwa słów. Przedmiotem zainteresowania są własności topologiczne tych sieci, wyrażone przez główne miary oferowane przez teorię sieci złożonych. Wyniki analizy danych empirycznych zestawione zostały z wynikami symulacji przeprowadzonych w oparciu o kilka różnych modeli, w tym autorskich. W ostatnim podrozdziale 4.3 przedstawiono wyniki badań tekstów literackich metodami analizy fraktalnej. Przedmiotem badań były szeregi czasowe długości zdań, a główny nacisk położony został na identyfikacje złożoności wyrażonej przez struktury multifraktalne.

W rozdziale 5 zawarto szereg krytycznych spostrzeżeń, które należy wziąć pod uwagę, dokonując rzetelnej analizy tekstów pisanych. Badanie języka naturalnego narzędziami wywodzącymi się z nauk ścisłych nie zawsze prowadzi bowiem do uzyskania wyników, których interpretacja jest jednoznaczna. Wskazana została potrzeba interdyscyplinarnej analizy przez specjalistów z różnych dziedzin nauki.

Rozdział 2

Język naturalny

2.1 Pochodzenie języka naturalnego

Jednym z najistotniejszych, a zarazem najtrudniejszych pytań dotyczących języka naturalnego jest geneza jego powstania [1]. Istnieją dwa aspekty tego zagadnienia: filogenetyczny i ontogenetyczny. Aspekt filogenetyczny dotyczy istoty powstania języka wraz z kształtowaniem się i rozwojem ludzkości, natomiast aspekt ontogenetyczny bada rozwój posługiwania się językiem u dzieci [2]. Przeprowadzone dotychczas badania nie dają jednoznacznej odpowiedzi, kiedy, w wyniku jakich okoliczności i dlaczego w ogóle język, którym ludzie posługują się na co dzień, powstał. Istnieje kilka hipotez, próbujących nadać naukowy i spójny charakter w obrazie darwinowskiej teorii ewolucji [3], którą powszechnie uznaje się za poprawną. Zagadnienia te już od dawna interesowały filozofów; Platon stawiał pytanie, czy język, jakim się posługują ludzie, jest *physei* czy *thesei*, naturalny czy stanowiony – czy język jest wrodzoną cząstką człowieczeństwa od początku jego istnienia, czy też jest umiejętnością nabytą w procesie socjalizacji. Z kolei Immanuel Kant stwierdził, że język *jest przypuszczalnym początkiem ludzkości* [4] i jest nieodzownie związany z umiejętnością komunikacji. Na gruncie stricte naukowym dociekanie prawdy jest jednak niezwykle trudne i obarczone niedostatkami metodologicznymi. Brak jakichkolwiek próbek bądź zachowanych przykładów języka mówionego w początkach jego formowania oraz szczątkowa wiedza odnośnie wczesnych form języka pisanego utrudniają przeprowadzanie dokładnych i rzetelnych badań nad *arche* języka naturalnego [5, 6].

Aby móc mówić o naukowym charakterze badań, dociekanie musi się zawierać w przestrzeni metodologii naukowej, wymuszającej m.in kryterium *falsyfikowalności*. W tym przypadku nie może być ono w pełni spełnione ze względu na brak jakiejkolwiek możliwości odtworzenia warunków początkowych panujących na wczesnym etapie kształtowania się języka. Współczesna wiedza opiera się jedynie o szczątkowe oraz trudne do identyfikacji dane archeologiczne [7], zawierające niekompletne informacje na temat prymitywnych kultur i zalążków życia społecznego [8, 9]. Dane tego rodzaju dostarczają istotnych informacji odnośnie budowy anatomicznej ówczesnych ludzi, na podstawie których można wysuwać wnioski odnośnie potencjalnych możliwościach komunikacji werbalnej [6]. Umiejętność posługiwania się mową może być bezpośrednio badana jedynie na podstawie budowy i umiejscowienia krtani oraz organów odpowiedzialnych za wydawanie dźwięków [7].

Istnieje silny związek pomiędzy umiejętnością posługiwania się językiem a tworzeniem kolektywnych struktur społecznych, mimo że wzajemne implikacje tych zjawisk są przedmiotem ciągłych analiz [10]. Poddając analizie obecne tempo rozprzestrzeniania się języka oraz poziom jego zdywersyfikowania można – używając metod statystycznych – estymować przedział czasu, w jakim mógł się on pojawić [11]. Wedle tych zgrubnych szacunków, początki jego rozwoju są datowane na 50 000 – 100 000 lat wstecz, natomiast rozwój pisma, jako już wtórnej umiejętności językowej, na 7000 lat wstecz [12, 13]. Znaczna rozbieżność bliska rzędowi wielkości wydaje się być naturalna – bo o ile datowanie początków języka jest jedynie aproksymacją, to w przypadku pisma[1] można polegać na danych utrwalonych w skamieniałościach.

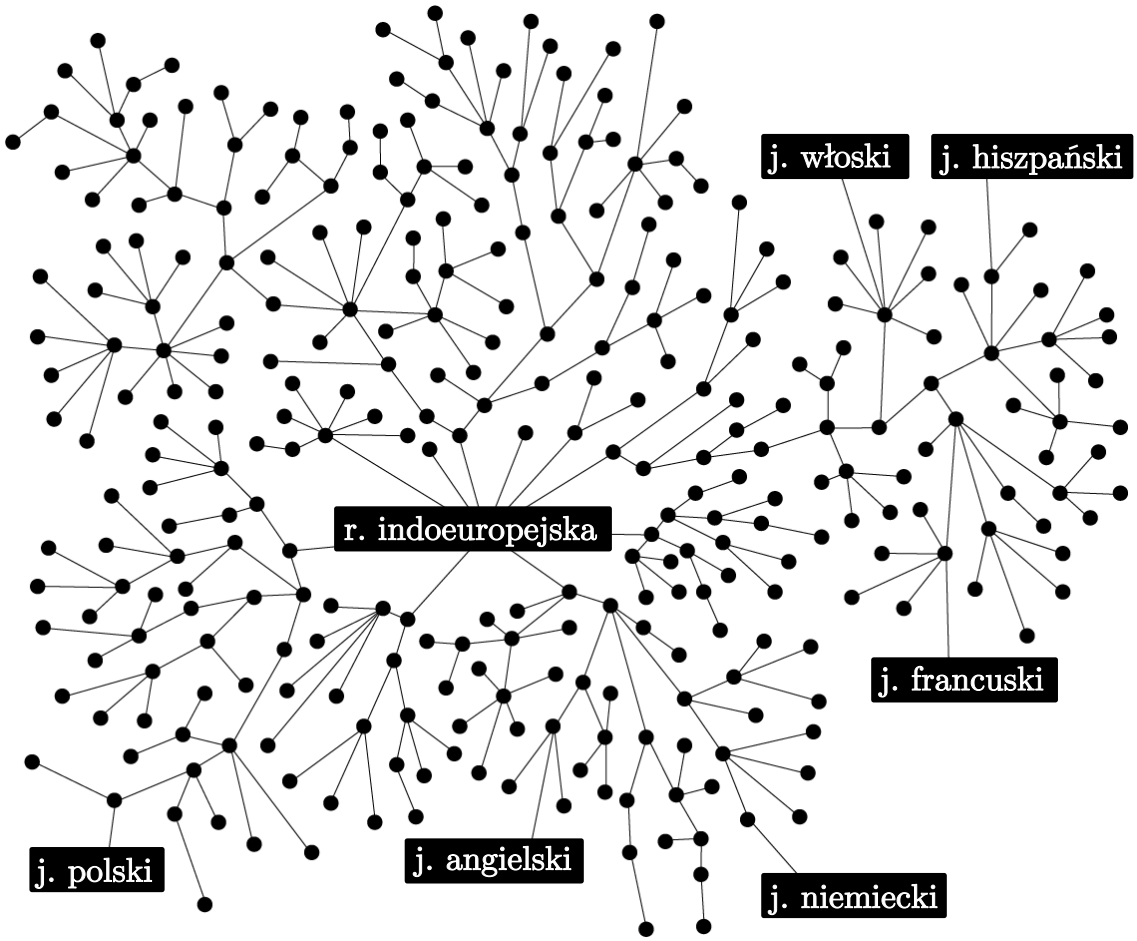
Na podstawie badań opisujących zmienność genetyczną i morfologiczną naszych przodków można również określić położenie geograficzne początków formowania się języka naturalnego [15]. Okazuje się, że jest ono związane ze stanowiskami występowania *homo sapiens* na terenach subsaharyjskich [7, 16]. Zmienność ta przekłada się na zróżnicowanie morfologiczne języka, gdzie najmniejszą liczbę fonemów[2] – 11 – zidentyfikowano w niektórych językach indo-pacyficznych, natomiast najwięcej, bo aż 141, w !Kung – języku używanym w Afryce Południowej [17]. Fakt ten wydaje się potwierdzać tezę monogenezy języka naturalnego, która miała mieć miejsce w środkowej Afryce [18]. Wraz ze stopniową migracją człowieka z Afryki w stronę Eurazji następowało stopniowe zacieranie zróżnicowania fonetycznego języka [19]. Mapy ukazujące rozkład zróżnicowania genetycznego i fenotypowego ludzi w ścisły sposób pokrywają się z mapami odzwierciedlającymi różnorodność fonetyczną mowy [20]. Na rysunku 2.1 przedstawiono drzewo języków indoeuropejskich wskazujące na ich wspólny rdzeń, świadczący o homogeniczności językowej.

Istnieje kilka hipotez na temat genezy języka, powstawania słów czy świadomej wymiany informacji w początkowych etapach jego rozwoju, ale są one na tyle wybiórcze i wąskie w swoim zakresie, że nie stanowią istotnej podstawy do stworzenia ewentualnej całościowej teorii. Wydawanie dźięków, będących kombinacją nucenia prostych melodii oraz chrząkania, miało na celu wyrażenie emocji, pozytywnych i negatywnych, pomiędzy nadawcą a odbiorą komunikatu. Splot tych dwóch różnych procesów: wydawania melodii (samogłosek) oraz chrząkań (spółgłosek) mogło dać początek bardziej skomplikowanym formom artykułowania dźwięków, jakimi są fonemy, słowa i frazy [21, 22].

Wczesne spekulacje na temat próby opisania pochodzenia mowy zostały przedstawione w ramach kilku prymitywnych procesów (zachowań), jakie mogły mieć miejsce podczas tworzenia się społeczności ludzkich. Teoria onomatopeiczna (ang. *ow-wow theory*) głosi, że ludzka mowa miała swoje źródło w dźwiękonaśladownictwie przyrody, imitując dźwięki natury [23, 24]. Według następnej teorii – wykrzyknieniowej (ang. *pooh-pooh theory*) – język powstał jako proces wyrażania bólu, cierpienia, radości czy popędu seksualnego. Kolejną teorią jest teoria apelatywna (ang. *yo-he-ho theory*), wywodząca język z dźwięków wydawanych w czasie wspólnej, na ogół ciężkiej pracy [25, 26, 27].

[1] Jako pismo rozumiemy ścisły system znaków, będących reprezentacją zbioru obiektów i pojęć, w tym także abstrakcyjnych (np. pismo piktograficzne czy pismo ideograficzne) [14].

[2] Najmniejsza jednostka stosowana występująca w języku.



Rysunek 2.1: Drzewo języków w obrębie rodziny indoeuropejskiej, skonstruowane w oparciu o regularne podobieństwa, takie jak: występowanie wspólnych lub zbliżonych form wyrazów, podobieństwo morfologiczne i składniowe itp.

Wydaje się jednak, że zbiór możliwych słów utworzonych za pomocą tych procesów nie jest na tyle duży i zróżnicowany, by móc stać się punktem wyjścia do wyrażania jakichkolwiek innych, abstrakcyjnych myśli, niezwiązanych z rzeczywistym przedmiotem czy prostą czynnością posiadającą swoje słowne określenie [28]. Przedstawione dociekania stanowią jedynie czysto mechanistyczne podejście do procesu tworzenia się języka, dając odpowiedź jedynie na pytanie o etymologię niektórych z używanych słów, nie wyjaśniając jednak w choćby minimalnym zakresie ogromnego zróżnicowania słownictwa, charakterystycznego dla każdego z języków naturalnych [29, 30, 31].

Zasadniczo inne podejście w kontekście powyższych rozważań nosi nazwę teorii gestów, głosząc, że język rozwinął się z gestykulacji [32], którą się posługiwano we wczesnych etapach komunikacji interpersonalnej [33, 34]. Okazuje się, że jest ona istotnie powiązana z językiem werbalnym, gdyż za ich funkcjonowanie odpowiadają te same struktury nerwowe, tj. ten sam obszar kory mózgowej [35]. Ponadto „niemy" przodek człowieka mógł używać gestykulacji do wyrażania bądź przekazywania prymitywnej informacji, co wydaje się być naturalnym sposobem bezpośredniego wyrażania myśli i emocji w tych warunkach [36]. Człowiek do tej pory używa gestykulacji,

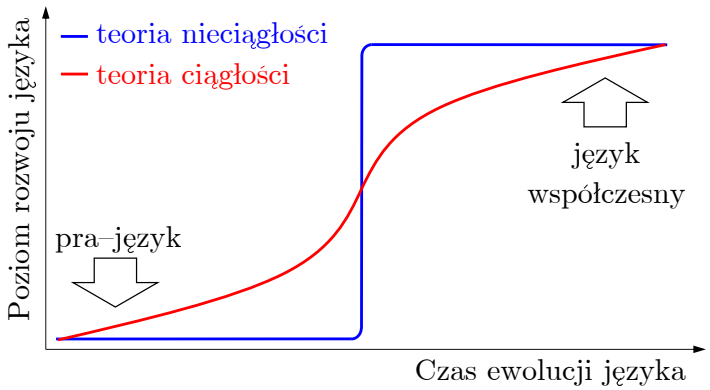
na ogół podświadomie, korzystając z niej w sytuacjach, kiedy możliwości językowe są niewystarczające. Mankamentem tej hipotezy jest fakt, że na ogół gestykulacja ma charakter nieświadomy, będący jedynie bezwładną manifestacją emocji. Język naturalny pod tym względem jest wyraźnie odmienną formą przekazywania informacji, podlegającą ścisłym regułom i konwencjom. Mimo iż w obrębie niego również istnieje możliwość nieświadomego i niekontrolowanego wydawania sygnałów (m.in alarmów, krzyków), tym niemniej teoria gestów jako rzeczywiste źródło pochodzenia języka, jest pod wieloma względami nie do przyjęcia. Tym niemniej możliwe są równoważne reprezentacje języka, niewykorzystujące kanału audytywnego, a kanał wzrokowy (np. język migowy) – jest to jednak wtórna umiejętność językowa, przypisująca określonym słowom bądź frazom odpowiednie symbole wizualne [37, 38].

Konkurencyjną tezę wysunął M.H. Christiansen [39], według której postać języka jest zdeterminowana charakterem (trybem) pracy mózgu. Język, jako organ, ewoluował w kierunku jak najefektywniejszego funkcjonowania w ramach pracy mózgu ludzkiego, modyfikując i optymalizując swoją postać. Inne (hipotetyczne) realizacje języka, nieposiadające żądanej formy, nie przetrwały w toku ewolucji, zostały jedynie te, które najlepiej dostosowały się do funkcjonowania organizmów, które się nim posługiwały. Prowadzone symulacje na sieciach neuronowych pokazały, że efektywność danego języka jest ściśle skorelowana z jego wewnętrzną strukturą i jedynie kilka form pozwala na optymalną jego naukę. Jest to wniosek niezwykle istotny, bo faktycznie, istnienie wielu języków o typowych strukturach (porządku części zdania) musi być konsekwencją pewnych preferencji decydujących o takiej a nie innej jego postaci.

Oprócz wyżej wymienionych teorii istnieje wiele innych, m.in.: teoria neuronów lustrzanych [33], teoria gramatykalizacji [40], czy teoria samoudomowionej małpy [19]. Wszystkie te podejścia nie są kompleksowymi, wewnętrznie spójnymi teoriami, jednak każda z nich rzuca pewne światło na badane zagadnienie. Wysoce prawdopodobne jest, że właściwe podejście nie powinno się ograniczać jedynie do jednej z nich, gdyż złożoność problemu jest na tyle wysoka, że naiwne by było szukać uzasadnienia wszystkich aspektów w ramach pojedynczej, odseparowanej od reszty teorii. Na ogół spekulatywny charakter tych hipotez może prowadzić do wniosku, że dokładny i gruntowny opis początków języka jest nadal niezwykle trudny, a może wręcz niemożliwy.

Istnieje w końcu stanowisko, reprezentowane przez szerokie grono naukowców, stojące w opozycji do samej koncepcji wyjaśnienia pochodzenia języka naturalnego [36]. Według tego podejścia nie ma istotnych powodów, aby traktować język jako osobny, niezależny byt, a jego istnienie należy raczej rozważać w sposób znacznie szerszy. Język nie jest odseparowaną, samoistną adaptacją na poziomie czysto biologicznym, ale jest częścią szeroko rozumianej kultury stworzonej przez człowieka, stanowiącą niezwykle istotną, integralną i nierozerwalną jej część. Ewentualne metody badawcze powinny zatem brać pod uwagę nie tylko sam proces komunikacji, ale również wszystkie inne możliwe do zaobserwowania czynności realizowane przez człowieka jako przejaw tych samych predyspozycji, świadczących o świadomości będącej cechą unikalną tylko dla rasy ludzkiej [41, 42].

Przy rozpatrywaniu samego tempa nabierania umiejętności posługiwania się językiem, możliwe są dwa różne podejścia. Jedno z nich, tzw. teoria ciągłości, głosi, że tak złożona struktura, jaką jest język naturalny, nie mogła powstać nagle i spontanicznie, ale musiała mieć swoje korzenie w jakimś prajęzyku, używanym przez naszych przodków. Zdolność komunikacji byłaby zdobywana stopniowo, w wyniku działania rozmaitych czynników zewnętrznych (początki kształtowania się życia społecznego, migracje, współzawodnictwo) i naturalnej ewolucji mózgu oraz organów odpowiedzialnych za umiejętność mowy. Czy stopniowe przystosowywanie się do nowej umiejętności było efektem, czy raczej przyczyną nabierania pionowej pozycji ciała – co w konsekwencji doprowadziło do obniżenia krtani, której coraz niższe położenie umożliwiało wydawanie zróżnicowanych dźwięków – jest nadal pytaniem pozostającym bez odpowiedzi [7]. Hipoteza ta nie daje również jednoznacznej odpowiedzi, dlaczego człowiek rozwinął tę umiejętność najbardziej spośród wszystkich gatunków żyjących na ziemi. Zakłada one jedynie, że inne potencjalnie istniejące stworzenia posługujące się językiem wymarły w wyniku braku innych cech gwarantujących przetrwanie [5].



Rysunek 2.2: Dwa hipotetyczne scenariusze ewolucji języka naturalnego. Skokowa zmiana poziomu rozwoju języka jest utożsamiana z mutacją genu FOXP2.

Innym podejściem jest tzw. teoria nieciągłości, która zakłada pojawienie się języka w wyniku przypadkowej mutacji genetycznej, mającej się pojawić około 20 000 – 40 000 lat temu [43]. Spowodowała ona, że osobniki, które w wyniku tego procesu nabyły nową cechę, zyskały ogromną przewagę ewolucyjną nad innymi zwierzętami, związaną z możliwością przekazywania sobie informacji w sposób dużo bardziej efektywny i skuteczny. Według tej hipotezy język od razu stał się niemal „perfekcyjny" w swojej istocie, czyli prawie dokładnie taki, jakim człowiek posługuje się obecnie. Ta konkretna korzyść, która pojawiła się w toku ewolucji, spowodowała że człowiek mógł nieporównywalnie szybciej od innych gatunków nabrać umiejętności adaptacyjnych, zasiedlając niemal wszystkie szerokości geograficzne, gromadząc się w grupach o znacznie lepszej organizacji wewnętrznej [44, 45, 46].

Według Noama Chomsky'ego umiejętność mowy stała się wrodzonym atrybutem, przekazywanym z pokolenia na pokolenie w formie *gramatyki uniwersalnej*, zakodowanej w mózgu każdego człowieka [47]. Te wrodzone kompetencje językowe pozwalają na bardzo szybki rozwój językowy u dzieci, nie wymagając nadmiernego

wysiłku, będąc zarazem niezwykle skutecznymi. Teoria ta może być słuszna jedynie pod warunkiem, iż każdy język ma na swoim podstawowym poziomie identyczną strukturę gramatyczną, którą można ujawnić, stosując odpowiednio głęboką analizę języka. To pociąga za sobą istnienie pewnych niezmienników gramatycznych, które mogą być przejawem homogeniczności języka naturalnego. Badania przeprowadzane wśród ludów nie mających jakichkolwiek kontaktów z innymi cywilizacjami (Tasmańczycy, Andamańczycy) pokazują, że mimo braku wpływów świata zewnętrznego oraz potwierdzonego braku migracji, posiadają własny język o porównywalnej złożoności jak języki społeczeństw wysoko cywilizowanych, co również może być przejawem istnienia wrodzonych predyspozycji językowych.

Kolejnym faktem potwierdzającym słuszność tego podejścia okazało się odkrycie genu FOXP2, który bezpośrednio odpowiada za proces posługiwania się językiem. Gen ten, obserwowany nie tylko wśród ludzi, w bezpośredni sposób odpowiada za komunikacje werbalną, a jego uszkodzenie powoduje natychmiastowe zaburzenia mowy oraz prowadzi do istotnych nieprawidłowości poznawczych [41]. Ponadto nie stwierdzono istotnej różnicy pomiędzy innymi genami współczesnych ludzi a genami ludzi pierwotnych, np. neandertalczyków, co jest kolejnym argumentem za skokową zmiana poziomu rozwoju języka [48]. Zauważono jednak niewielką różnicę w strukturze genu FOXP2 u szympansa i człowieka, mimo niezwykle dużego, bo 97-procentowego, podobieństwa ich genotypów [49]. W świetle powyższych faktów należy uznać za wysoce prawdopodobny scenariusz nabywania umiejętności językowych przez człowieka w wyniku mutacji genetycznej, prowadzącej do jakościowego skoku w kontekście umiejętności komunikowania się pomiędzy sobą. Zarówno postęp technologiczny, jak i poznawczy pozwalają coraz bardziej zbliżać się do istoty problemu genezy języka [50, 51, 52]. Tym niemniej nadal istnieją poważne ograniczenia, mogące w istotny sposób utrudniać jego analizę, pozostawiając wiele pytań, póki co bez odpowiedzi [53, 54, 55].

2.2 Struktura języka naturalnego

2.2.1 Gramatyka formalna a gramatyka języka naturalnego

Język jest to pewien skończony zbiór symboli podlegający ścisłym regułom, czyli gramatyce, regulującej jego wewnętrzną logikę i strukturę. Język formalny nie jest uogólnieniem języka naturalnego, a mimo to jest on używany do jego opisu [56]. Język formalny wraz z generującą go gramatyką formalną stanowią niezwykle ważne pojęcia, posiadające ścisły i systematyczny charakter, co jednak stanowi pewną przeszkodę w ich efektywnym wykorzystaniu do poprawnego i skutecznego opisu języka naturalnego [57]. Właściwy opis w obrazie języków formalnych nie jest możliwy m.in. ze względu na *kontekstowość* języka naturalnego, czyli zależność semantyki (znaczenia) od kontekstu wypowiedzi. Język naturalny jest pod tym względem niezwykle złożonym obiektem, wytworzonym bez świadomego i planowego ustalenia reguł nim rządzących, a jednak pozwalający na skuteczną i w miarę precyzyjną wymianę informacji pomiędzy jego użytkownikami [58]. Język formalny natomiast jest systemem sztucznym, generowanym za pomocą ściśle określonych reguł gramatycznych, determinujących jego wewnętrzną topologię [59].

Ścisła definicja języka formalnego wymaga zdefiniowania pewnego alfabetu Ω, który jest niepustym i skończonym zbiorem symboli terminalnych. Słowa języka to skończone ciągi tych symboli, natomiast ich zbiór tworzy słownik danego języka. Podobnie język naturalny jest zbiorem słów, których odpowiednio skonstruowane ciągi pozwalają na tworzenie kompozycji coraz bardziej złożonych. Posiadają one określoną formę, zdeterminowaną przez użytą gramatykę. Stanowi ona w miarę ścisły i unormowany obraz języka naturalnego, regulując zarówno syntaktykę (strukturę), jak i semantykę (w miarę jednoznaczną relację pomiędzy rzeczywistym obiektem a abstrakcyjnym elementem języka). Zrozumienie komunikatu w danym języku nie jest tylko kwestią znajomości znaczeń poszczególnych słów, ale również związków pomiędzy nimi. Skądinąd poprawne zdania z punktu widzenia gramatyki bywają nonsensowne, nie mające odzwierciedlenia w rzeczywistości[3]. Istnienie odstępstw od reguł gramatycznych oraz kontekstowość języka naturalnego stanowią główną przeszkodę w jego ścisłym i jednoznacznym opisie. Charakter języków formalnych nakłada zbyt rygorystyczne warunki na ich formę, które czasami w języku naturalnym mogą być niezachowywane. Dlatego też różnorodność oraz złożoność języka naturalnego jest konsekwencją subtelnego balansu pomiędzy ścisłym porządkiem syntaktycznym a specyficznym chaosem uniemożliwiającym przekazywanie rozmaitych informacji.

Gramatyka języka naturalnego jest na tyle ścisła, iż umożliwia tworzenie poprawnych struktur, pozwalających na zrozumienie przekazu oraz jego właściwą interpretację, ale pozwala też na pewną swobodę w tworzeniu nowych form. Z drugiej strony języki formalne podlegają jedynie regułom syntaktycznym, które porządkują wzajemne relacje pomiędzy jego abstrakcyjnymi elementami i określają ich wzajemną transformację, co pozwala na tworzenie struktur coraz bardziej złożonych. Brak przyporządkowania funkcji semantycznych elementom języka prowadzi do większej swobody oraz dowolności w generowaniu rozmaitych jego struktur [60]. Wprowadzona przez Noama Chomsky'ego klasyfikacja gramatyk [47] umożliwia usystematyzowanie wewnętrznych regularności języka oraz daje jakościowy opis jego budowy. Wyróżnione zostały cztery ich typy, jednak możliwe jest tworzenie kolejnych poprzez dodawanie ograniczeń bądź uogólnień na już istniejące reguły i własności danej gramatyki.

Niech Ω oznacza skończony zbiór zmiennych będących symbolami terminalnymi, a Ω^* zbiór wszystkich słów, które można utworzyć na zbiorze Ω. Ogólnie gramatykę można przedstawić jako czwórkę:

$$G := (\Omega, \Sigma, S, P), \tag{2.1}$$

gdzie: Σ – skończony zbiór zmiennych będących symbolami nieterminalnymi, gdzie zachodzi: $\Sigma \cap \Omega = \emptyset$ (zbiór ten jest rozłączny z alfabetem), S – wyróżniony element startowy, gdzie $S \in \Sigma$, oraz P – skończony zbiór reguł produkcji, taki że:

$$(\Sigma \cup \Omega)^\star \Sigma (\Sigma \cup \Omega)^\star \longrightarrow (\Sigma \cup \Omega). \tag{2.2}$$

Gramatyka typu 0 to gramatyka bez ograniczeń, generująca język rekurencyjnie przeliczalny wedle reguły $\alpha \rightarrow \beta$, który jest rozpoznawany przez maszynę Turinga.

[3]Słynne zdanie autorstwa Noama Chomskiego: *Colorless green ideas sleep furiously* jest poprawne z punktu widzenia gramatyki, natomiast jest pozbawione jakiegokolwiek sensu.

Gramatyka typu 1 to gramatyka kontekstowa, w której reguły przyjmują następującą postać: $\alpha A\beta \rightarrow \alpha\gamma\beta$, gdzie $A \in \Sigma$, natomiast α, β, γ są ciągami symboli terminalnych i nieterminalnych oraz $\gamma \notin \emptyset$. Język kontekstowy, generowany według tego typu gramatyki, jest rozpoznawalny przez niedeterministyczną maszynę Turinga.

Gramatyka typu 2 to gramatyka bezkontekstowa, wszystkie jej produkcje są postaci $A \rightarrow \gamma$, gdzie $A \in \Sigma$, a γ jest ciągiem składającym się z symboli terminalnych i nieterminalnych. Gramatyki bezkontekstowe są równoważne niedeterministycznym automatom ze stosem[4] i są podstawą większości języków programowania [61].

Gramatyka typu 3 to gramatyka regularna (liniowa) w której reguły produkcji przyjmują dwie następujące postaci: $X \rightarrow aY$ lub $X \rightarrow a$, gdzie $X, Y \in \Sigma$ oraz $a \in \Omega$. Gramatyka ta jest równoważna automatom skończonym[5].

Oznaczając jako R_i rodzinę wszystkich języków generowanych za pomocą gramatyki i-tego typu, słuszna jest następująca inkluzja:

$$R_0 \supset R_1 \supset R_2 \supset R_3. \tag{2.3}$$

Język generowany za pomocą gramatyki o mniejszym indeksie zawiera wszystkie języki powstałe w wyniku stosowania reguł gramatyk o indeksie większym. Powyższe inkluzje są ostre, ponieważ istnieją języki rekurencyjnie przeliczalne, które nie są kontekstowe, kontekstowe, które nie są bezkontekstowe oraz bezkontekstowe, które nie są regularne.

Konstrukcja gramatyk formalnych nie pozwala bezpośrednio tworzyć jawnych struktur językowych, które możemy badać, poddając analizie język naturalny, ale jest ściśle związana z jego strukturą głęboką. Struktura ta odpowiada za warstwę semantyczno-logiczną, tworząc ze zbioru zewnętrznych obserwabli i informacji już posiadanych przez użytkownika formę zdania. Forma ta, jako wynik procesów kognitywno-analitycznych w mózgu, zostaje zadana za pośrednictwem zastosowania odpowiednich rekurencyjnych operacji na zbiorze zakodowanych elementów terminalnych. Proces ten jest wspólny dla wszystkich języków naturalnych, będąc następnie przetwarzany w strukturze powierzchniowej, odpowiadającej za warstwę fonetyczno-fizyczną. Końcowy produkt tego złożonego procesu jest obserwowaną formą języka naturalnego, którego bezpośrednia i powierzchowna analiza nie może dać pełnej informacji o jego rzeczywistej strukturze.

2.2.2 Konstrukcja języka naturalnego

Struktura języka naturalnego, jako bezpośrednio obserwowany przejaw kognitywnej natury mózgu ludzkiego, musi przejawiać w swojej istocie pewne prawidłowości, pozwalające na właściwą percepcję, transformację i generowanie wiadomości. Język naturalny jest uniwersalnym i powszechnym atrybutem ludzi, będąc obiektem ewoluującym na pograniczu biologii, psychologii, matematyki i socjologii. Język rozważany na odpowiednio głębokim poziomie przyjmuje uniwersalną formę, wspólną

[4] Abstrakcyjny, matematyczny, iteracyjny model zachowania pewnego systemu opartego na macierzy dyskretnych przejść pomiędzy kolejnymi jego elementami, który dodatkowo może korzystać ze stosu do przechowywania danych.

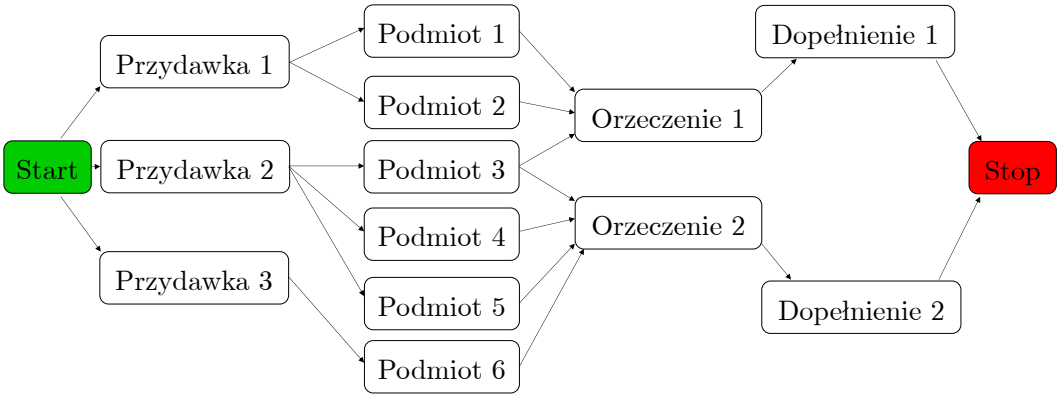
[5] Abstrakcyjny, matematyczny, iteracyjny model zachowania pewnego systemu opartego na macierzy dyskretnych przejść pomiędzy kolejnymi jego elementami.

dla wszystkich jego realizacji, ponadto rola jaką posiada, nie ogranicza się jedynie do komunikacji między osobami. Okazuje się, że właściwa i trafna jego analiza nie może się ograniczać tylko do jednej płaszczyzny, ale równocześnie musi uwzględniać wszystkie inne: ekspresję myśli, wyrażanie osądów i opinii czy kategoryzację postrzeganego świata. Fenomen języka posiada więc daleko idące implikacje w różnych dziedzinach życia. Pojęcie lingwistyki nie jest wyłącznie terminem z zakresu nauk humanistycznych, może zatem być również przedmiotem badań z zakresu biologii i fizyki jako przejaw emergentnej własności ludzkiego mózgu. Wieloaspektowość języka powoduje konieczność rozważania go nie tylko w kontekście wrodzonych zdolności językowych i mechanizmów psychofizycznych, ale również jego funkcjonalności oraz filogenezy.

Język naturalny jest osobliwą i symptomatyczną cechą ludzką, reprezentujący unikalne cechy pracy ludzkiego umysłu: jak zdolność myślenia abstrakcyjnego czy posiadanie świadomości. Owe zdolności językowe można rozpatrywać w dwojaki sposób: w sensie wąskim i szerokim. Pierwszy z nich, tzw. FLN (ang. *faculty of language – narrow sense*) – jest abstrakcyjną reprezentacją języka, będącą czymś w rodzaju systemu obliczeniowego mózgu, wrodzonej predyspozycji umożliwiającej algorytmiczne i rekursywne operacje logiczne. Pozwala to na generowanie nieograniczonej, twórczej i (na ogół) spójnej semantycznie informacji z dyskretnej bazy pojęć. Wynik tych działań manifestuje się w ramach szerszego systemu, tzw. FLB (ang. *faculty of language – broad sense*), który można rozważać jako zdolność mózgu do przetwarzania rozmaitych informacji wejściowych, ściśle zespolonych z systemami sensomotorycznymi i konceptualno-intencjonalnymi organizmu ludzkiego.

Ważnym pojęciem konstytuującym i regulującym pracę języka jest jego gramatyka, odpowiedzialna za generowanie odpowiednio poprawnych struktur językowych, nadająca mu ścisłą i wzajemnie zrozumiałą formę. Reguły te opierają się na założeniu, że język jest strukturą dwuwarstwową; jego tworzenie polega na wygenerowaniu informacji bazowych, podstawowych (struktura głęboka zdania), oraz transformacji ich na informację wyjściową (warstwa powierzchowna zdania). Szczególny nacisk kładzie się tutaj na warstwę głęboką, odpowiadającą za semantyczną percepcję informacji wejściowych, a następnie interakcję z już istniejącymi i generowanie nowych informacji. Warstwa ta, ściśle zespolona ze świadomością oraz umiejętnością myślenia abstrakcyjnego, jest unikatowa, natomiast warstwa powierzchowna jest już obserwowana u zwierząt, gdzie jest powszechnie utożsamiana z ich wzajemną komunikacją. Za poprawną formą zdań wyrażanych w danym języku stoi generująca je gramatyka, jako zbiór skończonych operacji przeprowadzanych na skończonym zbiorze elementów terminalnych i nieterminalnych. Ze względu na mechanizm jej działania można dokonać uściślenia, wyróżniając gramatykę skończenie stanową, gramatykę struktur frazowych i ich superpozycję – gramatykę generatywno-transformacyjną [47].

Struktura syntaktyczna zdania tworzonego za pomocą gramatyki skończenie stanowej jest zakodowana w postaci słów oraz ich wzajemnego porządku w tym zdaniu (rysunek 2.3). W zależności od specyfikacji danego języka (pozycyjny czy fleksyjny), wiodącą rolę stanowi bądź sama struktura zdania (odpowiednie następstwo wyrazów po sobie), bądź fleksja wyrazów. Faktycznymi składnikami zdania są elementy terminalne (słowa), natomiast symbole pomocnicze (stanowiące zbiór elementów nieterminalnych) są wykorzystywane do formowania struktur zdania. Niestety, za



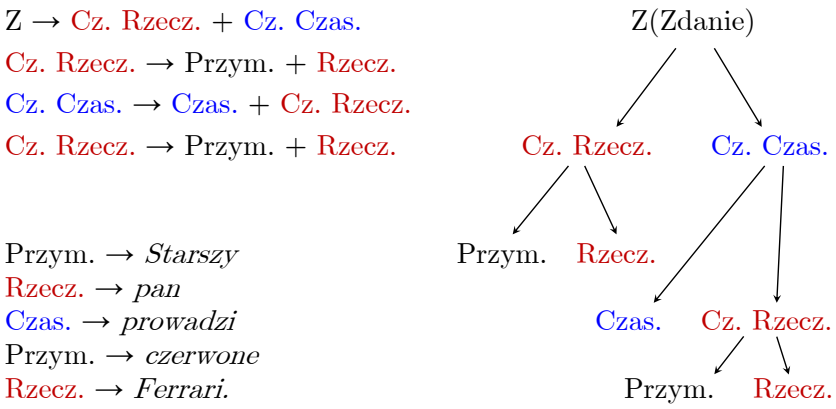
Rysunek 2.3: Struktura liniowa zdania generowana w ramach gramatyki skończenie stanowej. Dana myśl jest formułowana poprzez dobór odpowiednich słów, a zawarcie całościowej informacji możliwe jest w oparciu o odpowiednią ich sekwencję, nadającą kontekst dla sporządzanej wypowiedzi.

pomocą tego procesu nie można odtworzyć wszystkich możliwych zdań w danym języku, a jedynie takie, które posiadają zlinearyzowaną budowę, czyli przypadki, w których istnieje określony i jednoznaczny semantycznie porządek występujących terminów. Wieloznaczność składniowa, będąca wadą liniowej struktury gramatyki skończenie stanowej, dyskwalifikuje ją jako model ogólny, aczkolwiek jest pomocna i używana jako model szczególny w osobliwych sytuacjach. Istnieje zatem konieczność zastosowania innego modelu, eliminującego powyższe komplikacje, a ponadto posiadającego ścisły i ogólny charakter.

Gramatyka struktur frazowych jest gramatyką bezkontekstową, bezpośrednio odwołującą się do struktury głębokiej zdania, a jej forma jest analogią tzw. *nawiasowania*, stosowanego w matematyce lub logice symbolicznej. Wyrażenie postaci $a(b+c)$ posiada na ogół inną wartość niż $ab+c$, tożsamą z $(ab)+c$. Mimo zachowania tej samej struktury liniowej, ma się do czynienia z zupełnie innymi wynikami, co, używając formalizmu gramatyki skończenie stanowej – prowadzi do homonimii konstrukcyjnej (wieloznaczności strukturalnej). Pierwotne zdanie (informacja) w strukturze głębokiej jest materializowane w wyniku rekursywnego stosowania odpowiednich reguł gramatycznych (tutaj tzw. reguł przepisywania) do momentu, aż zostanie osiągnięty odpowiedni szereg terminalny, tj. zdanie wyjściowe. Widać, że i ta koncepcja jest obarczona pewnymi nieścisłościami, nie jest więc kompletnym i adekwatnym opisem, gdyż często nie wyraża intuicji użytkownika danego języka.

Z kolei idea gramatyki struktur frazowych polega na zdefiniowaniu ścisłych reguł generowania, co może być odtworzone za pomocą derywacji przykładowego zdania (rysunek 2.4). Pomimo że struktura jest opisana w sposób dokładny (w ramach tej gramatyki), semantyka zdań nie jest jednoznacznie określona[6]. Modyfikacja tej gramatyki o pewien komponent zaczerpnięty z gramatyki skończenie stanowej spełni założenia dotyczące gramatyki poprawnie opisującej strukturę wszystkich zdań możliwych w ramach danego języka.

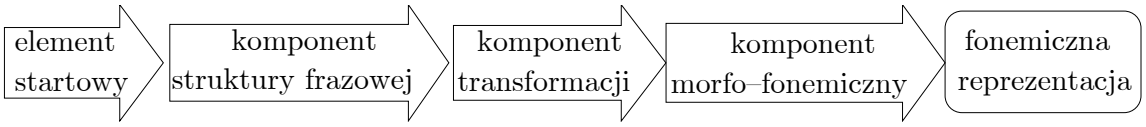
[6]Generowanie zdań w stronie czynnej i biernej jest możliwe za pomocą tej gramatyki, ale opisy te nie są równoważne, pomimo semantycznej równoważności. Zdania takie posiadają odmienną strukturę frazową, posiadając jednak podobne znaczenie.



Rysunek 2.4: Derywacja zdania w ramach gramatyki struktur frazowych. Po lewej stronie zaprezentowano zbiór reguł posiadających zawsze strukturę $X \to Y$, gdzie strzałka oznacza tzw. regułę przepisywania, tj. zastępowanie jednego elementu odpowiednim szeregiem jedno- lub wieloelementowym.

Gramatyka generatywno-transformacyjna, będąca złożeniem dwóch poprzednich gramatyk, jest dwuetapowym procesem, polegającym na tworzeniu zdań bazowych (ang. *kernel-sentences*) będących szeregami wyjściowymi oraz transformowaniu ich na zdania wyjściowe (rysunek 2.5). Dodatkowo uwidacznia się, że owa dwupoziomowość języka naturalnego niesie ze sobą o wiele bardziej złożone relacje niż zakładała to logika formalna stosowana w gramatyce formalnej[7].

Reasumując, kompleksowy i wyczerpujący opis wiernie oddający sposób generowania języka jest niezwykle trudny. Różne koncepcje starające się wyjaśniać to zagadnienie w sposób jak najszerszy, wprawdzie redukują poprzednie, posiadają jednak nowe założenia natury formalnej. Potrzebne są zatem dodatkowe informacje odnośnie subtelności struktury i formy języka, aby prowadzone dociekania mogły być weryfikowalne, a jednocześnie spójne z wiedzą dostarczoną w wyniku ilościowej analizy języka.



Rysunek 2.5: Schemat blokowy struktury generowania zdań za pomocą gramatyki generatywno-transformacyjnej.

[7]Opis pewnej klasy zdań, m.in. tzw. *zdań zanurzonych*, jest możliwy jedynie za pośrednictwem gramatyki generatywno-transformacyjnej.

Rozdział 3

Systemy złożone

3.1 Złożoność – fizyka a język naturalny

3.1.1 Identyfikacja złożoności

Złożoność jest terminem, który stał się powszechnie używany do określania rozmaitych układów rzeczywistych, wykazujących nietrywialną strukturę i subtelne zachowanie pomiędzy jego elementami [62, 63]. Układy te powszechnie występują na wielu skalach[1], są identyfikowane w naukach przyrodniczych jako struktury, których opis, mimo zastosowanej wnikliwej analizy badawczej, ciągle jest niekompletny, cierpi z powodu ograniczeń w stosowanej w nauce klasycznej terminologii i narzędzi poznawczych [64, 65].

Mówiąc o złożoności, należy wskazać obiektywne miary, które mogą ją wyrażać ilościowo, a nie tylko jakościowo, co jednak na ogół przysparza sporo problemów. Zaproponowano wiele definicji złożoności, opisującej różne jej aspekty, takich jak: pojęcie złożoności opartej na teorii informacji [66, 67] (np. złożoność algorytmiczna, entropia informacyjna), termodynamice (np. głębokość termodynamiczna)[2], geometrii (fraktalność, multifraktalność), strukturze oddziaływań. Miary te zostaną bliżej omówione w dalszej części rozdziału, tutaj warto zaznaczyć, że to ostatnie podejście jest najpowszechniej stosowane w fizyce. Zgodnie z nim, układ jest złożony, gdy składa się z wielkiej liczby elementów składowych, oddziałujących w sposób silnie nieliniowy i w jego dynamice można zaobserwować efekty emergentne. Przez emergentne rozumie się grupę zjawisk, które obserwuje się na poziomie makroskopowym struktury lub aktywności układu, a których nie da się opisać wyłącznie na gruncie znajomości oddziaływań elementów składowych na poziomie mikroskopowym. Tak rozumiane układy złożone mogą być w związku z tym opisywane wyłącznie uwzględniając wszystkie poziomy ich organizacji, zarówno mikroskopowe, jak i makroskopowe. Oznacza to, że klasyczne, redukcjonistyczne podejście naukowe, które stara się opis całego świata sprowadzić do kilku oddziaływań fundamentalnych napotyka tutaj na kres swojej stosowalności i musi zostać zastąpione przez wspomniane po-

[1] Same w sobie występują na różnych skalach (mikro-, makroświat), bądź występują jako ten sam układ na wielu skalach jednocześnie.

[2] Fizyczny odpowiednik głębokości logicznej, wyrażający minimalną entropię procesu pozwalającego zrealizować dany układ termodynamiczny.

dejście holistyczne [68]. Układy złożone zostały zidentyfikowane w wielu obszarach, począwszy od naturalnych, jak reakcje chemiczne, powierzchnia Ziemi, atmosfera ziemska, poprzez organizmy żywe, zaczynając od pojedynczych komórek a nawet ich organelli, poprzez ich ekosystemy, po układy społeczne, takie jak zbiorowości organizmów (ludzi, mrówek, pszczół itd.), skończywszy na rynkach finansowych, czy nawet języku naturalnym i wielu innych [69]. Należy oczywiście zdawać sobie sprawę, że skomplikowana struktura bądź stochastyczny charakter pewnych układów nie stanowi dowodu, że są to układy złożone, w takim rozumieniu jaki będzie przedstawiony w tej pracy. Zaawansowane układy elektroniczne czy konstrukcje mechaniczne są niewątpliwie skomplikowane w swojej budowie i funkcjonowaniu, tym niemniej ich opis zawiera się w ramach dobrze zdefiniowanej teorii, odpowiednio elektromagnetyzmu i mechaniki klasycznej [70].

Struktury noszące miano układów złożonych tworzą się spontanicznie, nie są wynikiem zaplanowanego działania konstruktora. Ma się wówczas do czynienia z samoorganizacją [71, 72]. Przykładem może tu być Internet, który wprawdzie jest wytworem inżynierii, wykazuje pewne cechy układów złożonych, w tym samoorganizacji. Gdyby Internet był centralnie sterowany, prawdopodobnie nie przejawiałby cech złożoności. Układy złożone pomimo ich różnorodności wykazują wiele wspólnych cech: samoorganizacja, efekty emergentne takie jak zjawiska kolektywne, czy krytyczność, otwartość, hierarchiczność struktury manifestująca się w jej bezskalowości. Z tego względu w badaniach nad układami złożonymi optymalne jest podejście interdyscyplinarne.

Fizyka jako nauka szukająca uniwersalnych praw przyrody i łącząca w sobie formalizm matematyczny z efektywną metodologią opisu świata na wielu poziomach jego abstrakcji, wydaje się adekwatnym narzędziem, którym możemy badać naturę układów złożonych oraz nadać rządzącym nimi prawom ścisłą, zmatematyzowaną formę. Ponadto silny wzrost możliwości obliczeniowych pozwolił na efektywniejsze badanie systemów złożonych, choć nadal nie są one na tyle wystarczające, by już dzisiaj móc sformułować kompleksowy opis układów złożonych [73].

Należy również zdawać sobie sprawę, iż dokładny opis jakiegokolwiek zjawiska fizycznego jest z góry ograniczony konsekwencjami wywodzącymi się z mechaniki kwantowej, w tym zasady nieoznaczoności Heisenberga i probabilistycznej natury pomiaru oraz teorii chaosu i związanej z nią czułością ewolucji układów na warunki początkowe [74].

3.1.2 Język naturalny jako system złożony

Język naturalny jest niewątpliwie przykładem układu, w którym wiedza o elementach, czyli w tym wypadku słowach, i zależnościach pomiędzy nimi (gramatyka, styl) nie pozwala wyczerpująco wyjaśnić pełnionych przez niego funkcji wyższego rzędu: społecznej i kulturotwórczej [12]. Język naturalny przejawia ponadto inne właściwości typowe dla układów złożonych: otwartość, hierarchiczność, samoorganizację i związaną z nią adaptowalność do zmieniających się warunków [75]. Hierarchiczność języka polega w tym wypadku na tym, że przy pomocy elementów składowych języka można tworzyć nowe formy, które wykazują nowe właściwości i niosą dodatkową informację. Na przykład fonem jest elementarnym dźwiękiem nie niosącym

żadnej istotnej informacji, ale już składając się na poszczególne morfemy (elementarne rdzenie słowotwórcze), uzyskuje znaczenie. Z morfemów powstają słowa, które stanowią już odzwierciedlenie konkretnych pojęć, stanów i obiektów. Kolejną jednostką strukturalną wyższego rzędu staje się fraza, które nadaje właściwy kontekst występującym w niej słowom, wykorzystując do tego celu m.in gramatykę. Frazy tworzą zdania proste lub złożone, które stanowią podstawowy element niosący treść intencjonalnie przekazywaną przez nadawcę. Przejścia pomiędzy tymi poziomami struktury są kluczowe dla języka, ujawniając złożoność procesów, które musiały zajść w mózgu osoby nim się posługującej. Kolejnymi, wyższymi poziomami organizacyjnymi są np. w wersji pisanej akapity, rozdziały, wreszcie formy literackie, natomiast w wersji mówionej wypowiedzi i dialogi. Te elementy strukturalne mogą nieść dodatkowe emergentne cechy – ekspresję, styl czy przesłanie. Hierarchiczność języka odbija się już w samej dyspersji dyscyplin naukowych, jakie zajmują się poszczególnymi poziomami jego organizacji, gdzie poziom najniższy jest obszarem badań biologi i fizjologii, poziomy wyższe lingwistyki, a najwyższe – teorii informacji (patrz podrozdział 2.2.2), psychologii, socjologii i literaturoznawstwa.

Terminem często występującym równolegle z hierarchicznością jest bezskalowość, czyli brak wyróżnionej skali, która byłaby charakterystyczna dla zjawisk występujących w układzie [76]. Bezskalowość jest związana z zależnościami potęgowymi poprzez własność funkcji potęgowej: $f(\lambda x) = \lambda^{\alpha} f(x)$, co sprawia, że funkcja ta wygląda podobnie w każdej skali. W układach złożonych brak charakterystycznej skali może dotyczyć tak rozkładu danej wielkości fizycznej w przestrzeni, jak i w czasie. W przypadku czysto geometrycznym, obiektami które spełniają tę zależność, są np. fraktale, w przypadku sygnałów – bezskalowe fluktuacje mierzonej wielkości w czasie. Brak wyróżnionej skali może też dotyczyć korelacji w przestrzeni i w czasie, tak jak ma to miejsce w zjawiskach krytycznych. Brak wyróżnionej skali ma zwykle poważne konsekwencje, jeśli chodzi o ewolucję układu, gdyż umożliwia zachodzenie w nim zjawisk o dowolnym rozmiarze, ograniczonym jedynie przez ziarnistość jego struktury mikroskopowej i wielkość makroskopową [77].

Bezskalowość jest bardzo rozpowszechniona w naturze, ponieważ może być konsekwencją wielu mechanizmów, m.in takich jak *procesy Yule'a* [78] czy preferencyjne przyłączanie [50]. Oba mechanizmy są wykorzystywane jako modele zjawisk związanych z językiem naturalnym, pierwszy z nich jako tzw. model Simona, a drugi jako model Barabási'ego-Albert (po stosownej adaptacji, o której będzie mowa w dalszej części pracy). Modele te służą do opisu wzrostu zasobów słownictwa w pojedynczym tekście lub korpusie będący ich zbiorem. Statystyczny rozkład częstotliwości słów był pierwszą przesłanką, wskazującą na istnienie bezskalowości w strukturze języka. Empirycznym przejawem tego zjawiska okazują się powszechnie obserwowane w tekstach pisanych *rozkłady Zipfowskie*, które zostały zinterpretowane jako skutek zasady *najmniejszego wysiłku* [77, 79].

Występowanie samoorganizacji jest kolejną ważną cechą układów złożonych i języka. W najbardziej ogólny sposób można ją zdefiniować jako proces zmiany stanu układu pod wpływem oddziaływania z otoczeniem, przy czym celem tej zmiany jest takie dostosowanie się układu do nowych warunków, aby jego stan był optymalny pod względem energetycznym i/lub maksymalizował szanse na dalsze istnienie. Taki stan jest zazwyczaj dynamiczną mieszanką porządku i nieporządku, które mogą

w siebie wzajemnie przechodzić. Procesy samoorganizacji przyczyniają się zwykle do wzrostu złożoności struktury podlegającego im układu. Podobnie, dzisiejsza struktura i funkcjonowanie języka naturalnego jest przykładem długotrwałego procesu samoorganizacji (co jest prawdą, nawet jeśliby przyjąć za Chomskim jego teorię gramatyki uniwersalnej), w którym początkowo prymitywna reprezentacja (protojęzyk) w wyniku ewolucji przekształciła się (a proces ten trwa nadal) w znacznie bardziej optymalną (i bardziej złożoną) formę języka współczesnego [80].

Przedstawione powyżej cechy, charakteryzujące systemy złożone, korespondują z budową języka naturalnego. Podobnie jak w przypadku innych układów złożonych, analiza języka wymaga sprzężenia ze sobą wielu często odległych dziedzin wiedzy, co czyni ją skomplikowaną, nie tylko za naukowego, ale również z praktycznego punktu widzenia. Z pomocą przychodzi metodologia rozwinięta na potrzeby opisu złożoności i układów złożonych. Korzyść z jej stosowania jest obiektywnie bezsporna, gdyż pozwala wyjaśnić bądź usystematyzować różne aspekty języka. Uzyskana w ten sposób wiedza może dać wymierne skutki w postaci efektywniejszej pracy nad sztuczną inteligencją bądź automatycznym przetwarzaniem języka naturalnego.

Wśród definicji złożoności pierwszą, będącą w niektórych dziedzinach podstawową, jest definicja oparta na *złożoności algorytmicznej*. Została ona wprowadzona niezależnie przez Solomonoffa (1964) [81], Chaitina (1969) [82] i Kołomogorowa (1968) [83]. Definiuje się ją jako wyrażenie losowego ciągu za pomocą jak najprostszego algorytmu zrozumiałego przez komputer. Złożoność algorytmiczna pewnego ciągu symboli $\{a\} = a_1, a_2, ..., a_i$ jest najkrótszym algorytmem $Alg : A \rightarrow a_i$, pozwalającym na jego wierne odtworzenie. Jeśli ciąg $\{a\}$ składa się z N różnych elementów, złożoność może być wyrażona jako $\log_N d\{a\}$, gdzie $d\{a\}$ to długość danego ciągu. Złożoność reprezentacji liczby w zapisie binarnym jest równa logarytmowi przy podstawie 2 z liczby znaków binarnych użytych do reprezentacji tejże liczby, natomiast złożoność danego słowa można wyrazić za pomocą $\log_N d$, gdzie N to długość alfabetu, a d to długość tego wyrazu (często wielkość ta jest oznaczana jako koszt użycia danego słowa w tekście).

Inny, praktyczny opis złożoności, wprowadziła fizyka statystyczna do opisu zagadnień z zakresu termodynamiki, fizyki ciała stałego i innych. Według tego podejścia wygodnie jest opisywać zjawisko złożoności w kategoriach *entropii*, jako pewnej miary bezładu panującego w rozważanym układzie, jego nieokreśloności bądź stochastyczności. Rozważmy pewien układ U, przyjmujący jeden spośród N możliwych stanów z prawdopodobieństwem p_i, gdzie $i = 1, 2, 3, ..., N$. *Entropia Shannona* definiowana jest jako suma:

$$H(U) = -\sum_{i=1}^{N} p_i \log p_i. \tag{3.1}$$

Poddając analizie stan układu, dokonuje się pomiaru jakiejś wielkości fizycznej i tym samym zmniejszamy entropię układu: stan układu staje się lepiej określony. Za pomocą powyższej definicji możemy np. określić entropię danego słowa w tekście. Przyporządkowując każdemu słowu prawdopodobieństwo wystąpienia w danym tekście $p_i = f_i/l$, gdzie f_i to częstość wystąpienia i-tego słowa, a l to długość tekstu, można określić entropię informacyjną. Iloraz f_i/l maksymalizuje się dla słów występujących bardzo często, stanowiących rdzeń języka. Z drugiej strony rzeczywista informacja nie jest przekazywana wyłącznie za pomocą pojedynczych słów, ale za pośrednic-

twem odpowiednio wygenerowanych ich ciągów – zdań, stąd użyteczność entropii wydaje się ograniczona. Pomimo że pomiar ilości informacji w formie entropii jest intuicyjny i naturalny, to jednak przyjmuje ona najwyższą wartość dla układów zupełnie przypadkowych, sprawiając, że nie jest to odpowiednie narzędzie do opisu złożoności (przypadkowe sekwencje liczb czy znaków nie mogą być złożone).

Ten konkretny problem można wyeliminować, wprowadzając inną miarę, tzw. *złożoność efektywną* [64]. Opisuje ona stopień złożoności sekwencji X poprzez minimalną złożoność algorytmiczną „dobrej" teorii, która tę sekwencję wyjaśnia. Przez „dobrą" teorię rozumie się taką, dla której sekwencja X jest „typowa" (tzn. prawdopodobieństwo jej uzyskania na gruncie tej teorii nie jest zbyt małe, a rozkład prawdopodobieństwa stowarzyszony z teorią ma małą entropię informacyjną), a sama teoria jest zarazem możliwie prosta algorytmicznie w sensie Kołmogorowa. Zaletą złożoności efektywnej jest rozpatrywanie tylko regularności w sekwencji X, a pomijanie losowości.

Przechodząc od sekwencji znaków do układów naturalnych i związanych z nimi procesów, stopień złożoności może być zdefiniowany jako minimalna długość (liczba zdarzeń) procesu termodynamicznego, który doprowadzić może do odtworzenia danego układu. Jest to tzw. *głębokość termodynamiczna* – wielkość, która teoretycznie może być stosowana w fizyce, choć w praktyce z jej użyciem wiąże się skomplikowany problem braku wiedzy o historii badanego układu i efektywna niemożność zastosowania tej definicji w rutynowych analizach [66]. Z tego powodu do ilościowego wyrażania stopnia złożoności układów w fizyce stosuje się inne podejście. Z pomocą przychodzi tu powszechność występowania struktur bezskalowych w przyrodzie. Wówczas geometryczna złożoność takich struktur opisywana jest w obrazie geometrii fraktalnej. Geometria fraktalna pozwala na ilościowy opis analizowanego układu poprzez podanie wymiaru fraktalnego, pod warunkiem, że układ ten ma dobrze zdefiniowaną samopodobną lub samoafiniczną strukturę [84]. W tym kontekście obiektami o największej złożoności fraktalnej są $multifraktale$, a więc obiekty, których struktura zawiera w sobie mieszankę wielu rożnych fraktali. Mimo powszechności występowania takich struktur w przyrodzie [85], fraktalność nie jest cechą wszystkich układów złożonych, a zatem nie może być uniwersalnym narzędziem definiującym zjawisko złożoności [86, 87, 88]. Niemniej jednak opis fraktalny jest możliwy również w odniesieniu do języka naturalnego, gdzie identyfikuje się multifraktalność pewnych obserwabli w tekstach literackich [89, 90, 91], o czym będzie mowa w kolejnych rozdziałach pracy.

3.2 Sieci złożone

Wzajemne relacje pomiędzy elementami danego układu wygodnie i pod wieloma względami korzystnie jest rozpatrywać jako sieć [92, 93]. Podejście to dostarcza istotnych wiadomości o strukturze i dynamice układu oraz pozwala przenieść ich opis na poziom bardziej abstrakcyjny, redukując przy tym ilość istotnej informacji w stosunkowo nieznacznym stopniu. Abstrakcja opisu ma olbrzymią zaletę, gdyż pozwala porównać ze sobą całkiem odmienne pod względem fizycznym układy, a mimo to znaleźć między nimi pewne uniwersalne cechy [94]. Jest to szczególnie istotne w kontekście interdyscyplinarnych badań nad układami złożonymi, gdzie podobień-

stwa dotyczyć mogą układów biologicznych, społecznych, technicznych, komunikacyjnych, a także naturalnych układów reprezentujących przyrodę nieożywioną [95]. W ten sposób teoria sieci, która przez kilka dekad była niedostatecznie atrakcyjna dla nauk przyrodniczych, doświadczyła w ostatnich 15 latach ogromnego rozwoju motywowanego dodatnim sprzężeniem zwrotnym, jakie powstało na skutek odkrycia jej użyteczności do opisu wielu zjawisk i układów [96, 97, 98]

Historycznie zastosowanie terminologii sieci (w obecnym rozumieniu tego słowa) miało miejsce już stosunkowo dawno temu, bo w XVIII wieku, przez wybitnego szwajcarskiego matematyka Leonarda Eulera, który korzystając z teorii grafów rozstrzygnął zagadnienie mostów królewieckich [99]. Gwałtowny rozwój tej dziedziny był jednak możliwy dopiero dzięki zaistnieniu komputerów, jako narzędzi odpowiednich do analizy dużej ilości danych. Nie bez znaczenia pozostaje też fakt, że wiele sieci rzeczywistych ukształtowało się dopiero w XX-wiecznym boomie cywilizacyjnym: sieci komunikacyjne, współpracy naukowej, kontaktów telefonicznych i wiele innych.

Matematyczna sieć to zbiór wierzchołków, połączonych ze sobą krawędziami. Na gruncie badań empirycznych węzłami są elementy składowe układu, a krawędziami ich wzajemne interakcje. Konsekwencją takiej definicji jest fakt, że każdy wierzchołek należący do danej sieci musi posiadać choć jedno połączenie z innym wierzchołkiem tej sieci. Liczba wszystkich krawędzi, którą posiada dany wierzchołek, nazywana jest jego stopniem i oznaczana jako k. Świadczy on o istotności węzła, gdyż jest bezpośrednią miarą jego interakcyjności z pozostałymi wierzchołkami. Okazuje się, że rozkład wzajemnych połączeń wewnątrz sieci stanowi kluczową cechę, implikując jej dynamikę, funkcjonalność czy nawet stabilność. Oddziaływania te mogą być dodatkowo parametryzowane poprzez podanie wag dla poszczególnych krawędzi, w tzw. *sieciach ważonych*, a same krawędzie mogą przyjąć określoną orientację względem wierzchołków, które łączą, w tzw. *sieciach skierowanych*. Tego typu sieci mogą stanowić już dobrą reprezentację układów złożonych.

Jedną z najistotniejszych, globalnych charakterystyk sieci jest rozkład krotności wierzchołków $P(k)$[3], będący rozkładem prawdopodobieństwa, że losowo wybrany wierzchołek będzie posiadał stopień k. Jeśli przez N oznaczymy liczbę wszystkich wierzchołków sieci, a przez N_{k_i} – tych, które posiadają określony stopień k_i, to prawdopodobieństwo $P(k_i)$ będzie oczywiście wynosić:

$$P(k_i) = \frac{N_{k_i}}{N}. \tag{3.2}$$

W praktyce bardzo często stosuje się zmodyfikowaną formę tego rozkładu, kumulując rozkład stopni; wtedy N_{k_i} jest liczbą wierzchołków o stopniu wyższym lub równym k, a związek pomiędzy skumulowanym a różniczkowym rozkładem wyraża się przez

$$P(k \geqslant k_i) = \int_{k_i}^{k} P(k)dk. \tag{3.3}$$

Zdefiniowanie stopnia bywa czasami trudne i niejednoznaczne, szczególnie dla sieci ważonych i skierowanych, w których występują sparametryzowane lub wierzchołkowo zorientowane krawędzie, co może prowadzić do różnej interpretacji rozkładów

[3]Rozkład krotności i rozkład stopni wierzchołków w tej pracy będzie używany zamiennie.

krotności wierzchołków. Tym niemniej rozkłady te są jawną manifestacją wewnętrznej topologii sieci i bywają znamienną charakterystyką, określającą jej strukturalny charakter oraz zbieżność z adekwatnym modelem teoretycznym.

Jednym z elementarnych zastosowań omawianej terminologii są *sieci regularne*, składające się z pewnej liczby elementów, które są powiązane bądź oddziaływają ze sobą na niewielkich odległościach. Wierzchołki tej sieci oddziaływują na swoich najbliższych sąsiadów, zatem przestrzenny rozkład połączeń ma w takiej sieci charakter lokalny. Każdy z wierzchołków ma ściśle określony, ten sam stopień, a więc odpowiadający im rozkład krotności przyjmuje trywialną postać, rysunek 3.1. Sieci regularne (siatki) znalazły zastosowanie np. w fizyce fazy skondensowanej do opisu struktury niektórych ciał stałych. *Sieci Bravais'go* są użytecznym obrazem w wyjaśnianiu różnych zjawisk, jak: rozprzestrzenianie się dyslokacji, przewodnictwo termiczne i elektryczne, dyfrakcja fal czy właściwości fizyko-chemiczne. Z kolei w dwuwymiarowym modelu Isinga [100] oraz w jego uogólnieniu – modelu Pottsa, użytecznym przy opisie oddziaływań spinowych i przejść fazowych, wykorzystuje się siatki prostokątne [101].

W naturze nie obserwuje się tak ścisłych regularności. Istniała zatem realna potrzeba wprowadzenia innych modeli, nie przejawiających tak wysokiego porządku. W latach 60. ubiegłego wieku dwaj węgierscy matematycy Paul Erdős i Alfred Rényi zaproponowali model, w którym graf (sieć) został przestawiony jako rezultat stochastycznego procesu polegającego na losowym łączeniu się skończonej, wcześniej ustalonej liczby wierzchołków [102]. Procedura konstrukcyjna tak zdefiniowanych grafów przypadkowych przekształca N-elementowy zbiór niepołączonych ze sobą wierzchołków w sieć, w której połączenie każdej z $\binom{N}{2}$ par jest realizowane z ustalonym wcześniej prawdopodobieństwem p. Dla prawdopodobieństwa $p \approx 0$, graf jest niespójny, tzn. składa się z niepołączonych ze sobą komponentów, które są odseparowane od siebie. W miarę wzrostu prawdopodobieństwa niezależne do tej pory klastry łączą się i dla krytycznej wartości $p = p_c$ (zwanej progiem perkolacji) graf staje się spójny, tj. pomiędzy dwoma dowolnymi wierzchołkami istnieje *ścieżka*[4]. Zjawisko perkolacji dobrze opisywane jest w obrazie przejść fazowych (tutaj drugiego rodzaju), gdzie punktem krytycznym parametru kontrolnego p jest wartość p_c, stanowiąca granicę pomiędzy fazą nieuporządkowaną, a fazą uporządkowaną z klastrem perkolacyjnym. W przypadku granicznym, gdy $p \to 1$, wszystkie jego wierzchołki stają się ze sobą nawzajem powiązane, tworząc już graf zupełny, rysunek 3.1.

Na podstawie tej definicji można zauważyć, że rozkład prawdopodobieństwa uzyskania przez dowolny węzeł stopnia k na $N-1$ możliwych połączeń jest tożsamy z uzyskaniem k sukcesów w $N-1$ próbach, gdy prawdopodobieństwo sukcesu wynosi p, co opisuje przez rozkład dwumianowy:

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{3.4}$$

Dla $p \ll 1$ rozkład może być przybliżany przez rozkład Poissona:

$$P(k) = \frac{e^{-\langle k \rangle} \langle k \rangle^{\langle k \rangle}}{k!}, \tag{3.5}$$

[4]Ścieżka jest to nieprzerwana sekwencja krawędzi, pozwalająca na przejście z każdego wierzchołka na dowolny inny.

gdzie przez $\langle k \rangle = p(N-1) \simeq pN$ oznaczamy średni stopień wierzchołków w sieci. Kolejną charakterystyczną cechą tego modelu jest dość mała dyspersja stopni wierzchołków, gdzie wariancja wartości średniej k dla rozkładu Poissona wynosi:

$$\sigma_k^2 = \langle k^2 \rangle - \langle k \rangle^2 = \langle k \rangle. \tag{3.6}$$

Zatem odchylenie standardowe dla stopni wierzchołków $\sigma_k = \sqrt{\langle k \rangle}$ świadczy, iż wierzchołki o stopniu istotnie różnym niż wartość średnia nie występują, co w świetle empirycznych danych nie jest częste. Z tego też powodu klasyczne sieci przypadkowe są czasem traktowane jako przybliżenie sieci regularnych o tym samym stopniu $k_i = \langle k \rangle$. Tym niemniej model ten, na cześć węgierskich matematyków został nazwany *modelem ER*, i jego rozszerzenia funkcjonują nadal w literaturze – nie tylko ze względów historycznych, ale przede wszystkim stanowiąc wygodny model odniesienia w rozmaitych hipotezach zerowych.

Modelem pokrewnym dla sieci ER jest *model konfiguracyjny*, w którym pierwotnie przyjmuje się skończony zbiór wierzchołków wraz z ich krotnością, a następnie dopuszcza się do losowego nawiązywania połączeń. Powtarzając tę procedurę wiele razy, dostaje się stochastycznie generowaną rodzinę grafów o dobrze określonym rozkładzie krotności, posiadającą również zdeterminowane parametry strukturalne: liczbę krawędzi, współczynnik gronowania i korelacje międzywęzłowe [103]. Należy tutaj wyraźnie zaznaczyć, że wszystkie charakterystyki opisujące strukturę generowanych grafów otrzymywane są z własności, które zostały uśrednione po wszystkich realizacjach tego modelu.

Stosunkowo nieliczne zastosowania modelu ER i modeli pokrewnych w praktyce (model dobrze opisuje np. sieci drogowe) spowodowały poszukiwanie innych modeli, pozwalających odtworzyć sieci występujące w rzeczywistym świecie. Jedną z ciekawszych konstrukcji, zaproponowaną przez Duncana Wattsa i Stevena Strogatza w 1998 roku, jest *model WS*, w którym początkowo regularna sieć z wierzchołkami o określonym stopniu k podlega procesowi losowego przełączania („przezwojenia") istniejących krawędzi, w wyniku czego sieć zmienia swoją topologię i rozkład krotności wierzchołków staje się poissonowski [104]. Najistotniejsze w tym modelu jest to, że pod pewnymi względami jego charakterystyki strukturalne stają się bliskie empirycznym danym. Jedną z nich jest gronowanie (klasteryzacja), świadcząca o występowaniu wewnątrz sieci lokalnych triad, czyli trzech wzajemnie połączonych wierzchołków, co globalnie można przedstawić za pomocą *współczynnika gronowania* C. Współczynnik ten jest uśrednioną po wszystkich węzłach wartością lokalnego współczynnika C_i, zdefiniowanego jako:

$$C_i = \frac{2E_i}{k_i(k_i-1)}, \tag{3.7}$$

gdzie E_i to liczba istniejących krawędzi między sąsiadami i-tego wierzchołka. Korzystając z lematu o uściskach dłoni, można wykazać, że wartość współczynnika zawiera się w przedziale $0 \leqslant C_i \leqslant 1$. Współczynnik ten jest dobrze określony na sieciach nieważonych i nieskierowanych, jednak w sytuacji, kiedy sieć ma charakter ważony lub jest skierowana, istnieje kilka konkurujących ze sobą definicji.

O ile dla modelu ER współczynnik gronowania był niski, to dla modelu Wattsa-Strogatza jego wartość jest wysoka i podobna do tych obserwowanych w rzeczywistych strukturach, np. w sieciach społecznych. Inną wielkością charakteryzującą sieć

jest średnia odległość pomiędzy dwoma losowo wybranymi wierzchołkami [105, 106]. Stanowi ona niezwykle ważne pojęcie, pozwalające m.in. określić efektywny rozmiar sieci i czas rozprzestrzeniania się sygnału w jej wnętrzu w ramach istniejących połączeń. Ściśle rzecz biorąc, średnia długość ścieżki jest definiowana jako:

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d(i,j), \tag{3.8}$$

gdzie $d(i,j)$ jest najkrótszą ścieżką pomiędzy dwoma wierzchołkami i oraz j. Jak się okazuje, sieci powstałe w wyniku procesu opisywanego przez model WS posiadają bardzo niską wartość ℓ, której przybliżenie wynosi:

$$\ell \sim \frac{\ln N}{\ln \langle k \rangle}. \tag{3.9}$$

Co najciekawsze, właściwość ta jest osiągana w wyniku przełączenia tylko znikomej liczby istniejących krawędzi w sieci regularnej, w której średnia odległość pomiędzy wierzchołkami $\ell \sim N^{1/d}$, gdzie d – to wymiar sieci, np. $d = 2$ dla siatki kwadratowej. Ze względu na niską wartość ℓ sieci takie określa się mianem *sieci małego świata*, co odzwierciedla niewielką odległość pomiędzy wszystkimi wierzchołkami i jest echem słynnego eksperymentu Stanleya Milgrama [107], który dowiódł, że pod względem bezpośrednich znajomości średnia odległość pomiędzy dwojgiem losowo wybranych ludzi jest zaskakująco niska.

Kolejnym charakterystycznym parametrem opisującym poszczególne węzły jest tzw. *pośrednictwo* [108, 109, 110]. Miara ta, obok wspomnianej wyżej krotności k, należy do grupy miar centralności węzłów, opisujących istotność danego węzła w strukturze sieci. Formalnie miara ta dla i-tego wierzchołka jest zdefiniowana jako:

$$b_i = \sum_k \sum_{j>k} \frac{\delta_{jk}^{(i)}}{\delta_{jk}}, \tag{3.10}$$

gdzie δ_{jk} jest liczbą najkrótszych ścieżek łączących węzły j i k , natomiast $\delta_{jk}^{(i)}$ określa liczbę tych spośród nich, które przechodzą przez i-ty węzeł.

Model WS, choć znacznie bardziej realistyczny niż modele typu ER, także nie posiada wszystkich cech sieci rzeczywistych. Taką znamienną i bardzo często występującą cechą rzeczywistych sieci jest silnie asymetryczny rozkład prawdopodobieństwa krotności wierzchołków, znacznie bardziej asymetryczny niż wynikałoby to z własności rozkładu Poissona. Okazuje się, że liczba wierzchołków o niskim stopniu jest niewspółmiernie większa od liczby tych o stopniu największym. Poza tym na ogół rzeczywiste sieci charakteryzują się pewną dynamiką związaną z ich ewolucją w czasie: zmienia się liczba wierzchołków, krawędzi lub obu tych elementów jednocześnie.

Pierwszym modelem sieci, który uwzględnił powyższe przesłanki, była konstrukcja przedstawiona przez Alberta-Laszló Barabási'ego i Rékę Albert (*model BA*) [94], opierająca się na dwóch prostych założeniach:

– *wzrostu sieci*: w każdym kroku t dodawany jest do sieci nowy wierzchołek z możliwością przyłączenia się do m innych, istniejących już wierzchołków;

– *preferencyjnego przyłączania*: nowy wierzchołek przyłącza się do już istniejącego z prawdopodobieństwem proporcjonalnym do stopnia tego drugiego:

$$p_i = \frac{\eta_i k_i^\mu}{\sum_j \eta_j k_j^\mu}. \tag{3.11}$$

Dla pierwotnej wersji modelu parametry η i μ równe są jedności.

W ramach modelu BA możliwe są rozmaite modyfikacje, uwzględniające tempo przyłączania się nowych wierzchołków, jak również pojawianie się bądź zanikanie krawędzi w już istniejącej sieci. Najważniejszą własnością sieci wygenerowanej w taki sposób jest potęgowy rozkład wierzchołków, rysunek 3.1:

$$P(k) \propto k^{-\gamma} \tag{3.12}$$

z charakterystycznym wykładnikiem $\gamma = 3$. Nieliniowość preferencyjnego przyłączania pojawia się, jeśli $\mu \neq 1$, co dla $\mu < 1$ prowadzi do wykładniczego obcięcia tego rozkładu, natomiast dla $\mu > 1$ początkowy węzeł staje się na tyle istotny, że skupia prawie wszystkie krawędzie występujące w sieci i połączenia występujące w sieci kondensują się na właśnie na nim. Dodatkowy parametr η cechuje pewną różnorodność adaptacyjną węzłów, modyfikując rolę krotności k_i. Jest to uzasadnione, gdyż spodziewać się można, że w rzeczywistych strukturach sama informacja odnośnie stopnia wierzchołka nie jest informacją pełną.

Model BA i jego warianty, w których uzyskuje się bardziej zbliżone do rzeczywistości wartości wykładnika γ, pozwalają dokładniej odzwierciedlić ewolucję sieci tworzonych w oparciu o dane empiryczne. To samo odnosi się do innych charakterystyk sieciowych, takich jak szybkość wzrostu średniej długości najkrótszej ścieżki. Jest ona bezpośrednio związana z wykładnikiem γ:

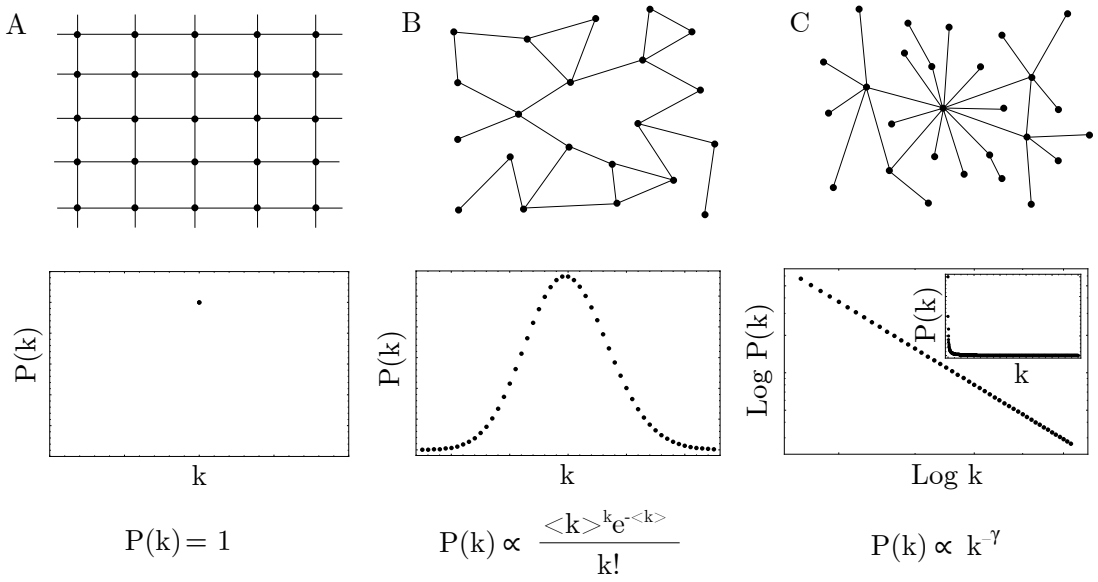
$$3 > \gamma > 2 \qquad \ell \propto \ln\ln N \tag{3.13}$$

$$\gamma = 3 \qquad \ell \propto \frac{\ln N}{\ln\ln N} \tag{3.14}$$

$$\gamma > 3 \qquad \ell \propto \ln N, \tag{3.15}$$

gdzie N jest rozmiarem sieci wyrażonym w liczbie wierzchołków.

Opisany tutaj minimalny model BA nie jest odpowiedni do opisu wszystkich ewoluujących sieci o potęgowym rozkładzie krotności wierzchołków. Sieci powstałe wedle tego opisu posiadają m.in. zbyt niską wartość współczynnika gronowania, gdzie dla parametru $m = 1$ współczynnik ten jest w ogóle równy zeru. Ponadto w sieciach rzeczywistych obserwuje się istnienie korelacji międzywęzłowych, w których węzły mają tendencję do łączenia się z węzłami o podobnych własnościach (np. krotności), model BA pozwala zaś na generowanie jedynie sieci, w których nie istnieją wspomniane korelacje. Konieczne jest więc wprowadzenie uogólnień modelu BA, takich jak wprowadzony przez Dorogowcewa i Mendesa model sieci o przyśpieszonym wzroście (*model DM*) [98]. W niniejszej pracy zostanie szerzej rozpatrzony przypadek tej właśnie rodziny modeli, umożliwiającej współistnienie dwóch mechanizmów wzrostu sieci: poprzez dodawanie nowych wierzchołków oraz poprzez dodawanie krawędzi pomiędzy już istniejącymi wierzchołkami. Jak widać, modele DM dobrze opisują topologię sieci, jaką można otrzymać, rozpatrując próbki języka naturalnego [111, 112, 113].



Rysunek 3.1: Graficzna reprezentacja trzech typów sieci: sieć regularna (A), sieć losowa (B) i sieć bezskalowa (C) (na górze) oraz odpowiadające im rozkłady krotności wierzchołków $P(k)$ (na dole).

3.3 Fraktale i multifraktale

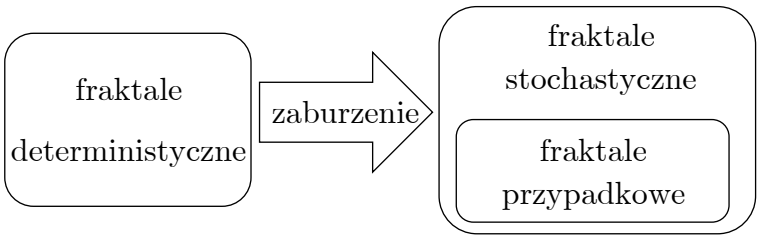
Pierwsze prace bezpośrednio związane z pojęciem fraktali były dziełem francuskich matematyków: Gastona Julii i Pierre'a Fautou [114]. Wykorzystując formalizm zespolonych funkcji wymiernych, opisywali oni własności układów dynamicznych, analizując zachowanie odpowiednio długiej iteracji. Prace te nie od razu stały się punktem zwrotnym w omawianej dziedzinie, będąc przez ponad pół wieku teorią nie posiadającą większego znaczenia praktycznego. Dopiero w latach 70. ubiegłego wieku Benoit Mandelbrot, wsparty technikami komputerowymi[5], usystematyzował oraz uogólnił tę teorię, nadając występującym w niej obiektom geometrycznym miano *fraktali*[6] [84].

Obecność struktur fraktalnych w układach rzeczywistych jest konsekwencją nieliniowości w dynamice tych układów na rozmaitych poziomach ich organizacji. Sprzężenia i konkurencja pomiędzy różnymi, często przeciwstawnymi procesami, mogą w konsekwencji prowadzić do krytyczności i związanych z nią zależności potęgowych [115].

Fraktale są obiektami, których definicja jest nie do końca ścisła i określa je jako obiekty samopodobne o bardzo szczegółowej i „chropowatej" strukturze, nie do końca dające się opisać w języku tradycyjnej geometrii euklidesowej, np. pod względem wymiaru. Z jednej strony fraktalami nazywa się ściśle określoną klasę obiektów powstałych w wyniku stosowania nieskończonej iteracji na jakiejś rodzinie funkcji (tzw. *fraktale deterministyczne*); choć co prawda w jawny sposób nie występują w przyrodzie, są inspirującą idealizacją, która w istotny sposób może

[5]Mandelbrot przez długi okres był pracownikiem naukowym IBM.
[6]Z łac. *fractus* – podzielny, ułamkowy, cząstkowy.

ułatwić teoretyczne rozważania dotyczące rzeczywistych obiektów. Z drugiej strony tak samo określa się procesy, których własności statystyczne są identyczne na różnych skalach (tzw. *fraktale stochastyczne*). Fraktale stochastyczne są niewątpliwie lepszą reprezentacją naturalnych struktur, które mogą być niekiedy określane jako fraktale deterministyczne zawierające pewien element przypadku, np. szum, błąd strukturalny, ograniczenia czasowo-przestrzenne, itd [85]. Taka stochastyczna modyfikacja może prowadzić do sytuacji, w której jakościowy charakter danego fraktala deterministycznego nie ulegnie zmianie i dalej zachowa on w przybliżeniu swoją pierwotną naturę, ale jego struktura może też ulec tak istotnej modyfikacji, że przejdzie on już do klasy fraktali stochastycznych. W tym drugim wypadku może dojść do tak znacznej deformacji pierwotnej struktury fraktalnej, że w konsekwencji pojawią się błędy w ilościowym opisie obiektu [116]. Stanowi to podstawową trudność w analizie fraktalnej, przede wszystkim układów rzeczywistych. W układach takich stochastyczne składowe mogą być tak silne, że pomimo jakościowej fraktalności, ich ilościowy opis może być poważnym wyzwaniem metodologicznym.



Rysunek 3.2: Jakościowy podział fraktali ze względu na ich charakter.

Podstawową wielkością określającą fraktale jest ich wymiar samopodobieństwa, który jak się okazuje może przyjmować wartości spoza zbioru liczb naturalnych. Jego wartość może też być interpretowana jako jedna z miar opisujących złożoność tych obiektów. Jeśli przez N oznaczymy liczbę części, na które został podzielony jakiś obiekt, a przez s rozmiar skali, to prawdziwa jest zależność $N(s) = s^{-d_\mathrm{T}}$, gdzie wspomniany wymiar samopodobieństwa to:

$$d_\mathrm{T} = -\frac{\log N(s)}{\log s}. \tag{3.16}$$

Dla obiektów ściśle opisywanych przez geometrię euklidesową wymiar samopodobienstwa staje się wymiarem topologicznym i tak na przykład podzielenie odcinka na trzy części ($N = 3$) odpowiada skali $s = 1/3$, dając w rezultacie $d_\mathrm{T} = 1$. Nieco inaczej wygląda sytuacja w przypadku zbioru Cantora, gdzie podzielenie odcinka na trzy części prowadzi do uzyskania jedynie dwóch kopii, stąd $d_\mathrm{T} = \frac{\log 2}{\log 3} \approx 0.6309$. W praktycznych zastosowaniach używa się tzw. *wymiaru pudełkowego* d_C, który może być otrzymany w prostej implementacji ściśle związanej z powyżej przedstawioną procedurą. Niech S będzie pewnym zbiorem w przestrzeni metrycznej (X, d), w szczególności w przestrzeni euklidesowej $\mathbb{R}^n$. Zakładając, że $N(\varepsilon)$ będzie liczbą kul o średnicy ε pokrywających zbiór S, to jego wymiar przyjmie postać:

$$d_\mathrm{C}(S) = \lim_{\varepsilon \to 0} -\frac{\log N(\varepsilon)}{\log \varepsilon}. \tag{3.17}$$

Jeśli powyższa granica nie istnieje, rozważa się górny $d_{\mathrm{C}}^{\mathrm{up}}$ bądź dolny $d_{\mathrm{C}}^{\mathrm{down}}$ wymiar pudełkowy[7], odpowiadający granicy odpowiedniego kresu dla powyższego ilorazu; stąd wymiar d_{C} jest dobrze określony wtedy, gdy $d_{\mathrm{C}}^{\mathrm{up}} = d_{\mathrm{C}}^{\mathrm{down}}$.

Kolejną, jedną z bardziej formalnych definicji jest *wymiar Hausdorffa*, opisany przy pomocy miary Hausdorffa. Wymiar ten może być stosowany do szerokiej klasy obiektów, a jedynie w wyjątkowych przypadkach istnieje konieczność rozróżnienia pomiędzy nim a wymiarami opisanymi poprzednio. Niech X będzie przestrzenią metryczną pokrytą rodziną zbiorów S, gdzie $S \subseteq X$, i $\delta > 0$, to δ-wymiarową miarą Hausdorffa zbioru S jest wielkość:

$$H^{\delta}(X) = \lim_{\varepsilon \to 0} \inf_{S} \{ \sum_{i=1}^{\infty} \phi(S_i)^{\delta} \quad \text{gdzie} \quad S = \cup_i S_i, \quad \phi(S_i) \leqslant \varepsilon \}, \tag{3.18}$$

gdzie infimum jest brane po wszystkich otwartych pokryciach S_i o średnicy nie większej niż ε, a wymiar Hausdorffa jest zdefiniowany jako

$$d_{\mathrm{H}}(S) = \inf\{\delta : H^{\delta}(X) = 0\} \quad = \quad \sup\{\delta : H^{\delta}(X) = \infty\}. \tag{3.19}$$

Związek pomiędzy przedstawionymi wymiarami jest postaci:

$$d_{\mathrm{C}} \geqslant d_{\mathrm{H}} \geqslant d_{\mathrm{T}}, \tag{3.20}$$

ale dla szerokiej klasy fraktali wymiary te są sobie tożsame. Pozwalają określić, jak szybko zmieniają się własności metryczne obiektu (długość, pole, objętość bądź jakakolwiek inna miara) w funkcji stosowanej w analizie skali.

3.3.1 Formalizm multifraktalny

Istnieją obiekty, tzw. *multifraktale*, dla których podanie samego wymiaru fraktalnego jest pewnym uproszczeniem. W obiektach tych nie tylko nośnik miary jest bowiem fraktalem, ale również sama miara ma charakter fraktalny. Multifraktale mogą być intuicyjnie rozumiane jako pewien splot wielu różnych struktur o własnych wymiarach fraktalnych. Tym samym ich właściwy opis nie może być realizowany przez jeden, uśredniony po wszystkich poziomach struktury parametr, jak to jest w przypadku omówionych wyżej wymiarów [117]. Ponieważ miara fraktalna jest niejednorodna, istnieje wiele możliwych jej rozkładów na danym nośniku [118]. Rodzi się zatem konieczność wprowadzenia dodatkowego formalizmu, który pozwoli na ilościowe rozróżnienie pomiędzy obiektami o różnym rozkładzie miary. Należy w tym celu poddać analizie lokalne własności obiektów [119].

Niech dla pewnego obiektu zachowanie wielkości S wokół jakiegoś punktu x_0 w skali ε będzie opisane przez lokalne prawo potęgowe:

$$S(x_0 + \varepsilon) - S(x_0) = \varepsilon^{\alpha(x_0)} \tag{3.21}$$

lub, równoważnie, w postaci:

$$\mu(S_{x_0}(\varepsilon)) = \varepsilon^{\alpha(x_0)}, \tag{3.22}$$

[7]Wymiar $d_{\mathrm{C}}^{\mathrm{up}}$ jest znany również jako *entropia Kołomogrowa* natomiast $d_{\mathrm{C}}^{\mathrm{up}}$ jako *wymiar dolny Minkowskiego*.

gdzie wielkość $\mu(S_{x_0}(\varepsilon))$ może być interpretowana jako rozkład gęstości miary wokół punktu x_0, a $\alpha(x_0)$ jest tzw. *wykładnikiem osobliwości*, opisującym zachowanie się tej miary wokół punktu x_0. Wtedy

$$\alpha(x_0) = \lim_{\varepsilon \to 0} \frac{\log \mu(S_{x_0}(\varepsilon))}{\log \varepsilon}. \tag{3.23}$$

Dla coraz bardziej osobliwego zachowania się miary $\mu(S)$ wykładnik $\alpha(x_0)$ maleje, osiągając zero dla dyskretnej reprezentacji $\mu(S)$[8].

Aby opisać globalny rozkład wartości $\alpha(x)$ dla danego obiektu, definiuje się dodatkowo *widmo osobliwości* $f(\alpha)$, które w typowych sytuacjach można interpretować jako wymiar Hausdorffa zbioru osobliwości o takim samym wykładniku $\alpha(x) = \alpha$:

$$f(\alpha) = d_{\mathrm{H}}(S(x)) \qquad \mathrm{dla}\{x : x \in \sup\mu(S), \alpha(x) = \alpha\}. \tag{3.24}$$

Widmo osobliwości osiąga maksimum przy najczęściej występującej wartości α. Wyznaczanie $f(\alpha)$ opiera się w praktyce na globalnych wielkościach, takich jak *funkcje rozdziału*:

$$Z(q, \varepsilon) = \sum_{i=1}^{N} \mu_i^q(\varepsilon), \tag{3.25}$$

gdzie N to liczba komórek histogramu widm osobliwości α dla $\varepsilon \to 0$, a $q \in \mathbb{R}$. Poprzez dobór odpowiednich wartości parametru q możliwe jest rozłożenie struktury ze względu na wartość gęstości miary $\mu(S)$. Jak pokazano wcześniej, miara probabilistyczna μ jest związana z wykładnikiem osobliwości α tak, że $\mu_i \equiv \varepsilon^{\alpha_i}$ (wzór 3.22), co daje rozkład wartości α dla przyjętej skali ε w postaci $\varrho(\alpha)\varepsilon^{-f(\alpha)}$. Podstawiając te wartości do powyższego wzoru, otrzymać można:

$$Z(q, \varepsilon) \cong \int_{x_{\min}}^{x_{\max}} \varrho(\alpha)\varepsilon^{q\alpha - f(\alpha)} d\alpha, \tag{3.26}$$

gdzie funkcja rozdziału $Z(q, \varepsilon)$ w granicy $\varepsilon \to 0$ ma charakter potęgowy:

$$Z(q, \varepsilon) \sim \varepsilon^{\tau(q)}, \tag{3.27}$$

a $\tau(q)$ to *uogólniony wykładnik skalowania*. Główny wkład do całki pochodzi od takiej wartości α, dla której wyrażenie $q\alpha - f(\alpha)$ przyjmuje wartość najmniejszą. Zatem $\tau(q)$ przyjmuje postać:

$$\tau(q) = \min_{\alpha}(q\alpha - f(\alpha)). \tag{3.28}$$

Stosując odwrotną transformację Lagrange'a, $f(\alpha)$ przybiera wartość:

$$f(\alpha) = \min_{q}(q\alpha - \tau(q)). \tag{3.29}$$

Konsekwencją (3.28) i (3.29) są następujące zależności:

$$\alpha(q) = \frac{d\tau(q)}{dq} \qquad \text{oraz} \qquad q = \frac{df(\alpha)}{d\alpha}. \tag{3.30}$$

[8]W takim przypadku miara $\mu(S)$ przybiera w punkcie x_0 kształt bliski funkcji Diraca $\delta(x - x_0)$.

Znając funkcje rozdziału $Z(q,\varepsilon)$ można wprowadzić tzw. *uogólniony wymiar fraktalny, wymiar Rényi'ego* D_q, zależny od ciągłego parametru q:

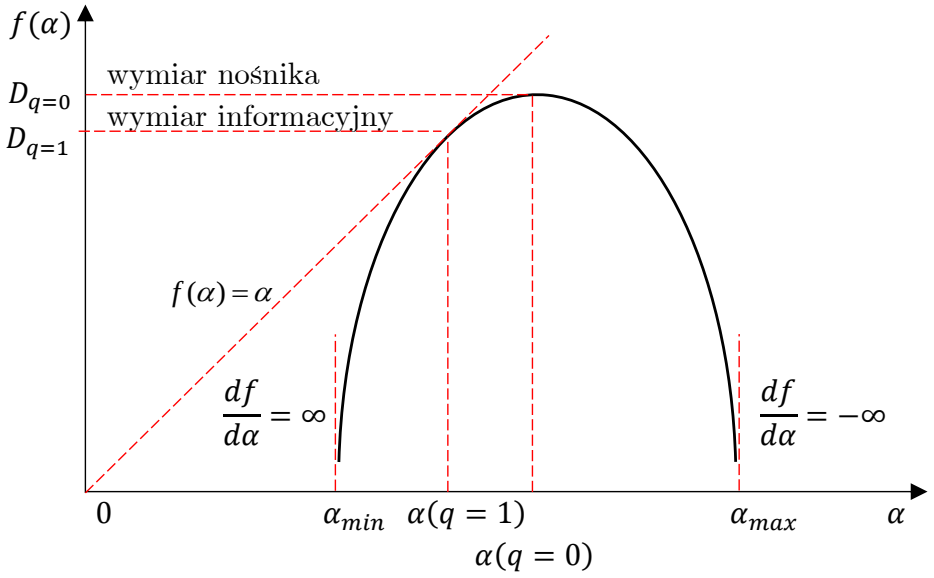
$$D_q = \frac{-1}{q-1} \lim_{\varepsilon \to 0} \frac{\log Z(q,\varepsilon)}{\log \varepsilon}, \tag{3.31}$$

co przy znajomości funkcji rozdziału można uprościć do postaci:

$$D_q = \frac{1}{q-1}(q\alpha - f(\alpha)) \qquad \text{czyli} \qquad D_q = \frac{\tau(q)}{q-1} \tag{3.32}$$

Analizując powyższą równość można dla odpowiednich wartości q wskazać pewne charakterystyczne wartości D_q. Dla $q \to \pm\infty$ funkcja rozdziału $Z(q,\varepsilon)$ jest zdominowana przez ekstremalne wartości $\mu(\varepsilon)$, dlatego brzegowe wartości α, dla których istnieje widmo osobliwości $f(\alpha)$, określają najsłabszą i najsilniejszą osobliwość. Dla $q = 1$ zachodzi równość: $\alpha = f(\alpha)$, a odpowiadająca temu wartość $D_{q=1}$ nazywana jest *wymiarem informacyjnym*. Z kolei dla $q = 0$ wartość $f(\alpha)$ osiąga maksimum, a związana z nią wartość $D_{q=0}$ jest wymiarem nośnika miary μ.

Szerokość widma $f(\alpha)$ świadczy o stopniu heterogeniczności analizowanego układu, czyli pewnej klasie multifraktalności. W przypadku istnienia jedynie pojedynczej osobliwości α całe powyższe widmo redukuje się do punktu $(\alpha_0, f(\alpha_0))$, a badany układ jest jednorodny, czyli monofraktalny.



Rysunek 3.3: Przykładowe spektrum multifraktalne $f(\alpha)$, z podaniem charakterystycznych rzutów na wartości argumentu α i funkcji $f(\alpha)$.

Rozdział 4

Charakterystyki złożoności języka naturalnego

4.1 Statystyczne charakterystyki złożoności języka naturalnego

4.1.1 Prawo Zipfa i inne prawa potęgowe

Rozkłady potęgowe zajmują uprzywilejowane miejsce w opisie różnych układów złożonych [120, 121]. Ich istnienie było obserwowane od dawna, ale dopiero powiązanie ich ze zjawiskami krytycznymi i związaną z nimi uniwersalnością sprawiło, że znalazły się w obszarze szczególnego zainteresowania. Istniejąc obok dobrze znanych z fizyki statystycznej rozkładów wykładniczych[1] $P(x) \propto e^{-\beta x}$, gdzie $a, x > 0$, rozkłady typu $P(x) \propto x^{-\alpha}$ znacznie lepiej oddają charakter wielu naturalnych układów bądź procesów. Jednym z takich układów okazał się być również język naturalny [123].

Dobrze znaną empiryczną prawidłowością dotyczącą statystycznych własności języka jest związek częstości słów f i ich rangi R, zwany *prawem Zipfa*. Ranga przyporządkowuje kolejne liczby naturalne elementom występującym z coraz mniejszą częstością [124]. Ilościowa analiza typowej próbki języka o odpowiednio dużej statystyce wykazuje potęgową relację ranga-częstość:

$$f(R) \propto \frac{1}{R^{\alpha}}, \tag{4.1}$$

gdzie α jest wykładnikiem, który przybiera uniwersalną wartość (≈ 1) dla większości języków posiadających zapis fonetyczny, natomiast w przypadku języków posługujących się zapisem ideograficznym, np. w j. chińskim, prawidłowości takiej aż tak wyraźnie się nie obserwuje [125]. Następnie można określić rozkład prawdopodobieństwa dla słów występujących z odpowiednią częstością f, który ma potęgową postać:

$$P(f) \propto f^{-\alpha'}. \tag{4.2}$$

Wykorzystując funkcyjną zależność pomiędzy częstością f a rangą R (wzór 4.1) oraz fakt, że gęstości rozkładów są takie same, czyli $P(f)df = P(R)dR$, otrzymać

[1] Rozkłady tego typu zostały wprowadzone do równowagowej fizyki statystycznej w celu opisu rozkładów prawdopodobieństwa mikrostanów [122].

można prawdopodobieństwo wystąpienia słowa o randze R:

$$P(R) = P(f)\frac{df}{dR}, \tag{4.3}$$

na podstawie zależności (4.1), (4.2) i (4.3) można wyznaczyć:

$$P(R) \propto R^{-\alpha} \tag{4.4}$$

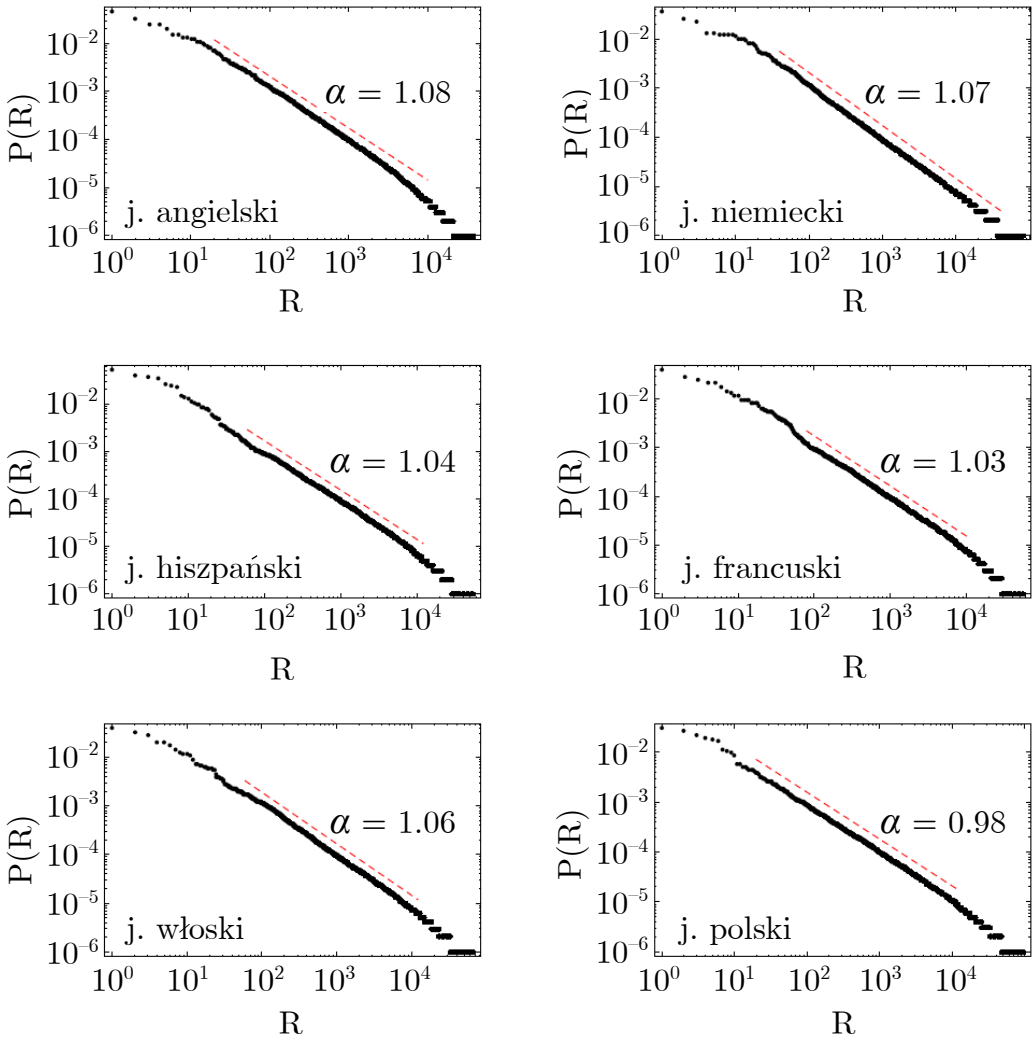
gdzie związek pomiędzy wykładnikami dla obu rozkładów to $\alpha' = 1+1/\alpha$. Rozkłady $P(R)$ oraz $P(f)$ są znane, kolejno, jako prawo Zipfa i odwrotne prawo Zipfa [126].

Jak się okazuje, rozkłady tego typu są obecne także poza językiem w obszarach, w których występują zdarzenia ekstremalne. Do takich zależności należą m.in.: rozkład wielkości miast – *prawo Gibrata*, rozkład cytowań artykułów naukowych – *rozkład Lotki* czy rozkład dochodów ludzi w gospodarce wolnorynkowej – *prawo Pareto* [123, 127].

Na rysunku 4.1 zaprezentowano rozkłady Zipfa dla sześciu korpusów, stworzonych w oparciu o zbiory tekstów, zawierających ok. 10^6 słów. Korpusy te powstały w wyniku scalenia reprezentatywnej próbki literatury napisanej w danym języku, wykazanej w dodatku A. Dla każdego przedstawionego języka obserwowany jest potęgowy rozkład prawdopodobieństwa z wykładnikiem α bliskim jedności.

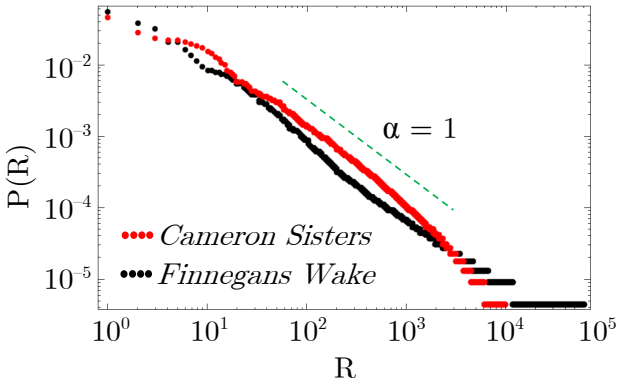
Powyższe rozkłady są przejawem hierarchiczności, obecnej w hiperbolicznej zależności ranga – częstość. Stosunkowo mały procent słów o najniższych rangach (np. słowa o rangach $1 \leqslant R \leqslant 20$ stanowią około 1/5 objętości całego języka) odgrywa kluczową rolę w organizacji języka, chociaż typowo nie są to jednostki o charakterze znaczeniowym, to stanowią one nośnik niektórych reguł gramatycznych [128]. Słowa odpowiadające coraz wyższym rangom obserwuje się coraz rzadziej, niosą ze sobą konkretne znaczenie, określając obiekty, czynności, pojęcia. Dla bardzo rzadkich słów o rangach $R > 10^4$ widzimy, że następuje załamanie jednorodnego skalowania, co jest efektem typowym dla korpusów tekstów, gdzie wymieszanych jest wiele różnych próbek języka [129]. Efekt ten wynika z różnic w słownictwie, wykorzystywanym w różnych próbkach i najczęściej bywa tłumaczony jako istnienie dwóch zbiorów słownictwa: podstawowego, wspólnego dla wszystkich autorów, obejmującego kilka tysięcy słów, oraz indywidualnego, zawierającego słowa specjalistyczne bądź charakterystyczne dla stylu konkretnego autora, neologizmy, nazwy własne, czy nawet zwykłe błędy, nierozpoznane przez narzędzia automatycznej analizy [130, 131].

Odstępstwa od typowego dla danego języka skalowania w obszarze podstawowym (skalowania Zipfa) są interpretowane w zależności od wartości wykładnika potęgowego. Nadreprezentacja słów występujących dosyć rzadko, co odpowiada $\alpha_{\text{tekst}} < \alpha_{\text{jezyk}}$, może być uznana za przejaw bogactwa słownictwa i kunsztu piszącego, o ile nie jest to wynik umyślnego działania, natomiast sytuacja przeciwna, tj. $\alpha_{\text{tekst}} > \alpha_{\text{jezyk}}$ może świadczyć o problemach rozwojowych lub określonej dysfunkcji mózgu np. schizofrenii [132]. Należy jednak zwrócić uwagę, że zaprezentowane na rysunku 4.1 rozkłady nie są identyczne dla wszystkich języków. Powodem jest istnienie fleksji, która zwiększa liczbę różnych słów, nie zwiększając de facto zasobu słownictwa [133]. Mimo że język angielski, uznawany za najbogatszy słownikowo [134], jest stanowiony przez $\approx 10^6$ różnych słów, w obrazie analizy zipfowskiej może być bardzo podobny do innego, ale już fleksyjnego języka, np. polskiego [135].



Rysunek 4.1: Rozkłady Zipfa $P(R)$, opisujące prawdopodobieństwo $P(R)$ wylosowania słowa o randze R, dla wybranych języków europejskich. Czerwoną przerywaną linią wyznaczono nachylenie, określone wykładnikiem α.

Na rysunku 4.2 przedstawiono rozkład Zipfa dla dwóch porównywalnych objętościowo tekstów literackich, zawierających po ponad $2{\cdot}10^4$ różnych słów. Jednym z nich jest książka *Cameron Sisters* Cathy Maxwell, będąca przykładem literatury lekkiej (literatura romansu), natomiast drugim – książka *Finnegans Wake* Jamesa Joyce'a, o wybitnie eksperymentalnym charakterze. Co zwraca uwagę, to odmienne zachowanie rozkładów dla obu książek [136]. Pierwszy z nich posiada nachylenie wyraźnie większe niż na ogół obserwowane dla języka angielskiego, z małą liczbą rzadkich słów.



Rysunek 4.2: Rozkłady Zipfa dla dwóch porównywalnych objętościowo książek: kolor czerwony - *Cameron Sisters* C. Maxwell, kolor czarny - *Finnegans Wake* J. Joyce'a. Różne zachowanie rozkładów $P(R)$ świadczy o bogactwie słownictwa znajdującego się w konkretnej książce.

Świadczy to o potocznym, zubożonym słownictwie, charakterystycznym dla reprezentowanego gatunku, niosącym ze sobą niską wartość ideową oraz artystyczną. Brak skalowania się rozkładu $P(R)$ jest manifestacją stosunkowo równowagowej (niepotęgowej) statystyki słów, która już na tym etapie analizy pozwala ocenić charakter badanego tekstu. Zupełnie inną reprezentację zipfowską posiada dzieło Jamesa Joyce'a, które jest jednym z najciekawszych ale i kontrowersyjnych dzieł światowego pisarstwa, o ogromnym, niespotykanym w utworach o podobnej objętości bogactwie leksykalnym, w tym zwłaszcza słów użytych w tekście tylko raz. Mimo rozmaitych analiz literaturoznawczych, utwór ten ciągle umyka jednoznacznej interpretacji, ujawniając skomplikowaną strukturę związana z techniką *strumienia świadomości* [137, 138].

Charakterystyczną cechą rozkładów potęgowych, np. typu Zipfa jest brak określonych momentów dla tych rozkładów, gdzie wartości pierwszego i drugiego:

$$\langle R\rangle = \int_1^\infty RP(R)dR = C\int_1^\infty R^{-\alpha+1}dR \tag{4.5}$$

$$\langle R^2\rangle = \int_1^\infty R^2P(R)dR = C\int_1^\infty R^{-\alpha+2}dR, \tag{4.6}$$

mają nieskończoną wartość, co oznacza, że określenie średniej rangi $\langle R\rangle$ dla takiego rozkładu czy odchylenia standardowego $\sigma^2 = \langle R^2\rangle - \langle R\rangle^2$ jest pozbawione sensu[2].

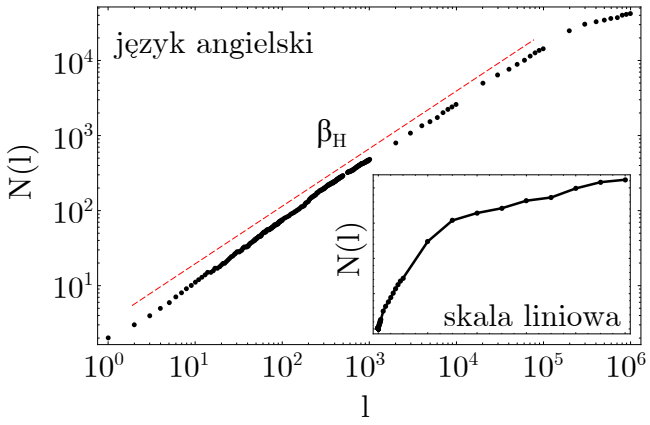
[2]Mimo że pierwszy moment rozkładów typu $P(x) \propto x^{-\alpha}$ dla $\alpha > 2$ ma określoną wartość, tym niemniej niepewność jego oszacowania nadal jest nieskończona, $\sigma \to \infty$.

Uogólniając, wszystkie momenty $\langle R^n \rangle$, gdzie $n \geqslant \alpha - 1$ dla rozkładów typu $P(R) \propto x^{-\alpha}$ są nieskończone. Cecha ta jest immanentną własnością wielu układów złożonych, stanowiącą o ich strukturze, zasięgu oddziaływania, skalach czasowo-przestrzennych.

Prawo Zipfa nie jest jedyną potęgową zależnością obserwowaną w języku naturalnym. Inną powszechnie spotykaną zależnością tego typu jest potęgowy wzrost liczby różnych słów w miarę zwiększania się rozmiaru tekstu, znaną jako *prawo Heapsa* lub *prawo Herdana* [139, 140] którą można przedstawić w postaci:

$$N(l) = gl^{\beta_{\mathrm{H}}}, \tag{4.7}$$

gdzie $N(l)$ jest liczbą różnych słów w tekście o danej długości l, a $0.5 \leqslant \beta_{\mathrm{H}} \leqslant 0.8$. Potęgowy reżim w prawie Heapsa jest utrzymywany (na ogół) przez 3 – 4 rzędy wielkości wartości l, jednak ograniczony zbiór słów funkcjonujących w danym języku lub znanych autorowi prowadzi do powolnego wysycenia, tak że dla $l \to \infty$ istnieje efektywny zanik tempa dodawania nowych słów, $dN(l)/dl \to 0$ (patrz rysunek 4.3). W ten sposób w miarę wzrostu objętości tekstu, pojawianie się kolejnych słów jest realizowane coraz bardziej w przestrzeni tych, które już choć raz wystąpiły do chwili l. Rozkład częstości słów w tekście ma na ogół charakter stacjonarny, wybór l nie wpływa na charakter tego rozkładu. Prawo Heapsa do niedawna było uważane jako niezależna od prawa Zipfa własność języka, dziś już jednak wiadomo, że jest ono bezpośrednią konsekwencją prawa Zipfa [141].



Rysunek 4.3: Prawo Heapsa dla języka angielskiego. Zależność pomiędzy długością tekstu, a liczbą występujących w nim unikalnych słów jest zależnością potęgową, z określonym wykładnikiem β_H. Dla dużych wartości l widoczne jest wysycenie związane ze skończoną liczba słów. Wykres w wstawce prezentuje tę samą zależność w skali liniowo-liniowej.

Mechanizm odpowiedzialny za potęgowy rozkład prawdopodobieństwa w języku naturalnym nie został dotąd jednoznacznie zidentyfikowany. Istnieje kilka hipotez, pozwalających jakościowo odtworzyć obserwowane charakterystyki, uwzględniając różne czynniki, które mogą prowadzić do takich charakterystyk. Odkrywca tej zależności G.K. Zipf wyjaśniał ją jako bezpośrednią konsekwencję *zasady najmniejszego wysiłku* [77]. Wedle niej rozkład częstości używania słów przez osobę tworzącą wypowiedź jest efektem jej optymalizacji pod kątem spełnienia dwóch przeciwstawnych warunków: uczynienia komunikatu możliwie wygodnym do stworzenia

(przez użycie możliwie małej liczby słów) oraz wystarczająco precyzyjnym do poprawnego odbioru (przez użycie wystarczająco dużej liczby pojęć). Zasada ta, sformułowana przez Zipfa czysto jakościowo, została wyrażona matematycznie przez Benoita Mandelbrota jakiś czas później [142, 143, 144], a niedawno powiązana formalnie z prawem Zipfa [145]. Zostały też zaproponowane inne teoretyczne mechanizmy, które mogą prowadzić do pojawienia się rozkładów Zipfa w języku naturalnym [146, 147, 148, 149].

Rozumowanie podane przez Mandelbrota łatwo odnieść do procesu stochastycznego opartego na probabilistycznym mechanizmie tzw. sporadycznej ciszy (ang. *intermittent silence*), zaproponowanym przez G.A. Millera [150]. Proces generowania tekstu polega w nim na losowym pojawianiu się liter z ustalonego alfabetu o długości $N > 1$, z jednakowym prawdopodobieństwem $(1 - p_{\rm sp})/N$ pojawienia się każdej litery i z określonym prawdopodobieństwem $p_{\rm sp}$ użycia spacji. Wszystkie zdarzenia są niezależne, stąd jest to realizacja *łańcucha Markova zerowego rzędu*. Pojawienie się słowa o długości d dane jest zatem przez:

$$p_d = p_{\rm sp} \left(\frac{1 - p_{\rm sp}}{N} \right)^d , \tag{4.8}$$

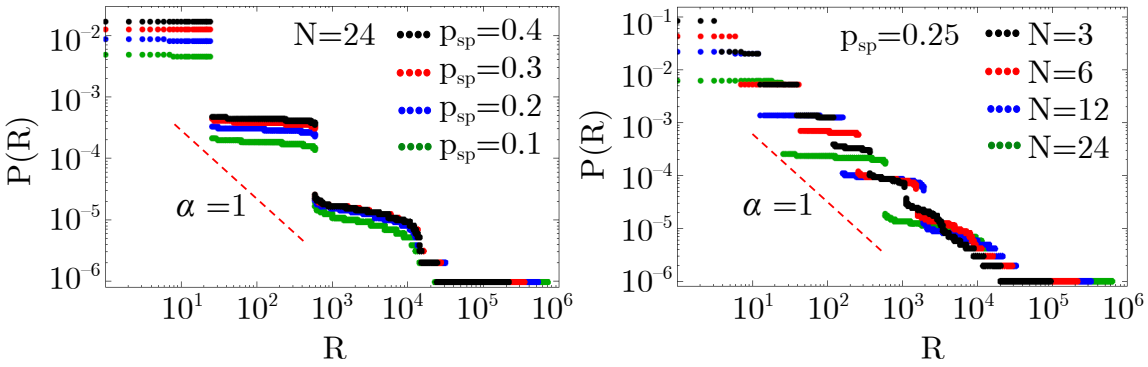
gdzie wszystkich słów o długości d jest N^d. Zgodnie z tym słowo o długości d posiada range R z przedziału:

$$\frac{N^d + N - 2}{N - 1} \leqslant R \leqslant \frac{N^{d+1} - 1}{N - 1}. \tag{4.9}$$

Przyjmując dla $N \gg 1$ przybliżenie $R \approx N^d$, otrzymuje się:

$$P(R) \approx p_{\rm sp} \left(\frac{1 - p_{\rm sp}}{N} \right)^{\log_N R} \quad \text{lub równoważnie} \quad P(R) \approx p_{\rm sp} R^{\log_N (1 - p_{\rm sp}) - 1}, \tag{4.10}$$

co jest tożsame rozważanemu prawu Zipfa.



Rysunek 4.4: Rozkład Zipfa dla sztucznie generowanych tekstów. Rozkłady $P(R)$ w prawym panelu, opisują tekst będący sekwencją liter losowanych z 24 - elementowego alfabetu, z określonym prawdopodobieństwem pojawienia się spacji p_{sp}. W prawym panelu analogiczny rozkład $P(R)$ dla tekstów będących sekwencją liter losowanych z N - elementowego alfabetu, z określonym prawdopodobieństwem $p_{sp} = 0.25$.

Na rysunku 4.4 przedstawiono symulacje opisanego procesu, generującego tekst o długości $l = 10^6$ wyrazów. Prawdopodobieństwo wystąpienia słów o długości d wynosi $P_d = \prod_{i=1}^{d} p_i p_{\mathrm{sp}}$, gdzie $p_i = N^{-1}$. Przyjęcie rzeczywistego (lewy wykres) i ograniczonego (prawy wykres) rozmiaru alfabetu prowadzi do dyskretnych wartości oczekiwanego prawdopodobieństwa wystąpienia wyrazów jedno-, dwu- i N - literowych. Mimo że model nie uwzględnia jakichkolwiek czynników słowotwórczych (np. możliwość wymowy, skończona długość wyrazów), dla obu przypadków skalowanie jest utrzymane w rygorze prawa Zipfa [146]. Choć jakościowo rozkład $P(R)$ przejawia potęgowy charakter, to otrzymanie empirycznej wartości wykładnika α obserwowanego dla języka naturalnego, gdzie $N = 26$, jest bardzo trudne. Wykładnik ten przyjmie bowiem wartość jedności, gdy:

$$\log_{N=26}(1 - p_{\mathrm{sp}}) = 0 \Leftrightarrow p_{\mathrm{sp}} \approx 0. \tag{4.11}$$

W związku z tym tylko dla bardzo małych wartości p_{sp} model byłby zbieżny z rzeczywistością, tymczasem prawdopodobieństwo „spacji" dla tekstów rzeczywistych jest istotnie większe i wynosi $0.15 \leqslant p_{\mathrm{sp}} \leqslant 0.3$. Proces ten jest czysto mechanistyczną realizacją, bazującą jedynie na zbiorze istniejących znaków z założonym a priori stałym prawdopodobieństwem, nie uwzględniając żadnych innych czynników mogących mieć istotny wpływ na strukturę tekstu. Słowa mogą mieć dowolną długość, dozwolona jest każda permutacja liter w obrębie danej sekwencji pojawiających się znaków. Dlatego też model ten często jest określany mianem „małpy przy klawiaturze" [151].

Z punktu widzenia nadawcy, język powinien być realizowany w jak najbardziej ekonomiczny sposób, zawierając minimalną liczbę słów, charakteryzujących się niskim kosztem użycia, zdefiniowanym jako $C_R = \log_N R$. Biorąc pod uwagę zależność $R = N^d$, koszt użycia słowa $C_R \sim d_R$, gdzie d_R to długość słowa o randze R, utworzonego z N-literowego alfabetu [152].

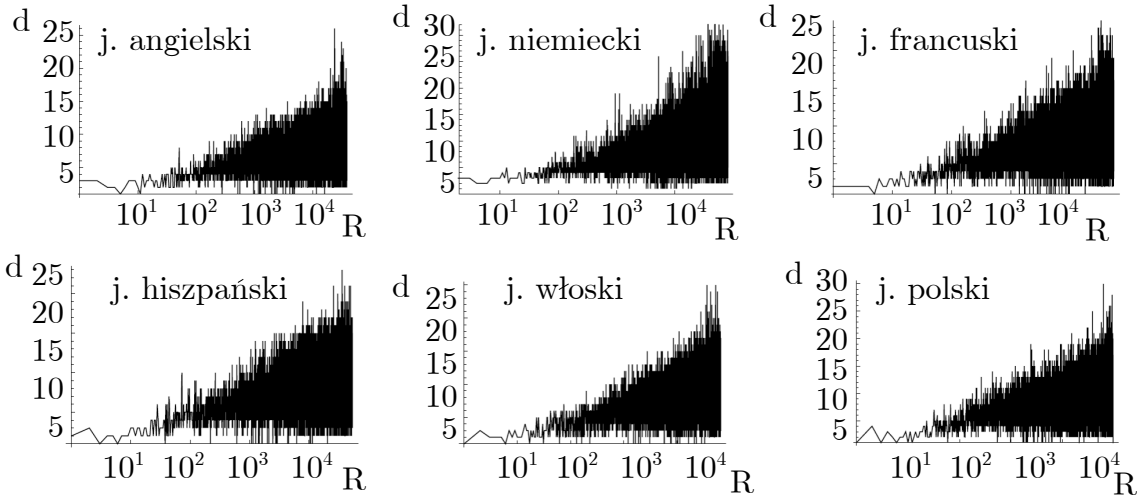
Faktycznie, taka zależność jest obserwowana (rysunek 4.5) i dla modelu Millera, i dla języka naturalnego. Uśredniając po całym słowniku, średni koszt poniesiony na słowo będzie równy: $C = \sum_{R=1}^{\max} p_R C_R$. Z punktu widzenia odbiorcy komunikat powinien zawierać jak największą ilość informacji, co może być wyrażone za pomocą entropii Shannona: $H = -\sum_{R=1}^{\max} p_R \log p_R$. Każde kolejno pojawiające się słowo niesie ze sobą pewną nową informację, zmniejszając entropię układu, czyli uściślając przekazywany komunikat. Przyjęcie obu powyższych założeń jest możliwe jedynie w sytuacji, w której nie występują inne istotne czynniki, mogące wpłynąć na strukturę wypowiedzi. Do takich czynników należą sytuacje ekstremalne (np. zagrożenie życia nadawcy) – wówczas koszt wypowiedzi będzie nadmiernie minimalizowany lub komunikaty matka→dziecko – wówczas przekazywana informacja jest nadmiarowa.

Biorąc pod uwagę typowy charakter przekazywanej informacji pomiędzy nadawcą i odbiorcą, optymalizacja języka oznacza, że iloraz C/H będzie jak najmniejszy:

$$\frac{\partial}{\partial p_R}\left(\frac{\sum_{R=1}^{\max} p_R C_R}{-\sum_{R=1}^{\max} p_R \log p_R}\right) = 0, \tag{4.12}$$

w wyniku czego dostaje się: $p_R = Ae^{-HC_R/C}$. Uwzględniając, że $C_R = \log_N R$, dostaje się związek:

$$p_R \sim R^{-(H \log_N 2)/C}. \tag{4.13}$$



Rysunek 4.5: Długość słów występujących w danym języku w zależności od rangi R. Dla każdego języka obserwowany jest statystyczny wzrost długości słowa w funkcji wartości jego rangi R, co jest zgodne z tezą optymalizacji kosztów ponoszonych w procesie nadawania komunikatu.

Zależność ta jest jakościowo tożsama z rozkładem Zipfa o wykładniku $\alpha = \frac{H}{C} \log_N 2$. Rozmaite warianty powyższego modelu w dobrym stopniu odzwierciedlają zachowanie języka naturalnego, a wprowadzane modyfikacje (np. uwzględnienie wag pomiędzy oboma procesami) prowadzą do zmiany nachylenia rozkładu $P(R)$.

Inną grupą procesów prowadzących do rozkładów zipfowskich są procesy oparte na mechanizmie przyłączeń uprzywilejowanych, zwanych *procesami Yule'a-Simona*. Powstający tekst jest wynikiem dodawania kolejnych elementów (wyrazów), wybieranych na dwa różne sposoby: z pewnym prawdopodobieństwem $p_n > 0$ dodawane jest nowe słowo do tekstu, natomiast z prawdopodobieństwem $p_s = 1 - p_n$ słowo, które już zostało dodane wcześniej. Dodatkowym założeniem narzuconym na p_s jest uwzględnienie częstości już zaistniałych słów, stąd prawdopodobieństwo wybrania i-tego słowa w kroku l to:

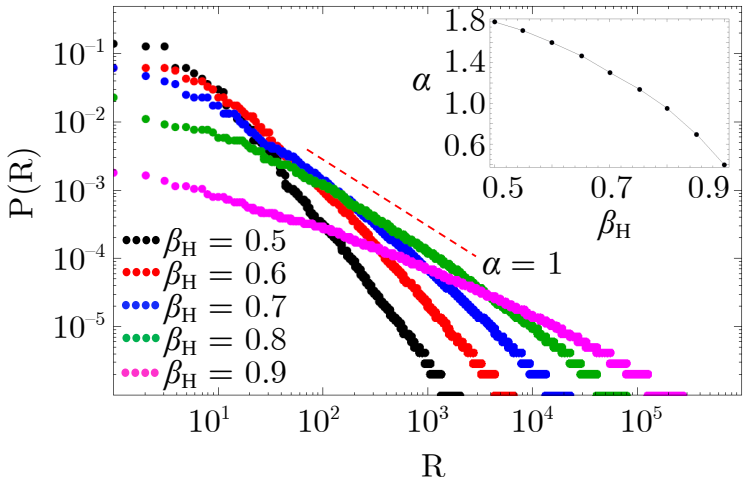
$$p_s(f_i) = \frac{f_i}{\sum_{i=1}^{l} f_i}, \tag{4.14}$$

gdzie mianownik jest równy liczbie wykonanych kroków l. W pierwotnej postaci tego procesu, zaproponowanej przez H.A. Simona, prawdopodobieństwo p_n dodawania nowych słów było stałe [153], ale jest to w oczywisty sposób sprzeczne z faktem wysycania się zasobów słownikowych języka w miarę wzrostu długości tekstu. Prawdopodobieństwo dodania nowego słowa musi więc być pochodną po czasie prawa Heapsa, $dN(l)/dl$, co jest tożsame z tempem pojawiania się nowych słów. Uwzględniając wzory 4.7 i 4.14, odpowiednie prawdopodobieństwa wynoszą:

$$p_s(l) = f_i/l \qquad p_n(l) = g\beta_{\mathrm{H}} l^{\beta_{\mathrm{H}}-1}. \tag{4.15}$$

Dla wartości $\beta_{\mathrm{H}} < 1$ granica $\lim_{l\to\infty} p_n(l) = 0$, a prawdopodobieństwo pojawienia się kolejnego słowa jest proporcjonalne do jego frakcji w rozważanym tekście.

Przedstawiony proces podlegający opisanym mechanizmom podlega rozkładom typu $P(f) \propto f^{-\alpha'}$, gdzie znowu wykorzystując fakt istnienia stałej gęstości prawdopodobieństwa $P(f)df = P(R)dR$ – bezpośrednio otrzymuje się rozkład Zipfa. Mimo że proces Yule'a-Simona-Heapsa opiera się na kilku prostych założeniach, nie uwzględniając żadnych efektów, np. związanych z pamięcią prowadzi do uzyskania potęgowych rozkładów, o nachyleniach zbliżonych do obserwowanych w rzeczywistych tekstach.



Rysunek 4.6: Rozkład Zipfa sporządzony w oparciu o sztucznie generowane teksty wedle mechanizmu Yule'a-Simona-Heapsa. Parametr β_H określa tempo dodawania nowych słów do tekstu, gdzie wraz ze wzrostem jego wartości udział nowych słów jest coraz większy. Wstawka w prawym górnym rogu opisuje relację pomiędzy wykładnikiem Heapsa β_H a wykładnikiem α, opisującym rozkład Zipfa.

Na rysunku 4.6 przedstawiono rozkład Zipfa dla sztucznych tekstów o długości $l = 10^6$ wyrazów każdy, powstałych w wyniku przeprowadzenia symulacji, w której przyjęto określoną wartość β_{H}. W prawym górnym rogu tego rysunku przedstawiono zależność pomiędzy wykładnikiem rozkładu Zipfa α w zależności od wykładnika β_{H}, związanego z rozkładem Heapsa. Dla wartości $\beta_{\mathrm{H}} \approx 0.7 \div 0.8$ obserwowany jest rozkład o wykładniku α zbliżonym do obserwowanego w rzeczywistych tekstach. Mimo że proces ten jest oparty wyłącznie o dwa proste złożenia, dobrze oddaje statystyczny charakter języka. W oparciu o ten model zachowane zostało również ugięcie wykresu dla małych wartości R, będące modyfikacją rozkładu zipfowskiego dokonaną przez Mandelbrota:

$$P(R) \propto (R + \rho)^{-\alpha}, \tag{4.16}$$

gdzie wartość czynnika $\rho \approx 1 \div 3$, a dla $\rho \ll R$ rozkład staje się pierwotnym prawem Zipfa. Odchylenie to jest konsekwencją istnienia małej liczby równoważnych słów spełniających rolę stricte gramatyczną (np. w angielskim są to przedimki *a, an, the*, a w niemieckim – rodzajniki *der, die, das*).

Istnieje szereg innych procesów prowadzących do prawa Zipfa [121], co może świadczyć, że sam rozkład nie daje istotnej informacji o mechanizmach rządzących strukturą języka naturalnego. Rozkład jest niezmienniczy względem dowol-

nej permutacji elementów (słów), co istotnie spłyca wartość informacyjną dla języków pozycyjnych[3]. Istnieje zatem potrzeba wprowadzenia narzędzi, które nie tylko uwzględnią tak rozumianą hierarchiczność, ale również wzajemne relacje pomiędzy słowami [154, 155]. Relacje te, narzucone przez reguły gramatyczne, nie tylko sprawiają, iż język staje się efektywnym narzędziem komunikacji, ale również pozwalają tworzyć nieskończoną mnogość wyrafinowanych form ze skończonej liczby elementów [156]. Gramatyka odpowiada również za oddziaływania między słowami, co tym bardziej wskazuje na konieczność rozszerzenia analizy.

4.2 Sieci lingwistyczne

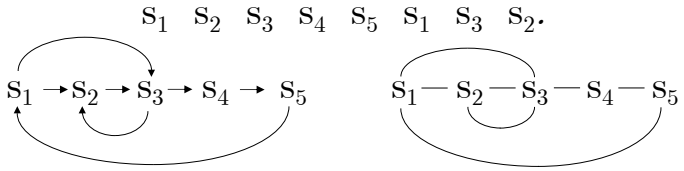
4.2.1 Modele dynamiki sieci ekspandujących

Interesującym podejściem byłby model, który równocześnie uwzględniałby dwie istotne własności języka: statystykę poszczególnych słów oraz relacje pomiędzy nimi. Okazuje się, że dopiero na poziomie fraz, czy dalej zdań, ujawniają się nieredukowalne informacje dotyczące znaczenia i kontekstu danego przekazu [157]. Wygodnym obrazem uwzględniającym te właściwości może być reprezentacja sieciowa. Pozwala ona nie tylko odzwierciedlać zróżnicowanie słownikowe, ale również poprzez swoją topologię – zawrzeć informacje o strukturze języka, co nie było możliwe na podstawie czysto statystycznej analizy zipfowskiej.

Jedna z możliwych topologii takich sieci oparta jest na strukturze liniowej zdań, rysunek 2.3. W ramach gramatyki skończenie stanowej sekwencja słów jest formą użytej gramatyki, niosącą ze sobą informację, której na ogół nie można jednoznacznie wyekstrahować ze zbioru wyrazów tworzących dane zdanie[4]. Dla sekwencji słów s_i, tworzącej np. pojedyncze zdanie, można skonstruować sieć przedstawioną na rysunku 4.7. Stopień wierzchołków w takiej sieci zależy od przyjętej definicji krawędzi. Jeśli orientują one pary wierzchołków względem siebie, mówimy o, odpowiednio, stopniu wejściowym k_{in} i wyjściowym k_{out}, natomiast dla sieci o takich samych, ale nieskierowanych połączeniach, stopień wierzchołka ma tylko jeden typ: $k = k_{\text{in}} + k_{\text{out}}$. Wraz ze wzrostem wielkości analizowanej próbki języka może pojawić się sytuacja, że dwa uprzednio sąsiadujące ze sobą słowa pojawiły się raz kolejny. Prowadzi to albo do konieczności wprowadzenia krawędzi wielokrotnych, a otrzymana w ten sposób sieć będzie miała charakter ważony, albo do traktowania relacji pomiędzy wierzchołkami binarnie, tzn. kolejne pojawienie się danej pary sąsiadujących wyrazów nie powoduje przyrostu stopni tych wierzchołków. Pierwsze podejście jest tożsame z analizą zipfowską, gdyż taki jest wówczas rozkład krotności wierzchołków, natomiast podejście drugie ma odmienny charakter. W pracy przyjęto konwencję rozróżniającą słowo ze względu na jego fleksję. Uczyniono tak z dwóch powodów: czysto praktycznego, bo idealna redukcja tekstu do wyrazów w formach podstawowych jest trudna do zalgorytmizowania, oraz semantycznego, bo słowo odmienione niesie ze sobą na ogół inną informacje niż jego forma podstawowa.

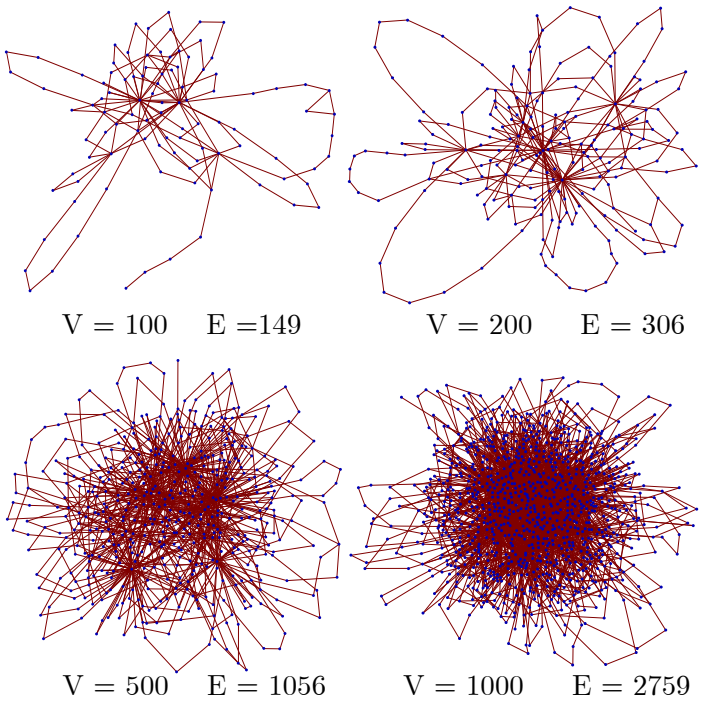
[3] Język pozycyjny to język, w którym gramatyka warunkuje porządek słów w zdaniu; przykładem jest np. język angielski.

[4] Mimo oczywistej różnicy, dla wygody terminy: *słowo* i *wyraz* będą używane zamiennie.



Rysunek 4.7: Sieć skierowana oraz sieć nieskierowana, zbudowane na postawie sekwencji wyrazów w ośmiowyrazowym zdaniu. Pojawienie się nowego wyrazu, nie obserwowanego w sekwencji s_i powoduje dodanie nowego elementu do sieci. Pojawienie się słowa po raz kolejny, skutkuje pojawieniem się połączeń pomiędzy nim a sąsiadującymi z nim słowami.

Wobec tego punktem wyjścia jest analiza języka istniejącego w formie rzeczywistej, a nie zredukowanej do postaci słownikowej. Na rysunku 4.8 przestawiono graficzną postać sieci, powstałą w wyniku przetworzenia fragmentu tekstu zaczerpniętego z książki *Ulysses* Jamesa Joyce'a. Wraz ze wzrostem liczby wierzchołków rośnie liczba krawędzi, ale co ciekawe, nie jest to wzrost liniowy, lecz przyspieszony. Jest to naturalna konsekwencja wcześniej wspomnianego prawa Heapsa, wiążącego liczbę różnych słów z długością tekstu: dla coraz dłuższej sekwencji wyrazów sieć staje się coraz bardziej wypełniona (pojawia się coraz więcej wewnętrznych połączeń). Takie wypełnianie luk w połączeniach międzywęzłowych jest charakterystyczne dla pewnej klasy sieci ewoluujących (o zmiennej topologii), rozważanych w literaturze nie tylko w kontekście lingwistycznym.



Rysunek 4.8: Wzrost sieci nieskierowanej składająca się z V wierzchołków i E krawędzi. Kolejne etapy ekspansji sieci pokazują przyspieszony charakter wzrostu.

Jak się okazuje, istnieje dużo szerszy zakres zjawisk, mogących mieć podobną dynamikę, w której występuje więcej niż jeden czynnik odpowiedzialny za zmiany strukturalne sieci. W takich przypadkach sieci ewoluujące są znacznie bliższą reprezentacją układów naturalnych niż ich statyczne odpowiedniki. W celu opisu dynamiki sieci lingwistycznych rozważmy dwa modele sieciowe. Pierwszy z nich to powszechnie znany, zaproponowany przez Dorogowcewa i Mendesa model sieci o przyśpieszonym wzroście, oparty ma idei preferencyjnego przyłączania [98]. Mechanizm ten, wywodzący się z formalizmu zaproponowanego przez Barabásiego, zakłada że podczas wzrostu sieci istnieje możliwość dodawania krawędzi nie tylko za pośrednictwem nowych wierzchołków, ale również pomiędzy wierzchołkami już istniejącymi.

W modelu minimalnym tempo dodawania nowych krawędzi jest stałe w czasie (dyskretny upływ czasu jest mierzony poprzez dodanie nowego wierzchołka do sieci). Niech w każdym kroku dodawany jest nowy węzeł preferencyjnie łączący się z już istniejącymi poprzez m krawędzi. W tym samym czasie pojawia się ct nowych krawędzi łączących stare wierzchołki z prawdopodobieństwem $\pi_s = k_i k_j$. Stosując podejście czasu ciągłego, średni stopień wierzchołka[5], który pojawił się w kroku s, zmienia się w czasie t wedle równania:

$$\frac{\partial \overline{k}(s,t)}{\partial t} = (m + 2cmt)\frac{\overline{k}(s,t)}{\int_0^t \overline{k}(u,t)du}, \tag{4.17}$$

z warunkiem początkowym $\overline{k}(0,0) = 0$ oraz brzegowym $\overline{k}(t,t) = m$. Suma wszystkich stopni po czasie t wynosi:

$$\int_0^t \overline{k}(u,t)du = 2mt + cmt^2. \tag{4.18}$$

Podstawiając tę całkę do mianownika równania (4.17) i rozwiązując równanie, otrzymuje się średni stopień wierzchołka:

$$\overline{k}(s,t) = m\left(\frac{t}{s}\right)^{\frac{1}{2}}\left(\frac{2+ct}{2+cs}\right)^{\frac{3}{2}}, \tag{4.19}$$

przy czym rozkład prawdopodobieństwa związanego z jego wystąpieniem w sieci:

$$P(k,t) = \frac{1}{ct}\frac{cs(2+cs)}{1+2cs}\frac{1}{k}, \tag{4.20}$$

gdzie $s = s(k,t)$ jest rozwiązaniem równania (4.19). Uzyskany rozkład $P(k,t)$ nie jest stacjonarny i dla małych wartości stopni k, gdzie $s \approx t$, rozkład przybiera postać $P(k) \approx \frac{1}{2}k^{-3/2}$. Dla starych wierzchołków, o dużych wartościach stopni, gdzie $s \ll t$, rozkład ma postać: $P(k) \simeq \frac{1}{4}(ct)^3 k^{-3}$. Charakterystyczna wartość stopnia k, przy której jeden rozkład przechodzi w drugi, można uzyskać, porównując je ze sobą. W wyniku tego dostaje się: $k_\times \approx \sqrt{ct}(2+ct)^{3/2}$.

W porównaniu z modelem BA, model ten pozwala na łączenie się już istniejących wierzchołków, co bezpośrednio prowadzi do zwiększenia się stopni tych wierzchołków, które posiadają już istotną krotność. Opisany proces potęguje jeszcze bardziej

[5]Wprowadzenie dla i-tego wierzchołka średniego stopnia k_i jest tutaj możliwe ze względu na wykładniczy rozkład prawdopodobieństwa [158].

zjawisko preferencyjnego przyłączania, które było jednym z istotnych założeń narzuconych na modele generujące sieci. Otrzymany w ten sposób potęgowy rozkład stopni wierzchołków $P(k)$ lepiej odzwierciedla strukturę układów naturalnych, przejawiających na ogół złożoną dynamikę ekspansji [159]. Jedynym wolnym parametrem w rozważanym modelu jest tempo wzrostu liczby krawędzi, jest to zatem w tym sensie model minimalny: przyspieszony wzrost stopni wierzchołków jest konsekwencją przyjętego tempa procesu, a nie wynika z zastosowania innych modyfikacji preferencyjnego przyłączania.

Uogólnieniem powyższego modelu jest model pozwalający na dowolny potęgowy, a nie tylko liniowy, przyrost stopni wierzchołków. Uogólnienie to zostało również zaproponowane przez Dorogowcewa i Mendesa i w oryginalnej pracy znalazło ono zastosowanie do opisu sieci modelujących rozkłady bogactwa w gospodarce nieinterwencyjnej (kapitalistycznej) [98]. Na potrzeby niniejszej pracy model niech nosi oznaczenie DM-AG (ang. *accelerated growth*). Wzrost sieci jest w nim indukowany przez te same dwa mechanizmy, co w wersji liniowej: dodawanie nowych wierzchołków i dodawanie krawędzi pomiędzy już istniejącymi wierzchołkami. Kluczowa jest relacja pomiędzy nimi, która jest zmienna w czasie: pierwszy mechanizm dominuje na początku formowania się sieci, natomiast drugi zaczyna dominować po odpowiednio długim czasie. Tempo wzrostu średniego stopnia wierzchołka, który pojawił się w kroku s, opisuje w chwili t następujące równanie:

$$\frac{\partial \overline{k}(s,t)}{\partial t} = ct^{\alpha} \frac{\overline{k}(s,t)}{\int_0^t \overline{k}(u,t)du}, \tag{4.21}$$

gdzie $\alpha > 0$. W odróżnieniu od poprzedniej wersji modelu, tempo wzrostu wierzchołków dodanych w chwili t wzrasta wykładniczo. Prowadzi to do wniosku, że dla wierzchołków dodanych wcześniej preferencyjne przyłączanie ma jeszcze bardziej uprzywilejowany charakter. Przyspieszony wzrost niesie ze sobą istotne konsekwencje w samej strukturze sieci, które zostaną zaprezentowane w dalszej części pracy.

Przyjmując, że $\overline{k}(0,0) = 0$ i $\overline{k}(t,t) = 1$, suma wszystkich stopni wierzchołków sieci w kroku t to:

$$\int_0^t \overline{k}(u,t)du = \frac{c}{\alpha+1} t^{\alpha+1}. \tag{4.22}$$

Dla przejrzystości dalszych rozważań przyjąć można oznaczenie:

$$\alpha + 1 = 1/\beta_{\mathrm{H}}. \tag{4.23}$$

Podstawiając całkę (4.22) do równania (4.21) i następnie rozwiązując je, otrzymuje się proste wyrażenie na średni stopień wierzchołka:

$$\overline{k}(s,t) = \left(\frac{t}{s}\right)^{\alpha+1}. \tag{4.24}$$

Zakładając ciągłość czasu, możemy wyznaczyć rozkład stopni wierzchołków za pomocą równania:

$$P(k,t) = \frac{1}{t}\int_0^t \delta(k - \overline{k}(s,t))ds = -\frac{1}{t}\left(\frac{\partial \overline{k}(s,t)}{\partial s}\right)^{-1}_{s=k(s,t)}, \tag{4.25}$$

gdzie pochodna cząstkowa:

$$\frac{\partial \bar{k}(s,t)}{\partial s} = -(\alpha+1)\frac{t}{s^2}\left(\frac{t}{s}\right)^{\alpha}. \tag{4.26}$$

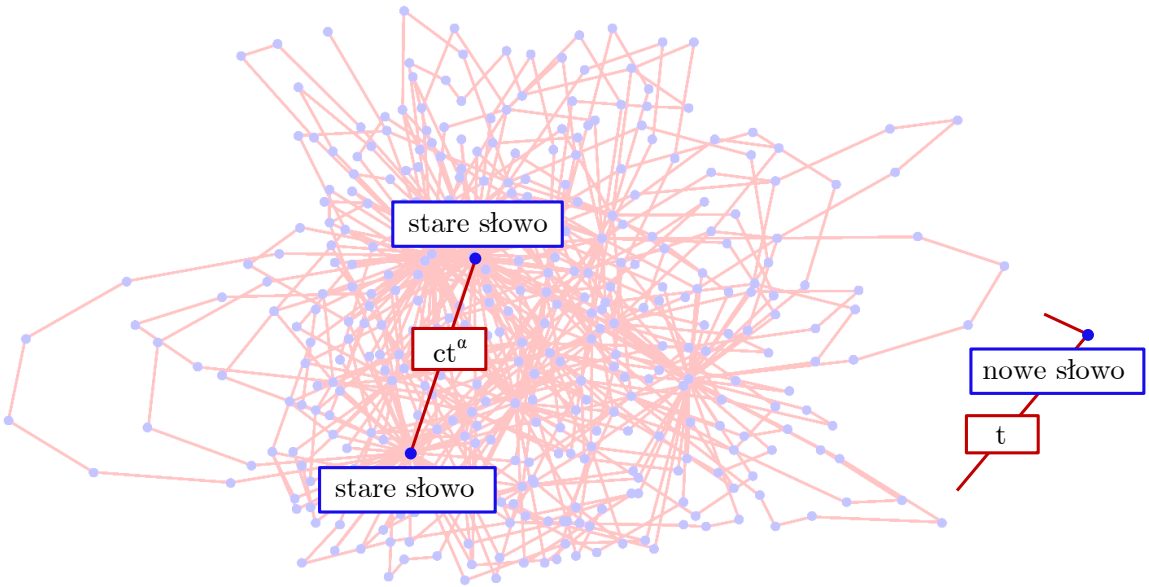
Podstawiając (4.26) do równania (4.25) przy uwzględnieniu równania (4.24), dostaje się rozwiązanie, będące stacjonarnym rozkładem krotności:

$$P(k) = \frac{1}{\alpha+1}k^{-(1+1/(\alpha+1))}. \tag{4.27}$$

Przyjmując oznaczenie (4.23), powyższy rozkład można zwięźle zapisać w postaci:

$$P(k) = \beta_{\mathrm{H}} k^{-(1+\beta_{\mathrm{H}})}. \tag{4.28}$$

Model z przyspieszonym wzrostem stopni wierzchołków jest, podobnie jak liniowy model DM, charakteryzowany jest przez potęgowy rozkład $P(k)$, jednak dodatkowo rozkład ten jest funkcją parametru β_{H}. Mimo to model jest nadal minimalny, w tym sensie, że nie wprowadza żadnych dodatkowych parametrów do równania (4.21), prócz stałej określającej tempo ekspansji (inaczej niż np. w przypadku *modelu fitness* [160]). Dla $\alpha = 1$ model DM-AG jest tożsamy z modelem liniowym DM, dając w wyniku ten sam rozkład $P(k) \propto \frac{1}{2}k^{-3/2}$. Zasadniczą różnicą pomiędzy nimi jest to, że w modelu DM-AG istnieje zależność funkcyjna na wartość wykładnika $\gamma(\beta_{\mathrm{H}})$, opisującego rozkład krotności stopni wierzchołków w sieci, natomiast w modelu DM wartości wykładników skalowania rozkładów $P(k)$ są stałe.



Rysunek 4.9: Model wzrostu sieci za pomocą dwóch współistniejących mechanizmów. Wraz z kolejnym krokiem t do sieci dodawany jest nowy wierzchołek (słowo). Równocześnie, z określonym tempem ct^{α} pojawiają się połączenia wewnątrz sieci.

4.2.2 Dynamika sieci lingwistycznej vs. model DM-AG

W kontekście powyższych rozważań należy uważnie zbadać dynamikę sieci lingwistycznych opartych o sąsiedztwo słów w tekście. Jak już wspomniano w podrozdziale 4.1.1, dla tekstów obowiązuje empiryczne prawo Heapsa, wyrażające potęgową zależność pomiędzy długością tekstu a liczbą zawartych w nim unikalnych słów. Jeśli przez l oznaczmy długość tekstu wyrażoną w liczbie wszystkich słów, natomiast przez $N(l)$ – liczbę unikalnych słów, to dla większości języków naturalnych [134, 140] istnieje związek w postaci $N(l) = gl^{\beta_{\mathrm{H}}}$, gdzie wykładnik β_{H} jest charakterystyczny dla konkretnego języka (w przypadku dużych korpusów tekstów) lub konkretnego twórcy, a nawet dzieła, natomiast g jest współczynnikiem proporcjonalności. Zmienność wykładnika na ogół zawiera się w przedziale $0.5 < \beta_{\mathrm{H}} < 0.9$, gdzie niska wartość świadczy o ubogim słownictwie, natomiast wysoka jest związana z różnorodnością (a w ekstremalnych sytuacjach – z nienaturalną nadreprezentacją rzadkich słów).

Dokonując inwersji prawa Heapsa (wzór 4.7), otrzymać można zależność:

$$l = g^{-1/\beta_{\mathrm{H}}} N(l)^{1/\beta_{\mathrm{H}}}, \tag{4.29}$$

gdzie równanie opisuje długość tekstu w funkcji liczby unikalnych słów. Rozpatrując powyższy przypadek w kontekście sieci sąsiedztwa słów, można zauważyć, że liczba unikalnych słów $N(l)$ to liczba wierzchołków w sieci. W analizowanym modelu sieci zakłada się, że w każdym nowym kroku t do sieci dodawany jest nowy wierzchołek, stąd można przyjąć, że $t = N(l)$, natomiast długość tekstu l jest równa liczbie krawędzi e, jakie ta sieć posiada[6]. W związku z tym, równanie (4.29) można przepisać do postaci:

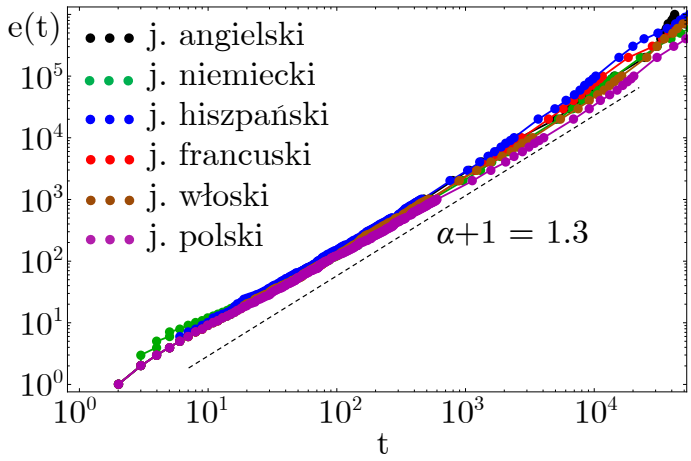
$$e(t) = g^{-1/\beta_{\mathrm{H}}} t^{1/\beta_{\mathrm{H}}}, \tag{4.30}$$

gdzie $e(t)$ to liczba krawędzi w chwili t. Na rysunku 4.10 przedstawiono zależność liczby krawędzi istniejących w sieci od czasu t, gdzie wykładnik $\alpha + 1 = 1/\beta_{\mathrm{H}}$ charakteryzuje tempo dodawania nowych wierzchołków. Jakościowy charakter wzrostu $e(t)$ jest zachowany dla wszystkich języków o nachyleniu większym od jedności. Empiryczne wartości wykładnika Heapsa β_{H} były wielokrotnie wyznaczane w szeregu pracach lingwistyki opisowej, a jego wartość dla wszystkich rozważanych języków została wyznaczona niezależnie [133, 141]. Analizując rysunek 4.10, dostrzec można, że lokalne nachylenie coraz silniej rośnie dla coraz dłuższych tekstów i dla $t > 10^4$ słów dalsza realizacja tekstów odbywa się już w znacznym stopniu w przestrzeni istniejących wyrazów.

Każda nowo dodana do sieci krawędź zwiększa całkowitą liczbę stopni wierzchołków o 2, co oznacza, że podwójna liczba wszystkich istniejących krawędzi w każdej chwili t jest równa sumie stopni wszystkich wierzchołków. Zatem, jeśli $2e(t) = \sum_j k_j(t)$, to korzystając z równania (4.30) – otrzymać można wyrażenie opisujące sumę wszystkich stopni w sieci:

$$\sum_j k_j(t) = 2g^{-1/\beta_{\mathrm{H}}} t^{1/\beta_{\mathrm{H}}}. \tag{4.31}$$

[6]Precyzyjnie, $e(t) = l(t) - 1$, jednak tę różnicę można zaniedbać dla $t \gg 1$.



Rysunek 4.10: Wzrost liczby krawędzi $e(t)$ w funkcji czasu t charakterystyczny dla modelu DM-AG. Rożnymi kolorami opisano zmianę wartości $e(t)$ dla poszczególnych języków.

Uśredniając rozkład krotności po wszystkich węzłach, średni stopień i-tego wierzchołka w chwili t można zapisać jako[7]:

$$\overline{k}(s,t) \propto \delta_{s,t} \sum_j k_j(t). \tag{4.32}$$

Wykorzystując równanie (4.31), można jakościowo zapisać:

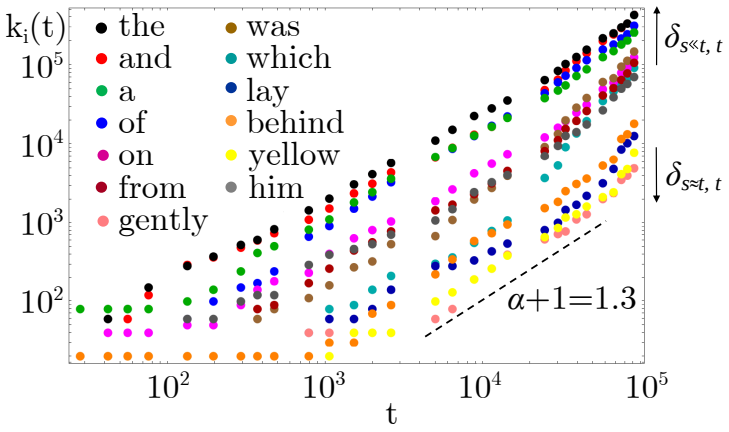
$$\overline{k}(s,t) \propto 2\delta_{s,t} g^{-1/\beta_{\mathrm{H}}} t^{1/\beta_{\mathrm{H}}}. \tag{4.33}$$

Wprowadzenie parametru $\delta_{s,t} = (t-s)/t$ odzwierciedla średni przyrost stopni wierzchołków dodanych w różnych chwilach czasu do sieci. Dla wierzchołków dodanych na początku $s \ll t$ oraz $\delta_{s,t} \to 1$, co prowadzi do maksymalizacji ich średniego stopnia $\overline{k}(s,t)$, natomiast dla wierzchołków dodanych później $s \approx t$ oraz $\delta_{s,t} \to 0$, stąd $\overline{k}(s,t) \approx 0$.

Na rysunku 4.11 przedstawiono wzrost stopni wybranych wierzchołków sieci sąsiedztwa słów, sporządzonej dla języka angielskiego. Słowa, które pojawiły się na początku ekspansji sieci, takie jak: *the, and, a*, stają się kandydatami na *huby*, a wraz z upływem czasu ich zdolność przyłączania do siebie innych wierzchołków rośnie ze ściśle określonym tempem $\alpha + 1 \approx 1.3$, gdzie dla dużych wartości t wzrost ten jest jeszcze szybszy. Słowa, które pojawiły się później, np. *behind, yellow, him*, posiadają znacznie mniejszą krotność, choć tempo przyrostu stopnia jest takie samo. Jest to bezpośrednią konsekwencją wzoru (4.33), w którym wprowadzony parametr $\delta_{s,t}$ powinien być rozumiany jako wektor przesunięcia rozkładu w wartościach krotności wierzchołków.

Tempo wzrostu liczby wszystkich krawędzi w rozważanej sieci może być łatwo wyznaczone poprzez zróżniczkowanie równania (4.31). Analogicznie można przedstawić tempo przyrostu sumy liczby stopni w sieci, różniczkując równanie (4.32), bądź,

[7]Równoważnie można zapisać, że $\overline{k}(s,t) \propto \sum_{k=0}^{t} kp(k,t)$, gdzie $p(k,t)$ jest rozkładem dwumianowym zajścia k sukcesów w t próbach.



Rysunek 4.11: Wzrost stopni wybranych wierzchołków w sieci lingwistycznej oparte o tekst w języku angielskim. Szybszy wzrost krotności w obszarze $t \sim 10^5$ związany jest z wysyceniem słownictwa.

wykorzystując równanie (4.33), można łatwo wyznaczyć tempo wzrostu średniego stopnia wierzchołka:

$$\frac{\partial \overline{k}(s,t)}{\partial t} \propto 2g^{-1/\beta_{\mathrm{H}}}\beta_{\mathrm{H}}^{-1}\delta'_{t,s}t^{1/\beta_{\mathrm{H}}-1}. \tag{4.34}$$

Wyrażenia te, uzyskane w drodze jakościowo-ilościowej analizy prawa Heapsa, wyrażające sumę stopni w sieci (równanie (4.31)) oraz tempo przyrostu średniego stopnia wierzchołka (równanie (4.34)), można uprościć, stosując podstawienie (4.23), do postaci:

$$\sum_j k_j(t) = \frac{2}{g^{\alpha+1}}t^{\alpha+1} \qquad \frac{\partial \overline{k}(s,t)}{\partial t} \propto \frac{2(\alpha+1)}{g^{\alpha+1}}\delta'_{t,s}t^{\alpha} \tag{4.35}$$

Konfrontując powyższe, empirycznie wyprowadzone wzory z równaniami opisującymi model DM-AG, widać, że są one tożsame. Przyrównując całkę (4.22), wyrażającą sumę stopni wierzchołków po czasie t, z sumą $\sum_j k_j(t)$ otrzymuje się:

$$\frac{c}{\alpha+1}t^{\alpha+1} = \frac{2}{g^{\alpha+1}}t^{\alpha+1}. \tag{4.36}$$

Równania te są jawnie tożsame, jeśli współczynnik c wynosi:

$$c = \frac{2(\alpha+1)}{g^{\alpha+1}}. \tag{4.37}$$

Wykorzystując uzyskaną postać współczynnika c, możemy ostatecznie uprościć zależności (4.35) do postaci:

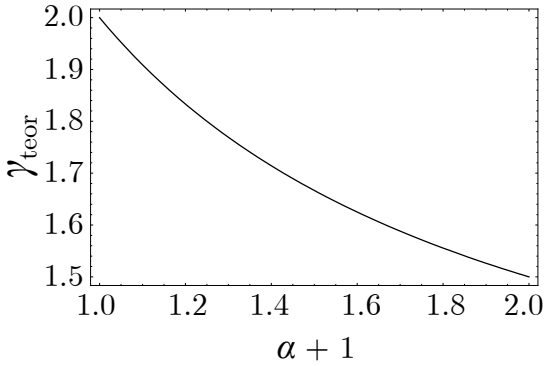
$$\sum_j k_j(t) = \frac{c}{\alpha+1}t^{\alpha+1} \qquad \frac{\partial \overline{k}(s,t)}{\partial t} \propto ct^{\alpha}\delta'_{t,s}, \tag{4.38}$$

gdzie parametr $\delta'_{t,s} \equiv k_i/\sum_j k_j$ należy rozumieć, w kontekście prowadzonej analizy, jako człon wyrażający preferencyjne przyłączanie.

Jak zostało to pokazane, obie wersje modelu DM, liniowa i potęgowa, są z różnym powodzeniem wykorzystywane do opisu sąsiedztwa słów w tekście w reprezentacji

sieciowej. Liniowy model DM, zakładający stałe tempo tworzenia się nowych połączeń, przedstawiony równaniem (4.17), nie uwzględnia szeroko tutaj opisanego prawa Heapsa. Ponadto przyjęcie liniowego tempa wzrostu sieci w postaci $m + 2mct$ prowadzi również do błędnego oszacowania sumy stopni wszystkich wierzchołków, którą to sumę można łatwo uzyskać poprzez wycałkowanie po czasie zdefiniowanego tempa wzrostu sieci. Jej wartość, $2mt + cmt^2$, jest na tyle szybko zmienna w czasie, iż nawet w małym zakresie t nie odzwierciedla rzeczywistej całkowitej krotności wierzchołków w sieci[8]. Co więcej, uzyskany w modelu liniowym niestacjonarny rozkład stopni wierzchołków $P(k)$ nie zależy, oprócz c, od jakichkolwiek innych parametrów, co w kontekście naturalnego różnicowania językowego jest nienaturalne. W świetle przytoczonych faktów liniowy model DM staje się niewystarczający.

Model DM-AG zakłada przyspieszony charakter ekspansji sieci. Analiza tempa dodawania nowych połączeń w sieci wygenerowanej według tego modelu, przeprowadzona z uwzględnieniem empirycznych przesłanek, dość dobrze oddaje charakter dynamiki sieci lingwistycznych. Zarówno tempo wzrostu sieci, jak i suma wszystkich stopni wierzchołków w określonej chwili t mają swoje ilościowe uzasadnienie w statystycznych własnościach języka. Otrzymany rozkład krotności $P(k)$ zależy od charakteru ekspansji sieci, co w kontekście prowadzonej tutaj ilościowej charakterystyki języków naturalnych, stanowi konkretną miarę struktury badanego języka.



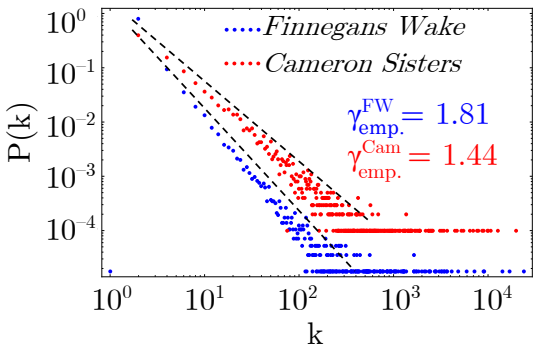
Rysunek 4.12: Zmiana wykładnika γ_{teor} w funkcji wykładnika $\alpha + 1$. Coraz większe tempo dodawania nowych wierzchołków do sieci skutkuje coraz mniejszym nachyleniem rozkładu krotności wierzchołków $P(k)$.

Na rysunku 4.12 przedstawiono zmianę wartości wykładnika γ_{teor}, przewidzianego przez model DM-AG, w zależności od określającego tempo ekspansji sieci wykładnika $\alpha + 1$. Coraz większa jego wartość charakteryzuje teksty o coraz uboższym słownictwie, co przekłada się w obrazie ewolucji sieci na coraz intensywniejsze tempo dodawania krawędzi w kolejnym kroku t. W konsekwencji musi to prowadzić do zmian w rozkładzie krotności stopni wierzchołków $P(k)$. Coraz większy stosunek liczby krawędzi do liczby wierzchołków prowadzi do sytuacji, w której wierzchołki o wyższym stopniu będą coraz bardziej liczne w stosunku do wierzchołków o stopniu

[8]Jedynie przyjęcie za c = 0 daje, w zakresie małych t, zgodność z danymi empirycznymi, jednak przyjęcie tego założenia prowadzi do zmiany modelu na zwykły model BA.

niższym. Zatem prawdopodobieństwo zaobserwowania węzłów o małej krotności jest coraz mniejsze dla coraz większych wartości $\alpha + 1$.

Jaskrawym przykładem, jakim można zilustrować tę sytuację, jest przedstawienie w reprezentacji sieciowej dwóch różnych książek: *Finnegans Wake* Jamesa Joyce'a oraz *Cameron Sisters* Cathy Maxwell. Bogate słownictwo pierwszej z nich i ubogie drugiej daje w konsekwencji inne zachowanie w kontekście prawa Heapsa; wykładniki potęgowe są różne: $\alpha^{\mathrm{FW}}+1 = 1.07$ i $\alpha^{\mathrm{Cam}}+1 = 1.68$. Wykładniki te determinują z kolei wartości wykładników skalowania rozkładów krotności $P(k)$, których teoretyczna wartość, w oparciu o model DM-AG to, odpowiednio $\gamma^{\mathrm{FW}}_{\mathrm{teor}} = 1.93$ i $\gamma^{\mathrm{Cam}}_{\mathrm{teor}} = 1.59$.



Rysunek 4.13: Rozkład krotności wierzchołków $P(k)$ dla sieci lingwistycznych opartych o dwie książki: *Finnegans Wake* J. Joyce'a oraz *Cameron Sisters* C. Maxwell. Zgodnie z proponowanym modelem DM-AG, większe tempo dodawania nowych słów, związane z parametrem $\alpha = \beta_H^{-1} - 1$, prowadzi do stromszych rozkładów krotności $P(k)$.

Na rysunku 4.13 przedstawiono empiryczne rozkłady krotności wierzchołków sporządzone na podstawie dwóch analizowanych książek. Zwraca uwagę różne nachylenie rozkładów, co zostało przewidziane w modelu z przyspieszonym wzrostem wraz z podaniem teoretycznych wartości wykładników skalowania, które z niewielkim błędem odpowiadają ich rzeczywistym wartościom. Rozkład ten znacznie lepiej oddaje zróżnicowanie słownikowe zawarte w prezentowych utworach literackich niż sam rozkład Zipfa, przedstawiony na rysunku 4.2.

W tablicy 4.1 zebrano wyznaczone wykładniki Heapsa dla kilku rozpatrywanych języków, wyznaczone dla korpusów tekstów o łącznej długości 10^6 słów. Na podstawie zaprezentowanego modelu sieci z przyspieszonym wzrostem podano wartości wykładników $\gamma_{\mathrm{teor}} = 1+1/(\alpha+1)$. Język angielski w tym zestawieniu charakteryzuje się największym wykładnikiem $\alpha+1$, co ma związek z brakiem rozbudowanej fleksji oraz charakterem wybranych tekstów literackich (dostępność literatury w ramach jednego stylu literackiego w tym języku jest zacznie obszerniejsza niż dla innych). Najmniejszym wykładnikiem $(\alpha + 1)$ charakteryzuje się język polski. Jest to odzwierciedlenie różnorodności słownikowej, w dużej mierze związanej z fleksją, oraz z koniecznością zespolenia ze sobą kilku stylów literackich (książek pochodzących z różnych epok).

Tablica 4.1: Wartości wykładników charakterystycznych dla poszczególnych języków, β_{H} – wykładnik Heapsa, c i α – empiryczne parametry określające tempo pojawiania się krawędzi w rzeczywistej sieci lingwistycznej, γ_{teor} – wykładnik skalowania $P(k)$ dla modelu DM-AG określonego przez parametry c i α, γ_{emp} – wykładnik skalowania $P(k)$ rzeczywistych sieci lingwistycznych.

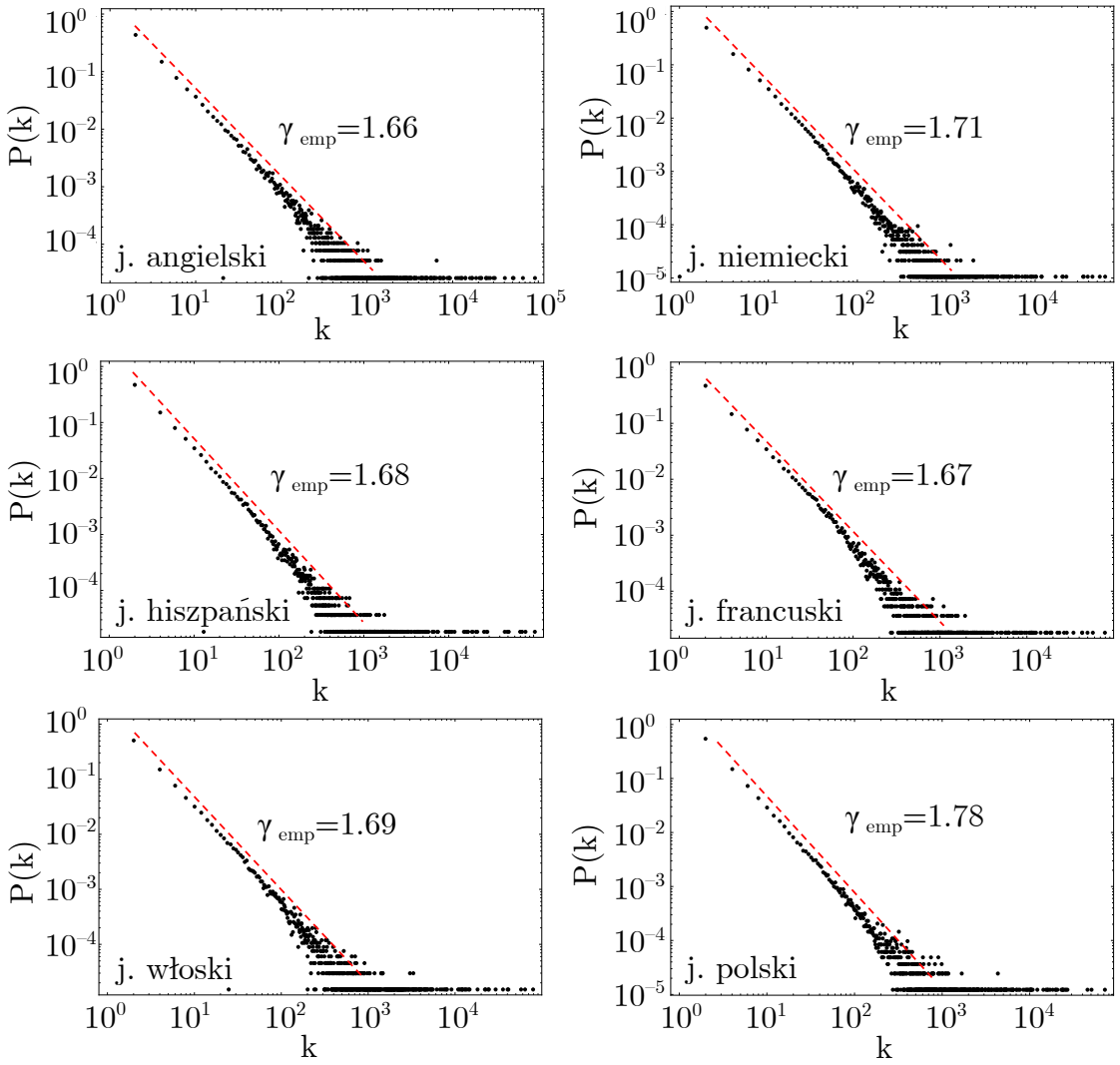
Język	β_{H}	c	$\alpha+1$	γ_{teor}	γ_{emp}
j. angielski	0.70	0.49	1.42	1.71	1.66
j. niemiecki	0.79	0.33	1.26	1.79	1.71
j. hiszpański	0.76	0.39	1.32	1.75	1.68
j. francuski	0.75	0.48	1.34	1.74	1.67
j. włoski	0.77	0.36	1.29	1.77	1.69
j. polski	0.80	0.32	1.24	1.81	1.78

4.2.3 Rozkłady krotności wierzchołków $P(k)$ dla sieci lingwistycznych

Na rysunku 4.14 zestawiono różniczkowe rozkłady krotności $P(k)$, sporządzone dla 6 języków europejskich. Dla każdego z nich rozkład $P(k)$ ma charakter potęgowy, różniący się nachyleniem w zależności od języka. Na podstawie sporządzonych rozkładów doskonale widać relację pomiędzy rozkładem Heapsa obowiązującym dla danego języka, a związanym z nim rozkładem $P(k)$. Znaczna dyspersja krotności wierzchołków o niskich wartościach $P(k)$ związana jest z charakterem sieci, która powstaje w wyniku uwzględnienia wszystkich par sąsiadujących słów, a więc także takich, które są częścią utartych zwrotów, idiomów i wielowyrazowych nazw własnych, co może prowadzić do arbitralnego przyrostu stopni. Różnica pomiędzy wartością wykładnika γ_{teor} a γ_{emp} jest natomiast konsekwencją ograniczonej stosowalności prawa Heapsa, związanego z wysyceniem słownikowym.

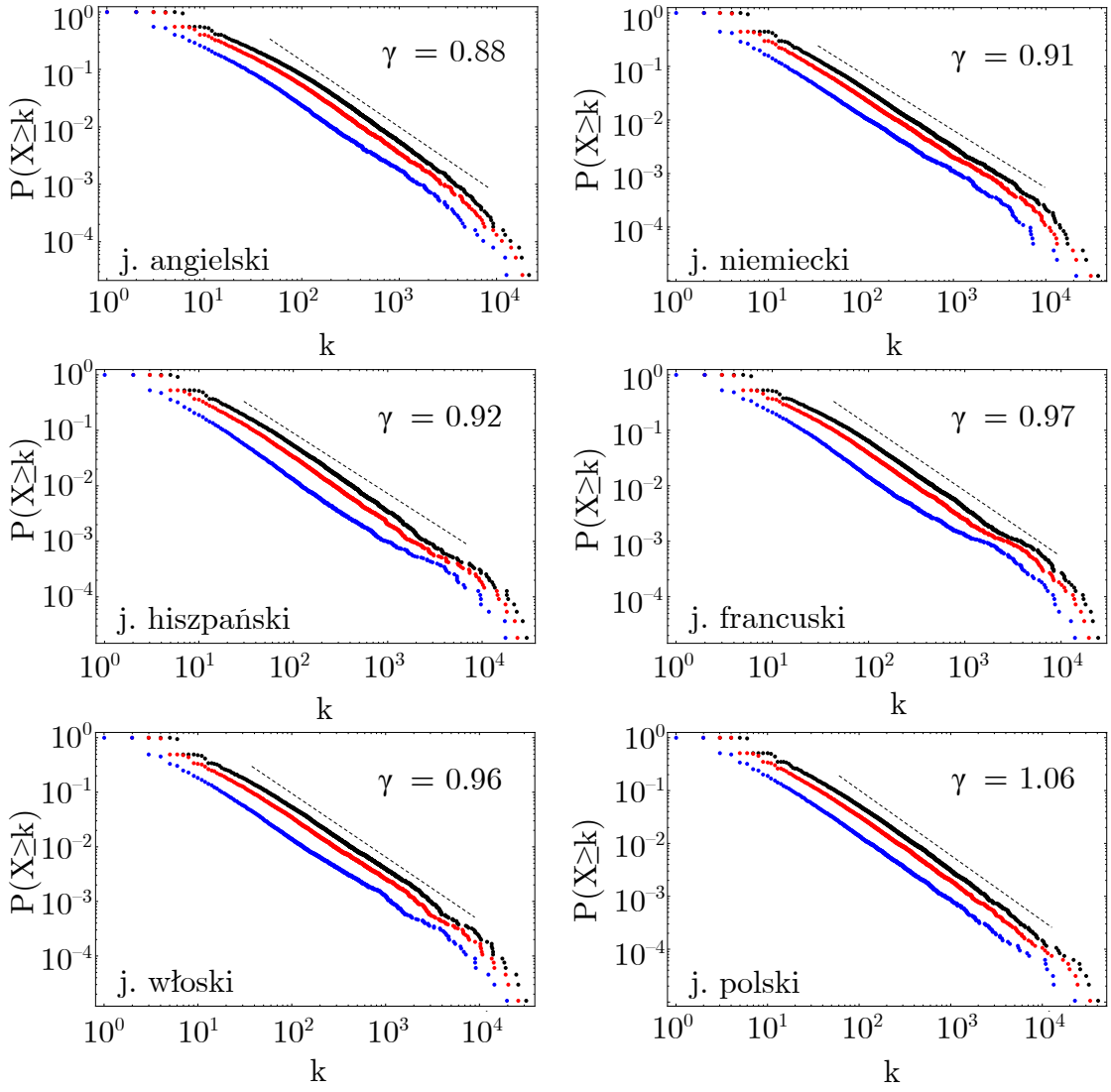
Skumulowane rozkłady krotności wierzchołków dla rozważanych języków przedstawiono na rysunku 4.15. Czarnym kolorem przedstawiono rozkłady dla sieci ważonych, w których stopień poszczególnych wierzchołków wzrastał, jeśli tylko odpowiadające mu słowo pojawiło się w tekście. W związku z tym, że otrzymana w ten sposób informacja jest tożsama z informacją uzyskaną z rozkładu Zipfa, warto stworzyć sieć opartą na binarnych (nieważonych) połączeniach pomiędzy wierzchołkami. Rozkłady krotności dla tego rodzaju sieci zostały przedstawione kolorem czerwonym. Dla obu typów sieci wyraźnie widoczne jest załamanie rozkładu w okolicach $k \approx 10^4$. Z wykresów trudno określić, czy skalowanie węzłów o dużej krotności (hubów) jest również obecne, ale jeśli tak, to ze znacznie większą wartością wykładnika γ_{hub} niż ma to miejsce dla węzłów o mniejszej krotności.

Innym ujęciem relacji pomiędzy ważonym a nieważonym charakterem sieci mogą być drzewa minimalnie napinające (MST, ang. *minimum spanning tree*). Podejście to zostanie zaprezentowane w dalszej części pracy. Wykresy na rysunku 4.15 to skumulowane rozkłady $P(X \geqslant k)$ dla analizowanych próbek języków. Różnymi kolorami przedstawiono rozkłady sporządzone dla sieci, których połączenia pomiędzy wierzchołkami były realizowane w nieco inny sposób.



Rysunek 4.14: Rozkład różniczkowy krotności wierzchołków dla sieci nieważonej. Czerwona, przerywana linia wyznacza nachylenie rozkładów $P(k)$, dla wszystkich analizowanych tej rozprawie języków, stwierdzono skalowanie krotności wierzchołków $P(k)$ przez co najmniej trzy dekady zmienności stopnia wierzchołka k.

Kolorem niebieskim przedstawiono rozkłady dla sieci, w których połączenie pomiędzy wierzchołkami istniało wówczas, gdy odpowiadające im słowa były bezpośrednimi sąsiadami w tekście. Kolorem czerwonym oznaczono analogiczne rozkłady dla sieci, w których połączenie pomiędzy wierzchołkami istniało, jeśli odpowiadające im słowa znajdowały się maksymalnie jedno, a kolorem czarnym – dwa słowa od siebie. Uwzględnienie par słów niebędących bezpośrednimi sąsiadami jest konsekwencją faktu, iż często logicznie związane ze sobą wyrazy są rozłączone poprzez spójnik lub inny wraz pełniący funkcję gramatyczną. Na rysunkach zwraca uwagę jednakowe skalowanie wszystkich rozkładów, a zwiększanie zasięgu oddziaływania pomiędzy wyrazami przynosi jedynie zwiększenie liczby połączeń. Z tego powodu uwzględnienie słów znajdujących się w większych odległościach (np. w obrębie całych zdań) nie wniosłoby żadnej nowej informacji do rozkładów $P(X \geqslant k)$.

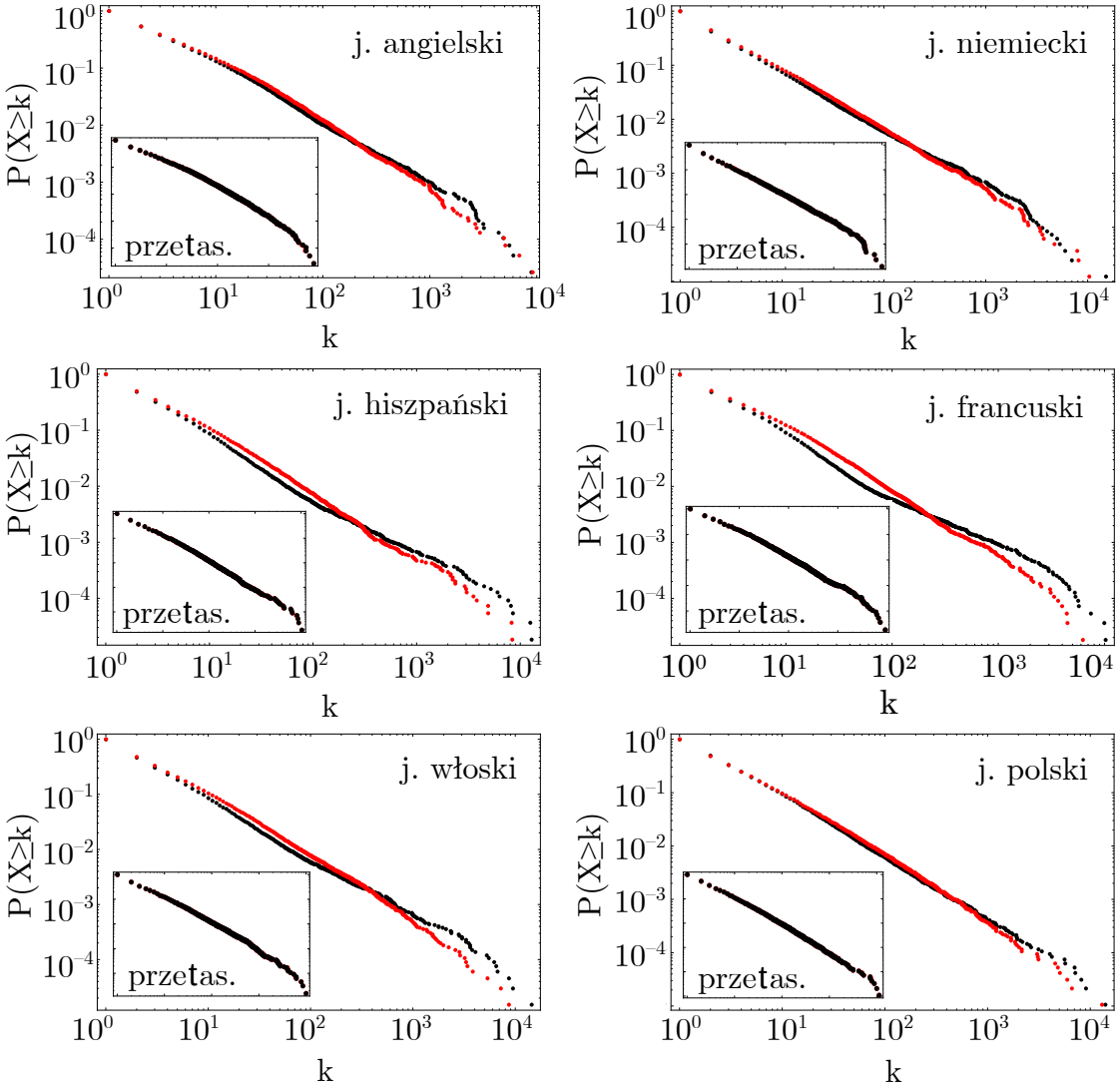


Rysunek 4.15: Rozkład skumowany dla sieci z różnym zasięgiem oddziaływań. Czarna, przerywana linia oznacza nachylenie rozkładu skumulowanego $P(X \geqslant k)$. Kolorem niebieskim oznaczono rozkłady dla sieci, gdzie wyraz jest połączony z innym gdy znajdują się obok siebie w choćby jednym zdaniu, kolorem czerwonym oznaczono rozkłady dla sieci opartych o sąsiedztwo czterech najbliższych słów znajdujących się względem danego słowa, natomiast kolorem czarnym oznaczono rozkłady oparte o sąsiedztwo determinujące połączenie zostało rozszerzone jeszcze o dwa słowa – znajdujące się kolejno trzy słowa przed i po względem danego słowa.

Zebrane na rysunku 4.16 rozkłady skumulowane $P(X \geqslant k)$ sporządzono, rozpatrując kierunkowy charakter sieci. Kolorem czerwonym zaprezentowano rozkłady stopni wejściowych, pochodzących od połączeń ze słowami użytymi wcześniej (z lewej strony danego słowa), natomiast czarnym – wyjściowych, pochodzących od połączeń ze słowami użytymi później (z prawej strony).

Dla wszystkich analizowanych języków, prócz j. hiszpańskiego i j. francuskiego, przebieg rozkładów jest bardzo podobny. Dla języka hiszpańskiego, a szczególnie francuskiego, rozkłady te mają nieco inne skalowania, co może być konsekwencją

kilku okoliczności. Język francuski jest językiem o dosyć skomplikowanej morfologii, gdzie porządek słów ma nie tylko charakter gramatyczny (jak np. w języku angielskim), ale również słowotwórczy, ponadto, w odróżnieniu od innych współczesnych języków posiadający zróżnicowanie ze względu na mnogość stosowanych czasów i trybów. Ponadto na uzyskany rozkład miał wpływ charakter rozważanych tekstów które mogły obfitować w nazwy własne, oraz zwroty które wpłynęły na prezentowane wyniki. W wyniku losowego wymieszania kolejności słów w tekstach zaobserwowano zniwelowanie różnic w rozkładach $P(X \geqslant k)$, co pokazują wstawki na rysunkach 4.16.



Rysunek 4.16: Rozkład skumowany dla sieci skierowanej w normalnej reprezentacji (wykresy główne), czerwonym kolorem oznaczono rozkłady sieci opartych o sąsiedztwo słów poprzedzających dane słowo, kolorem czarnym – następujących po danym słowie. We wstawkach pokazano rozkłady dla języka z przetasowaną kolejnością występujących słów.

4.2.4 Generatywne modele języka naturalnego

W tym podrozdziale omówione zostaną własności sieci, stworzone na bazie trzech modeli odzwierciedlających dynamikę tekstów:

– modelu sztucznych tekstów generowanych przez procesy Yule'a-Simona z uwzględnieniem prawa Heapsa,

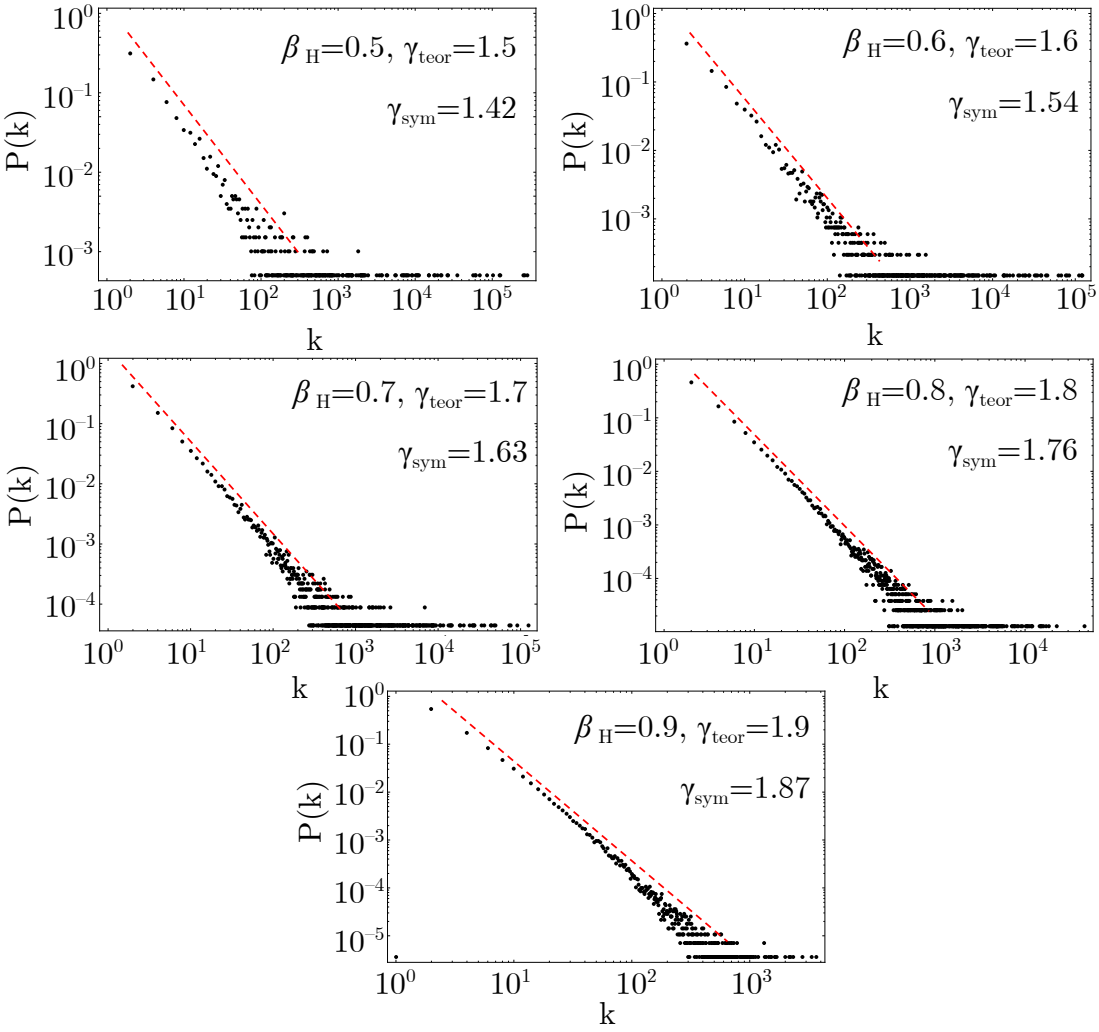
– potęgowy model sieci o przyspieszonym wzroście (model DM-AG),

– model błądzenia po sieci bezskalowej o dwóch wariantach:

  – bez pamięci,

  – z pamięcią.

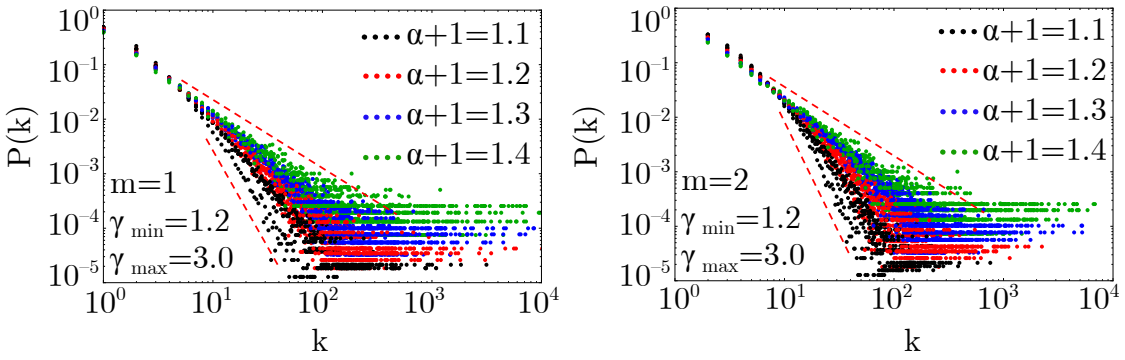
Modelowe procesy Yule'a-Simona uwzględniające prawo Heapsa YSH, równanie (4.15) zostały przedstawione w podrozdziale 4.1.1. Przy pomocy tych procesów można wygenerować szereg czasowy, a następnie przetransformować go w sieć, zgodnie z przyjętą metodą łączenia wierzchołków odpowiadających sąsiednim słowom.



Rysunek 4.17: Rozkład różniczkowy krotności wierzchołków sieci $P(k)$ stworzonych dla danych z procesów Yule'a-Simona z uwzględnieniem prawa Heapsa. Czerwoną, przerywaną linią oznaczono nachylenie rozkładów $P(k)$ dla sieci wygenerowanych wedle modelu.

Jest rzeczą oczywistą w tym przypadku, że rozkład $e(t)$ (a zatem i rozkład stopni wierzchołków) posiada postać potęgową z określonym parametrem β_{H}. Pod względem mechanizmu tworzenia, sieci te są więc pokrewne empirycznym sieciom sąsiedztwa słów. Na rysunku 4.17 przedstawione są rozkłady krotności $P(k)$ takich sieci, opisane wykładnikiem γ_{sym}. Na podstawie własności modelu YSH wiadomo, że powinna istnieć prosta relacja pomiędzy nachyleniem rozkładu $P(k)$, a wykładnikiem w prawie Heapsa: $\gamma_{\mathrm{teor}} = 1 + \beta_{\mathrm{H}}$. Wraz ze wzrostem β_{H} rośnie nachylenie rozkładu, co jest naturalną konsekwencją coraz bardziej zróżnicowanych wartości szeregu – w takim przypadku sieć składa się z coraz mniejszej liczby hubów. Warto odnotować, że przewidziana przez model YSH wartość wykładnika γ_{teor} różni się o ok. 5% od wartości uzyskanej w symulacjach; w każdym przypadku wartość przewidziana przez model jest nieznacznie większa. Wynikać to może z efektu skończoności próbki, w wyniku którego nachylenie uzyskanych rozkładów odbiega od teoretycznego, lub z błędu w oszacowaniu nachylenia (ze wskazaniem na tę drugą przyczynę). Pomimo tej niezgodności można przyjąć, że, z punktu widzenia otrzymanych rozkładów $P(k)$, wyniki są zgodne z przewidywaniami analitycznymi z podrozdziału 4.1.1.

Model DM-AG, zdefiniowany w podrozdziale 4.2.1 przez formułę (4.21), opiera się na preferencyjnym przyłączaniu wierzchołków, ale dodatkowo przyspieszony wzrost stopni wierzchołków wymaga dodatkowego mechanizmu, jakim jest dodawanie krawędzi pomiędzy już istniejącymi węzłami. Relacja pomiędzy tymi procesami jest opisana funkcyjnie za pomocą członu ct^{α}, co prowadzi w konsekwencji do uzyskania potęgowej relacji pomiędzy liczbą krawędzi a liczba wierzchołków. Do wygenerowania sieci przyjęto takie wartości parametrów c i α, które są obserwowane w ewolucji empirycznych sieci lingwistycznych: $c \in \{0.1, 0.2, 0.3, 0.4\}$ oraz $(\alpha + 1) \in \{1.1, 1.2, 1.3, 1.4\}$. Znając relację $\gamma_{\mathrm{teor}} = 1 + 1/(\alpha + 1)$, można skonfrontować przewidywania z wynikami symulacji. Przedział zmienności wykładnika γ_{teor} dla przyjętych wartości α zawiera się w granicach $1.7 < \gamma_{\mathrm{teor}} < 1.9$. Zgodnie z wynikami analitycznymi, otrzymane rozkłady nie powinny zależeć od stałej c, natomiast ich nachylenie jest zdeterminowane wartością parametru α.



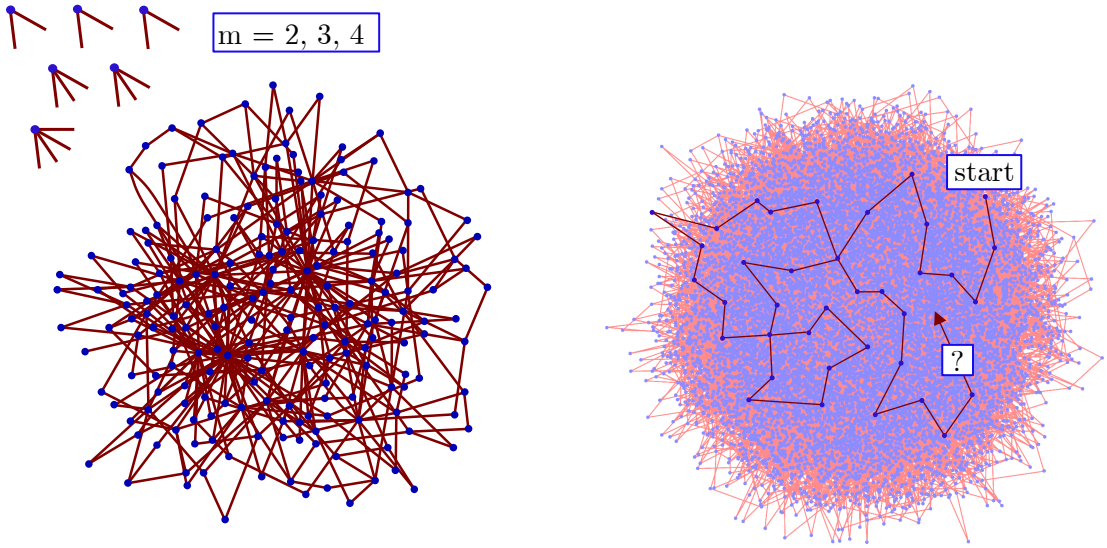
Rysunek 4.18: Rozkłady krotności wierzchołków $P(k)$ dla sieci wygenerowanych wedle modelu DM-AG.

W poszczególnych oknach rysunku 4.18 zestawiono rozkłady $P(k)$ uzyskane w oparciu o wygenerowane sieci z określonym parametrem m. Każde okno prezentuje po 16 różnych rozkładów, będących wynikiem przyjętych kombinacji parametrów c i α. Przyjęty parametr m, będący początkowym stopniem każdego nowo dodanego wierz-

chołka, nie wpływa istotnie na zachowanie się rozkładów $P(k)$, podobnie jak w innych modelach uwzględniających preferencyjne przyłączanie. Wyniki te świadczą o zgodności wyników symulacji modelu DM-AG z przewidywaniami jego teoretycznej konstrukcji.

Procesy YSH oraz model DM-AG opierają się na jawnie zastosowanych mechanizmach preferencyjnego przyłączania. Jednak możliwe jest także tworzenie sieci o podobnych własnościach, gdy preferencyjne przyłączanie jest wykorzystywane jedynie pośrednio. Ma to miejsce, gdy rozpatruje się model błądzenia po sieci. Niech zadana będzie sieć rosnąca wedle modelu BA o ustalonym wcześniej parametrze $2 \leqslant m \leqslant 4$. Będzie to sieć o krawędziach nieskierowanych i o rozkładzie krotności $P(k) \sim k^{-\gamma}$, gdzie $\gamma = 3$. Po wygenerowaniu odpowiednio dużej statystyki $V = 10^6$, losowo wybierając jakiś wierzchołek jako początkowy, można przeprowadzić błądzenie w przestrzeni istniejących wierzchołków sieci.

Rysunek 4.19 przedstawia ideę prowadzonej symulacji, gdzie w pierwszym kroku zostaje wygenerowana pierwotna sieć bezskalowa z określonym początkowym stopniem dodawanego wierzchołka $k_0 = m$, natomiast w drugim kroku zostaje wygenerowana sieć wtórna, będąca rezultatem przekształcenia sekwencji wierzchołków, otrzymanych w procesie błądzenia, na sieć ich wzajemnego sąsiedztwa (rysunek 4.20).

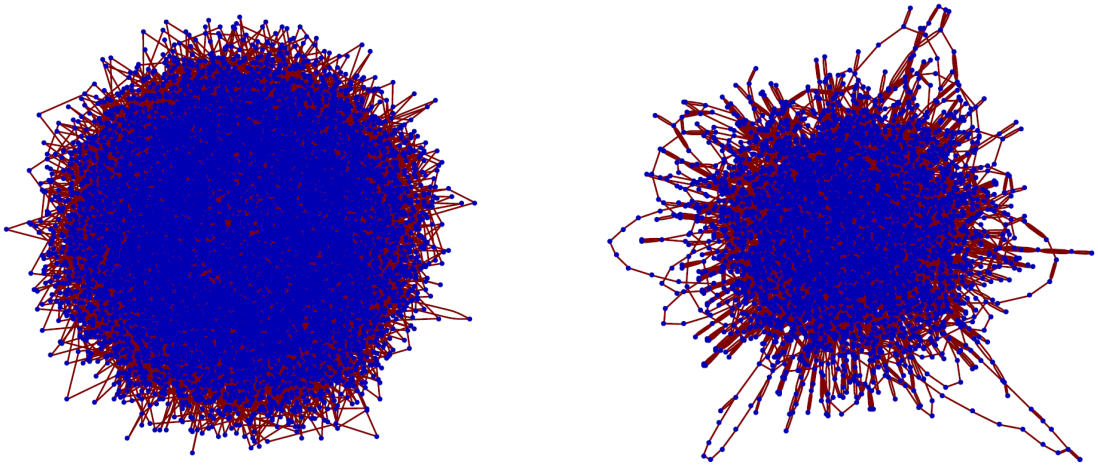


Rysunek 4.19: Schemat błądzenia losowego po sieci bezskalowej.

Parametr $m = 2$ wydaje się naturalny, gdyż każdy nowo dodawany wierzchołek (słowo) posiada w momencie inicjacji stopień $k_0 = 2$. Dla $m = 1$ sieć przyjmuje strukturę drzewiastą – jakościowo odmienną od obserwowanych w sieciach lingwistycznych, po której błądzenie będzie się charakteryzowało korelacjami krótkozasięgowymi, co w rzeczywistych warunkach nie jest obserwowane[9]. Wprowadzenie wyższych wartości parametru m prowadzi z kolei do wysokich stopni $k_i \gg 1$ wszystkich wierzchołków, co również nie jest obserwowane w rzeczywistości. Błądzenie po sieci będzie realizacją *łańcucha Markova*, gdzie prawdopodobieństwo wybrania i-tego

[9]Korelacja ta może wystąpić, gdy zostanie obrana pewna gałąź w sieci i po dotarciu do jej końca, nastąpi powrót po wcześniej odwiedzonych węzłach.

wierzchołka jest określone przez rozkład $P(k)$. Znormalizowana macierz sąsiedztwa (ang. *adjacency matrix*), $\sum_j a_{ij} = 1$, staje się wtedy macierzą przejścia. Prawdopodobieństwo przejścia od słowa i do słowa j opisuje prawdopodobieństwo warunkowe $p_{i,j} = P(X_{n+1} = j | X_n = i)$, gdzie uzyskanie odpowiedniej sekwencji wierzchołków (słów) jest równe iloczynowi prawdopodobieństw.

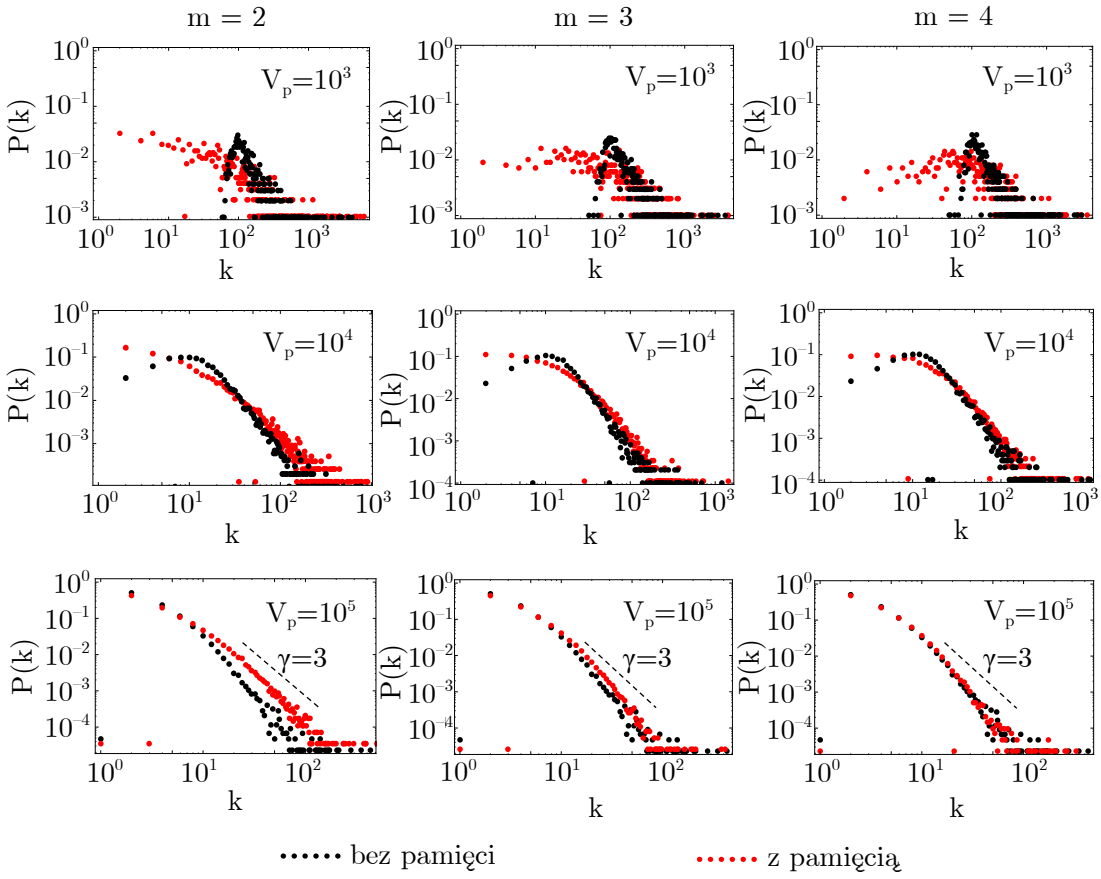


Rysunek 4.20: Pierwotna sieć bezskalowa (z lewej), po której odbywa się błądzenie, i sieć wtórna (z prawej), otrzymana z przekształcenia sekwencji odwiedzonych wierzchołków w sieć sąsiedztwa.

Rozpatrzone zostaną dwa przypadki realizacji swobodnego błądzenia w sieci: w pierwszym przypadku błądzenie odbywa się w przestrzeni istniejących połączeń i nie powoduje zmian w strukturze sieci, w drugim błądzenie jest realizowane w podobny sposób, ale po każdorazowym przejściu z wierzchołka v_i do v_j dodawana jest kolejna krawędź między nimi. W ten sposób tworzy się sieć ważona, a sam proces błądzenia zależy już nie tylko od warunków początkowych, ale również od historii procesu. Wybór wierzchołka, który zostanie odwiedzony w następnym kroku, zależy od jego krotności w sieci pierwotnej i od statystyki jego wystąpienia w utworzonej już sekwencji. Z lingwistycznego punktu widzenia jest to mechanizm bardzo ciekawy, bo stopień wierzchołka w sieci pierwotnej odzwierciedla początkową atrakcyjność słów jeszcze nie dodanych do tekstu. W obu przypadkach wzrost sieci wtórnej ma przyspieszony charakter: błądzenie będzie na ogół realizowane przez krawędzie rozpięte pomiędzy wierzchołkami już odwiedzonymi. Dokładnie taka sytuacja ma miejsce w sieciach lingwistycznych. Efekt ten będzie spotęgowany w drugim przypadku, gdzie dodatkowym czynnikiem jakim jest efekt pamięci błądzenia.

W przeprowadzonych symulacjach najpierw wygenerowano sieci wedle modelu BA z różnymi wartościami parametru $m = 2, 3, 4$, dla których ustalono rozmiar rzędu $V = 10^3, 10^4$ oraz 10^5 węzłów. Następnie na tak zadanych sieciach przeprowadzono proces błądzenia, ustalony dla wszystkich realizacji na $t = 10^5$ kroków. Otrzymana sekwencja wierzchołków zamieniana jest następnie na sieć sąsiedztwa. Odpowiednie rozkłady krotności sporządzone dla tych sieci przedstawiono na rysunku 4.21. Przy niskim rozmiarze sieci pierwotnej, wygenerowana sieć wtórna nie ma bezskalowego charakteru rozkładu $P(k)$, szczególnie przy realizacji błądzenia bez pamięci. Wraz ze wzrostem rozmiaru sieci pierwotnej rozkłady te stają się grubo-

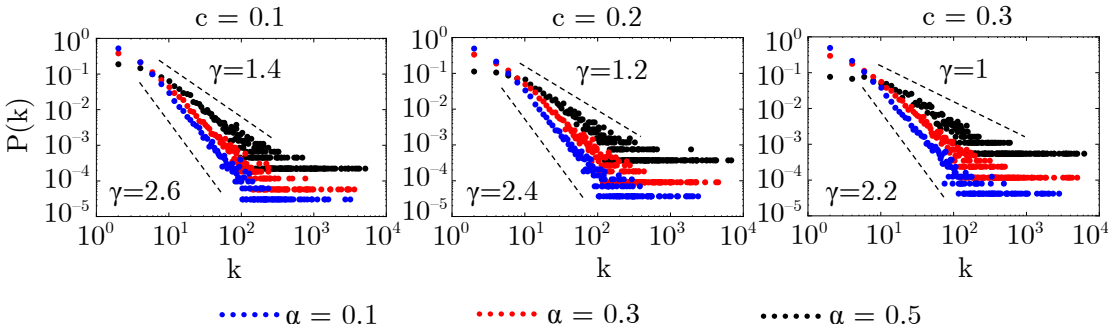
ogonowe, aby ostatecznie stać się bezskalowymi po osiągnięciu przez sieć rozmiaru ok. 10^5 wierzchołków. W odróżnieniu od sieci lingwistycznych rozkłady te cechuje znaczne nachylenie, bliskie rozkładom $P(k)$ dla sieci pierwotnych. Wyniki te świadczą, że zdefiniowany powyżej proces błądzenia po sieci bezskalowej nie jest właściwy do modelowania sieci lingwistycznych, mimo pozornej jego atrakcyjności. Ponadto przy coraz większym rozmiarze sieci pierwotnej, różnice pomiędzy błądzeniem z pamięcią i bez zacierają się w w świetle rozkładów krotności węzłów.



Rysunek 4.21: Rozkład krotności wierzchołków dla modelu błądzenia po sieci bezskalowej z pamięcią (kolor czerwony) i bez pamięci (kolor czarny) dla róznej liczby wierzchołków V i różnych wartości początkowej krotności wierzchołków m.

Proces błądzenia po sieci może zostać rozwinięty na przypadek, gdy błądzenie odbywa się równolegle w czasie względem wzrostu samej sieci pierwotnej. W tym modelu w chwili początkowej sieć pierwotna składa się z pewnej liczby wierzchołków połączonych w pierścień, a w chwilach późniejszych kolejne węzły są dodawane według reguł opisanych w modelu DM-AG. Ma miejsce więc sytuacja, w której trajektoria ruchu penetruje rozszerzającą się przestrzeń dostępnych stanów. Celem zapewnienia, że trajektoria nie będzie zawierała krótkich, powtarzalnych sekwencji wierzchołków (np. $...a \rightarrow b \rightarrow c \rightarrow a...$), które są raczej niespotykane w przypadku rzeczywistych sieci lingwistycznych, wprowadzony został dodatkowy warunek, że trajektoria nie może ponownie odwiedzić wierzchołków, które były wybrane w s krokach wstecz.

Warunek ten wiąże się z koniecznością wprowadzenia jeszcze jednego warunku, aby nowo przyłączany wierzchołek nie mógł wytworzyć krawędzi do wierzchołków, będących jego sąsiadami do rzędu $s-1$ (przy czym sąsiedzi pierwszego rzędu to swoi najbliżsi sąsiedzi). W przeciwnym wypadku trajektoria ruchu mogłaby zostać uwięziona w jakimś wierzchołku, nie mogąc go opuścić. Uzyskane w symulacjach tego modelu rozkłady krotności wierzchołków sieci wtórnej przedstawia rysunek 4.22. Nie różnią się one jakościowo od tych z rysunku 4.21, uzyskanych dla przypadku ze statyczną siecią pierwotną.



Rysunek 4.22: Rozkład krotności $P(k)$ dla błądzenia po sieci o przyspieszonym wzroście (model DM-AG). Wykresy sporządzono dla różnych wartości parametrów przyspieszonego wzrostu c i α. W każdym oknie podane zostały wartości wykładników skalowania γ dla dwóch skrajnych przypadków. Preferencyjne przyłączanie nowych wierzchołków realizowano dla $m = 2$, a rozmiar sieci rósł do wartości $V = 10^5$.

4.2.5 Ilościowe charakterystyki sieci

4.2.5.1 Drzewa MST

Sieciom powstałym na podstawie sąsiedztwa słów można nadać ważoną postać, rozpatrując wielokrotny charakter połączeń pomiędzy jej wierzchołkami i nadając tym połączeniom wagę proporcjonalną do krotności połączenia. W celu wizualizacji podstawowych cech takiej sieci bez konieczności zmagania się z nadmiarem informacji, można przedstawić ją w postaci spójnego dendrogramu, zwanego minimalnym drzewem rozpinającym (ang. *minimal spanning tree*, MST) [111]. Formalnie, minimalne drzewo rozpinające jest zdefiniowane w następujący sposób. Niech zadany będzie pewien graf G:

$$G := (V, E, w)\,, \tag{4.39}$$

gdzie: V to zbiór wierzchołków grafu, $E \subseteq \{\{u, v\} : u, v \in V\}$ – zbiór krawędzi, natomiast w_e oznacza wagę krawędzi $e \in E$. Minimalnym drzewem rozpinającym jest graf T, taki że:

$$T := (V, D)\,, \tag{4.40}$$

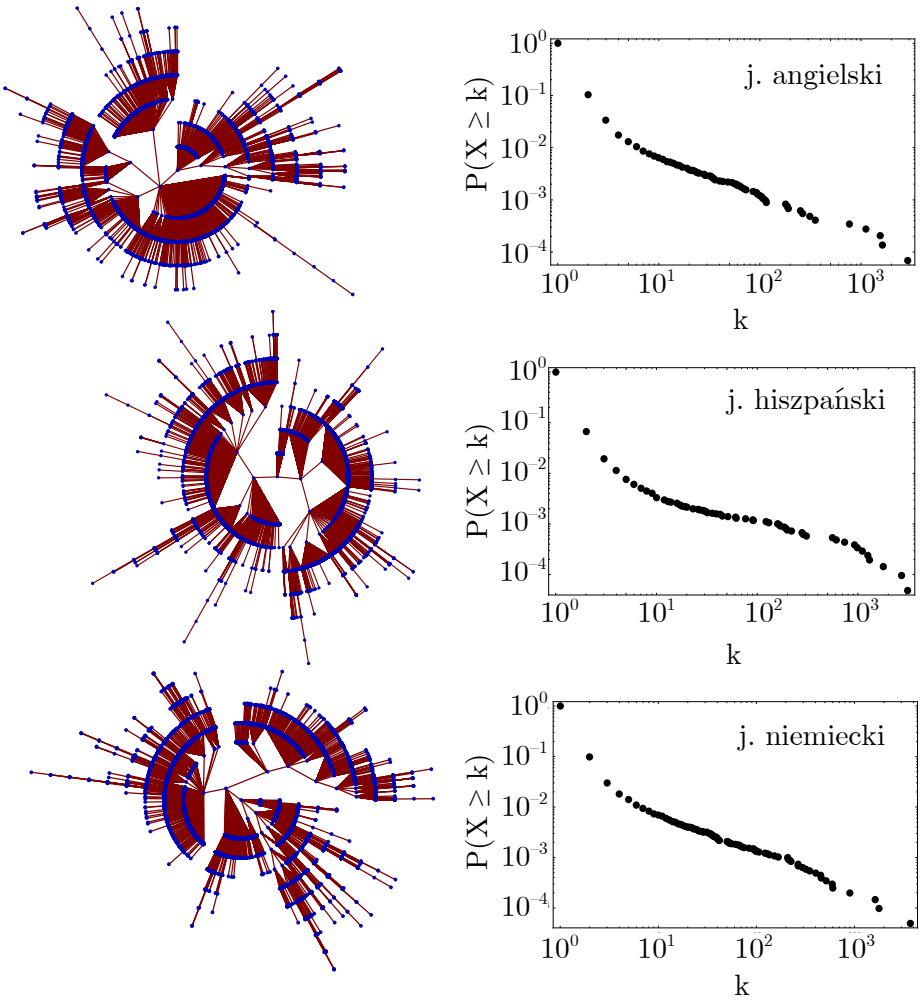
w którym $D \subseteq E$, przy założeniu, że suma wag krawędzi $\sum_{e\in T} w_e$ jest najmniejsza z możliwych. Dendrogram T jest grafem spójnym, acyklicznym oraz dla dowolnych wierzchołków $u, v \in V$ istnieje tylko jedna ścieżka między nimi. Relacje pomiędzy

MST a pełną siecią są nadal przedmiotem badań, jednak wiadomo na pewno, że niektóre własności oryginalnej sieci są zachowywane wraz z opisującymi je miarami (np. średnia długość najkrótszej ścieżki, pośrednictwo), natomiast niektóre nie (np. bezskalowość sieci nie implikuje podobnego rozkładu $P(k)$ dla dendrogramu MST, choć na ogół otrzymane drzewo również ma hierarchiczną strukturę) [69].

Najistotniejszą informację o strukturze sieci zawierają najsilniejsze połączenia między wierzchołkami, a więc krawędzie o największych wagach. By móc wyrazić analizowaną sieć w postaci MST, konieczne jest więc przetransformowanie istniejących wag w_e, w których siła połączenia jest proporcjonalna do wartości wagi, na wagi, w których ta zależność byłaby odwrotna. Z tego powodu wprowadza się miarę odległości między wierzchołkami, zdefiniowaną jako:

$$d_e = \sqrt{2\,(1 - w_e)}, \tag{4.41}$$

która spełnia dodatkowo warunki metryki.



Rysunek 4.23: Empirycznie uzyskane minimalne drzewo rozpinające dla j. angielskiego, hiszpańskiego i niemieckiego. Pomimo hierarchicznego charakteru dendrogramów po lewej stronie, skalowanie stopni wierzchołków $P(X \geqslant k)$ nie jest zachowane, szczególnie dla małych wartości stopni.

Przedstawienie sieci lingwistycznej jako dendrogramu MST wymaga przyjęcia kilku modyfikacji. Po pierwsze, aby ograniczyć rozmiar problemu, sieć na podstawie której zostanie skonstruowane drzewo, zostaje obcięta o wierzchołki, których stopień jest mniejszy lub równy 2. Z przyjętej definicji sieci sąsiedztwa wynika, że stopień taki cechuje wierzchołki odpowiadające słowom pojawiającym się incydentalnie, można więc je traktować jako lokalną fluktuację, nie wpływającą na własności sieci. Na rysunku 4.23 przedstawiono otrzymane drzewa dla trzech wybranych języków (struktura drzew dla pozostałych języków jest bardzo podobna). Dendrogramy posiadają hierarchiczną strukturę, w której rolę centralnych węzłów spełniają słowa będące również hubami w sieciach binarnych. Uzyskane rozkłady $P(k)$ świadczą jednak o braku skalowania we wszystkich przedziałach wartości krotności, a ich lokalne nachylenie jest typowo mniejsze niż $1/k$. Struktura drzew ukazuje również niewielką średnicę otrzymanej sieci (w stosunku do liczby wszystkich wierzchołków), gdyż największa odległość dla każdego języka zawiera się do 20 połączeń między wierzchołkami. Właściwość ta nosi nazwę efektu „małego świata" i zostanie szerzej omówiona w podrozdziale 4.2.5.3.

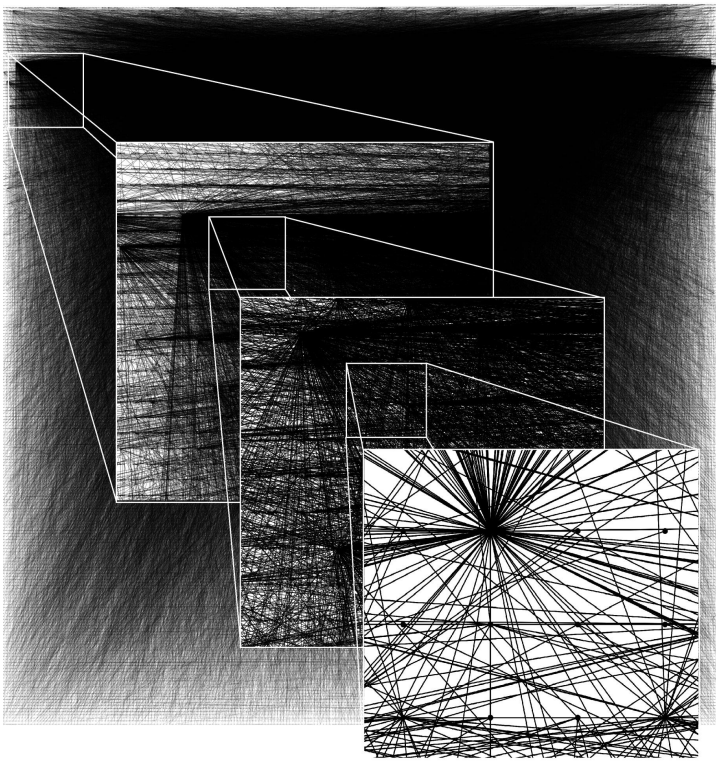
4.2.5.2 Gronowanie i pośrednictwo

Sieci oparte na sąsiedztwie słów nie tylko odzwierciedlają wzajemne interakcje pomiędzy nimi, ale również niosą informację odnośnie podstruktur w ramach współtworzonej sieci. Nie ulega wątpliwości, że pewne słowa występują często w otoczeniu innych specyficznych słów, jak: zwroty, struktury gramatyczne, opisy ściśle zdefiniowanych obiektów/problemów, motywy czy styl pisarstwa. Konsekwencją takich zjawisk jest pewna modularność języka, wyrażająca się w hierarchiczności struktur sieciowych. Rysunek 4.24 przedstawia schemat hierarchicznej budowy sieci, gdzie na każdym z poziomów jej struktury można dostrzec węzły dominujące i węzły peryferyjne, słabo połączone z innymi. W sieciach o topologii hierarchicznej stosunkowo niewielkie grupy (grona) silnie połączonych ze sobą węzłów tworzą coraz większe struktury o odpowiednio niższej liczbie połączeń. Powyższy opis przypomina jakościowo ideę fraktalności, czyli samopodobieństwa struktur na różnych poziomach organizacji.

Miarą określającą stopień modularności sieci jest współczynnik gronowania (ang. *clustering coefficient*), który wyraża ilościowo wzajemność lokalnych powiązań wierzchołków w sieci. Może on być zdefiniowany lokalnie dla pojedynczego węzła, ale też wyrażony jako miara globalna, po uśrednieniu po wszystkich wierzchołkach (wzór (3.7)). Wyrażona w ten sposób struktura sieci hierarchicznej prowadzi do potęgowej relacji skalowania [96]:

$$C_i(k) \sim k^{-(0.75 \div 1)}. \tag{4.42}$$

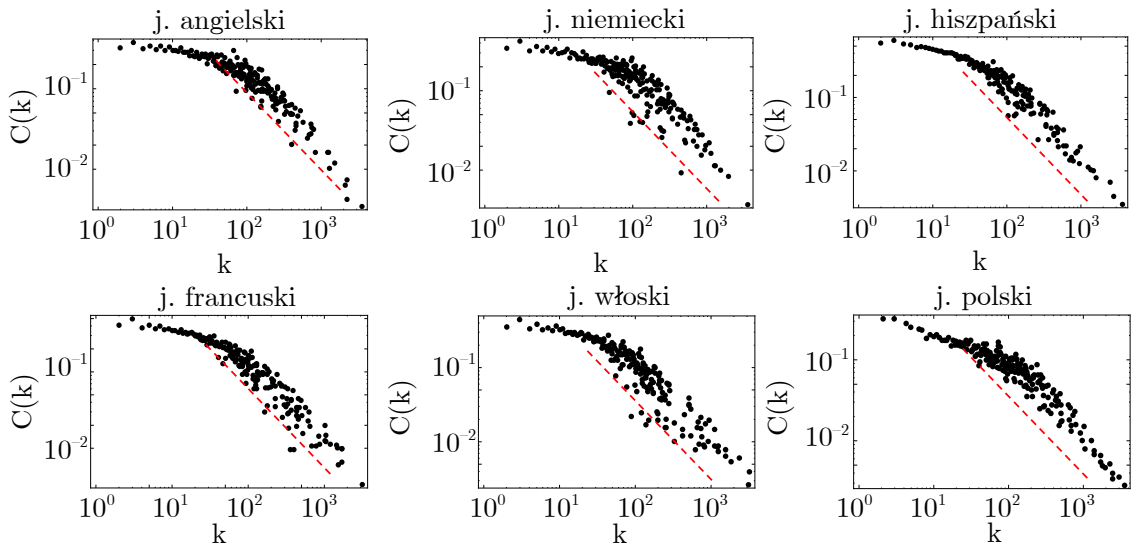
Okazuje się, że powyższa relacja nie jest uniwersalna dla wszystkich sieci o bezskalowych rozkładach stopni wierzchołków. Bez trudu można dostrzec, że istnieje pewna klasa sieci (dendrogramy), w których nie istnieją zamknięte cykle, a przez to nie istnieje w nich jakikolwiek przyczynek do współczynnika gronowania. Na ogół są to sieci planarne, czyli sieci, które mogą być zanurzone w płaszczyźnie (tzn. nie będą wtedy posiadać żadnych przecięć krawędzi poza węzłami), takie jak np. Internet na poziomie routerów czy sieci elektryczne.



Rysunek 4.24: Hierarchiczna struktura języka. Stopniowo powiększając fragmenty sieci, uzyskuje się obraz, który pod względem liczby wierzchołków o dużych stopniach i słabo połączonych jest samopodobny.

Sieci oparte na sąsiedztwie słów nie posiadają powyższego ograniczenia, istnieje w nich nawet nadreprezentacja krawędzi w stosunku do rozmiarów sieci, wynikająca z własności języka. Słowa są w tych sieciach w interakcji, choćby tylko incydentalnie pojawiły się razem, a zatem koszt nawiązywania połączenia pomiędzy nimi nie jest wysoki. Na rysunku 4.25 zestawiono rozkłady współczynników gronownaia $C_i(k)$ dla badanych języków. W każdym z przypadków zaobserwowano skalowanie uniwersalne dla rozważanej klasy sieci. Otrzymane rozkłady świadczą o hierarchicznej budowie sieci w zakresie wierzchołków, których krotność jest większa od 100. Przerywaną linią koloru czerwonego wyznaczono nachylenie zgodne ze wzorem (4.42). Efekt braku skalowania dla słabo połączonych (małe k) węzłów w sieci jest obserwowany dla większości sieci rzeczywistych, których rozkłady również nie posiadają skalowania w pełnym zakresie [96].

Rola węzłów w sieciach złożonych nie jest jednakowa, a ich znaczenie można rozpatrywać w rozmaity sposób. Jeśli usunięcie danego węzła spowoduje jakościowe zmiany w samej strukturze sieci bądź parametrów ją opisujących, można przypuszczać, że odgrywał on istotną rolę, w przeciwieństwie do innych, których usunięcie nie wiąże się ze znacznymi zmianami. Istnieje szereg parametrów, które określają istotność danego węzła w sieci, a za najbardziej pierwotny uznaje się jego stopień. Zachodzą jednak szczególne przypadki, w których sama krotność wierzchołka nie daje pełnej informacji o jego istotności [110]. W takim wypadku należy wykorzystać dodatkowe miary, m.in pośrednictwo [108]. Dla pojedynczego węzła wyraża ono stosunek liczby najkrótszych ścieżek przez niego przechodzących do wszystkich



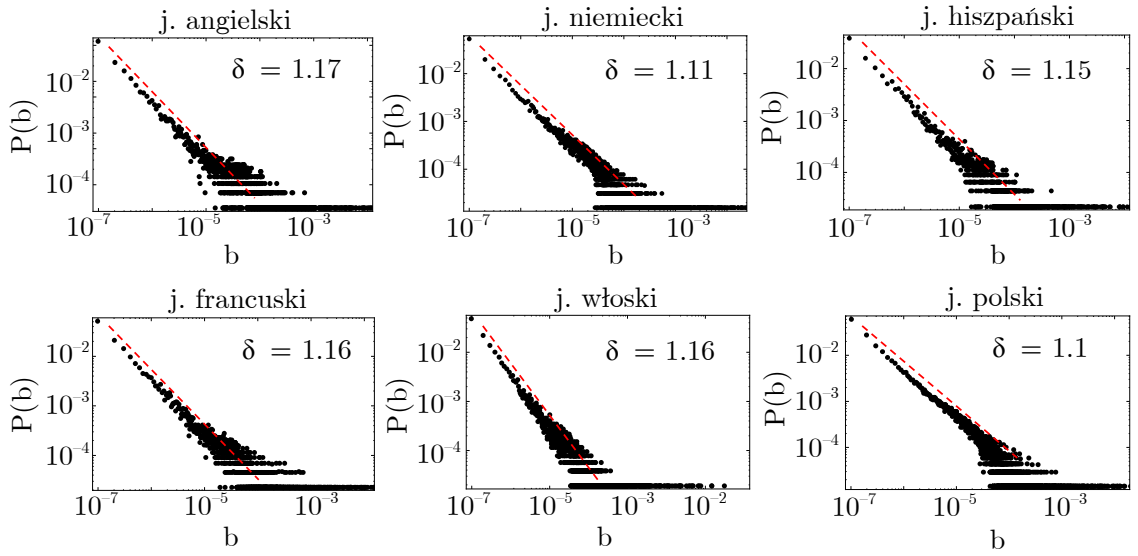
Rysunek 4.25: Rozkład współczynnika klasteryzacji dla sieci lingwistycznych. Czerwona, przerywana linia wyznacza nachylenie rozkładu $C(k)$ o wartości -1. Dla każdego z analizowanych języków, istnieje skalowanie przewidziane przez teorię, szczególnie dla dużych wartości stopni wierzchołków k.

najkrótszych ścieżek pomiędzy parami węzłów w sieci. Na ogół w sieciach o potęgowym rozkładzie krotności rolę tę pełnią huby, stąd też istnieje silne skorelowanie ich stopnia z pośrednictwem. Dlatego też dla sieci o bezskalowym rozkładzie stopni obserwuje się podobnie bezskalowy rozkład pośrednictwa, gdzie rozkład prawdopodobieństwa wystąpienia węzła o danym pośrednictwie b wynosi:

$$P(b) \sim b^{-\delta}. \tag{4.43}$$

Dla sieci bezskalowych, w których $\gamma = 3$, wykładnik δ pozostaje z wykładnikiem określającym rozkład krotności wierzchołków w relacji o postaci: $\delta \geqslant (\gamma + 1)/2$, a rozkład pośrednictwa wynosi $P(b) \sim b^{-2}$ [109].

Sieci lingwistyczne charakteryzują się dużo mniejszym wykładnikiem γ, który – w zależności od języka – przyjmuje wartość w zakresie $1.66 \leqslant \gamma \leqslant 1.78$ [111]. W związku z tym, odpowiadający im teoretyczny zakres zmienności δ zawiera się w przedziale $1.33 \div 1.39$. Na rysunku 4.26 przedstawiono rozkłady $P(b)$ sporządzone dla kilku analizowanych języków. Rozkłady te skalują się w zakresie kilku dekad zmienności b, z odpowiednimi wykładnikami δ różniących się od przewidzianej teoretycznie wartości o ok. 13%. Trzeba jednak wziąć pod uwagę rzeczywisty – a nie modelowy – charakter sieci, który może ukrywać dodatkowe parametry wpływające na otrzymany rozkład. Otrzymane rozkłady dla współczynników: grownowania i pośrednictwa świadczą, że sieci oparte o sąsiedztwo słów wykazują znamiona sieci złożonych o charakterze bezskalowym. Spójność uzyskanych wyników w kontekście proponowanych modeli teoretycznych świadczy o użyteczności stosowanego podejścia, czyli opisu języka naturalnego w formalizmie teorii sieci.



Rysunek 4.26: Rozkłady pośrednictwa $P(b)$ dla empirycznych sieci sąsiedztwa słów, uzyskanych dla zbiorów tekstów napisanych w wybranych językach europejskich. Czerwona, przerywana linia określa nachylenie rozkładu $P(b)$, określone przez wykładnik δ.

4.2.5.3 Średnia długość najkrótszej ścieżki

Jednym z najszerzej dyskutowanych właściwości sieci bezskalowych jest niewielka odległość pomiędzy dwoma losowo wybranymi wierzchołkami. Efekt ten jest możliwy dzięki istnieniu wierzchołków pośredniczących, których wyeliminowanie może prowadzić do znacznego zwiększenia odległości międzywęzłowej. Średnia długość najkrótszej ścieżki ℓ może być rozumiana jako efektywny rozmiar sieci i różni się od *średnicy* sieci, definiowanej jako długość najdłuższej spośród najkrótszych ścieżek pomiędzy wszystkimi parami węzłów w sieci. Jest to miara bardzo często używana do statystycznego określenia wzajemnej bliskości węzłów. Mała wartość ℓ jest znamienna dla wielu sieci, takich jak Internet (gdzie determinuje szybkość przesyłania informacji bądź rozprzestrzeniania się wirusów), sieci komunikacyjne (gdzie służy do optymalizacji przebywanej drogi) czy społecznych (określa szybkość rozprzestrzeniania się chorób zakaźnych).

Istnieje wiele podejść do analitycznego wyznaczenia średniej długości najkrótszej ścieżki dla sieci o zadanej topologii, biorących pod uwagę jedynie klasę sieci (sieci losowe ER, sieci bezskalowe) lub też dodatkowo odpowiednie parametry (np. wykładnik skalowania γ dla rozkładów $P(k)$) [106]. Dla sieci losowych o dobrze określonej średniej wartości $\langle k \rangle$, np. sieci typu ER i WS, liczbę wszystkich wierzchołków można zgrubnie oszacować: jeśli liczba najbliższych sąsiadów jakiegoś wierzchołka wynosi k_1, wówczas około k_1^ℓ węzłów jest w odległości mniejszej bądź równej ℓ. Stąd, jeśli $N \sim k_1^\ell$, to $\ell \sim \ln N / \ln k_1$ (co jest praktyczną wersją wzoru (3.9)).

Sieci o bezskalowych rozkładach krotności $P(k)$ charakteryzują się jeszcze mniejszą wartością średniej długości najkrótszej ścieżki, której wartość w przypadku mo-

delu BA ($\gamma = 3$) została analitycznie wyznaczona jako [105]:

$$\ell(N) \sim \ln N / \ln \ln N. \tag{4.44}$$

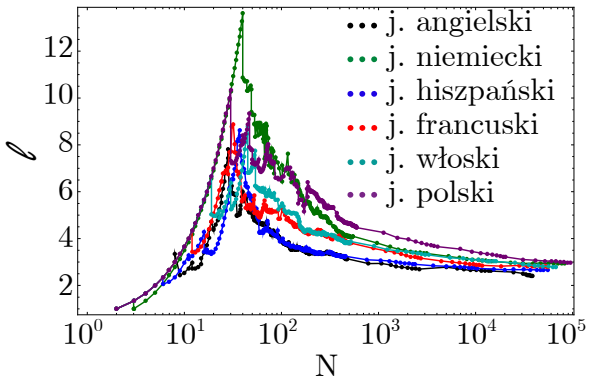
Jak można dostrzec, nie jest to uniwersalna formuła dla całej klasy sieci o rozkładach krotności typu, $P(k) \sim k^{-\gamma}$, ponieważ wraz ze spadkiem wartości wykładnika γ (co jest równoważne wzrostowi udziału hubów w sieci) zmniejsza się też wartość średniej długości najkrótszej ścieżki. Dla niskich wartości wykładnika skalowania $2 < \gamma < 3$ wartość tej wielkości może być wyrażona przez [161]:

$$\ell(N) \sim \ln \ln N \tag{4.45}$$

i osiąga dla odpowiednio dużych N saturację, której poziom zalezy od γ [106]:

$$\lim_{N \to \infty} \ell = 2/(3-\gamma) + 1/2. \tag{4.46}$$

Powyższe wzory dobrze przybliżają wartości ℓ osiągane w symulacjach. Wyraźnie wolniejszy wzrost ℓ niż ten przewidziany dla modelu WS (wzór 3.9) powoduje, że sieci takie zostały określane mianem *ultramałych światów*.

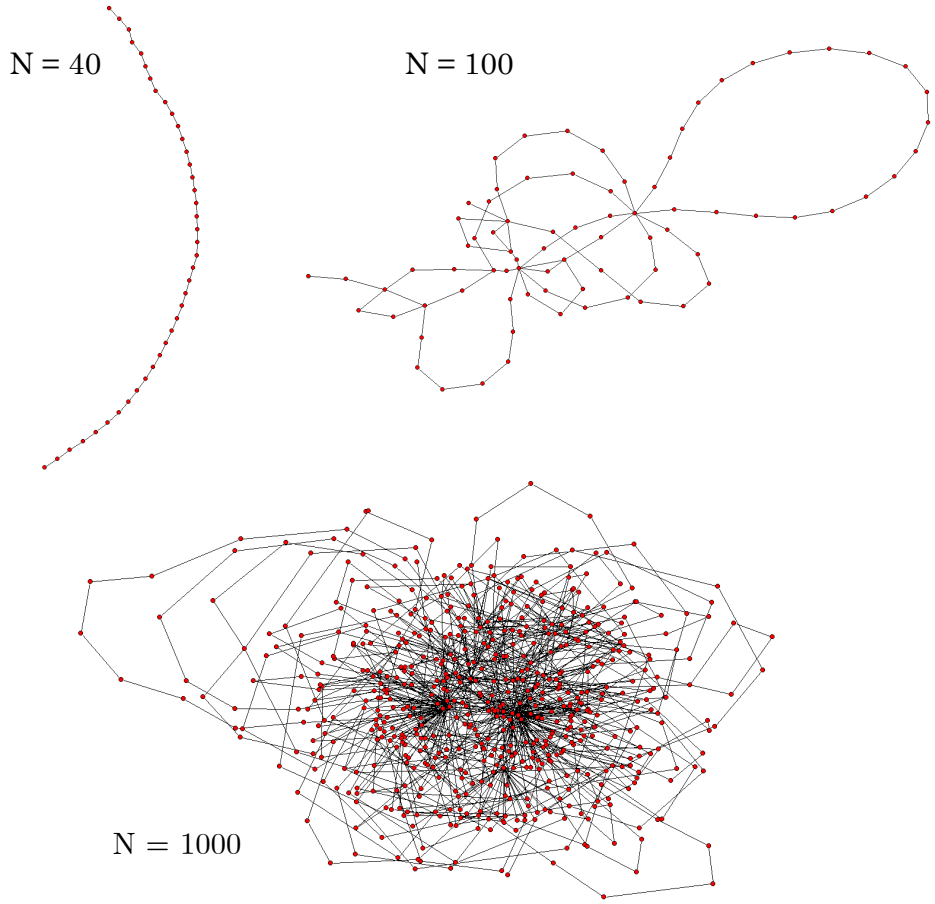


Rysunek 4.27: Średnia długość najkrótszej ścieżki dla sieci sąsiedztwa słów, zbudowanej z korpusów tekstów literackich w kilku językach europejskich. Obserwowany szybki wzrost średniej długości ścieżki jest bezpośrednią konsekwencją istnienia reguł gramatycznych, określających porządek słów w tekście.

Empiryczne sieci sąsiedztwa słów charakteryzują się niższą wartością wykładnika skalowania γ (większa rola hubów) od tych w powyżej omawianych modelach, co ma swoje odzwierciedlenie w wartościach średniej długości najkrótszej ścieżki. Zachowanie tej wielkości w zależności od rozmiaru sieci N przedstawiono na rysunku 4.27. Przebieg uzyskanych zależności $\ell(N)$ jest znamienny dla tego rodzaju sieci, gdzie wartość maksymalna ℓ jest osiągana bardzo szybko, a następnie spada, asymptotycznie dążąc do wartości minimalnej.

Znacznym wyzwaniem w tej sytuacji jest zaproponowanie takiego modelu tworzenia sieci, za pomocą którego będzie można odtworzyć tę własność sieci sąsiedztwa słów. Przedstawione powyżej wzory są rosnącymi funkcjami rozmiaru sieci N, stąd nie nadają się do opisu danych empirycznych. Głównym czynnikiem prowadzącym do tej rozbieżności jest specyficzny charakter wzrostu sieci sąsiedztwa, na którą są

nałożone więzy związane z gramatyką i stylem. Po pierwsze, sieci te posiadają na ogół pamięć o kilkunastu/kilkudziesięciu odwiedzonych ostatnio węzłach i istnieje „okres karencji", kiedy taki węzeł nie może być odwiedzony ponownie (po to, aby uczynić zadość gramatyce lub po to, aby tego samego słowa nie używać zbyt często ze względów stylistycznych). O ile wpływ tego efektu na ℓ jest pomijalny, gdy sieć jest już wystarczająco duża, o tyle może całkowicie determinować topologię sieci, gdy jest ona w fazie wczesnego wzrostu. Wiele słów jest wówczas używanych po raz pierwszy, powodując tworzenie się długich, swobodnych łańcuchów i pętli, co znacznie zwiększa długości ścieżek między wierzchołkami. W miarę wzrostu sieci nowe węzły pojawiają się coraz rzadziej (prawo Heapsa), a ewolucja odbywa się głównie przez zagęszczanie krawędzi, co w naturalny sposób skraca ℓ. Przykład rosnącej sieci sąsiedztwa dla rzeczywistego tekstu przedstawia rysunek 4.28. Wprawdzie wzrost *dojrzałej* sieci sąsiedztwa słów przypomina jakościowo sieci o przyspieszonym wzroście, to jednak w modelach Yule'a-Simona-Heapsa czy Dorogowcewa-Mendesa uzyskanie podobnych struktur, mimo teoretycznej dopuszczalności, w praktyce nie jest możliwe. Na przeszkodzie stoi fakt, że oba modele nie mają pamięci i traktują wzrost sieci jako sekwencję niezależnych zdarzeń, determinowanych jedynie określonymi rozkładami prawdopodobieństwa wystąpienia.



Rysunek 4.28: Typowy obraz wczesnych faz wzrostu sieci sąsiedztwa N słów dla rzeczywistego tekstu (*Lalka* B. Prusa).

Biorąc pod uwagę bezskalowy charakter sieci, średnia długość najkrótszej ścieżki $\ell(N)$ musi być funkcją parametru α, wyrażającego przyspieszony wzrost (wzór (4.21)). Mając na uwadze, że nawiązywanie połączeń pomiędzy istniejącymi węzłami również jest oparte o preferencyjne przyłączanie, średnia długość najkrótszej ścieżki powinna być w odwrotności do wzrastającego tempa przyrostu. Aby jakościowo wyjaśnić asymptotyczne zmniejszanie się wartości $\ell(N)$ ze wzrostem N, przyjąć można dla uproszczenia, że słuszna dla sieci sąsiedztwa jest zależność $\ell(N)$ taka, jak dla sieci przypadkowych ER, dana wzorem (3.9), gdzie $\langle k \rangle = 2e(N)/N$:

$$\ell(N) \sim \frac{\ln N}{\ln \frac{2e(N)}{N}}. \tag{4.47}$$

Jest to zależność słuszna dla $\gamma > 3$ [106], jednak w rzeczywistym przypadku $\gamma < 3$, co oznacza, że dla sieci sąsiedztwa słów powyższe założenie da górne ograniczenie prawdziwej zależności. Podstawiając teraz wielkość określającą liczbę krawędzi w sieci o przyspieszonym wzroście (ze wzoru (4.21)):

$$e(N) = \frac{c}{(\alpha+1)} N^{\alpha+1}, \tag{4.48}$$

otrzymuje się:

$$\ell(N) \sim \frac{\ln N}{\ln \frac{2c}{\alpha+1} + \alpha \ln N}. \tag{4.49}$$

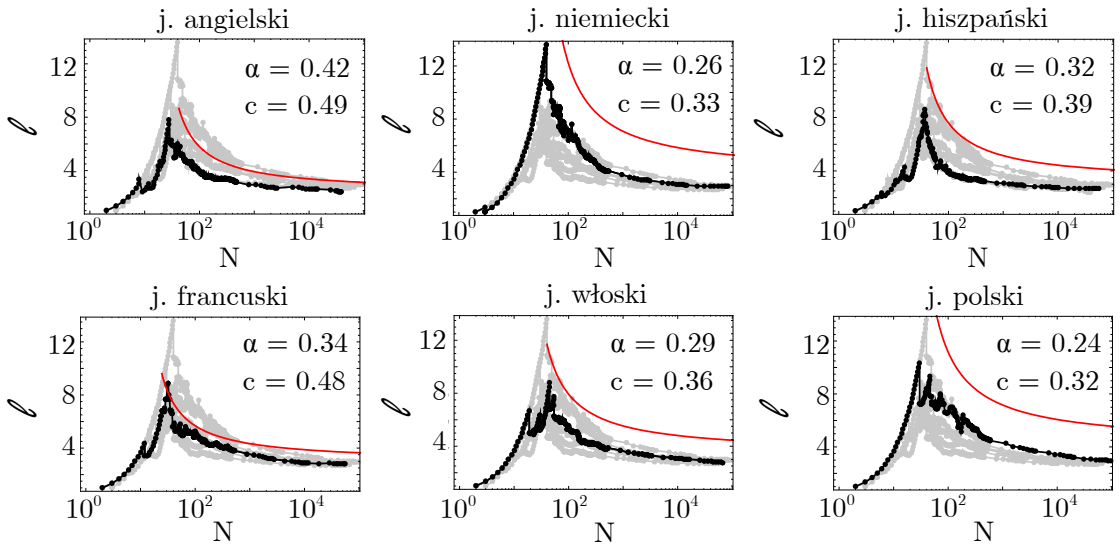
Dla $c < (\alpha+1)/2$ powyższe wyrażenie jest malejącą funkcją liczby węzłów N, której asymptotyczne zachowanie opisuje granica:

$$\lim_{N\to\infty} \ell(N) = \frac{1}{\alpha}. \tag{4.50}$$

Można przypuszczać, że sieci empiryczne będą wykazywać asymptotyczną wartość $\ell(N)$ mniejszą od tej danej powyższą granicą (ze względu na małą wartość γ). Przedstawione rozumowanie ma przede wszystkim charakter heurystyczny, a nie analityczny, pozwala jednak na jakościowe wytłumaczenie zachowania się średniej długości najkrótszej ścieżki dla sieci lingwistycznych.

Na rysunku 4.29 przedstawiono reprezentatywne rozkłady $\ell(N)$ dla analizowanych przykładów sieci, powstałych w oparciu o strukturę języków naturalnych. W każdym z tych przypadków można zaobserwować maksymalną wartość $\ell_{\max}$, po osiągnięciu której wartości średniej długości najkrótszej ścieżki asymptotycznie maleją. Kolorem czerwonym przedstawiono przebieg funkcji (4.49) z wartościami parametrów c i α uzyskanymi z dopasowania modelu YSH do danych empirycznych (tabela 4.1 na str. 54). Rozbieżność pomiędzy rzeczywistym a teoretycznym przebiegiem zmienności ℓ ma charakter stały, co prawdopodobnie wynika z upraszczającego założenia, że funkcyjna zależność długości najkrótszej ścieżki dana jest przez wzór (3.9).

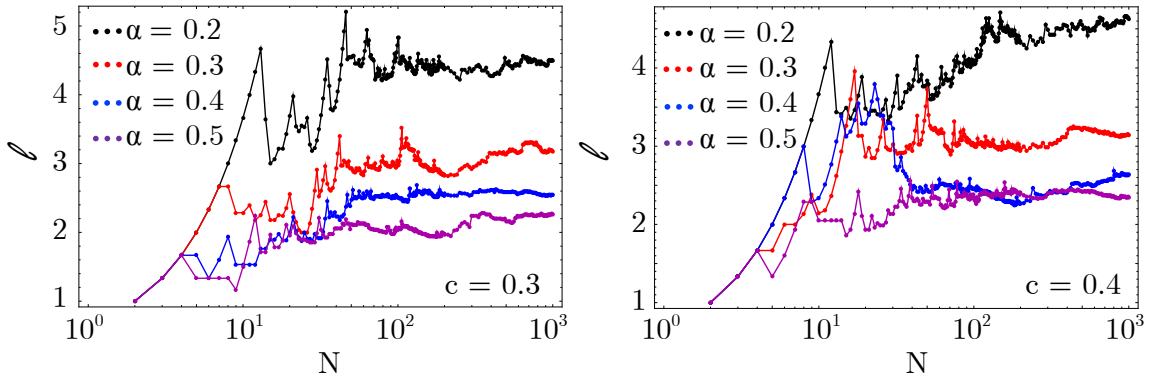
W celu dalszej weryfikacji adekwatności przedstawionych modeli sieci sąsiedztwa słów, dla każdego z nich wyznaczona została numerycznie zależność $\ell(N)$ i porównana z danymi empirycznymi. Wyniki dla sieci wygenerowanej wedle modelu opartego o mechanizm Yule-Simona-Heapsa, opisany w podrozdziale 4.1.1, przedstawiono na rysunku 4.30. Wartości parametrów α i c dobrano w zakresie obserwowanym w danych empirycznych.



Rysunek 4.29: Średnia długość najkrótszej ścieżki dla fragmentów tekstów napisanych w poszczególnych językach wraz z dopasowaną (kolorem czerwonym) niezależnie w każdym przypadku modelową funkcją (4.49), określoną empirycznie uzyskanymi parametrami parametrami α i c (patrz tabela 4.1).

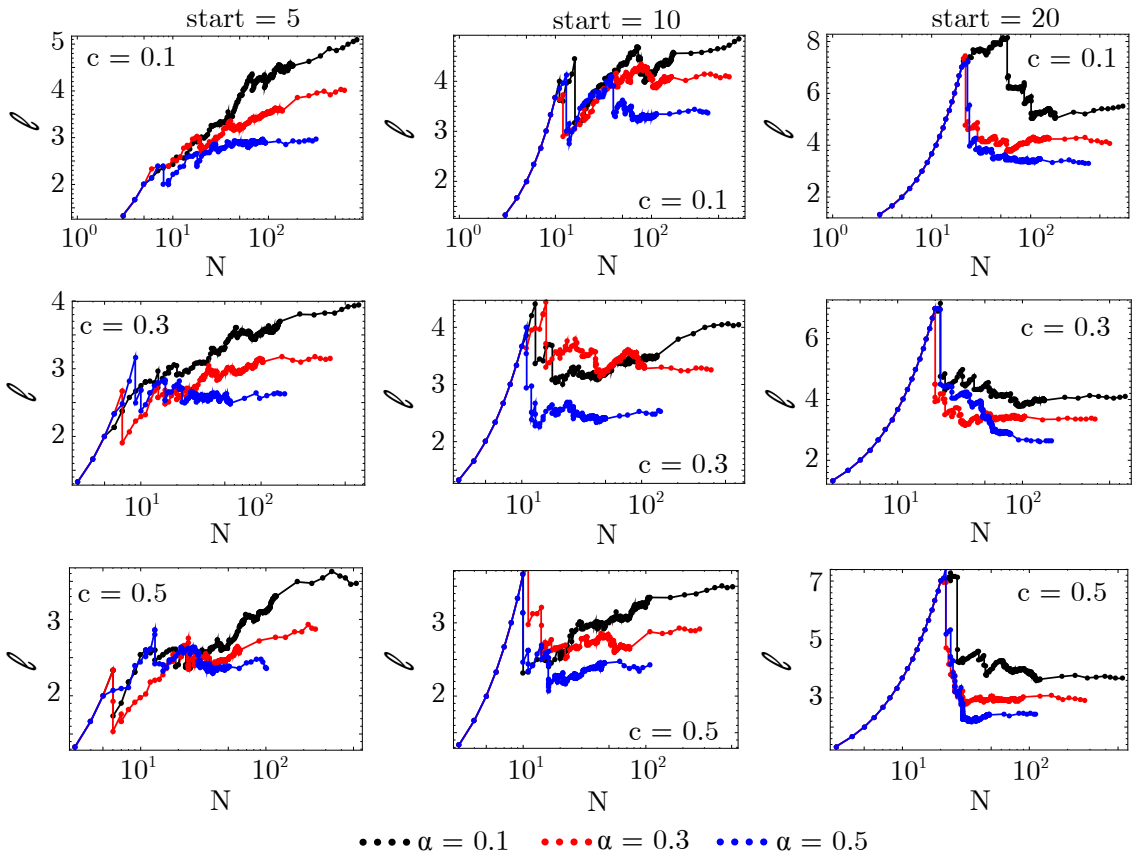
Dla $\alpha > 0.2$ i dużych wartości N obserwuje się mniej więcej stały poziom $\ell(N)$, natomiast dla $\alpha \leqslant 0.2$ funkcja $\ell(N)$ rośnie (co koresponduje z niewielkim tempem przyrostu krawędzi wewnątrz sieci i jej ewolucji głównie poprzez przyłączanie nowych wierzchołków).

Przebieg średniej długości najkrótszej ścieżki wyznaczony dla sieci generowanych wedle modelu DM-AG został przedstawiony na rysunku 4.31. Dla każdej z symulacji przyjęto pewne warunki początkowe, będące łańcuchami o długości $N_0 = 5$, 10 i 20 połączonych ze sobą wierzchołków. Ponadto do każdej symulacji przyjęto inne wartości stałych α i c, tak dobrane, by odpowiadały swoim empirycznym odpowiednikom. Wybór warunku początkowego N_0 spełnia istotną rolę w zachowaniu się średniej długości najkrótszej ścieżki.



Rysunek 4.30: Średnia długość najkrótszej ścieżki $\ell(N)$ dla modelu Yule'a-Simona-Heapsa z różnymi wartościami parametrów α oraz c zaznaczonych na rysunku odpowiednimi kolorami.

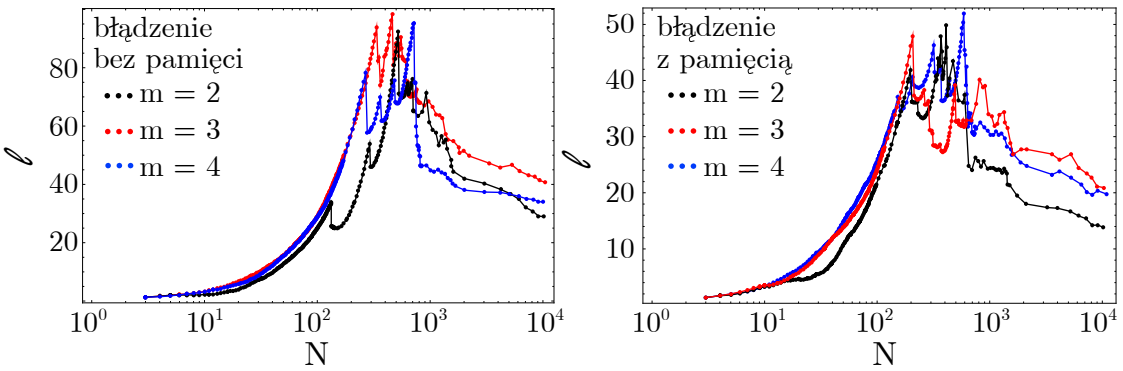
W przypadku, gdy łańcuch jest krótki, sieć rozrasta się od początku w sposób przewidziany przez model, czyli wraz ze wzrostem parametru α, wartość $\ell(N)$ szybko się wysyca. Natomiast im wartość N_0 jest większa, tym początkowy wzrost $\ell(N)$ staje się silniejszy; po jakimś czasie długość ścieżki osiąga maksimum i zaczyna przejściowo spadać (w wyniku zapętlenia się początkowego łańcucha), dochodząc w końcu do zachowania przewidzianego przez model. Efekt ten jest tego samego rodzaju, co opisany w modelu Wattsa i Strogatza, gdzie niewielka, losowa zamiana krawędzi w sieci regularnej o dużym ℓ prowadzi do znacznego zmniejszenia odległości między początkowo odległymi wierzchołkami.



Rysunek 4.31: Średnia długość najkrótszej ścieżki dla modelu DM-AG z różnymi wartościami parametrów c i α i różną długością początkowego łańcucha wierzchołków. We wszystkich realizacjach średnia długość najkrótszej ścieżki osiąga niewielką wartość, $\ell_{\max} < 8.09$.

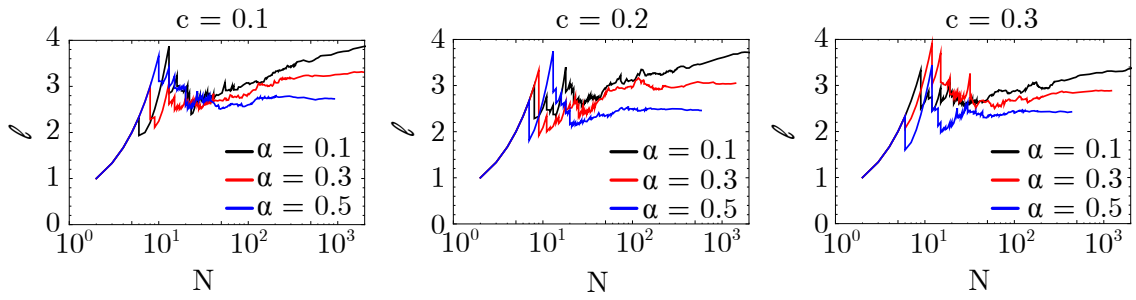
Następnym przykładem jest model błądzenia po sieci bezskalowej. Przebieg zmian wartości $\ell(N)$ dla powstałej w wyniku tego mechanizmu sieci przedstawiono na rysunku 4.32. W przypadku nieuwzględnienia pamięci błądzenia, $\ell(N)$ szybko rośnie do znacznych wartości, jednak w momencie przejścia przez już wcześniej odwiedzony w procesie wierzchołek wzrost średniej długości najkrótszej ścieżki ulega załamaniu i jej wartość zaczyna sukcesywnie maleć wraz ze wzrostem sieci. Przy uwzględnieniu efektu pamięci przebieg jest podobny, jednak maksimum osiągane przez $\ell(N)$ jest o połowę mniejsze.

W obu przypadkach obserwuje się zmniejszanie wartości $\ell(N)$ po osiągnięciu przez sieć odpowiedniego rozmiaru. Wartości parametru m, określającego początkowy stopień węzła w sieci pierwotnej, nie wpływa zasadniczo na zachowanie się $\ell(N)$ dla sieci wtórnej. Analizowany model błądzenia po sieci charakteryzują więc odmienne wartości średniej długości najkrótszej ścieżki oraz rozmiaru sieci, przy którym jest osiągane maksimum $\ell(N)$.



Rysunek 4.32: Średnia długość najkrótszej ścieżki dla modelu błądzenia po sieci bezskalowej z pamięcią (po prawej) i bez pamięci (po lewej) dla rożnych wartości parametru m, określającego początkowy stopień wierzchołka przyłączanego do sieci.

Ostatni z rozważanych modeli to błądzenie po sieci o przyspieszonym wzroście. Wyniki symulacji przedstawione zostały na rysunku 4.33. W porównaniu z błądzeniem po statycznej sieci bezskalowej (rysunek 4.32) widać, że tym razem spadek $\ell(N)$ ze wzrostem N jest tylko przejściowy, do momentu aż zaniknie pamięć o przyjętym warunku początkowym w postaci pierścienia wierzchołków. Później zachowanie się $\ell(N)$ w sieci wtórnej zaczyna przypominać swój odpowiednik dla modelu DM-AG, rysunek 4.31. Topologia sieci wtórnej jest zatem odbiciem topologii sieci pierwotnej. Oznacza to, że z punktu widzenia modelowania sieci sąsiedztwa słów rozpatrywany model błądzenia nie posiada dodatkowych atutów w stosunku do modelu DM-AG.



Rysunek 4.33: Średnia długość najkrótszej ścieżki $\ell(N)$ dla błądzenia po sieci o przyspieszonym wzroście (model DM-AG) dla różnych wartości parametrów α i c. Każdy przyłączany wierzchołek ma krotność $m = 2$, a docelowa sieć składa się z $V = 10^5$ wierzchołków.

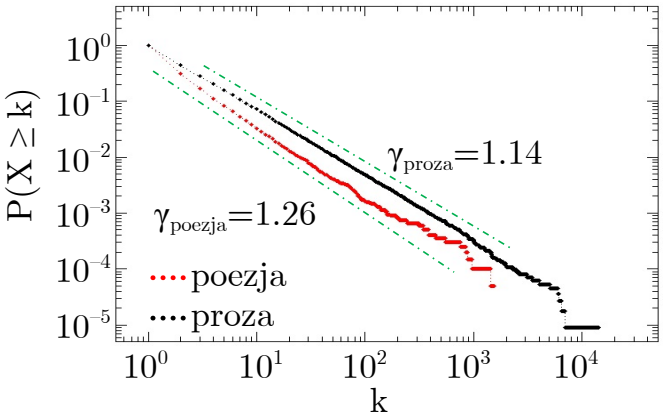
Sieci sąsiedztwa słów charakteryzują się specyficznym przebiegiem średniej długości najkrótszej ścieżki, która po osiągnięciu przez sieć odpowiedniego rozmiaru zaczyna asymptotycznie maleć. Oznacza to, że wraz z dojrzewaniem sieci, staje się ona coraz bardziej efektywna komunikacyjnie. Podobny charakter wzrostu sieci może być obserwowany w niektórych innych układach rzeczywistych, np. w sieciach społecznych [162]. Można byłoby się intuicyjnie spodziewać, że to zjawisko będzie obserwowane także w sieciach stron WWW, gdyż wraz z rozwojem tej sieci jej rozmiar (liczba wierzchołków) zwiększył się tak bardzo, że w grę mogą zacząć wchodzić efekty skończonej liczby osób i instytucji, które te strony prowadzą. Przez to w pewnym momencie wzrost sieci WWW mógłby się realizować głównie przez dodawanie wewnętrznych połączeń, a nie przez dodawanie nowych wierzchołków. Tymczasem w rzeczywistości obserwowane są jedynie efekty powolnej saturacji wzrostu $\ell(N)$, a nie zmniejszania się jej wartości [163, 164]. Zjawisko to można prawdopodobnie wytłumaczyć przez strukturę stron WWW, w których duże dokumenty HTML są dzielone na mniejsze, celem ich szybszego transferu i łatwiejszej nawigacji. W ten sposób rozwój sieci ma miejsce ciągle w pierwszym rzędzie przez dodawanie nowych wierzchołków (dokumentów), a nie odnośników pomiędzy już istniejącym wierzchołkami.

4.2.6 Charakterystyki sieciowe literatury światowej

Zróżnicowanie słownictwa języków naturalnych jest rzeczą oczywistą, jednak spojrzenie na ich reprezentację sieciową może dostarczyć nowych informacji o własnościach języka. Wyniki uzyskane na drodze analizy sieci sąsiedztwa słów wskazały na dość uniwersalny obraz struktury takich sieci, którą można zakwalifikować do klasy sieci z przyspieszonym wzrostem. Mimo tego jakościowego podobieństwa sieci dla różnych próbek języka, ich ilościowe miary nie są już takie same. Charakterystyczne konstrukcje gramatyczne, stylistyka i warsztat pisarski oraz kontekst i wykorzystywane motywy mogą mieć swoje odbicie w charakterystykach sieci lingwistycznych [165].

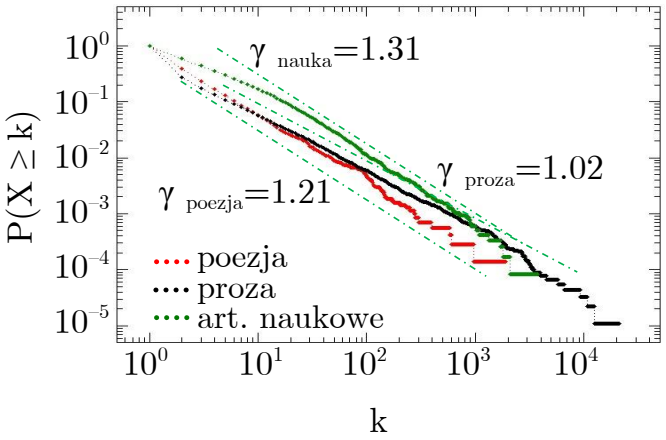
W celu porównania statystycznych właściwości stopni wierzchołków dla wielu typów tekstów i języków stworzono pięć różnych korpusów zawierających: angielską prozę, angielską poezję, angielskie prace naukowe, polską prozę oraz polską poezję. Dla każdego z tych korpusów wyznaczono skumulowany rozkład wierzchołków $P(X \geqslant k)$. Na rysunkach 4.34 oraz 4.35 przedstawiono te rozkłady wraz z odpowiadającymi im nachyleniami. Biorąc pod uwagę prozę, rozkład krotności dla języka angielskiego posiada mniejsze nachylenie niż odpowiedni rozkład krotności sporządzony dla prozy w języku polskim. Rozkłady sporządzone dla korpusów będących poezją również się różnią, posiadając znacznie większe nachylenie, szczególnie obserwowane dla języka angielskiego. Zgrubnie rzecz ujmując, zbiory poezji i prozy, napisane w języku polskim, w świetle prowadzonej analizy wyglądają podobnie, natomiast obserwowane są wyraźnie różnice w przypadku języka angielskiego.

Rozkład krotności $P(X \geqslant k)$ dla korpusu, zawierającego artykuły naukowe, ujawnia największe nachylenie z $\gamma_{\text{nauka}} = 1.31$. Nie jest to zaskakujące, biorąc pod uwagę charakter tekstów naukowych, zawierających sporo specjalistycznych zwrotów i wyrażeń, często nie posiadających synonimicznych odpowiedników, stąd ich

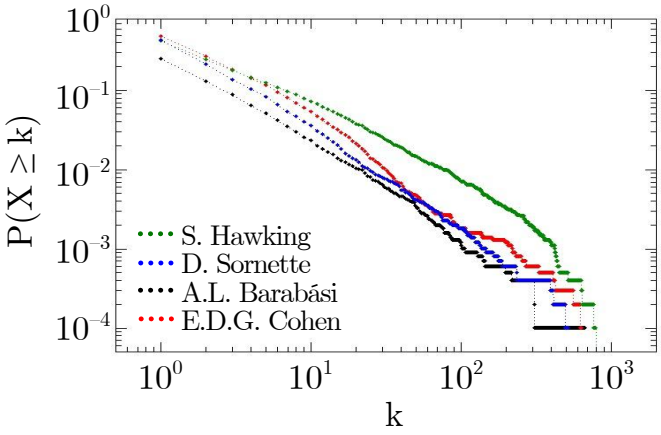


Rysunek 4.34: Rozkłady krotności wierzchołków dla sieci lingwistycznej, opartej o teksty napisane w języku polskim. Kolorem czerwonym przedstawiono rozkład $P(X \geqslant k)$ dla tekstów będących zbiorem poezji, kolorem czarnym – teksty które zostały napisane prozą.

istnienie prowadzi do zauważalnego zubożenia słownikowego w porównaniu do prozy, gdzie nie obserwuje się tak dalekich ograniczeń ze względu na dobór słów. Biorąc pod uwagę rozkłady krotności $P(X \geqslant k)$ sporządzone dla zbiorów zawierających artykuły naukowe konkretnych autorów (rysunek 4.36), nie zawsze obserwuje się skalowanie stopni wierzchołków. Istnieje silna zależność pomiędzy używaną terminologią [166] a rozkładami krotności wierzchołków, im mniej „zmatematyzowanego" i uściślonego słownictwa, tym bardziej zachowanie $P(X \geqslant k)$ ma charakter rozkładów bezskalowych.



Rysunek 4.35: Rozkłady krotności wierzchołków dla sieci lingwistycznej, opartej o teksty napisane w języku anielskim. Kolorem czerwonym przedstawiono rozkład $P(X \geqslant k)$ dla tekstów będących zbiorem poezji, kolorem czarnym – teksty, które zostały napisane prozą, a kolorem zielonym – teksty będące zbiorem artykułów naukowych z zakresu fizyki.



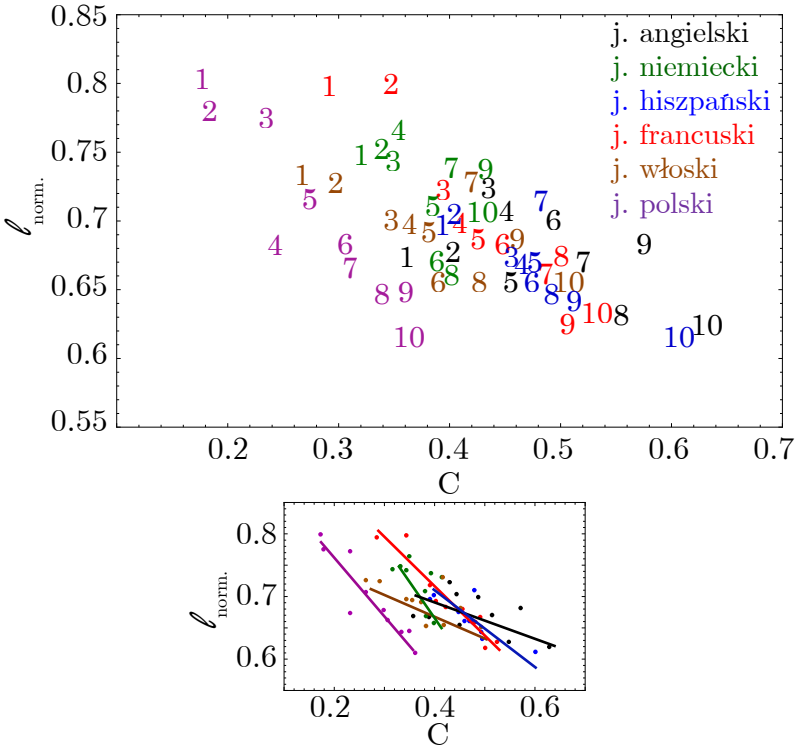
Rysunek 4.36: Rozkłady krotności wierzchołków dla sieci lingwistycznej, opartej o artykuły naukowe napisane przez różnych autorów. Odpowiednimi kolorami oznaczono rozkłady odpowiadające pracom fizyków: S. Hawking, D. Sornette, A.L. Barabási, E.D.G. Cohen

Inną ciekawą reprezentacją pozwalającą na ujawnienie właściwości językowych jest zostawienie globalnych parametrów sieciowych sporządzonych dla konkretnych tekstów pisanych. Do analizy wzięto pod uwagę po 10 reprezentatywnych utworów literackich napisanych w 6 językach europejskich. Jako naturalny wybór jest branie pod uwagę utworów napisanych prozą, która wyraża mowę pozbawioną stałego wzorca rytmicznego, codzienną lub przynajmniej pozbawioną słownictwa specjalistycznego. Utwory te przedstawiono w reprezentacji sieciowej sąsiedztwa słów oraz wyznaczono parametry charakteryzujące topologię tej sieci: średnią długość najkrótszej ścieżki, współczynnik gronowania i pośrednictwo. Ze względu na występująca zależność funkcyjną $\ell(N)$, szerzej opisaną w podrozdziale 4.2.5.3, i różne długości badanych tekstów literackich, wielkość tę znormalizowano do logarytmu:

$$\ell_{\text{norm}} = \ell / \ln N, \tag{4.51}$$

gdzie N jest całkowitą liczbą różnych słów występujących w konkretnym tekście. Wyniki uzyskane na drodze analizy ilościowej przedstawiono na rysunkach 4.37 oraz 4.38.

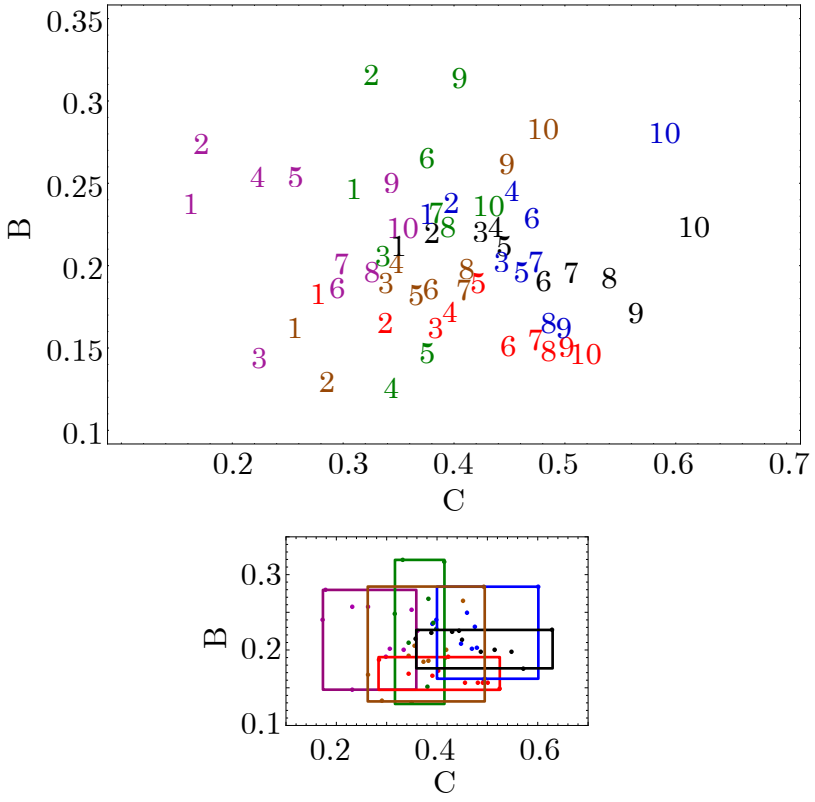
Wielkości ℓ_{norm} i C okazują się być skorelowane ze sobą. Wraz ze wzrostem współczynnika gronowania średnia długość najkrótszej ścieżki maleje, co ma miejsce dla każdego z rozważanych języków. Ta zależność jest naturalną konsekwencją faktu, że coraz większa gęstość krawędzi w sieci, z jednej strony, sprzyja silniejszemu gronowaniu i wzrostowi C, a z drugiej powoduje skracanie odległości między wierzchołkami. Mimo niewielkiej statystyki wyników pokazanych na rysunku 4.37, można zauważyć pewne ilościowe różnice pomiędzy językami. Najbardziej znacząca okazuje się ta dla języka polskiego, gdzie wartości współczynników gronowania są na ogół niższe niż dla pozostałych języków. Różnicę tę można tłumaczyć w kontekście usytuowania ich w różnych podgrupach rodziny języków indoeuropejskich. Badane języki reprezentują trzy zasadnicze grupy: germańską (język niemiecki, język angielski), italocelticką (język włoski, język francuski, język hiszpański) oraz bałtosłowiańską (język polski).



Rysunek 4.37: Znormalizowana do logarytmu średnia długość najkrótszej ścieżki ℓ_{norm} vs. współczynnik gronowania C. Każdy punkt odpowiada jednemu utworowi literackiemu, a kolor odpowiada językowi jego oryginalnego tekstu. Na dolnym wykresie przedstawiono te same punkty, co na górnym, ale z dopasowanymi metodą najmniejszych kwadratów zależnościami liniowymi $\ell_{\text{norm}}(C)$.
Literatura angielska: 1 – Finnegans Wake, 2 – Ulysses, 3 – Dubliners, 4 – 1984, 5 – Moby Dick, 6 – Adventure of Sherlock Holmes, 7 – Oliver Twist, 8 – Gone with the Wind, 9 – Emma, 10 – Lord of the Ring. **Literatura niemiecka**: 1 – Das Perfum, 2 – Der Steppenwolf, 3 – Der Moloch, 4 – Der Prozess, 6 – Der Untertran, 7 – Buddenbrooks, 8 – Effi Briest, 9 – Also Sprach Zarathustra, 10 – Danziger. **Literatura hiszpańska**: 1 – Rayuela , 2 – La Tournee de Dios, 3 – El Club Dumas, 4 – Cien Anos de Soledad, 5 – La Sombra del Viento, 6 – El Juego del Angel, 7 – El Beso de la Mujer Arana, 8 – La reina descalza, 9 – La Mano de Fatima, 10 – Don Quijote. **Literatura francuska**: 1 – La Disparition, 2 – L'etranger, 3 – La Peste, 4 – Madame Bovary, 5 – Nana, 6 – La Reine Margot, 7 – Le Collier de la reine, 8 – Les Trois Mousquetaires, 9 – Les Miserebles, 10 – Le Comte de Monte Christo. **Literatura włoska**: 1 – Il Trionfo della Morte, 2 – Mastro don Gesualdo, 3 – Il Pendolo di Foucault, 4 – Il Nome della Rosa, 5 – L'isola del Giorno Prima, 6 – Mastro Don Gesualdo, 7 – Discorsi sopra la prima Deca di Tito Livio, 8 – Il Disprezzo, 9 – Il Gattopardo, 10 – Decameron. **Literatura polska**: 1 – Inny świat, 2 – Przedwiośnie, 3 – Granica, 4 – Ziemia obiecana, 5 – Opowieści o pilocie Pirxie, 6 – Lalka, 7 – Sława i chwała, 8 – Noce i dnie, 9 – Chłopi, 10 – Ogniem i mieczem.

Charakterystyki odpowiadających im sieci sąsiedztwa słów wskazują na odmienną strukturę języka polskiego, a zatem ujawniają inny charakter jego wewnętrznej organizacji niż języków należących do pozostałych grup. Na podstawie tych analiz nie można stwierdzić znaczącej różnicy pomiędzy grupami: germańską i italoceltycką. Z jednej strony może to wiązać się z niedostateczną czułością stosowanych metod analizy, z drugiej zaś strony języki te posiadają wspólną własność: o wiele mniej niż np. w języku polskim rozwiniętą fleksję, co może wpływać na podobieństwo topologii związanych z nimi sieci [167].

Na rysunku 4.38 zestawiono przeciętną wartość pośrednictwa B (uśrednionych po wszystkich wierzchołkach indywidualnych pośrednictw b_i, danych wzorem (3.10)) i współczynnika gronowania C dla różnych utworów. W tym przypadku jedyna zaobserwowana różnica międzyjęzykowa dotyczy języka francuskiego, którego pośrednictwo systematycznie przyjmuje stosunkowo niskie i zbliżone do siebie wartości. Trudno powiedzieć, czy ta różnica ma swoje źródło w specyfice struktury języka francuskiego, czy też jest wynikiem takiego, a nie innego doboru analizowanych utworów.



Rysunek 4.38: Uśrednione po wszystkich wierzchołkach sieci pośrednictwo B vs. współczynnik gronowania C dla analizowanych utworów (liczby i kolory przyporządkowane do konkretnych utworów są identyczne jak na rysunku 4.37. Na dolnym wykresie te same punkty zostały ujęte w prostokąty, by ułatwić optyczne określenie obszaru ich rozrzutu.

Reprezentacja utworów literackich w obrazie sieci lingwistycznych pozwala na identyfikację tekstów, pod kątem stylu czy specjalistycznego charakteru, jakie ono posiada. W pracy [111], przeanalizowano utwory wywodzące się z różnych gatunków literackich, m.in poezja i proza, jak i również teksty będące pracami naukowymi. Uzyskane charakterystyki sieciowe pozwoliły z zadowalającą dokładnością je zidentyfikować, wskazując również na zachodzące relacje pomiędzy samą naturą tych tekstów a uzyskanymi wartościami ℓ_{norm} , C, B. W przypadku utworów eksperymentalnych i niejednoznacznych w ocenach literaturoznawczych, jak *Finnegans Wake* czy *Rayuela* wyznaczone charakterystyki przybierają wartości ekstremalne.

Uwidacznia się ponadto, że stosowane metody mogą mieć charakter stylometryczny, charakteryzując styl pisarski danego autora. Osobliwy styl może zostać zidentyfikowany poprzez analizę wybranych tekstów jednego autora na tle utworów innych autorów. Przykładem może być tutaj charakterystyczny styl reprezentowany przez Jamesa Joyce'a, którego trzy utwory znalazły się w swoim bezpośrednim sąsiedztwie na rysunku 4.37.

Stosowana metodologia może być w związku z tym użyta również w celu przypisania konkretnej osobie tekstów o nieznanym autorstwie, tropienia plagiatów itd., stanowiąc dodatkowe narzędzie badawcze. Dogłębna analiza uzyskanych wyników w kontekście stylometrii wykracza jednak poza zakres prowadzonych w niniejszej pracy rozważań.

4.3 Język naturalny w obrazie analizy multifraktalnej

W tej części pracy próbki języków naturalnych poddane zostaną analizie multifraktalnej celem uzyskania informacji o stopniu złożoności i różnorodności ich struktury. Analiza multifraktalna okazała się bardzo użyteczna do opisu złożoności danych empirycznych pochodzących z rożnych układów, wśród których należy wymienić zwłaszcza rynki finansowe [168, 169, 170], muzykę [171] organizmy żywe [172], rozkład galaktyk [173], atmosferę ziemską i klimat oraz skorupę ziemską [174], a także takie zjawiska fizyczne, jak turbulencja [175, 176] i agregacja ograniczona dyfuzją [177]. W literaturze dostępne są też wyniki pierwszych prób zastosowania formalizmu multifraktalnego do analizy różnych aspektów języka naturalnego.

W pracy [178] próbki tekstów literackich zostały przetworzone na szeregi czasowe odległości pomiędzy wystąpieniami konkretnych kombinacji liter, a następnie poddane analizie pod kątem istnienia struktury wieloskalowej. Okazało się, że teksty w takiej reprezentacji wykazują korelacje dalekiego zasięgu oraz skalowanie multifraktalne. Podobne własności zostały odkryte w tekstach wyrażonych w postaci szeregów rang słów i długości słów, przy czym zaobserwowano różnice pomiędzy językiem naturalnym, reprezentowanym przez język angielski, a językiem sztucznym, reprezentowanym przez esperanto [91].

4.3.1 Wybór optymalnej reprezentacji języka

Język naturalny w postaci pisanej reprezentowany jest przez słowa, których odpowiednia sekwencja przedstawia myśl, jaką nadawca chce przekazać odbiorcy w sposób najbardziej zrozumiały i (na ogół) jednoznaczny. Najmniejszą jednostką struktury tekstu, która zawiera informację, jest zdanie. Słowo samo w sobie jest jedynie odniesieniem do pewnej rzeczy, stosunku do niej, wyrażać może czynność, określać przymiot lub być czystą formą gramatyczną, nie istniejącą samodzielnie. W momencie ułożenia słów w odpowiednim szyku stają się one razem nowym jakościowo tworem, za pomocą którego wyrażana jest spójna myśl, opis jakiegoś zjawiska, czyli konkretna informacja. Większe jednostki strukturalne tekstów, takie jak akapity i rozdziały, nie są już w tym samym stopniu spójne: informacja w nich zawarta może odnosić się do różnych obiektów. Zdanie można zatem uznać za podstawowy element strukturalny tekstów pisanych, materializujący znaczenie poszczególnych wyrazów w zależności od użytego szyku czy formy.

Użyteczną miarą zawartości informacyjnej zdania może być jego długość: dłuższe zdania są – intuicyjnie – bogatsze w informację od krótkich. Długość zdania może być rozumiana jako liczba występujących w nim znaków, słów czy zdań składowych, jeśli to zdanie jest złożone. W bieżącej analizie rozpatrywana będzie liczba słów składających się na zdania, ponieważ liczba ta wydaje się lepiej określać ilość informacji zawartej w zdaniu w porównaniu do liczby znaków czy liczby zdań składowych. Bierze się to stąd, że średnia długość pojedynczego słowa w zdaniu jest ujemnie skorelowana z liczbą słów tworzących to zdanie (o czym stanowi jedna z wersji prawa Menzeratha-Altmanna [179]), a więc ta sama długość zdań wyrażona w znakach może odpowiadać bądź bogatszym informacyjnie zdaniom z większą liczbą krótszych słów, bądź uboższym zdaniom z mniejszą liczbą dłuższych słów. Z kolei wzięcie pod uwagę liczby zdań składowych spowodowałoby niewielką zmienność miary opisującej długości tych zdań i, co za tym idzie, niewielką czułość tej miary.

Podział wypowiedzi na zdania jest jedną z istotnych cech tekstów pisanych (w języku mówionym podział na zdania jest na ogół trudniejszy do ustalenia, co nie oznacza, że nie istnieje), wpływającą na klarowność przekazu i jego zrozumienie przez potencjalnego odbiorcę. Można wskazać sytuacje, w których długość generowanych zdań jest silnie uzależniona od zewnętrznych czynników czy trybu formułowania wypowiedzi. Zdania, będące rozkazami na polu walki, są krótkie, ponadto posiadają znacznie mniejszą dyspersję długości aniżeli zdania stanowiące monolog (w tym monolog wewnętrzny).

Na ogół jednak długość zdań jest wynikiem zachowania balansu pomiędzy ich zrozumiałą formą a oddaniem złożoności sytuacji czy wyrażeniem form stylistycznych, intencjonalnie użytych przez autora. Czytelność danego zdania jest (statystycznie) funkcją jego długości, gdyż wraz z jej wzrostem, informacje zawarte w zdaniu są coraz trudniej przetwarzane przez odbiorcę m.in. na skutek ograniczonej pojemności jego pamięci roboczej. Jak się okazuje, długość zdań i ich rozkład są jednymi z cech, które mogą charakteryzować autora tekstu. W konsekwencji może to stanowić, obok klasycznych metod analitycznych, dodatkową, ilościową miarę stylometryczną (oraz estetyczną), określającą dany tekst.

4.3.2 Zastosowana metodologia analizy multifraktalnej

Szereg czasowy długości zdań utworzony na bazie rozpatrywanego tekstu można poddać analizie pod kątem występowania struktur multifraktalych [180]. Ze względu na nieunikniony w tym przypadku brak stacjonarności sygnałów, zastosowanie funkcji rozdziału (danej wzorem 3.25) do wyznaczania wymiarów fraktalnych D_q jest niestety mało efektywne. Problem ten pozwalają ominąć dwie konkurencyjne metody wyznaczania funkcji $f(\alpha)$[10]. Są to: metoda multifraktalnej analizy fluktuacji detrendowanych (ang. *Multifractal Detrended Fluctuation Analysis*, MF-DFA) [181] oraz metoda maksimów modułu transformaty falkowej (ang. *Wavelet Transform Modulus Maxima*, WTMM) [182]. Pierwsza z nich jest uogólnioną wersją metody DFA, służącej do analizy skalowania fluktuacji sygnału na różnych skalach czasowych [183], a druga identyfikuje obszary wielokrotnego skalowania w widmach transformaty falkowej sygnału. Obie metody, choć oparte na różnych narzędziach matematycznych, są z powodzeniem wykorzystywane do analizy struktur multifraktalnych w szeregach czasowych.

4.3.2.1 Metoda MF-DFA

Multifraktalna wersja metody DFA składa się z dwóch etapów: najpierw eliminuje się wpływ trendu na sygnał, następnie wyznacza funkcje opisujące zachowanie się fluktuacji na różnych skalach czasowych, a na końcu oblicza widmo uogólnionych wykładników Hursta lub widmo osobliwości $f(\alpha)$. Niech będzie zadany jednowymiarowy szereg czasowy $x(t_i)$ reprezentujący wyniki pomiarów wielkości X w chwilach $i = 1, ..., N$. Można wtedy wyznaczyć jego profil $Y(t_j)$, zdefiniowany jako:

$$Y(t_j) = \sum_{i=1}^{j}(x(t_i) - \langle x \rangle), \qquad \text{gdzie} \qquad \langle x \rangle = \frac{1}{N}\sum_{i=1}^{N} x(t_i), \tag{4.52}$$

który zostaje następnie dwukrotnie podzielony na segmenty o długości s, zaczynając podział raz od początku a raz od końca szeregu. Wyznaczanie segmentów w ten sposób pozwala uniknąć pominięcia punktów, znajdujących się na początku albo na końcu serii. Za każdym razem otrzymuje się $M_s = N/s$ rozłącznych segmentów. Każdy z powstałych segmentów ν pozbawiany jest lokalnego trendu poprzez odjęcie od profilu $Y(t_j)$ wielomianu $Z_\nu^{(k)}(t_j)$ o zadanym stopniu k, jednakowym dla wszystkich segmentów. Wariancja fluktuacji dla każdego z przedziałów ν jest dana funkcją:

$$F^2(\nu, s) = \frac{1}{s}\sum_{i=1}^{s}[Y((\nu - 1)s + i) - Z_\nu^{(k)}(t_i)]^2. \tag{4.53}$$

Ponieważ usuwany trend jest zależny od stopnia zastosowanego wielomianu Z, prowadzi to do istnienia kilku wariantów tej metody: dla $k = 1$ to MF-DFA1, dla $k = 2$ to MF-DFA2 itd. Wyliczając potęgi wariancji dla poszczególnych przedziałów i uśredniając je po wszystkich przedziałach, można wyznaczyć tzw. funkcję fluktuacji, daną wzorem:

$$F_q(s) = \frac{1}{2M_s}\left\{\sum_{\nu=1}^{2N_s}[F^2(\nu, s)]^{q/2}\right\}^{1/q}, \quad q \in \mathbf{R} \setminus \{0\}, \tag{4.54}$$

[10]Funkcja ta wyczerpująco jest opisana w podrozdziale 3.3.1.

w którym q spełnia rolę *parametru Rényi'ego*, nadającego efektywne wagi poszczególnym przedziałom ν w zależności od ich wariancji $F^2(\nu, s)$. Dla $q \ll 0$ przyczynek do $F_q(s)$ będzie pochodził głównie od segmentów o małej wariancji, natomiast przy $q \gg 0$ wkład do funkcji $F_q(s)$ będzie pochodził przede wszystkim od segmentów o dużej wariancji. W szczególnym przypadku, gdy $q = 2$, metoda MF-DFA redukuje się do zwykłego DFA [183].

Najważniejszym z praktycznego punktu widzenia etapem metody MF-DFA jest analiza zachowania $F_q(s)$ dla różnych skal s. Pozwala to na odróżnienie sygnałów o charakterze multifraktalnym od sygnałów monofraktalnych i w ogóle pozbawionych fraktalności. Dla sygnału o charakterze fraktalnym, zarówno mono-, jak i multifraktalnego, obserwuje się potęgową zależność:

$$F_q(s) \sim s^{h(q)}, \tag{4.55}$$

gdzie $h(q)$ jest rodziną wykładników, zwanych *uogólnionymi wykładnikami Hursta* [184]. W przypadku braku takiej zależności potęgowej sygnał nie ma charakteru fraktalnego. Zmienność wartości wykładnika $h(q)$ pozwala określić rodzaj fraktalności analizowanego szeregu: dla szeregów monofraktalnych jego wartość nie ulega zmianie, $h(q) = H$ dla każdego q, natomiast zmienność $h(q)$ oznacza, że skalowanie szeregu czasowego ma charakter multifraktalny. Stałą wartość H można utożsamić ze znanym z klasycznej analizy fluktuacji wykładnikiem Hursta [185] (zachodzi bowiem związek $H \equiv h(2)$). Znając wykładniki $h(q)$, możliwe jest wyznaczenie spektrum osobliwości $f(\alpha)$ z następujących związków:

$$\alpha = h(q) + qh'(q), \qquad f(\alpha) = q(\alpha - h(q)) + 1. \tag{4.56}$$

4.3.2.2 Metoda WTMM

Drugą metodą pozwalającą wyznaczyć spektrum osobliwości jest WTMM [182]. Opiera się ona na dekompozycji sygnału na składowe odpowiadające wybranej funkcji falkowej na różnych skalach czasowych, a następnie zbadaniu, czy otrzymane widma współczynników transformaty falkowej mają charakter multifraktalny. Niech dana będzie transformata falkowa w wersji ciągłej:

$$T_\Psi[x](t_0, s) = \frac{1}{s} \int \Psi\left(\frac{t - t_0}{s}\right) x(t) dt, \tag{4.57}$$

gdzie Ψ jest falką, t_0, s to, odpowiednio, parametry przesunięcia w czasie i skali czasowej, a $x(t)$ to analizowany sygnał. Lokalizacja w czasie i częstotliwości falki Ψ jest możliwa, jeśli istnieje skończona całka:

$$\int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty, \tag{4.58}$$

gdzie $\hat{\psi}(\omega)$ jest transformatą Fouriera falki macierzystej $\Psi(t)$. W analizie WTMM można wykorzystać każdą falkę macierzystą, która spełnia warunek (4.58), do celów analizy przedstawionej tutaj wygodny jest jednak wybór rodziny falek Hermite'a, będących pochodnymi funkcji Gaussa:

$$\psi^m(t) = \frac{d^m}{dt^m} e^{-\frac{t^2}{2}}. \tag{4.59}$$

Wybór ten jest optymalny w przypadku sygnałów niestacjonarnych, ponieważ falki Hermite'a są przydatne do usuwania trendów wielomianowych aż do stopnia $(m-1)$ włącznie. Opisując analizowany sygnał $x(t)$ za pomocą wielomianu Taylora oraz wykorzystując ortogonalność $\Psi(t)$ do wielomianu stopnia $(m-1)$, wzór (4.57) przyjmuje postać:

$$T_\Psi[x](t_0, s) = C|s|^{\alpha(t_0)} \int |x|^{\alpha(t_0)} \psi(t')dt', \tag{4.60}$$

a stąd:

$$T_\Psi[x](t_0, s) \sim s^{\alpha(t_0)}. \tag{4.61}$$

Relacja ta pozwala identyfikować osobliwości, jeśli występują pojedynczo, natomiast w przypadku ich nakładania się skalowanie jest niestabilne. Pomocna w tym przypadku jest identyfikacja lokalnych maksimów T_Ψ w celu wyliczenia funkcji rozdziału (danej wzorem (3.25 na str. 33)):

$$Z(q, s) = \sum_{l \in L(s)} |T_\psi(n_l(s), s)|^q, \tag{4.62}$$

gdzie $L(s)$ jest zbiorem wszystkich maksimów T_Ψ dla skali s, a $n_l(s)$ jest pozycją konkretnego maksimum. Oznaczenie $[x]$ zostało w powyższym wzorze pominięte dla uproszczenia notacji. Dla zachowania monotoniczności rodziny funkcji $Z(q, s)$ wprowadza się dodatkowy warunek ograniczający w postaci:

$$Z(q, s) = \sum_{l \in L(s)} (\sup_{s' \leqslant s} |T_\psi(n_l(s), s)|)^q. \tag{4.63}$$

W przypadku sygnałów fraktalnych wykładnik $\tau(q)$ opisuje potęgowe zachowanie funkcji rozdziału:

$$Z(q, s) \sim s^{\tau(q)}. \tag{4.64}$$

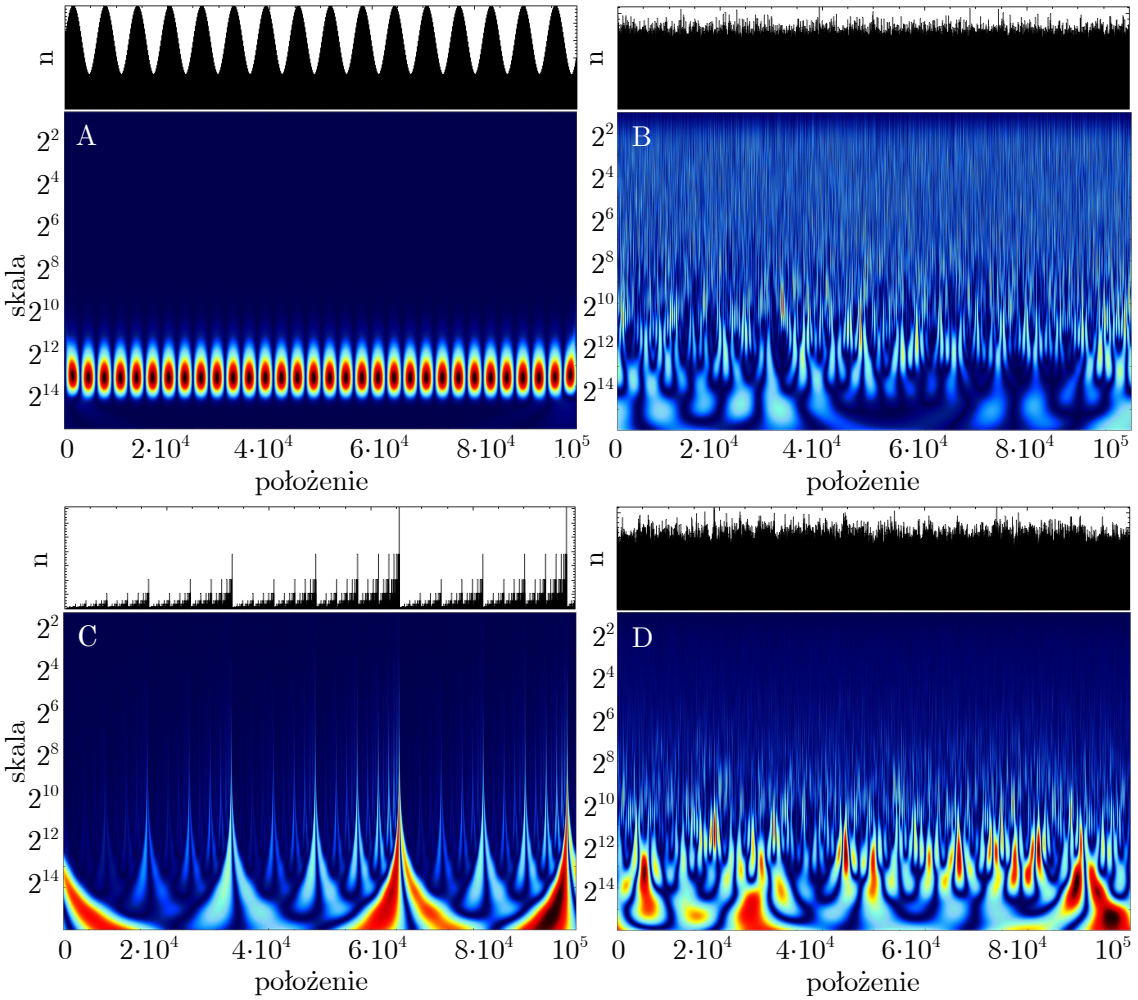
Wykładnik $\tau(q)$ w przypadku sygnałów multifraktalnych jest nieliniową funkcją q. Widmo osobliwości otrzymuje się poprzez transformację Legendre'a funkcji $\tau(q)$ wedle formuł:

$$\alpha = \tau'(q), \qquad f(\alpha) = q\alpha - \tau(q). \tag{4.65}$$

Wiedząc, że falka (4.57) posiada oscylacje, poprzez parametr s można dokonywać ich zagęszczenia tj. zwiększenia częstotliwości. Parametr t_0 jest z kolei związany z przestrzenną rozdzielczością transformaty, odpowiadając za translację falki wzdłuż przebiegu badanego sygnału. Przyjście do coraz mniejszych skal prowadzi do ujawnienia (o ile w rzeczywistości istnieją) jeszcze drobniejszych szczegółów w badanym sygnale, stąd transformata falkowa jest często określana jako „matematyczny mikroskop".

Współczynniki tej transformaty można przedstawić w graficzny sposób na dwuwymiarowej mapie położenie t_0 – skala s, obrazując kolorami ich wartości. Przykładowe widma współczynników transformaty falkowej zostały przedstawione na rysunku 4.39, obrazując wyniki dla sygnałów posiadających strukturę: regularną (sinusoida, okno A), monofraktalną (szum gaussowski, okno B) oraz multifraktalną (kaskada dwumianowa, okno C, i sygnał będący realizacją stochastycznego procesu

ARFIMA[11] z wykładnikiem Hursta $H = 0.5$, okno D). Dla sygnału regularnego istnieje jedna skala, dla której częstotliwość falki Ψ jest zbieżna z częstotliwością serii sinusoidalnej. Dla sygnału losowego nie istnieje znacząca dyspersja wartości współczynników, natomiast w przypadku sygnałów multifraktalnych C i D mapa ukazuje szeroki zakres takiej dyspersji.



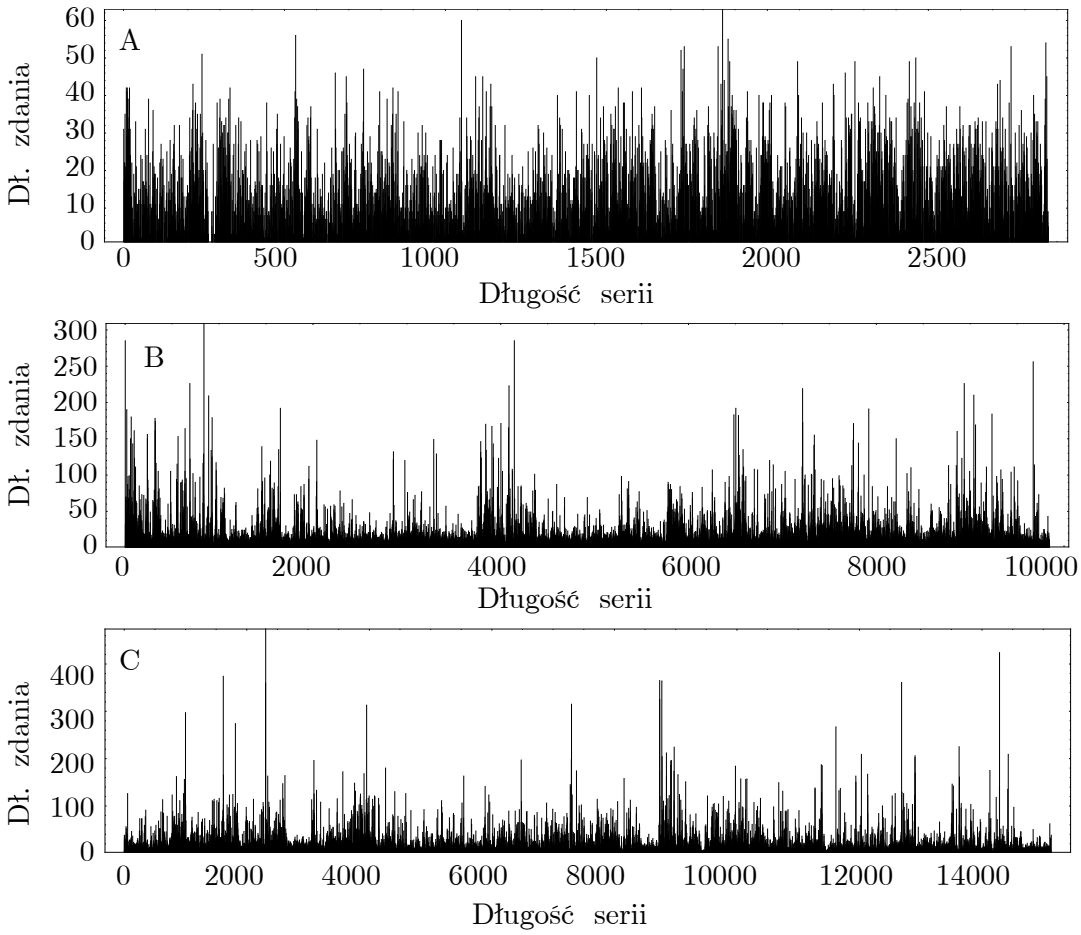
Rysunek 4.39: Skalogramy transformaty falkowej dla przykładowych sygnałów: A – sygnału sinusoidalnego, B – sygnału monofraktalnego (szum gaussowski) oraz sygnałów multifraktalnych: C – kaskady dwumianowej, D – sygnału ARFIMA z zadanym wykładnikiem Hursta $H = 0.5$.

4.3.3 (Multi)fraktalna natura języka naturalnego

Oba przedstawione narzędzia analizy multifraktalnej, tj. MF-DFA i WTMM, pozwalają m.in. na detekcję korelacji długozasięgowych [187], które są jednym z głównych źródeł fraktalności i multifraktalności [189, 190]. Ich występowanie w tekstach jest

[11]Procesy ARFIMA (ang. *autoregressive fractionally integrated moving average*) są procesami o grubych ogonach rozkładów fluktuacji i długiej pamięci, wprowadzonymi do modelowania danych z rynków finansowych [186].

niezwykle intrygujące z lingwistycznego punktu widzenia, niniejszy podrozdział będzie zatem poświęcony temu zagadnieniu. Na rysunku 4.40 pokazano jakościowy przebieg długości zdań dla trzech wybranych utworów literackich, charakteryzujących się różną strukturą.

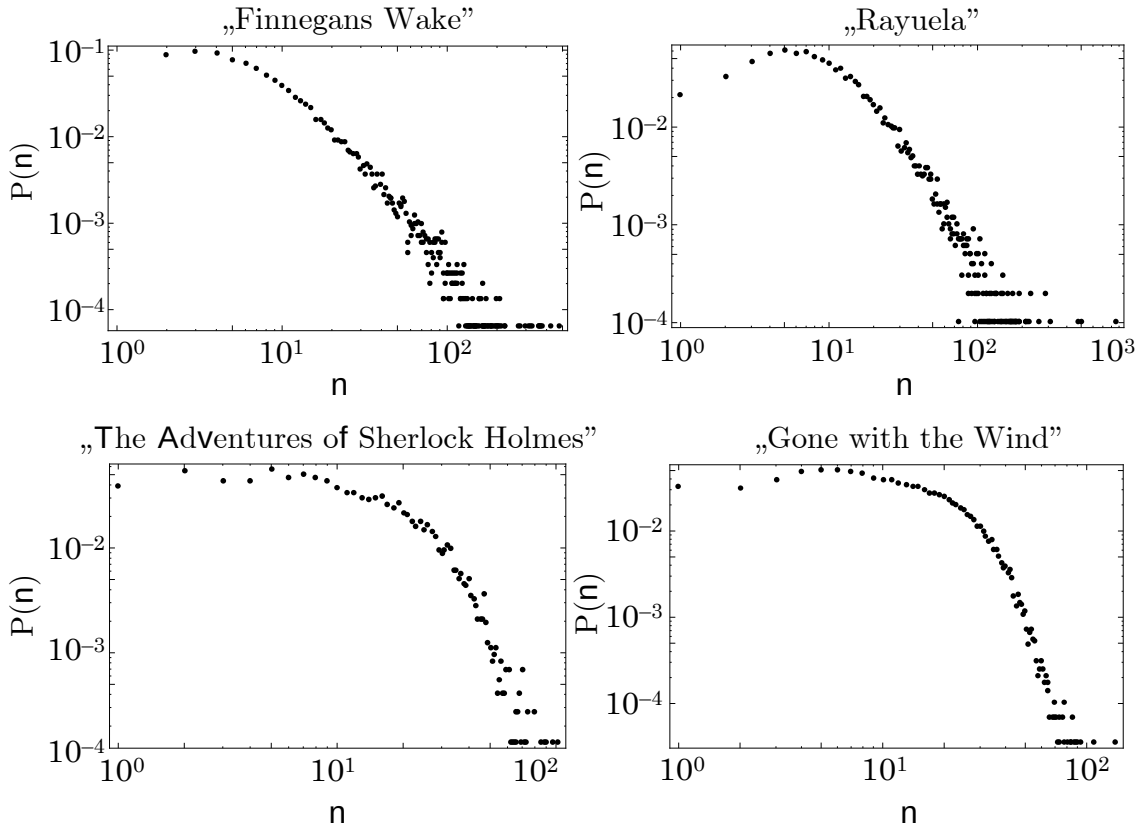


Rysunek 4.40: Wykresy długości zdań dla książek: A – *Gone with the Wind* M. Mitchell, B – *Rayuela* J. Cortázara, C – *Finnegans Wake* J. Joyce'a. Należy zwrócić uwagę na różnice w skalach osi pionowej pomiędzy poszczególnymi książkami.

Na rysunku 4.41 przedstawiono rozkłady prawdopodobieństwa wystąpienia zdania składającego się z n wyrazów dla czterech wybranych książek, napisanych w języku angielskim i hiszpańskim. Wraz ze wzrostem długości zdania maleje prawdopodobieństwo $P(n)$ jego wystąpienia w tekście (zwiększa się nachylenie rozkładów $P(n)$), ale istota tego zaniku jest charakterystyczna dla konkretnych utworów. Analizując powyższe rozkłady, widać znaczną nadreprezentację długich zdań w przypadku książki *Finnegans Wake* J. Joyce'a, czego nie można dostrzec w pozostałych dwóch pozycjach.

Złożoność budowy poszczególnych utworów literackich analizowana będzie za pomocą metody MF-DFA w wersji z detrendującym wielomianem drugiego stopnia $Z_\nu^{(2)}$. Metoda MF-DFA oferuje bowiem wyniki, które są wiarygodne dla szerszej klasy sygnałów niż w przypadku metody WTMM [187]. Transformata falkowa będzie natomiast wykorzystywana głównie w celu ilustracji wieloskalowej struktury

sygnałów w czasie (skalogramy), co nie jest możliwe za pomocą metody MF-DFA. Zastosowanie wielomianów drugiego stopnia wynika z kolei z faktu, że znajdują się one w obszarze stabilności ($1 \leqslant k \leqslant 4$), w którym zmiana ich stopnia nie wpływa zasadniczo na wyniki analizy multifraktalnej.

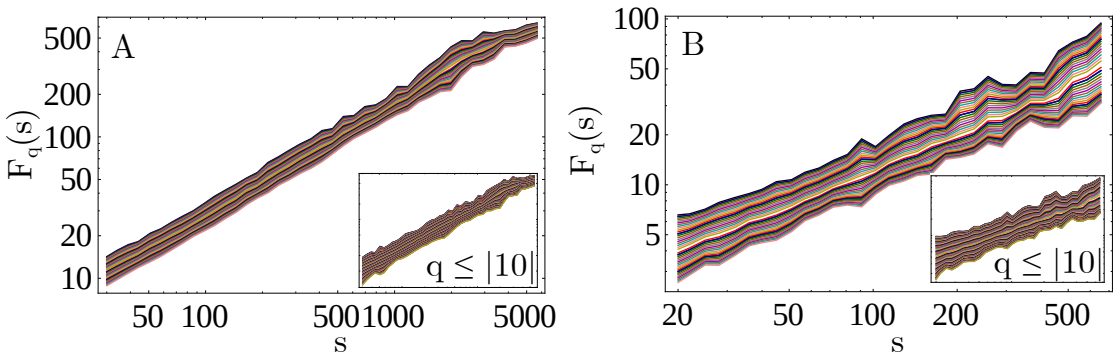


Rysunek 4.41: Rozkłady prawdopodobieństwa $P(n)$ wystąpienia zdania o długości n dla wybranych tekstów literackich.

W związku z brakiem istotnej korelacji wśród rozkładów $P(n)$ dla książek napisanych w tym samym języku, miara ta nie ma charakteru porównawczego pomiędzy nimi, a jedynie dla poszczególnych książek lub autorów. Ze względu na wymagający statystycznie charakter prowadzonych analiz wzięto pod uwagę jedynie te książki, które posiadają odpowiednio dużą liczbę zdań, tak aby długości reprezentujących je szeregów spełniały warunek $T \geqslant 5000$. To spowodowało, że w badaniu z konieczności pominięto wiele utworów literackich. Poddanie ich analizie multifraktalnej byłoby bardzo interesujące z uwagi na ich przemyślaną, nietrywialną strukturę (chodzi tu zwłaszcza o słynne utwory eksperymentalne, takie jak np. *House of Leaves* M. Danielewskiego czy *La Vie mode d'emploi* G. Pereca).

Posługując się metodą MF-DFA, istnienie i klasę fraktalności badanego szeregu czasowego można pierwszoplanowo zidentyfikować za pomocą rodziny funkcji fluktuacji $F_q(s)$, wyliczonych dla różnych wartości parametru q. Ponieważ rozpatrywane szeregi czasowe mają często charakter niegaussowski o wolniejszym zaniku ogonów rozkładów ich wartości niż w przypadku rozkładu normalnego. Zanik ten nie ma jednak typowo charakteru potęgowego, lecz wykładniczy, co zostało zgrubnie pokazane

na rysunku 4.41. Niektóre teksty literackie wykazują jednak quasi-potęgowe zachowanie $P(n)$, dlatego zmienność wykładnika q powinna być ograniczona z obu stron. W tym przypadku przyjęto, że $-4 \leqslant q \leqslant 4$, przy czym w obliczeniach wprowadzono krok $\Delta q = 0.2$.

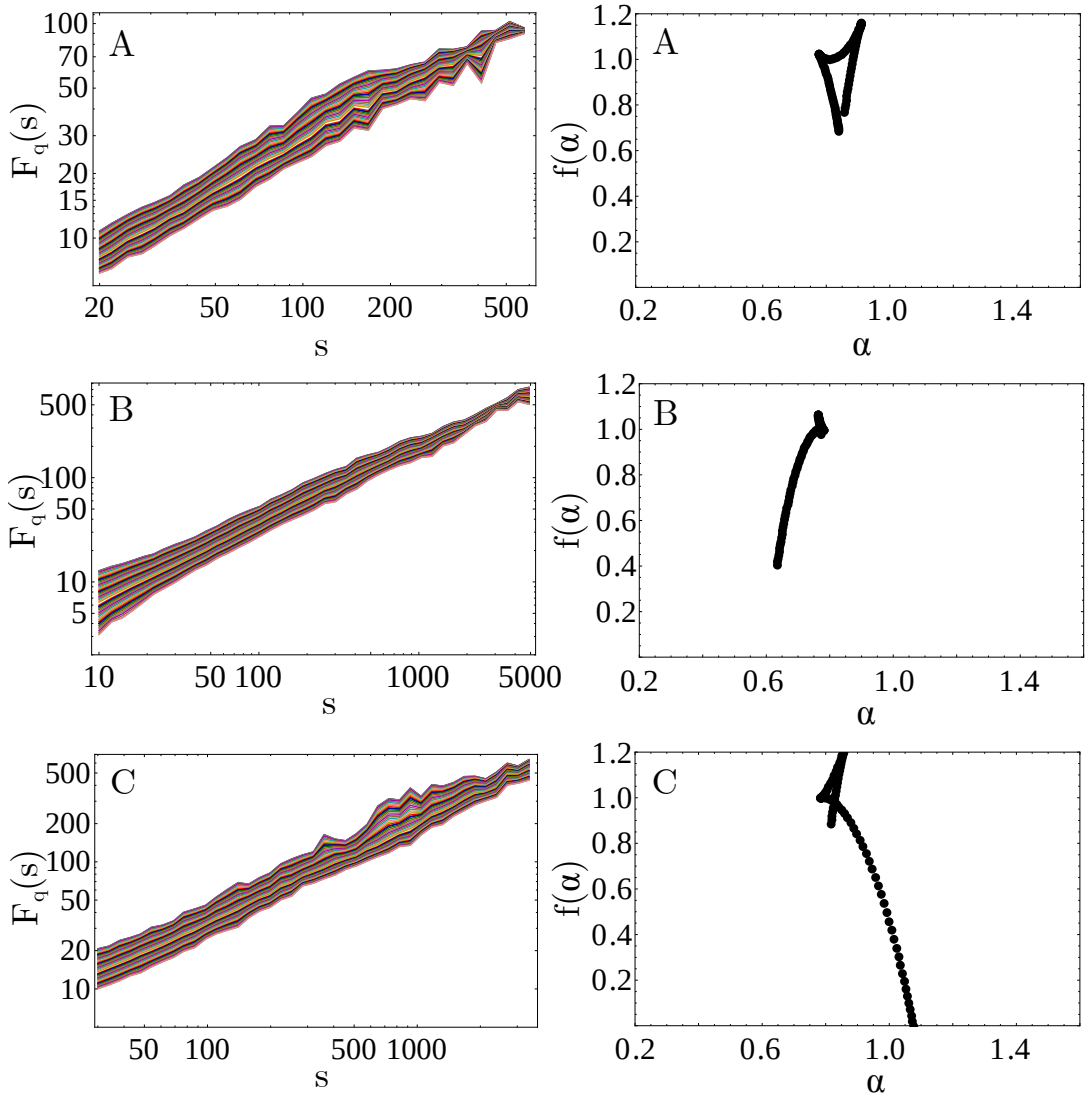


Rysunek 4.42: Funkcje fluktuacji $F_q(s)$ dla monofraktalnych szeregów czasowych, sporządzonych w oparciu o długości zdań w książkach : A – *Gone with the Wind* M. Mitchell oraz B – *Fräulein Else* A. Schnitzlera. Główne wykresy przedstawiają funkcje dla $-4 \leqslant q \leqslant 4$, natomiast wstawki pokazują te same funkcje przy zwiększonym zakresie wartości parametru $-10 \leqslant q \leqslant 10$. Różne zakresy pokazanych skal s i różne szerokości wiązek funkcji $F_q(s)$ w przypadku obu tekstów wynikają z ich odmiennych długości.

Na rysunku 4.42 przedstawiono zachowanie funkcji fluktuacji $F_q(s)$ sporządzonych dla szeregów czasowych odpowiadających dwóm wybranym książkom. Potęgowe zachowanie funkcji fluktuacji jest utrzymane w całym zakresie zmienności skal s i, co ważne, ten charakter jest taki sam pomimo różnych zakresów wartości, jakie przyjmuje parametr q na głównych wykresach i w odpowiednich wstawkach. Świadczy to o stabilności wyników bez względu na to, czy rozpatrywane są małe fluktuacje długości zdań n ($q \ll 0$), czy duże fluktuacje ($q \gg 0$). Takie zachowanie oznacza monofraktalność analizowanych szeregów. Tak jednorodne zachowanie $F_q(s)$ nie jest jednak powszechnie obserwowane w danych empirycznych opartych o długości zdań.

Zgodnie z równaniem (4.55), w przypadku szeregów o własnościach multifraktalnych wykładniki skalowania funkcji fluktuacji są inne dla różnych wartości q, co oznacza, że funkcje $F_q(s)$ nie biegną w formie równoległej wiązki, jak to było na rysunku 4.42. Oczywiście, nie w każdym przypadku identyfikacja skalowania multifraktalnego jest bezdyskusyjna i istnieje wiele sytuacji pośrednich, gdy sygnały trudno jednoznacznie zakwalifikować do mono- lub multifraktali. Podobnie w ramach sygnałów multifraktalnych można wskazać takie o umiarkowanej i takie o silnej charakterystyce multifraktalnej. Różne przykłady tego typu pokazane zostały na rysunkach 4.43, 4.44 i 4.45.

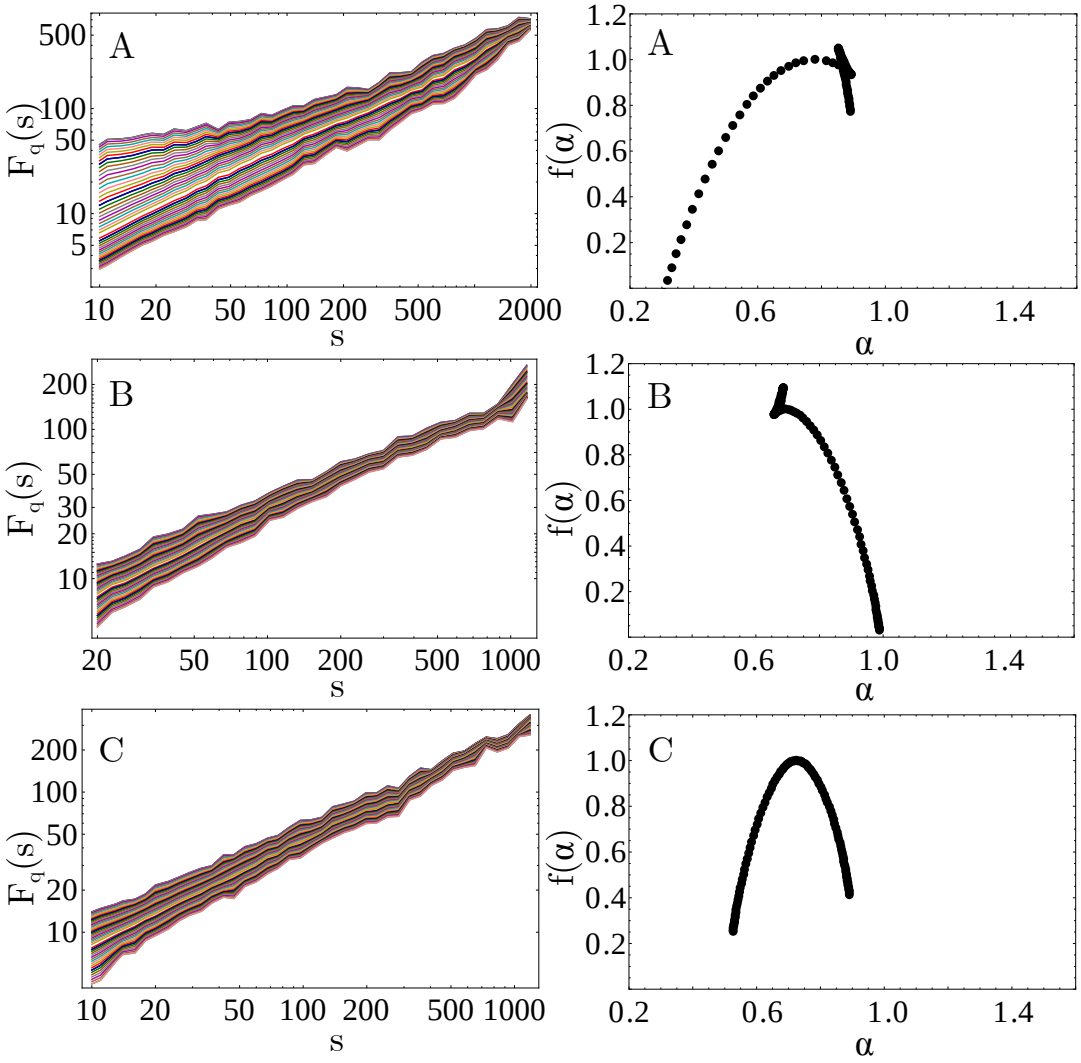
Wykresy na rysunku 4.43 przedstawiają funkcje fluktuacji i widma osobliwości dla szeregów długości zdań dla trzech utworów. Wszystkie charakteryzują się ubogim (wąskim) widmem osobliwości $f(\alpha)$, odpowiadającym strukturze monofraktalnej. Na każdym z wykresów można zaobserwować obszary anomalnego zachowania $f(\alpha)$, gdy jest ono wypukłe; towarzyszy temu również nietypowy charakter funkcji $F_q(s)$, dla których wykładniki skalowania $h(q)$ rosną ze wzrostem q.



Rysunek 4.43: Funkcje fluktuacji $F_q(s)$ (lewa kolumna) i widma osobliwości $f(\alpha)$ (prawa kolumna) sporządzone w oparciu o szeregi czasowe długości zdań dla wybranych tekstów literackich nie wykazujących struktury multifraktalnej (języki oryginalne): A – *A Study in Scarlet* A.C. Doyle'a, B – *Bleak House* C. Dickensa, C – *La Collier da la reine* A. Dumasa. Szerokość widma $f(\alpha)$ odzwierciedla bogactwo struktury multifraktalnej.

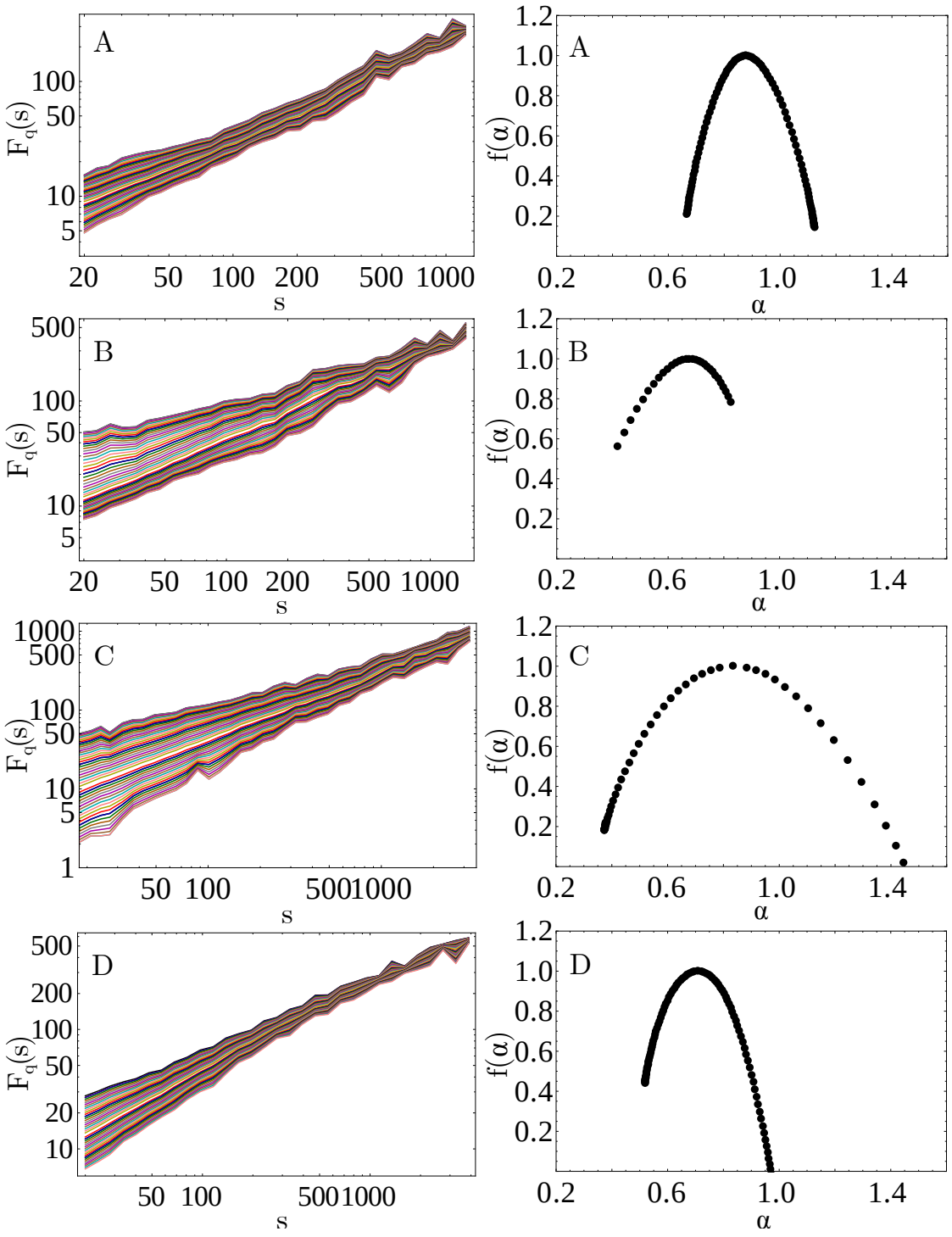
Tego typu efekty są niefizyczne i można je utożsamić z artefaktami procedury numerycznej. Wykresy na rysunku 4.44 posiadają szerszy zakres $f(\alpha)$ niż na rysunku wcześniejszym, co świadczy o większej różnorodności osobliwości spotykanych w sygnałach. Silnie multifraktalne funkcje fluktuacji i widma $f(\alpha)$ zostały przedstawione na rysunku 4.45. Funkcje $F_q(s)$ skalują się tu przez prawie cały zakres zmienności s, co jest dodatkowym argumentem, świadczącym o rzeczywistym efekcie, a nie potencjalnym artefakcie.

W przypadku książek *Sense and Sensibility* (rysunek 4.44), *Ferdydurke*, *Gravity's Rainbow* i *Finnegans Wake* (rysunek 4.45) widma mają pełny, symetryczny kształt zbliżony do paraboli, podobny do widm teoretycznych dla modeli multifraktalnych



Rysunek 4.44: Funkcje fluktuacji $F_q(s)$ (lewa kolumna) i widma osobliwości $f(\alpha)$ (prawa kolumna) sporządzone w oparciu o szeregi czasowe długości zdań dla wybranych tekstów literackich o umiarkowanie silnej multifraktalności (języki oryginalne): A – *Rayuela* J. Cortázara, B – *The Adventures of Tom Sawyer* M. Twaina, C – *Sense and Sensibility* J. Austen. Szerokość widma $f(\alpha)$ odzwierciedla bogactwo struktury multifraktalnej.

procesów stochastycznych [187]. Z drugiej strony bywają także widma o wyraźnej asymetrii, np. dla książek *Rayuela* oraz *The Adventures of Tom Sawyer* (rysunek 4.44). Lewostronne ramię wykresu A (związane z $q > 0$) jest silniej nachylone niż jego prawostronny odpowiednik (związany z $q < 0$). W takich przypadkach dla długich zdań multifraktalność jest wyraźniejsza niż dla krótkich. W przypadku B jest na odwrót. Efekt ten można zrozumieć, biorąc pod uwagę specyficzny charakter analizowanych sygnałów. Najniższe dopuszczalne ich wartości wynoszą $n = 1$, co odpowiada zdaniu składającemu się z jednego słowa, natomiast najwyższe wartości nie są teoretycznie ograniczone. W szczególnych przypadkach, jak np. w różnych eksperymentalnych utworach w rodzaju *Ulissesa* J. Joyce'a czy *Bram raju* J. Andrzejewskiego najdłuższe zdania wypełniają całe rozdziały, a nawet cały utwór.



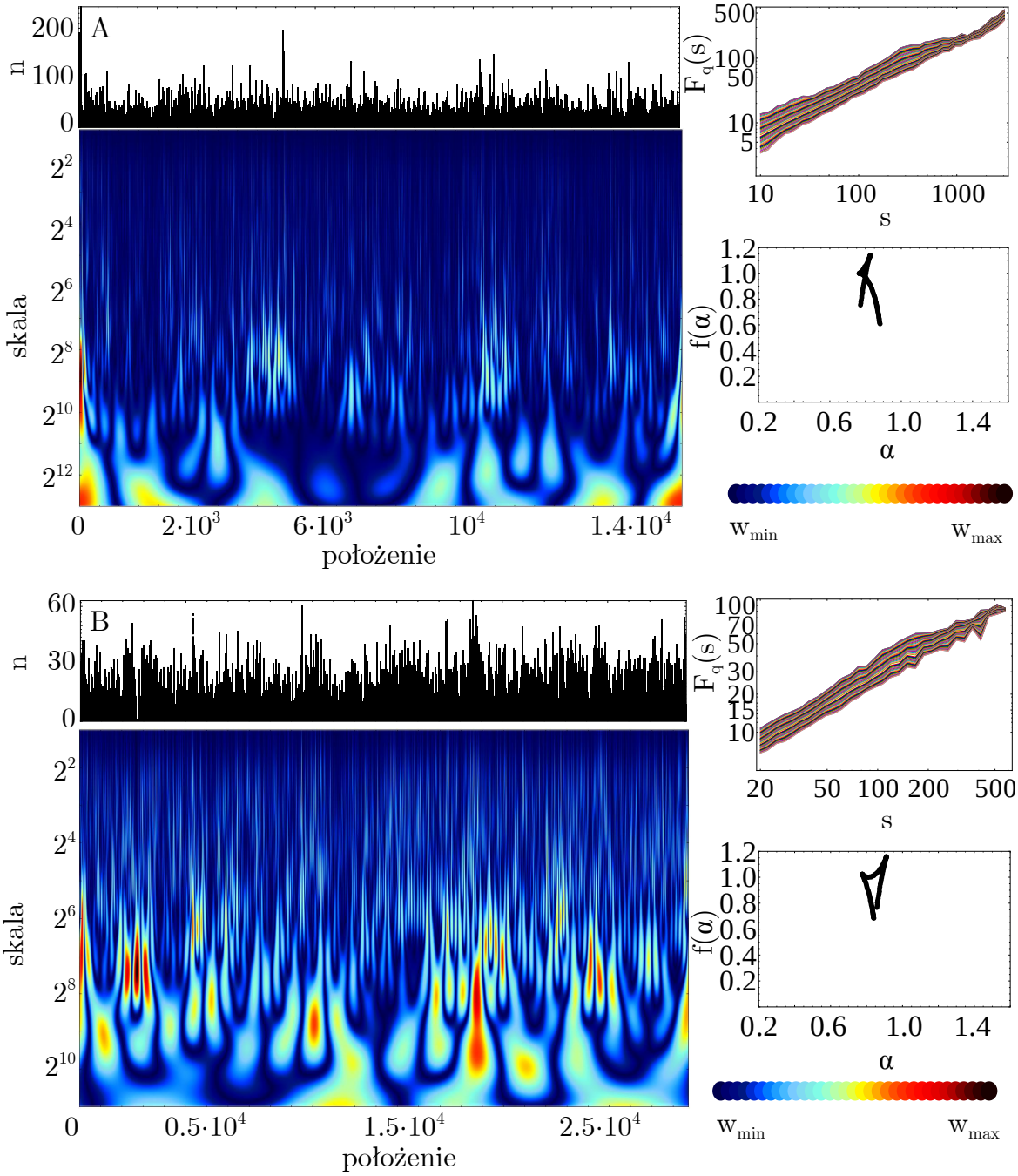
Rysunek 4.45: Po lewej: funkcje fluktuacji $F_q(s)$ uzyskane dla szeregów czasowych długości zdań odpowiadających wybranym tekstom literackim o jednoznacznie multifraktalnym charakterze: A – *Ferdydurke* W. Gombrowicza, B – *Manhattan Transfer* J. Dos Passosa, C – *Finnegans Wake* J. Joyce'a, D – *Gravity's Rainbow* T. Pynchona. Po prawej: odpowiednie wykresy widm osobliwości $f(\alpha)$.

Stąd nawet dane empiryczne mogą zawierać wartości $n \gg 1000$. Ta asymetria pomiędzy dopuszczalnymi wartościami małych i dużych fluktuacji w szeregach czasowych długości zdań odróżnia je od np. niektórych rodzajów danych finansowych, takich jak fluktuacje cen, wartości wskaźników giełdowych czy czasów międzytransakcyjnych. Może to prowadzić do istnienia pewnych asymetrii w skalowaniu funkcji fluktuacji, gdzie dla $q > 0$ zachodzi pożądany proces wzmacniania dużych wartości elementów serii i w konsekwencji także przyczynków do $F_q(s)$, natomiast wybór $q < 0$ powoduje słabszy wzrost odpowiednich przyczynków do funkcji fluktuacji. Tak więc dla różnych wartości q zachowanie funkcji niewiele się różni, przypominając sytuację dla monofraktali. Obie prezentowane na rysunkach miary, tj. funkcja fluktuacji i widmo osobliwości, dają wzajemnie komplementarną informacje o cechach fraktalnych sygnałów. Funkcja fluktuacji wskazuje, czy badany sygnał jest fraktalny i czy związane z fraktalnością skalowanie obejmuje odpowiednio szeroki zakres skal, natomiast widmo $f(\alpha)$ pozwala określić charakter sygnału, czy jest on mono- czy multifraktalny (ewentualnie bifraktalny, jeśli występują tylko dwie istotne wartości wykładnika α [188]).

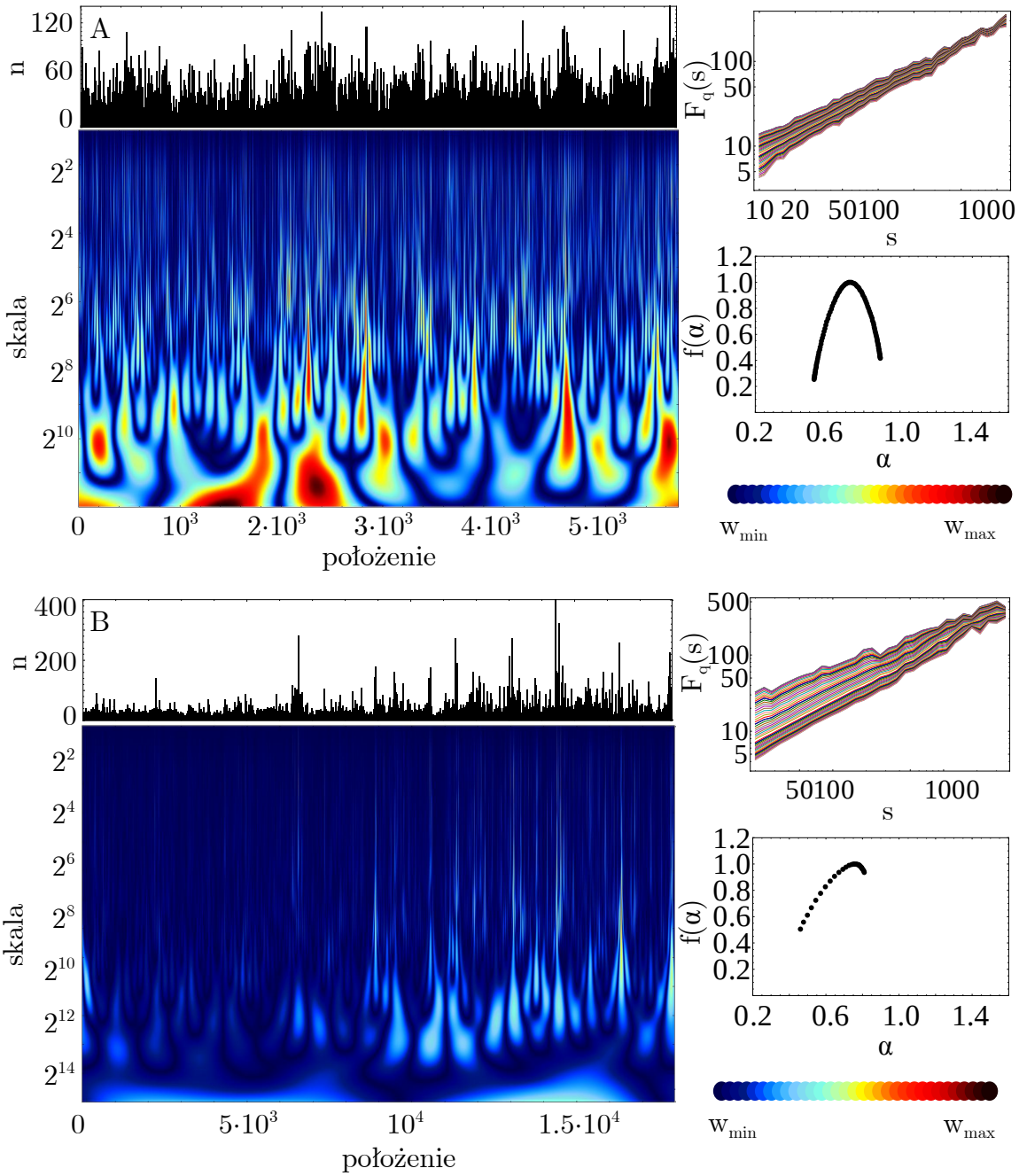
Dodatkową informację, tym razem o lokalnej strukturze sygnału, niemożliwą do otrzymania w metodzie MF-DFA, oferuje transformata falkowa, wykorzystywana w metodzie WTMM. Dzięki obrazowi sygnału utworzonemu w skalogramie tej transformaty można np. zobaczyć, czy własności fraktalne sygnału są stabilne przez całą jego długość, czy też zmieniają się w czasie. Na rysunku 4.46 pokazano skalogramy dla szeregów czasowych o strukturze monofraktalnej. Monofraktalność manifestuje się tu w postaci stosunkowo jednorodnej struktury maksimów i minimów współczynników transformaty falkowej, która nie zmienia się na przestrzeni całego tekstu. Znacznie ciekawsze są skalogramy uzyskane dla szeregów multifraktalnych oraz takich, których własności były niejednoznaczne w formalizmie metody MF-DFA. Teksty multifraktalne posiadają bowiem bardziej zróżnicowane skalogramy, gdzie praktycznie na każdym poziomie skal można zidentyfikować niejednorodność serii, co jest naturalnym rezultatem multifraktalności. Przykłady zostały przedstawione na rysunkach 4.47 i 4.48. We wszystkich przypadkach typ fraktalności jest jakościowo podobny w całych tekstach.

Osobny przypadek stanowią teksty o niejednorodnym charakterze, których jaskrawym przykładem jest przedstawiony na rysunku 4.49 skalogram dla *Ulissesa*. Widać na nim, że w okolicach połowy bardzo silnie zmienia się charakter tekstu, przy czym jest to jakościowo przejście od struktury o słabej multifraktalności do struktury silnie multifraktalnej. Ta niejednorodność pośrednio manifestuje sie też w funkcji fluktuacji przedstawionej w górnym oknie bocznym, która ma zupełnie inne skalowanie dla krótkich skal i dla długich ($s > 10^3$).

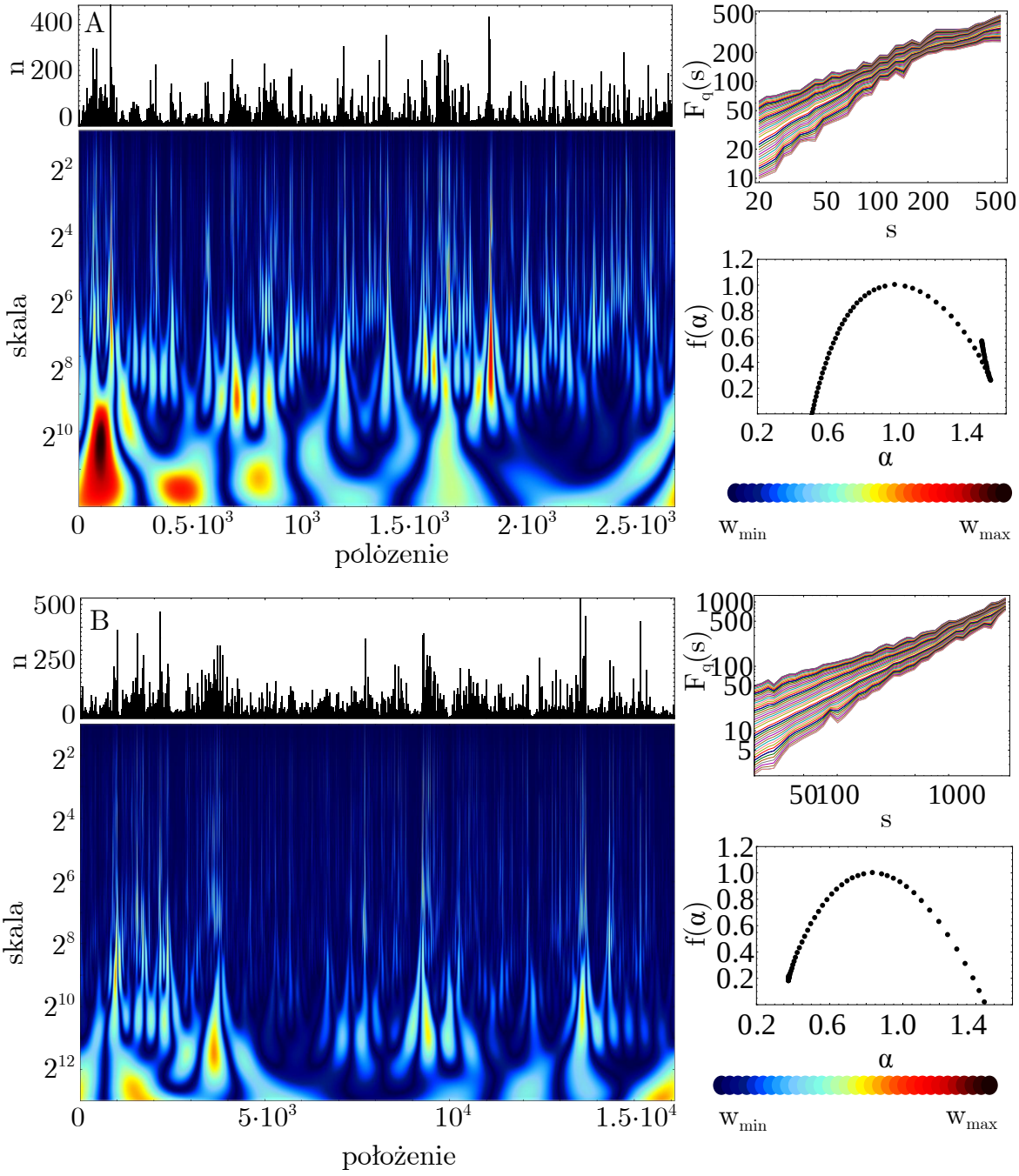
Książka, zawierająca wiele stylów literackich, eksperymentów stylistycznych, szczególnie mających miejsce w jej drugiej połowie. W obrazie skalogramu doskonale widać tę właściwość, gdzie do połowy nie wyróżnia się jakimś szczególnym zachowaniem, natomiast po połowie staje się niezwykle zróżnicowana, zarówno pod kątem różnorodności w skalach, jak i współczynnikach transformaty falkowej. Właściwości multifraktalne tejże serii okazują się być niezwykle osobliwe, jak np. zbiór funkcji fluktuacji $F_q(s)$ (prawa wstawka na rysunku 4.49), gdzie jej zachowanie jest jakościowo inne niż obserwowane dla innych utworów literackich.



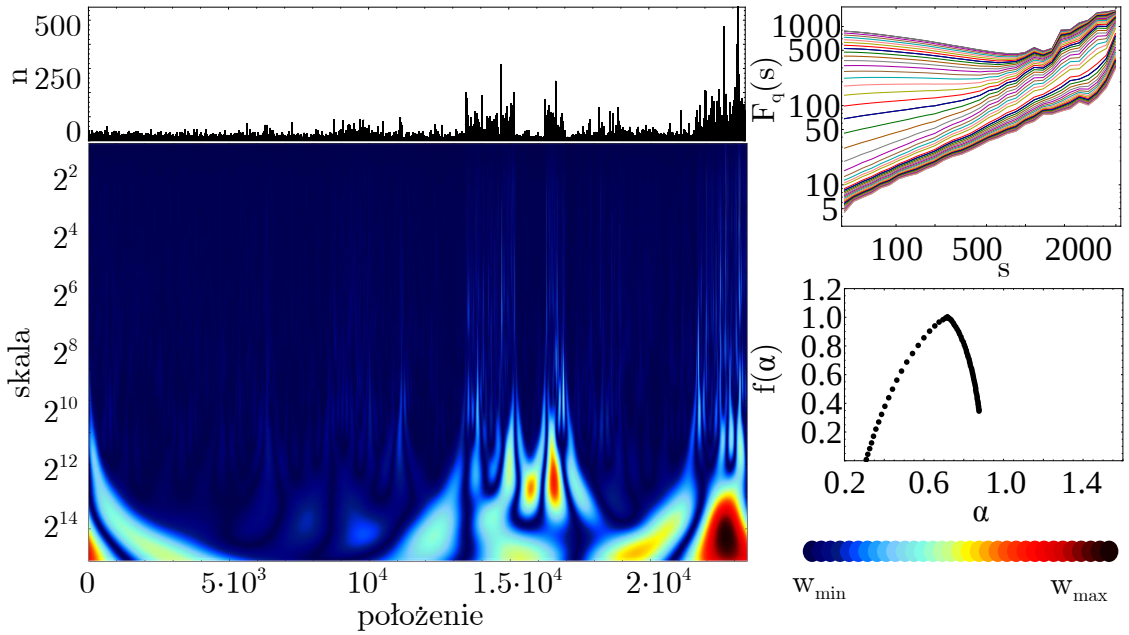
Rysunek 4.46: Po lewej: Skalogramy dla serii monofraktalnych, po prawej: funkcja fluktuacji $F_q(s)$ oraz spektrum osobliwości $f(\alpha)$ dla: A – *La Reine Margot* A. Dumasa, B – *A Study in Scarlet* A.C. Doyle'a.



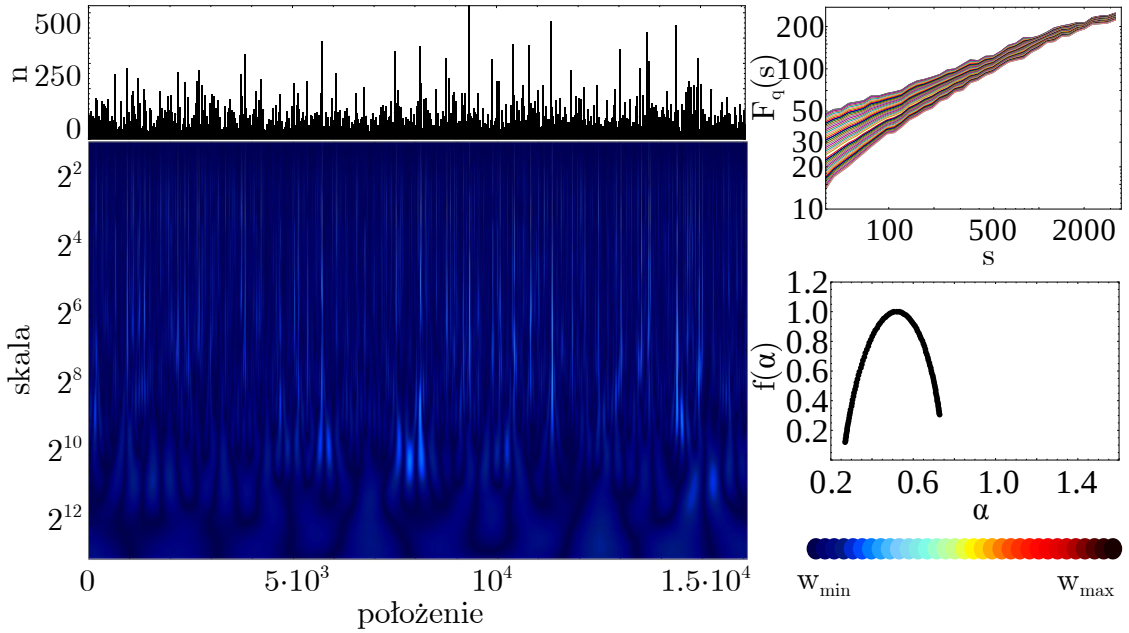
Rysunek 4.47: Po lewej: Skalogramy dla serii multifraktalnych, po prawej: funkcja fluktuacji $F_q(s)$ oraz spektrum osobliwości $f(\alpha)$ dla: A – *Sense and Sensibility* J. Austen, B – *Mort á crédit* L.F. Céline'a.



Rysunek 4.48: Po lewej: Skalogramy dla serii multifraktalnych, po prawej: funkcja fluktuacji $F_q(s)$ oraz spektrum osobliwości $f(\alpha)$ dla: A – *62 Modelo para armar* J. Cortázara, B – *Finnegans Wake* J. Joyce'a.



Rysunek 4.49: Po lewej: Skalogramy dla serii multifraktalnych, po prawej: funkcja fluktuacji $F_q(s)$ oraz spektrum osobliwości $f(\alpha)$ dla książki *Ulysses* J. Joyce'a.



Rysunek 4.50: Po lewej: Skalogramy dla serii multifraktalnych, po prawej: funkcja fluktuacji $F_q(s)$ oraz spektrum osobliwości $f(\alpha)$ dla *Finnegans Wake* J. Joyce'a w wersji przetasowanej.

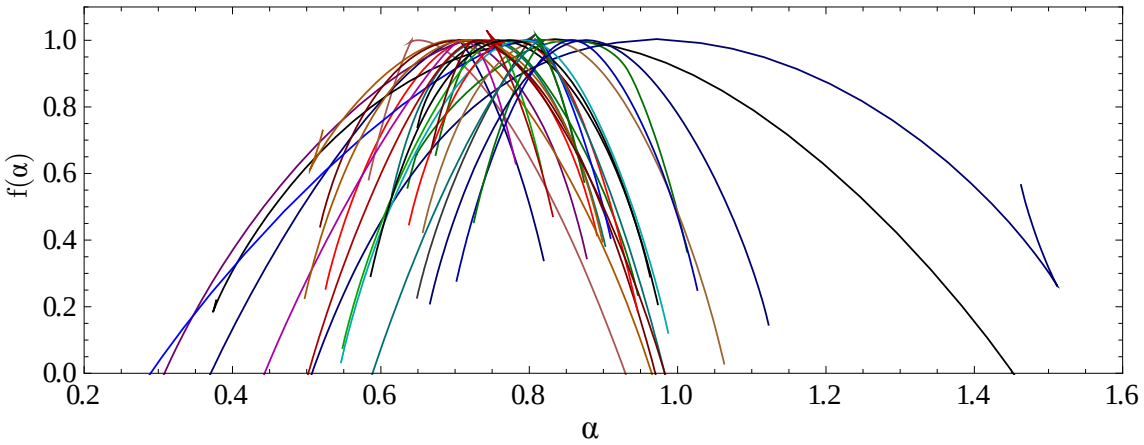
Warto też zauważyć, że źródła multifraktalności analizowanych danych należy upatrywać w długozasięgowych korelacjach o nieliniowym charakterze, a nie w leptokurtycznych rozkładach wartości, które czasem w literaturze podaje się jako możliwe źródło multiskalowania, a które w rzeczywistości prowadzą tylko do efektów związanych ze skończonością próbki [188]. Aby się o tym przekonać, wystarczy porównać skalogramy falkowe, funkcje fluktuacji i widma osobliwości dla oryginalnych danych z danymi wymieszanymi (wymieszanie niszczy korelacje, ale pozostawia nienaruszone rozkłady wartości).

Na rysunku 4.50 zamieszczone są odpowiednie wykresy dla książki *Finnegans Wake*, w której wszystkie zdania zostały losowo wymieszane. Rzut oka na wykres 4.48 pozwala się przekonać, że dane losowe mają zupełnie inne własności, ze szczególnym uwzględnieniem znacznie mniejszego bogactwa osobliwości (węższe widmo $f(\alpha)$).

Traktując szerokość widma osobliwości jako miarę zróżnicowania strukturalnego badanego sygnału, zakres tego zróżnicowania może być łatwo określony ilościowo:

$$\Delta\alpha = \alpha_{\max} - \alpha_{\min} = \alpha(q_{\min}) - \alpha(q_{\max}), \tag{4.66}$$

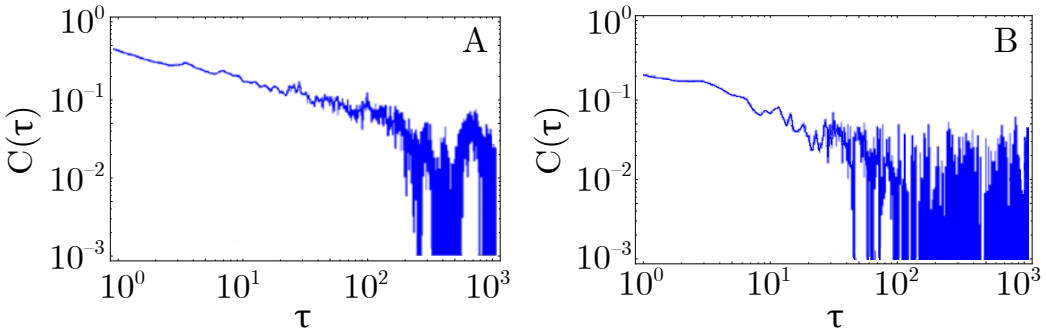
gdzie q ma ustalony zakres zmienności, zależny od typu rozkładu $P(n)$. Ze względu na fakt, że szeregi czasowe długości zdań mają rozkłady leptokurtyczne (rysunek 4.41), w analizie przyjęto zakres $-4 \leqslant q \leqslant 4$. Jako kryterium braku istotnej multifraktalności przyjęto $\Delta\alpha \leqslant 0.1$, natomiast szeregi multifraktalne charakteryzują się $\Delta\alpha \geqslant 0.3$. Zastosowano tu zatem bardziej restrykcyjne kryterium niż w pracach [170, 187].



Rysunek 4.51: Widma osobliwości $f(\alpha)$ dla szeregów czasowych długości zdań, odpowiadających analizowanym tekstom o własnościach multifraktalnych ($\Delta\alpha \geqslant 0.3$).

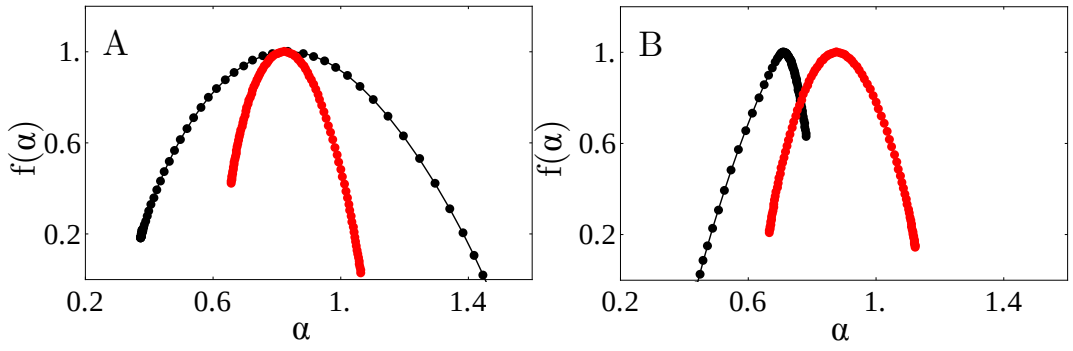
Na rysunku 4.51 zostało zebranych kilkanaście widm, odpowiadających tym utworom, których struktura jest multifraktalna. Cechują się one na ogół dużą szerokością $f(\alpha)$, gdzie $0.3 \leqslant \alpha_{\min} \leqslant 0.7$ oraz $0.8 \leqslant \alpha_{\max} \leqslant 1.1$, natomiast w niektórych przypadkach zachodzi nawet $\alpha_{\max} \approx 1.5$. Liczby te świadczą o dużym bogactwie struktury, zwłaszcza w porównaniu do danych pochodzących z innych układów, które też posiadają naturę multifraktalną: danych finansowych [169], zapisów rytmu serca [172], czy rozkładów galaktyk [173]. Duży zakres zmienności otrzymywanych charakterystyk może być cenną własnością analizowanych danych, pozwalającą na

ilościowe klasyfikowanie tekstów. Znaczna wartość $\Delta\alpha$ dla niektórych tekstów jest zarówno konsekwencją występowania korelacji długozasięgowych w analizowanych szeregach, jak i występowania dużych fluktuacji, praktycznie niemożliwych w przypadku danych opisywanych rozkładami Gaussa. Na rysunku 4.52 przedstawiono zanik funkcji autokorelacji $C(\tau)$, gdzie jej zachowanie może mieć potęgowy charakter, typu $C(\tau) \sim \tau^{-\delta}$ dla τ w zakresie $10^0 - 10^2$, co odpowiada mniej więcej kilku stronom typowego tekstu.



Rysunek 4.52: Autokorelacja $C(\tau)$ dla szeregów czasowych długości zdań: A – *Moby Dick* H. Melville'a, B – *Sense and Sensibility* J. Austen.

Znajomość wykładnika δ pozwala określić wartość wykładnika Hursta, zgodnie ze wzorem: $H = 1 - 0.5\delta$. W przypadku powolnego zaniku $C(\tau)$ wykładnik potęgi przyjmuje niewielką wartość ($\delta \ll 1$), co prowadzi do dużych wartości wykładnika Hursta: $0.5 < H < 1.0$. Zatem ma się tutaj do czynienia z sygnałami o silnej persystencji. Silne autokorelacje w przypadku szeregów długości zdań oznaczają, że całe długie fragmenty takiego tekstu mają podobny styl, który typowo nie zmienia się od zdania do zdania. I rzeczywiście, doświadczenie czytelnicze podpowiada, że obszerne fragmenty tekstu mogą zawierać rozbudowane opisy, które sprzyjają używaniu tylko długich zdań, podobnie rzecz się ma z fragmentami, w których akcja toczy się powoli.



Rysunek 4.53: Przykładowe różnice pomiędzy widmami osobliwości $f(\alpha)$ dla szeregów długości zdań odpowiadających wybranym książkom. A – widma różniące się szerokością $\Delta\alpha$: *Finnegans Wake* J. Joyce'a (czarny) i *Le Trois Mousquetaires* A. Dumasa (czerwony), B – widma różniące się położeniem α: *Alice Adventures in Wonderland* L. Carrolla (czarny) i *Der Zauberberg* T. Manna (czerwony).

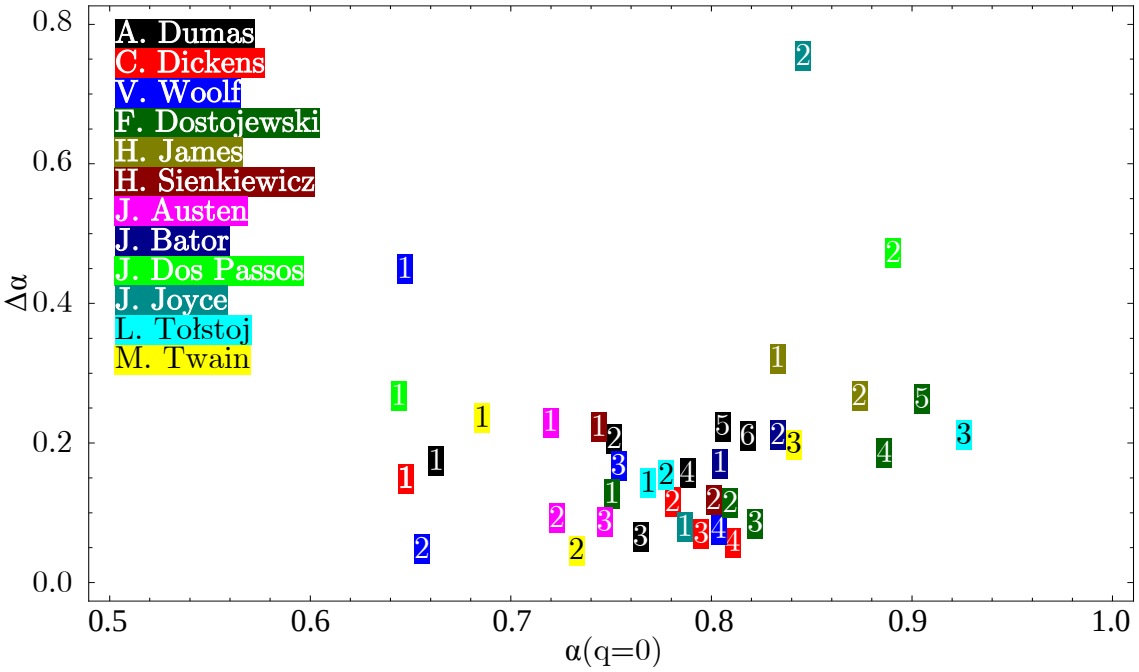
Z kolei w przypadku długich partii dialogowych czy przyspieszonej akcji, autorzy mają tendencję do wybierania raczej krótszych zdań niż dłuższych. Oprócz szerokości, widma $f(\alpha)$ są także charakteryzowane przez przesunięcie ich maksimum względem $\alpha = 0.5$. Jeśli najwyższa wartość występuje dla $\alpha > 0.5$, wiąże się ono z korelacjami długozasięgowymi, które identyfikuje wykładnik Hursta $H \equiv h(2)$. Wynika to ze wzoru (4.56), gdyż $h(q)$ jest funkcją nierosnącą, a więc dla $q = 2$ zachodzi relacja: $\alpha(2) \leqslant h(2)$. Ilustrację różnic pomiędzy widmami osobliwości przedstawia rysunek 4.53. W przypadku A spektra $f(\alpha)$ istotnie różnią się rozpiętością $\Delta\alpha$, natomiast w przypadku B widma są podobnej szerokości i kształtu, ale leżą w innych zakresach α.

Przypisując analizowanemu tekstowi dwa parametry określające widmo $f(\alpha)$, tzn. jego szerokość $\Delta\alpha$ oraz położenie maksimum $\alpha(q = 0)$, własności multifraktalne tekstu można przedstawić graficznie, umieszczając wyznaczone wartości na dwuwymiarowej mapie $(\alpha(0), \Delta\alpha)$. Na rysunku 4.54 przedstawiono mapę dla tekstów literackich sporządzonych przez kilku autorów, którzy napisali dwa lub więcej utworów wziętych pod uwagę w tej pracy. Natomiast na rysunku 4.55 przedstawiono mapę dla kilkudziesięciu utworów literackich, napisanych w 6 różnych językach europejskich. Pomimo, że nie widać systematycznych różnic pomiędzy językami, widoczna jest silna dyspersja położenia poszczególnych utworów. W większości dominują książki, dla których szerokość $f(\alpha)$ przyjmuje wartości $\Delta\alpha \leqslant 0.3$, a jego maksimum zawiera się w przedziale $0.7 \leqslant \alpha \leqslant 0.9$. W tak określonych granicach zmienności parametrów znajduje się 4/5 przebadanych pozycji (szare tło), natomiast pozostałą 1/5 z nich można podzielić na dwie zasadnicze grupy. Ta grupa tekstów, która znajduje się na rysunku pod przerywaną linią, nie spełnia warunków multifraktalności w kontekście przyjętego kryterium, jednak odpowiadające jej widma osobliwości są stosunkowo słabo przesunięte ($0.5 < \alpha < 0.7$), czyli nie wykazują silnej persystencji. Książki z tej grupy w badanej reprezentacji stosunkowo najmniej różnią się od szumu. Pozostałe 9 książek, dla których wartości $\Delta\alpha$ leżą nad przerywaną linią, posiadają widma bez wątpienia multifraktalne. W większości przypadków charakteryzują się one silną persystencją, będącą konsekwencją wyjątkowo skomplikowanej budowy i silnej zmienności długości zdań tych pozycji.

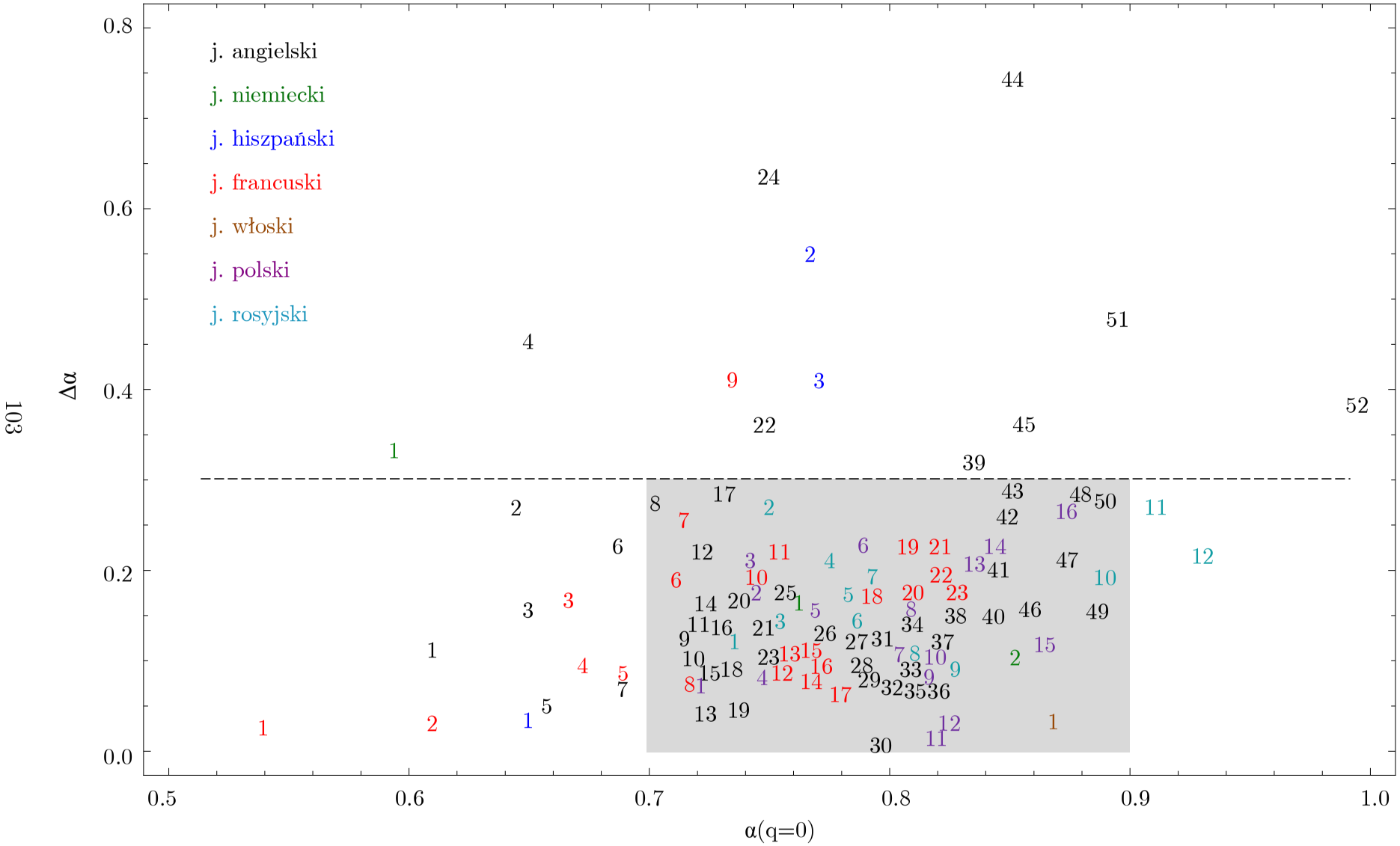
Odnosząc się do położenia przedstawionych utworów literackich, widać silną korelację pomiędzy bogactwem typów osobliwości występujących w sygnałach a literackim charakterem badanego dzieła. Teksty posiadające symetryczne i szerokie widma $f(\alpha)$ są opisywane przez krytykę literacką jako dzieła eksperymentalne, wykorzystujące nowatorskie techniki narracyjne [180, 191]. Znaczna część analizowanych pozycji literatury wykazała strukturę multifraktalną, jednak o różnym stopniu heterogeniczności w świetle szerokości widma $f(\alpha)$.

Istotnym z punktu prowadzonej analizy staje się pytanie o źródło zaobserwowanych sygnałów multifraktalnych. Odpowiedź na to pytanie nie jest jednoznaczna ani oczywista. Utwory literackie powstają w wyniku skomplikowanych reakcji świadomości (umysłu) autora i środowiska, w jakim się on znajduje w momencie tworzenia tekstu. Znane są przypadki, w których sam tekst powstawał w wyniku silnych emocji, doznań erotycznych, bądź wpływu substancji psychotropowych jak alkohol czy narkotyki.

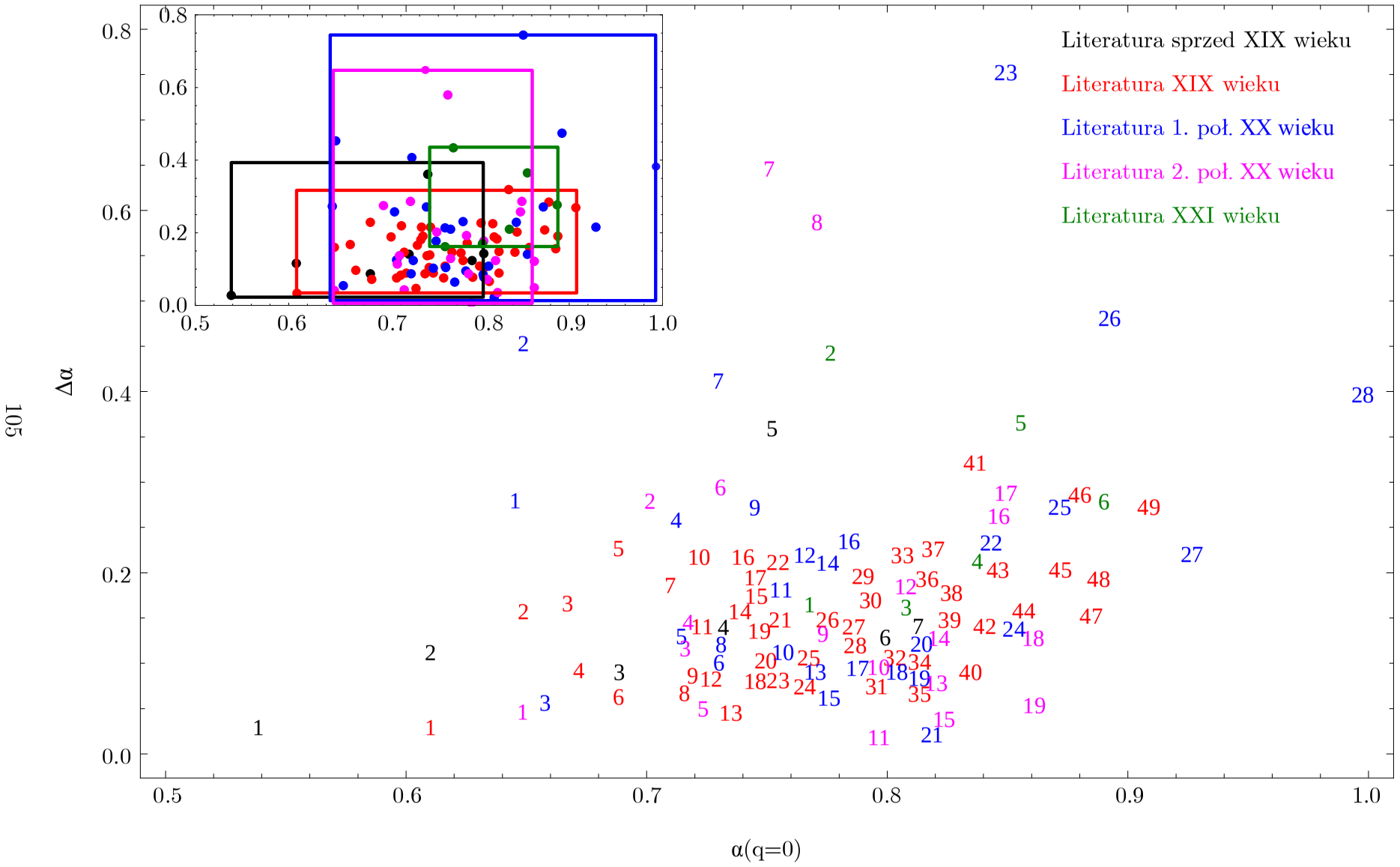
Tym niemniej głównym przyczynkiem do możliwości zaistnienia wykazanych efektów są nieliniowe procesy zachodzące w mózgu autora tekstu, będące naturalnymi oddziaływaniami pomiędzy jego elementami: neuronami, ośrodkami mózgowymi itd. Dogłębna analiza tych efektów jest możliwa jedynie wówczas, gdy praca samego mózgu będzie dobrze i skrupulatnie opisana, co przy obecnym stanie rzeczy jest niemożliwe.



Rysunek 4.54: Mapa przedstawiająca wartości $\Delta\alpha$ oraz $\alpha(q=0)$ dla tekstów literackich napisanych przez różnych autorów. Każda liczba odpowiada konkretnemu utworowi. **A. Dumas**: 1 – *Vingt ans après*, 2 – *Le Comte de Monte Christo*, 3 – *La Reine Margot*, 4 – *La Collier de la reine*, 5 –*Le Victome de Bragelonne* , 6 – *Les Trois Mousquetaires*; **C. Dickens**: 1 – *David Copperfield*, 2 – *Great Expectation*, 3 – *Bleak House*, 4 – *A Tale of Two Cities*; **V. Woolf**: 1 – *The Waves*, 2 – *The Years*; **F. Dostojewski**: 1 – *Дневник писателя*, 2 – *Преступление и наказание*, 3 – *Братья Карамазовы*, 4 – *Идиот*, 5 – *Бесы*; **H. James**: 1 – *The Portrait of a Lady*, 2 – *What Maisie Knew*; **H. Sienkiewicz**: 1 – *Trylogia*, 2 – *Quo vadis*; **J. Austen**: 1 – *Sense and Sensibility*, 2 – *Emma*, 3 – *Pride and Prejudice*; **J. Bator**: 1 – *Ciemno, prawie noc*, 2 – *Piaskowa góra*; **J. Dos Passos**: 1 – *Manhattan Transfer*, 2 – *U.S.A trilogy*; **J. Joyce**: 1 – *Ulysses*, 2 – *Finnegans Wake*; **L. Tołstoj**: 1 – *Анна Каренина*, 2 – *Война и миръ*, 3 – *Воскресение*; **M. Twain**: 1 – *The Adventures of Tom Sawyer*, 2 – *A Connecticut Yankee in King Arthur's Court*, 3 – *Life on the Mississippi*.


j. angielski
j. niemiecki
j. hiszpański
j. francuski
j. włoski
j. polski
j. rosyjski
Δα
10³
α(q=0)
0.0
0.2
0.4
0.6
0.8
0.5
0.6
0.7
0.8
0.9
1.0

Rysunek 4.55: Mapa przedstawiająca wartości $\Delta\alpha$ oraz $\alpha(q = 0)$ dla tekstów literackich napisanych w jednym z 6 języków europejskich. Pozioma przerywana linia oddziela zakres multifraktalny ($\Delta\alpha \geqslant 0.3$) od zakresu stricte monofraktalnego ($\Delta\alpha < 0.1$) i pośredniego ($0.1 \leqslant \Delta\alpha < 0.3$). Każda liczba odpowiada konkretnemu utworowi. **Jęz. angielski**: 1 – *Memoirs of a Woman of Pleasure*; 2 – *Manhattan Transfer*; 3 – *David Copperfield*; 4 – *The Waves*; 5 – *The Years*; 6 – *The Adventures of Tom Sawyer*; 7 – *Dracula*; 8 – *Gravity's Rainbow*; 9 – *The Lord of the Rings*; 10 – *Trainspotting*; 11 – *The Butcher Boy*; 12 – *Sense and Sensibility*; 13 – *The Illuminatus! Trilogy*; 14 – *Oliver Twist*; 15 – *Emma*; 16 – *Leviathan*; 17 – *One Flew Over the Cuckoo's Nest*; 18 – *The Secret Adversary*; 19 – *A Connecticut Yankee in King Arthur's Court*; 20 – *The Adventures of Sherlock Holmes*; 21 – *Moby Dick*; 22 – *The Life and Opinions of Tristram Shandy*; 23 – *Pride and Prejudice*; 24 – *A Heartbreaking Work of Staggering Genius*; 25 – *The Voyage Out*; 26 – *The Vampire Chronicles*; 27 – *Great Expectations*; 28 – *Dubliners*; 29 – *The Stand: The Complete and Uncut Edition*; 30 – *Atlas Shrugged*; 31 – *Clarissa, or, the History of a Young Lady*; 32 – *Bleak House*; 33 – *Night and Day*; 34 – *Robinson Crusoe*; 35 – *Gone with the Wind*; 36 – *A Tale of Two Cities*; 37 – *Suttree*; 38 – *The Picture of Dorian Gray*; 39 – *The Portrait of a Lady*; 40 – *The Bostonians*; 41 – *Life on the Mississippi*; 42 – *Catch-22*; 43 – *Infinite Jest*; 44 – *Finnegans Wake*; 45 – *The Goldfinch*; 46 – *Middlemarch: A Study of Provincial Life*; 47 – *Beloved*; 48 – *What Maisie Knew*; 49 – *The Jungle*; 50 – *The Luminaries*; 51 – *U.S.A. trilogy*; 52 – *The Ambassadors*. **Jęz. niemiecki**: 1 – *Berlin Alexanderplatz*; 2 – *Der Zauberberg*. **Jęz. hiszpański**: 1 – *Cien años de soledad*; 2 – *Rayuela*; 3 – *2666*. **Jęz. francuski**: 1 – *Artamène ou le Grand Cyrus*; 2 – *Le Rouge et le noir*; 3 – *Vingt ans après*; 4 – *Le Petit Chose*; 5 – *Les Liaisons dangereuses*; 6 – *Les Rougon-Macquart*; 7 – *La Condition humaine*; 8 – *Madame Bovary*; 9 – *Mort à credit*; 10 – *Bel-Ami*; 11 – *Le Comte de Monte-Cristo*; 12 – *Aventures extraordinaires d'un savant russe*; 13 – *À la recherche du temps perdu*; 14 – *La reine Margot*; 15 – *La Comédie humaine*; 16 – *Le roman de Tristan et Iseut*; 17 – *Voyage au bout de la nuit*; 18 – *Le Collier de la reine*; 19 – *Le Vicomte de Bragelonne ou Dix ans plus tard*; 20 – *Molloy* + *Malone meurt* + *L'Innommable* (łącznie); 21 – *Les Trois Mousquetaires*; 22 – *Les Misérables*; 23 – *Les Mystères de Paris*. **Jęz. włoski**: 1 – *Il nome della rosa*. **Jęz. polski**: 1 – *Ziemia obiecana*; 2 – *Nad Niemnem*; 3 – *Trylogia*; 4 – *Lalka*; 5 – *Ostatnie rozdanie*; 6 – *Chłopi*; 7 – *Quo vadis: Powieść z czasów Nerona*; 8 – *Ciemno, prawie noc*; 9 – *Trylogia kosmiczna*; 10 – *Popioły*; 11 – *Noce i dnie*; 12 – *Widnokrąg*; 13 – *Piaskowa Góra*; 14 – *Trylogia księżycowa*; 15 – *Tysiąc spokojnych miast* + *Pod mocnym aniołem* + *Miasto utrapienia* (łącznie); 16 – *Ferdydurke*. **Jęz. rosyjski**: 1 – *Архипелаг ГУЛАГ*; 2 – *Тихий Дон*; 3 – *Дневник писателя*; 4 – *Петербургъ*; 5 – *Анна Каренина*; 6 – *Война и миръ*; 7 – *Мёртвые души*; 8 – *Преступление и наказание*; 9 – *Братья Карамазовы*; 10 – *Идиот*; 11 – *Бесы*; 12 – *Воскресение*.


Literatura sprzed XIX wieku
Literatura XIX wieku
Literatura 1. poł. XX wieku
Literatura 2. poł. XX wieku
Literatura XXI wieku
Δα
105
α(q=0)
0.0
0.2
0.4
0.6
0.8
0.5
0.6
0.7
0.8
0.9
1.0

Rysunek 4.56: Mapa przedstawiająca wartości $\Delta\alpha$ oraz $\alpha(q = 0)$ dla tekstów literackich napisanych w różnych epokach. Każda liczba odpowiada konkretnemu utworowi. Wykres we wstawce wskazuje na zakres zmienności prezentowanych parametrów w obrębie jednej epoki. **Literatura sprzed XIX wieku:** 1 – *Artamène ou le Grand Cyrus*; 2 – *Memoirs of a Woman of Pleasure*; 3 – *Les Liaisons dangereuses*; 4 – *Leviathan*; 5 – *The Life and Opinions of Tristram Shandy*; 6 – *Clarissa, or, the History of a Young Lady*; 7 – *Robinson Crusoe*. **Literatura XIX wieku:** 1 – *Le Rouge et le Noir*; 2 – *David Copperfield*; 3 – *Vingt ans après*; 4 – *Le Petit Chose*; 5 – *The Adventures of Tom Sawyer*; 6 – *Dracula*; 7 – *Les Rougon-Macquart*; 8 – *Madame Bovary*; 9 – *Ziemia obiecana*; 10 – *Sense and Sensibility*; 11 – *Oliver Twist*; 12 – *Emma*; 13 – *A Connecticut Yankee in King Arthur's Court*; 14 – *The Adventures of Sherlock Holmes*; 15 – *Nad Niemnem*; 16 – *Trylogia*; 17 – *Bel-Ami*; 18 – *Lalka*; 19 – *Moby Dick*; 20 – *Pride and Prejudice*; 21 – *Дневник писателя*; 22 – *Le Comte de Monte-Cristo*; 23 – *Aventures extraordinaires d'un savant russe*; 24 – *La Reine Margot*; 25 – *La Comédie humaine*; 26 – *Анна Каренина*; 27 – *Война и миръ*; 28 – *Great Expectations*; 29 – *Мёртвые души*; 30 – *Le Collier de la reine*; 31 – *Bleak House*; 32 – *Quo vadis: Powieść z czasów Nerona*; 33 – *Le Vicomte de Bragelonne ou Dix ans plus tard*; 34 – *Преступление и наказание*; 35 – *A Tale of Two Cities*; 36 – *Les Trois Mousquetaires*; 37 – *Les Misérables*; 38 – *Les Mystères de Paris*; 39 – *Братья Карамазовы*; 40 – *The Picture of Dorian Gray*; 41 – *The Portrait of a Lady*; 42 – *The Bostonians*; 43 – *Life on the Mississippi*; 44 – *Middlemarch: A Study of Provincial Life*; 45 – *Beloved*; 46 – *What Maisie Knew*; 47 – *The Jungle*; 48 – *Идиот*; 49 – *Бесы*. **Literatura 1. poł. XX wieku:** 1 – *Manhattan Transfer*; 2 – *The Waves*; 3 – *The Years*; 4 – *La Condition humaine*; 5 – *The Lord of the Rings*; 6 – *The Secret Adversary*; 7 – *Mort à crédit*; 8 – *Архипелаг ГУЛАГ*; 9 – *Тихий Дон*; 10 – *À la recherche du temps perdu*; 11 – *The Voyage Out*; 12 – *Berlin Alexanderplatz*; 13 – *Le Roman de Tristan et Iseut* ; 14 – *Петербургъ*; 15 – *Voyage au bout de la nuit*; 16 – *Chłopi*; 17 – *Dubliners*; 18 – *Night and Day*; 19 – *Gone with the Wind*; 20 – *Popioły*; 21 – *Noce i dnie*; 22 – *Trylogia księżycowa*; 23 – *Finnegans Wake*; 24 – *Der Zauberberg*; 25 – *Ferdydurke*; 26 – *U.S.A. trilogy*; 27 – *Воскресение*; 28 – *The Ambassadors*. **Literatura 2. poł. XX wieku:** 1 – *Cien años de soledad*; 2 – *Gravity's Rainbow*; 3 – *Trainspotting*; 4 – *The Butcher Boy* ; 5 – *The Illuminatus! Trilogy*; 6 – *One Flew Over the Cuckoo's Nest*; 7 – *A Heartbreaking Work of Staggering Genius*; 8 – *Rayuela*; 9 – *The Vampire Chronicles*; 10 – *The Stand: The Complete and Uncut Edition*; 11 – *Atlas Shrugged*; 12 – *Molloy* + *Malone Meurt* + *L'Innommable* (łącznie); 13 – *Trylogia kosmiczna*; 14 – *Suttree*; 15 – *Widnokrąg*; 16 – *Catch-22*; 17 – *Infinite Jest*; 18 – *Tysiąc spokojnych miast* + *Pod mocnym aniołem* + *Miasto utrapienia* (łącznie); 19 – *Il nome della rosa*. **Literatura XXI wieku:** 1 – *Ostatnie rozdanie*; 2 – *2666*; 3 – *Ciemno, prawie noc*; 4 – *Piaskowa góra*; 5 – *The Goldfinch*; 6 – *The Luminaries*.

Rozdział 5

Analiza rezultatów pracy

Przedmiotem przeprowadzonych badań był język naturalny, a dokładniej jego zapisana forma – teksty literackie. Wybór takiej reprezentacji języka był zdeterminowany przez kilka czynników. Proza, jako popularny rodzaj literatury, jest najbardziej naturalną formą „zmaterializowania" języka naturalnego w formie próbki, mogącej być przedmiotem ilościowych analiz. Ponadto istnieje stosunkowo łatwy dostęp do tego typu reprezentacji języka oraz wygodny i efektywny sposób jej przetwarzania. Analiza języka mówionego, pod pewnymi względami bliższego właściwemu językowi naturalnemu, wymagałaby znacznie wyższego nakładu pracy związanej z przetransformowaniem sygnału akustycznego (mowa) w sygnał, który mógłby być poddany ilościowej analizie (pismo). Analiza mowy byłaby ciekawa ze względu na jej spontaniczny charakter (pomijając oczywiście wypowiedzi wcześniej przygotowane, pokrewne językowi pisanemu), a przez to ścisły związek z aktualnymi procesami zachodzącymi w mózgu mówiącego. Forma pisana języka jest pod tym względem bardziej finalnym produktem długich procesów myślowych i redakcji niż zapisem chwilowego stanu umysłu (choć i to może się zdarzyć). Wychodząc jednak z założenia, że forma mówiona i pisana jest określona tymi samymi regułami (gramatyką) oraz realizuje się za pomocą tych samych słów, ze względów praktycznych całą analizę oparto tylko o teksty pisane.

Interdyscyplinarne podejście do badania układu, jakim jest język naturalny, pozwoliło na użycie wielu dobrze sprawdzonych w praktyce narzędzi analizy danych empirycznych, prowadząc do ujawnienia w języku licznych właściwości wspólnych dla układów złożonych. Mimo że sama materia języka posiada specyficzny charakter, to znalezienie odpowiedniej reprezentacji języka i przetransformowanie jego próbek na szeregi czasowe umożliwia analizę ilościową i, co za tym idzie, na obiektywne scharakteryzowanie jego struktury i dynamiki.

Przeprowadzone analizy statystyczne jednoznacznie wskazały na uniwersalizm potęgowego rozkładu występowania słów w danym korpusie języka. Jednakowy, zipfowski charakter każdego z analizowanych języków, świadczą nie tylko o oparciu struktury opisywanych języków o ten sam mechanizm generatywny, ale również wskazują na subtelny charakter statystyki występujących w nim elementów, charakterystyczny dla układów złożonych. Należy zaznaczyć, że analiza rozkładu częstotliwości słów została wykonana na korpusie opartym o teksty, które posiadają słownictwo charakterystyczne dla epoki, w której zostały stworzone, indywidualnego

stylu autora oraz poruszanej tematyki. W związku z tym należy wziąć pod uwagę, że otrzymane rozkłady Zipfa określają charakterystykę danego korpusu słownikowego, a nie języka jako takiego. Z drugiej zaś strony, rozmiar przyjętej statystyki (ok. 10^6 słów, uzyskany po złożeniu ~ 60 reprezentatywnych książek) jest w miarę bliskim obrazem danego języka, uśredniającym w sobie indywidualne zachowanie poszczególnych książek.

Opis języka, uwzględniający nie tylko częstość występowania różnych słów, ale również związek jaki istnieje pomiędzy nimi, był możliwy dzięki zastosowaniu analizy sieciowej. Modelowanie języka naturalnego w obrazie sieci było oparte tylko o jedną reprezentację, związaną z sąsiedztwem słów w tekście. Przyjęta konstrukcja sieci zakłada a priori, że słowa stojące obok siebie oddziałują ze sobą, co nie jest oczywiście zawsze prawdziwe. Bez trudu można wskazać sytuacje, że dwa słowa stojące obok siebie nie posiadają istotnego związku ze sobą, jak np. para składająca się ze słowa zamykającego zdanie i słowa otwierającego zdanie następne; podobnie bywa czasem w przypadku sąsiednich słów, wchodzących w skład dwóch różnych zdań składowych zdania złożonego. Efekty te nie są jednak na tyle istotne statystycznie, by w znaczący sposób wpływać na strukturę otrzymywanych sieci. Ponadto wykonano analizę, uwzględniającą oddziaływanie dalszych (niesąsiednich) słów ze sobą, nie stwierdzając żadnych istotnych zmian jakościowych w otrzymywanych sieciach sąsiedztwa słów.

Wskazanie odpowiedniego mechanizmu stojącego za dynamiką sieci sąsiedztwa słów pozwoliło nie tylko na lepsze zrozumienie ich struktury, ale również na wyjaśnienie osobliwego zachowania niektórych miar opisujących topologie tych sieci. Modele sieci z przyspieszonym wzrostem wydają się być z tego punktu widzenia niezwykle atrakcyjne i mieć zastosowanie do opisu szerokiej klasy układów ewoluujących przez przyłączanie nowych wierzchołków, jednak – aby uzyskać zgodność z danymi empirycznymi – wymagają wprowadzenia modyfikacji do generycznych swych wersji. Przykładem jest tutaj średnia długość najkrótszej ścieżki, która w przypadku sieci sąsiedztwa słów – zamiast rosnąć ze wzrostem sieci, podobnie, jak się to obserwuje w generycznym modelu sieci rosnących (DM-AG), ma – po krótkim, przejściowym okresie – tendencję do monotonicznego zmniejszania swej wartości. Brak pełnej zgodności wyników symulacji z wynikami analizy empirycznej wskazuje, że wymagane modyfikacje modelu prawdopodobnie powinny też odzwierciedlać subtelne własności języka, takie jak ograniczenia powtarzalności słów narzucone przez reguły gramatyki czy przyjętego stylu narracji. Nieco bliższa topologii języka niż modele sieci o przyspieszonym wzroście wydaje się być topologia sieci utworzonych przez przetworzenie na reprezentacje sieciowe procesów stochastycznych naśladujących tworzenie tekstów literackich.

Rezultaty uzyskane na drodze analizy sieciowej stały się kolejnym argumentem przemawiającym za istnieniem złożoności w strukturze języka naturalnego. Potęgowe zachowanie się rozkładów stopni wierzchołków, jak również potęgowy charakter pośrednictwa i współczynnika gronowania obserwowane w sieciach sąsiedztwa słów są specyficzne dla sieci bezskalowych, które posiadają topologię bardzo często identyfikowaną w sieciowych reprezentacjach układów złożonych. Interesującym wnioskiem z przeprowadzonych badań jest to, że zróżnicowanie topologii takich sieci ze względu na obrany język jest mniejsze niż zróżnicowanie ze względu na zawartość

tematyczną. Specjalizacja słownictwa stanowi zatem znacznie silniejszy generator struktury języka, co można ujawnić właśnie w obrazie sieciowym. Zakres tego typu badań przeprowadzonych na potrzeby tej pracy nie pozwala jednak na wyciągnięcie bardziej szczegółowych wniosków, daje mimo to asumpt do prowadzenia dalszych analiz w przyszłości.

Omówione w pracy modele generatywne języka są tylko realizacją pewnych procesów stochastycznych, nie uwzględniających żadnych innych zjawisk (np. zachodzących w mózgu nadawcy komunikatu) i mogących mieć wpływ na przebieg. Ścisłe podejście do generowania sygnału będącego próbką języka naturalnego jest przedmiotem nieco innych, bardziej skomplikowanych analiz – z zakresu sztucznej inteligencji (np. opracowywanie algorytmów generujących sygnał, potrafiący przejść test Turinga). Przedstawione modele miały na celu odtworzenie w kontrolowanych warunkach tych statystycznych własności tekstów, które były przedmiotem zainteresowania w tej rozprawie, jak np. rozkładów Zipfa czy charakterystyk ich sieciowych reprezentacji – rozkładów krotności wierzchołków i średniej długości najkrótszej ścieżki.

Metodologia związana z formalizmem fraktalnym wymagała istnienia odpowiednio długich sygnałów, tak aby uzyskane wyniki były pozbawione wpływu warunków brzegowych, co było kluczowe dla odpowiedniej ich interpretacji. To wymuszało wzięcie pod uwagę tylko tekstów literackich o odpowiedniej długości, nawet jeśli zbadanie własności pewnych nieuwzględnionych tekstów mogło potencjalnie być bardziej interesujące z poznawczego punktu widzenia (przykładem mogą być np. *Bramy raju* J. Andrzejewskiego czy słynny ostatni rozdział *Ulissesa*, zawierające jedynie 1-2 zdania). Analiza była prowadzona na różnej długości szeregach czasowych (związanych z długością poszczególnych książek), czego w innych tego typu badaniach się unika (np. w analizie danych z rynków finansowych dobiera się odpowiednie okna czasowe). To ograniczenie nie mogło jednak rzutować na jakościowe rezultaty przeprowadzonej analizy, choć wyniki ilościowe mogły być podatne na wpływ długości w przypadku sygnałów silnie niestacjonarnych.

Wyniki tej analizy pozwoliły na ujawnienie istnienia multiskalowania na poziomie fluktuacji długości zdań niektórych utworów literackich. Na podstawie najpopularniejszych metod analizy multifraktalnej szeregów czasowych, tj. MF-DFA i WTMM, zaobserwowano intrygującą korelację pomiędzy bogactwem multifraktalności (parametryzowanym przez szerokość widma osobliwości) a wykorzystaniem techniki narracyjnej zwanej strumieniem świadomości. Utwory, będące sztandarowymi dziełami tego typu, takie jak *Finneganów tren* J. Joyce'a, *Gra w klasy* J. Cortázara czy *Fale* V. Woolf wykazywały też najbogatszą multifraktalność, podczas gdy dzieła literatury XIX-wiecznej i wcześniejsze, a także współczesne, ale o tradycyjnej narracji, miały raczej strukturę monofraktalną bądź nieokreśloną (niefraktalną). Wynik ten stanowi niezwykle ciekawy przykład, jak pojęcia pozornie przynależne wyłącznie humanistyce można ująć w sposób czysto ilościowy, pozbawiony subiektywności właściwej dla nauk humanistycznych. To wskazuje, że analiza multifraktalna może być cennym narzędziem także dla analiz czysto literaturoznawczych.

Z uwagi na skomplikowaną naturę wielu aspektów języka, praca ta pokazuje, że potrzebne są dalsze badania nad jego strukturą i dynamiką. Silne zapotrzebowanie ze strony rozwijanych technologii komunikacyjnych może stać się dodatkowym im-

pulsem do dalszego, szybkiego rozwoju tego typu badań. W związku z tym istnieje potrzeba wprowadzenia nowych, efektywniejszych narzędzi analitycznych. Proces ten powinien równolegle indukować teoretyczną analizę języka naturalnego, pozwalającą na odpowiednie kategoryzowanie występujących tu pojęć, oraz przyczyniając się do lepszego zrozumienia jego natury i roli, jaką pełni w życiu indywidualnym i społecznym człowieka.

Dodatek A

Spis wykorzystanej literatury

Wszystkie teksty zostały pozyskane z ogólnodostępnych cyfrowych bibliotek za pośrednictwem poniższych stron internetowych:

Projekt Gutenberg – *www.gutenberg.org*
Free-Ebooks.net – *www.free-ebooks.net*
Spiegel Online Kultur – *www.gutenberg.spiegel.de*
Bibliomania – *www.bibliomania.com*
Classic Reader – *www.classicreader.com*
World Public Library – *www.worldlibrary.net*
The Public's Library and Digital Archive – *www.ibiblio.org*
Wirtualna Biblioteka Literatury Polskiej – *www.univ.gda.pl/literat/books.htm*
19th-Century German Stories – *www.germanstories.vcu.edu*
ABU: La Bibliothèque Universelle – *www.abu.cnam.fr*
ATHENA – *www.athena.unige.ch*
The Online Book Page – *www.onlinebooks.library.upenn.edu*
Library of Spanish Books – *www.donquijote.org*
Open LIBRARY – *www.openlibrary.org*
Literature.org – *www.literature.org*

Pobrane książki zostały oczyszczone z treści nie mających związku z zamieszczonym tekstem literackim, takie jak: informacje edytorskie, spisy treści, przedmowy, posłowia, numery stron, oznaczenia rozdziałów itd. Podczas formatowania tekstu zachowano międzynarodowy standard kodowania znaków *Unicode*, zdolny do wiernego zapisu niemal wszystkich języków świata.

Na następnej stronie zebrano całą literaturę, w oparciu o którą stworzono korpusy dla badanych języków. W nawiasach podano polskie tłumaczenia tytułów oraz rok wydania oryginalnego tekstu. Jeśli nie istnieje polskie przełożenie, bądź jest ono identyczne z oryginalnym tytułem, podano jedynie rok wydania. W przypadku serii tekstów, zwartych pod jednym tytułem, podano przedział, w którym były one publikowane.

Literatura angielska:

J. Austen:

Sense and Sensibility (*Rozważna i romantyczna*, 1811)
Pride and Prejudice (*Duma i uprzedzenie*, 1813)
Mansfield Park (1813)
Emma (1816)
Persuasion (*Perswazje*, 1817)
Northanger Abbey (*Opactwo Northanger*, 1817)

E. Bell:

Wuthering Heights (*Wichrowe wzgórza*, 1847)

K. Brooks:

Kissing the Rain (*Całując deszcz*, 2004)

F.H. Burnett:

The Secret Garden (*Sekretny ogród*, 1911)

A. Christie:

The Secret Adversary (*Tajemniczy przeciwnik*, 1922)
Murder on the Orient Express (*Zabójstwo w Orient Expressie*, 1934)
Towards Zero (*Godzina zero*, 1944)
Endless Night (*Noc i ciemność*, 1967)

J. Cleland:

Memoirs of a Woman of Pleasure (*Pamiętniki Fanny Hill*, 1748)

E. Catton:

The Luminaries (2013)

C. Dickens:

Oliver Twist (1838)
David Copperfield (1850)
Bleak House (*Samotnia*, 1853)
A Tale of Two Cities (*Opowieść o dwóch miastach*, 1859)
Great Expectations (*Wielkie nadzieje*, 1861)

D. Defoe :

Robinson Crusoe (*Przypadki Robinsona Crusoe*, 1719-1720)

A.C. Doyle:

A Study in Scarlet (*Studium w szkarłacie*, 1888)
Adventure of Sherlock Holmes (*Przygody Sherlocka Holmesa*, 1892)

G. Eliot:

The Mill on the Floss (*Młyn nad Flossą*, 1860)
Middlemarch: A Study of Provincial Life (*Miasteczko Middlemarch*, 1871 - 1872)

D. Eggers:

A Heartbreaking Work of Staggering Genius (*Wstrząsające dzieło kulejącego geniusza*, 2000)

H. Fielding:

Bridget Jones's Diary (*Dziennik Bridget Jones*, 1996)

J. Fowles:

The French Lieutenant's Woman (*Kochanica Francuza*, 1969)

E. Gaskell:

North and South (*Północ i Południe*, 1855)

W. Golding:

Lord of the Flies (*Władca much*, 1954)

J. Heller:

Catch 22 (*Paragraf 22*, 1975)

J. Hilton:

Goodbye, Mr. Chips (*Żegnaj Chips*, 1934)

T. Hobbes:

Leviathan (*Lewiatan*, 1651)

H. James:

The Portrait of a Lady (*Portret damy*, 1881)
The Bostonians (*Bostończycy*, 1886)
What Maisie Knew (*O czym wiedziała Maisie*, 1897)

H. Jones:

The Ambassadors (*Ambasadorowie*, 1903)

J. Joyce:

Dubliners (*Dublińczycy*, 1914)
Ulysses (*Ulisses*, 1922)
Finnegans Wake (*Finneganów Tren*, 1939)

S. King:

The Stand: The Complete and Uncut Edition (*Bastion*, 1978)

K. Kesey:

One Flew Over the Cuckoo's Nest (*Lot nad kukułczym gniazdem*, 1962)

C. McCabe:

The Butcher Boy (*Chłopak rzeźnika*, 1979)

C. McCarthy:

Suttree (1979)

H. Melville:

Moby Dick (*Wieloryb*, 1851)

M. Mitchell:

Gone with the Wind (*Przeminęło z wiatrem*, 1936)

T. Morrison:

Beloved (*Pokochać*, 1987)

G. Orwell:

Animal Farm (*Folwark zwierzęcy*, 1945)
1984 (*Rok 1984*, 1949)

J. Dos Passos:

Manhattan Transfer (1925)
U.S.A trilogy (1938)

T. Pynchon:

Gravity's Rainbow (*Tęcza grawitacji*, 1973)

A. Rand:

Atlas Shrugged (*Atlas zbuntowany*, 1957)

A.C. Rice:

The Vampire Chronicles (*Kroniki wampirów*, 1977-2014)

S. Richardson:

Clarissa, or, the History of a Young Lady (1748)

J.K. Rowling:

Harry Potter (1997-2007)

R. Shea, R.A. Wilson:

The Illuminatus! Trilogy (*Trylogia Illuminatus!*, 1997-2007)

U. Sinclair:

The Jungle (*Dżungla*, 1906)

L. Sterne:

The Life and Opinions of Tristram Shandy (*Życie i myśli JW Pana Tristrama Shandy*, 1759)

B. Stoker:

Dracula (*Drakula*, 1897)

D.F. Tartt:

The Goldfinch (*Szczygieł*, 2013)

J.R.R. Tolkien:

The Lord of the Rings (*Władca pierścieni*, 1954-1955)

M. Twain:

The Adventures of Tom Sawyer (*Przygody Tomka Sawyera*, 1876)
Life on the Mississippi (*Życie na Missisipi*, 1883)
The Adventures of Huckleberry Finn (*Przygody Hucka Finna*, 1884)
A Connecticut Yankee In King Arthur's Court (*Jankes na dworze króla Artura*, 1889)

D.F. Wallace:

Infinite Jest (1996)

I. Welsh:

Trainspotting (*Trainspotting. Ślepe tory*, 1993)

O. Wilde:

The Picture of Dorian Gray (*Portret Doriana Graya*, 1906)

V. Woolf:

The... Voyage Out (*Podróż w świat*, 1915)
The Waves (*Fale*, 1931)
The Years (*Lata*, 1937)

Literatura niemiecka:

F. Dahn:

Abermals krähte der Hahn (*I znowu zapiał kur*, 1962)
Ein Kampf um Rom (*Walka o Rzym*, 1878)

K. Deschner:

Der Moloch (*Moloch*, 2002)

A. Döblin:

Berlin Alexanderplatz (1929)

M. Ende:

Die unendliche Geschichte (*Niekończąca się historia*, 1979)

V.C. Felscherinow:

Wir Kinder vom Bahnhof Zoo (*My, dzieci z dworca ZOO*, 1978)

T. Fontane:

Irrungen, Wirrungen (*Rozdroża, bezdroża*, 1888)
Effi Briest (1894)

F. Gerstäcker:

Tahiti (1854)
Nach Amerika (1855)

J.W. von Goethe:

Italienische reise (*Włoska podróż*, 1816-1817)
Faust (1773-1832)
Die Wahlverwandtschaften (1809)

G. Grass:

Danziger Trilogie (*Trylogia gdańska*, 1959-1963)
Das Treffen in Telgte (*Spotkanie w Telgte*, 1979)
Ein weites feld (*Szerokie pole*, 1995)
Beim Häuten der Zwiebel (*Przy obieraniu cebuli* , 2006)

W. Hauff:

Der Mann im Mond (1824)
Complete mitteilungen aus den memoiren des Satan (1825)
Lichtenstein (1826)

J.P. Hebel:

Schatzkaestlein des rheinischen Hausfreundes (1811)

G.W.F. Hegel:

Phenomenologie des Geistes (*Fenomenologia ducha*, 1807)

H. Hesse:

Der Steppenwolf (*Wilk stepowy*, 1927)
Das Glasperlenspiel (*Gra szklanych paciorków*, 1943)

E.T.A. Hoffamn:

Nachtstücke (1816-1817)
Lebens-Ansichten des Katers Murr (*Kota Mruczysława poglądy na życie*, 1819)

F. Kafka:

Der Prozess (*Proces*, 1914)
Der Verschollene (*Ameryka*, 1927)

I. Kant:

Kritik der reinen Vernunft (*Krytyka czystego rozumu*, 1781)

G. Keller:

Die Leute von Seldwyla (*Ludzie z Seldwili*, 1853-1855)

P. Keller:

Ferien vom ich (1915)

B. Kellermann:

Der Tor (1909)
Des Meer (*Morze*, 1910)
Der Tunnel (*Tunel*, 1913)
Totentanz (*Taniec umarłych*, 1948)

P. Lagarde:

Gesammelte Abhandlungen (1923)

H. Mann:

Der Untertan (*Poddany*, 1918)

T. Mann:

Buddenbrooks (*Buddenbrookowie*, 1901)
Der Zauberberg (*Czarodziejska góra*, 1924)
Mario und der Zauberer (*Mario i Czarodziej*, 1930)
Doktor Faustus (1947)

F.W. Nietzsche:

Also Sprach Zarathustra (*Tako rzecze Zaratustra*, 1883–1885)

G. Rosemarie:

Deutsche Maerchen und Sagen (*Niemieckie baśnie i legendy*, 1977)

A. Sapper:

Das kleine Dummerle und andere Erzählungen (1904)

P. Suskind:

Das Parfum (*Pachnidło*, 1997)

R. Walser:

Geschwister Tanner (*Rodzeństwo Tanner*, 1906)

J. Wassermann:

Der Wendekreis (1920)

E. von Wolzogen:

Der Dichter in Dollarica (1912)

Literatura hiszpańska:

J.S. Adalid:

La luz del oriente (*Światło wschodu*, 2005)
Felix de Lusitania (*Felix Lusitania*, 2006)
La tierra sin mal (*Ziemia bez zła*, 2008)

I. Allende:

La casa de los espíritus (*Dom duchów*, 1982)
La suma de los días (*Suma naszych dni*, 2008)

G.T. Ballester:

Filomeno, a mi pesar (1988)

R, Bolaño

2666 (2004)

J.L. Borges:

El Aleph (*Alef*, 1949)

E. Calderón

La bailarina y el inglés (2009)

S.F. Cardenas:

En tela de juicio (*W zawieszeniu*, 1964)
Los peces (*Ryba*, 1968)
Segundo sueño (*Drugi sen*, 1976)
Los desfiguros de mi corazón (*Maski mojego serca*, 1983)

A. Carpentier:

Los pasos perdidos (*Podróż do źródeł czasu*, 1953)
El siglo de las luces (*Eksplozja w katedrze*, 1962)

C.J. Cela:

La cruz de San Andrés (*Krzyż świętego Andrzeja*, 1994)

M. Cervantes:

Don Quijote (*Don Kichot*, 1605)

J. Cortázar

Rayuela (*Gra w klasy*, 1963)
Todos los fuegos el fuego (*Dla wszystkich ten sam ogień*, 1966)
62 / modelo para armar (*62. Model do składania*, 1968)

A.B. Echenique:

El huerto de mi amada (*Ogród z moją ukochaną*, 2002)

L. Esquivel:

Como agua para chocolate (*Przepiórki w płatkach róży*, 1989)

I. Falcones:

La catedral del mar (*Katedra w Barcelonie*, 2006)
La mano de Fatima (*Ręka Fatimy*, 2009)
La reina descalza (*Bosonoga królowa*, 2014)

E. Freire:

Melocotones helados (*Mrożone brzoskwinie*, 1999)

C. Fuentes:

Terra nostra (1975)

M.V. Llosa:

El pez en el aqua (*Jak ryba w wodzie*, 1993)
El paraiso en la otra esquina (*Raj tuż za rogiem*, 2003)

G.G. Márquez:

Cien años de soledad (*Sto lat samotności*, 1967)
El otoño del patriarca (*Jesień patriarchy*, 1975)
Crónica de una muerte anunciada (*Kronika zapowiedzianej śmierci*, 1981)
El amor en los tiempos del cólera (*Miłość w czasach zarazy*, 1985)

J.M. Mendiluce:

Pura vida (*Pełnia życia*, 1998)

J.J. Millas:

El mundo (*Świat*, 2007)

A. Neuman

El viajero del siglo (*Wieczny podróżnik*, 2009)

A. Perez-Reverte

El Club Dumas (*Klub Dumas*, 1993)

E.J. Poncela:

La tournée de Dios (1932)

E. Poniatowska:

La piel del cielo (2001)

M. Puig:

El beso de la mujer araña (*Pocałunek kobiety pająka*, 1976)

L. Restrepo:

Delirio (*Delirium*, 2004)

S. Roncagliolo:

Abril rojo (*Czerwony kwiecień*, 2006)

J. Sempruna:

Le mort qu'il faut (2001)

J. Sierra:

Las ouertas templarias (*Bramy templariuszy*, 2000)

F.L. Ubeda:

Libro de entretenimiento de la Pícara Justina (*Księga rozrywek łotrzycy Justiny)*, 1605)

X. Velasco:

Diablo guardián (2003)

M. Vicent:

Pascua y naranjas (*Wielkanoc i pomarańcze*, 1966)

C.R. Zafón:

La trilogía de la niebla (*Trylogia mgły*, 1995)
La sombra del viento (*Cień wiatru*, 2001)
El juego del ángel, (*Gra anioła*, 2008)
El prisionero del cielo (*Więzień nieba*, 2011)

Literatura francuska:

H. Balzac:

La peau de chagrin (*Jaszczur*, 1831)
Les Proscrits (*Wygnańcy*, 1831)
Le Père Goriot (*Ojciec Goriot*, 1835)
La Comédie humaine (*Komedia ludzka*, 1830-1856)
Illusions perdues (*Stracone złudzenia*, 1836-1843)

S. Beckett:

Molloy (1951)
Malone meurt (1951)
L'Innommable (1953)

J. Bédier:

Le roman de Tristan et Iseut (*Tristian i Izolda*, 1902-1905)

H. Beyle:

Le Rouge et le Noir (*Czerwone i czarne*, 1830)

A. Camus:

L'Étranger (*Obcy*, 1942)
La Peste (*Dżuma*, 1947)
La Chute (*Upadek*, 1952)

L-F. Céline:

Voyage au bout de la nuit (*Podróż do kresu nocy*, 1932)
Mort à crédit (*Śmierć na kredyt*, 1936)

A. Daudet:

Le Petit Chose (1878)

A. Dumas (ojciec):

Georges (1843)
Le Comte de Monte Christo (*Hrabia Monte Christo*, 1844)
Les Trois Mousquetaires (*Trzej muszkieterowi*, 1844)
La Reine Margot (*Królowa Margot*, 1845)
Vingt ans après (*Dwadzieścia lat później*, 1845)
Le Chevalier de Maison-Rouge (1846)
La Dame de Monsoreau (*Pani Monsoreau*, 1846)
Le Collier de la reine (*Naszyjnik królowej*, 1850)

Le Vicomte de Bragelonne (*Wicehrabia de Bragelonne*, 1850)
Ange pitou (*Anioł Pitou*, 1850-1851)
Les Compagnons de Jéhu (*Towarzysze Jehudy*, 1857)

G. Flaubert:

Madame Bovary (*Pani Bovary*, 1856)
L'Éducation sentimentale (*Szkoła uczuć*, 1869)

H. de Graffigny, G. Le Faure:

Aventures extraordinaires d'un savant russe (1888)

V. Hugo:

Han d'Islande (*Han z Islandii*, 1823)
Notre dame de Paris (*Dzwonnik z Notre-Dame*, 1831)
Les Miserebles (*Nędznicy*, 1862)
Quatrevingt-treize (*Rok dziewięćdziesiąty trzeci*, 1873)
Actes et parole (1875-1876)

P.C. de Laclos:

Les Liaisons dangereuses (*Niebezpieczne związki*, 1782)

M. Leblanc:

Les dents du tigre (*Zęby tygrysa*, 1921)

A. Malraux:

La Condition humaine (*Dola człowiecza*, 1933)

G. de Maupassant:

Bel-Ami (1885)

G. Perec:

La Disparition (*Zniknięcie*, 1969)

M. Proust:

À la recherche du temps perdu (*W poszukiwaniu straconego czasu*, 1913)

G. Sand:

La Comtesse de Rudolstadt (*Hrabina Rudolstadt*, 1843)
Les maîtres sonneurs (1853)
La Daniella (1857)
Ces beaux messieurs de Bois-Doré (1858)

G. Scudéry

Artamène ou le Grand Cyrus (1649-1659)

E. Sue:

Les mystères de Paris (*Tajemnice Paryża*, 1842-1843)

E. Zola:

Les Rougon-Macquart (*Rougon-Macquartowie*, 1871-1893)
Les Trois Villes (*Trzy miasta*, 1893-1898)
La Débâcle (*Klęska*, 1892)
Germinal (1885)
L'Assommoir (*W matni*, 1877)
La Bête humaine (*Bestia ludzka*, 1890)

Literatura włoska:

A. Abati:

Delle frascherie, fasci tre (1651)

D. Alighieri:

Divina Commedia (*Boska Komedia*, 1555)

E. Amicis:

Cuore (*Serce*, 1886)

G. Annunzio:

Il trionfo della morte (*Triumf śmierci*, 1894)

T. Avoledo:

Le radici del cielo (*Korzenie niebios*, 2011)

G. Bazzoni:

Falco della rupe O La guerra di Musso (2011)

L. Bianchini:

Io che amo solo te (*Kocham tylko ciebie*, 2013)

L. Blissett:

Q (*Taniec śmierci*, 1999)

G. Boccaccio:

Elegia di Madonna Fiammetta (1343-1344)
Decameron (*Dekameron*, 1350-1353)

G. Bruno:

Spaccio de la bestia trionfante (1584)

D. Buzzati:

Il deserto dei Tartari (*Pustynia Tatarów*, 1940)

I. Calvino:

La giornata d'uno scrutatore (*Długi dzień Ameriga*, 1963)
Ti con zero (*T Zero*, 1967)
Il castello dei destini incrociati (*Zamek krzyżujących się losów*, 1969)
Se una notte d'inverno un viaggiatore (*Jeśli zimową nocą podróżny*, 1979)

L. Capuana:

Giacinta (*Hiacynta*, 1879)

G. Carofiglio:

Testimone inconsapevole (*Świadek mimo woli*, 2002)
Ad occhi chiusi (*Z zamkniętymi oczami*, 2003)
Ragionevoli dubbi (*Ponad wszelką wątpliwość*, 2006)
Le perfezioni provvisorie (*Ulotna doskonałość*, 2010)

G. Deledda:

Elias Portolu (1900)
Annalena Bilsini (1927)

U. Eco:

Il pendolo di Foucault (*Wahadło Foucaulta*, 1988)
Il nome della rosa (*Imię róży*, 1980)
L'isola del giorno prima (*Wyspa dnia poprzedniego*, 1994)
La misteriosa fiamma della regina Loana (*Tajemniczy płomień królowej Loany*, 2004)
Il cimitero di Praga (*Cmentarz w Pradze*, 2010)

O. Fallaci:

Lettera a un bambino mai nato (*List do nienarodzonego dziecka*, 1975)

G.T. Lampedusa:

Il Gattopardo (*Lampart*, 1958)

N. Machiavelli:

Discorsi sopra la prima Deca di Tito Livio (*Dyskursy o Liwiuszu*, 1513)
Dell'arte della guerra (*Sztuka wojny*, 1520)

V.M. Manfredi:

Chimaira, (2001)
L'ultima legione (*Ostatni legion*, 2002)

A. Moravia:

Il disprezzo (*Pogarda*, 1954)
1934 (1982)

F. Petrarca:

Canzoniere (*Sonety do Laury*, 1336)

G. Rossi:

Larici senza confini (2013)

L. Sciascia:

Todo modo (1974)

R. Monaldi, F. Sorti:

Veritas (2006)

S. Tamaro:

Rispondimi (*Odpowiedź*, 2010)

T. Tasso:

Gerusalemme liberata (*Jerozolima wyzwolona*, 1581)

G. Verga

I Malavoglia (*Rodzina Malavogliów*, 1881)
Mastro Don Gesualdo (*Mastro-don Gesualdo*, 1889)
Tutte le novella (*Wszystkie wiadomości*, 1887-1922)

Literatura polska:

J. Abramow-Newerly:

Alinaci (1987)

J. Bator:

Ciemno, prawie noc (2012)
Piaskowa góra (2009)

W. Berent:

Próchno (1903)
Ozimina (1911)

Z. Białas:

Korzeniec (2011)

K.O. Borchardt:

Znaczy kapitan (1960)
Szaman morski (1986)

M. Bujko:

Złoty pociąg (2006)

M. Dąbrowska:

Noce i dnie (1931-1934)

J. Gerhard:

Luny w Bieszczadach (1959)

W. Gombrowicz:

Ferdydurke (1937)

G. Herling-Grudziński:

Inny świat (1953)

K. Irzykowski:

Pałuba (1903)

J. Iwaszkiewicz:

Sława i chwała (1956-1962)

J. Kaden-Bandrowski:

Generał Barcz (1923)

R. Kapuściński:

Podróże z Herodotem (2004)

Z. Kossak-Szczucka:

Pożoga (1922)

A. Kowalska:

Pestka (1996)

W. Kuczok:

Senność (2008)

A. Lange:

Stypa (1911)

S. Lem:

Opowieści o pilocie Pirxie (1968)

K. Boruń, A. Trepka

Zagubiona przyszłość (1954)
Proxima (1956)
Kosmiczni bracia (1959)

W. Łoziński:

Zaklęty dwór (1859)

A. Mickiewicz:

Pan Tadeusz (1834)

H. Mniszkówna:

Trędowata (1909)

M. Moczar:

Barwy walki (1962)

W. Myśliwski:

Widnokrąg (1997)
Ostatnie rozdanie (2013)

Z. Nałkowska:

Granica (1935)

E. Orzeszkowa:

Meir Ezofowicz (1878)
Nad Niemnem (1888)

S. Pasecki:

Żywot człowieka rozbrojonego (1962)

E. Paukszta:

Pogranicze (1961)

J. Pilch:

Tysiąc spokojnych miast (1997)
Pod mocnym aniołem (2000)
Miasto utrapienia (2004)

B. Prus:

Lalka (1887–1889)
Emancypantki (1894)
Faraon (1897)

W. Raymont:

Ziemia obiecana (1899)
Chłopi (1904)

R. Sadaj:

Prowincjusz (1979)

H. Sienkiewcz:

Trylogia (1884-1888)
Krzyżacy (1900)
Bez dogmatu (1891)
Rodzina Połanieckich (1894)

Quo vadis (1896)
Wiry (1910)

A. Struga:

Dzieje jednego pocisku (1910)

D. Terakowska:

Ono (2003)

T. Tryzna:

Panna Nikt (1994)

Z. Urbanowska:

Róża bez kolców (1903)

M. Wańkowicz:

Ziele na kraterze (1951)

W. Wasilewska:

Ojczyzna (1935)

S. Żeromski:

Syzyfowe prace (1897)
Ludzie bezdomni (1900)
Popioły (1902)
Przedwiośnie (1924)

W. Żukrowski:

Kamienne tablice (1966)

J. Żuławski

Trylogia księżycowa (1903-1911)

Literatura rosyjska:

Andriej Biały:

Петербургъ (*Petersburg*, 1909)

Fiodor Dostojewski:

Идиотъ (*Idiota*, 1867)
Бесы (*Biesy*, 1872)
Дневникъ писателя (*Dziennik pisarza*, 1877)
Братья Карамазовы (*Bracia Karamazow*, 1880)
Преступление и наказание (*Zbrodnia i kara*, 1887)

Nikolaj Gogol:

Мёртвые души (*Martwe dusze*, 1842)

Aleksander Sołżenicyn:

Архипелагъ ГУЛАГ (*Archipelag GUłag*, 1869)

Michaił Szołochow:

Тихий Донъ (*Cichy Don*, 1928)

Lew Tołstoj:

Война и миръ (*Wojna i pokój*, 1869)
Анна Каренина (*Anna Karenina*, 1877)
Воскресение (*Zmartwychwstanie*, 1899)

Bibliografia

[1] M.H. Christiansen, S. Kirby. *Language evolution: the hardest problem in science?* Oxford University Press, 2003

[2] M. Cassandro, P. Collet, A. Galves, C. Galves. *A statistical-physics approach to language acquisition and language change.* Physica A: Statistical Mechanics and its Applications 263.1, 427-437, 1999

[3] C. Darwin. *On the Origin of Species.* London, 1859

[4] I. Kant. *Muthmasslicher Anfang der Menschengeschichte.* Philosophische Ansichten der Geschichte, 1786

[5] S. Pinker, P. Bloom. *Natural Language and Natural Selection.* Behavioral and Brain Sciences 13.04, 707-727, 1990

[6] B. Arensburg, A.M. Tillier, B. Vandermeersch, H. Duday, L.A. Schepartz, Y. Rak. *A middle palaeolithic human hyoid bone.* Nature 338.6218, 758-760, 1989

[7] A. Bass, E. Gilland, R. Baker. *Evolutionary Origins for Social Vocalization in a Vertebrate Hindbrain–Spinal Compartment.* Science 321.5887, 417–421, 2008

[8] M. Sigman, G.A. Cecchi. *Global organization of the wordnet lexicon.* Proceedings of the National Academy of Sciences of the United States of America 99.3, 1742–1747, 2002

[9] S. Pinker. *The language instinct: how the mind creates language.* Harper Perennial Modern Classics, 2007

[10] M.A. Nowak, N.L. Komarova, P. Niyogi. *Computational and evolutionary aspects of language.* Nature 417.6889, 611-617, 2002

[11] V. Deoliviera, M. Gomesand, I. Tsang. *Theoretical model for the evolution of the linguistic diversity.* Physica A: Statistical Mechanics and its Applications 361.1, 361-370, 2006

[12] J. Nicolas. *The origin and dispersal of languages: Linguistic evidence.* California Academy of Sciences, 1998

[13] C. Perreault, S. Mathew. *Dating the origin of language using phonemic diversity.* PloS One 7.4, e35289, 2012

[14] M. Serva, F. Petroni. *Indo-european languages tree by levenshtein distance.* Europhysics Letters 81.6, 68005, 2008

[15] P. Liebermanand, E.S. Crelin. *On the speech of Neandertal Man.* Linguistic Inquiry 2.2, 203-222, 1971

[16] E. Silva, V. Deoliveira. *Evolution of the linguistic diversity on correlated landscapes.* Physica A: Statistical Mechanics and its Applications 387.22, 5597-5601, 2008

[17] B. Heine, H. Honken. *The Kx'a Family: A New Khoisan Genealogy.* Journal of Asian and African Studies 79, 5–36, 2010

[18] D. Falk. *Prelinguistic evolution in early hominins: whence motherese?* Behavioral and Brain Sciences 27.04, 491-503, 2004

[19] R. Burling. *The talking ape: how language evolve.* Oxford University Press, 2005

[20] G. Baxter, R. Blythe, W. Croft, A. McKane. *Utterance selection model of language change.* Physical Review E 73, 046118, 2006

[21] S.R.H. Steklis, D. Horst, J. Lancaster. *Origins and Evolution of Language and Speech.* Annals of the New York Academy of Sciences 280, 1976

[22] P. Lieberman. *On the nature and evolution of the neural bases of human language.* American Journal of Physical Anthropology 119, 36-62, 2007

[23] M.D. Hauser, N. Chomsky, W.T. Fitch. *The faculty of language: what is it, who has it, and how did it evolve?* Science 298.5598, 1569-1579, 2002

[24] R. Allott. *The Motor Theory of Language: Origin and Function.* Language Origin: A Multidisciplinary Approach 74, 105-119, Springer 1992

[25] S. Mithen. *The Singing Neanderthals.* Harvard University Press, 2006

[26] J. Aitchison. *The Seeds of Speech.* Cambrige University Press, 2000

[27] R. Botha, C. Knight. *The prehistory of language.* Oxford University Press, 2009

[28] E. Bakalis, A. Galani. *Modeling language evolution: aromanian, an endangered language in greece.* Physica A: Statistical Mechanics and its Applications 391.20, 4963-4969, 2012

[29] N. Chomsky. *Three factors in language design.* Linguistic Inquiry 36.1, 1-22, 2005

[30] C. Knight. *The origins of symbolic culture.* Homo novus: a human without illusions, 193–211, Springer-Verlag 2010

[31] C.F. Hackett. *The Origin of Speech.* Scientific American 203, 88-96, 1960

[32] A. Leroi-Gourhan. *Gesture and Speech.* MIT Press, 1993

[33] J.R. Skoyles. *Gesture, Language Origins, and Right Handedness.* Psycoloquy 11.024, 2000

[34] M. Corballis. *From Hand to Mouth.* Priction University Press, 2002

[35] N.C. Capone, K.K. McGregor. *Gesture development: A review for clinical and research practices.* Journal of Speech, Language, and Hearing Research 47.1, 173-186, 2004

[36] I.M. Roca. *Logical Issues in Language Acquisition.* Fortis Publication, 1990

[37] L.J. Boe, Louis, S. Maeda, J.L. Heim. *Neandertal man was not morphologically handicapped for speech.* Evolution of Communication 3.1, 49-77, 1999

[38] F. Cucker, S. Smale, D-X. Zhou. *Modeling language evolution.* Foundations of Computational Mathematics 4.3, 315-343, 2004

[39] M.H. Christiansen. *Language as shaped by the brain.* Behavioral and Brain Sciences 31.05, 489-509, 2008

[40] M. Brune. *On human self-domestication, psychiatry, and eugenics.* Philosophy, Ethics, and Humanities in Medicine 2.1, 21, 2007

[41] S.E. Fisher et. al. *Localisation of a gene implicated in a severe speech and language disorder.* Nature Genetics 18.2, 168–170, 1998

[42] C. Beckner, R. Blythe, J. Bybee, M.H. Christiansen, W. Croft, N.C. Ellis, J. Holland, J. Ke, D. Larsen-Freeman, T. Schoenemann. *Language is a complex adaptive system.* Language Learning 59.s1, 1-26, 2009

[43] J. Krause, et al. *The Derived FOXP2 Variant of Modern Humans Was Shared with Neandertals.* Current Biology 17.21, 1908-1912, 2007

[44] K.C. Diller, R.L. Cann. *Evidence Against a Genetic-Based Revolution in Language 50 000 Years Ago.* Oxford University Press, 2009

[45] M. Musso, A. Moro, V. Glauche, M. Rijntjes, J. Reichenbach, C. Büchel, C. Weiller. *Broca's area and the language instinct.* Nature Neuroscience 6.7, 774-781, 2003

[46] J.P. Pinasco, L. Romanelli. *Coexistence of languages is possible.* Physica A: Statistical Mechanics and its Applications 361.1, 355-360, 2006

[47] N. Chomsky. *Language and Mind.* Cambrige University Press, 2006

[48] D. Loritz. *How the Brain Evolved Language.* Oxford University Press, 1999

[49] M.C. King, A.C. Wilson. *Evolution at two levels in humans and chimpanzees.* Science 11.188, 107-116, 1975

[50] G. Yule. *Study of Language.* Cambrige University Press, 2010

[51] R. Hegger, H. Kantz, L. Matassini. *Denoising human speech signals using chaoslike features.* Physical Review Letters 84.14, 3197-3200, 2000

[52] M. Dunn, S.J. Greenhill, S.C. Levinson, R.D. Gray *Evolved Structure of Language Shows Lineage-Specific Trends in Word-order Universals.* Nature 473.7345, 79-82, 2011

[53] M. Tomasello. *Origins of Human Communication.* MIT Press, 2008

[54] N. Chomsky. *Powers and Prospects. Reflections on human nature and the social order.* South End Press, Boston 1996

[55] V. Loreto, L. Steels. *Social dynamics: emergence of language.* Nature Physics 3.11, 758-760, 2007

[56] C. Shannon. *A mathematical theory of communication.* ACM SIGMOBILE Mobile Computing and Communications Review 5.1, 3-55, 2001

[57] M.A. Nowak, J.B. Plotkin, V.A.A. Jansen. *The evolution of syntactic communication.* Nature 404.6777, 495-498, 2000

[58] W.G. Mitchener, M.A. Nowak. *Chaos and language.* Proceedings of the Royal Society of London-B, 271.1540, 701-404, 2004

[59] M.A. Nowak, N.L. Komarova, P. Niyogi. *Evolution of universal grammar.* Science 291.5501, 114-118, 2001

[60] L. White. *Second Language Acquisition and Universal Grammar.* Cambridge University Press, 2003

[61] S. Wolfram. *Universality and complexity in cellular automata.* Physica D: Nonlinear Phenomena 10.1, 1-35, 1984

[62] M. Gell-Mann. *What is Complexity?* Complexity and industrial clusters, 13-24, Physica-Verlag HD 2002

[63] P.W. Anderson. *More is different.* Science 177.4047, 393-396 1972

[64] M. Gell-Mann, S. Lloyd. *Effective Complexity.* Nonextensive entropy, 387-398, 2004

[65] L.P. Kadanoff, W. Gotze, D. Hamblen, R. Hecht, E.A.S. Lewis, V.V. Palciauskas, M. Rayl, J. Swift, D. Aspnes, J.W. Kane. *Static Phenomena Near Critical Points: Theory and Experiment.* Reviews of Modern Physics 39.2, 395, 1967

[66] S. Lloyd, H. Pagels. *Complexity as thermodynamic depth.* Annals of Physics 188.1, 186-213, 1988

[67] N. Ay, M. Muller, A. Szkoła. *Effective complexity and its relation to logical depth.* Information Theory, IEEE Transactions on 56.9, 4593-4607, 2010

[68] P. Schuster. *How does complexity arise in evolution?* COMPLEXITY-NEW YORK 2, 22-30, 1996

[69] J. Kwapień, S. Drożdż. *Physical approach to complex systems.* Physics Reports 515.3, 115-226, 2012

[70] D. Chu, R. Strand, R. Fjelland. *Theories of complexity.* Complexity 8.3, 19-30, 2003

[71] P. Bak, C. Tang, K. Weisenfeld. *Self-organized criticality.* Physical Review A 38.1, 364, 1988

[72] D.L. Turcotte. *Self-organized criticality.* Reports on Progress in Physics 62.10, 1377, 1999

[73] J.M. Carlsson, J. Doyle. *Complexity and robustness.* Proceedings of the National Academy of Sciences 99.1, 2538-2545, 2002

[74] D. Sornette. *Critical Phenomena In Natural Sciences.* Springer Science & Business, 2006

[75] V. Schwämmle, P.M. Castro de Oliveira. *A simple branching model that reproduces language family and language population distributions.* Physica A: Statistical Mechanics and its Applications 388.14, 2874-2879, 2009

[76] A. Baronchelli, E. Caglioti, V. Loreto. *Artificial sequences and complexity measures.* Journal of Statistical Mechanics: Theory and Experiment 2005.04, P04002, 2005

[77] G.K. Zipf. *Human behavior and the principle of least effort.* 1949

[78] C. Cattuto, V. Loreto, V.D.P. Servedio. *A yule-simon process with memory.* Europhysics Letters 76.2, 208-214, 2006

[79] R. Ferrer i Cancho, R.V. Sole. *Least effort and the origins of scaling in human language.* Proceedings of the National Academy of Sciences of the United States of America 100.3, 788-791, 2003

[80] W. Marslen-Wilson. *Linguistic structure and speech shadowing at very short latencies.* Nature 244, 1973

[81] R. Solomonoff. *A Formal Theory of Inductive Inference.* Information and Control 7.1, 1-22, 1964

[82] J.G. Chaitin. *On the Simplicity and Speed of Programs for Computing Infinite Sets of Natural Numbers.* The Journal of Alternative and Complementary Medicine 16.3, 407–422, 1969

[83] A. Kolmogorov. *Logical basis for information theory and probability theory.* IEEE Transactions on Information Theory 14.5, 662–664, 1968

[84] B.B. Mandelbrot. *The Fractal Geometry of Nature.* New Scientist 127.1734, 38-43, 1990

[85] H.O. Peitgen, H. Jurgens, D. Saupe. *Chaos and Fractals: new frontiers of science.* Springer, 2004

[86] A. Bunde, S. Havlin. *Fractals and Disordered Systems.* Springer-Verlag New York Incorporation, 1991

[87] S. Lloyd. *Measures of complexity: A non-exhaustive list.* IEEE Control System Magazine 21.4, 7-8 2001

[88] J. Perello, J. Masoliver, A. Kasprzak, R. Kutner. *A model for interevent times with long tails and multifractality in human communications: An application to financial trading.* Physical Review E 78.3, 036108, 2008

[89] L.L.L.B. Gonalves. *Fractal power law in literary english.* Physica A: Statistical Mechanics and its Applications 360.2, 557-575, 2006

[90] C.X. Huang, Sh.L Peng. *Fractals of forbidden words and approximating their box dimensions.* Physica A: Statistical Mechanics and its Applications 387.2, 703-716, 2008

[91] M. Ausloos. *Generalized Hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series.* Physical Review E 86.3, 031108, 2012

[92] Z. Burda, A. Krzywicki, O.C. Martin, Z. Tabor. *From simple to complex networks: inherent structures, barriers and valleys in the context of spin glasses.* Physical Review E 73, 031108, 2006

[93] A.L. Barabási, R. Albert. *Emergence of scaling in random networks.* Science 286.5439, 509-512, 1999

[94] R. Albert, A.L. Barabási. *Statistical mechanics of complex networks.* Reviews of Modern Physics 74.1, 47, 2001

[95] N.L. Biggs, E.K. Lloyd, R.J. Wilson *Graph Theory 1736-1936.* Clarendon Press, 1986.

[96] E. Ravasz, A.L. Barabási *Hierarchical organization in complex network.* Physical Review E 67.2, 026112, 2003

[97] L. Bogacz, Z. Burda, B. Waclaw. *Homogeneous complex networks.* Physica A: Statistical Mechanics and its Applications 366, 587-607, 2006

[98] S.N. Dorogovtsev, J.F.F. Mendes. *Accelerated growth of networks.* Handbook of Graphs and Networks: From the Genome to the Internet, 2006

[99] G. Alexanderson. *Euler and Konigsberg's bridges: a historical view.* Bulletin of the american mathematical society 43.4, 67-573, 2006

[100] S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes. *Ising Model on Networks with an Arbitrary Distribution of Connectionsm.* Physical Review E 66.1, 016104, 2002

[101] S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes. *Potts model on complex networks.* The European Physical Journal B-Condensed Matter and Complex Systems 38.2, 177-182, 2004

[102] P. Erdős, A. Rényi *On the evolution of random graphs.* Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5, 17–61, 1960

[103] M. Kaiser. *Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks.* New Journal of Physics 10.8, 083042, 2008

[104] D.J. Watts, S.H. Strogatz, *Collective dynamics of 'small-world' networks.* Nature 393.6684, 440–442, 1998

[105] B. Bollobas, O. Riordan. *The diameter of a scale-free random graph.* Combinatorica 24.1, 5-34, 2004

[106] A. Fronczak, P. Fronczak, J.A. Holyst. *Average path length in random networks.* Physical Review E 70.5, 056110, 2004

[107] J. Travers, S. Milgram. *An Experimental Study od the Small World Problem.* Sociometry 32, 425-443, 1969

[108] M. Barthelemy. *Betweeness Centrality in Large Complex Network.* The European Physical Journal B-Condensed Matter and Complex Systems 38.2, 163-168, 2004

[109] K.I. Goh, B. Kahng, D. Kim. *Universal behaviour of load distribution in scale-free networks.* Physical Review Letters 87.27, 279701, 2001

[110] D. Chen, L. Lüb, M-S. Shang, Y-C. Zhang, T. Zhoua. *Identifying influential nodes in complex networks.* Physica A: Statistical Mechanics and its Applications 391.4, 1777-1787, 2012

[111] I. Grabska-Gradzińska, A. Kulig, J. Kwapień, S. Drożdż. *Complex network analysis of literary and scientific texts.* International Journal of Modern Physics C 23.07, 2012

[112] L. Sheng, Ch. Li. *English and chinese languages as weighted complex networks.* Physica A: Statistical Mechanics and its Applications 388.12, 2561-2570, 2009

[113] M. Markosova. *Network model of human language.* Physica A: Statistical Mechanics and its Applications 387.2, 661-666, 2008

[114] T. Lei. *Similarity between the Mandelbrot set and Julia Sets.* Communications in Mathematical Physics 134.4, 587–617, 1990

[115] P. Meakin. *Fractals, Scaling and Growth far from Equilibrium.* Cambridge University Press, 1998

[116] A.Z. Górski, S. Drożdż, A. Mokrzycka, J. Pawlik. *Accuracy analysis of the box counting algorithm.* Acta Physica Polonica A 121.2B, B28-B30, 2012

[117] H.E. Stanley. *Scaling, universality, and renormalization: Three pillars of modern critical phenomena.* Reviews of Modern Physics 71.2, 358, 1999

[118] A.O. Gogolin, A. Nersesyan, A.M. Tsvelik. *Bosonization and strongly correlated systems.* Cambridge University Press, 2004

[119] H.E. Stanley, P. Meakin *Multifractal phenomena in physics and chemistry.* Nature 335.6189, 405–409, 1988

[120] G. Troll, P. Graben. *Zipf's law is not a consequence of the central limit theorem.* Physical Review E 57.2, 1347-1355, 1998

[121] M.E.J. Newman. *Power laws, Pareto distributions and Zipf 's law.* Contemporary Physics 46.5, 323-351, 2005

[122] K. Huang. *Introduction to statistical physics.* CRC Press 2001

[123] G. Nicolis, C. Nicolis, J.S. Nicolis. *Chaotic dynamics, markov partitions, and zipf's law.* Journal of Statistical Physics 54.3, 915-924, 1989

[124] G.K. Zipf. *Selective studies and the principle of relative frequency in language.* MIT Press, 1932

[125] W.Li. *Fitting chinese syllable-to-character mapping spectrum by the beta rank function.* Physica A: Statistical Mechanics and its Applications 391.4, 1515-1518, 2012

[126] H. Zhou, G.W. Slater. *A metric to search for relevant words.* Physica A: Statistical Mechanics and its Applications 329.1, 309-327, 2003

[127] B. Corominas-Murtra, J. Fortuny, R.V. Solé. *Emergence of zipf's law in the evolution of communication.* Physical Review E 83.3, 36115, 2011

[128] R. Ferrer i Cancho. *Zipf's law from a communicative phase transition.* The European Physical Journal B 47.3, 449-457, 2005

[129] W. Dahui, L. Menghui, D. Zengru. *True reason for zipf's law in language.* Physica A: Statistical Mechanics and its Applications 358.2, 545-550, 2005

[130] R. Ferrer i Cancho, R.V. Solé. *The small world of human language.* Proceedings of the Royal Society B: Biological Sciences 268.1482, 2261-2265, 2001

[131] M.A. Montemurro. *Beyond the zipf-mandelbrot law in quantitative linguistics.* Physica A: Statistical Mechanics and its Applications 300.3, 567-578, 2001

[132] M.J. Berryman, A. Allinson, D. Abbott. *Statistical techniques for text classification based on word recurrence intervals.* Fluctuation and Noise Letters 3, L1-L10, 2003

[133] I. Eliazar. *The growth statistics of zipfian ensembles: beyond heaps' law.* Physica A: Statistical Mechanics and its Applications 390, 3189-3203, 2011

[134] A.M. Petersen, J.N. Tenenbaum, S. Havlin, H.E. Stanley, M. Perc. *Languages cool as they expand: Allometric scaling and the decreasing need for new words.* Scientific Reports 2, 2012

[135] J. Kwapień, S. Drożdż, A. Orczyk. *Linguistic complexity: English vs. Polish, text vs. corpus.* Acta Physica Polonica A 117.4, 716-720, 2010

[136] S. Havlin. *The distance between zipf plots.* Physica A: Statistical Mechanics and its Applications 216.1, 148-150, 1995

[137] R. Humphrey. *Stream of Consciousness in the Modern Novel*, University of California Press, 1965

[138] O. Sachs. *In the River of Consciousness.* New York Review of Books 51.1, 41-45, 2004

[139] M. Gerlach, E.G. Altmann. *Stochastic model for the vocabulary growth in natural languages.* Physical Review X 3.2, 21006, 2013

[140] L. Egghe. *Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments.* Journal of the American Society for Information Science and Technology 58.5, 702-709, 2007

[141] F. Font-Clos, G. Boleda, Á. Corral. *A scaling law beyond zipf's law and its relation to heaps' law.* New Journal of Physics 15.9, 93033, 2013

[142] B. Mandelbrot. *Structure formelle des textes et communication.* Word 10, 1-27, 1954

[143] R. Ferrer i Cancho, R.V. Solé. *Two regimes in the frequency of words and the origins of complex lexicons: zipf's law revisited.* Journal of Quantitative Linguistics 8.3, 165-173, 2001

[144] D.Y. Manin. *Mandelbrot's Model for Zipf's Law: Can Mandelbrot's Model Explain Zipf's Law for Language.*, Journal of Quantitative Linguistics 16.3, 274-285, 2009

[145] R. Ferrer i Cancho, R.V. Solé. *Zipf's law and random texts.* Advances in Complex Systems 5.01, 1-6, 2002

[146] W. Li. *Random texts exhibit zipfs-law-like word frequency distribution.* IEEE Transactions on Information Theory 38.6, 1842-1845, 1992

[147] C. Biemann. *A random text model for the generation of statistical language invariants.* Human Language Technologies, 105-112 2006

[148] V.V. Bochkarev, E.Y. Lerner, A.V. Shevlyakova. *Deviations in the zipf and heaps laws in natural languages.* Journal of Physics: Conference Series 490, 2014

[149] K. Kosmidis, A. Kalampokis, P. Argyrakis. *Statistical mechanical approach to human language.* Physica A: Statistical Mechanics and its Applications 366, 495-502, 2006

[150] G. Miller. *Some effects of intermittent silence.* The American Journal of Psychology 70, 311–314, 1957

[151] M.A. Montemurro, D.H. Zanette. *Entropic analysis of the role of words in literary texts.* Advances in Complex Systems 5.01, 7-17, 2002

[152] G. Stephens, W. Bialek. *Statistical mechanics of letters in words.* Physical Review E 81.6, 066119, 2010

[153] H.A. Simon. *A Behavioral Model of Rational Choice.* The Quarterly Journal of Economics 69, 99-118, 1955

[154] K. Kechedzhi, O. Usatenko, V. Yampolski. *Rank distributions of words in correlated symbolic systems and the zipf law.* Physical Review E 72.4, 046138, 2005

[155] A.E. Allahverdyan, W. Deng, Q.A. Wang. *Explaining zipf's law via a mental lexicon.* Physical Review E 88.6, 062804, 2013

[156] H. Darooneh, B. Rahmani. *Finite size correction for fixed word length zipf analysis.* The European Physical Journal B-Condensed Matter and Complex Systems 70.2, 287-291, 2009

[157] S.N. Dorogovtsev, J.F.F. Mendes. *Language as an evolving word web.* Proceedings of the Royal Society B: Biological Sciences 268.1485, 2603-2606, 2001

[158] P.L. Krapivsky, S. Redner. *Connectivity of growing random networks.* Physical Review Letters 85.21, 4629, 2000

[159] S.N. Dorogovtsev, J.F.F. Mendes. *Transition from small to large world in growing networks.* Europhysics Letters 81.3, 30004, 2008

[160] M. Stella, M. Brede. *A k-deformed Model of Growing Complex Networks with Fitness.* Physica A: Statistical Mechanics and its Applications: Statistical Mechanics and its Applications 409, 360-368 2014

[161] R. Cohen, S. Havlin. *Scale-free networks are ultrasmall.* Physical Review Letters 90.5, 058701, 2003

[162] S.P. Borgatti, A.A. Mehra, D.J. Brass, G. Labianca. *Network Analysis in the Social Sciences.* Science 323, 892–895, 2009

[163] *http://www.worldwidewebsize.com/*

[164] J. Zhao, P.Zhu, X. Lu, L. Xuan. *Does the Average Path Length Grow in the Internet?* Towards Ubiquitous Networking and Services, 183-190. Springer Berlin Heidelberg 2008

[165] R.M Roxas, G. Tapang. *Prose and poetry classification and boundary detection using word adjacency network analysis.* International Journal of Modern Physics C 21.04, 503-512, 2010

[166] C. Schulze, D. Stauffer. *Computer simulation of language competition by physicists.* Econophysics and Sociophysics: Trends and Perspectives, 307-332, 2006

[167] D.R. Amancio, E.G. Altmann, O.N. Oliveira, L. Da Fontoura Costa. *Comparing intermittency and network measurements of words and their dependence on authorship.* New Journal of Physics 13.12, 123024, 2011

[168] A.Z. Gorski, S. Drożdż, J. Speth. *Financial multifractality and its subtleties: an example of DAX.* Physica A: Statistical Mechanics and its Applications 316.1, 496-510, 2002

[169] R. Kutner, F. Świtała. *Remarks on the possible universal mechanism of the non-linear long-term autocorrelations in financial time-series.* Physica A: Statistical Mechanics and its Applications 344.1, 244-251, 2004

[170] P. Oświęcimka, J. Kwapień, S. Drożdż. *Components of multifractality in high-frequency stock returns.* Physica A: Statistical Mechanics and its Applications 350.2, 466-474, 2005

[171] P. Oświęcimka, J. Kwapień, I. Celińska, S. Drożdż, R. Rak. *Computational approach to multifractal music.* arXiv preprint arXiv:1106.2902, 2011

[172] D. Makowiec, A. Dudkowska, R. Galaska, A. Rynkiewicz. *Multifractal analysis of normal RR heart-interbeat signals in power spectra ranges.* arXiv preprint q-bio/0702047, 2007

[173] F.S Labini, M. Montuori, L. Pietronero.*Scale-invariance of galaxy clustering.* Physics Reports 293.2, 61-226, 1998

[174] Z. Yu, Y. Leung, Y. Chen, Q. Zhang, V. Anh, Y. Zhou. *Multifractal analyses of daily rainfall time series in Pearl River basin of China.* Physica A: Statistical Mechanics and its Applications 405, 193-202, 2004

[175] C. Meneveau, K.R. Sreenivasan *The multifractal nature of turbulent energy dissipation.* Journal of Fluid Mechanics 224.180, 429-484, 1990

[176] E. Leonardis, S.C. Chapman, W. Daughton, V. Roytershteyn, H. Karimabadi. *Identification of intermittent multifractal turbulence in fully kinetic simulations of magnetic reconnection.* Physical Review Letters 110.20, 205002, 2013

[177] E. Canessa. *Multifractal Properties of Diffusion-Limited Aggregates and Random Multiplicative Processes.* Physica Status Solidi B 166.1, 97-103 1991

[178] A.N. Pavlov, W. Ebeling , L. Molgedey, A.R. Ziganshin, V.S. Anishchenko. *Scaling features of texts, images and time series.* Physica A: Statistical Mechanics and its Applications 300.1, 310-324, 2001

[179] L. Hérebâiécek. *Text Levels. Language Constructs, Constituents and the Menzerath-Altmann Law.* Wissenschaftlicher Verlag Trier 56, 1995

[180] I. Grabska-Gradzińska, A. Kulig, J. Kwapień, S. Drożdż. *Multifractal analysis of sentence lengths in English literary texts.* Global Journal on Technology 3, 2013

[181] J.W. Kantelhardt, S.A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, H.E. Stanley. *Multifractal detrended fluctuation analysis of nonstationary time series.* Physica A: Statistical Mechanics and its Applications 316.1, 87-114, 2002

[182] A. Arneodo, E. Bacry, J.F. Muzy. *The thermodynamics of fractals revisited with wavelets.* Physica A: Statistical Mechanics and its Applications 213.1, 232-275, 1995

[183] C.K. Peng. *Mosaic organization of DNA nucleotides.* Physical Review E 49.2, 1685, 1994

[184] A.L. Barabási, T. Vicsek. *Multifractality of self-affine fractals.* Physical Review A 44.4, 2730-2733, 1991

[185] H.E. Hurst. *Long-term storage capacity of reservoirs.* Transactions of the American Society of Civil Engineers 116, 770-808, 1951

[186] P.S. Kokoszka, M.S. Taqqu. *Infinite variance stable moving averages with long memory.* Journal of Econometrics 73.1, 79-99, 1996

[187] P. Oświęcimka, J. Kwapień, S. Drożdż. *Wavelet versus Detrended Fluctuation Analysis of multifractal structures.* Physical Review E 74.1, 016103, 2006

[188] S. Drożdż, J. Kwapień, P. Oświęcimka, R. Rak. *Quantitative features of multifractal subtleties in time series.* Europhysics Letters 88.6, 60003, 2009

[189] E. Alvarez-Lacalle, B. Dorow, J.P. Eckmann, E. Moses. *Hierarchical structures induce long-range dynamical correlations in written texts.* Proceedings of the National Academy of Sciences of the United States of America 103.21, 7956-7961, 2006

[190] W. Ebeling, A. Neiman. *Long-range correlations between letters and sentences in texts.* Physica A: Statistical Mechanics and its Applications 215.3, 233-241, 1995

[191] M.A. Montemuro, P.A. Pury. *Long-range fractal correlations in literary corpora.* Fractals 10.04, 451-461, 2002