The Henryk Niewodniczański Institute of Nuclear Physics
Polish Academy of Sciences

Complex Systems Theory Department

Complexity characteristics of punctuation usage patterns in written language

Tomasz Stanisz

A thesis submitted for the degree of
Doctor of Philosophy

supervised by
prof. dr hab. Stanisław Drożdż

Kraków 2022

Abstract

Natural language has a number of features that allow it to be treated as a complex system. It has a complicated, hierarchical organization, and the properties and interactions appearing in its structures might not be directly deducible from the properties of the elements of which these structures are built. The subject of this thesis is the study of several aspects of natural language organization, namely the ones that can be characterized with the use of the formalism commonly applied in research on complex systems. The study focuses on written language, in the form of literary texts in several European languages (English, German, French, Italian, Spanish, Polish and Russian). The first part of the analysis discusses power-law distributions describing word frequencies in texts and investigates how the shapes of these distributions are changed when the frequencies of punctuation marks are taken into consideration. The next part focuses on representing texts in the form of time series constructed on the basis of text's partition into sentences or into pieces between consecutive punctuation marks. It turns out that these series have features often found in signals generated by complex systems - the presence of long-range correlations along with fractal or multifractal structures. Moreover, the analysis of the partition of texts into pieces determined by punctuation allows to observe that the distances between consecutive punctuation marks can be described using the discrete Weibull distribution; this can be considered a statistical regularity that applies to texts in all the languages studied in the thesis. The last part of the dissertation is devoted to linguistic networks (networks representing selected aspects of language organization) and focuses on word-adjacency networks - whose structure reflects the co-occurrence of words in texts. The results of word-adjacency network analysis indicate that the quantities characterizing such networks can be used for the classification of texts, for example in authorship attribution. In addition, methods routinely used in research on complex networks have been applied to linguistic networks of a different type, namely the so-called word-association networks, which are constructed on the basis of data collected in certain psycholinguistic experiments. This has allowed to reveal complex structures, significantly different from those observed in random networks. In all of the presented analyzes pertaining to written language, punctuation and its impact on the measurable traits of language is the key issue. In the analysis of word frequency distributions, punctuation marks are treated as words - this leads to a better agreement between the observed distributions and power-law distributions proposed by Zipf's law. Introducing punctuation into word-adjacency networks provides valuable information which can be used to significantly improve the effectiveness of identifying features distinguishing one text from another. The results of time series analysis show that the organization that punctuation introduces into language has both largely universal properties (common to many different texts) and certain features characteristic of particular texts - for example, texts in a specific language.

Streszczenie

Język naturalny posiada szereg specyficznych cech, które pozwalają traktować go jak układ złożony. Ma on skomplikowaną, hierarchiczną organizację, a właściwości i oddziaływania charakterystyczne dla poszczególnych jego struktur niekoniecznie wynikają wprost z właściwości elementów składających się na te struktury. Tematem pracy jest badanie tych aspektów organizacji języka naturalnego, które w użyteczny sposób można opisywać za pomocą formalizmu stosowanego do opisu układów złożonych. Przedmiotem analizy jest język w formie pisanej - którego próbkę stanowią teksty literackie w kilku językach europejskich (angielskim, niemieckim, francuskim, włoskim, hiszpańskim, polskim i rosyjskim). Pierwszą badaną kwestią są rozkłady potęgowe opisujące częstość występowania słów w tekstach oraz wpływ, jaki na kształt tych rozkładów ma uwzględnienie częstości występowania znaków interpunkcyjnych. Kolejnym zagadnieniem jest reprezentacja tekstów w postaci szeregów czasowych, skonstruowanych w oparciu o podział na zdania lub na fragmenty pomiędzy kolejnymi znakami interpunkcyjnymi. Okazuje się, że szeregi te posiadają cechy często spotykane w sygnałach generowanych przez układy złożone - obecność korelacji długozasięgowych i związanych z nimi odpowiednich struktur fraktalnych lub multifraktalnych. Co więcej, analiza podziału tekstów na fragmenty wyznaczone przez interpunkcję pozwala zaobserwować, że odległości pomiędzy kolejnymi znakami interpunkcyjnymi można opisać za pomocą dyskretnego rozkładu Weibulla; stanowi to pewną statystyczną prawidłowość, której podlegają teksty we wszystkich przebadanych w pracy językach. Ostatnia część rozprawy poświęcona jest sieciom lingwistycznym (sieciom złożonym reprezentującym wybrane aspekty organizacji języka) i koncentruje się na sieciach sąsiedztwa słów - których struktura odzwierciedla współwystępowanie słów w tekstach. Rezultaty badania sieci sąsiedztwa słów wskazują, że wielkości charakteryzujące takie sieci mogą być wykorzystywane do klasyfikacji tekstów, na przykład w rozpoznawaniu autorstwa. Dodatkowo, metody analizy sieci zostały zastosowane do sieci lingwistycznych innego typu, konkretnie do tak zwanych sieci skojarzeń pomiędzy słowami, skonstruowanych w oparciu o dane pochodzące z odpowiednich eksperymentów psycholingwistycznych. W sieciach tych zostały zidentyfikowane złożone struktury, istotnie różne od tych, które można zaobserwować w sieciach przypadkowych. We wszystkich przeprowadzonych analizach dotyczących języka pisanego kluczowym zagadnieniem jest interpunkcja i jej wpływ na mierzalne cechy języka. W analizie częstości występowania słów znaki interpunkcyjne są traktowane jak słowa, co prowadzi do zwiększenia zgodności rozkładu częstości z rozkładem potęgowym, określonym prawem Zipfa. Wzięcie pod uwagę interpunkcji w sieciach sąsiedztwa słów dostarcza użytecznej informacji, której uwzględnienie istotnie poprawia efektywność identyfikacji cech rozróżniających teksty. Z rezultatów analizy szeregów czasowych wynika, że organizacja, jaką do języka wprowadza interpunkcja, ma zarówno właściwości w znacznym stopniu uniwersalne (wspólne dla różnych tekstów), jak i pewne cechy charakterystyczne dla poszczególnych tekstów - na przykład dla tekstów w konkretnym języku.

Contents

Introduction

One of the reasons for which humans can be considered extraordinary among all species present on Earth is the ability to think in abstract categories and to efficiently communicate the results of such a thinking process. Although research indicates that some animals are capable of solving tasks appearing to require abstract reasoning and are able to communicate with each other, the complexity and sophistication of human cognitive and communication abilities are enormously superior to those possessed by any other known organisms. The ability to use language is a key factor that allowed for the development of civilization and culture, things that are considered unique for humans.

Language is such an important and multifaceted phenomenon that it draws the attention of a great variety of academic disciplines. To grasp the diverse properties of language and to be able to describe it possibly comprehensively, an interdisciplinary approach is required. Language is a set of symbols and rules, an organism's ability to generate sounds, a communication tool, a logical system of notions guiding the thinking process, as well as a social and cultural phenomenon. Therefore the fields actively studying subjects related to language range from humanities through social sciences to natural and formal sciences, each of them focusing on a different perspective.

Mathematics and physics offer tools that can be successfully applied to language study. A number of concepts originating in these sciences have found their use in the quantitative description of natural language. In physics, an approach that seems to be particularly fruitful is the one based on the notion of the so-called complex systems - a class of systems, typically consisting of a large number of constituents, whose general properties usually cannot be deduced only from the properties of those constituents. Complex systems can often be characterized by the phrase "the whole is something beside the parts". This seems to be well suited for the natural language, whose complicated multilevel structure cannot be simply reduced to a set of rules or laws. A number of traits that are commonly shared among complex systems can be found in natural language, for example hierarchical structure, fractality, or the presence of power laws.

In physics, the description of systems consisting of a great number of elements is naturally done with the use of statistical mechanics; this also applies to complex systems. Due to their generality, statistical physics' tools can be used to study a broad class of systems, ranging from "purely physical", through biological, to social and economic ones. Common statistical properties of these systems can often be grasped by models using stochastic processes characterized by heavy-tailed distributions, hierarchical structure and long-range correlations might be detected by fractal and multifractal analysis, and mutual relations among system's elements are usually conveniently represented by a complex network. These tools, among others, have been successfully applied to the challenging field of quantitative research on natural language. Studying language from physics' point of view gives insight into its complex structure, allows to investigate its usage and evolution, and can also have practical applications in designing tools and methods of automatic language

processing, which are probably going to have a growing impact on everyday life in the future.

The goal of this dissertation is to quantitatively describe several properties of natural language (primarily in its written form, as the data used in the analysis consists of corpora constructed from literary texts in a few different languages). The investigated subjects include: power-law distributions describing word frequencies in linguistic corpora, long-range correlations leading to fractal or multifractal structure of time series representing samples of written language, and the organization of networks illustrating relationships between words in texts. A key issue is how the properties of language seen from the mentioned perspectives are influenced by the presence and the specific usage of punctuation marks. The main theses of this dissertation can be summarized in the form of the following statements.

- Zipf-Mandelbrot law - stating that word frequency distributions in texts have the form of a power law with the exception of a few most frequent words, for which a deviation from a power law can be observed - can be modified to take punctuation marks into account (by treating them in the same way as words). Such a modification reduces the deviation from the power-law form of the rank-frequency relationship. In other words, including punctuation marks into word frequency analysis brings word frequency distribution closer to a power law.

- The structure of texts constituted by the partition into sentences is characterized by the presence of long-range correlations (in time series constructed from the lengths of consecutive sentences) and statistical self-similarity; certain types of texts even exhibit multifractality. Another possible partition of a text is given by all punctuation marks. Interestingly, the mentioned characteristics (long-range correlations, fractality, and multifractality) can also be identified in time series representing the distances (measured in the number of words) between consecutive punctuation marks. The two discussed types of partition lead to two types of time series which are distinct, but related - as sentences are marked by a subset of the set of all punctuation marks. The variability range of Hurst exponents and the degree of multifractality observed in time series representing distances between punctuation marks is systematically smaller than in series representing sentence lengths. This suggests that partitioning a text into pieces separated by punctuation marks is less diversified and leaves less freedom to the writer than partitioning into sentences. From such a point of view, sequences of words between consecutive punctuation marks can be considered text's "building blocks" of nature more fundamental than sentences.

- The distribution of distances between consecutive punctuation marks (measured in the number of words between them) can be approximated with the so-called discrete Weibull distribution. This seems to apply universally to large variety of texts (texts in various languages utilizing diverse styles of writing) and allows to use a simple model to characterize the process which determines where punctuation marks are placed in a text. The parameters of the distributions are to some degree specific to individual languages and therefore a quantitative comparison between statistical properties of punctuation in different languages is possible. Interestingly, in terms of probability distributions, sentence lengths do not behave as regularly as distances between consecutive punctuation marks. This suggests - in accordance with the idea mentioned above - that when compared to partition into sentences, the partition of a text determined by all punctuation marks yields pieces which seem to be more restricted by rules governing text composition and which behave in a more consistent way.

- Word-adjacency networks (networks representing the co-occurrence of words in texts), apart from exhibiting a number of traits which can be considered quite general and universal for language, are capable of identifying properties specific to particular texts. This allows to use the information carried by such networks in stylometric analysis - authorship attribution, for instance. Characteristics of the structure of a word-adjacency network, especially in their local variant (that is, describing the structure in the vicinity of a selected node) provide information that is complementary to the information contained in basic quantities routinely used in text classification (word frequencies, for example).

- If punctuation marks participate in the process of word-adjacency networks' construction, their characteristics in the resulting networks resemble the characteristics of words belonging to the same frequency range (high frequencies). This supports the idea of treating punctuation marks in the same way as words in certain types of statistical analysis of written language.

- Statistical characteristics of punctuation are an important factor in identifying the properties that distinguish one text from another. This is evidenced by the fact that the studied text classification procedures experience a significant decrease in accuracy when punctuation marks are removed from the analysis.

- The generality of network representation allows methods and concepts used to quantify the properties of word-adjacency networks to be applied also to other kinds of linguistic networks, like word-association networks (networks designed to represent how words are associated in human mind).

The thesis is organized as follows. In Chapter 1, a few perspectives on research on natural language are mentioned and the variety of scientific disciplines in which language is studied is presented. Some concepts and ideas which can be considered the foundations of computational approach to the description of language structure are given. The chapter briefly describes the notion of a complex system and discusses several ways of how quantifying complexity is approached. The final part of the chapter lists a few reasons for considering natural language a complex system.

Each of the Chapters 2, 3, 4 is devoted to one aspect of the analysis of language related to complexity - Chapter 2 focuses on power-law distributions describing word frequencies in texts, Chapter 3 discusses insights from time series analysis and fractal geometry applied to time series constructed from linguistic data, and Chapter 4 presents results of investigating the properties of language with the use of complex networks. Each of the chapters contains an introductory part, consisting of basic notions, definitions, and methods used in the analysis.

The key conclusions are collected in Summary. Appendix A contains a short description of one of the machine learning algorithms used in Chapter 4. The books used as the input data throughout this dissertation are listed in Appendix B.

Chapter 1

Natural language and complex systems

1.1 Studying natural language from various perspectives

Depending on the context and aspect of interest, language can be defined in multiple ways. One of obvious and natural ways of understanding the notion of language is to state that language is a structured system of communication. Human language, which spontaneously evolved with the development of human communities, is often referred to as *natural language*, as opposed to *formal language*, which is a mathematical object (a set of sequences derived from a finite set of symbols). Another opposing term is *constructed language* (sometimes called *conlang*). Like natural language, a constructed language is a language whose purpose is communication, but its structure is a result of a planned activity - it is a language with artificially designed vocabulary, grammar or phonology. Examples are Esperanto, Interlingua, or languages created by fantasy writer J. R. R. Tolkien.

The ability to communicate is not an extraordinary phenomenon among animals. Among well-known examples one may list birds singing to attract mates and to repel rivals [1], bees dancing to inform their nestmates about the distance and direction to food sources [2, 3], or dolphins whistling to recognize each other [4–6]. There are plenty of forms of communication between animals, with various types of signals: visual, auditory, olfactory, tactile, etc. [7–14].

However, human language is unique compared to communication systems of other animals. In 1960s, Charles Hockett defined a collection of essential characteristics of language [15], the so-called *design features*, potentially useful in setting language apart from animal communication. The original list evolved over time and has been modified. Although its practical use is in some cases limited [16], the idea of studying the proposed features of language strongly influenced linguistics. Among those features, it is worth to mention *displacement*, *productivity*, *cultural transmission*, *duality of patterning*, *learnability* and *reflexiveness*. Displacement is the possibility of referring to events remote in space and time, to objects that are not present in the immediate environment, or even do not exist. Productivity is the ability of language users to create and understand new expressions that can convey any message; productivity provides that the number of possible utterances in any human language is infinite. Cultural transmission is constituted by the fact that language is learned by interactions with individuals already capable of using it. Although some predisposition to be able to use language may be innate, the key factor in language acquisition is the social setting (it determines, for example, which

particular language is acquired as the first). Duality of patterning refers to the organization of language simultaneously on two levels: meaningless constituents (sounds, letters) are combined into units that have particular meaning (words); these units can be further combined into a complete message. Learnability means that a speaker of some language can learn other languages. Reflexiveness is the ability of language to describe itself. Humans can use language to define what language is, to discuss its structure, or to talk about its usage. Human language is the only known system in nature which exhibits all of the aforementioned features - animal communication systems either possess only some of them in a limited form, or do not have them at all [17, 18].

The uniqueness of human language poses a question about its origin. However, the development of the study on when and how human language came into existence has been severely limited by the lack of empirical evidence that could prove or disprove numerous hypotheses [19, 20]. The direct evidence of the existence of language can be obtained by discovering the earliest traces of writing (which are dated to about 3000 years BC), but speech is much older than written language [18, 21]. Contemporary research on this subject relies on indirect information supplied by paleontology, archaeology, biology, linguistics and cognitive science. Studying fossil record may reveal human ancestors' anatomical traits of potential relevance to language, for example brain size. However, this line of reasoning faces certain limitations, like the lack of possibility of reconstructing brain's internal structure, or the absence of data regarding the evolution of vocal tract [19]. A related approach concentrates on the artefacts left by early humans - depending on the level of sophistication, advanced tools or art might indicate the capability of abstract and symbolic thinking [22]. This can confirm certain cognitive skills at a given stage of human development, but is rather unhelpful in determining how and when exactly language started, as language could precede fossilizable art and advanced tools [23]. A different line of enquiry employs genetics and genomics to study the origin and migration of human populations, as well as to identify the points in time when the lineage of modern humans diverged from other species [19, 24, 25]; this allows, for example, to put constraints on the time period when the language was born. Studying human genome is also aimed at determining which genes are relevant for language capacity. However, due to language complexity, a precise and consistent view on how genes are related to the emergence of language has not yet been reached [20, 26, 27]. Another area of investigation is related to the research on animal cognition and communication. Its major direction is the study of language-related traits in non-human primates [19, 28], in particular chimpanzees, which are the closest living relatives to humans (the last common ancestor of the two species is estimated to have lived between 4 and 8 million years ago [24, 29–33]). Assuming that a trait that is present in all the species sharing a certain common ancestor was probably also present in that ancestor, one may attempt to determine which species preceding humans exhibited particular traits necessary for the development of language. Research on other, more distantly related species may also be informative; an important concept here is convergent evolution - a process of the independent development of a similar feature in a few different species whose last common ancestor did not have that feature (an example is the streamlined body shape shared by penguins, fish, and some aquatic mammals [34]). Convergent evolution is a result of adaptation to similar environment in similar ways. Therefore, studying selected animals' traits at least partially related to language (like vocal learning, which occurs in whales, dolphins, bats and some birds [35]) might be helpful in explaining the mechanisms that have driven the emergence of language [28]. The problem with language-related research concerning animals is that since human language has no similarly complex counterpart among

animals, it is hard to find and choose animals' features that can be unequivocally linked to language. Also, a number of results in this field (regarding, for example, the question whether the great apes' gestural communication possesses some characteristics of human language) are disputable and difficult to interpret [19, 36–39].

With all the methods of investigating the early history of natural language having their problems and limitations, the question about the details of language origin remains unanswered [18]. It is not known whether human traits relevant for language developed as an adaptation for early forms of communication, or whether they are a result or a byproduct of adaptation to other tasks, like tool-making or numerical reasoning [40]. It is not known when the emergence of language started and how long it took as well as whether language's spoken form was the first to appear, or whether it was preceded by gestural communication [22]. The answers of contemporary science to these questions are still to a large extent speculative; it has been stated that within this area of research "the richness of ideas is accompanied by a poverty of evidence" [20]. The constant development of research methods gives hope for the knowledge on language origin to become more and more detailed and precise in the future.

Language is not a static entity. Languages continually undergo gradual changes of lexical, phonological, syntactic and semantic nature. These changes are driven by a number of factors, like migration and language contact, the development of technology, or people's willingness to use the language that they associate with a certain degree of social prestige [18, 41]. Studying how language changes over time allows to get an insight into certain cultural and social processes [42–44]. Languages influence each other, some of languages die out, and new languages can be born. Therefore the number of living, actively used languages also changes over time. Currently, the number of living languages is estimated to around 7000 [45, 46] (the exact number depends, for example, on whether some varieties are classified as separate languages or other entities, like dialects). Research on language history aims to find laws that govern the process of language change, and to answer the question how particular languages are related. It often employs the method of comparing phonological, morphological, or syntactic features of different languages, as well as their lexicons. A noteworthy example of a tool used for such comparisons is the so-called Swadesh list [47]. The Swadesh list is a list consisting of 100 or 200 words (depending on the version; other numbers are also in use) [48], which are assumed to represent basic vocabulary (10 examples of the words in the English Swadesh list are: *water*, *hand*, *tree*, *tooth*, *rain*, *moon*, *long*, *cold*, *give*, *sleep*). Such list can be constructed for virtually any language, by identifying words corresponding to the given meanings in that language. Creating Swadesh lists for two different languages and determining how many words are cognates (words of common etymological origin) allows to analyse the lexical relationships between those languages. Quantitative description of such relationships (Figure 1.1) may be useful in answering the question whether two languages derive from a common parent language, and, if so, when their divergence from each other took place [49, 50].

One may ask not only about how language started and evolved for the whole humanity, but also about the mechanisms driving the development of language in every individual human. Two major schools of thought regarding language acquisition may be distinguished - one of them states that language is an ability that is learned in the way similar to other cognitive skills [52, 53], the other one (established primarily by Noam Chomsky) says that some features of language are innate [54, 55]. The proponents of the latter argue that the amount and the diversity of the information that children are exposed to is too small to properly acquire language skills from the basics. This argument, called the *poverty of the stimulus*, was the key reason

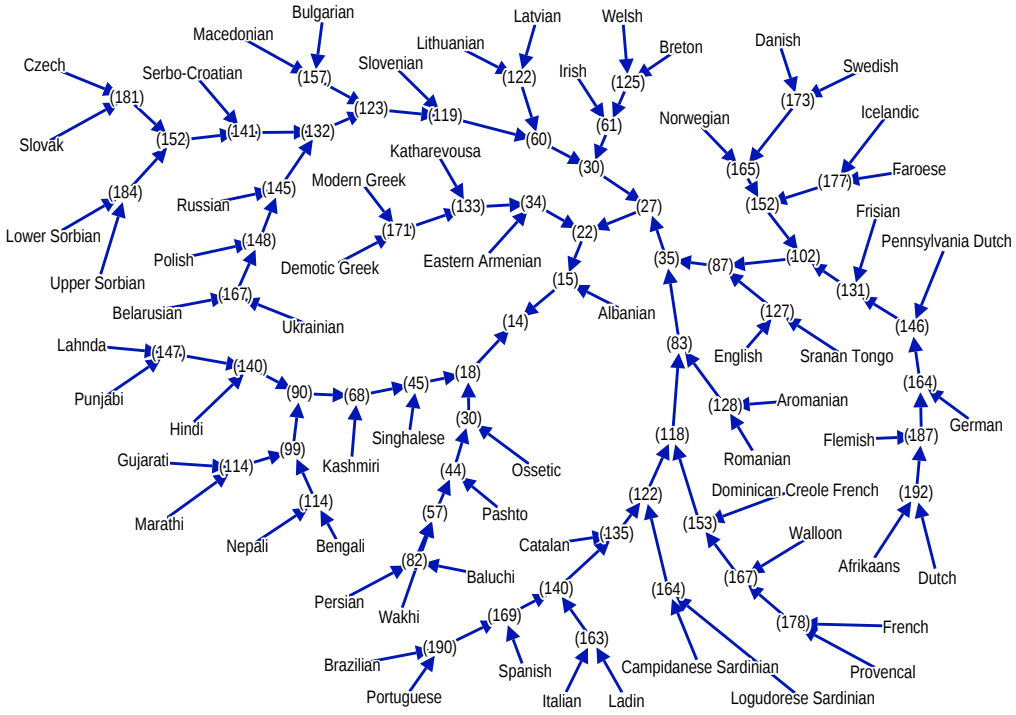


Figure 1.1. The network representing lexical relationships between 63 selected languages from the Indo-European family, based on the subset of data used by Dyen, Kruskal and Black [51]. The data is a multilingual Swadesh list with $N = 200$ entries. Each entry on the list corresponds to one meaning and consists of the words representing that meaning in the studied languages. Those words are associated into groups reflecting their possible common origin. As a result, each pair of words under a given entry is judged as "cognate", "doubtfully cognate", or "not cognate". One can define the proximity $n_c(l_1, l_2)$ between two languages l_1, l_2 as the total number of word pairs judged as "cognate" among all entries; consequently, the distance between l_1 and l_2 can be expressed as $d(l_1, l_2) = N - n_c(l_1, l_2)$. The network presented above is a directed tree representing hierarchical clustering of the studied languages, using the so-defined distances. Each leaf (a node with no ingoing edges) corresponds to one language, and each internal node (a node with ingoing edges) is a cluster of languages. Consecutive groupings into bigger and bigger clusters are represented by arrows (directed edges). Each cluster is labeled with its internal minimum proximity - if k is the number labeling the cluster, then the proximity n_c (the number of shared cognates) between any two languages belonging to that cluster is not smaller than k. More advanced methods of analyzing the distances between languages' lexicons may be useful in attempts to reconstruct evolutionary trees of languages, like in [49].

for introducing the concepts of *language acquisition device* - an innate, theoretical component of human mind responsible for certain linguistic skills - and *universal grammar* - the set of highly abstract rules and characteristics shared among all the world's languages, which are encoded in each human's brain. According to the theories utilizing these concepts, children acquire language relatively quickly (regarding the amount of the available "linguistic data"), because the core cognitive features required are known in advance; they only have to adjust the parameters that can vary among languages. However, the assumptions behind this point of view have been questioned and debated [56–59]; current research emphasizes the role of learning in the process of language acquisition, and, using the results from neuroscience, aims to discover the mechanisms driving this process [60].

One of important perspectives of studying natural language is focused on the relationship between the language and the functioning of the brain. It aims to reveal brain's internal mechanisms responsible for learning and processing language, and investigates which parts of the brain take part in language-related activities. This

field of study benefits greatly from the development of neuroimaging techniques [61, 62], which allow to conduct experiments and measurements able to verify hypotheses about how language is processed and represented in the brain. For a long time the prevalent view on that matter has been that language comprehension and production is to a large extent contained within two regions of cerebral cortex: Broca's area and Wernicke's area. These regions were identified as crucial for the ability to use language by 19th-century physicians, who worked with patients suffering impairment of language abilities caused by brain damage and who were able to link the symptoms with the damage to specific parts of the brain [63]. However, modern research found that treating language-related processes as dependent on only two brain regions is an oversimplification [64]. It turns out that the system of language comprehension and production in the brain constitutes a distributed network involving multiple brain regions. Moreover, that network cannot be easily divided into separate modules responsible for different tasks - a single task meant to interfere with one particular linguistic ability can activate multiple parts of the whole system [62, 65–68]. This shows that the system has a complicated structure and exhibits complex patterns of activity. Studying the way in which language is processed in the brain can potentially lead both to better understanding of human language and to practical contributions to other disciplines - for example, it can be helpful in treating people suffering from language disorders [69].

Another interesting area of investigation is the relationship between language and thinking. There are a number of theories contending that language affects at least some of the other aspects of cognition. There exists an idea that the process of thinking has the structure similar to the structure of language - it combines simple concepts into complex thoughts in the way analogous to the way that syntax combines words into sentences. This hypothetical structure has been named the language of thought, sometimes also called "mentalese" [55, 70, 71]. The hypothesis has been a subject of a debate. There has been, for example, a contrary line of argument, stating that thinking occurs in natural language - each human thinks in the language which he or she speaks [72–75]. Another influential concept, the so-called linguistic relativity (also known under the name of "Sapir-Whorf hypothesis") states that the particular language used by an individual influences their perception and way of thinking [76, 77]. Whorf's idea was based on the study of Hopi language, in which the conceptualization of time is different to the one usually appearing in world's languages. Whorf pointed out that the Hopi language lacks the word referring to time and that its verbs do not distinguish between present, past, and future - therefore the view on the surrounding reality possessed by a Hopi language native speaker may be different from the view typical for the user of languages like English [78, 79]. Although later studies showed that Whorf's conclusions about the inability to refer to time in the Hopi language could have been exaggerated [80], his work initiated a line of research on how particular languages may influence the way of thinking. A language also investigated in this context is the Pirahã language, which has no words for numerals and lacks the notion of counting; its native users seem to have extreme difficulties with acquiring even the most basic numeracy skills [81, 82]. Another example of research in this field are the experiments suggesting that the presence of grammatical gender in a language affects the way that its native speakers describe particular objects and also influences their ability to memorize names given to those objects [83]. Research on the relationships between language and cognitive skills like color perception, spatial orientation and numerical reasoning show that language, understood as a mental ability universal for humans, serves as a powerful tool in cognitive processes, and is particularly useful in transforming various pieces of information into a convenient representation [84].

Since the process of thinking often employs combining simpler concepts into more sophisticated ones, it benefits greatly from the system capable of representing virtually any concept in a standardized, usable manner. Language is such a system, and therefore it can be considered a human mind's resource of great importance, indispensable in complex intellectual activities.

1.2 Computational and quantitative approach to language

The origins of the usage of quantitative and computational methods to study language can be dated back to 1940s and 1950s. The development of this area of research was related to the desire to create systems capable of automatic processing of natural language; one of the first natural language processing tasks that gained much attention was machine translation [85]. After a period of optimism and enthusiasm, it became clear that not only that particular task was difficult, but modeling language in general was much more challenging than it initially appeared. Language has a multilevel, complex structure, and its processing typically requires multiple steps utilizing various tools. A language user needs to be able to extract and recognize a sequence of words from an audio signal and to transform a sequence of words into an audio signal; these actions require knowledge of phonology and phonetics - which describe what sounds are needed to pronounce each word and how these sounds are physically realized. Words may have various forms; using these forms correctly requires knowledge about morphology - which specifies how words can be divided into components and what information is carried by these components (the distinction between the singular and the plural form of a noun, for instance). Maintaining relationships between individual words (their order, for example) utilizes the knowledge of rules given by syntax. The knowledge of the meaning of words and their combinations - semantics - is needed both to understand an utterance and to generate one, as appropriate words have to be found to express a thought. In a conversation, one needs to be aware of the context and the situation in which the conversation takes place - in other words, one has to keep track of the discourse. This requires knowledge about linguistic units larger than a single utterance. Of course, language users do not necessarily have to be able to verbalize what kind of knowledge they are using; nevertheless, comprehending and generating utterances in natural language is inseparably connected to activities mentioned above.

Each of these activities can be considered a separate task. Such tasks need to be performed by a human or a machine using or processing language. Describing objectives of such tasks using mathematical formalism and developing and studying algorithms of carrying them out is the subject of the disciplines of computational linguistics and natural language processing. Currently, due to recent rapid increase in computing power availability and development of appropriate methods, natural language processing usually employs machine learning, especially deep learning [86, 87]. Before the time of machine learning prevalence, language processing systems were usually "rule-based" - they operated on a predefined set of rules designed to capture the structure and relationships in the modeled aspect of language. Since such rules are explicit, they are interpretable, in the sense that one can track how a system using them produces given result. This is unlike many machine learning methods, whose intermediate stages of operation are often unreadable to a human. However, systems of rules can be very complicated - in which case they might not be

easily comprehensible. Nevertheless, constructing rules describing even only a part of language structure can be useful from a theoretical point of view, as it can help in grasping the mechanisms driving various linguistic phenomena.

An important and influential concept in research on language is the idea of characterizing language structures with the use of formal languages, developed by Noam Chomsky in 1950s and 1960s [54, 88]. It treats the ability to use natural language as a certain sort of computational system, which uses a set of rules - defined by an appropriate grammar - to organize individual components into complete utterances. Formal language is a mathematical concept - it is a set whose elements are in a sense constructed from elements of some other set. Let Σ be a finite set. From the elements of Σ one can construct strings - sequences of arbitrary (but finite) length consisting of elements of Σ. Let Σ^* be the set of all such sequences. A formal language is a subset of Σ^*. In other words, it is a subset of the set of all finite strings that can be generated from the elements of Σ.

Specifying which strings in Σ^* belong to a formal language can be done in a few ways. A way that is particularly useful from the viewpoint of study on natural language is using a formal grammar. Formal grammars are mathematical objects providing sets of rules able to generate strings from other strings. Defining a formal grammar can be done in the following way. Let Σ again be a finite set, whose elements are now called terminal symbols. Let N be a finite set, whose elements are called nonterminal symbols, and which is disjoint with Σ. The distinction between terminal and nonterminal symbols is due to their role in strings - terminal symbols are the symbols that are present in the "final form" of a string, while nonterminal symbols are the ones that occur at intermediate stages of string construction. Let S be a distinguished symbol, belonging to N and called the start symbol. Finally, let P be a finite set of production rules; each element of P is a rule of the form: $\alpha \to \beta$, where α and β are strings consisting of terminal and nonterminal symbols; β is an arbitrary string (it can also be an empty string ε), while α must contain at least one nonterminal symbol. A formal grammar G can be defined as a 4-tuple:

$$G = (\Sigma, N, S, P)\,. \tag{1.1}$$

A formal grammar can be interpreted as a string rewriting system, transforming strings into other strings. Starting from a string consisting solely of the start symbol, one can rewrite strings in such a way that each rewriting introduces modifications given by a selected production rule (the notation $\alpha \to \beta$ represents replacing the substring α with the substring β). The set of all strings which contain only terminal symbols and can be constructed from the start symbol by applying some finite sequence of production rules is a formal language; such language is called a language generated by a given grammar.

As an example, let the following grammar be considered: $G = (\Sigma, N, S, P)$, where

$$\begin{aligned} &\Sigma = \{a, b\}\,, \\ &N = \{A, B\}\,, \\ &S = A, \\ &P = \{\ A \to abBaa,\ \ B \to bBaa,\ \ B \to a\ \}\,. \end{aligned} \tag{1.2}$$

Generating strings by this grammar is performed as follows. At the beginning, the string consists of one symbol, the start symbol A. The only rule that can be applied at this stage is the rule $A \to abBaa$, which rewrites A into $abBaa$; therefore, the string becomes $abBaa$. Now, any of the rules $B \to bBaa$ and $B \to a$ can be applied. If the first one is chosen, then B is replaced with $bBaa$ and the string becomes $abbBaaaa$. This new string also allows to apply both of the rules $B \to bBaa$ and

$B \to a$. Using the first one again (one or more times) expands the string and allows for further expansion. Using the rule $B \to a$ at any stage removes B from the string and inserts a; when this happens, no more operations on the string are possible. At this stage the string consists only of terminal symbols, and it can be considered a string belonging to a language generated by the grammar G. All strings in this language have a single a as their first symbol, then b is repeated n times ($n = 1, 2, 3, ...$), and then a is repeated $2n+1$ times; therefore, all such strings are of the form $ab^n a^{2n+1}$, where $n = 1, 2, 3, ...$

Many important properties of formal grammars depend on the constraints imposed on their production rules. Such constraints have an impact on grammar generality. Formal grammars can be divided into types pertaining to that generality. A widely known classification of grammars is the Chomsky hierarchy [89, 90], which distinguishes 4 types of grammars, labeled by numbers 0, 1, 2, and 3. Let α, β be arbitrary strings (possibly empty) of terminal and nonterminal symbols and let γ be a nonempty string of terminal and nonterminal symbols. Let A and B be nonterminal symbols, and let a denote a terminal symbol. The most general form of a production rule is:

$$\gamma \to \beta. \tag{1.3}$$

A grammar which does not have any additional constraints imposed on its production rules, is a type-0 grammar (also called an unrestricted grammar). All languages that can be generated by such a grammar are called recursively enumerable languages. A grammar whose all production rules are of the form:

$$\alpha A \beta \to \alpha \gamma \beta \tag{1.4}$$

is a type-1 grammar, also known under the name of a context-sensitive grammar. To make it possible for a context-sensitive grammar to generate empty strings, one additional rule is allowed: $S \to \varepsilon$, where S is the start symbol, and ε denotes an empty string. Each production rule of a context-sensitive grammar can be interpreted as a procedure transforming a single nonterminal symbol A into a nonempty string, with a condition that such transformation may be dependent on the "neighborhood" of A (the context). Languages generated by a context-sensitive grammar are called context-sensitive languages.

A grammar which only has rules of the form:

$$A \to \alpha \tag{1.5}$$

is called a type-2 grammar, or a context-free grammar. Languages generated by this type of grammar are context-free languages. Left-hand side of any production rule of a context-free grammar is a single nonterminal symbol. The name "context-free" reflects the fact that grammar's rules can be applied regardless of the context of a nonterminal symbol. Context-free grammars are a class of grammars particularly important in modeling and studying language. Their complexity is restricted enough to allow the construction of efficient parsing algorithms - algorithms determining whether a given string of symbols belongs to the language generated by a given grammar, and, if so, finding the sequence of rules leading to the generation of this string. Yet they are general enough to be useful in studying natural language; an important example of their use is syntax analysis. Context-free grammars are also often the backbone of programming languages.

Type-3 grammars, known as regular grammars, can be divided into two groups: left-regular grammars and right-regular grammars. A left-regular grammar is a

grammar having only the rules of one of the following types:

$$\begin{aligned} A &\to a \\ A &\to Ba \\ A &\to \varepsilon \end{aligned} \tag{1.6}$$

where ε denotes the empty string. A right-regular grammar has only the rules of one of the following forms:

$$\begin{aligned} A &\to a \\ A &\to aB \\ A &\to \varepsilon. \end{aligned} \tag{1.7}$$

Regular grammars generate regular languages. They are related to regular expressions - special strings that represent certain patterns. A regular expression specifies a set of strings that match the given pattern. Each language that can be generated by a regular grammar can also be specified by a regular expression and vice versa. Regular expressions are a concise way of describing a set of strings sharing some properties; they have found application in various text processing tools. Contemporary implementations of regular expressions (present, for example, in many programming languages) often extend their basic functionality and make them capable of specifying also the languages other than the ones generated by regular grammars.

Chomsky hierarchy puts the types of grammars into order related to their generality - consecutive types are more restricted than the previous ones. If L_i denotes the set of languages that can be generated by type-i grammars, then:

$$L_3 \subset L_2 \subset L_1 \subset L_0. \tag{1.8}$$

In other words, the set of context-sensitive languages is contained in the set of recursively enumerable languages, the set of context-free languages is contained in the set of context-sensitive languages, and the set of regular languages is contained in the set of context-free languages. All these inclusions are strict inclusions - each L_i contains languages that are not present in L_{i+1}.

There is a close relationship between formal languages and automata theory. For a given formal language, one can define an automaton (an abstract machine) capable of determining whether a given string belongs to that language. To do that, such an automaton (called an acceptor or a recognizer) starts from a starting state, reads the input string - symbol after symbol - in consecutive steps, and changes its internal state in each step. The combination of automaton's current state and the symbol being read determines the transition to the next state (if the automaton is deterministic) or the set of possible transitions (if the automaton is nondeterministic). If the input string is a sequence of symbols for which there exists a sequence of state transitions leading the automaton from the starting state to a predefined final state, then the automaton accepts the string. Otherwise, the string is rejected. The complexity of an acceptor depends on the type of language it is designed to recognize. Regular languages can be recognized by finite-state automata (an example of a finite-state automaton is presented in Fig. 1.2). Context-free languages can be recognized by nondeterministic pushdown automata (which can be thought of as nondeterministic finite-state automata with a stack capable of storing read symbols). Context-sensitive languages are recognized by linear bounded automata. The automaton needed to recognize a recursively enumerable language in a general case is a Turing machine. The more general the grammar, the more powerful automaton is required to recognize languages generated by that grammar.

Formal grammars, automata, and related concepts are widely applied in both theoretical and practical scientific approach to natural language. An important

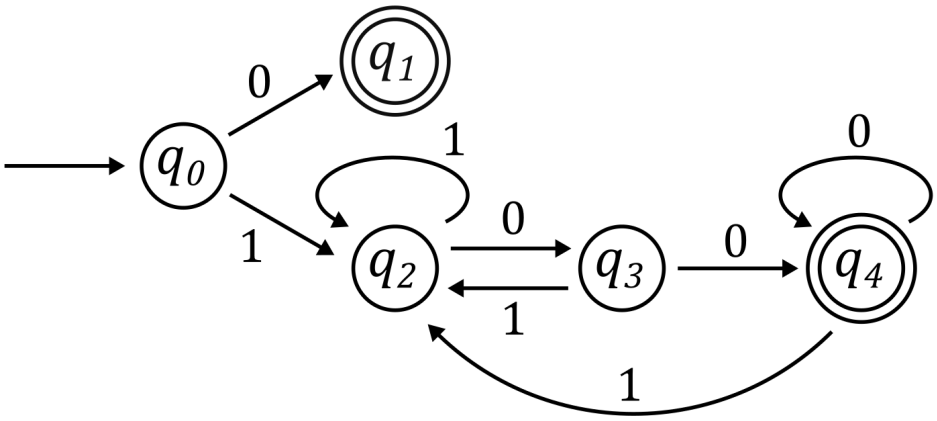


Figure 1.2. An example of a state diagram of a (deterministic) finite-state automaton. The automaton reads strings consisting of zeros and ones and recognizes (accepts) a string if and only if the string is a binary representation of a natural number divisible by 4 (without any padding with leading zeros). States of the automaton are represented by circles (accepting states are marked by a double circle) and arrows correspond to transitions between states. The automaton switches its state to the one pointed by an arrow when the symbol read from the input matches the symbol labeling the arrow. The automaton utilizes the fact that numbers divisible by 4 have at least two zeros at the end of their binary representation. The accepting states are: q_1, which deals with the case when the whole input string is just one zero, and q_4, in which the automaton stays if it reads two or more zeros in a row. The string is recognized if its reading is finished and the automaton is in an accepting state. Any other situation (including, for example, the inability to move to another state from the state q_0 when reading from the input is not finished) leads to the rejection of the string. The regular expression corresponding to the strings recognized by the presented automaton is "0|1+[01]*00".

example is syntactic analysis with the use of the so-called constituency grammars. Constituency grammars are grammars designed to express the syntactic structure by relations between constituents. A constituent is a word or a group of words that can be treated as a single unit in a larger grammatical construction. Constituents of the same type appear in similar syntactic environments, and can be treated as interchangeable - to some extent - form purely syntactic point of view. For example, in the sentence *"The glass that was on the table fell on the floor"*, all of the phrases: *the glass*, *the table*, *the floor*, as well as *the glass that was on the table* can be treated as noun phrases - constituents which perform the grammatical function of a noun. Swapping these phrases between one another could lead to a sentence which is semantically nonsensical, but syntactically correct. Constituency structure is hierarchical - constituents might consist of other constituents, all the way down to individual words. Assigning a constituency structure to a given sentence is done by constructing a parse tree (also named a derivation tree) - a graph whose structure corresponds to the relationships between constituents. An illustration of the concept of a parse tree is presented in Figure 1.3.

Constituency grammars are used to describe the phrase structure of sentences, therefore they are also called phrase structure grammars. Another relrated name is generative grammar. Generative grammar is a broad term, which serves as a common name to multiple theories; what these theories have in common is the usage of formal grammars to model the grammar of natural language [91]. It is worth noting that phrase structure grammars are only one of possible methods of the analysis of syntax. A line of inquiry often presented as opposing to constituency-based approach is concentrated on the so-called dependency grammars, which instead of relying on constituency relations, employ dependency relations - binary relations between individual words, not groups of words. Each kind of approach has its ad-

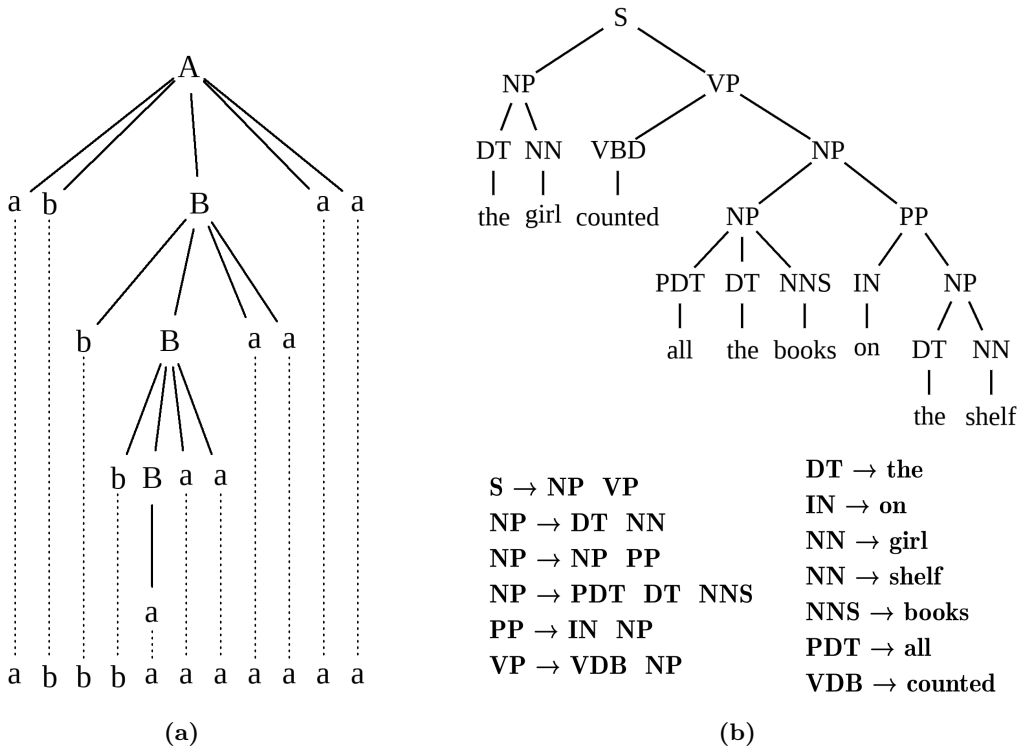


Figure 1.3. Examples of parse trees. The tree in (a) is the parse tree of the string *abbbaaaaaaa*, using the grammar defined in Eq. 1.2. The tree in (b) is the parse tree of the sentence *"The girl counted all the books on the shelf."*, using a context-free grammar modeling the syntax of English. The rules used to construct the tree are listed below the tree; the whole grammar has many more rules. Abbreviations of syntactic categories are as follows: S - start symbol, NP - noun phrase, VP - verb phrase, DT - determiner, NN - noun, VBD - verb in past tense, PP - prepositional phrase, PDT - pre-determiner, NNS - noun in plural form, IN -preposition or subordinating conjunction.

vantages and disadvantages in particular situations, and both are important tools of syntactic analysis.

Formal languages are quite general and abstract mathematical objects; therefore a number of concepts derived from formal language theory have found applications outside mathematics and linguistics. An interesting example are the so-called Lindenmayer systems (L-systems in short) - string rewriting systems, originally invented to model the development and growth of some organisms - one of the first organisms studied in that context were algae [92, 93]. The theory was then applied more generally, to various kinds of branching systems. An L-system G can be defined as [94]:

$$G = (V, \omega, P)\,, \tag{1.9}$$

where V is the set of symbols, ω is a nonempty string of symbols called the axiom, and P is the set of production rules. The definition is very similar to the definition of a formal grammar, with the exception that the distinction between terminal and nonterminal symbols is not necessary and that the start symbol is replaced by the start string - the axiom. L-systems are different from formal grammars in the way in which production rules are applied - rules of formal grammars are applied sequentially (one rule at a time), while L-systems apply their rules in parallel - at each iteration the rewriting is performed in all possible places in the string (in all places where a production rule can be applied). Like formal grammars, L-systems can be divided into types, according to the properties of their production rules. In that sense, an important class of L-systems are context-free L-systems, with all production rules of the form:

$$A \to \alpha, \tag{1.10}$$

where A is a single symbol from V and α is a string of symbols from V. If an L-system is designed to model the growth or the development of a certain object or system, then the assumption that the L-system is context-free can be related to the assumption that individual parts of the modeled object develop independently, without interactions between each other. If for each symbol in V there is exactly one rule which has that symbol on its left-hand side, then the system is deterministic. If any symbol appears as a left-hand side in more than one production rule (which means that it can be rewritten in multiple ways), then the system is called a stochastic L-system. In such a system, each time when multiple rules can be applied to a given symbol, the rule is chosen randomly from the set of possible rules; each rule in that set has some probability of being chosen.

How an L-system works can be illustrated by the following example. Let $G = (V, \omega, P)$, be an L-system where

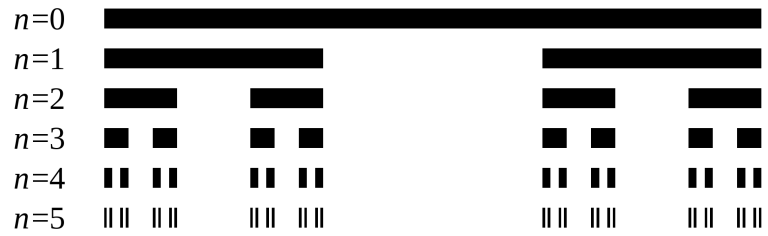
$$\begin{aligned} &V = \{A, B\}, \\ &\omega = A, \\ &P = \{A \to ABA,\ B \to BBB\}. \end{aligned} \tag{1.11}$$

String production in such an L-system proceeds as follows. At the beginning, the string contains one symbol, A. In the first iteration, the rule $A \to ABA$ produces the string ABA. In the second iteration, each A in the string is replaced by ABA, and each B (here only one) is replaced by BBB; hence, the string $ABABBBABA$ is obtained. The process is then repeated in consecutive iterations, up to the point when a predefined number of iterations is reached.

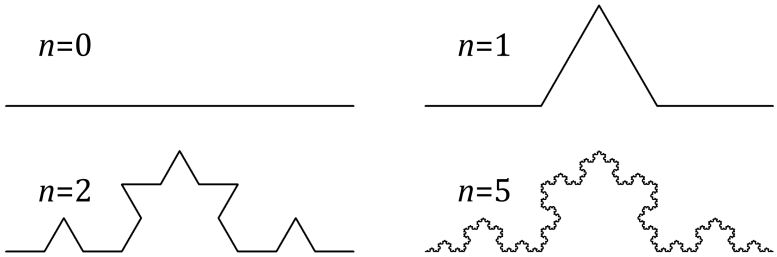
Strings generated by L-systems can be represented graphically, using turtle graphics - a method of creating graphics in which an imaginary object (called the turtle) moves around the drawing area according to a sequence of commands, and the trail left by this object is the desired output. If each symbol in a string generated by an L-system is treated as a command to a drawing device, then a graphical representation of this string can be plotted. Since L-systems are able to generate strings with recursively nested patterns, they are well suited to model self-similar structures. Therefore, they are used to generate fractals (discussed in Chapter 3) and other objects exhibiting similar properties; a noteworthy example of L-systems' application is creating models of plants, used in the field of computer graphics. Examples of images created with the use of L-systems are presented in Figure 1.4.

The fact that similar mathematical objects like formal grammars and L-systems are used to model self-similar, hierarchical systems as well as natural language, is more than a mere coincidence. Multiple aspects of language organization are of strongly hierarchical nature. In fact, the ability to generate multilevel, recursively nested structures is sometimes considered one of defining features of natural language [40]. This is one of the reasons why analytical tools designed to study such structures (like fractal geometry) are useful in research on natural language.

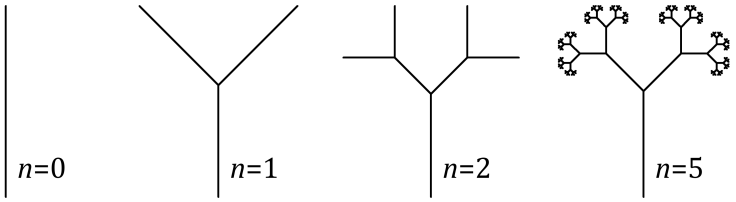
An important part of research on natural language is the subfield called quantitative linguistics, which investigates language using statistical methods. It concentrates on the properties of language which can be described in terms of probability distributions, statistical models, time series and related tools, and attempts to formulate linguistic laws pertaining to those properties. Studying such laws and the reasons of their presence allows to formulate hypotheses about cognitive processes behind language, and about language origin, evolution, and learning. It also has a practical purpose, as the knowledge about statistical patterns in language can be applied in designing natural language processing tools and methods. To propose linguistic laws and to verify them empirically, quantitative linguistics usually uses appropriately large samples of language; in case of written language, such a



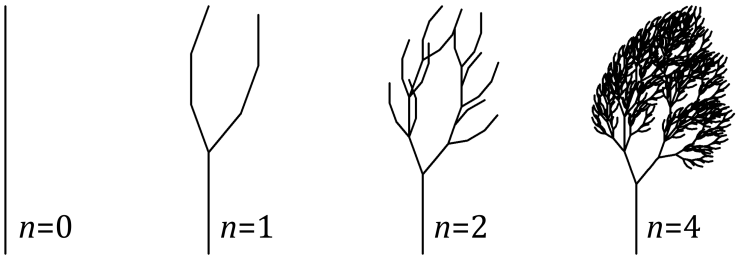
(a) Generating an image of the Cantor set. Symbol set: $V = \{A, B\}$, axiom: $\omega = A$, production rule set: $P = \{A \to ABA,\ B \to BBB\}$ (this is the L-system defined in Eq. 1.11). Commands assigned to symbols: A - move forward a fixed distance and draw a line, B - move forward a fixed distance without drawing.



(b) Generating the Koch curve. Symbol set: $V = \{F, L, R\}$, axiom: $\omega = F$, production rule set: $P = \{F \to FLFRRFLF\}$. Commands assigned to symbols: F - move forward a fixed distance and draw a line, L/R - turn 60 degrees left/right.



(c) Generating a binary tree. Symbol set: $V = \{F, G, L, R, <, >\}$, axiom: $\omega = G$, production rule set: $P = \{G \to F<LG><RG>,\ F \to FF\}$. Commands assigned to symbols: F or G - move forward a fixed distance and draw a line, L/R - turn 45 degrees left/right, $</>$ - push/pop current position and angle onto/from the stack (a LIFO queue allowing to save the state of the plotting device and restore it later).



(d) Generating a tree. Symbol set: $V = \{F, G, L, R, <, >\}$, axiom: $\omega = G$, production rule set: $P = \{G \to FF<LGRGRG><RRGLGLG>,\ \ F \to FF\}$. Commands assigned to symbols: F or G - move forward a fixed distance and draw a line, L/R - turn 20 degrees left/right, $< / >$ - push/pop current position and angle onto/from the stack (a LIFO queue allowing to save the state of the plotting device and restore it later).

Figure 1.4. Examples of images generated with the use of L-systems and turtle graphics. Each subfigure (a,b,c,d) presents several images; each image corresponds to a given number n of string generation iterations ($n = 0$ corresponds to the starting string - the axiom ω). Images within each subfigure are rescaled, so that they all have the same size. Symbol set V, axiom ω and production rule set P of each used L-system are given along with the plotting device operations assigned to each symbol.

sample - called a corpus (plural: corpora) - usually is a collection of texts in one language. The diversity and the number of used texts depends on particular application. A large enough corpus can be treated as representative for the whole studied language. Therefore, observations made with the use of a corpus can be generalized to particular language; analyzing multiple corpora in various languages makes it possible to draw conclusions about laws universal across languages.

Among the most famous linguistic laws is Zipf's law [95–97], which specifies the distribution of word frequencies in texts. Heaps' law (also known as Herdan's law) [98–100] describes how the number of different words in a text varies with the text's length. Menzerath-Altmann law [101, 102] states that the sizes of linguistic constructs are negatively correlated with the sizes of their constituents - for example, the sizes of sentences - measured by the number of clauses - are negatively correlated with the sizes of clauses - measured by the average number of words in a clause. These examples, among others [103–105], are a part of the field of active research, which attempts to observe general statistical patterns in language, describe them with appropriate formalism, and explain their origin. Two examples of widely known linguistic laws - Zipf's law and Heaps' law - are discussed in more detail in Chapter 2.

In this context, it is worth to note the distinction between spoken and written language. Natural languages are primarily spoken; writing is a complement to speech - an invented system of representing spoken language visually. Spoken language is, in some sense, more "natural" - it existed before writing and is acquired (in childhood) without specific instructions, while learning to write requires significant deliberate effort; some languages do not have a written form at all [17, 18]. Speech and writing differ in a number of traits. When writing down a spoken utterance, some information - conveyed by voice modulation, for instance - might be lost. Written language is often more formal than speech. Things like grammatical errors, hesitations, or repetitions are typically present in spoken, not written language. Analysis of one selected representation of language (spoken or written) is in many situations treated as sufficient to draw conclusions about language in general. However, it is important to be aware that some of the effects observed in such analysis may be specific to the chosen representation.

To investigate the properties of natural language utilizing statistical methods, samples of language of considerable size are needed. The type of the samples varies depending on specific area of study (clearly, different tools are needed to study spoken language and written language). For written language, a convenient source of linguistic data are literary texts written in prose. Obviously, it should be remembered that narrative texts constitute a specific type of data and do not account for whole language. However, among forms of language used in literary works, prose is the one which to the greatest extent mimics a natural flow of speech and uses grammatical structures typical for the language used for everyday communication. The fact that the language used in literary texts usually adheres more strictly to grammatical rules and uses more refined vocabulary than colloquial language can be an advantage - as it allows to study structures of certain degree of sophistication and complexity, which are often lacking in everyday language. Many narrative texts have the form of books; books are samples typically large enough to be subject to statistical analysis on their own. These traits of narrative texts make them highly useful in quantitative study of language.

1.3 Complexity and complex systems

Among the most important goals of science, especially physics, is to understand natural phenomena, to explain them using models, and to make predictions based on these models. The models are typically designed in such a way that they can grasp the relevant information about the studied systems, but they also remain as simple as possible, avoiding unnecessary intricacies. Therefore, an essential method of explaining and modeling natural phenomena is treating them as effects of other, more fundamental phenomena. In this view, the characteristics of a system are the result of the interactions between the elements of that system. Such a line of reasoning is known under the name of reductionism. An example which illustrates this kind of approach is the classical theory of electromagnetism. Within a range of scales appropriate for classical physics, the description of all electromagnetic phenomena in any system can be reduced to a few equations - the Maxwell equations. All such phenomena, regardless of how complicated the system is, are the effect of interactions whose complete description is given by a set of "fundamental" physical laws. Therefore the whole classical theory of electromagnetism can be viewed as a set of various applications of these laws. Such an approach, applied to describe various processes in nature, clearly led to a huge number of achievements, both enhancing humanity's understanding of the Universe, and allowing for the development of technology able to solve many practical problems. This is one of the reasons why it has become the dominant paradigm in scientific activity over the last few centuries, in which modern science has developed.

However, there exist systems, called complex systems, for which the description in purely reductionistic manner poses significant difficulties. These are the systems in which the relationship between macroscopic and microscopic properties might not be straightforward and direct. Complexity and complex systems do not have a precise definition; usually a sort of a working definition is used to identify systems whose properties can be attributed to complexity. Such a definition can be stated as follows: a complex system is a system consisting of a large number of nonlinearly interacting elements, which exhibits collective behavior, and, by interacting with its surroundings, is able to modify its internal structure and patterns of activity [106]. A common trait of complex systems is emergence - the presence of phenomena which cannot be reliably deduced or predicted only from the knowledge about the properties of system's constituents and their interactions; it is often summarized with phrases such as "more is different" [107] or "the whole is something beside the parts" [108]. Emergence occurs when interactions between system's elements in microscopic scale give rise to a spontaneous appearance of macroscopic order. This may happen when the interactions inside a system are nonlinear and can be propagated over long distances; in that case, local effects (occurring, for example, due to fluctuations) can be transformed into a collective behavior, depending on the system's interaction with the surroundings. This is possible under specific circumstances, particularly when the system is near the critical state [106, 109]. When a system approaches its critical point, one can observe the divergence of correlation length (a quantity representing typical, characteristic range of correlations between the states of individual elements interacting within the system). This leads to a situation where the maximal range of correlations is limited only by the size of the whole system. Fluctuations might then propagate to arbitrarily long distances and collective behavior can occur on all possible scales. Keeping a system in the vicinity of a critical state is often considered to require some effort to maintain delicate control over external parameters characterizing the environment (temperature, for instance); however, many complex systems seem to be able to evolve spontaneously towards a critical state. It is fre-

quently stated that one of properties characteristic of complex systems is that they operate "at the edge between order and chaos" [109–116].

The nontrivial relationship between the global characteristics of a complex system and the properties of its constituents gives rise to the need of the development of analytical tools designed specifically to study systems belonging to the class of complex systems. The research area concentrated on studying such systems is sometimes considered a new, separate scientific discipline, known under the name of complexity science. It is justified by the prevalence of the characteristics typical for complexity which can be observed in many, sometimes distant and seemingly unrelated systems in nature. Examples of phenomena involving effects related to complexity are: convection [117], phase transitions [109, 110, 118], formation of landforms and coastlines [119–122], organization of the Internet [123–126], population dynamics in ecosystems [127, 128], brain activity [129–138], speculative bubbles and functioning of financial markets [139–148], climate [149–154], epidemics [155–162], and organization of social systems [163–169]. These and many other processes share some aspects of complexity between each other, although not all properties typical for complex systems have to be present in every complex system. Among such properties are the presence of power laws, self-organization, criticality, long-range correlations, fractality, multilevel hierarchical structure, and the presence of complicated organization in network representation.

To be able to asses the complexity of various systems in one unified manner, it would be beneficial to use some kind of quantity able to measure the degree of complexity of an arbitrary system. There have been multiple attempts to construct such a quantity; each of proposed measures has its own rationale, but also has significant drawbacks, limiting its use. An important concept in this context is algorithmic complexity, introduced independently by Solomonoff, Kolomogorov and Chaitin [170–174]. Algorithmic complexity of a string (a sequence of symbols) can be viewed as the length of the description (formulated in some computationally universal language, that is, a language in which any Turing machine can be implemented) of the shortest possible algorithm generating that string. This quantity, although important in the fields of information theory and computability theory, has limited practical use due to its uncomputability for arbitrary sequences [175] and due to the fact that it loses its functionality when dealing with random data. The latter stems from the observation that to specify a truly random sequence one needs to give it explicitly - as there are no regularities to exploit - and therefore the description of the relevant algorithm has the length comparable to the sequence itself. A quantity related to algorithmic complexity is the so-called effective complexity [176, 177]. To avoid treating random sequences as complex, it is designed to measure only the complexity of the non-random contribution to a sequence. However, determining the extent to which a string is random involves some degree of arbitrariness [178, 179].

Another approach to quantifying complexity utilizes the notion of logical depth - which is also related to algorithmic complexity. Logical depth of a string can be interpreted as the time needed by a universal computer (a device capable of computing what can be computed by a Turing machine) to execute the algorithm which generates the string and has description as short as possible [179, 180]. According to this idea, complex objects (logically deep objects) are the objects which require large computational effort to be generated. It reflects the intuition that complex structures are often created by complicated processes; it also treats random data as relatively "shallow".

A concept originating in physics and often employed in quantifying complexity, is information entropy (Shannon entropy [181, 182]). Information entropy is a quantity measuring the level of uncertainty of a random variable. For a discrete random

variable which has n possible values x_i occurring with probabilities p_i ($i = 1, 2, ..., n$), information entropy is defined as:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i. \tag{1.12}$$

Depending of the choice of logarithm base in the above definition, entropy can be expressed in various units; if the logarithm base is 2, then entropy is given in bits; choosing the base e gives entropy in nats ("natural units"). Information entropy can be considered a generalization of the notion of entropy known from statistical physics. If in Equation 1.12 all probabilities p_i are equal ($p_i = 1/n$), then the equation simplifies to $H = \log_2 n$. If n denotes the number of possible microstates which can yield a given macrostate of some system, then after changing the logarithm base and introducing a multiplicative constant k_B (the Boltzmann constant), the formula can be recognized as Boltzmann's definition of entropy:

$$H = k_B \ln n. \tag{1.13}$$

For a discrete random variable, the lowest possible value of information entropy is 0; it is attained when the variable has only one possible value (other values either do not exist or have zero probability; for values with zero probability the product $p_i \log_2 p_i$ is assigned the value 0, in accordance with the limit: $\lim_{x\to 0^+} x \log_2 x = 0$). Information entropy of a random variable with a fixed number of possible values is maximized when probability is uniformly distributed over those values. Therefore entropy is interpreted as the degree of uncertainty or randomness inherent in a random variable. It can also be treated as an average amount of information contained in a single measurement of a quantity described by the considered variable. This view can be presented as follows: if some system can be in one of n states, and the probability is distributed approximately uniformly among the states - which means that the entropy of the system is high - then a single measurement revealing the system's state gives much information - because it would be difficult to make a correct assumption about the state before the measurement. Conversely, when the distribution of probability among states is highly nonuniform and the entropy is low, a measurement is not very informative - as it typically leads to an expected result - that the system is in one of the states of high probability.

The usefulness of information entropy as a direct measure of complexity is limited due to the fact that it attains the highest values for systems with the highest degree of randomness - and systems organized in a purely random fashion (random strings of symbols, for instance) cannot be considered complex. However, many methods aiming to quantify complexity employ entropy and related concepts. An example of a quantity rooted in statistical physics and intended to measure complexity is thermodynamic depth [183], which can be considered a physical counterpart of logical depth. It is based on the assumption that complex systems are the systems which are difficult to assemble or create. In that view, complexity is measured by the amount of information required to specify the trajectory (a history of system's past states) that the system followed to arrive at its present state. It can be expressed as the entropy of the distribution of trajectories leading to system's current state. Although it is in agreement with the intuitive comprehension of complexity (being a product of a complicated process), thermodynamic depth encounters serious problems with its practical application. One limitation is the fact that the knowledge about the whole history of a system is usually unavailable; another difficulty is arbitrariness involved in determining the trajectory followed by the system [114, 184].

Another measure of complexity, designed to study symbolic sequences and objects that can be described by such sequences, based on identifying repeating patterns, is Lempel-Ziv complexity [185]. There exist several definitions of Lempel-Ziv

complexity, but all rely on the same idea - iterative processing of the string and identifying patterns which are copies of patterns encountered at earlier stages. This became a backbone of the Lempel-Ziv algorithm - a lossless data compression algorithm, existing in multiple variants (LZ77, LZ78, LZW, and others [186–188]), and being of huge importance for computer science and practical applications of information theory [189]. The key part of the method proposed by Lempel and Ziv can be briefly presented as an appropriate string partition procedure. One of its variants relies on dividing a string S into substrings $S_1, S_2, ..., S_N$, called *phrases*, such that their concatenation is equal to S and that each consecutive phrase S_i is the shortest possible phrase different from each of the phrases $S_1, S_2, ..., S_{i-1}$ (except for the last one, S_N, which might not be unique). For example, according to that procedure, the string $AABABBBABAABABBBABBABBA$ is divided into $A|AB|ABB|B|ABA|ABAB|BB|ABBA|BB|A$ (vertical lines separate consecutive phrases $S_1, S_2, ..., S_N$). For a string containing many repeating substrings, the number of unique phrases grows with string's length more slowly than in case of a string in which symbol sequences are rarely repeated. Hence, string complexity can be measured in terms of the number of unique phrases. Using the presented scheme to compress the string relies on the observation that each of the consecutive phrases of length greater than 1 is a copy of some of the previously encountered phrases, concatenated with a single symbol (the symbol determining that the phrase is distinct from each of the previously encountered phrases). Therefore, instead of specifying the phrase explicitly, one can specify how it can be constructed from previous phrases: if S_j is a phrase which can be obtained by appending a single symbol to a phrase S_i already encountered in the string S, then S_j can be described by 3 parameters: the position in S at which S_i starts, the length of S_i, and the symbol that needs to be appended to S_i to get S_j. For a string of sufficient length, containing many repeated substrings, such an approach allows to represent the string in a significantly more compact form.

Measuring complexity using the idea proposed by Lempel and Ziv has certain advantages - like the fact that Lempel-Ziv complexity can be relatively easily computed for arbitrary strings - but in the context of complex systems it suffers from the same problem as information entropy - it assigns high complexity to random sequences (as they have no systematically repeating patterns) and randomness is different from complexity. In fact, Lempel-Ziv complexity is related to entropy - for example, procedures of identifying repeating patterns in strings, similar to the one presented above, are used in methods of estimating information entropy of symbolic sequences [190, 191].

Complexity can also be understood in relation to how system's parts interact with each other. In this view, the more relationships and connections are present between the individual elements of the system, the higher the system complexity. The occurrence of such connections can manifest itself by statistical dependence of variables representing the states of system's constituents. There exist a number of tools designed to measure dependence of this kind. A basic example is correlation function - which expresses how, on average, the values of variables pertaining to individual states of system's constituents are related to each other, depending on the distance (in space or time) between those constituents. Another example is mutual information, which is the difference between the sum of entropies $H(X)+H(Y)$ and the joint entropy $H(X,Y)$ of two variables X and Y. If the mutual information $I(X,Y) = H(X) + H(Y) - H(X,Y)$ is greater than zero, it means that observing the value of one variable allows to reduce the "uncertainty" of the other; in other words, some information is shared by X and Y, and they are mutually dependent. Statistical dependence and correlations, especially those of long-range and nonlinear

character, are common traits of complex systems; however, in some cases, their presence might be due to reasons more obvious and straightforward than complexity [106, 192]. For example, a multi-component system in which all components are in the same state and evolve in the same way, does exhibit very strong internal correlations, but such a system is not considered complex. Therefore, measuring complexity solely on the basis of correlations' strength is not a method that could be reliably applied to all types of systems.

Finally, complexity can be investigated using tools designed to study fractals and multifractals. Many systems in nature have hierarchical, multilevel organization which is self-similar or statistically self-similar. This means that some statistical properties specific to higher levels of organization are identical with the properties corresponding to lower levels; a structure exhibiting such behavior is sometimes called *scale-free*, due to the fact that it has similar characteristics regardless of the scale at which it is inspected. Such systems are often conveniently described with the use of fractal geometry [193–196]. Certain properties of fractals, like "rough", irregular shape, difficult to describe with the standard approach based on Euclidean geometry, or the presence of recursively nested patterns, can be intuitively interpreted as signs of complexity. Fractal geometry allows to identify those properties and to characterize them quantitatively. From that perspective, the most complex objects are the so-called multifractals, which can be thought of as systems consisting of many different fractals. However, despite the fact that fractality is quite abundant in nature, it is not necessarily present in all types of complex systems; there also exist systems in which fractality, although present, may be difficult to detect. Therefore the possibility of measuring complexity by identifying fractals and multifractals in system's structure or its characteristics is restricted to only a certain subclass of complex systems.

The methods of quantifying complexity presented above certainly do not exhaust all the ways in which complexity can be expressed or measured [197]. Instead, they show that virtually any kind of approach to measuring complexity has its limitations, either conceptual or practical. Due to the huge diversity of complex systems, each of the proposed methods, usually designed to specific class of systems or signals (symbol sequences, thermodynamic systems, geometrical objects, etc.), is insufficient to reliably characterize systems in which complexity is understood differently. Therefore, when studying complex systems, rather than applying one unified approach, one usually investigates a number of characteristics related to complexity. It is not necessary that all of such characteristics occur within one system; having only some of them is typically sufficient to identify the system as complex.

1.4 Aspects of language complexity

Natural language is clearly an example of a complex system. The properties of its multilevel, hierarchical structure can be considered as displaying emergence in multiple aspects. Higher levels of its organization usually cannot be simply reduced to the sum of the elements involved. For example, phonemes or letters basically do not have any meaning, but the words consisting of them are references to specific objects and concepts. Likewise, knowing the meaning of separate words does not necessarily provide the understanding of a sentence composed of them, as a sentence can carry additional information, like an emotional load or a metaphorical message. And the meaning of a sentence can be fully understood when analyzed

in an appropriate context, constituted by other sentences. The presence of many different types of relationships between various structures in language, each typical for a specific level of language organization, extends to higher and higher levels; for written language, examples of such levels are paragraphs, chapters or whole books. The complexity of language's hierarchical structure is reflected in the number of the academic disciplines being involved in research on natural language. The lowest levels of language organization are studied by biology and physiology, the higher ones - by linguistics and its various sub-fields, and the highest - by sociology, psychology and literary studies.

Language reveals its complexity also when the system of rules and laws governing the relationship between various elements of language is considered. On the one hand, the rules of grammar have to be precise enough to allow for generating utterances which can be understood by other language users; on the other hand, there is some degree of freedom in constructing an utterance - the rules allow for new forms and can also evolve over time. A large part of linguistic structures cannot be characterized by simple rules that are not subject to exceptions. Therefore language can be considered a system displaying a certain form of balance between regularity and irregularity. This is one of the reasons which render the description of natural language difficult and which make studying language from various perspectives particularly valuable.

Another perspective on language complexity is related to how language changes over time. The conditions that language has to satisfy (to remain an effective communication tool, for instance) and the way in which it evolves indicate that language is subject to self-organization. Self-organization of language is typically studied from two perspectives [198]. One perspective considers language as a system of thought expression individual for each human, and focuses on the evolution of that system towards its optimization with respect to the ease of its use and acquisition [199–201]. The other approach treats language as a system of communication between individuals belonging to some population. In this view, self-organization is a process driven by mutual interactions between language users, leading to continual language change and adaptation to changing conditions [202–205]. It is important to point out that language evolution is extremely difficult to describe quantitatively, as it is itself a complex process, driven by many factors, like the evolution of language users (humans), the influence that language users have on each other and the interactions they have with the environment.

Among traits typical for complex systems which are also observed in natural language, one can mention the presence of power laws and the most well-known example of such a law - Zipf's law, describing the distribution of word frequencies in texts. When language is treated as a signal, one can usually observe long-range correlations and scale-free fluctuations, also described by power laws; this corresponds to fractal or multifractal structure, being another sign of complexity. Studying various properties of language with the use of network analysis reveals complicated patterns of organization, some of them being typical for networks representing other complex systems. Insights from studying language with the use of power-law distributions, time series analysis, and complex network formalism are presented in subsequent chapters.

Chapter 2

Power laws

2.1 Basic properties of power laws

In many complex systems the distributions of certain quantities describing system's structure or behavior are given by power laws. This property is exhibited by a great variety of systems, including physical, biological, economic, and social ones [206–208]. A quantity x is distributed according to a *power-law distribution* (also called shortly just a *power law*, when the context is clear), if its probability density function (for a continuous variable) or its probability mass function (for a discrete variable) is of the form:

$$p(x) = Cx^{-\beta}, \tag{2.1}$$

where C is a normalization constant and $\beta > 0$. It is assumed that x is bounded from below by some positive constant x_{min}, being the lowest possible value of x, as for $x \to 0$ the function $x^{-\beta}$ diverges. It is common that the distribution of the quantity of interest adheres to a power law only in the tail - in such case x_{min} is a threshold above which the analysis of power-law behavior of x is relevant. The power law given above does not necessarily have to be obeyed exactly - asymptotic agreement is usually considered sufficient. If the support of the distribution is bounded from above, then β can be any positive number. However, in many typically encountered situations, the support is right-unbounded (x can take arbitrarily large values), and this is the case considered here. Then β has to be greater than 1, to allow for proper normalization. When the support is right-unbounded and x is discrete, that is, when $x \in \{x_{min}, x_{min}+1, x_{min}+2, ...\}$, the normalization is given by:

$$1 = \sum_{k=x_{min}}^{+\infty} Ck^{-\beta}. \tag{2.2}$$

For continuous x, the normalization is:

$$1 = \int_{x_{min}}^{+\infty} Cx^{-\beta}dx = \frac{C}{1-\beta}\Big[x^{-\beta+1}\Big]_{x=x_{min}}^{x=+\infty} \tag{2.3}$$

Both the series in Eq. 2.2 and the integral in Eq. 2.3 are convergent only when $\beta > 1$, hence the restriction of the possible values of β.

Since it is often the tail of the distribution that is under consideration when studying power laws, it is convenient to express a power-law distribution in terms of its *complementary cumulative distribution function* $\overline{F}$ (also called *survival function* or *tail distribution*). For a random variable X the survival function $\overline{F}$ can be defined as:

$$\overline{F}(x) = P(X \geq x), \tag{2.4}$$

where $P(X \geq x)$ denotes the probability that X takes on a value greater than or equal to x. Equivalently, $\overline{F}$ can be expressed as:

$$\overline{F}(x) = 1 - F(x), \tag{2.5}$$

where F is the cumulative distribution function. Depending on how exactly F is defined (as right-continuous or as left-continuous), the inequality in Eq. 2.4 can be strict or not (this distinction is important only for discrete distributions); here it is assumed that $\overline{F}(x) = P(X \geq x)$. Like cumulative distribution function, survival function fully specifies the studied distribution.

The survival function of a power-law distribution is a power function. For a continuous variable it is given by:

$$\overline{F}(x) = \int_x^{+\infty} Ct^{-\beta} dt = \frac{C}{1-\beta}\, x^{-\beta+1}. \tag{2.6}$$

For a discrete variable, the survival function is:

$$\overline{F}(x) = \sum_{k=x}^{+\infty} Ck^{-\beta}. \tag{2.7}$$

Although the sum above does not follow a power law exactly, it can be approximated for large x (using, for example, Euler's summation formula [209,210]) by an integral:

$$\sum_{k=x}^{+\infty} Ck^{-\beta} \approx \int_x^{+\infty} Ct^{-\beta} dt. \tag{2.8}$$

Therefore, it can be stated that for sufficiently large x both continuous and discrete power-law distributions have survival functions behaving like power functions:

$$\overline{F}(x) \propto x^{-\beta+1}. \tag{2.9}$$

One can introduce the notation: $\overline{F}(x) \propto x^{-\alpha}$, where $\alpha = \beta - 1$; both α and β can be called the *exponents* of a power law, depending on the context. Since power-law-like behavior of probability density function or probability mass function is closely related to the same type of behavior of survival function, the identification of a power-law distribution can be performed by observing that any of the mentioned functions is a power function. And due to the fact that certain sums and integrals can be asymptotically approximated by one another (like in Eq. 2.8), many characteristics of continuous power-law distributions are valid also for their discrete counterparts (calculations for one variant of the distribution might be much more tractable than for the other, however). For that reason, from now on, the presented properties of power-law distributions are given for their continuous variants.

Power laws belong to the class of the so-called *heavy-tailed distributions*. A distribution with survival function $\overline{F}$ has a (right) heavy tail [211], when for any $\lambda > 0$:

$$\limsup_{x \to \infty} \frac{\overline{F}(x)}{e^{-\lambda x}} = \infty, \tag{2.10}$$

that is, a distribution has a heavy tail when for $x \to \infty$ its survival function decays slower than any decreasing exponential function. There exists an important subclass of heavy-tailed distributions - the so-called *subexponential distributions*; most of typically encountered heavy-tailed distributions belong to the class of subexponential

distributions. All power-law distributions are subexponential. A distribution is subexponential [211, 212] when

$$\lim_{x\to\infty} \frac{\overline{F*F}(x)}{\overline{F}(x)} = 2. \tag{2.11}$$

In the above formula, $F*F$ is the convolution of the cumulative distribution function F with itself - which corresponds to the cumulative distribution function of the sum of two independent random variables distributed according to F. Consequently, $\overline{F*F}$ is the survival function of such a sum. Subexponentiality is equivalent to the following property. Let $X_1, X_2, ..., X_n$ be independent, identically distributed random variables with a subexponential distribution. Then

$$\lim_{x\to\infty} \frac{P\left(X_1 + X_2 + ... + X_n \geq x\right)}{P\left(\max(X_1, X_2, ..., X_n) \geq x\right)} = 1, \tag{2.12}$$

where $P\left(X_1 + X_2 + ... + X_n \geq x\right)$ denotes the probability that the value of the sum $X_1 + X_2 + ... + X_n$ is greater or equal to x and $P\left(\max(X_1, X_2, ..., X_n) \geq x\right)$ denotes the probability that the largest value among $X_1, X_2, ..., X_n$ is greater or equal to x. Equation 2.12 expresses the fact that for large enough x, the sum of values drawn independently from a subexponential distribution exceeds x with practically the same probability as the largest of those values does. In that sense, the behavior of the sum is to a large extent "determined" by the behavior of the largest value; this is known as the *single big jump principle*, and is a substantial characteristic of processes described by subexponential distributions [213, 214].

Power laws are often characterized as distributions which can span over several orders of magnitude, in contrast to, for example, normal distribution or exponential distribution, for which one usually can identify a typical range of values or a characteristic scale. Whether a distribution covers multiple orders of magnitude depends on the distribution's parameters and on the units in which the studied quantity is measured, but power laws found in many systems in nature are indeed associated with quantities considered to have a wide range of possible values. This is often related to the fact that the m-th raw moment of a power-law distribution:

$$\langle x^m \rangle = \int_{x_{min}}^{+\infty} C x^{-\beta+m} dx = \frac{C}{m+1-\beta} \Big[x^{-\beta+m+1} \Big]_{x=x_{min}}^{x=+\infty} \tag{2.13}$$

is finite only for $\beta > m+1$. This implies that the expected value $\langle x \rangle$ exists only for $\beta > 2$ and the variance $\langle x^2 \rangle - \langle x \rangle^2$ only for $\beta > 3$. For that reason, for distributions with $\beta \leq 3$, which are quite common in nature, there is no finite expected value, or - if it exists - the average squared deviation from the expected value is infinite. A related effect, attributed to power laws with appropriately low exponents, is the high degree of "non-uniformity". When some quantity, which can be interpreted as a certain kind of resource, is distributed over some population according to such a power law, then a large fraction of the overall amount of the resource is concentrated within a small fraction of the population. In some areas this phenomenon has been called *Pareto principle* or *80-20 rule*; the former name refers to Vilfredo Pareto - an economist who pioneered in using power laws to represent wealth distribution in society [215], the latter describes situations where 20% of some population holds 80% of some resource [216]. It should be noted, however, that exact numbers expressing that effect may vary; the relationship 80%-20% is obtained for a specific value of power law exponent β. In a continuous distribution with probability density function $p(x)$ of the form

$$p(x) = (\beta - 1) x_{min}^{\beta-1} \, x^{-\beta}, \tag{2.14}$$

which is known as the Pareto distribution, β has to be equal to $1 + \log_4 5 \approx 2.16$ to comply to 80-20 rule precisely.

An important property of power laws is scale invariance. For a power function

$$f(x) = Cx^{-\beta} \tag{2.15}$$

and a positive constant λ, the following condition is satisfied:

$$f(\lambda x) = C(\lambda x)^{-\beta} = \lambda^{-\beta} f(x), \tag{2.16}$$

which means that scaling the argument of the function by a constant λ results in scaling the value of the function by the constant $\lambda^{-\beta}$. Therefore, a function of that type does not have any characteristic scale - its properties are qualitatively the same in all possible scales. For that reason, all power functions with a particular exponent are in a sense equivalent, since they differ from each other only by a multiplicative constant. Scale invariance of power-law probability distributions can be interpreted as the presence of a certain kind of hierarchy - for a power-law distribution with probability density function $p(x) = Cx^{-\beta}$ and any x_1, x_2 contained in the interval in which the power-law relationship is valid, the densities $p(x_1)$, $p(x_2)$ are bound by:

$$\frac{p(x_2)}{p(x_1)} = \left(\frac{x_2}{x_1}\right)^{-\beta}. \tag{2.17}$$

Identifying power-law distributions in empirical data usually employs the fact that a relationship described by a power function $f(x) = Cx^{-\beta}$ can be transformed into a linear relationship, by taking the logarithm of both sides:

$$\log(f(x)) = \log C - \beta \log x. \tag{2.18}$$

Therefore, when $f(x)$ is presented on a log-log plot (which might be a plot in log-log scale or a graph of $\log(f(x))$ vs. $\log x$), observing a linear relationship allows to conclude that f is a power function of x. The exponent can be determined from the slope of the line.

To investigate if a sample comes from a power-law distribution, one can compute the empirical survival function $\widetilde{\overline{F}}$, defined as:

$$\widetilde{\overline{F}}(x) = \frac{N_{[x;\infty)}}{N}, \tag{2.19}$$

where $N_{[x;\infty)}$ is the number of observations greater or equal to x in the sample, and N is the total number of observations; $\widetilde{\overline{F}}$ is a step function, with steps at points corresponding to unique values in the sample, therefore it is usually computed only for x being such values. After $\widetilde{\overline{F}}$ is determined, one can plot the set of points $(\log x, \log \widetilde{\overline{F}}(x))$. If the points lie on a straight line for some range of x, then within that range $\widetilde{\overline{F}}$ is a power function of x, and since $\widetilde{\overline{F}}$ approximates the survival function of the underlying distribution, the distribution can be recognized as a power law. The exponent α, describing the behavior of the survival function ($\overline{F} \propto x^{-\alpha}$), which is usually of primary interest, can be obtained by determining the slope of the observed line. The slope is very often computed by using the least squares method to fit a linear relationship to $\widetilde{\overline{F}}(x)$. The advantage of this approach is simplicity; maximum likelihood estimation of α is superior in terms of accuracy and error estimation, but requires choosing carefully the range in which the power-law relationship holds and, depending on the type of the data, it might involve solving a transcendental equation [217].

A tool similar to survival function plots and often used in identifying power-law distributions in empirical data, is the so-called rank-size distribution. If some collection of values is presented in the form of ranking - that is, a list in which the first element is the largest observation, the second element is the second largest observation, and so on, then the rank-size distribution is the function which relates the value with its position in the ranking. If this function is a power function, the underlying probability distribution is a power-law distribution. This can be understood with the following line of reasoning. Let the values in a N-element sample drawn from some distribution be sorted into a non-increasing sequence $(x_1, x_2, ..., x_N)$. The rank $R(x_k)$ of an observation x_k $(k = 1, 2, ..., N)$ can be defined in a few ways, two possibilities are considered here:

$$(1) \quad R(x_k) = k \tag{2.20}$$

$$(2) \quad R(x_k) = \max\{j : x_j = x_k\} \tag{2.21}$$

In the first variant, the rank R of the observation x_k is the position of x_k in the ranking; in the second one, R is the number of observations greater or equal to x_k in the sample. They differ only when the values in the sample might repeat (which often happens for data coming from a discrete distribution); however, even when they do, the differences typically appear for large R, and it is the range of small R that is usually of interest, as it corresponds to the tail of the probability distribution. Therefore the distinction between definitions in Eq. 2.20 and Eq. 2.21 is rarely significant in practical calculations, but while the first is often used in the literature, the second one is more suited to the derivation of the formulas given here. With rank defined as in Eq. 2.21, one can relate each unique value x in the sample with its rank R; the function $x(R)$ is called the rank-size distribution (sometimes also the rank-frequency distribution, if the data represents the frequencies or counts). Stating that x has rank R is equivalent to stating that exactly R observations in the sample are larger or equal to x. It means that R/N is an estimate of $P(X \geq x)$ - the probability that a random variable X with the considered probability distribution takes on a value greater or equal to x. That is, R/N is equal to the value of the empirical survival function at point x:

$$\frac{R}{N} = \widetilde{\overline{F}}(x). \tag{2.22}$$

Therefore, rank-size distribution contains information sufficient to fully characterize a sample, same as empirical survival function. If the rank-size distribution $x(R)$ is a power law:

$$x \propto R^{-\gamma} \quad \text{for some } \gamma > 0, \tag{2.23}$$

then by raising both sides of that relationship to the power $-1/\gamma$, the inverse function $R(x)$ is obtained:

$$R \propto x^{-1/\gamma}. \tag{2.24}$$

Since N is a constant for a given sample, R/N behaves in the same way as R with respect to x, that is: $R/N \propto x^{-1/\gamma}$; using Eq. 2.22, one gets:

$$\widetilde{\overline{F}}(x) \propto x^{-1/\gamma}. \tag{2.25}$$

This shows that a power law in the rank-size distribution $x(R) \propto R^{-\gamma}$ corresponds to a power-law form of the empirical survival function $\widetilde{\overline{F}}(x) \propto x^{-\alpha}$, and the exponents α and γ are related by:

$$\alpha = \frac{1}{\gamma}. \tag{2.26}$$

Therefore, a power law in the rank-size distribution of some sample indicates that the sample is drawn from a power-law probability distribution.

It is important to note that the above-presented methods of identifying power laws based on studying the behavior of survival function or of rank-size distribution pertain to situations where the probability distribution function (or probability mass function) is a power function with the exponent β greater than 1. The formulas describing the functional form of the survival function and of the rank-size distribution cease to be valid for $\beta \leq 1$. For distributions with β below or close to 1, other methods of detecting the power-law relationship should be used; it should be pointed out that such distributions have to have a bounded support or a cutoff at some point to be normalizable. For example, one can use a histogram as a piecewise constant approximation of probability density function or of probability mass function, and study its behavior in log-log scale. Since power-law distributions, especially those with low exponents, typically span a wide range of values, it is often beneficial to construct the histograms of power laws with the use of bins of varying length.

2.2 Generating power laws

Transforming an exponentially distributed random variable

The abundance of power laws in nature raises a question about their origin. It turns out that they can be generated in multiple types of processes; selected examples are presented here. One of situations in which power laws appear is when some quantity is an exponential function of another, exponentially distributed quantity. Let X be a positive random variable with probability density function $p_X(x)$ being an exponential function:

$$p_X(x) \propto e^{ax}. \tag{2.27}$$

with some $a \neq 0$. If $a > 0$ then the distribution's support must have a finite upper bound, so that it can be normalized. Let Y be a random variable being an exponential function of X:

$$Y \propto e^{bX} \tag{2.28}$$

with $b \neq 0$. This means that if X takes on a value x, then Y takes on a value determined by the function $y(x) \propto e^{bx}$. The probability density function $p_Y(y)$ of Y can be expressed as:

$$p_Y(y) = p_X(x(y)) \left| \frac{d\,x(y)}{d\,y} \right|, \tag{2.29}$$

where $x(y)$ denotes the inverse function of $y(x)$. Based on Eq. 2.28, the behavior of $x(y)$ is given by:

$$x(y) \propto \frac{1}{b} \log(y); \tag{2.30}$$

where $\log(\cdot)$ denotes the natural logarithm. Substituting the relationship of this form to Eq. 2.29, one gets:

$$p_Y(y) \propto e^{(a/b)\log y} \left| \frac{1}{by} \right|. \tag{2.31}$$

Since y is positive, $|by| = |b|y$ and $p_Y(y)$ satisfies

$$p_Y(y) \propto y^{(a/b)-1}. \tag{2.32}$$

Therefore, the probability density function of Y is a power function:

$$p_Y(y) \propto y^{-\beta} \tag{2.33}$$

with exponent $\beta = 1 - (a/b)$. If $a/b < 0$ then $\beta > 1$ and the distribution can be supported on the interval $(0, +\infty)$; otherwise the support must have a finite upper bound to allow normalization.

Yule processes

Another mechanism of generating power laws are stochastic processes known under the names of *Yule processes*, *Yule-Simon processes*, or (especially in the context of complex networks) *preferential attachment*. Originally conceived for biological systems - to model the emergence of power-law distributions describing the number of species in genera, or more generally, the number of subtaxa in taxa - they have found application in many other areas [206, 218–221]. A Yule process models the behavior of a system composed of a collection of objects which have a certain positive quantity assigned to them, when both the number of objects and the total sum of the studied quantity in the system grow in consecutive time steps in a specific way. For illustrative purposes it is convenient to imagine the considered system as a collection of boxes, with balls inside them. Then one of the forms of the Yule process can be described as follows.

At each point in time, the system consists of a certain number of boxes; the i-th box has k_i balls in it. A single time step of the process starts from adding $m > 0$ new balls to the system and distributing them among boxes in the following way: m boxes are chosen randomly from the system, and one ball is added to each of them; the probability P_i of choosing a particular box is an increasing linear function of the number k_i of balls already present in that box:

$$P_i \propto (k_i + c), \tag{2.34}$$

where c is a real constant. After inserting balls into the boxes, one new box with $K_0 \geq 0$ balls inside is added to the system; as a consequence, the number of boxes present in the system increases by 1. The time step ends here; in the next step the presented procedure is repeated. The constants K_0 and c have to satisfy the condition $K_0 + c > 0$, which ensures that $k_i + c$ is positive for any possible k_i, because $k_i \geq K_0$ for all i.

One of the roles of the constant c in Eq. 2.34 is to allow the boxes added to the system to participate in the process of distributing new balls among boxes when $K_0 = 0$. More generally, it allows to make the relationship between P_i and k_i more flexible. If $c = 0$, then the probability of choosing a box is just directly proportional to the number of balls in that box: $P_i \propto k_i$.

Having defined what happens at each time step, the only thing that remains to be specified is the initial state of the system (the number of boxes and the number of balls in each of them at the beginning of the process). The influence of the initial state on the characteristics of the process becomes negligible in the limit of large number of time steps; therefore, the initial state is to some degree arbitrary. Here it is assumed that in the initial state there are at least m boxes and each of them has at least K_0 balls inside.

Based on the characterization given above, it can be concluded that a Yule process in the presented form is controlled by 3 parameters: a positive integer m, a non-negative integer K_0, and a real number c; the parameters c and K_0 must satisfy: $c + K_0 > 0$. Power laws are generated in Yule processes in the limit of large number of steps.

If n denotes the number of performed steps and $p(k)$ denotes the probability mass function of the distribution of the number of balls in a box, that is, the probability that a randomly chosen box has exactly k balls inside, then after large number of steps ($n \to \infty$) the distribution $p(k)$ is a power-law distribution [206]:

$$p(k) \sim C\, k^{-\xi} \quad \text{for } k \to +\infty \text{ and some constant } C. \tag{2.35}$$

An example of such a distribution is presented in Fig. 2.1. The value of the exponent ξ is given by [206]:

$$\xi = 2 + \frac{(K_0 + c)}{m}; \tag{2.36}$$

therefore, by tuning the values of the parameters m, K_0, and c, an arbitrary exponent greater than 2 can be obtained.

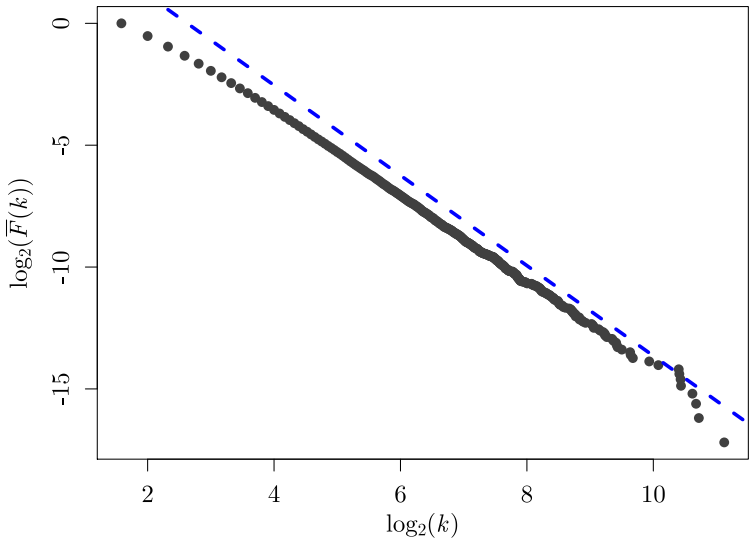


Figure 2.1. Survival function $\overline{F}(k)$ of the distribution of the number of balls in a box generated by a realization of the Yule process with $1.5 \cdot 10^5$ time steps. The parameters of the process are: $m = 5$, $K_0 = 3$, $c = 1.25$. Since both the argument of the function and its value are under logarithm, a straight line shape indicates the presence of a power law. The dashed line has the slope $-\alpha = -1.85$, corresponding to the limiting distribution.

The key property of Yule processes, allowing to generate power laws, is expressed by Eq. 2.34. It is the tendency to put new balls into boxes which already have many balls inside. It can be said that newly added items have the *preference* to be placed where their concentration is already high, hence the name preferential attachment. Other names and phrases used to describe this effect are "cumulative advantage", "rich get richer" or "Matthew effect" [222, 223]; the last name refers to a verse in the biblical Gospel of Matthew: "*For to every one who has will more be given, and he will have abundance; but from him who has not, even what he has will be taken away*". It should be noted, however, that only the first part of the verse correctly describes the considered effect, as Yule processes do not contain mechanisms removing or relocating items already present in the system.

Self-organized criticality

A different perspective on generating power laws is related to the concept of self-organized criticality [207, 224–228]. Self-organized criticality is an idea according to

which some systems naturally, spontaneously evolve in time in the way that keeps them near critical state. Interpreting what criticality exactly implies for a system depends on the system's type, but generally criticality manifests itself in the lack of finite characteristic scale (in space or time), leading to power-law behavior of quantities characterizing the system. In a system in a critical state the response to a small local perturbation can be of any size, limited only by the size of the whole system. The crucial aspect of self-organized criticality is the fact that keeping the system in critical state does not require fine-tuning of external control parameters; the dynamics of the system constantly drive it towards the critical state.

The archetypical model of a system displaying self-organized criticality is the Bak-Tang-Wiesenfeld model [224, 225], also known as the *abelian sandpile* model. A basic version of the model can be defined as a celluar automaton on a D-dimensional hypercubic lattice of linear size L. The cells are identified by their positions on the lattice; the position of a cell is given by a vector $\vec{r} = (r_1, r_2, ..., r_D)$ where $r_i \in \{1, 2, ..., L\}$ for all i. Let $\vec{e_i}$ denote the i-th basis vector of the lattice; this means that $\vec{r}$ can be represented as $\vec{r} = \sum_{i=1}^{D} r_i \vec{e_i}$. For example, for a widely studied two-dimensional version ($D = 2$), one can write: $\vec{r} = (r_1, r_2) = (x, y)$, $\vec{e_1} = \vec{e_x} = (1, 0)$ and $\vec{e_2} = \vec{e_y} = (0, 1)$. Let z be a dynamical variable associated with the system; $z(\vec{r})$ is the value of z in the cell at the position $\vec{r}$. In the one-dimensional version of the model ($D = 1$), the system can be imagined as a certain form of a sandpile (hence the name), and the values of z can be interpreted as the slope at particular positions (the higher the value of z, the steeper the slope). In higher dimensions, the interpretation is less straightforward [225]. The automaton is initiated with some initial distribution of z, such that $z(\vec{r}) \in \{0, 1, 2, ..., z_c - 1\}$ for all $\vec{r}$; z_c is a constant, and here it is assumed that $z_c = 2D$. One possible initial state is $z(\vec{r}) = 0$ for all $\vec{r}$. After initialization, the model evolves in discrete time. At each time step, one position $\vec{r}$ is randomly chosen (each position is chosen with the same probability) and the value of z at $\vec{r}$ is increased by 1:

$$z(\vec{r}) \longrightarrow z(\vec{r}) + 1. \tag{2.37}$$

This action is sometimes given the name of *perturbation*. If after the perturbation the value of $z(\vec{r})$ remains below z_c, the time step ends here. However, if $z(\vec{r}) \geq z_c$, then the cell at $\vec{r}$ becomes *unstable* and the process called *relaxation* or *toppling* takes place. The values of z in the cell at $\vec{r}$ and in the cells in its direct neighborhood (cells at $\vec{r} \pm \vec{e_i}$) are modified according to the following rules:

$$\begin{aligned} &z(\vec{r}) \longrightarrow z(\vec{r}) - 2D, \\ &z(\vec{r} \pm \vec{e_i}) \longrightarrow z(\vec{r} \pm \vec{e_i}) + 1 \quad \text{for } i = 1, 2, ..., D. \end{aligned} \tag{2.38}$$

The above relaxation rules apply to cells which are not at the boundary of the system, that is to cells for which all r_i satisfy $r_i \neq 0 \wedge r_i \neq L$. The behavior of the cells at the boundary is determined by boundary conditions. A basic and straightforward possibility is imposing $z = 0$ at the boundary of the system:

$$z(\vec{r}) = 0 \quad \text{if } r_i = 0 \vee r_i = L \text{ for any } i. \tag{2.39}$$

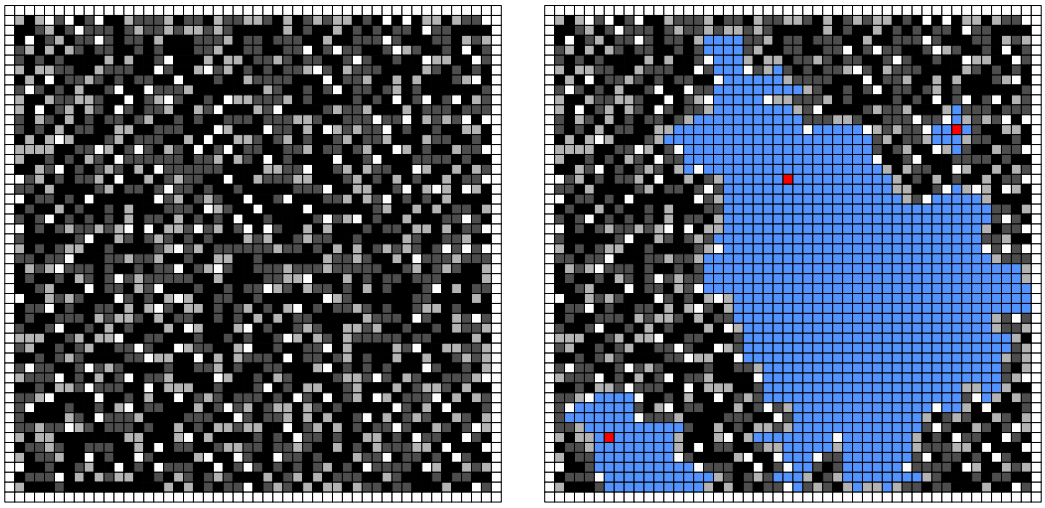
The relaxation of a cell might make neighboring cells unstable; in that case the affected cells also undergo relaxation. When at a given moment there is a certain nonempty set S_1 of unstable cells, all of those cells are relaxed; if this leads to a state of the system in which there is again a nonempty set S_2 of unstable cells, then all cells in S_2 are relaxed, and so on. This process continues until all the cells in the system are stable, that is until $z(\vec{r}) < z_c$ for all $\vec{r}$. When all cells reach stability, the time step is ended. The relaxations of the cells in one particular

configuration of unstable cells S_k can be thought of as being parallel, constituting a global "update" of the system's state. The number of such updates required to get rid of unstable cell states is interpreted as the *duration* of the avalanche. However, in the implementation of the model the relaxations for particular S_k can be performed sequentially; in fact, in the version of the sandpile model presented here, the order in which the relaxations for a particular S_k are performed has no influence on the result of the update (the state the system arrives at); this property is the reason for calling the model *abelian* [229, 230].

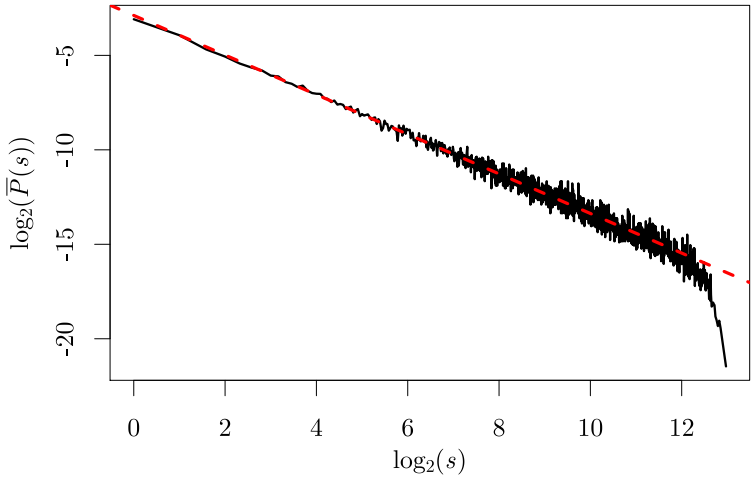
The system is constantly in an nonequilibrium state, as it exchanges the studied quantity z with the environment (the sum of z over the whole system is increased with each perturbation event; at the same time, it decreases with each relaxation taking place in the neighborhood of the system's boundary, since for the cells at the boundary z is fixed at $z = 0$). The key property of the system is the fact that regardless of the initial state, it spontaneously evolves towards a critical state in which a single perturbation can lead to an avalanche of any size, up to the size of the whole system. The distributions of quantities characterizing the system's behavior (like the avalanche's duration or the total number of cells relaxed one or more times during an avalanche) are power-law distributions. Therefore the system exhibits scale invariance (for $L \to \infty$). An illustration of the abelian sandpile model is presented in Fig. 2.2.

An important aspect of the relationship between power laws and self-organized criticality is that systems exhibiting self-organized criticality are able to generate signals in time domain which are characterized by some forms of power laws. In fact, the original motivation behind the sandpile model and the idea of self-organized criticality was to explain the fact that in many different natural systems certain time-varying quantities behave as signals belonging to a specific class of signals, known under the name of "$1/f$ noise" (described in more detail in Chapter 3). It was initially stated that a signal generated by the two-dimensional sandpile model (the changes of the total sum of z induced by consecutive avalanches) is $1/f$ noise [224]. It was soon realized it is in fact an example of a signal of different type, namely $1/f^2$ noise [231–233]. Nevertheless, self-organized criticality became to be considered one of possible mechanisms of generating $1/f$-type signals; some modified variants of the sandpile model do after all generate $1/f$ noise [226, 234, 235].

The presented mechanisms able to generate power laws - a specific form of dependence between random variables, preferential attachment, and self-organized criticality - definitely do not constitute an exhaustive list. They are examples of possible explanations for the emergence of power laws in many systems in nature; the diversity of such explanations gives an opportunity to investigate the properties and the behavior of complex systems from multiple perspectives.



(a) **(b)**



(c)

Figure 2.2. (a) An example of a state obtained in a two-dimensional abelian sandpile model ($D = 2$) of linear size $L = 50$, after 10^5 time steps. Colors mark the values of z in cells: white - 0, light gray - 1, dark gray - 2, black - 3. (b) Examples of relaxation clusters which can be observed when the state presented in (a) is perturbed. When a perturbation causes a relaxation, an avalanche starts; the relaxation cluster is the set of cells which undergo at least one relaxation during the avalanche and the number of such cells is the size of the avalanche. The clusters are colored blue and the cells at which the respective avalanches are initiated are colored red. (c) The distribution of cluster sizes, for a sandpile model defined on a 100×100 lattice. The plot presents $\log_2(\overline{P}(s))$ vs. $\log_2(s)$, where s is the cluster size and $\overline{P}(s)$ is a piecewise constant approximation of the probability mass function. Technically, $\overline{P}(s)$ is determined by a histogram with varying bin width (the width is such that the number of counts in each bin is at least 10) and the line representing $\overline{P}(s)$ is a linear interpolation between points corresponding to bin centers. The dashed line is a line fitted to the data; it has the slope $-\beta = -1.05$.

2.3 Power laws in natural language

2.3.1 Zipf's law and Heaps' law

A fundamental statistical property of natural language, first observed by J. B. Estoup [236], later systematically studied and popularized by G. K. Zipf [95–97] - and therefore called the Zipf's law - is the power-law distribution of word frequencies in texts, or more generally, in linguistic corpora (a corpus is a text or a set of texts put together one after another). Assuming that a "word" is a sequence of letters between whitespace characters, the Zipf's law can be summarized by a statement that the frequency (the number of occurrences) of a word in a text is inversely proportional to the rank of that word, where the rank is the position on the list of all different words appearing in the text, sorted by decreasing frequency. More precisely, if for some text R denotes the rank of a word, and ω is the number of times that word appears in the text, then

$$\omega \propto R^{-\alpha}, \tag{2.40}$$

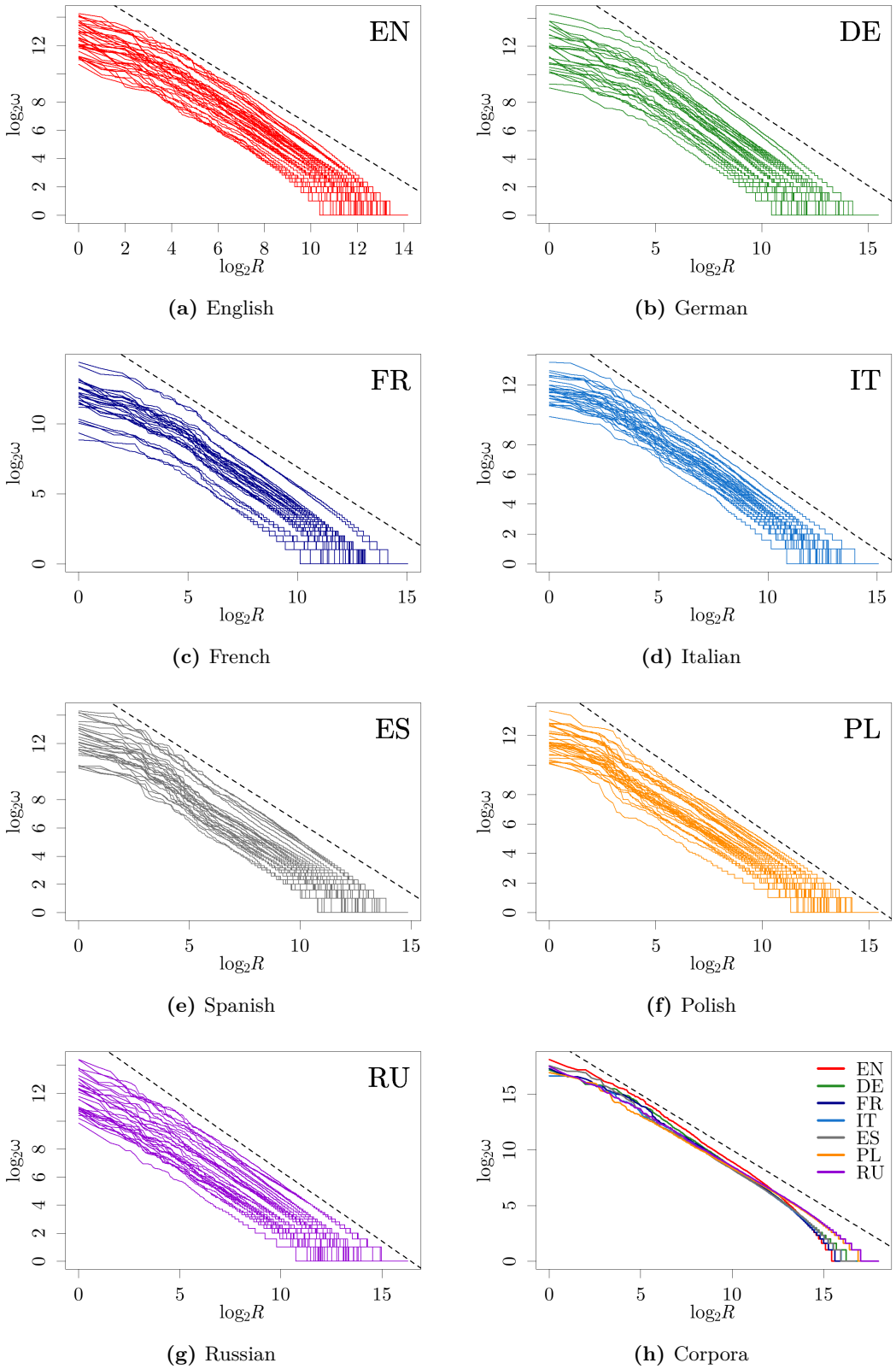
with $\alpha \approx 1$. Mathematically, Eq. 2.40 expresses a rank-size distribution; since the "size" here pertains to frequency, it is also called a rank-frequency distribution. The corresponding probability distribution can be characterized as follows. If V is the vocabulary of a text, that is, the set of all distinct words appearing in the text, then the probability $p_\omega(\omega)$ that a word randomly chosen from V has the frequency ω in the text is expressed by

$$p_\omega(\omega) \propto \omega^{-\beta}, \tag{2.41}$$

where $\beta = 1/\alpha + 1 \approx 2$. Both Eq. 2.40 and Eq. 2.41 are referred to as Zipf's law; the latter is sometimes called the *inverse Zipf's law* [237].

The Zipf's law is observed in the majority of languages studied with regard to this aspect, including artificial languages [238], and extinct languages [239]. Exceptions are languages using logographic writing systems (such as Chinese), but it has been shown that although Zipf's law might not hold for logograms, it might be exhibited in some other way, for example by combinations of logograms [240,241]. An illustration of Zipf's law in 7 different languages is presented in Figure 2.3, where rank-frequency distributions of books listed in Appendix B.1 (and of corpora constructed from those books) are shown.

Although the values of the exponents α and β are to a large degree universal, it is possible to observe deviations from Zipf's law with $\alpha \approx 1$ and $\beta \approx 2$ [242]. β greater than 2 is a sign that a text contains lots of rare words (words with low frequencies); this may be a result of covering a wide range of topics (each with specific vocabulary), or a consequence of the richness of vocabulary of particular author. Conversely, $\beta < 2$ indicates poor vocabulary, which may be specific to particular language user (for example $\beta < 2$ is observed in schizophrenia and in language used by very young children) or due to specific circumstances (for example, in military communication, where non-essential words tend to be avoided). It is also worth mentioning that even when Zipf's law holds in its basic form (with $\beta \approx 2$), linguistic analysis relying on Zipf's law may reveal properties specific to individual texts. For example, one can define a distance between Zipf plots of two texts (based on the ordering of the words in rank-frequency relationship), and it seems that such a distance is lower for pairs of texts being similar in some sense (belonging to the same author or genre, for instance) than for pairs of "unrelated" texts [243]. Another interesting observation derived from word frequency analysis pertains to how statistical regularities like Zipf's law are manifested in different parts of a text. For example, it has been observed that if a text is cut into two halves, there are



(a) English **(b)** German

(c) French **(d)** Italian

(e) Spanish **(f)** Polish

(g) Russian **(h)** Corpora

Figure 2.3. (a)-(g): Zipf's law in books in various European languages. Each line represents the log-log plot of the rank-frequency distribution (its continuity is only to make the graph more legible, since rank is a discrete variable) for a single book. The books are listed in Appendix B.1. Words are in their original form (they are not lemmatized). The differences of individual lines' vertical positions are due to different sizes of the books. (h): Zipf's law for corpora constructed from all the books in each language. In all the plots (a)-(h) the dashed line has the slope equal to -1.

statistically significant differences between some statistical properties of the first and of the second half [244].

Another linguistic law involving power-law relationships is Heaps' law (also called Herdan's law) [98–100]. It describes how the number of distinct words increases with the increasing size of a text. If N denotes the number of all words encountered up to some point in the text and $V(N)$ is the number of distinct words (the vocabulary size) up to that point, then Heaps' law can be formulated as:

$$V(N) \approx C_H N^{\eta}, \tag{2.42}$$

where η is a real number between 0 and 1, and $C_H > 0$ is a constant with respect to N; it might depend on language and on the specific text. The relationship between N and $V(N)$ given by Eq. 2.42 typically holds for a few orders of magnitude of N; for very long texts ($N \to \infty$), the increase of $V(N)$ becomes slower and slower, as there are less and less commonly used words in the set of words yet unused. For some time, Heaps' law was treated as a trait of language separate from Zipf's law, but it has been shown that it can be considered as related to Zipf's law - that is, it is possible to show that assuming that language is subject to Zipf's law leads to Heaps' law, under some (mild) additional assumptions [245–247]. Heaps' law is illustrated in Figure 2.4; the figure presents the log-log plots of $V(N)$ for the corpus constructed from English books listed in Appendix B.1.

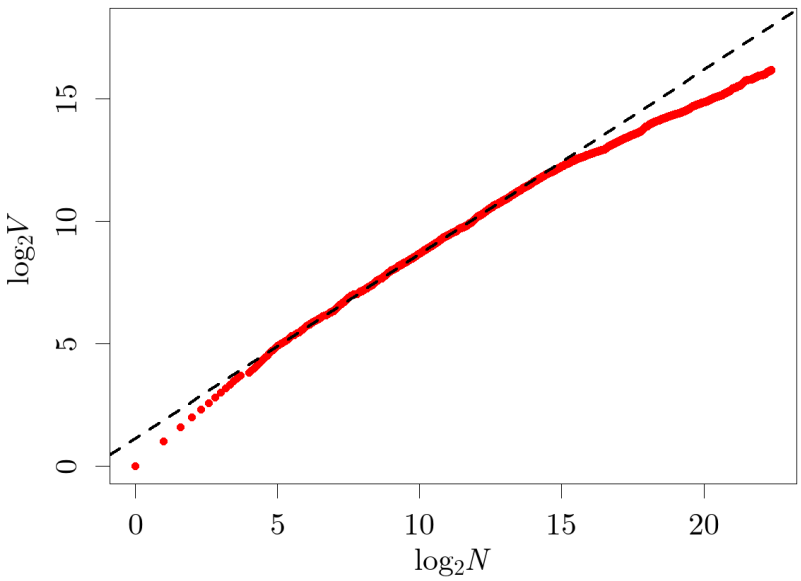


Figure 2.4. An illustration of the Heaps' law, using the corpus constructed from English books listed in Appendix B.1. The dots represent $V(N)$, the size of the vocabulary as a function of text length. The slope of the dashed line is equal to $\eta = 0.75$. The power-law regime holds for a few orders of magnitude. For small N, the relationship $V(N)$ is practically linear (as almost every consecutive word in the text expands the vocabulary); for large N, the lack of new, yet unencountered words makes $V(N)$ grow more slowly.

2.3.2 Attempts to explain the origin of Zipf's law

As stated before, there are multiple mechanisms which might serve as explanations of the presence of power laws in various systems. This applies also to natural language. There have been many attempts to explain the origin of Zipf's law, some of them contradictory to each other, but no universally agreed theory has been proposed. Zipf's original explanation was the *principle of least effort*. According to

this principle, the language optimizes the information transfer between the speaker (information source) and the listener (information receiver). The messages need to be as short as possible, but at the same time they have to contain enough information to be understandable. The idea of the principle can be roughly presented with the following line of reasoning [248]. Let words be sequences of symbols, taken from an n-element alphabet. The cost of using a word is equal to its length (the number of symbols it contains). Therefore the most frequently used words are the shortest ones. Assuming that all possible sequences of symbols are used to form words, it can be stated that the cost of the word of rank R can be approximated as $c_R \approx \log_n R$. This can be understood as follows: there are n^l words of length l, so for an arbitrary word of length l there is $m = \sum_{j=1}^{l} n^j$ words with lengths less or equal to l and therefore occurring with frequencies higher or equal to the frequency of the considered word. Hence, m can be interpreted as the rank of a word consisting of l symbols, $m = R$. Then to approximate R one can notice that

$$R = m = \sum_{j=1}^{l} n^j = n^l \sum_{j=0}^{l-1} \left(\frac{1}{n}\right)^j. \tag{2.43}$$

Since the sum on the right-hand side of the above equation is always smaller than the sum of the geometric series $\sum_{j=0}^{\infty}(1/n)^j = n/(n-1)$, R satisfies

$$n^l < R < n^l \frac{n}{n-1}. \tag{2.44}$$

The larger the n, the better the approximation of R by $R \approx n^l$. From that approximation one gets $l \approx \log_n R$. The cost of using a word is expressed by the length of that word, so the cost c_R of using the word of rank R can be expressed as: $c_R = l \approx \log_n R$. If p_R is the normalized frequency of the word with rank R (in other words, it is the probability that a word randomly chosen from the text is the word with rank R), then the average cost per word $\langle c \rangle$ is:

$$\langle c \rangle = \sum_{R=1}^{\max\{R\}} p_R c_R. \tag{2.45}$$

The average amount of information per word can be expressed by information entropy H (here $\log(\cdot)$ is the natural logarithm and the entropy is given in nats):

$$H = -\sum_{R=1}^{\max\{R\}} p_R \log p_R. \tag{2.46}$$

According to the principle of least effort, word frequency distribution in language is such that the transmission of information is cost-efficient, that is, it minimizes the quantity $\langle c \rangle / H$. The set of numbers p_R, which constitute the rank-frequency distribution, can be found by minimizing $\langle c \rangle / H$ with the normalization constraint imposed on p_R: $\sum_{R=1}^{\max\{R\}} p(R) = 1$. This can be done by Lagrange multipliers method - treating $\langle c \rangle$ and H as functions of p_R ($R = 1, 2, , ..., \max\{R\}$) specified by Eq. 2.45 and Eq. 2.46, one minimizes $\langle c \rangle / H$ by solving for each p_R:

$$\frac{\partial}{\partial p_R}\left(\frac{\langle c \rangle}{H} - \lambda \sum_{r=1}^{\max\{R\}} p_r\right) = 0, \tag{2.47}$$

where λ is the Lagrange multiplier. Calculating the derivative transforms the above equation into:

$$\frac{c_R}{H} + \frac{\langle c \rangle (\log p_R + 1)}{H^2} - \lambda = 0 \quad \text{for each } R, \tag{2.48}$$

from which one gets

$$p_R \;=\; \exp\left(\frac{\lambda H^2}{\langle c\rangle} - 1\right)\exp\left(-\frac{c_R H}{\langle c\rangle}\right) \;=\; A_\lambda R^{-H/(\langle c\rangle \log n)}, \tag{2.49}$$

where $\exp(\cdot)$ is the exponential function, $A_\lambda = \exp\left(\lambda H^2/\langle c\rangle - 1\right)$ serves as a normalization constant (which can be set by setting λ appropriately), and the last equality follows from expressing the word usage cost c_R in the form $c_R = \log_n R$. The above formula implies that minimizing $\langle c\rangle/H$ leads to a power-law rank-frequency distribution $p_R \propto R^{-\alpha}$, with exponent $\alpha = H/(\langle c\rangle \log n)$ (it is worth noting that this result does not give p_R explicitly; to obtain a closed-form solution one needs to explicitly determine $\langle c\rangle$ and H).

A model of text generation able to generate power-law rank-frequency distributions, based on a line of reasoning different than the one presented above, is the so-called model of *intermittent silence* (also referred to as *typewriting monkey* [206]), introduced by Miller [249]. It can be shown that under some general assumptions the basic idea of that model is in fact mathematically equivalent to the idea of the least effort principle [250], but the intuition behind is slightly different. Let a text be generated by adding one symbol at a time, each symbol being either a letter from an n-element alphabet or a space. The symbol to append at each step is chosen randomly, the space is chosen with probability p_s; if the chosen symbol is not the space, then it is a letter picked randomly from an uniform distribution, so the probability of each letter is equal to $(1-p_s)/n$. The choices of symbols are independent of each other. To generate a particular word, a specific sequence of symbols followed by a space must occur. The probability p_l of generating a specific word of length l is therefore given by:

$$p_l = \left(\frac{1-p_s}{n}\right)^l p_s = p_s \exp\left(l\,\log\frac{1-p_s}{n}\right) \tag{2.50}$$

where $\exp(\cdot)$ and $\log(\cdot)$ denote the exponential function and natural logarithm, respectively. Using the same approximate relation between the length l and the rank R of a word as before: $l \approx \log_n R = \log R/\log n$, one obtains the normalized frequency p_R of the word with rank R in the form:

$$p_R = p_l = p_s \exp\left(\frac{\log R}{\log n}\,\log\frac{1-p_s}{n}\right) = p_s R^{-\alpha} \quad \text{with } \alpha = 1 - \frac{\log(1-p_s)}{\log n}, \tag{2.51}$$

where $\exp(\cdot)$ and $\log(\cdot)$ denote the exponential function and natural logarithm, respectively. Setting $n = 26$ and $p_s = 0.18$, which are values used originally by Miller, taken from English language, one gets $\alpha \approx 1.06$, which is close to the exponent of the Zipf's law (Eq. 2.40). It is worth noting that the power-law distribution generated by intermittent silence model can be derived by considering an exponential function of an exponentially distributed variable (this line of reasoning is discussed in section 2.2). Since the number of distinct words of length l is n^l, in a finite vocabulary the probability $p(l)$ that a word randomly chosen from that vocabulary has length l is proportional to n^l, so $p(l) \propto e^{\,a\,l}$, where $a = \log n$. A particular word of length l has the frequency ω in the text proportional to $\left((1-p_s)/n\right)^l$, so $\omega(l) \propto e^{\,b\,l}$, where $b = \log\left((1-p_s)/n\right)$. Using Eqs. 2.27 - 2.33 one gets the probability that a word randomly chosen from the vocabulary has the frequency ω in the text: $p(\omega) \propto \omega^{-\beta}$, with β given by:

$$\beta = 1 - \frac{a}{b} = 1 - \frac{\log n}{\log(1-p_s) - \log n}. \tag{2.52}$$

Using the relationship between the exponent β of a probability distribution and the exponent α of a rank-frequency distribution $\alpha = 1/(\beta - 1)$, the rank-frequency distribution in the form given by Eq. 2.51 is obtained.

Other mechanisms generating power laws, like Yule processes and its modified variants, have also been used as possible explanations of Zipf's law. An example of such a mechanism is the model of random text production studied by Simon [220]. According to that model, a text initially consists of a single word and then words are added in consecutive steps (one word in one step) in the following way. With probability q a new word (a word not yet present in the vocabulary) is appended, and with probability $1-q$ the appended word is randomly chosen from the words already present in the text; the probability of choosing a specific word is proportional to its frequency (the number of times it has already occurred in the text); this determines that the model can be considered a variant of the Yule process. The model gives the word frequency distribution of the form $p_\omega(\omega) \propto \omega^{-\beta}$, with $\beta = 1 + 1/(1-q)$, in the limit of large number of steps. With q close to 0, β is close to 2, which corresponds to Zipf's law in the form given by Eq. 2.41.

Knowing that power-law distributions can be obtained in simple stochastic models, it may seem doubtful whether the fact that word frequencies in texts are described by power laws gives any significant information about language. However, the studied models are often clearly unrealistic and do not account for many essential traits of language. For example, the intermittent silence model does not take into account that words in natural language do not consist of random letters; only some letter sequences are allowed as others might even not be pronounceable. Also, the distribution of word lengths does not correspond to what is observed in natural language [237]. Nevertheless, the presented models and other procedures of similar type remain an important class of models showing it is possible to obtain a power-law distribution as a result of a rather simple process.

2.3.3 Modifying Zipf's law

In empirical language samples taken from real-world texts, some deviations from Zipf's law in its original form can be observed. One of such deviations is particularly typical for very big corpora, usually consisting of large numbers of texts. For large samples, the rank-frequency distribution $\omega(R)$ with exponent $\alpha \approx 1$ holds up to some rank R_c, and for ranks $R > R_c$ it breaks down and transforms into another power law, with exponent α', larger than α [251,252]. This is often explained by the existence of two types of vocabulary: one being a kind of core vocabulary, consisting of a few thousand words most frequently used in language, the other being more specialized, consisting of less common words which are specific to particular topics or circumstances.

Another frequently observed form of discrepancy between Zipf's law in its basic form and empirical data is the fact that usually words with lowest ranks have frequencies slightly lower than predicted by Zipf's law. Accordingly, Zipf's law holds for ranks above some rank R_Z, usually R_Z is on the order of 5 or 10. For $R < R_Z$, the frequencies $\omega(R)$ are below the frequencies given by exact power-law relationship between ω and R. To account for this effect, Mandelbrot introduced a correction to Zipf's rank-frequency relationship; the resulting formula, known as *Zipf-Mandelbrot law*, can be written as:

$$\omega(R) \propto (R + c)^{-\alpha}, \tag{2.53}$$

where R is the rank of a word, $\omega(R)$ is the frequency of the word of rank R in the text, and c is a non-negative constant. For $c = 0$, Zipf-Mandelbrot law reduces to Zipf's law. Nonzero values of c in the equation describing the rank-frequency relationship introduce the flattening of $\omega(R)$ for small values of R, and therefore allow for more accurate description of empirical data. An illustration of how the shape of $\omega(R)$ given by Zipf-Mandelbrot law depends on the value of c is shown in Figure 2.5.

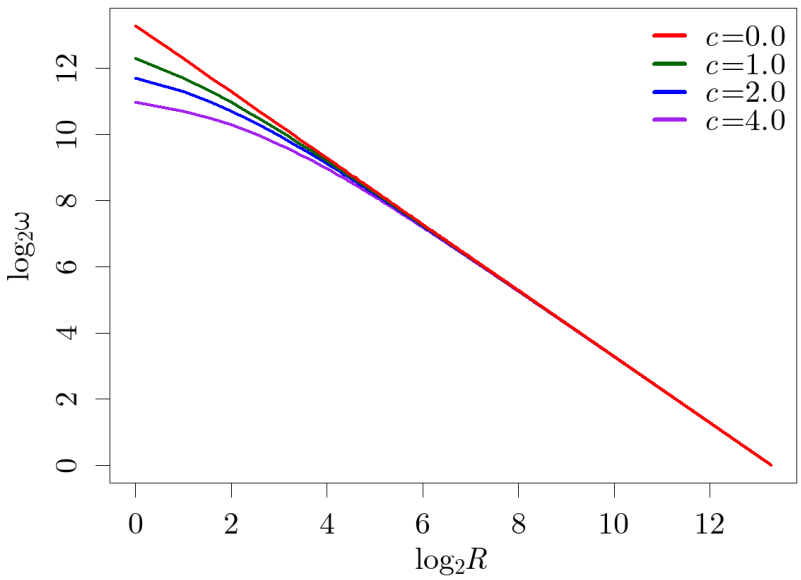


Figure 2.5. Log-log plots of exemplary functions $\omega(R)$ given by Zipf-Mandelbrot law (Eq. 2.53), with $\alpha = 1$ and different values of c.

Zipf's law applies not only to the distribution of words - which in written language can be understood as sequences of non-whitespace characters surrounded by whitespaces from both ends - but also to some other distributions characterizing language. One can study, for example, the rank-frequency distribution of words after lemmatization, that is after reducing each word in the text to its basic, dictionary form, called lemma. It turns out that lemmatized corpora also conform to Zipf's law [253]. In some situations, this allows for more meaningful comparison between languages, especially between languages with different degrees of inflection usage. Without lemmatization, the size of vocabulary can be artificially inflated in inflected languages (languages utilizing inflection to specify words' grammatical features), as various inflected forms of the same lemma are then counted as separate words.

Another possible type of word frequency analysis is investigating how subsets of words behave in terms of frequency distributions. For example, the set of all words in a corpus can be partitioned into subsets corresponding to different parts of speech, and the frequency distribution within each subset can be studied [97, 106]. An example of results of word frequency analysis utilizing such an approach is presented in Fig. 2.6. When performed for corpora in English, the analysis of this type reveals various types of rank-frequency distributions, depending on part of speech under consideration; some parts of speech are subject to power-law rank-frequency distributions, some exhibit different type of behavior. This is related to the role of individual word classes in language. Words whose role is mostly grammatical (like conjunctions, prepositions, pronouns or articles in English) are the ones used most frequently - and their rank-frequency distribution can be considered a power law for some range of low ranks, above which their frequencies quickly decay towards zero. For words being references to specific objects and notions (nouns, for example), the agreement with power-law relationship is typically better outside the range of low ranks. The distinctive behavior of verbs, which seem to conform to a power law in general (in a wide range of ranks), might be related to the fact that verbs play in a sense a dual role in language - most of verbs are associated with some kind of action or state, but a group of verbs in English have special grammatical uses (like *be, have* or *will*).

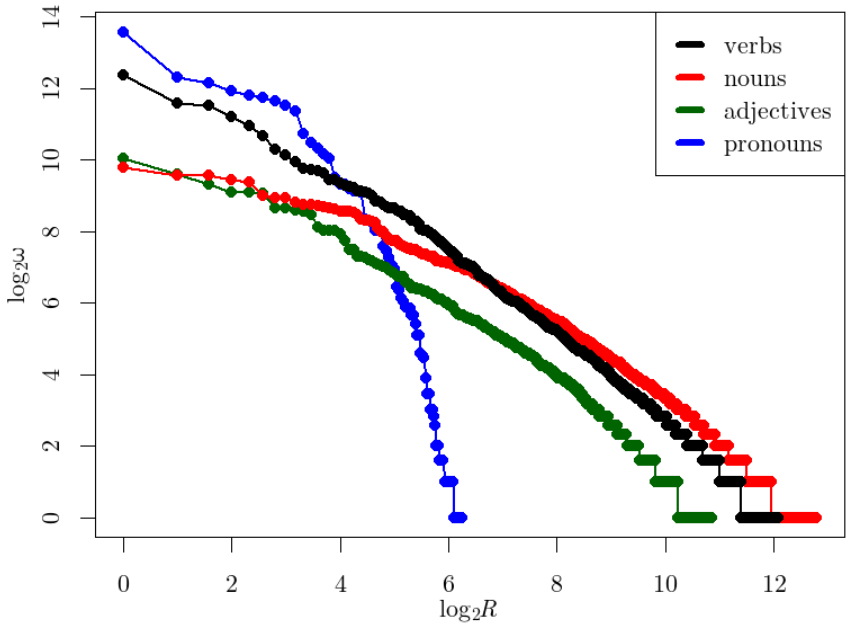
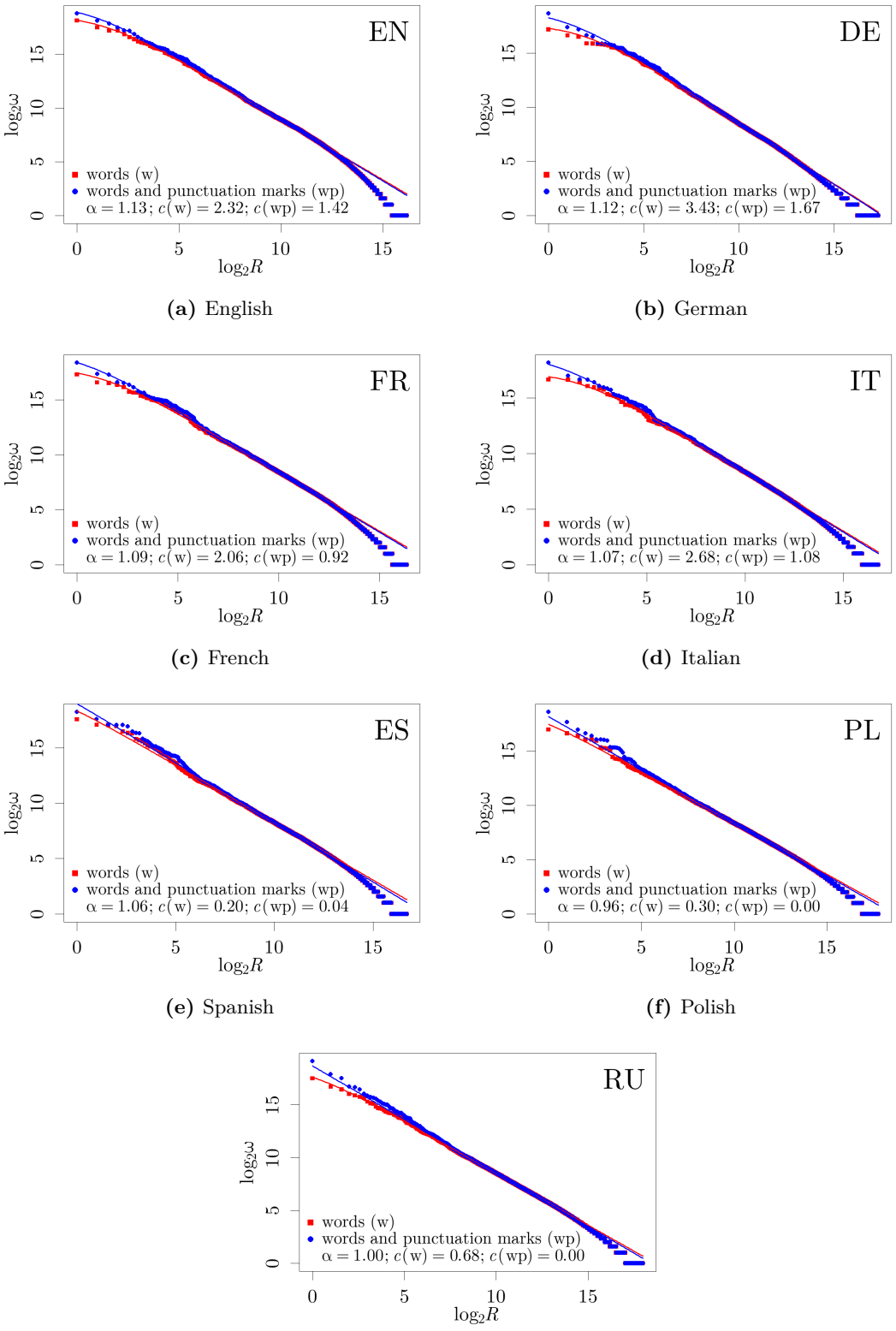


Figure 2.6. Log-log plots of rank-frequency distributions $\omega(R)$ determined separately for selected parts of speech in a book (*David Copperfield* by Charles Dickens). Words are not lemmatized. The power-law relationship is most closely followed by verbs.

Word frequency analysis can also be generalized to other entities occurring in written language. An interesting result has been obtained for punctuation marks treated as words and included into rank-frequency distributions of literary texts [254]. Some of punctuation marks have frequencies comparable with the frequencies of the most frequent words. In fact, a comma and a period often occupy ranks between 1 and 3. It turns out that while rank-frequency distributions determined for words only are typically described by Zipf-Mandelbrot law, the distributions for words together with punctuation marks are closer to the regime given by "pure" Zipf's law. This effect is demonstrated in Figure 2.7. It can be stated that treating punctuation marks as words decreases the flattening of $\omega(R)$ for small R. A few European languages have been studied in that context; all of them have been identified as displaying the presented effect, but with varying intensity. Quantitatively, the decrease in the rank-frequency distribution's flattening for small R corresponds to the decrease of the value of the constant c in Eq. 2.53, expressing Zipf-Mandelbrot law. Among the studied languages, Germanic languages (English and German) have the weakest tendency to restore Zipf's law when punctuation is included (c decreases, but remains significant); for Slavic languages (Polish and Russian) including punctuation into analysis results in c dropping close to 0. Therefore, punctuation marks and words can be considered to fit into the same Zipfian regime of frequency distribution. Together with results from other methods of text analysis [254], this fact allows to state that although punctuation marks and words are clearly different objects, their statistical properties are in many respects comparable. This conclusion leads to a hypothesis that at least in some aspects punctuation carries information in a way similar to words. It also justifies taking punctuation into account when computing word frequency characteristics in statistical analysis of language and its practical applications.



(a) English **(b)** German

(c) French **(d)** Italian

(e) Spanish **(f)** Polish

(g) Russian

Figure 2.7. Rank-frequency distributions for corpora constructed from books in selected European languages (the books are listed in Appendix B.1). Red squares represent the distribution for words, and blue dots - the distribution for words and punctuation marks treated as words; in both cases words are not lemmatized. Zipf-Mandelbrot law (Eq. 2.53) fitted to the data is denoted by solid lines. The power law exponents α, relevant for both distributions (words and words with punctuation marks), are given under each graph, along with the values of c.

Chapter 3

Natural language and time series analysis

3.1 Time series complexity

In physics, describing a system or a phenomenon usually relies on determining various quantities, relationships between them, and the way in which they change over time. In some situations such quantities might be viewed as signals, which can be studied with the use of appropriate tools developed by mathematics and physics. An important class of signals are *time series*, which are sequences of values of some quantity measured in consecutive points in time. Since virtually any time-changing quantity can be represented (precisely or approximately) in the form of a time series, methods of time series analysis are of great importance and wide applicability. Applying such methods to signals encountered in the study of complex systems reveals that many signals of that type share some common traits. An example of such a trait is the presence of long-range correlations, which suggests that an event might significantly influence the occurrence of other events even when they are distant in time. Certain types of correlations in time series result in the emergence of diverse self-similar structures - fractals and multifractals - whose properties can be described with the use of fractal geometry. A notion important in the context of signals with long-range correlations is $1/f$ noise. Signals belonging to the class of $1/f$ noises, or, more generally, $1/f^{\beta}$ noises, can be considered instances of power laws, as they are characterized by power functions in the frequency domain. Such a characteristic indicates the presence of a specific structure and allows to quantify the character of correlations by determining the exponent of a corresponding power law. A time series usually has the form of a sequence of numbers, but it can also be a sequence of symbols, representing consecutive values of some categorical variable. Series of this type can also be subject to methods of time series analysis, which allow to identify patterns of their organization.

Sections 3.2, 3.3 and 3.4 present several concepts useful in studying time series from the perspective of complexity. Section 3.2 briefly presents how the notion of information entropy can be applied to symbolic sequences to quantify the degree of their unpredictability. The sections that follow (3.3 and 3.4) discuss some elementary concepts used to identify and study $1/f$ noises and introduce the basics of fractal geometry and its applications in the analysis of time series. The remaining part of the chapter presents the results of applying the introduced concepts to the analysis of linguistic time series.

3.2 Entropy of symbolic sequences

Let $\{X(t)\}$ be a stationary sequence of categorical random variables (here it is assumed that the possible values of each of the variables are symbols from some fixed set), indexed by time $t \in \{0, \pm 1, \pm 2, ...\}$; stationarity means that for any t, s, and τ the following condition is satisfied:

$$F\Big(X(t), X(t+1), ..., X(t+s)\Big) = F\Big(X(t+\tau), X(t+1+\tau), ..., X(t+s+\tau)\Big), \quad (3.1)$$

where F denotes the joint cumulative distribution function - for some variables $X_1, X_2, ..., X_n$, the function $F(X_1, X_2, ..., X_n)$ is the joint cumulative distribution function of $X_1, X_2, ..., X_n$. The above condition expresses that any block of variables of arbitrary and fixed size has joint probability distribution which does not depend on time.

Stochastic processes constituted by sequences of random variables like the one defined above can differ in how values in different points in time depend on each other. The extent to which variables in a sequence are dependent on each other can be quantified with the use of information entropy. The entropy rate H_X (or less precisely, the entropy) of a stationary process $\{X(t)\}$ with values in some set of symbols can be defined in terms of conditional entropy:

$$\begin{aligned} H_X &= \lim_{n\to\infty} H\Big(X(t)|X(t-1), X(t-2), ..., X(t-n)\Big) = \\ &= \lim_{n\to\infty} \mathbf{E}\Big[-\log_2 P\Big(X(t)|X(t-1), X(t-2), ..., X(t-n)\Big)\Big], \end{aligned} \quad (3.2)$$

where t is an arbitrary point in time, $P\big(X(t)|X(t-1), ..., X(t-n)\big)$ is the conditional probability distribution of $X(t)$ given $X(t-1), X(t-2), ..., X(t-n)$, and $\mathbf{E}[\cdot]$ is the expectation operator; in the above definition entropy is expressed in bits (per one symbol). Entropy rate of a process measures the average degree of uncertainty about the value at some time point, when the values at prior time points are known.

Practical calculation of the entropy of a symbolic sequence can be approached in few ways. One of possible methods of wide applicability is using an estimator based on the idea utilized in Lempel-Ziv algorithm - identifying repeating sequences in the series. For a sequence of symbols $x_0, x_1, x_2, ...$ (treated as a realization of some stochastic process $\{X(t)\}$), let l_i be the length of the longest subsequence $x_i, x_{i+1}, ..., x_{i+l-1}$ starting at position i which also appears somewhere between positions 0 and $i-1$. Let L_i be defined as $l_i + 1$; L_i is the length of the shortest subsequence starting at position i which does not appear anywhere up to position $i-1$ in the studied sequence. Let $\hat{H}_n$ be defined as follows:

$$\hat{H}_n = \left(\frac{1}{n}\sum_{i=2}^{n}\frac{L_i}{\log_2 i}\right)^{-1}. \quad (3.3)$$

It can be shown [190,191] that under some mild assumptions regarding the behavior of the underlying process in large time scales, $\hat{H}_n$ converges to the value of process' entropy H_X (in bits per symbol) when $n \to \infty$. It is a consequence of the fact that in a process with high entropy consecutive symbols are highly "unpredictable" and repeating long symbol sequences is rare; this corresponds to low values of L_i. On the other hand, low entropy processes generate sequences with frequent repetitions - which correspond to high values of L_i. The relationship between $\hat{H}_n$ and H_X can be used to estimate the entropy of a finite symbol sequence - the entropy can be approximated by $\hat{H}_n$ computed for possibly large n (limited by sequence length). It should be noted, however, that the obtained results might be prone to non-negligible errors, as the convergence of $\hat{H}_n$ to H_X is relatively slow [191].

3.3 Long-range correlations in time series

Dependencies introducing complex patterns of organization into signals can have various forms, depending on signal type. For signals of different type, different methods of identifying such dependencies and patterns are used. For data having the form of a sequence of numbers, basic methods often utilize the analysis in frequency domain and studying the properties of the autocovariance function. Let (x_n) for $n \in \{0, \pm 1, \pm 2, ...\}$ be a sequence of real numbers, interpreted as values of some quantity measured in equally spaced points in time. The sequence (x_n) might represent a signal defined both on infinite and on finite time interval; in the latter case one can set all the values outside the studied interval to 0. If the signal has *finite energy*, which means that

$$\sum_{n=-\infty}^{+\infty} |x_n|^2 < \infty, \tag{3.4}$$

then its spectral density $S(f)$ can be defined as the squared modulus of its discrete-time Fourier transform [255]:

$$S(f) = \left| \sum_{n=-\infty}^{+\infty} x_n \, e^{-i(2\pi f)n} \right|^2 . \tag{3.5}$$

For signals with finite energy, spectral density is called the *energy spectral density* or *energy spectrum*. However, many signals do not have finite energy; also, time series are often realizations of random (stochastic) processes, whose properties cannot be fully characterized by studying just one realization. For infinite-energy signals and stochastic processes a useful notion is the *power spectral density*, also defined in terms of Fourier transform. For an infinite sequence $x_0, x_1, x_2, ...$ power spectrum $S(f)$ can be defined as [255]:

$$S(f) = \lim_{N \to +\infty} \mathbf{E} \left[\frac{1}{N} \left| \sum_{n=0}^{N} x_n \, e^{-i(2\pi f)n} \right|^2 \right], \tag{3.6}$$

where $\mathbf{E}[\cdot]$ denotes the expectation operator, which averages over the ensemble of possible realizations, if (x_n) represents a stochastic process. For (x_n) being a single sequence of real numbers, the averaging can be omitted. Signals typically encountered in practice are finite sequences $x_0, x_1, x_2, ..., x_{N-1}$; for such time series one can define spectral density, also referred to as the *periodogram*, in terms of squared modulus of the discrete Fourier transform (DFT):

$$S(k) = C \left| \sum_{n=0}^{N-1} x_n \, e^{-i(2\pi k/N)n} \right|^2 , \tag{3.7}$$

where C is a constant and $k = 0, 1, 2, ..., N-1$ is the variable indexing harmonic frequencies of DFT's fundamental frequency (corresponding to one cycle per whole sequence). If the time series represents some process generating signals with finite power, then $C = 1/N$ and the periodogram can be interpreted as an approximation of the power spectral density of the underlying process. If the series is treated as a signal with finite energy, then $C = 1$. When comparing spectra of multiple signals it is sometimes useful to normalize the signals before computing the periodograms - for example to unit energy or to unit average power (total energy divided by series' length), depending on signals' type. If the series $x_0, x_1, x_2, ..., x_{N-1}$ represents a sequence of measurements taken with the sampling period T_s, then $S(k)$ can be interpreted as the value of spectral density corresponding to the frequency $f =$

$k/(NT_s) = kf_b$, where $f_b = 1/(NT_s)$ is the fundamental frequency (here it is worth remembering that due to DFT's periodicity, all "significant" values of $S(k)$ are inside the range of k given by: $0 \leq k \leq N/2$). Thus, spectral density S can be expressed as a function of $f/f_b = k$ or as a function of f, with f having discrete values: $f = kf_b$. If the indices of the series' consecutive values do not have a specific interpretation of points in time, T_s can be set to 1 and then the fundamental frequency is expressed by $f_b = 1/N$. From a technical point of view, it should be pointed out that sometimes determining a reliable estimation of the spectrum from finite-size data requires additional steps [255], like dividing the time series into windows, computing the periodogram for each window, and then averaging the results.

Spectral density of a signal describes how much a given frequency or frequency band contributes to signal's total variability. It also gives insight into temporal correlations, just as the autocovariance function, which is related to spectral density when the signal or stochastic process is of certain type. For a stochastic process being a collection of real-valued random variables $\{X(t)\}$ indexed by time t, the autocovariance function is defined as:

$$\begin{aligned} R_{XX}(t_1, t_2) &= \mathbf{E}\Big[\Big(X(t_1) - \mathbf{E}[X(t_1)]\Big)\Big(X(t_2) - \mathbf{E}[X(t_2)]\Big)\Big] = \\ &= \mathbf{E}\Big[X(t_1)X(t_2) - \mathbf{E}[X(t_1)]\,\mathbf{E}[X(t_2)]\Big] \end{aligned} \tag{3.8}$$

(again, $\mathbf{E}[\cdot]$ denotes the expectation operator). It is worth mentioning that there exists another naming convention in which the above function is called the *autocorrelation function*. Autocovariance function has a simplified form for *weakly stationary* processes. A process $\{X(t)\}$ is weakly stationary (also referred to as *covariance stationary* or *wide-sense stationary*), if:

$$\begin{aligned} &\mathbf{E}[X(t_1)] = \mathbf{E}[X(t_2)] \quad \text{for all } t_1, t_2, \\ &\mathbf{E}\Big[X(t)^2\Big] = \sigma^2 < \infty \quad \text{for all } t, \\ &R_{XX}(t_1, t_2) = R_{XX}(0, t_2 - t_1) = R_X(t_2 - t_1) = R_X(\tau) \quad \text{for all } t_1, t_2. \end{aligned} \tag{3.9}$$

The last condition expresses the fact that the autocovariance function of a weakly stationary process depends on only one variable, $\tau = t_2 - t_1$, hence the notation $R_X(\tau)$. Autocovariance function is related to spectral density: for a weakly stationary process with zero mean ($\mathbf{E}[X(t)] = 0$ for all t), spectral density is equal to the Fourier transform of the autocovariance function [255].

The form of the autocovariance function $R_X(\tau)$ describes the type of correlations characteristic for the process. For example, a weakly stationary, zero-mean process is said to have *long-range correlations* or *long memory* when its autocovariance function decays so slowly that its sum or integral to infinity is divergent [256, 257]. For a process in discrete time, this can be written as:

$$\lim_{N \to \infty} \sum_{\tau=0}^{N} R_X(\tau) = \infty. \tag{3.10}$$

A signal whose spectral density $S(f)$ satisfies

$$S(f) \propto \frac{1}{f^\beta} \tag{3.11}$$

for sufficiently wide range of frequency f, is called a $1/f^\beta$ noise (or $1/f^\alpha$ noise, since the exponent in Eq. 3.11 is also often denoted by α). The range of β is typically

between 0 and 2. The case of $1/f$ noise corresponds $\beta = 1$, but sometimes this name is used to refer to $1/f^\beta$ noise in general. Recognizing a signal as $1/f^\beta$ noise implies the presence of a specific structure of correlations. For example, it can be shown [258] that a weakly stationary, zero-mean process in discrete time whose autocovariance function for large τ has the form:

$$R_X(\tau) \propto \tau^{-\alpha}, \quad \alpha \in (0;1)\,, \tag{3.12}$$

(which makes the process a long-memory process), has spectral density also behaving like a power function for small f (that is, for large time scales; for finite time series minimal f is determined by the series' length):

$$S(f) \propto \frac{1}{f^{1-\alpha}}. \tag{3.13}$$

Therefore, $1/f^\beta$ spectrum for β between 0 and 1 can be directly related to long-range correlations. The most elementary characteristics of $1/f^\beta$ noises with β between 1 and 2 can be illustrated with the use of an important and widely studied example of a process able to generate such signals - the so-called *fractional Brownian motion* [259, 260]. One of the ways in which it can be defined is using another process, the *fractional Gaussian noise*. In its discretized variant, fractional Gaussian noise can be characterized as a collection of variables $\{B'_H(t)\}$ indexed by discrete time t, in which all the variables $B'_H(t)$ have the same normal distribution with zero mean and standard deviation σ, and the autocovariance function is given by:

$$R_{B'_H}(\tau) = \frac{\sigma^2}{2}(\,|\tau+1|^{2H} - 2\,|\tau|^{2H} + |\tau-1|^{2H}\,). \tag{3.14}$$

The autocovariance function of fractional Gaussian noise depends on the parameter $H \in (0;1)$ called *Hurst parameter*, *Hurst index* or *Hurst exponent*. The value of H determines the character of the correlations; for $H = \frac{1}{2}$ the variables $\{B'_H(t)\}$ are independent. For $H < \frac{1}{2}$ the process is *antipersistent*, that is, its values in consecutive time steps are negatively correlated; for $H > \frac{1}{2}$ it is *persisent* - which means that its consecutive values are correlated positively. It can be shown [256] that $R_{B'_H}(\tau)$ behaves as a power function with exponent $2H-2$ for large τ:

$$R_{B'_H}(\tau) \propto \tau^{2H-2}. \tag{3.15}$$

Spectral density $S(f)$ of fractional Gaussian noise for small f satisfies [256, 259, 260]:

$$S(f) \propto \frac{1}{f^{2H-1}}. \tag{3.16}$$

Therefore, for H greater than $\frac{1}{2}$ fractional Gaussian noise is a process exhibiting long-range correlations.

Fractional Brownian motion can be defined by specifying its increment process - fractional Brownian motion B_H with Hurst exponent H is a process whose increment process is fractional Gaussian noise with Hurst exponent H [257, 261]. For processes in discrete time this can be written as:

$$B_H(t) = \sum_{s=0}^{t} B'_H(s). \tag{3.17}$$

A time series representing the fractional Brownian motion can be considered a realization of a correlated random walk (starting at zero, according to the characterization given above). The correlations of the increments are determined by the Hurst

exponent H; for $H = \frac{1}{2}$ the process reduces to classical Brownian motion. Fractional Brownian motion is a non-stationary process, its autocovariance function depends on two variables. Spectral density $S(f)$ of a fractional Brownian motion with Hurst exponent H holds [257, 259, 260]:

$$S(f) \propto \frac{1}{f^{2H+1}}. \tag{3.18}$$

Examples of fractional Gaussian noises and fractional Brownian motions with different values of H are shown in Figure 3.1.

One of important properties of fractional Brownian motion is related to the behavior of the variance - for any t_0 and $t > 0$ [256, 262]:

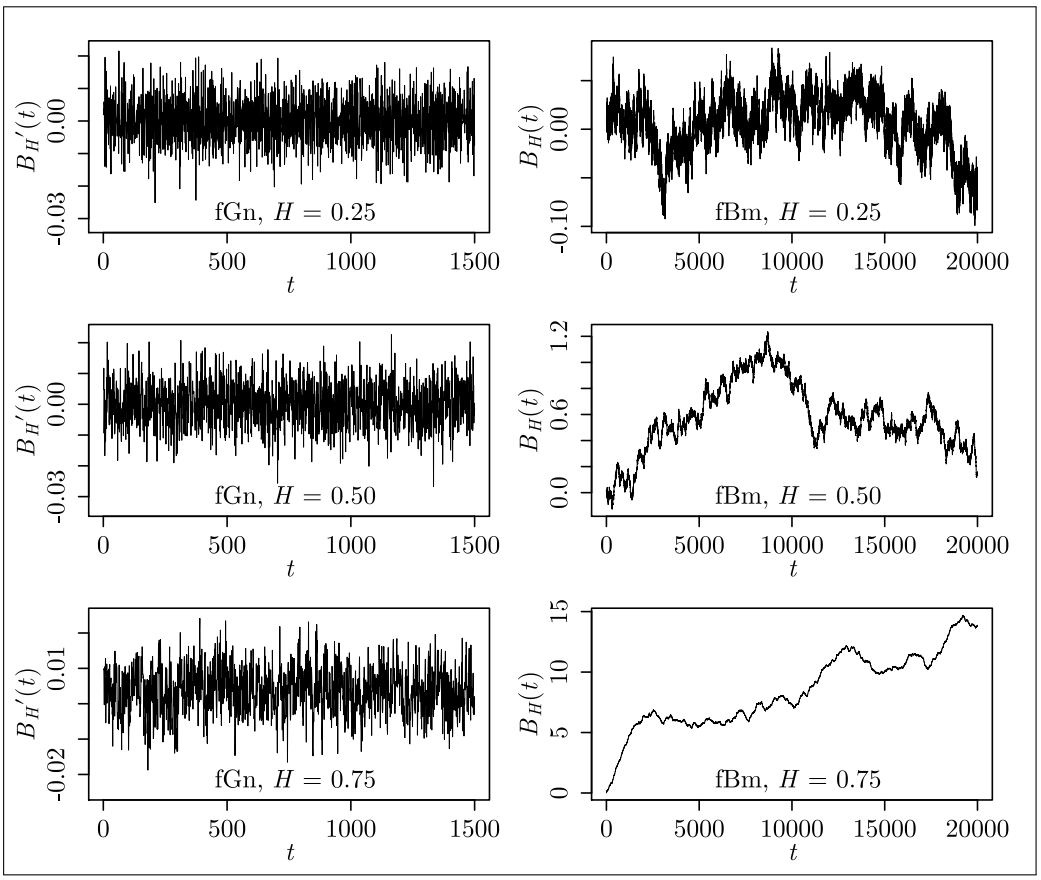
$$\mathbf{E}\Big[\big(B_H(t_0+t) - B_H(t_0)\big)^2\Big] = \sigma^2 t^{2H}, \tag{3.19}$$

where σ^2 is the variance of the increments of B_H (constant in time, since the increments are stationary). This relationship can be used to define Hurst exponent in a more general sense - for processes other than fractional Gaussian noise and fractional Brownian motion. If $\{Y(t)\}$ is a process whose increments are given by a (weakly stationary) process $\{X(t)\}$, and $\{Y(t)\}$ satisfies (for any t_0 and $t > 0$):

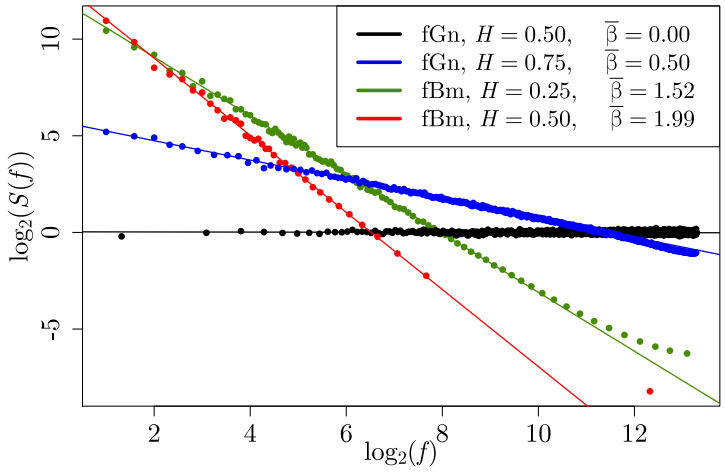
$$\mathbf{E}\Big[\big(Y(t_0+t) - Y(t_0)\big)^2\Big] \propto t^{2H}, \tag{3.20}$$

then H can be called the Hurst exponent of $\{X(t)\}$. It characterizes the process by quantifying how the quantity expressed by the left-hand side of Eq. 3.20 scales with time, compared to a random walk process. Knowing that $H = \frac{1}{2}$ is observed when $\{Y(t)\}$ is a random walk with uncorrelated increments, observing $H > \frac{1}{2}$ or $H < \frac{1}{2}$ allows to identify, respectively, the persistence or the antipersistence of $\{X(t)\}$. It is worth mentioning that the notion of the Hurst exponent occasionally pertains to the "cumulative" process $\{Y(t)\}$ instead of the increments $\{X(t)\}$ (the case of fractional Gaussian noise and fractional Brownian motion both using the name "Hurst exponent" to refer to their parameters is an example) - the exact meaning might be dependent on the context.

It can be concluded from Eq. 3.16 and Eq. 3.18 that using fractional Gaussian noise or fractional Brownian motion one can generate $1/f^\beta$ noises with β both below and above 1. White noise - a signal with no correlations and flat spectral density (proportional to $1/f^0$) - is obtained from fractional Gaussian noise with $H = \frac{1}{2}$. Fractional Gaussian noise with $H \in (\frac{1}{2}; 1)$ produces a long-memory signal of $1/f^\beta$ type with $\beta \in (0; 1)$. Fractional Brownian motion with $H \in (0; \frac{1}{2})$ is a $1/f^\beta$ signal with $\beta \in (1; 2)$. Classical Brownian motion, obtained by setting $H = \frac{1}{2}$ in fractional Brownian motion, generates $1/f^2$ noise (Brownian noise). Apart from summarizing the relationships between power laws, correlations, and spectral properties of a certain class of signals important in studying many natural phenomena, the presented statements allow to illustrate why $1/f$ noises (here understood as $1/f^\beta$ noises with β equal to or close to 1) are in a sense special. $\beta = 1$ is at the interface between two regimes; one corresponding to (correlated) random-walk-like processes ($1 < \beta \leq 2$), the other describing the processes behaving like the increments of a correlated random walk ($0 \leq \beta < 1$). For that reason in some contexts $\beta = 1$ is considered to be a case of particular interest, corresponding to signal's maximum complexity [263, 264].



(a)



(b)

Figure 3.1. (a) Time series being realizations of fractional Gaussian noise (fGn) and fractional Brownian motion (fBm) with Hurst exponents $H = 0.25$, $H = 0.5$ and $H = 0.75$. The length of the series is 1500 for fGn and 20000 for fBm. The effect of H on the behavior of the series can be clearly seen in fBm - the higher the value of H, the greater the smoothness of the curve representing the series. (b) A log-log plot of the spectral densities of fGn and fBm for selected values of H; each spectrum is computed by averaging and smoothing spectral densities of 100 time series with $2 \cdot 10^4$ points, generated by the relevant process. The signals are normalized to have the same average power. The power-law behavior of the spectra allows to identify the underlying signals as $1/f^{\beta}$ noises, with $\beta \in [0; 2]$. The lines drawn on the plot are fitted by least squares method (in the range where the spectra are approximately linear); the values of $\overline{\beta}$, which represent the slopes of the lines with the minus sign, are given alongside the parameters of the processes used to generate the signals.

3.4 Fractals and multifractals

3.4.1 Elementary concepts in fractal geometry

The origins of the study of fractals can be dated back to 1960s and 1970s, when B. Mandelbrot developed foundations of *fractal geometry* [119, 193]. A number of concepts which are today incorporated into fractal geometry had appeared from time to time in works of mathematicians before the unified theory was established; however, those concepts were usually considered highly abstract and having little practical application. It was only after the systematization of the relevant ideas initiated by Mandelbrot that fractal geometry became an important tool in understanding and describing many shapes and patterns formed by nature.

The notion of a fractal is typically associated with a geometrical shape having the property of *self-affinity*, which can be loosely understood as being composed of parts that are downscaled and possibly rotated or reflected copies of the whole shape. More precisely, self affinity is defined in terms of affine transformations. An affine transformation $S : \mathbb{R}^n \to \mathbb{R}^n$ is a transformation of the form:

$$S(v) = T(v) + b, \tag{3.21}$$

where v and b are vectors in $\mathbb{R}^n$, and $T : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation, usually represented as a $n \times n$ matrix; linearity of T means that T satisfies $T(v_1 + v_2) = T(v_1) + T(v_2)$ and $T(\lambda v_1) = \lambda T(v_1)$ for all $v_1, v_2 \in \mathbb{R}^n$ and all $\lambda \in \mathbb{R}$. An affine transformation can therefore consist of scaling, rotation and translation (reflection, which is sometimes mentioned in this context, can be treated as scaling with the scale of -1). Scaling is in general anisotropic, that is, the scaling factors for different directions can be different. If all the scaling factors have the same value, then the scaling is isotropic. An affine transformation in which scaling is isotropic is a similarity transormation - therefore, affine transformations can be considered generalizations of similarity transformations. If there exists a number $c \in (0; 1)$ such that for all $v_1, v_2 \in \mathbb{R}^n$ an affine transformation S satisfies

$$\|S(v_2) - S(v_1)\| \leq c \|v_2 - v_1\|, \tag{3.22}$$

where $\|\cdot\|$ denotes the norm of a vector, then S is called a *contraction mapping* (or shortly a *contraction*). If $S_1, S_2, ...S_m$ is a sequence of contraction mappings in $\mathbb{R}^n$, then there exists a closed and bounded set F such that [265]

$$F = \bigcup_{i=1}^{m} S_i(F). \tag{3.23}$$

This means that if each of the transformations S_i is applied to F, the union of their images is F; F is therefore a fixed point of the system of contraction mappings $S_1, S_2, ...S_m$. The existence and uniqueness of F is guaranteed by Banach fixed point theorem (also known as contraction mapping theorem). Another consequence of the theorem is that an approximation of F can be obtained by applying the sequence of transformations $S_1, S_2, ...S_m$ recursively, starting from a practically arbitrarily chosen, closed and bounded set. If $S(A)$ denotes the union of the images of some set $A \in \mathbb{R}^n$ under $S_1, S_2, ...S_m$, then the sequence (F_k), defined as

$$F_k = \begin{cases} A, & \text{for } k = 0 \\ S(F_{k-1}), & \text{for } k = 1, 2, 3, ... \end{cases} \tag{3.24}$$

tends to F as $k \to \infty$, provided that A is nonempty, closed, bounded, and satisfies $S_i(A) \subset A$ for all $i = 1, 2, ..., m$. This fact is the basis of an important method of

generating fractals, the so-called *Iterated Function Systems* (IFS). The method, in its basic form, is a straightforward application of the fact that if a fractal set F is a fixed point of some known contraction mappings system, then by applying the contractions over and over again, one can obtain successive approximations of F. Figure 3.2 shows an example of the identification of the contraction mapping system whose fixed point is a fractal set (Sierpiński triangle).

If a set is a fixed point of a contraction mapping system, it is called *self-affine*. *Self-similarity* is the case where all the contractions in the system are similarity transformations. However, the term self-similarity is often used to describe both self-similarity in strict sense and self-affinity; such a convention can be used when it does not lead to confusion.

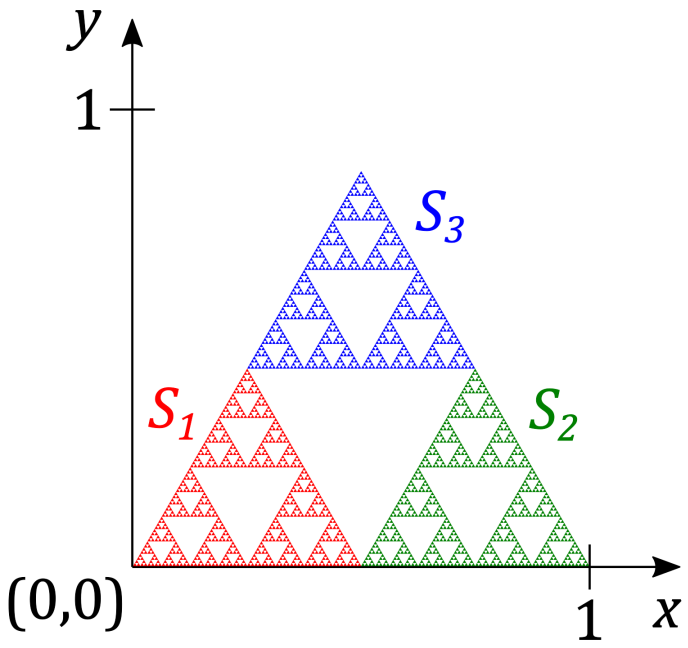


Figure 3.2. Self-similarity of the Sierpiński triangle. The whole triangle can be divided into three parts - red, green and blue - each of which is an image of the whole triangle under one of similarity transformations S_1, S_2, S_3. If $(x, y)^T$ is a column vector representing the coordinates of a point on the plane, the transformations S_1, S_2, S_3 can be written as: $S_1\big((x, y)^T\big) = \frac{1}{2}(x, y)^T$, $S_2\big((x, y)^T\big) = \frac{1}{2}(x, y)^T + (\frac{1}{2}, 0)^T$, $S_3\big((x, y)^T\big) = \frac{1}{2}(x, y)^T + (\frac{\sqrt{3}}{4}, 0)^T$. This shows that the Sierpiński triangle is a fixed point of the contraction mapping system constituted by S_1, S_2, S_3.

Although self-similarity is a typical property of fractals, it is not sufficient to define a fractal. There are objects which do exhibit self-similarity and are not fractals. For example, a square in $\mathbb{R}^2$ is self-similar (it can be easily seen that a square can be divided into four smaller squares, each being an image of the original square under a similarity transformation), and yet, it is not identified as a fractal. In fact, the notion of a fractal does not have a precise definition; fractals are usually characterized with a collection of statements regarding their properties. These can be summarized as follows [265].

- A fractal has a fine structure, that is, detail on arbitrarily small scales.
- Fractals have irregular, "rough" shape, which makes it difficult to describe their properties in terms of "traditional" geometry.
- They can typically be defined as objects generated by recursively repeated (and often relatively simple) procedures.

- Fractals exhibit some form of self-similarity (or self-affinity), possibly in approximate or statistical sense.
- A fractal usually has fractal dimension greater than its topological dimension (fractal dimensions are discussed in subsection 3.4.2).

The fact that fractals are often produced by a simple procedure applied recursively many times can be exemplified by Iterated Function Systems, L-systems (Figure 1.4) and some other methods of generating fractals (Figure 3.3). It also allows to intuitively understand the abundance of fractal patterns in nature - repeating the same action recursively over and over again is a kind of process which can be identified in many natural phenomena (some examples of fractal shapes observed in nature are presented in Figure 3.4). Self-similarity, which can be exact for idealized mathematical objects, in case of shapes and patterns encountered in nature is typically approximate. Self-similarity can also be understood in statistical sense - a fractal can be an object whose certain statistical properties are the same in all scales (or in a wide range of scales). In order to refer to different types of self-similarity, fractals are sometimes distinguished into *deterministic* fractals and *random* (or *stochastic*) fractals. Self-similarity implies lack of characteristic scale, and is therefore expressed by power laws.

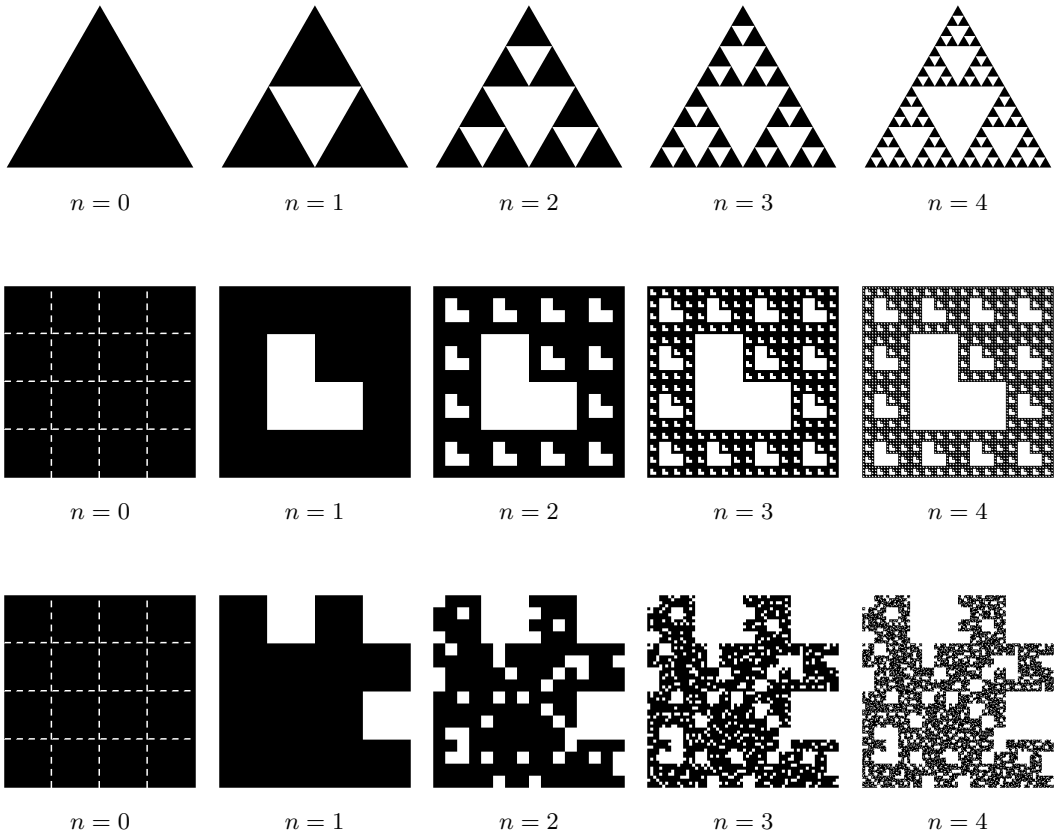
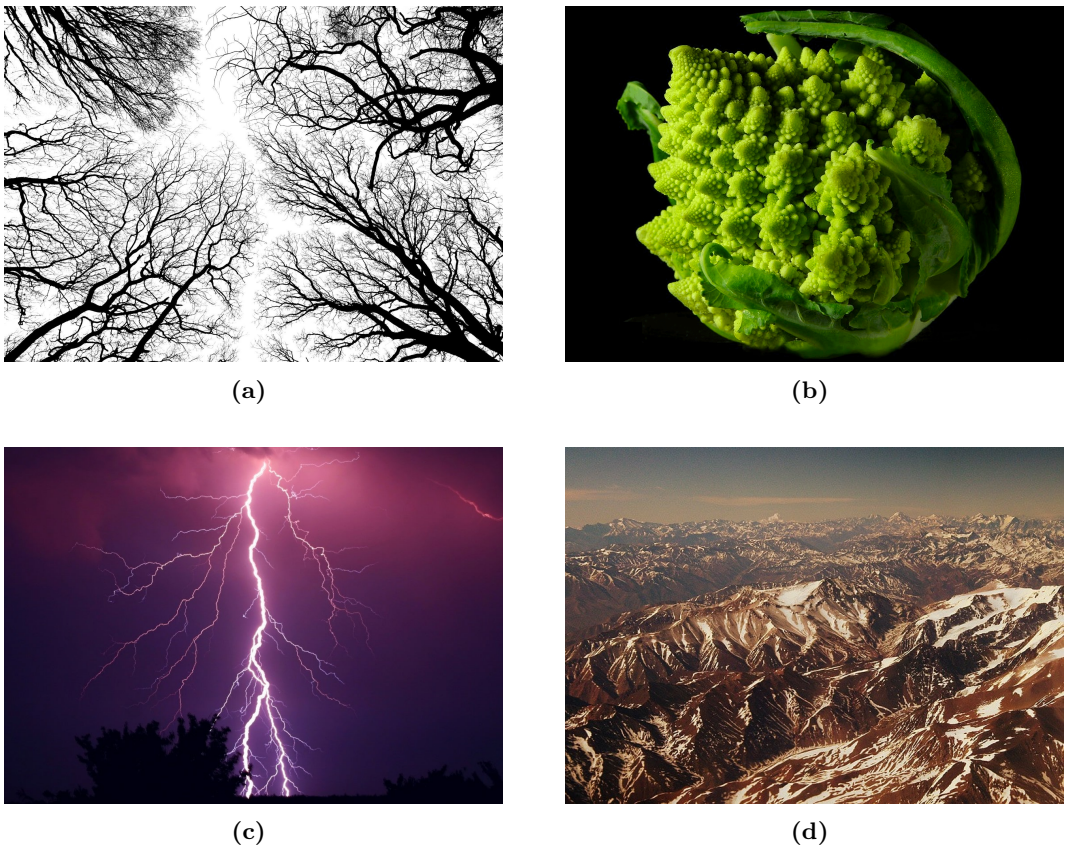


Figure 3.3. Examples of fractals generated by recursive removal of selected parts of a shape. The number of iterations is denoted by n. **First row:** The Sierpiński triangle generated by dividing an equilateral triangle into four identical parts, removing the central part, and repeating this step in the remaining triangles recursively. **Second row:** a fractal obtained by dividing a square into 16 identical parts, removing three of them in the way shown in the picture corresponding to $n = 1$, and applying this procedure recursively to the remaining parts. **Third row:** an example of a stochastic fractal, generated by the procedure same as the one producing the fractal in the second row, with the difference being the random choice of the three squares to remove at each stage of the process. Although the shapes generated in the second and in the third row look different, they have the same box-counting dimension (defined in subsection 3.4.2), equal to $\log(13)/\log(4) \approx 1.85$.



(a) **(b)**

(c) **(d)**

Figure 3.4. Examples of fractal patterns in nature - tree branches (a), Romanesco broccoli (b), the paths emerging at an electrical breakdown (c), mountain ridge system (d).

3.4.2 Fractal dimension

Self-similarity dimension

Fractals are often characterized with the use of fractal dimension. There are a few ways of defining fractal dimension, but all the variants express the same idea - they characterize the power law describing the structure of an object. A strictly self-similar object is composed of smaller copies of itself. If N is the number of such copies of a given size and $s < 1$ is the stretching factor of the similarity transformation which transforms the whole object into one of those copies, then the following relation holds:

$$N = (1/s)^{d_S}, \tag{3.25}$$

where d_S, the exponent of the power law describing the relationship between N and s, is called the *self-similarity dimension* of the studied object. Hence, d_S can be written as

$$d_S = \frac{\log N}{\log(1/s)}. \tag{3.26}$$

This quantity differs from topological dimension in that it can be a non-integer number, which is often (but not always) the case for fractals. For example, Sierpiński triangle consists of $N = 3$ copies of itself, each of the copies being the image of the whole triangle under similarity transformation with stretching factor $s = 1/2$, and therefore the self-similarity dimension of Sierpiński triangle is $d_S = \log(3)/\log(2) \approx 1.59$. For self-similar objects which are not fractals, self-similarity dimension is equal to topological dimension. For example, a square with side of length L can be divided into $N = 4$ squares with side of length $L/2$; the smaller squares are similar to the bigger one, and the stretching coefficient s is equal to $1/2$. Therefore, self-similarity

dimension of a square is $d_S = \log(4)/\log(2) = 2$, which coincides with its topological dimension.

Hausdorff dimension

A notion which is more general and applicable to objects that are not necessarily strictly self-similar, is *Hausdorff dimension*. Let F be a nonempty set in a metric space. Let $\mathrm{diam}(U)$ denote the diameter of a set U, that is

$$\mathrm{diam}(U) = \sup\{\rho(x,y) : x,y \in U\}; \quad \mathrm{diam}(\emptyset) = 0, \tag{3.27}$$

where ρ is the distance function (the metric), and $\emptyset$ is the empty set. For any $\delta > 0$, let the *δ-cover* of F be a countable collection of sets $\{U_i\}$ such that $F \subseteq \bigcup_i U_i$ and $0 < \mathrm{diam}(U_i) \leq \delta$ for all i. For any $s \geq 0$, the following quantity can be defined:

$$\mathcal{H}^s_\delta(F) = \inf\left\{\sum_i (\mathrm{diam}(U_i))^s : \{U_i\} \text{ is a } \delta\text{-cover of } F\right\}. \tag{3.28}$$

Obtaining $\mathcal{H}^s_\delta(F)$, requires covering F with sets U_i of diameters at most δ, and then finding the infimum of the sum of the s-th powers of diameters of U_i. $\mathcal{H}^s_\delta(F)$ is a non-increasing function of δ, because for any $\delta_1, \delta_2 \in \mathbb{R}_+$ such that $\delta_1 < \delta_2$, the set of all possible δ_2-covers of F contains the set of all possible δ_1-covers of F, and therefore the infimum in Eq. 3.28 for δ_2 can only be smaller or equal to infimum for $delta_1$. As a consequence, when δ decreases, $\mathcal{H}^s_\delta(F)$ does not decrease, and therefore when $\delta \to 0$, $\mathcal{H}^s_\delta(F)$ approaches a limit, finite or not. That limit, denoted by $\mathcal{H}^s(F)$, is called the *s-dimensional Hausdorff measure of F*:

$$\mathcal{H}^s(F) = \lim_{\delta \to 0} \mathcal{H}^s_\delta(F). \tag{3.29}$$

For sufficiently large s, the value of s-dimensional Hausdorff measure of F is equal to 0. The *Hausdorff dimension* of F (also commonly referred to as the *Hausdorff-Besicovitch dimension*), denoted by d_H, is defined as:

$$d_H = \inf\{s : s \geq 0 \wedge \mathcal{H}^s(F) = 0\}. \tag{3.30}$$

This can be understood as follows. For almost all choices of s, the s-dimensional Hausdorff measure of a set is either infinite or equal to zero. There exists a critical value of s, at which $\mathcal{H}^s(F)$ changes its value from ∞ to 0. This value of s, $s = d_H$, is the Hausdorff dimension. So it can be stated that for a set F, the Hausdorff dimension d_H is the number such that:

$$\mathcal{H}^s(F) = \begin{cases} \infty, & \text{for } s < d_H \\ 0, & \text{for } s > d_H. \end{cases} \tag{3.31}$$

The value of $\mathcal{H}^s(F)$ for $s = d_H$ is usually a finite number different from zero, but it also can be infinite or equal to zero; it is the point at which $\mathcal{H}^s(F)$ switches from infinity to zero that is important for the definition. The exception is the case in which $\mathcal{H}^s(F)$ is equal to zero for $s = 0$; since s is a non-negative number, this implies that there are no values of s such that $\mathcal{H}^s(F) = \infty$, and the Hausdorff dimension is then equal to zero.

Box-counting dimension

Hausdorff dimension is an important mathematical tool allowing to formalize the description of fractals. However, in practical applications, like describing the properties of fractals encountered in nature, another definition of fractal dimension is

widely used - the definition of the so-called *box-counting dimension*, also known as *capacity dimension* or *Minkowski dimension*. Let F denote a bounded nonempty set in a metric space; stating that F is bounded means that it is contained inside some (hyper)ball in the considered space. Let $N(\delta)$ denote the smallest possible number of balls of diameter δ needed to cover F; the cover of F is again understood as a collection of sets such that their union contains F, and the diameter of a ball is equal to its radius multiplied by 2. The box-counting dimension of F is defined as:

$$d_C = \lim_{\delta \to 0} \frac{\log\left(N(\delta)\right)}{\log(1/\delta)} \tag{3.32}$$

provided that the limit exists. When it does not exist, it is sometimes helpful to consider limit inferior and limit superior, which lead to the definitions of lower Minkowski dimension $\underline{d_C}$ and upper Minkowski dimension $\overline{d_C}$, respectively. The need to distinguish between $\overline{d_C}$, $\underline{d_C}$, and $\overline{d_C}$ happens rarely in practical applications, and usually d_C is used to characterize the studied object. The coverings of F do not necessarily have to be done with balls, other types of sets with given diameter or characteristic size directly related to diameter can be used. When the considered space is $\mathbb{R}^n$, one can use (hyper)cubes with the side of length δ. The sets used to cover the studied set F are often called *boxes*. The name - box-counting dimension - comes from the fact that d_C describes how the number of boxes $N(\delta)$ needed to cover the considered object changes with the size δ of the boxes. From Eq. 3.32 it can be seen that d_C is the exponent of the power law of the form:

$$N(\delta) \sim C(1/\delta)^{d_C} \quad \text{as } \delta \to 0 \tag{3.33}$$

where C can be treated as a constant (technically, Eq. 3.32 allows C to be a function of δ varying sufficiently slowly, namely a function $C = C(\delta)$ such that $\log(C(\delta))/\log(1/\delta) \to 0$ as $\delta \to 0$). The importance of box-counting dimension is due to the relative ease of its use in numerical computations - to estimate the fractal dimension of an object, one needs to cover the object with a (possibly small) number of boxes of given size δ, repeat the procedure for various δ, and determine the exponent of the power-law relationship (holding for small δ) between the size δ of a box and the number $N(\delta)$ of the boxes used (Eq. 3.33). It should be emphasized, however, that estimating the box-counting dimension of objects not being idealized mathematical constructions (like fractals occurring in nature) might be prone to quite large numerical errors, which can be a result of the limited amount of relevant data (an insufficiently detailed representation of an object, for instance) or the slow convergence of $\log(N(\delta))/\log(1/\delta)$. An illustration of the procedure of estimating box-counting dimension is shown in Figure 3.5.

Similarly to other presented fractal dimensions, box-counting dimension can have a non-integer value and it gives results agreeing with intuition when applied to non-fractal sets in $\mathbb{R}^n$ - objects like a line segments, polygons, or polyhedrons have box-counting dimension equal to their topological dimension.

Correspondence between different fractal dimensions

It is worth noting that self-similarity dimension can be considered a simplified variant of box-counting dimension, suited to strictly self-similar objects. A strictly self-similar object can be covered by boxes in such a way that each box contains one piece being the image of the whole object under similarity transformation with stretching factor $s < 1$; in such a case, the number of boxes required for covering is equal to the number of the pieces. If δ_0 is a constant such that the whole studied shape fits in the box of size δ_0, then a piece of the whole shape being its copy scaled

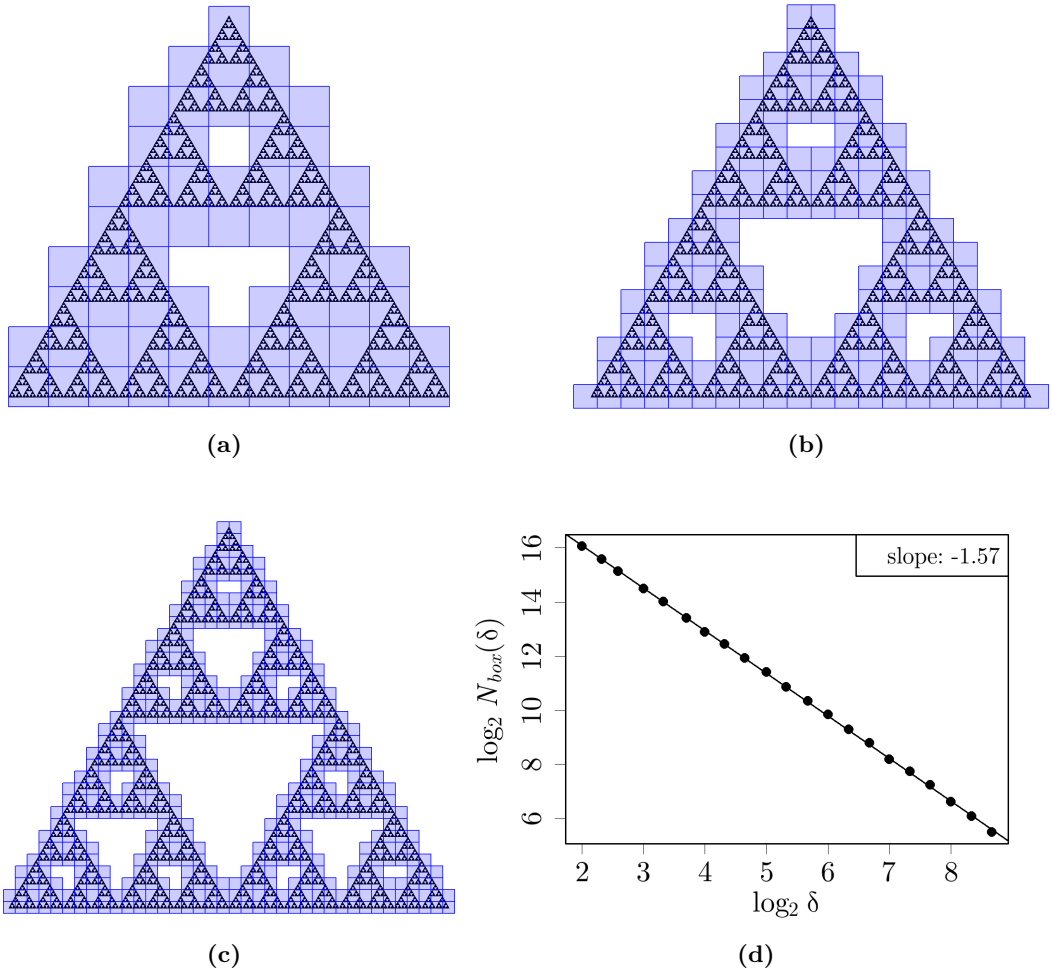


Figure 3.5. Estimating the box counting dimension of the Sierpiński triangle. (a), (b), (c) - three examples of coverings with boxes of different sizes (boxes are marked in blue); (d) - the log-log plot of the number of boxes as a function of box size; the dimension estimated from the slope of the fitted line, equal to 1.57, is close to the true value of the fractal dimension, $d_C = \log(3)/\log(2) \approx 1.59$.

by the stretching factor $s = s_0^k$, where $s_0 < 1$ and $k = 1, 2, 3, ...$, fits in the box of size $\delta_0 s_0^k$. Therefore, to calculate the box-counting dimension of a strictly self-similar object, one can use the sequence of coverings with box sized given by a sequence δ_k of the form:

$$\delta_k = \delta_0 s_0,\ \delta_0 s_0^2,\ \delta_0 s_0^3, ... = \delta_0 s_0^k \quad \text{for } k = 1, 2, 3, ... \tag{3.34}$$

The number of downscaled copies of the whole shape obtained by applying similarity transformation with stretching factor s_0^k, can be determined from Eq. 3.25, by inserting $s = s_0^k$. This number is equal to the number $N(\delta_k)$ of boxes of size δ_k needed to cover the shape. Hence $N(\delta_k) = (1/s_0^k)^{d_S}$, where d_S is the self-similarity dimension. The ratio $\log(N(\delta_k))/\log(1/\delta_k)$ can therefore be expressed as:

$$\begin{aligned} \frac{\log\left(N(\delta_k)\right)}{\log\left(1/\delta_k\right)} &= \frac{\log\left((1/s_0)^{k\,d_S}\right)}{\log\left(1/(\delta_0 s_0^k)\right)} = d_S\,\frac{k\log(1/s_0)}{k\log\left(1/(\delta_0^{1/k} s_0)\right)} = \\ &= d_S\,\frac{\log(1/s_0)}{\log(1/s_0) + (1/k)\log(1/\delta_0)}, \end{aligned} \tag{3.35}$$

where in the second equality the identity $\delta_0 = \left(\delta_0^{1/k}\right)^k$ is used. The limit of the above expression as $k \to \infty$ (which corresponds to $\delta_k \to 0$), defines the box-counting dimension d_C. Since the value of that limit is equal to d_S, one can write: $d_C = d_S$.

So, the self-similarity dimension of a strictly self-similar object can be interpreted as the box-counting dimension determined with the specific choice of boxes which simplifies the calculation.

In most cases relevant from the standpoint of practical computations, box-counting dimension d_C and Hausdorff dimension d_H are equal and referred to simply as "fractal dimension", but in principle they satisfy $d_H \leq d_C$. They both might exceed topological dimension d_T, therefore in general d_T, d_H and d_C satisfy [266]:

$$d_T \leq d_H \leq d_C. \tag{3.36}$$

Fractal dimension can be thought of as a way of expressing the information about how the characteristics of an object change when inspected at different scales - for example, the box counting dimension describes how the number of boxes needed to cover the object changes with the changing size of the boxes. Fractal dimension is often summarized as a quantity expressing the "complexity" of a shape, understood as "roughness", or the capacity to fill the space that the shape is embedded in (for example, comparing two curves in $\mathbb{R}^2$ having different fractal dimensions, one can typically observe that the one with the lower fractal dimension looks more smooth while the one with the higher fractal dimension is more "wiggled" or "jagged').

3.4.3 Multifractals

Although fractal dimension can give insight into the properties of various shapes and patterns, in many situations the information provided by examining just the fractal dimension itself is insufficient. An obvious reason for that is that fractal dimension does not give information about the structural details or about the process that generated the object; many different fractals can have the same fractal dimension. Also, fractal dimension characterizes a shape as a whole; however, there exist objects whose different parts have different local properties. To describe the so-called *multifractals*, more general tools are required. Multifractals can be thought of as objects in which multiple different fractal structures are entangled. With the use of appropriate formalism, these structures can be identified and their contribution to the structure of the whole multifractal can be quantified [194, 265–268]. While the mathematical description of fractals utilizes the notion of a set, the description of mutltifractals is formalized in terms of *measures* (a measure can be thought of as a function specifying how some non-negative quantity is distributed over some space).

Let μ be a measure in $\mathbb{R}^n$; μ can represent a quantity like mass, probability (provided that μ is normalized to 1), or electric charge (technically, it should be taken into account that electric charge can be negative and a measure is always non-negative). Let $\mathrm{supp}(\mu)$ denote the support of the measure μ, that is, the largest possible set in the considered space such that every open neighborhood of every point of the set has positive measure. Let μ be distributed in such a way that around an arbitrary point $x_0 \in \mathrm{supp}(\mu)$ it satisfies:

$$\mu\left(K(x_0, \varepsilon)\right) \sim C_\mu \, \varepsilon^{\alpha(x_0)} \quad \text{for } \varepsilon \to 0, \tag{3.37}$$

where $K(x_0, \varepsilon)$ is the (hyper)cube of side length ε centered at x_0, $\mu\left(K(x_0, \varepsilon)\right)$ is the measure of $K(x_0, \varepsilon)$, $\alpha(x_0)$ is a non-negative real number, and C_μ is a constant (independent of x_0 and ε). This means that for $\varepsilon \to 0$, the distribution of measure μ around x_0 is given by a power law with exponent $\alpha(x_0)$. This exponent, called *singularity exponent* or *Hölder exponent*, describes the "strength" of singularity of μ around x_0 - the lower the exponent, the more singular the measure is; the limit $\alpha(x_0) = 0$ corresponds to a behavior similar to a Dirac delta at x_0. Conversely, the

greater the $\alpha(x_0)$, the more uniform the measure is around x_0. From Eq. 3.37 it can be seen that the singularity exponent can be defined as:

$$\alpha(x_0) = \lim_{\varepsilon \to 0} \frac{\log \mu(K(x_0, \varepsilon))}{\log \varepsilon}. \tag{3.38}$$

An equivalent definition can be obtained for $K(x_0, \varepsilon)$ replaced with another type of set, for example a (hyper)ball with diameter ε, centered at x_0. The values of the singularity exponent may be different for different points in space, and therefore α is a function on x_0. For any value of α one can define a set E_α being a subset of $\mathrm{supp}(\mu)$ such that all the points in E_α have the singularity exponent equal to α:

$$E_\alpha = \{x : x \in \mathrm{supp}(\mu) \wedge \alpha(x) = \alpha\}. \tag{3.39}$$

For each α, the set E_α can be characterized by its Hausdorff dimension; this dimension is denoted by $f(\alpha)$:

$$f(\alpha) = d_H(E_\alpha) \tag{3.40}$$

(here, $d_H(\cdot)$ denotes the operation of determining the Hausdorff dimension of a set). So $f(\alpha)$ is a function which assigns to each α the fractal dimension of the set of points having singularity exponent equal to α. The set of pairs $(\alpha, f(\alpha))$ for all $\alpha \in (\alpha_{\min}; \alpha_{\max})$ (that is, for all α occurring in a system) is called the *singularity spectrum*. It shows how the whole structure of the studied quantity's distribution is composed of multiple intertwined components, each of them having its own singularity exponent and fractal properties. When the whole system is characterized by the same singularity exponent α_0, then the singularity spectrum reduces to a single point: $(\alpha_0, f(\alpha_0))$; the value of $f(\alpha_0)$ depends of the fractal dimension of the support of the measure. A singularity spectrum reduced to a single point corresponds to a measure that is called *homogeneous* or *monofractal*. Conversely, a measure whose singularity spectrum consists of many points - indicating that there is some range of singularity exponents - is called *multifractal*. Objects and quantities described by measures of the presented types - are referred to as *monofractals* and *multifractals*, respectively. A term often appearing in relation to multifractality is *multiscaling*; it refers to the fact that different parts of a multifractal object exhibit different types of scaling behavior. In typically studied cases, the singularity spectrum is a concave function, spanned between some finite values $\alpha_{\min}$ and $\alpha_{\max}$, having the maximum value $\max\{f(\alpha)\}$ equal to $d_H(\mathrm{supp}(\mu))$, and assuming a shape resembling "an inverted U" (Figure 3.6) [194, 196, 265–267]; this is the case considered here, but it is worth mentioning that it is possible to consider objects with singularity spectra of different shapes [269–272]. The width of the singularity spectrum $\Delta\alpha = \alpha_{\max} - \alpha_{\min}$ expresses the variety of the singularity exponents - and therefore a wide spectrum is often a sign of a certain kind of complexity.

Detecting and quantifying multifractality in empirical data based on the definitions given above typically suffers from large errors. Therefore an approach based on the so-called *partition function* is utilized. The space $\mathbb{R}^n$ can be divided into (hyper)cubic cells of side length ε. The cells containing any points belonging to $\mathrm{supp}(\mu)$ are then numbered by $i = 1, 2, 3, ..., N(\varepsilon)$; the measure μ contained in the i-th cell is denoted by $\mu_i(\varepsilon)$. The partition function $Z(q, \varepsilon)$ for $q \in \mathbb{R}$ is defined as:

$$Z(q, \varepsilon) = \sum_{i=1}^{N(\varepsilon)} \mu_i(\varepsilon)^q. \tag{3.41}$$

The behavior the partition function in the limit $\varepsilon \to 0$ and with fixed q is given by a power law:

$$Z(q, \varepsilon) \sim C_Z \, \varepsilon^{\tau(q)} \quad \text{for } \varepsilon \to 0 \text{ and fixed } q. \tag{3.42}$$

Here C_Z is a constant (independent of ε) and $\tau(q)$ is an exponent in general dependent on q, called the *(generalized) scaling exponent* or the *mass exponent*. To relate q and $\tau(q)$ to singularity spectrum, a following line of reasoning can be utilized.

Determining $Z(q,\varepsilon)$ from the definition given above relies on calculating the q-th power of μ in each of the boxes and summing the results. But $Z(q,\varepsilon)$ can be expressed in another way. It can be informally stated, that to calculate $Z(q,\varepsilon)$ one can calculate the q-th power of the measure $\mu(K(x,\varepsilon))$ contained in a cell $K(x,\varepsilon)$ of size ε centered at an arbitrary point $x \in \mathrm{supp}(\mu)$ such that its Hölder exponent is equal to α, multiply the result by the number of cells characterized by the Hölder exponent equal to α, and then integrate over all possible α. More precisely, $Z(q,\varepsilon)$ can be approximated as follows:

$$Z(q,\varepsilon) \approx \int_{\alpha_{\min}}^{\alpha_{\max}} C_1\, \mu(K(x,\varepsilon))^q\, \widetilde{\rho}(\alpha,\varepsilon)\, d\alpha, \tag{3.43}$$

where C_1 is a constant and $\widetilde{\rho}(\alpha,\varepsilon)$ can be interpreted as the distribution of α, inspected at scale ε; in other words, $\widetilde{\rho}(\alpha,\varepsilon)d\alpha$ is the probability that a randomly chosen cell $K(x,\varepsilon)$ with non-zero measure has the *coarse Hölder exponent* between α and $\alpha + d\alpha$ (the coarse Hölder exponent $\widetilde{\alpha}$ of a cell $K(x,\varepsilon)$ of size ε is defined as $\widetilde{\alpha} = \log \mu(K(x,\varepsilon))/\log \varepsilon$). The support of the measure consists of sets E_α with singularity exponents α and fractal dimensions $f(\alpha)$, that is $\mathrm{supp}(\mu) = \bigcup_\alpha E_\alpha$, where α runs over the continuum of possible values between $\alpha_{\min}$ and $\alpha_{\max}$. The number of cells of size ε needed to cover E_α behaves as $\varepsilon^{-f(\alpha)}$. Hence, $\widetilde{\rho}(\alpha,\varepsilon)$ can be written as $\widetilde{\rho}(\alpha,\varepsilon) \propto \rho(\alpha)\varepsilon^{-f(\alpha)}$, where $\rho(\alpha)$ can be interpreted as a function assigning weights to sets E_α according to their contribution to $\mathrm{supp}(\mu)$; in other words, $\rho(\alpha)d\alpha$ can be understood as the probability that a randomly chosen set $E_{\alpha'}$ has the singularity exponent α' between α and $\alpha + d\alpha$. By inserting $\widetilde{\rho}(\alpha,\varepsilon) \propto \rho(\alpha)\varepsilon^{-f(\alpha)}$ and $\mu(K(x,\varepsilon)) \approx C_\mu \varepsilon^\alpha$ (Eq. 3.37), into Eq. 3.43, one gets:

$$Z(q,\varepsilon) \approx \int_{\alpha_{\min}}^{\alpha_{\max}} C_2 \rho(\alpha) \varepsilon^{\alpha q - f(\alpha)} d\alpha, \tag{3.44}$$

where C_2 is a constant. In the limit $\varepsilon \to 0$, the above integral is dominated by the values of the integrand corresponding to the lowest values of the exponent of ε. Therefore the behavior of $Z(q,\varepsilon)$ for $\varepsilon \to 0$ and fixed q can be summarized as:

$$Z(q,\varepsilon) \sim C_3 \varepsilon^{\alpha q - f(\alpha)} \quad \text{with } \alpha \text{ minimizing } (\alpha q - f(\alpha)), \tag{3.45}$$

where C_3 is a constant. The value of α minimizing $(\alpha q - f(\alpha))$ can be found as the solution of the equation:

$$\frac{\partial}{\partial \alpha}\left(\alpha q - f(\alpha)\right) = 0, \tag{3.46}$$

from which one concludes that $\alpha(q)$, the searched value of α, satisfies:

$$\left.\frac{df}{d\alpha}\right|_{\alpha=\alpha(q)} = q. \tag{3.47}$$

Comparing Eq. 3.42 with Eq. 3.45, gives

$$\tau(q) = q\alpha(q) - f(\alpha(q)). \tag{3.48}$$

This equation, along with Eq. 3.47, allows to derive formulas expressing $(q,\tau(q))$ in terms of $(\alpha, f(\alpha))$ and expressing $(\alpha, f(\alpha))$ in terms of $(q,\tau(q))$:

$$\begin{cases} (1)\ q(\alpha) = \dfrac{df}{d\alpha} \\ (2)\ \tau(q(\alpha)) = \alpha q(\alpha) - f(\alpha) \end{cases} \qquad \begin{cases} (3)\ \alpha(q) = \dfrac{d\tau}{dq} \\ (4)\ f(\alpha(q)) = q\alpha(q) - \tau(q). \end{cases} \tag{3.49}$$

The interchangeability of the descriptions of a multifractal in terms of $(\alpha, f(\alpha))$ and in terms of $(q, \tau(q))$ facilitates practical calculations; $\tau(q)$ is often easier to compute numerically. Another way of characterizing a multifractal employs the so-called *generalized fractal dimensions* D_q, which can be defined as:

$$D_q = \begin{cases} \dfrac{\tau(q)}{q-1}, & \text{for } q \neq 1 \\ \lim\limits_{q\to 1} \dfrac{\tau(q)}{q-1} = \lim\limits_{\Delta q \to 0} \dfrac{\tau(1+\Delta q)}{\Delta q} = \left.\dfrac{d\tau}{dq}\right|_{q=1}, & \text{for } q = 1. \end{cases} \tag{3.50}$$

A number of properties can be demonstrated using the above presented equations. If the singularity spectrum consists of a single point (the measure is monofractal), then $d\tau/dq$ is constant, and $\tau(q)$ is linear; nonlinearity of $\tau(q)$ indicates multifractality. For large q, the value of $Z(q,\varepsilon)$ is dominated by the contributions coming from cells i with high μ_i; conversely, low q in principle "selects" the cells with low μ_i. Therefore, for different values of q, $\tau(q)$ characterizes the scaling behavior within the cells of different measure. For $q = 0$, $Z(q,\varepsilon)$ is the number of boxes required to cover $\mathrm{supp}(\mu)$; therefore using Eq. 3.41, Eq. 3.42, and Eq. 3.50 one can identify $-\tau(0) = D_0$ as the fractal dimension of the support of the measure $d_C(\mathrm{supp}(\mu))$, which is also the maximum value of $f(\alpha)$:

$$-\tau(0) = D_0 = d_C(\mathrm{supp}(\mu)) = \max\{f(\alpha)\}. \tag{3.51}$$

The concavity of $f(\alpha)$ can be seen by noticing that $\alpha(q)$ is found as the value minimizing $(\alpha q - f(\alpha))$ in Eq. 3.44; this implies that the second derivative of $(\alpha q - f(\alpha))$ is greater than 0; setting $\partial^2/\partial\alpha^2(\alpha q - f(\alpha)) > 0$ yields $d^2f/d\alpha^2 < 0$ for any $\alpha(q)$, which shows that $f(\alpha)$ is concave. This also allows to notice that $q(\alpha)$ is a decreasing function of α; differentiating the first formula in Eq. 3.49 with respect to α yields $dq/d\alpha = d^2f/d\alpha^2$, which is less than 0; using the derivative of an inverse function, one concludes that also $\alpha(q)$ is a decreasing function of q.

The generalized fractal dimensions D_q for some specific values of q have specific interpretations. As mentioned above, D_0 is the fractal dimension of the support of the measure. For $q = 1$, $Z(q,\varepsilon) = \sum_{i=1}^{N(\varepsilon)} \mu_i = \mu(\mathrm{supp}(\mu))$; therefore $\tau(1) = 0$. Inserting this into the first and the second formula in Eq. 3.49 gives $1 = df/d\alpha|_{\alpha(q=1)}$ and $f(\alpha(q{=}1)) = \alpha(q{=}1)$, respectively. Using the definition of $\tau(q)$, that is $\tau(q) = \lim_{\varepsilon\to 0}\left((\log\sum_{i=1}^{N(\varepsilon)}\mu_i(\varepsilon))/\log\varepsilon\right)$, to calculate $d\tau/dq$, one obtains:

$$D_1 = \alpha(q{=}1) = \left.\frac{d\tau}{dq}\right|_{q=1} = \lim_{\varepsilon\to 0} \frac{\sum\limits_{i=1}^{N(\varepsilon)} \mu_i(\varepsilon)\log\mu_i(\varepsilon)}{\log\varepsilon}. \tag{3.52}$$

If the measure is normalized ($\mu(\mathrm{supp}(\mu)) = 1$), it can be interpreted as probability, and then the expression $-\sum_{i=1}^{N(\varepsilon)}\mu_i(\varepsilon)\log\mu_i(\varepsilon)$ is the *entropy* of the measure distributed over cells of size ε. Thus, D_1 describes how the entropy changes with the change of ε, and is often called the *information dimension*. Setting $q = 2$, gives D_2, known as the *correlation dimension*:

$$D_2 = \tau(q{=}2) = \lim_{\varepsilon\to 0} \frac{\log\sum\limits_{i=1}^{N(\varepsilon)} \mu_i(\varepsilon)^2}{\log\varepsilon}. \tag{3.53}$$

If the measure is normalized and represents the density of points in space - meaning that the probability that a randomly chosen point belongs to the i-th cell of size ε is proportional to $\mu_i(\varepsilon)$, then the expression $\sum_{i=1}^{N(\varepsilon)}\mu_i(\varepsilon)^2$ can be interpreted as the

probability that two randomly chosen points belong to the same cell of size ε. So, D_2 describes the behavior of that probability for varying ε. Dividing both sides of the second formula in Eq. 3.49 by $(q-1) \neq 1$ yields:

$$\frac{\tau(q)}{q-1} = \alpha(q)\frac{q}{q-1} - \frac{f(\alpha)}{q-1}. \tag{3.54}$$

Since $f(\alpha)$ is bounded both from above and below, taking the limits $q \to +\infty$ and $q \to -\infty$ gives:

$$\begin{aligned} D_{+\infty} &= \alpha(q{=}+\infty) = \alpha_{\min}, \\ D_{-\infty} &= \alpha(q{=}-\infty) = \alpha_{\max}, \end{aligned} \tag{3.55}$$

which means that $D_{+\infty}$ and $D_{-\infty}$ characterize the weakest and the strongest singularity, respectively. An exemplary shape of a singularity spectrum of a multifractal object, with some characteristic points described above, is shown in Figure 3.6.

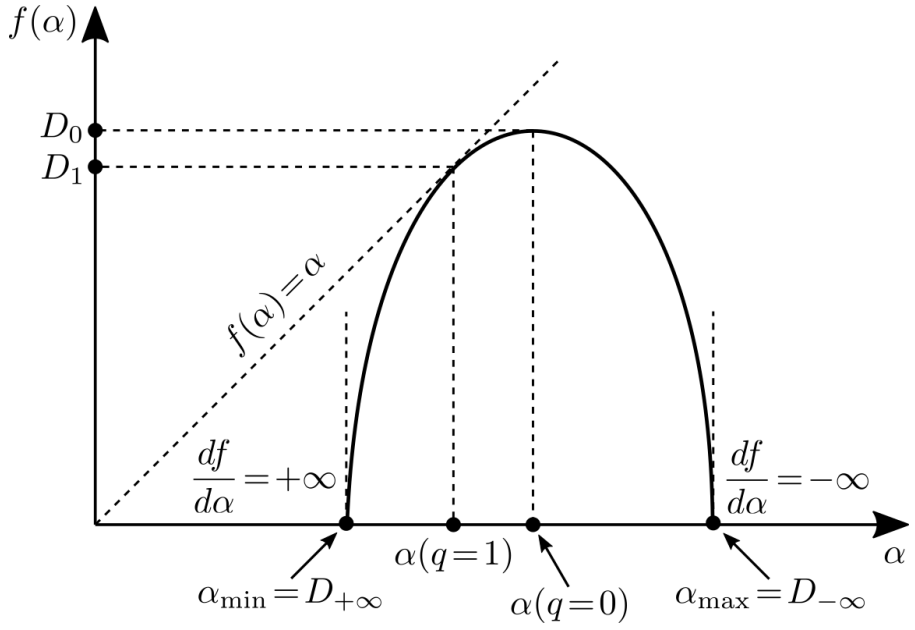


Figure 3.6. A sketch of a typical singularity spectrum of a multifractal object, with some characteristic points marked on the plot.

3.4.4 Fractals and multifractals in time series

Hurst exponent and fractality

Among the examples of objects which are often studied with the use of (multi)fractal analysis are signals and time series. The complexity of signals can be manifested by the presence of properties which can be attributed to fractality or multifractality. Identifying such properties in a signal typically allows to characterize some aspects of the process generating that signal, for example multifractality is often linked to heavy-tailed distributions and the presence of nonlinear correlations (correlations that are not captured by autocovariance function or spectral density) [148, 273–275].

One of the methods of investigating the behavior of a signal in various scales utilizes the notion of the Hurst exponent. Hurst exponent is related to fractality in a few ways. For example, the fractal dimension d_C of a fractional Brownian motion with Hurst exponent H is given by: $d_C = 2 - H$ (d_C can be interpreted as the fractal dimension of the graph of the function - it describes how the number of boxes required to cover the line representing the trajectory of the process changes

with changing box size; it is worth noting that there are also other ways of defining the fractal dimension of a signal) [193, 276]. According to Eq. 3.20, the Hurst exponent of a time series is the exponent of the power law describing how the variability of the series behaves in various time scales; as the relationship has the form of a power law, it can be said to be scale-free. Lack of characteristic scale can also be considered in the following sense. If, for any positive constant λ, a process represented by a function $X(t)$ satisfies [256, 276]:

$$X(t_0 + \lambda t) - X(t_0) \stackrel{d}{=} \lambda^c (X(t_0 + t) - X(t_0)), \tag{3.56}$$

where "$\stackrel{d}{=}$" denotes the equality of probability distributions, then the process is called *self-affine* or (less precisely) *self-similar*; c is called *self-similarity index* or the Hurst exponent, as its value is equal to the Hurst exponent (defined as in Eq. 3.20) for certain processes - fractional Brownian motion, for instance. The interpretation of the above equation can be expressed as follows: if a process represented by a signal $X(t)$ is self-similar and has self-similarity index c, then rescaling the argument t by some factor λ together with rescaling the values of $X(t)$ by the factor $1/\lambda^c$ gives a process statistically indistinguishable from $X(t)$.

Multifractal Detrended Fluctuation Analysis

There are a few methods of detecting and quantifying fractality and multifractality in time series. The method used in this work is Multifractal Detrended Fluctuation Analysis (MFDFA) [277]. MFDFA is a generalization of a simpler method - Detrended Fluctuation Analysis (DFA) [278, 279], designed to estimate the Hurst exponent of a series; MFDFA contains DFA as a special case. It allows to estimate both the singularity spectrum and the Hurst exponent of a time series. An important feature of MFDFA is that it allows to analyze non-stationary series; it removes trends from the data and focuses on the fluctuations around the trends. The method can be divided into a few steps, which are listed below.

Let $x(1), x(2), x(3), ..., x(N)$ be a time series with real values. The first step of MFDFA is computing the *profile* of the series, that is, the cumulative series:

$$y(k) = \sum_{i=1}^{k} x(i). \tag{3.57}$$

The second step is dividing the range of the variable representing time into $N_s = \lfloor N/s \rfloor$ non-overlapping segments of equal length s (the notation $\lfloor \cdot \rfloor$ represents the floor function; a segment of length s starting at some time step k consists of time series' indices $k, k+1, k+2, ..., k+s-1$). To avoid disregarding any piece of the data, two partitions are done - one starting from $k = 1$ and one starting from the opposite end of the series; this gives $2N_s$ (overlapping) segments in total. The segments are then numbered by $\nu = 1, 2, 3, ..., 2N_s$; the set of indices corresponding to the segment number ν is denoted by I_ν.

The third step starts from *detrending* - a local polynomial trend p_ν is computed for each segment I_ν, using the least squares method and the detrended cumulative series $y(k) - p_\nu(k)$ is constructed. The order of the chosen polynomial influences the shape of the trend that can be removed from the data; common choices are linear, quadratic, and cubic polynomials. After detrending, the quantity $F^2(\nu, s)$, called variance, is computed for each segment I_ν:

$$F^2(\nu, s) = \frac{1}{s} \sum_{k \in I_\nu} (y(k) - p_\nu(k))^2 . \tag{3.58}$$

In the fourth step, a single value of the *q-th order fluctuation function* $F_q(s)$ for a given s is computed as the power mean of order q of the square roots of the variances $F^2(\nu, s)$:

$$F_q(s) = \begin{cases} \left(\frac{1}{2N_s} \sum_{\nu=1}^{2N_s} \left(F^2(\nu, s) \right)^{q/2} \right)^{1/q}, & \text{for } q \neq 0 \\ \exp\left(\frac{1}{4N_s} \sum_{\nu=1}^{2N_s} \log\left(F^2(\nu, s) \right) \right), & \text{for } q = 0 \end{cases} \tag{3.59}$$

(here $\exp(\cdot)$ and $\log(\cdot)$ denote the exponential function and natural logarithm, respectively). This is done for multiple values of q; typically one uses an equally spaced sequence of q values in some range centered at 0, for example in the interval $[-4; 4]$. $F_q(s)$ characterizes the fluctuations of given magnitude at a given scale; the main contribution to $F_q(s)$ for strongly negative q comes from small fluctuations, and for large q the largest fluctuations are "amplified".

The steps from the second to the fourth need to be repeated for different segment lengths s chosen from some range $[s_{\min}; s_{\max}]$; the choice depends on the studied data, but it is often suggested that $s_{\min}$ should be not less than 10 and $s_{\max}$ not greater than $N/5$. With that procedure, the set of values of $F_q(s)$ for different q and s is obtained. Then the scaling behavior of $F_q(s)$ is investigated; if within the studied range of s the q-th order fluctuation function is described by a power law of the form:

$$F_q(s) \approx C s^{h(q)}, \tag{3.60}$$

where C is a constant, then $h(q)$ is the *generalized Hurst exponent*, dependent on q. For $q = 2$, $h(q)$ is identical to the "ordinary" Hurst exponent, $h(2) = H$. If the only considered value of q is $q = 2$, MFDFA reduces to DFA (Detrended Fluctuation Analysis). The sequence of generalized Hurst exponents $h(q)$ characterizes the scaling of the fluctuations of different magnitudes. If h is independent of q (that is, $h(q)$ is constant), the series can be characterized by a single scaling relationship; if there is a significant dependence of h on q, then the series exhibits multiscaling.

The relationship between the generalized Hurst exponents and the description of multifractals based on the partition function and singularity spectrum can be demonstrated on an example of a stationary, normalized series with non-negative values. If $x(1), x(2), x(3), ..., x(N)$ is a time series generated by a covariance stationary process, satisfying $x(i) \geq 0$ and $\sum_{i=1}^{N} x(i) = 1$, then the analysis of fluctuations does not require detrending, as there is no trend to remove. In such a case, quantifying the fluctuations within a segment I_ν can be done with a simplified definition of variance $F^2(\nu, s)$:

$$F^2(\nu, s) = \left(y(\max I_\nu) - y(\min I_\nu) \right)^2. \tag{3.61}$$

In the above definition, $\min I_\nu$ is the smallest index in I_ν (the left end of the segment), and $\max I_\nu$ is the largest index in I_ν (the right end of the segment). Here $F^2(\nu, s)$ is the squared sum of the time series' values in the segment I_ν. Inserting this form of $F^2(\nu, s)$ into the definition of fluctuation function yields:

$$F_q(s) = \left(\frac{1}{2N_s} \sum_{\nu=1}^{2N_s} |y(\max I_\nu) - y(\min I_\nu)|^q \right)^{1/q} \quad \text{for } q \neq 0. \tag{3.62}$$

Identifying the expression $\sum_{\nu=1}^{2N_s} |y(\max I_\nu) - y(\min I_\nu)|^q$ as the partition function $Z(q, s)$ of a certain measure and assuming scaling of $F_q(s)$ in the form given by

Eq. 3.60 allows to state that

$$\left(\frac{1}{2N_s}Z(q,s)\right)^{1/q} \sim Cs^{h(q)}. \tag{3.63}$$

Noticing that $N_s \approx N/s$ and using Eq. 3.42 gives

$$C_1 s^{\tau(q)} \sim C_2 s^{qh(q)-1}, \tag{3.64}$$

where C_1 and C_2 are constants. Hence, the relationship between $\tau(q)$ and $h(q)$ is given by: $\tau(q) = qh(q) - 1$. Using this result and Eq. 3.49, one can derive the equations for computing the singularity spectrum from generalized Hurst exponents:

$$\begin{cases} \alpha = h(q) + q\dfrac{dh}{dq} \\ f(\alpha) = q(\alpha - h(q)) + 1. \end{cases} \tag{3.65}$$

3.5 Entropy in written language

Samples of written language can be naturally represented in terms of symbol sequences. In the most straightforward approach, a text can be treated as a sequence of letters (and spaces). Another approach is to consider words as individual symbols - a text becomes then a sequence of words. Determining the entropy of such a sequence gives an insight into how much, on average, the occurrence of a word in a text is determined by the specific word sequence preceding the considered word. However, instead of using the entropy itself, one can also study the difference between the entropy H_{orig} of the original text and the entropy H_{rand} of the same text, but with words shuffled randomly. In a symbolic sequence in which the order of the symbols is random, the entropy is determined purely by the distribution of symbol frequencies. Hence, the difference $H_{rand} - H_{orig}$ provides information about the decrease in text's entropy caused by the specific order of words, compared to the entropy which would be observed if words were placed at random in the text. The usefulness of the quantity $H_{rand} - H_{orig}$, which in the considered context is referred to as *relative entropy*, is due to the fact that it allows to remove the influence of purely frequency-based effects. For example, if words are not lemmatized in the analysis (as is the case here), texts in languages with extensive use of inflection typically have more unique words (symbols) than texts in languages in which inflection is less developed; this influences the frequency distribution - and consequently, the entropy - but since the distribution is the same in the original and in the randomized text, the relative entropy can be anticipated to capture only the effects related to word ordering. Relative entropy has been reported in the literature [280, 281] to be approximately independent of language (this was tested on corpora in a number of languages), with values ranging from about 3 bits per word to about 4 bits per word. This suggests that despite the differences between the grammars and the vocabularies of individual languages, the amount of "order" contained in how words are placed with respect to each other is to a certain degree universal across languages.

Figure 3.7 shows the entropy (estimated using the estimator given by Eq. 3.3) of texts from dataset specified in Appendix B.1. Each text is considered in both its original form and in a randomized variant (with random word order), giving two values: the entropy of the original text H_{orig} and the entropy of the randomized text H_{rand}. The figure also shows the entropy for the same texts computed with

punctuation marks taken into account and included into the analysis on the same terms as words. Data sequences including punctuation marks are in two variants as well - the original and the randomized one - giving rise to two values of entropy, H_{orig}^{punct}, H_{rand}^{punct}, respectively. Additional quantities: $H_{rand} - H_{orig}$, $H_{rand}^{punct} - H_{orig}^{punct}$, $H_{orig}^{punct} - H_{orig}$, $H_{rand}^{punct} - H_{rand}$, are also shown in the figure. The narrow distributions of $H_{rand} - H_{orig}$ and $H_{rand}^{punct} - H_{orig}^{punct}$ confirm the mentioned result reported in the literature - that the values of relative entropy of word ordering are concentrated in the range between 3 and 4 bits per word. The distribution of $H_{orig}^{punct} - H_{orig}$ (all values slightly below 0) indicates that taking punctuation marks results in a decrease of text's entropy. This could lead to a conclusion that punctuation organizes written language in a manner that lowers the "randomness" of a text; however, since practically the same effect is observed for randomized texts ($H_{rand}^{punct} - H_{rand}$ is also slightly below 0), the decrease of entropy can be attributed to changes in frequencies of individual symbols introduced by including punctuation marks into the analysis. This can be understood with the help of results regarding how punctuation influences the shape of word frequency distribution in texts (Fig. 2.7). Treating punctuation marks as words brings the shape of word frequency distribution closer to the one specified by a power law - the frequencies of the most frequent symbols (words or punctuation marks) become higher than in the original distribution. The distribution becomes thus less "uniform" - as the discrepancy between the highest and the lowest frequencies increases. This results in a decrease of distribution's entropy (as entropy is maximized for maximally uniform distributions) which affects the entropy of a sequence consisting of symbols coming from the considered distribution. Hence, the behavior of the entropy caused by introducing punctuation marks into analysis can be considered a consequence of the influence that punctuation has on Zipf-Mandelbrot law describing word frequencies in texts.

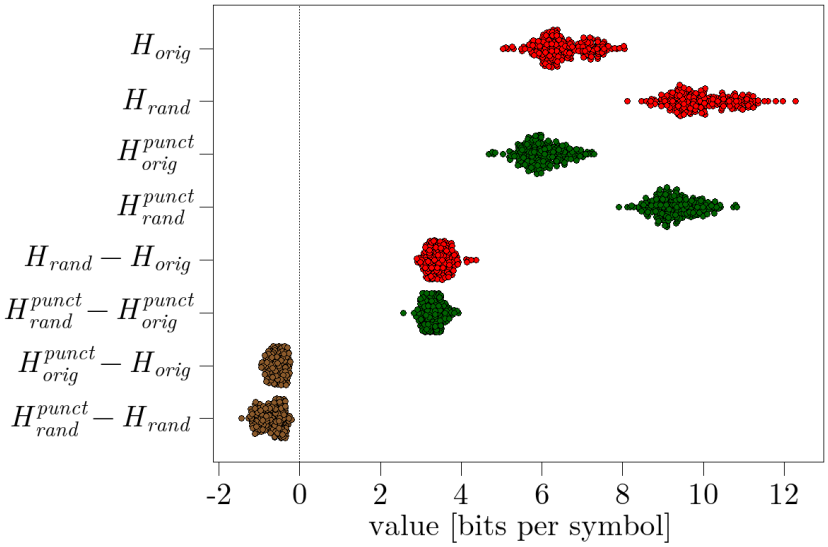


Figure 3.7. Distributions of the values of several quantities constructed from entropy rates computed for texts from dataset specified in Appendix B.1. The values of entropy rates are obtained with the use of the estimator given in Eq. 3.3. Each point on the plot corresponds to one text. Different positions along the vertical axis correspond to different quantities. The studied quantities are: entropy rate of a text treated as a sequence of words (H_{orig}), entropy rate of a text treated as a sequence of words and punctuation marks (H_{orig}^{punct}), entropy rate of a randomly shuffled text treated as a sequence of words (H_{rand}) , entropy rate of a randomly shuffled text, treated as a sequence of words and punctuation marks (H_{rand}^{punct}). Four additional quantities are constructed from the listed ones and presented in the plot: $H_{rand} - H_{orig}$, $H_{rand}^{punct} - H_{orig}^{punct}$, $H_{orig}^{punct} - H_{orig}$, $H_{rand}^{punct} - H_{rand}$.

3.6 Time series constructed from sentence lengths

3.6.1 Long-range correlations

Time series analysis is a tool well-suited to the study of natural language, as in certain situations language can be treated as a signal, often having the form of time series. There are multiple ways of representing a language sample as a signal in time domain; different approaches allow to focus on different properties. Spoken language takes the form of auditory signal, which can therefore be considered a basic, "raw" representation of language. Extracting individual sounds (phones), words, sentences etc. allows to construct higher-level representations. This applies also to written language (with the distinction that on the most basic level information is carried by appropriate symbols instead of sounds).

Studying linguistic data having the form of a time series might give an opportunity to reveal patterns of organization which can be universal for language or specific to particular language samples (samples of language typical for particular situations, for instance). Investigating the behavior of the quantities like grammatical distances between words, word recurrence times, or word lengths (as a function of their positions in text) allows to identify certain statistical regularities which are useful in attempts of characterizing the processes governing language usage [280–287].

An interesting example of a signal constructed from linguistic data is a time series representing the lengths of sentences in a text, measured by the number of words. It is a sequence of numbers in which the k-th number is the number of words in the k-th sentence (for practical purposes, sentence can be understood as a sequence of words between punctuation marks belonging to the following group: period, question mark, exclamation mark, ellipsis; usually the text needs to be appropriately pre-processed - to remove periods denoting abbreviations, for instance). Such a time series allows to investigate organization of language on a level higher than the one corresponding to individual words. Sentences are structures in which the complexity of syntax is manifested and in which words fully acquire their meanings. The content of a sentence is usually linked to the content of neighboring sentences, which constitute the context. But it turns out that the correlations typically have range larger than a few closest sentences; this effect can be captured by analyzing time series representing sentence lengths.

Figure 3.8a shows spectral densities of time series representing sentence lengths, for 239 books in 7 languages. The books are listed in Appendix B.2. The series seem to behave as $1/f^{\beta}$ signals, with β depending on the text. The histogram of the values of β, obtained by fitting lines to log-log plots of spectra $S(f)$, is shown in Figure 3.8b; typical values of β lie between 0.2 and 0.8. The presence of long-range correlations is confirmed by observing power-law behavior of fluctuation functions (Fig. 3.8c), yielding Hurst exponents H greater than 0.5 (Fig. 3.8d). The correspondence between Hurst exponents H and the exponents of spectral densities β is presented in Fig. 3.8e; it can be seen that the data conforms to an approximate relationship $\beta = 2H - 1$. That relationship, mentioned before for fractional Gaussian noises (Eq. 3.16), can be considered a more general result, holding approximately for a wider class of signals [261].

Assessing the compliance of fluctuation functions $F_2(s)$ with power-law behavior used to determine Hurst exponents can be done by inspecting the linearity of relevant log-log plots; to present all of them in a single figure one can use a linear transformation which makes each of the plots fit in the square $[0, 1] \times [0, 1]$ and in which the linearity of original plot is transformed into the linearity with slope 1 and intercept 0. The transformation having those properties is defined as follows. Let $y(x)$ be the relationship between finite sets of values x and y whose linearity is

investigated. Let (x_{min},x_{max},y_{min},y_{max}) be the minimum and the maximum values of x and y, respectively. Let $y = a + bx$ be the equation describing the assumed linear relationship. The boundaries of the rectangle enclosing the $y(x)$ plot, denoted by $x_{plot.min}$, $x_{plot.max}$, $y_{plot.min}$, $y_{plot.max}$, are defined as follows:

$$\begin{aligned} x_{plot.min} &= \min\left\{x_{min},\, \frac{y_{min} - a}{b}\right\}, \\ x_{plot.max} &= \max\left\{x_{max},\, \frac{y_{max} - a}{b}\right\}, \\ y_{plot.min} &= \min\left\{y_{min},\, a + b\, x_{min}\right\}, \\ y_{plot.max} &= \max\left\{y_{max},\, a + b\, x_{max}\right\}. \end{aligned} \tag{3.66}$$

The transformation from (x, y) to the normalized coordinates $(\widetilde{x}, \widetilde{y})$, which fit in the unit square and in which $y = a + bx$ is transformed into $\widetilde{y} = \widetilde{x}$, is given by:

$$\begin{aligned} \widetilde{x} &= \frac{x - x_{plot.min}}{x_{plot.max} - x_{plot.min}}, \\ \widetilde{y} &= \frac{y - y_{plot.min}}{y_{plot.max} - y_{plot.min}}. \end{aligned} \tag{3.67}$$

When multiple data sets (x, y), transformed with the given equations, are all presented in one $\widetilde{y}$ vs. $\widetilde{x}$ plot, and all the data points lie close to the line $\widetilde{y} = \widetilde{x}$, then all the original sets (x, y) can be considered to approximately conform to linear relationships (with possibly different slopes and intercepts). So a collective plot of $\widetilde{y}$ vs. $\widetilde{x}$ for multiple linear fits can serve as a tool for a qualitative assessment of the linear relationships' detection validity. The presented idea is applied in Fig. 3.8c; the log-log plots of the fluctuation functions $F_2(s)$ computed for the studied texts are transformed to normalized coordinates $(\widetilde{x}, \widetilde{y})$ by setting $x = \log s$ and $y = \log\left(F_2(s)\right)$ in the procedure given above.

3.6.2 Sentence lengths' multiscaling

Fluctuation scaling having the form of a power law with Hurst exponent $H > 1/2$ indicates that sentence lengths are arranged into a specific scale-free structure. However, Hurst exponent provides information which is in some sense averaged over the whole text. Complex patterns of organization in some texts can be investigated in more detail with the use of multifractal formalism. Multifractality of sentence lengths in literary texts has been studied in [285]; the analysis has revealed that while fractality is a rather general property, the degree of multifractality is more rare and specific to individual texts. Among the studied books, the ones with the richest multifractal structure are quite often the ones that use the narrative technique known as the stream of consciousness. On one hand, this technique can be considered as "natural" in certain sense, as it attempts to mimic the natural flow of thoughts and feelings passing through a character's mind, which often results in the presence of incomplete thoughts, sensory impressions, unusual grammar, and in general, certain degree of disorganization. On the other hand, it can be considered "unnatural", as it is clearly different from how the majority of written language looks like. The distribution of sentence lengths can be highly inhomogeneous, with intermittent bursts of long sentences clustered together. This effect can be captured by multifractal analysis.

Multifractality of a time series, which is usually quantified in terms of the width $\Delta\alpha$ of series' singularity spectrum $f(\alpha)$, is typically associated with two factors: heavy-tailed distributions and nonlinear correlations. While the presence of

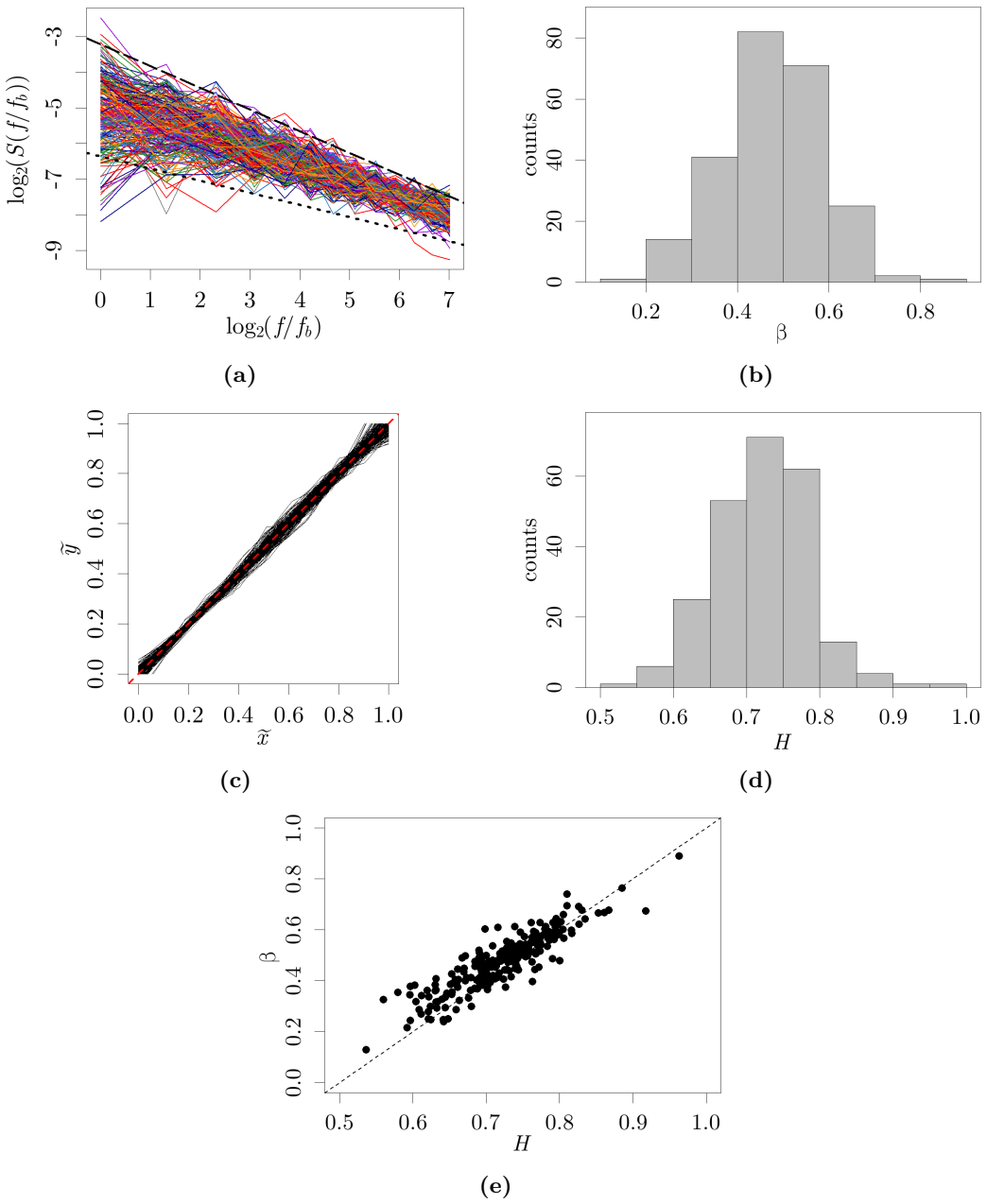


Figure 3.8. The properties of time series representing sentence lengths, quantified by spectral densities and Hurst exponents, for the books listed in Appendix B.2. (a) Spectral densities $S(f/f_b)$ for each of the studied texts, plotted in the range of small f in the form of a log-log plot (f_b denotes the fundamental frequency of the DFT). Each solid line corresponds to one text. Determining $S(f/f_b)$ for a given series involves splitting the series into 3 segments of equal length, computing the periodograms within the segments, averaging the results, smoothing, and restricting the range of f/f_b to the one presented in the plot. Series are normalized to have the same power in the considered range of frequencies. It can be seen that for small f/f_b (spanning more than two orders of magnitude), the signals exhibit a $1/f^\beta$ behavior. The dotted and the dashed line represent, respectively, the slopes determined by the 10th and the 90th percentile of the distribution of estimated β (taken with minus sign); the values of the percentiles are 0.34 and 0.61. (b) The histogram of the spectral density exponents β, obtained by fitting linear relationships to log-log plots of $S(f/f_b)$. (c) A plot demonstrating power-law behavior of fluctuation functions $F_2(s)$, computed with DFA. The plot presents $\widetilde{y}(\widetilde{x})$, where $\widetilde{x}$ and $\widetilde{y}$ are normalized coordinates, obtained by setting $x = \log s$ and $y = \log\left(F_2(s)\right)$ in Eqs. 3.66 and 3.67. Each solid line represents one text; its deviation from the $\widetilde{y} = \widetilde{x}$ relationship (dashed line) corresponds to a deviation from a power law. (d) The histogram of the Hurst exponents H. (e) The relationship between H and β. Each point represents one text; the dashed line is given by the equation $\beta = 2H - 1$.

correlations of specific type is essential for the emergence of a multifractal structure, a heavy-tailed distribution of the series' values might increase the width of the spectrum both in case when the series is truly multifractal and in case when the nonzero width of the spectrum is an artifact being a result of the finite length of the series. The latter case is related to the fact that a spectrum of a finite-length, uncorrelated random series can spuriously indicate multifractality, although such a series does not have any specific organization and in the limit of infinite length is either monofractal or bifractal (having two distinct values of singularity exponents) - depending on the distribution of series' values [273]. To clarify whether a spectrum of nonzero width is indeed related to series' multiscaling, it is profitable to confirm that the values of the series are correlated with each other in some way, using tools like autocovariance function [273]. To demonstrate that the correlations responsible for multifractality are of nonlinear character, one can investigate the singularity spectrum of a specifically constructed surrogate series - a series which is randomized in a way that removes all correlations except the linear ones. Constructing such a surrogate relies on phase randomization of the Fourier transform: for a series $x_n = x_0, x_1, x_2, ...x_{N-1}$ the (discrete) Fourier transform $\widehat{x}_k = \widehat{x}_0, \widehat{x}_1, \widehat{x}_2, ...\widehat{x}_{N-1}$ is computed; then the phases $\arg(\widehat{x}_k)$ of the transform's coefficients are randomized, by multiplying each of them by an uniformly distributed random number from the interval $[0; 2\pi]$. Then the inverse Fourier transform gives the desired surrogate. Such a surrogate has the same spectral density as the original series (as spectral density depends only on the modulus of the Fourier transform), and, consequently, the same linear correlations. Correlations of other types are destroyed. It is worth noting that this procedure in general alters the distribution of the series' values. The expected result of the multifractal analysis of such a surrogate series is a singularity spectrum practically reduced to one point, corresponding to singularity exponent determined by linear correlations.

Figure 3.9 shows the run charts $x(t)$, the autocovariance functions $R_X(\tau)$, the fluctuation functions $F_q(s)$ and the singularity spectra $f(\alpha)$, for sentence lengths in three books: *As I Lay Dying*, *Finnegans Wake*, and *Quo Vadis* (each in its original language - English or Polish). Power-law decay of the books' autocovariance functions confirms the presence of long-range correlations, also detected by the analysis of spectral density and of the Hurst exponents (Fig. 3.8). The first two of the books are examples of works utilizing stream of consciousness writing style, and *Finnegans Wake* is additionally known for its highly experimental, unusual language with uncommon grammar and vocabulary. The last of the books is more typical in terms of narrative techniques and does not rely on experimental linguistic constructions. The books exhibit different kinds of fluctuation scaling; while *Quo Vadis* is clearly a monofractal (as evidenced by singularity spectrum $f(\alpha)$ collapsed to a narrow range of α), *As I Lay Dying* and *Finnegans Wake* have a multifractal structure. *As I Lay Dying* and *Finnegans Wake* are in a sense extreme - their singularity spectra are wider than the spectra of typical examples of literary texts in prose (as mentioned before, this is to some degree characteristic of some forms of experimental writing). Sentence lengths in a literary text in prose are often either monofractal or have only some trace of multifractality, manifested by spectra of moderate width.

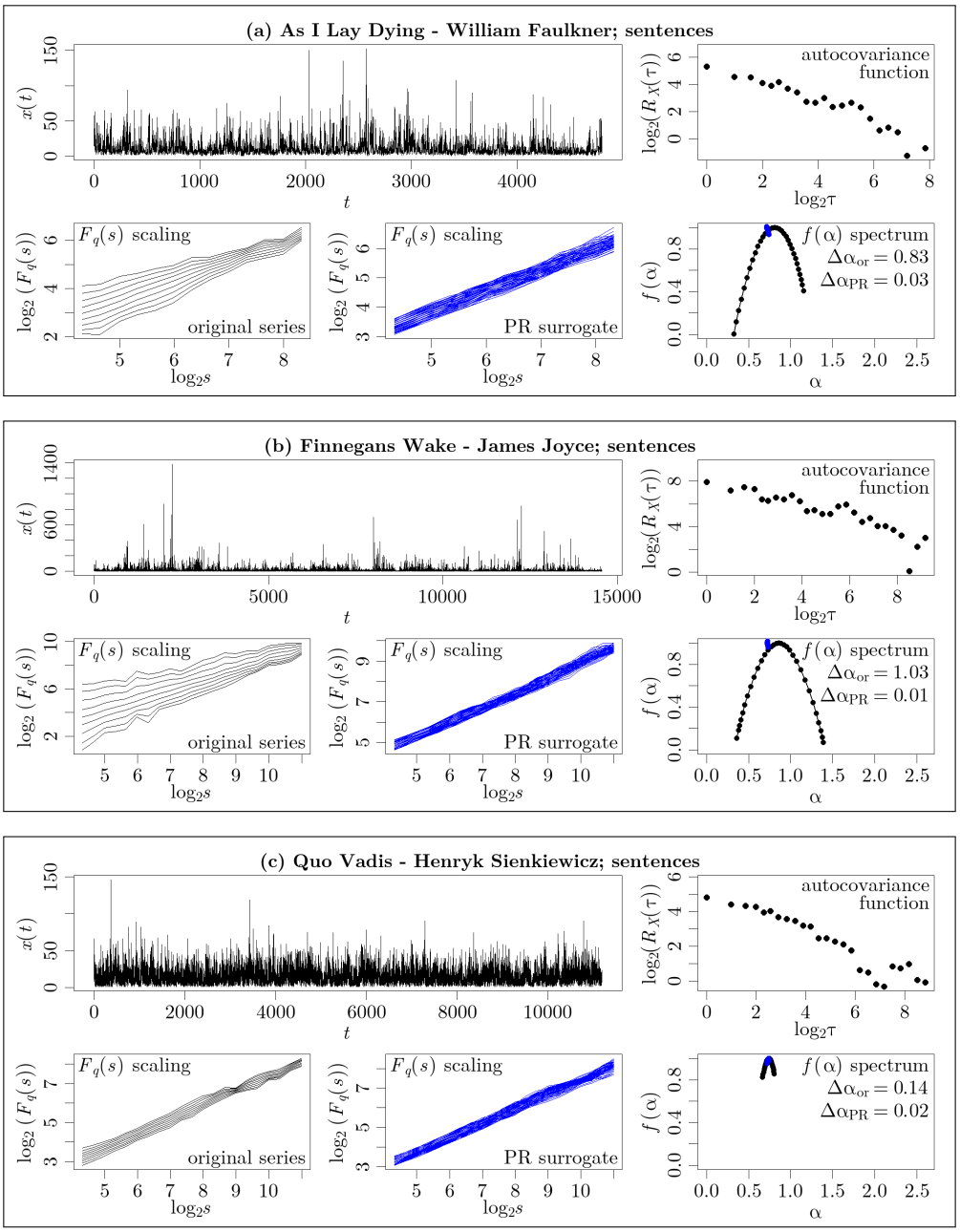


Figure 3.9. Results of multifractal analysis of the time series representing sentence lengths, for three books - *As I Lay Dying* (a), *Finnegans Wake* (b), and *Quo Vadis* (c). The analysis is performed with the use of the MFDFA method. Each box corresponds to one book and consists of two rows; the upper row presents the run chart $x(t)$ of the series and the log-log plot of the autocovariance function $R_X(\tau)$. In the lower row, there are three plots: the first two are the log-log plots of fluctuation functions $F_q(s)$ with integer q, for the original series and for the phase-randomized ("PR") surrogate series (five independent realizations). The third plot shows the singularity spectra $f(\alpha)$ of the original series (black), and for the PR surrogate series (blue). The overall range of q is $[-4; 4]$ and s is in the range in which the log-log plots of $F_q(s)$ are approximately linear, and which is a subinterval of $[20; N/5]$, with N being the length of the time series. The width of singularity spectra $\Delta\alpha_{or}$ and $\Delta\alpha_{PR}$ (for the original series, and for the PR surrogate series, respectively) are given in the upper right corner. Estimating $f(\alpha)$ for surrogate series involves averaging the results obtained in five independent realizations; for each realization the fluctuation functions $F_q(s)$ and the generalized Hurst exponents $h(q)$ are computed, and then the value of $h(q)$ averaged over the realizations is used in computing $f(\alpha)$. As expected, the resulting spectra of PR surrogate series are nearly collapsed to single points, indicating that removing nonlinear correlations destroys patterns of organization within the series that are responsible for multifractality.

3.7 Punctuation waiting times

3.7.1 Correlations, Hurst exponents, and multiscaling

Dividing a text into sentences seems quite natural - a sentence typically constitutes a complete, closed structure, capable of expressing a concrete thought. Such a partition is also meaningful from a quantitative point of view. A sentence can be treated as a sequence of words between two appropriate punctuation marks. So the length of a sentence can be interpreted as the "waiting time" for the next such mark, right after the previous one is encountered; here, "time" is measured by the number of words. If, for example, instead of punctuation marks used to end sentences, one considers some selected words as the delimiters of the sequences of other words, the waiting times are no longer multifractal [285]. This result in a sense confirms the significance of the multifractal analysis of sentence lengths, as it indicates that their multifractality is not a spurious effect.

Another way of partitioning a text into word sequences and representing it as a signal is based on considering all punctuation marks, instead of only the ones used to end a sentence. A time series can be formed of the punctuation waiting times, that is, the lengths of the word sequences between consecutive punctuation marks. Although it may seem somewhat artificial, from a certain point of view a time series of punctuation waiting times can constitute a representation encoding useful information. The historical origins of the use of punctuation in written language are related to the attempts to split texts into pieces in order to make reading in public more manageable [288]; punctuation was less specialized and less standardized than today. The classification of punctuation marks and the rules of their usage have been established in modern times. Therefore, an approach in which all punctuation marks are treated as symbols indicating the presence of some kind of a pause seems justified. The "pauses" do not have to be related to reading out loud - they might be necessary to keep the logical consistency of the text or to avoid ambiguity, for instance. So it can be postulated that punctuation marks act as boundaries for word sequences which are separated from others logically, grammatically, or in the way that facilitates comprehension and reading.

As is the case with sentence lengths, punctuation waiting times in literary texts exhibit long-range correlations and behave as $1/f^{\beta}$ signal. Figure 3.10 shows the spectral densities $S(f/f_b)$, the histogram of the spectra's exponents β, the scaling of the fluctuation functions $F_2(s)$ (in normalized coordinates defined by Eq. 3.66 and Eq. 3.67), and the relationship between the values of β and Hurst exponents H, for the books listed in Appendix B.2. The punctuation marks taken into consideration are: period, question mark, exclamation mark, ellipsis, comma, dash, semicolon, colon, left parenthesis and right parenthesis. Symbols not present on that list and not being words (quotation marks, for instance) are removed from the texts. The time series are formed of non-zero waiting times (a waiting time equal to zero occurs when two punctuation marks are placed next to each other, for example when a question mark is followed by an exclamation mark; since such cases correspond to a single "pause" in a text, they can be disregarded in the construction of the series). It can be observed that punctuation waiting times usually have the Hurst exponent H lower than the Hurst exponent of sentence lengths in the same text (Fig. 3.11). Nevertheless, punctuation waiting times have the value of H still above 0.5, which indicates their persistence.

In terms of fluctuation scaling, punctuation waiting times typically display more uniform behavior than sentence lengths (Fig. 3.12) - their singularity spectra are usually significantly narrower than the spectra of sentence lengths. However, some degree of multifractality can sometimes be observed, especially when the range of

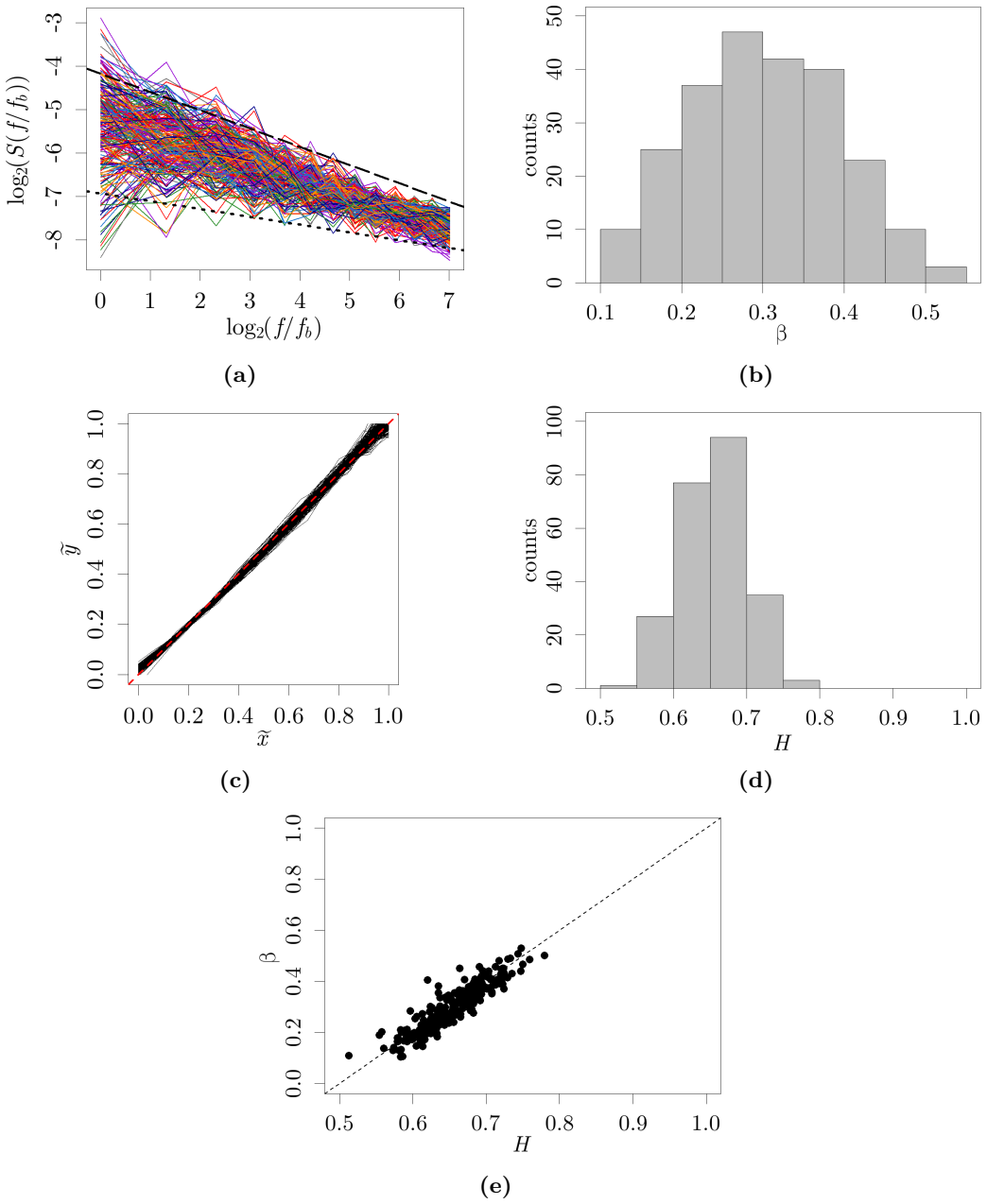


Figure 3.10. The properties of time series representing punctuation waiting times (numbers of words between consecutive punctuation marks), quantified by spectral densities and Hurst exponents, for the books specified in Appendix B.2. (a) Spectral densities $S(f/f_b)$ for each of the studied texts, plotted in the range of small f in the form of a log-log plot (f_b denotes the fundamental frequency of the DFT). Each solid line corresponds to one text. Determining $S(f/f_b)$ for a given series involves splitting the series into 3 segments of equal length, computing the periodograms within the segments, averaging the results, smoothing, and restricting the range of f/f_b to the one presented in the plot. Series are normalized to have the same power in the considered range of frequencies. It can be seen that for small f/f_b (spanning more than two orders of magnitude), the signals exhibit a $1/f^\beta$ behavior. The dotted and the dashed line represent, respectively, the slopes determined by the 10th and the 90th percentile of the distribution of estimated β (taken with minus sign); the values of the percentiles are 0.18 and 0.42. (b) The histogram of the spectral density exponents β, obtained by fitting linear relationships to log-log plots of $S(f/f_b)$. (c) A plot demonstrating power-law behavior of fluctuation functions $F_2(s)$, computed with DFA. The plot presents $\widetilde{y}(\widetilde{x})$, where $\widetilde{x}$ and $\widetilde{y}$ are normalized coordinates, obtained by setting $x = \log s$ and $y = \log(F_2(s))$ in Eqs. 3.66 and 3.67. Each solid line represents one text; its deviation from the $\widetilde{y} = \widetilde{x}$ relationship (dashed line) corresponds to a deviation from a power law. (d) The histogram of the Hurst exponents H. (e) The relationship between H and β. Each point represents one text; the dashed line is given by the equation $\beta = 2H - 1$.

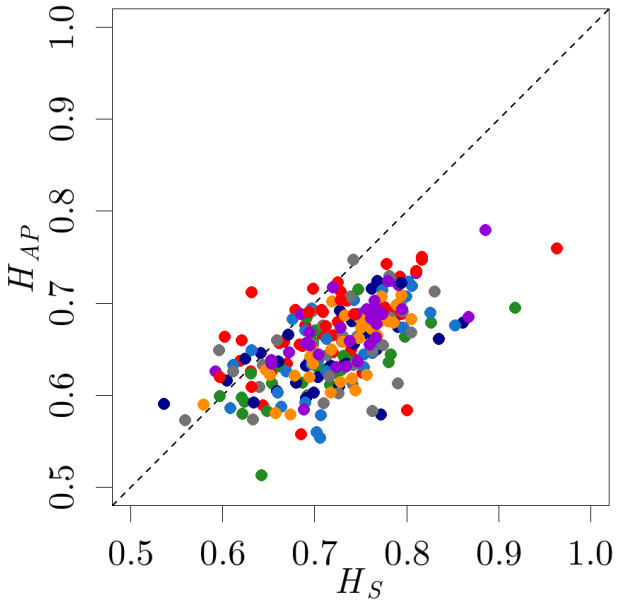


Figure 3.11. The scatterplot of the Hurst exponent of punctuation waiting times H_{AP} versus the Hurst exponent of sentence lengths H_S, for the books specified in Appendix B.2. Each point represents one text; colors correspond to languages: red - English, green - German, dark blue - French, light blue - Italian, gray - Spanish, orange - Polish, purple - Russian. The dashed line has the equation $H_{AP} = H_S$. The Pearson correlation coefficient between H_S and H_{AP} in the whole dataset is equal to 0.59.

the values present in the series is wide enough (as is the case for *As I Lay Dying*, for instance). The histogram in Figure 3.13 illustrates how much the singularity spectrum width decreases when the representation of a text changes from sentence lengths to punctuation waiting times. The quantity presented in the histogram is the relative change of singularity spectrum width, that is $(\Delta\alpha_{AP} - \Delta\alpha_S)/\Delta\alpha_S$, where $\Delta\alpha_S$ is the spectrum width for sentence lengths and $\Delta\alpha_{AP}$ is the spectrum width for punctuation waiting times (the subscript "AP" is derived from "all punctuation marks"). The books used in this part of the analysis are the books which are given in Appendix B.2 and which satisfy three additional conditions. Firstly, they have no less than 3000 sentences each; secondly, the log-log plots of their sentence lengths' fluctuation functions $F_q(s)$ for all $q \in [-4; 4]$ are approximately linear for s in the range $[20; N/5]$, where N is the overall number of sentences in a text, and thirdly, the width of their singularity spectra of sentence lengths is not less than 0.2. The presented conditions aim to ensure that the books are sufficiently long and have a range of $F_q(s)$ scaling wide enough to provide reasonable amount of data for the automated estimation of multifractal properties, and that their sentence lengths exhibit at least weak multifractality.

The Hurst exponents of sentence lengths H_S and the Hurst exponents of punctuation waiting times H_{AP} have values above 0.5 and are correlated (the Pearson correlation coefficient between H_S and H_{AP} is equal to 0.59), as evidenced in Fig. 3.11. This raises a question about how the properties of these two types of series are related, and whether the relationship between their Hurst exponents is a consequence of the way in which the series are constructed, or whether it can be attributed to other factors. Since sentence-ending punctuation marks constitute a subset of all punctuation marks used in written language, it seems natural that the properties of sentence lengths and of punctuation waiting times are not entirely independent. To way of approaching that issue quantitatively is to investigate the behavior of both types of series, randomized in the way that keeps the other series unchanged. A ran-

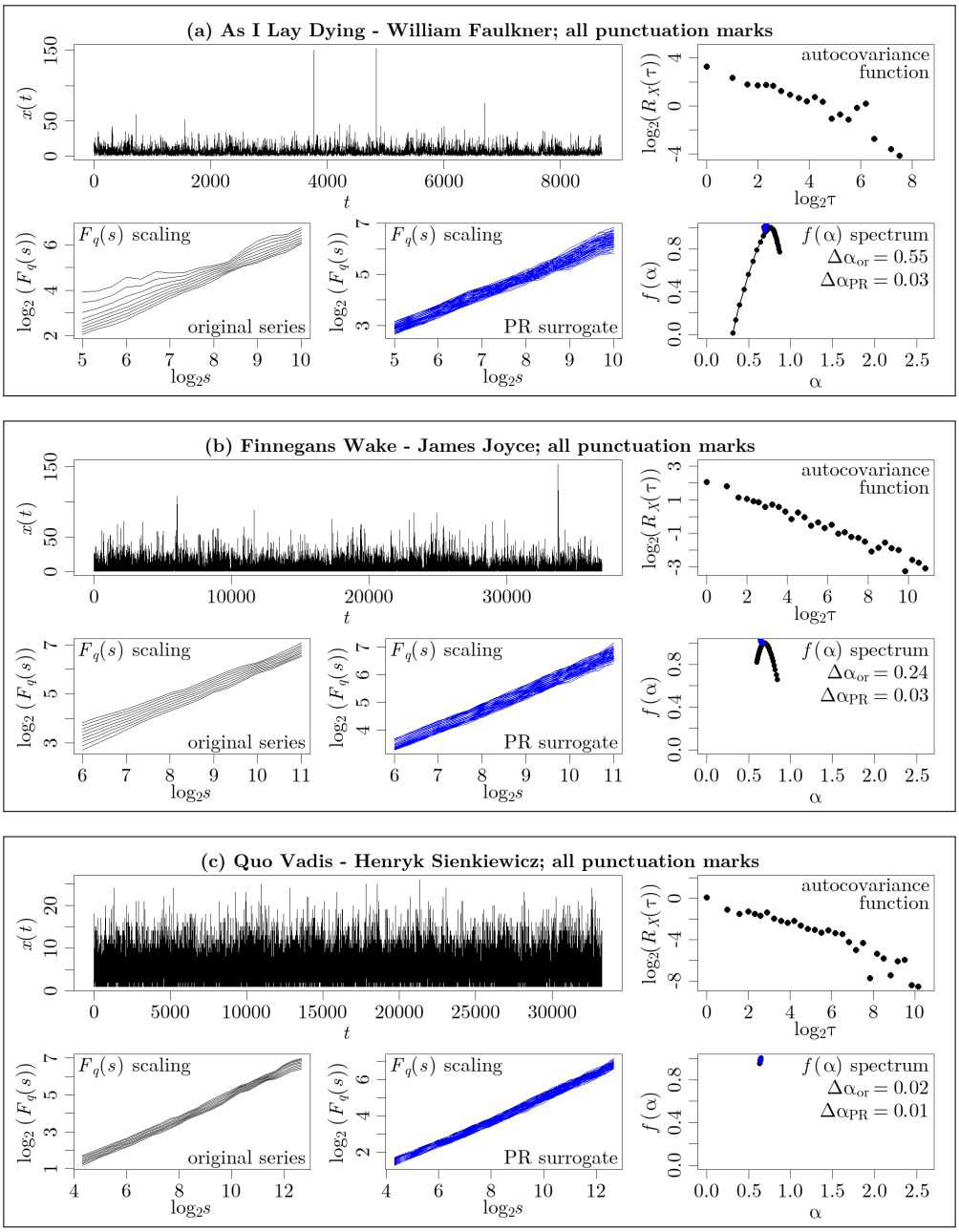


Figure 3.12. Results of multifractal analysis of the time series representing punctuation waiting times, for three books - *As I Lay Dying* (a), *Finnegans Wake* (b), and *Quo Vadis* (c). The analysis is performed with the use of the MFDFA method. Each box corresponds to one book and consists of two rows; the upper row presents the run chart $x(t)$ of the series and the log-log plot of the autocovariance function $R_X(\tau)$. In the lower row, there are three plots: the first two are the log-log plots of fluctuation functions $F_q(s)$ with integer q, for the original series and for the phase-randomized ("PR") surrogate series (five independent realizations). The third plot shows the singularity spectra $f(\alpha)$ of the original series (black), and for the PR surrogate series (blue). The overall range of q is $[-4; 4]$ and s is in the range in which the log-log plots of $F_q(s)$ are approximately linear, and which is a subinterval of $[20; N/5]$, with N being the length of the time series. The width of singularity spectra $\Delta\alpha_{or}$ and $\Delta\alpha_{PR}$ (for the original series, and for the PR surrogate series, respectively) are given in the upper right corner. Estimating $f(\alpha)$ for surrogate series involves averaging the results obtained in five independent realizations; for each realization the fluctuation functions $F_q(s)$ and the generalized Hurst exponents $h(q)$ are computed, and then the value of $h(q)$ averaged over the realizations is used in computing $f(\alpha)$. As expected, the resulting spectra of PR surrogate series are nearly collapsed to single points, indicating that removing nonlinear correlations destroys patterns of organization within the series that are responsible for multifractality.

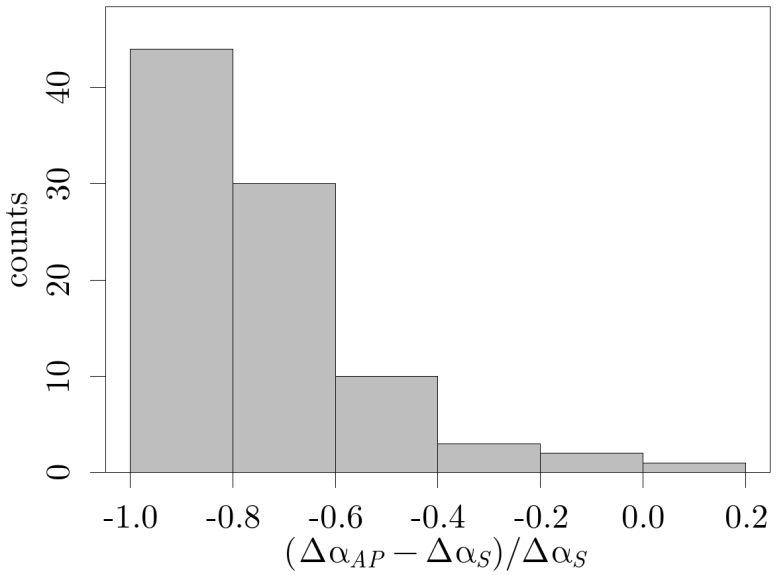


Figure 3.13. The histogram of $(\Delta\alpha_{AP} - \Delta\alpha_S)/\Delta\alpha_S$, where $\Delta\alpha_S$ and $\Delta\alpha_{AP}$ are the widths of singularity spectra for sentence lengths and punctuation waiting times, respectively. The studied texts constitute a subset of the books given in Appendix B.2; these are the books of at least 3000 sentences, having $\Delta\alpha_S$ greater or equal to 0.2. The considered quantity $(\Delta\alpha_{AP} - \Delta\alpha_S)/\Delta\alpha_S$ represents the relative change of singularity spectrum width observed when changing the representation of a text from sentence lengths to punctuation waiting times.

domization of text's sentence lengths which does not alter the series of punctuation waiting times can be done by permuting randomly the positions of all the punctuation marks present in the text. The set of punctuation marks' positions remains unchanged, but which mark occupies which position is decided by chance. Therefore, sentences lose their original structure, as sentence-ending marks are located at random positions allowed by the overall arrangement of punctuation marks. This method of randomization models the situation in which punctuation is placed as in the original text, but sentence lengths, apart from satisfying the condition that they are delimited by symbols belonging to an appropriate subset of punctuation, are completely random.

To perform randomization the other way round, a procedure described below can be used. The sentences in a text can be grouped into "buckets", each bucket corresponding to particular range of lengths. For example, in such a partition, one bucket might consist of all the sentences in the text which have lengths equal to 1 or 2, another bucket can contain all the sentences of length 3, yet another bucket can be composed of all the sentences with lengths between 12 and 18, and so on. Each sentence in the text needs to be assigned to a (single) bucket; the range of lengths covered by a bucket should be as narrow as possible, provided that each of the buckets contains at least a few (3-5) sentences. Randomization consists of assigning sentences to buckets, and permuting the positions of sentences inside each bucket. This means that sentences randomly swap positions with other sentences belonging to the same bucket (having the same or similar length). Consequently, the series representing sentence lengths is approximately the same as the original one (the exact level of agreement depends on the length ranges used to define buckets), but the contents of sentences (including punctuation) become randomly scattered across the text. However, it should be noted that the resulting the series of punctuation waiting times can be affected by statistical relationships binding the structure of punctuation inside a sentence with sentence length. An example of such a relationship is the one expressed by Menzerath-Altmann law; for sentences, the law can summarized by

the following statement: the longer a sentence, the smaller the average size of the constituents it is composed of. Under the assumption that sentences can be divided into constituents separated by punctuation marks, Menzerath-Altmann law results in a tendency of punctuation waiting times to be short in regions where sentences are long, and to be long in the parts of texts in which sentences are short.

Figure 3.14 presents how the Hurst exponents of sentence lengths and of punctuation waiting times change when the randomization procedures given above are performed on the texts presented in Appendix B.2. Typically, the Hurst exponents of the appropriately randomized series are substantially lower than the exponents of the corresponding original series, but their value is usually still above 0.5. This means that the persistence of sentence lengths and the persistence of punctuation waiting times can be partially explained by each other - when for a given text one type of series is randomized and the other is kept unchanged, the randomized one exhibits some degree of persistence due to the persistence of the other one. Also, performing randomization of any of the presented types does not remove the correlations between the Hurst exponents of sentence lengths and of punctuation waiting times - that is, texts with high Hurst exponents describing sentence lengths H_S also tend to have high Hurst exponents pertaining to punctuation waiting times H_{AP}. Conversely, low H_S typically co-occurs with low H_{AP}. Pearson correlation coefficients between H_S and H_{AP}, describing that effect, is equal to 0.87 for the randomization of sentence lengths (preserving H_{AP}) and equal to 0.62 for the randomization of punctuation waiting times (preserving H_S). So even when one of the two series is random, it is correlated with the other one, provided that the conditions making sentence lengths and punctuation waiting times consistent with each other are satisfied. Therefore, the correlation between H_S and H_{AP} can be seen as an effect caused by the fact that sentence-ending punctuation marks constitute a subset of all punctuation marks.

Comparing sentence lengths and punctuation waiting times in terms of long-range correlations, fractality, and multifractality provides an insight into the significance of punctuation's role in language. In a sense, the properties of punctuation waiting times seem more universal than the corresponding properties of sentence lengths - for example, the variability of Hurst exponents among different texts is lower when all punctuation marks are considered instead of only the marks which divide texts into sentences. Also, in terms of fluctuation scaling, punctuation marks treated collectively determine a structure more homogeneous than the one constituted by sentences; this fact is reflected by a stronger inclination of punctuation waiting times towards monofractality. The presented results can be viewed as being in agreement with a common intuition that the division of a text into sentences involves some degree of arbitrariness. A message or thought which is expressed by a long sentence, composed of several components, usually can also be expressed by a few short sentences, each of which corresponds to some component of the long sentence. Therefore, the number and the lengths of the used sentences depend on author's choice. However, when the first of the two options is chosen (one long sentence), the components of the sentence usually have to be separated by punctuation marks (comma, for instance). Hence, a certain number of such marks has to appear inside the sentence and punctuation waiting times are not arranged entirely freely.

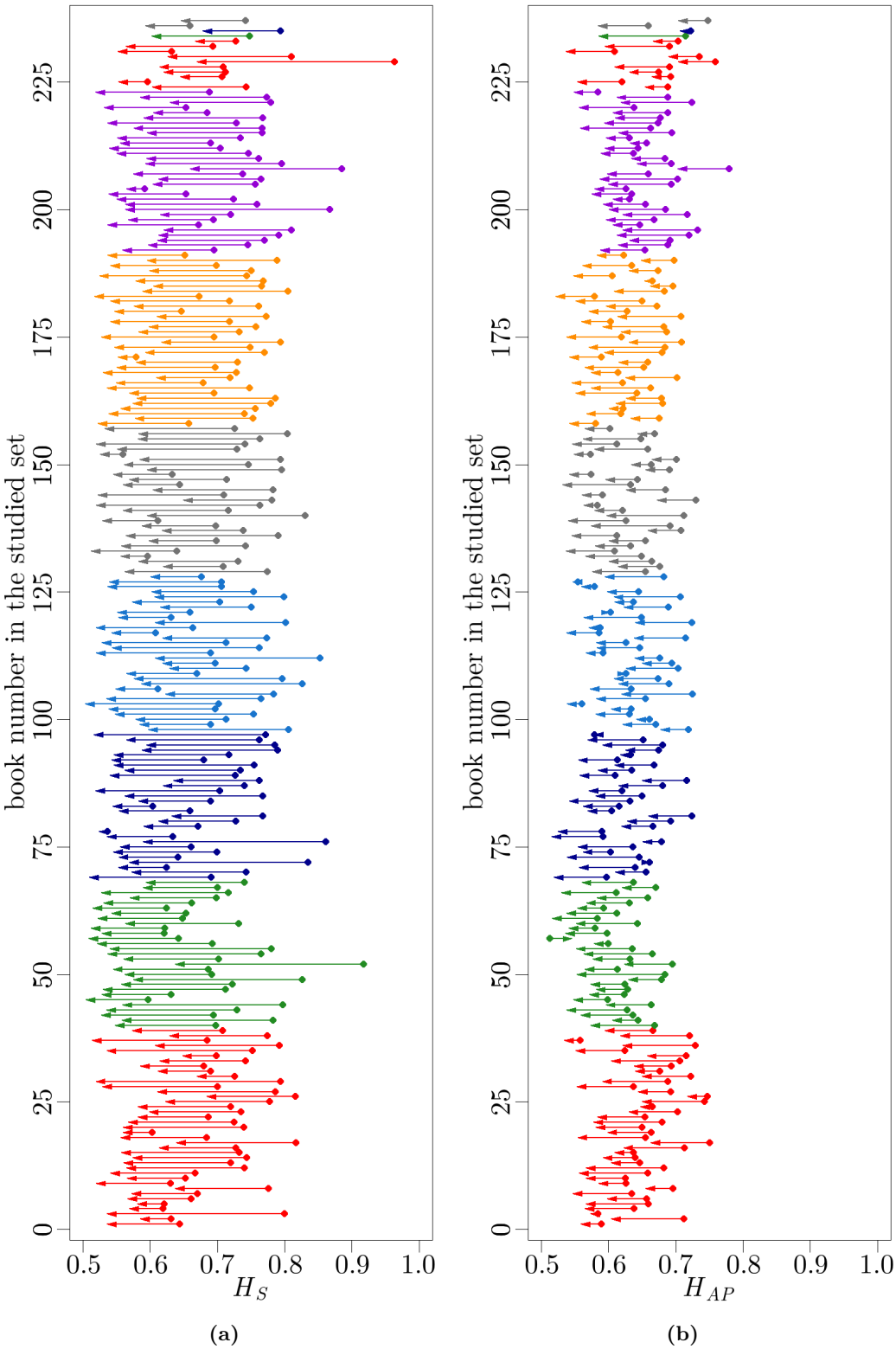


Figure 3.14. The Hurst exponents of sentence lengths H_S and of punctuation waiting times H_{AP}, computed for the original and the randomized time series, for the books specified in Appendix B.2. Fig. (a) pertains to the randomization of sentence lengths (preserving H_{AP}); Fig. (b) pertains to the randomization of punctuation waiting times (preserving H_S). Arrows mark the change of the Hurst exponent induced by randomization - dots denote the Hurst exponents of the original series, arrow heads denote the Hurst exponents of the randomized series (computed as an average over 5 independent randomizations). Consecutive dot-arrow pairs represent consecutive books from the dataset specified in Appendix B.2; colors correspond to languages: red - English, green - German, dark blue - French, light blue - Italian, gray - Spanish, orange - Polish, purple - Russian.

3.7.2 Distributions of punctuation waiting times

The conclusion that time series constructed from punctuation waiting times behave in a more "consistent" way compared to series representing sentence lengths can also be supported by analyzing the probability distributions of the values of the two types of series. It turns out that the distribution of punctuation waiting times in texts can be characterized by two numbers, being the parameters of the so-called discrete Weibull distribution. The distribution can be introduced with the help of the following reasoning. When a text is considered a sequence of words and punctuation marks occurring between some of them, it can be assumed that distributing punctuation marks across text is governed by some process deciding for each consecutive word whether a punctuation mark is to be placed after that word or not. Assuming that the process is random and it puts a punctuation mark after a word with some constant probability p, each such decision is a Bernoulli trial with p being the probability of success. In such a case, the punctuation waiting time (the number of words between consecutive punctuation marks) is the number k of trials required to obtain the first success, after the last one observed ($k = 1, 2, 3, ...$). The number of trials until the first success in Bernoulli process follows the geometric distribution. Observing a waiting time longer than k is equivalent to not observing a success in the first k trials; therefore, one can write:

$$1 - F(k) = (1 - p)^k \,, \tag{3.68}$$

where F is the cumulative distribution function ($F(k)$ is defined as the probability that a waiting time is less than or equal to k). The above relationship pertains to situation when punctuation marks are placed independently of each other, with constant probability. However, it is reasonable to anticipate that the probability of placing a punctuation mark after a particular word depends on the sequence of words and punctuation marks preceding the considered word. Hence, a distribution more general than the geometric distribution is required. One way of generalizing the geometric distribution is introducing an exponent, $\beta > 0$, into its survival function:

$$1 - F(k) = (1 - p)^{k^\beta} \,. \tag{3.69}$$

A distribution specified in such a way is a discrete analogue of the Weibull distribution, therefore called the discrete Weibull distribution [289]. Due to its flexibility, Weibull distribution, especially in its continuous version, is widely applied in various fields of science and engineering, for instance in survival analysis, in medicine and health sciences, or in modelling natural phenomena like wind speed or rainfall intensity [290]. Interestingly, it has also been employed in studies on natural language, namely in investigating the distribution of word recurrence times in textual data [284].

The parameter β of the discrete Weibull distribution determines the deviation from the geometric distribution, which is recovered for $\beta = 1$. It describes how the probability of obtaining a success depends on how many trials have been performed since the last success observed. This dependence can be characterized by the so-called *hazard function* $h(k)$. Hazard function can be defined as the conditional probability that a success occurs on the k-th trial, given that it has not occurred in the preceding $k-1$ trials. With $P(k)$ denoting the probability mass function, the hazard function of the discrete Weibull distribution is given by:

$$h(k) = \frac{P(k)}{1 - F(k-1)} = \; 1 - (1 - p)^{k^\beta - (k-1)^\beta} \,. \tag{3.70}$$

For $\beta < 1$, the hazard function is a decreasing function - the probability of observing a success becomes smaller as the waiting time gets longer. For $\beta > 1$, it is increasing

with time. For $\beta = 1$, the hazard function is a constant - one obtains the geometric distribution, which is therefore said to be *memoryless*. The parameter p of the discrete Weibull distribution also can be intuitively interpreted - it is the probability of observing a success in the first trial. The plots presenting the discrete Weibull distribution for selected values of p and β are shown in Figure 3.15.

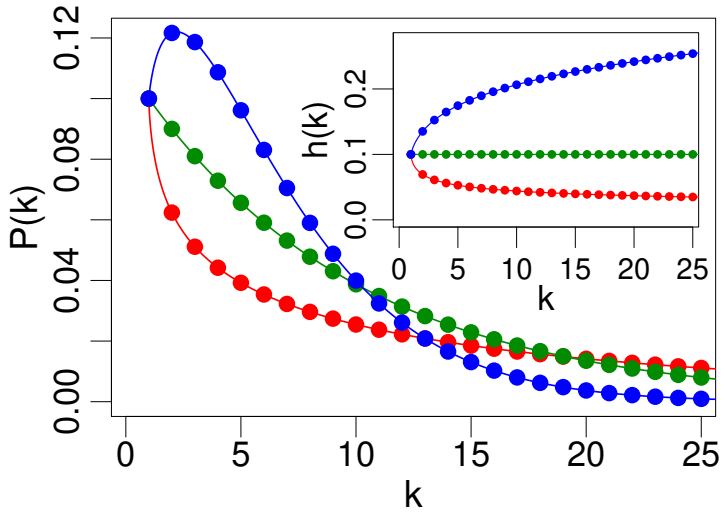


Figure 3.15. The probability mass function $P(k)$ of the discrete Weibull distribution, for $p = 0.1$ and three different values of β: $\beta = 0.75$ (red), $\beta = 1$ (green), $\beta = 1.25$ (blue). The corresponding hazard functions are presented in the inset. Since the distribution is discrete, $P(k)$ and $h(k)$ are defined only at integer k and their values are represented by dots; the connecting lines are only guides for the eye and do not indicate continuity.

An easy way of assessing how well a given data set fits to a Weibull distribution (both in continuous and discrete case) is constructing the so-called Weibull plot. It can be shown that Eq. 3.69 can be rewritten as:

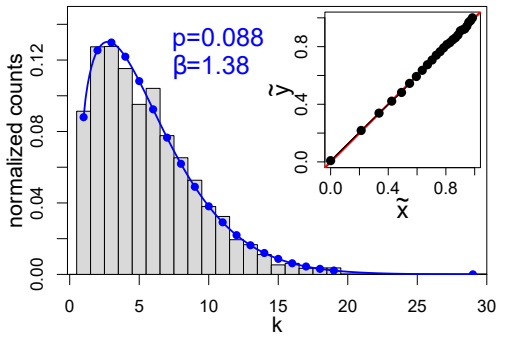
$$\log(-\log(1-F(k))) = \beta \log k + \log(-\log(1-p)) \,. \tag{3.71}$$

Therefore, if the data comes from the discrete Weibull distribution with parameters (p, β), then when plotting the empirical cumulative distribution function $F_{emp}(k)$ in coordinates (x, y), where

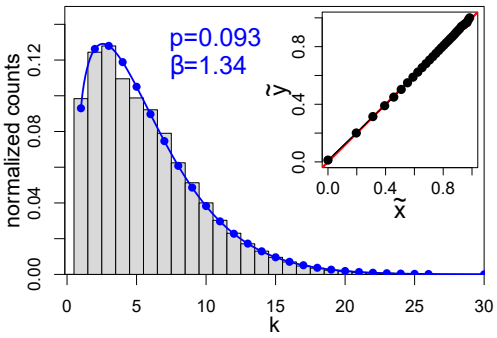
$$\begin{aligned} x &= \log k \\ y &= \log\left(-\log\left(1 - F_{emp}(k)\right)\right), \end{aligned}$$

one should observe a straight line with slope β and intercept $\log(-\log(1-p))$. To make comparison between fits to different Weibull distributions easier, one can use the transformation analogous to the one given by Eq. 3.66 and Eq. 3.67 to rescale the coordinates (x, y) to $(\widetilde{x}, \widetilde{y})$, fitting in the square $[0, 1]\times[0, 1]$. In a plot in rescaled coordinates (here referred to as a *rescaled Weibull plot*), the deviation from the Weibull distribution is observed as the deviation from the line $\widetilde{y} = \widetilde{x}$.

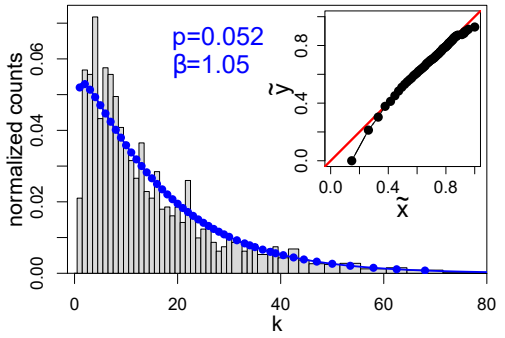
Figure 3.16 shows the empirical distributions of punctuation waiting times and of sentence lengths, for two books: *Alice's Adventures in Wonderland* by Lewis Carroll and *David Copperfield* by Charles Dickens. Discrete Weibull distribution is fitted to the data - maximum likelihood estimation (MLE) is used to find the parameters of the distribution. It can be seen that punctuation waiting times in both of the books are well described by discrete Weibull distribution, but in case of sentence lengths one of the books (*Alice's Adventures in Wonderland*) exhibits considerably worse agreement between the empirical and the proposed model distribution. It turns out that this applies also to other texts - while the distribution of punctuation waiting



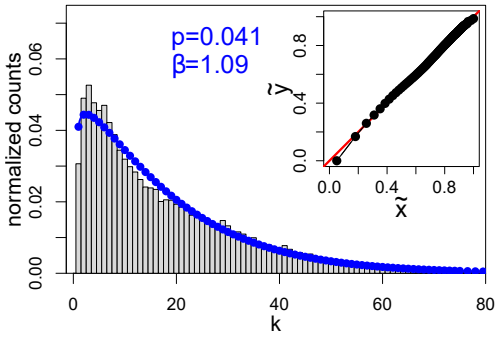
(a) *Alice's Adventures in Wonderland*, punctuation waiting times



(b) *David Copperfield*, punctuation waiting times



(c) *Alice's Adventures in Wonderland*, sentence lengths



(d) *David Copperfield*, sentence lengths

Figure 3.16. Examples of the distributions of punctuation waiting times and of sentence lengths, for two books: *Alice's Adventures in Wonderland* by Lewis Carroll and *David Copperfield* by Charles Dickens. In each figure, histogram represents the empirical distribution and blue dots represent the discrete Weibull distribution fitted using maximum likelihood estimation (the obtained parameters p and β are given above the plots). Insets show the corresponding rescaled Weibull plots, in which deviations from the line $\widetilde{y} = \widetilde{x}$ correspond to discrepancies between the fitted and the empirical distribution.

times can almost universally be modeled by discrete Weibull distribution, the distribution of sentence lengths might either be of the same type or of more "irregular" nature (meaning that it is much harder to find a distribution with relatively simple functional form that would accurately represent the data). This fact is demonstrated in Figure 3.17, which presents rescaled Weibull plots of punctuation waiting times and of sentence lengths for 223 books in 7 languages (books from Appendix B.1). The deviations from the line $\widetilde{y} = \widetilde{x}$ in the rescaled Weibull plots of sentence lengths tend to be significantly larger than the ones observed in the rescaled Weibull plots of punctuation waiting times.

From the viewpoint considering only the probability distribution characterizing punctuation, the process of writing a text can be thought of in terms of a simple mathematical model, based on the properties of the discrete Weibull distribution. The model assumes that a text is generated word by word, and a punctuation mark can be placed after each word, with some probability $h(k)$ which depends only on k, the number of words that occurred since the last placed punctuation mark. The relationship between h and k is of the form given in the Eq. 3.70. The resulting distribution of distances between punctuation marks in the text is the discrete Weibull distribution. By adjusting the parameters p and β in the function $h(k)$, one can obtain a distribution observed in real texts. The parameters are easily interpreted: p is the probability that a punctuation mark appears right after the first

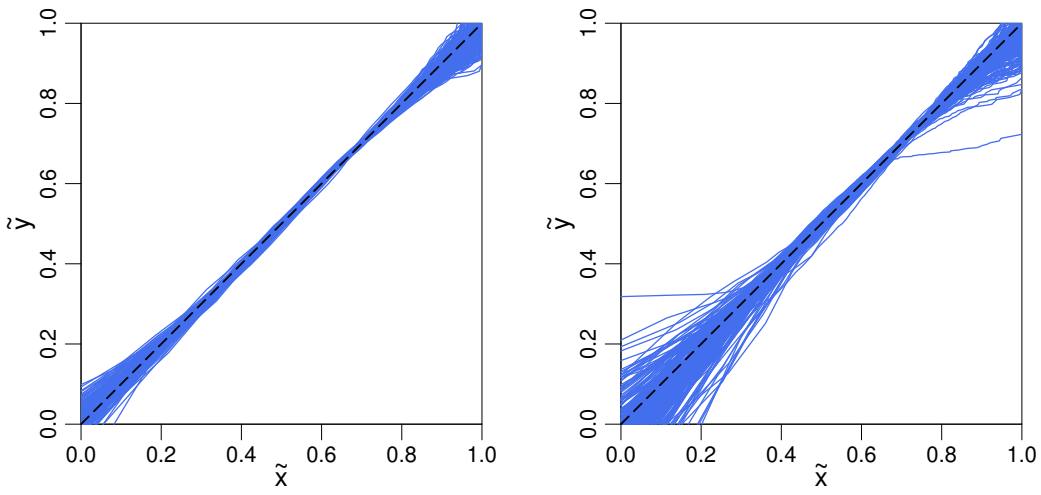


Figure 3.17. The rescaled Weibull plots of punctuation waiting times (a) and sentence lengths (b), for books listed in Appendix B.1. Each curve on a plot corresponds to one book; the dashed line $\widetilde{y} = \widetilde{x}$ represents the ideal fit to the discrete Weibull distribution.

word since the last punctuation mark; β describes how fast the probability of the punctuation mark occurrence changes with the growing number of words appearing since the last punctuation mark observed. The assumption that the probability of a punctuation mark occurrence depends only on k, which is equivalent to the statement that word sequences between punctuation marks are generated independently, is of course idealized. In real texts, it is obviously violated by the presence of long-range correlations, for instance. However, when only the probability distribution is considered, correlations between punctuation waiting times can be neglected.

Discrete Weibull distributions characterizing punctuation waiting times in all of the studied texts have the value of β between 1 and 2; this means that $h(k)$ is an increasing function. With $k \to \infty$, it converges to 1. It seems to be a reasonable result - the sequences of words without punctuation should not be infinitely long. The values of p are typically below 0.2. Interestingly, the parameters of the distributions (determining their shape) seem to be to some degree specific to particular languages. When the values of p and β related to each book are plotted on a plane (each point represents one book), one can distinguish regions occupied mainly by the texts in the same language (Figure 3.18). Average values of p and β for each language can be calculated to determine the corresponding hazard functions $h(k)$ (Figure 3.19). Using the concept of random process underlying the arrangement of punctuation in texts, these functions characterize the dynamics of the process; they provide information how "urgent" it is to place a punctuation mark in order to finish an uninterrupted word sequence, depending on the length of that sequence.

It can be noticed in Figure 3.19 that within the range of waiting times between 1 and 15 (which corresponds to more than 80% of all observed waiting times in each of the studied texts), the Slavic languages have the highest values of the averaged hazard functions among the studied ones, therefore being the most inclined towards short word sequences between consecutive punctuation marks. Regarding the punctuation distribution properties, two of the Romance languages considered, French and Italian, turn out to be quite similar to each other. They have close average values of p and β and their dispersions are overlapping. The averaged hazard function for German, having the lowest p and the highest β, starts from a low value and increases quickly. The most slowly-varying averaged hazard functions belong

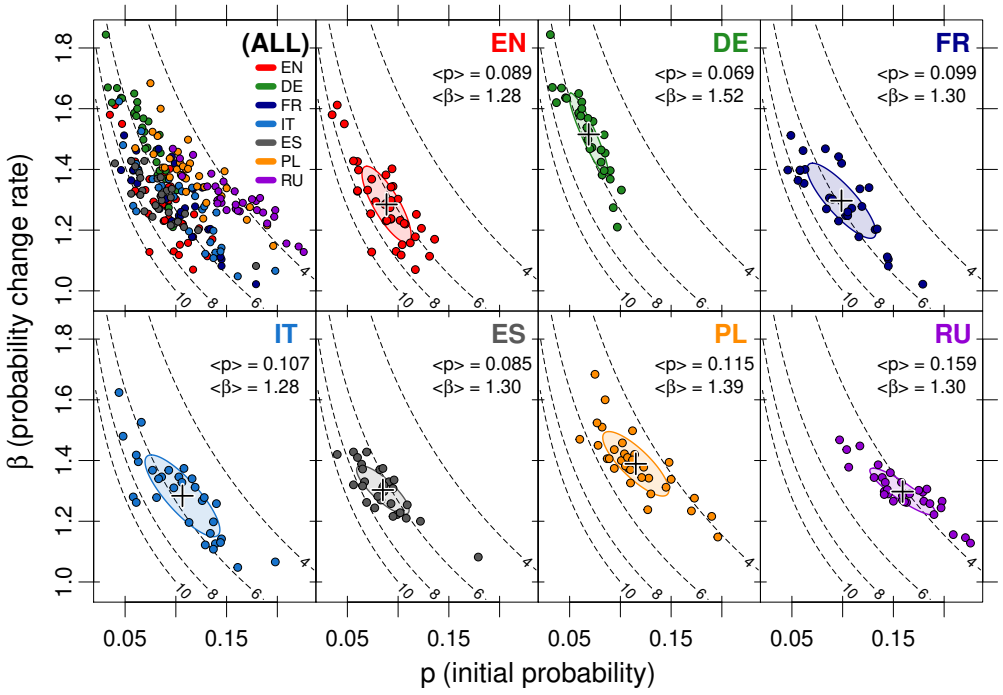


Figure 3.18. The parameters p and β of the discrete Weibull distributions fitted to punctuation waiting times, for the books listed in Appendix B.1. The chart in the upper left corner (ALL) pertains to all the studied languages collectively, the remaining ones present results for the individual languages. In each plot, a text is represented by a point (p, β). All the plots are in the same scale. The dashed lines are isolines of constant expected value of the discrete Weibull distribution - all distributions with (p, β) along one such line have the same expected value. In each plot pertaining to a single language, the quantities $\langle p \rangle$ and $\langle \beta \rangle$, the average values of p and β, are given, and the centroid of the point cloud, $(\langle p \rangle, \langle \beta \rangle)$ is marked by "+". The ellipses characterize the distributions of points - the semi-axes of each ellipse are the principal components of the point set in the given language. The major semi-axis of the ellipse gives the direction of the greatest variance and its length is the square root of that variance. The length of the minor semi-axis is the square root of the variance in the perpendicular direction. The ellipses for each language are shown collectively in Figure 3.19.

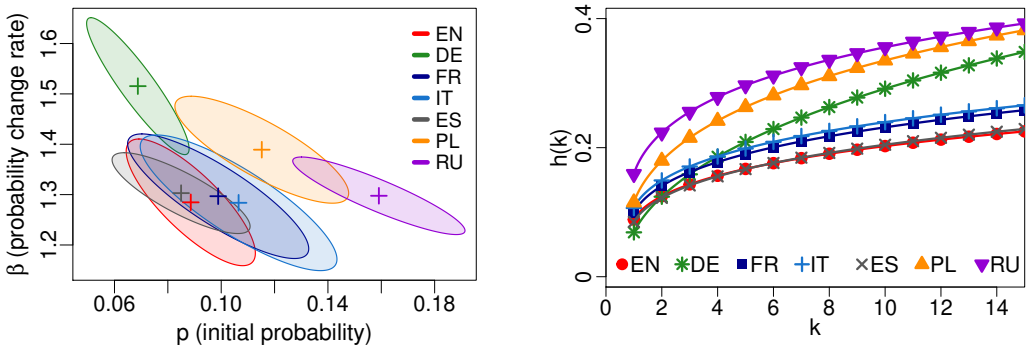


Figure 3.19. (a) The ellipses characterizing the distributions of (p, β), the parameters of the discrete Weibull distributions describing punctuation waiting times in texts, for the languages considered in the dataset specified in Appendix B.1 (these are the same ellipses as in Figure 3.18, collected in a single chart). The centroids of (p, β) for each language are marked by "+". (b) The hazard functions of discrete Weibull distributions with parameters (p, β) corresponding to the centroids of the ellipses presented in (a).

to English and Spanish, suggesting that long sequences of words between pauses indicated by punctuation marks are more natural for those languages than for the others. However, a comprehensive description of such properties would require a more detailed investigation. For example, the above-mentioned tendency of Polish and Russian to favour short intervals between punctuation marks may be caused by the lack of articles in these languages; in other languages studied here, articles are present. Although they are not stand-alone words, they are treated just as the other ones in the analysis, and therefore they lengthen the sequences of words appearing between punctuation marks.

The analysis of linguistic time series using methods discussed in this chapter gives access to information about certain fundamental properties of language. When it comes to the distribution of punctuation, for example, it allows to observe general statistical regularities (applying to various texts in various languages), as well as to identify differences between languages. It gives an insight into the character and the origin of long-range correlations in texts and allows to compare how different levels of language organization (sentences, sentence components) behave when treated as a signal. Finally, it can detect complex, multifractal structures and relate their presence to specific styles of writing. The results obtained with the use of the presented methods might be used to quantitatively characterize samples of written language (as they make it possible to express the differences between texts in terms of measurable quantities, for instance) or to hypothesize about rather general problems in the study of natural language, like the significance of partitioning a text into sentences and the dependence between such a partitioning and other ways of breaking a text into parts or components.

Chapter 4

Linguistic networks

4.1 Basic concepts in network theory

A multitude of systems in nature can be characterized by a very general statement, that they consist of a large number of constituents interacting with each other. The nature of the interaction is system-specific. However, if from the standpoint of studied properties it is sufficient to treat a system as a set of some objects and a set of pairwise relations between these objects, then often the system can be represented as a network. The advantage of network representation is the fact that it allows to describe diverse systems with the use of unified, abstract formalism. This makes it possible to look for common traits of various systems and formulate universal laws describing their structure or behavior. Although the study of networks is rooted in the concepts of the graph theory, it has evolved into a theory which can be considered a separate field of research, sometimes called *network theory* or *network science*. The wide applicability of the network theory resulted in the development of research on systems such as social networks, biological networks, networks representing financial dependencies, the structure of the Internet, or the organization of transportation systems [106, 142, 144, 167, 291–301].

From a mathematical point of view, a network is the same as a graph, therefore the terms "network" and "graph" are often used interchangeably. Formally, a graph is a pair of sets $G = (V, E)$, where V (called the *vertex set*) is some set (here assumed to be finite) and E (the *edge set*) is a set of two-element subsets of V (in other words, the elements of E are pairs of distinct vertices). The elements of vertex set V are called *vertices* or *nodes*; the elements of edge set E are called *edges* or (sometimes) *links*. The vertices belonging to an edge are called the *endpoints* of that edge. An edge and a vertex being one of the endpoints of that edge are said to be *incident* to each other or be *touching* each other. Vertex set and edge set of a graph G are sometimes denoted by, respectively, $V(G)$ and $E(G)$.

According to the above definition, a graph is a set of some objects (nodes) together with a set of connections between these objects (edges). Connections are binary in their nature, that is, a pair of distinct vertices is either connected by an edge or not. A graph defined in such a way is a *simple graph*; a number of modifications can be introduced to generalize that concept. If pairs of vertices are allowed to be connected by more than one edge (multiple edges connecting the same pair of vertices are sometimes called *parallel edges*) or edges are allowed to connect a vertex with itself (such edges are called *loops*), then a *multigraph* is obtained. If edges are assigned numbers, called *weights* (often used to represent the strength of individual connections), then the graph becomes a *weighted graph*; to avoid ambiguity, graphs whose edges do not have weights are sometimes called *unweighted graphs*. If, instead of being two-element subsets of V (unordered pairs), edges are ordered pairs, then

a direction can be assigned to each of them, and the graph is a *directed graph*; when there is a need to specify explicitly that a graph does not have edge directions, then a term *undirected graph* is used.

A graph can be completely described by the so-called *adjacency matrix*. Let N be the number of nodes of a graph G and let the nodes be numbered by consecutive positive natural numbers; this means that $V(G) = \{1, 2, 3, ..., N\}$. The adjacency matrix A of the graph G is a $N \times N$ matrix with elements a_{ij} $(i, j = 1, 2, 3, ..., N)$, defined as:

$$a_{ij} = \begin{cases} 1, & \text{if } \{i, j\} \in E(G), \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

So, each element of the adjacency matrix of a graph expresses whether a certain connection exists or not. The adjacency matrix of a simple graph is a binary symmetric matrix with zeros on the diagonal. In a weighted graph, the elements of the adjacency matrix represent edge weights, and therefore they can be different from 0 and 1. In a directed graph, the adjacency matrix does not have to be symmetric, as in such a graph the presence of an edge from i to j does not imply the presence of an edge from j to i. If loops are allowed, then the diagonal of the adjacency matrix might contain elements not equal to zero.

An important concept in network theory is the notion of *connected graph*. Let $G = (V, E)$ be a graph. A *path* from some vertex $u \in V$ to some other vertex $v \in V$ is a sequence $(e_1, e_2, e_3, ..., e_n)$ of edges of G such that:

- u is the first endpoint of e_1;
- v is the second endpoint of e_n;
- for each $k = 1, 2, 3, ..., n$, the second endpoint of e_k is the first endpoint of e_{k+1}.

The distinction between the first and the second endpoint of an edge is important only in a directed graph; in an undirected graph, they are interchangeable. A path which starts and ends in the same vertex is called a *cycle*. A graph is a connected graph when any vertex can be reached from any other vertex, that is, when for any pair of vertices (u, v) there exists a path from u to v.

Graphs can be considered abstract objects, as few restrictions are imposed on the nature of graph's vertices and edges. Therefore, the notion of a graph is general enough to be useful in the description of a wide range of systems; this is the reason for which complex networks gain interest in various fields of research. A complex network is often defined as a graph with a nontrivial structure, different from the one observed in random graphs or in graphs with regular, repeatable patterns of organization. The presence of such a structure - occurring in multiple systems whose organization or behavior can in certain aspects be modeled with the use of networks - is often related to system's complexity. Therefore, networks constitute an important tool in studying complex systems.

4.1.1 Network characteristics

Studying complex networks often involves studying a number of network characteristics - quantities describing various properties of networks. Some characteristics are global - describing a network as a whole, some are local - they pertain to a single node. Below, selected quantities often used in the analysis of complex networks are presented. They are given in their basic form designed for unweighted networks, along with their variants generalized to weighted networks; generalizations onto directed networks are not presented, since the networks studied in this work are almost exclusively undirected networks. It is worth mentioning that unweighted

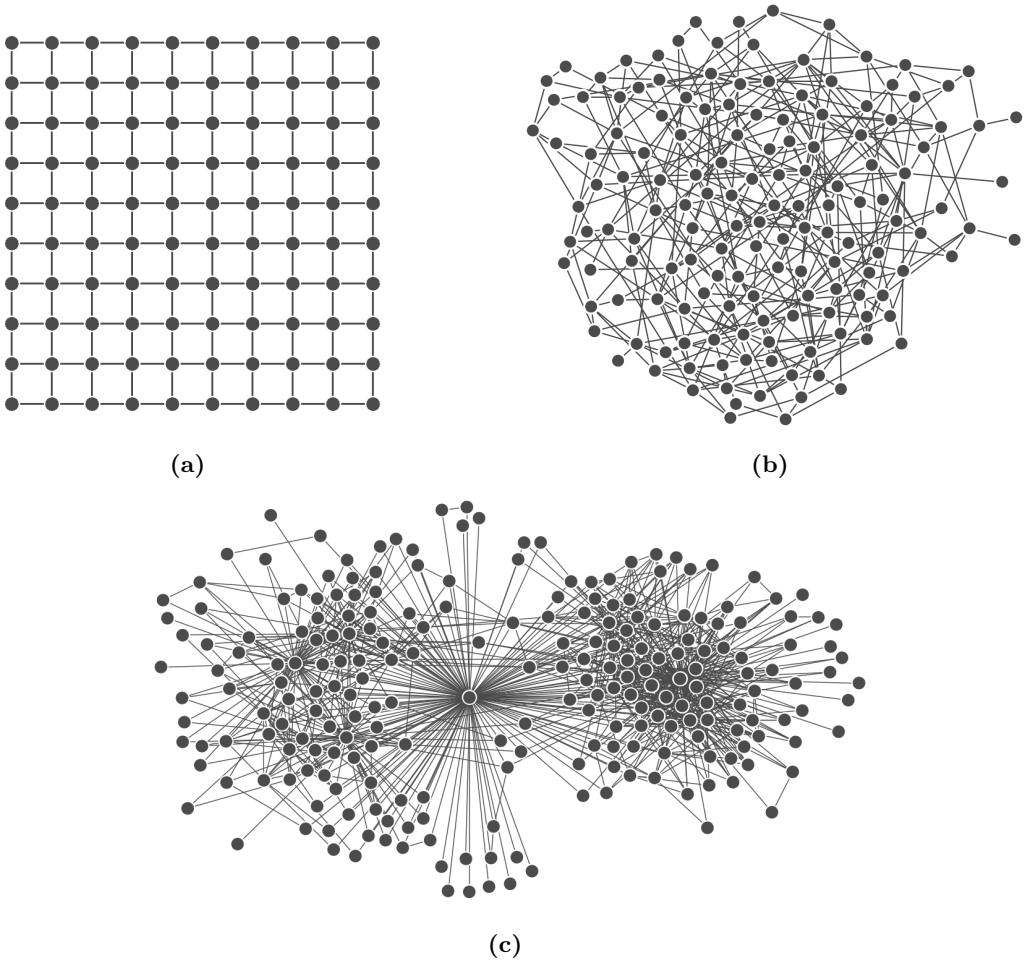


Figure 4.1. Examples of networks: a network with simple, regular structure (a), a random network (b), a network with nontrivial organization (c).

characteristics are not reserved to unweighted networks; they can be determined also for weighted networks - edge weights are then ignored.

Vertex degree and strength

Degree is a quantity describing individual vertices. In an unweighted network, the degree of a node v is the number of edges incident to v, that is, the number of edges that v belongs to. The degree of v is denoted by $\deg(v)$. In weighted networks, the degree can be generalized to weighted degree (also called *strength* and denoted by $\mathrm{str}(v)$), which is the sum of weights of the edges incident to v. An important relationship regarding the (unweighted) node degrees is the degree sum formula, sometimes called *the handshaking lemma*; it states that the sum of all nodes' degrees equals twice the number of edges in the network. With V denoting the vertex set and M denoting the number of edges in the network, this is written as:

$$\sum_{v \in V} \deg(v) = 2M. \tag{4.2}$$

Clustering coefficient

In unweighted networks, the clustering coefficient of a given vertex represents the probability that two randomly chosen direct neighbors of that vertex are also direct neighbors of each other. A *direct neighbor* of a vertex v is here understood as a vertex

connected with v by an edge. Let m_v be the number of edges in the network that link the direct neighbors of v with other direct neighbors of v. Then the clustering coefficient C_u (the subscript "u" comes from the word "unweighted") of the node v is given by:

$$C_u(v) = \frac{2m_v}{\deg(v) \cdot (\deg(v) - 1)}. \tag{4.3}$$

An example of determining node's clustering coefficient in an unweighted network is presented in Fig. 4.2.

Generalization of the clustering coefficient onto the weighted networks can be done in multiple ways. In this work a definition proposed by Barrat et al. [302] is used. Let $S(v)$ denote the set of direct neighbors of a vertex v, and let w_{uv} denote the weight of the edge connecting vertices u and v (if there is no such edge, then $w_{uv} = 0$). Let a_{uv} denote an unweighted adjacency matrix element, i.e. a number defined in Eq. 4.1. The weighted clustering coefficient of v is written as:

$$C_w(v) = \frac{1}{\mathrm{str}(v) \cdot (\deg(v) - 1)} \sum_{u,t \in S(v)} \frac{w_{vu} + w_{vt}}{2} a_{vu} a_{ut} a_{tv}, \tag{4.4}$$

where summation is over all pairs (u, t) of neighbors of v. It is worth noting that if $\deg v = 0$ or $\deg v = 1$, the clustering coefficient cannot be determined from the above-given formulas; in such cases, it is often assumed to be equal to 0.

The above definitions pertain to individual vertices of a network. Global clustering coefficient can be defined in more than one way; in this work the approach based on averaging local clustering coefficients is applied. If V stands for the vertex set of a network and N is the number of elements in V, then the global clustering coefficient of the network is given by:

$$C = \frac{1}{N} \sum_{v \in V} C(v). \tag{4.5}$$

Here the subscript "u" or "w", indicating the unweighted or weighted network, is omitted, because the formula is identical in both cases.

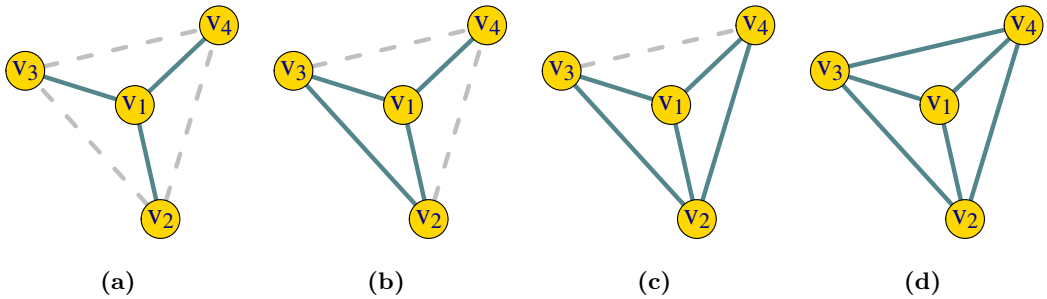


Figure 4.2. An example of computing clustering coefficient in an unweighted network. Node v_1 has 3 neighbors. There are 3 possible connections between the neighbors of v_1; these possible connections are marked with dashed lines. The clustering coefficient of v_1 is equal to the number of edges existing between the neighbors of v_1 divided by the number of possible connections; hence, in (a), (b), (c) and (d) it is equal to 0, $\frac{1}{3}$, $\frac{2}{3}$ and 1, respectively.

Average shortest path length

In unweighted networks, the length of a path between two vertices is the number of edges constituting that path. In weighted networks, the length of a path can be defined as the sum of the reciprocals of edge weights on that path. The length of

the shortest path between vertices u and v is also called the *distance* between u and v and is denoted by $d(u, v)$.

The average shortest path length $\ell(v)$ of a vertex v is the average distance from v to every other vertex in the network. It is one of the measures of the centrality of a vertex in the network, and is given by the formula:

$$\ell(v) = \frac{1}{N-1} \sum_{u \in V \setminus \{v\}} d(v, u), \tag{4.6}$$

in which V is the network's vertex set, and N is the number of elements in V.

The quantity defined above has finite values only in connected networks. If there are at least two vertices that are not connected by any path, the distance between them is not defined; usually it is treated as infinite, and therefore $\ell(v)$ cannot be calculated.

Global average shortest path length is a quantity describing the whole network; it is the average distance between all pairs of vertices. If local average distances $\ell(v)$ for all v in V are given, then the global mean distance in the whole network can be expressed by:

$$\ell = \frac{1}{N} \sum_{v \in V} \ell(v). \tag{4.7}$$

Equations 4.6 and 4.7 apply both to unweighted and weighted networks; the difference between the unweighted and the weighted average shortest path length arises as a consequence of different definitions of distance in those two types of networks.

Assortativity

Assortativity is a global characteristic of a network, describing the preference of vertices to attach to others that have similar degree. A network is called assortative, if vertices with high degree tend to be directly connected with other vertices with high degree, and low-degree vertices are typically directly connected to vertices which also have low degree. In disassortative networks, the high-degree nodes are typically directly connected to the nodes with low degree.

In unweighted networks, the assortativity coefficient can be defined as the Pearson correlation coefficient between the degrees of nodes that are connected by an edge. Let (u, v) denote an ordered pair of vertices that are connected by an edge. Since edges are undirected and the pair (u, v) is ordered, two such pairs can be assigned to each edge in the network. For each pair one can determine the degrees of vertices u and v, and form a pair $(\deg(u), \deg(v))$. The set of all pairs $(\deg(u), \deg(v))$ for all edges can be treated as the set of values of a certain two-dimensional random variable (X, Y). With such notation, the assortativity coefficient r_u is expressed by the Pearson correlation coefficient of variables X and Y:

$$r_u = \operatorname{corr}(X, Y). \tag{4.8}$$

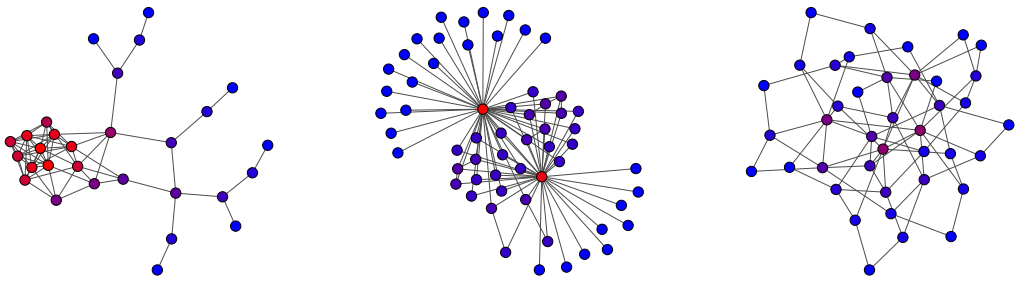
The generalization of the above formula to weighted networks used in this work is done by replacing the degrees of vertices by their strengths, and calculating weighted correlation coefficient instead of the usual one. Let (X, Y) be a two-dimensional random variable whose values are pairs $(x, y) = (\operatorname{str}(u), \operatorname{str}(v))$ for all pairs of vertices (u, v) connected by an edge. Let w be a function that assigns to each pair $(x, y) = (\operatorname{str}(u), \operatorname{str}(v))$ the weight of an edge connecting u and v. Then the weighted assortativity coefficient r_w can be written as:

$$r_w = \operatorname{wcorr}(X, Y; w), \tag{4.9}$$

where $\mathrm{wcorr}(X, Y; w)$ denotes the weighted Pearson correlation coefficient of variables X and Y with the weighing function w. Some of the definitions encountered in literature, for example in [303], are equivalent to the one given above.

In this work, one more variant of assortativity coefficient is introduced and used in addition to the one presented above. It is defined with the use of Spearman correlation coefficient instead of Pearson correlation coefficient; to avoid ambiguity it is referred to as *rank assortativity coefficient* and denoted by ρ. The reason behind such an idea is the fact that while Pearson correlation coefficient measures only linear correlations, Spearman correlation coefficient - which is equal to Pearson correlation coefficient computed for ranks of the original variables - allows to detect monotonic relationships (whether linear or not). To obtain unweighted and weighted rank assortativity coefficients ρ_u and ρ_w, it is sufficient to replace X and Y with their ranks, $R(X)$, $R(Y)$, in Equations 4.8 and 4.9, respectively. As a technical remark, it is worth mentioning that the definition of rank used in Spearman correlation coefficient might assign fractional ranks to repeating observations: identical values are assigned ranks equal to the average of their positions in the sorted sequence of values.

Since assortativity coefficient is expressed by correlation coefficient, it has values between -1 and 1. Networks with positive r are assortative, while networks with negative r are disassortative. Examples of networks with different values of assortativity coefficient are presented in Fig. 4.3.



(a) **(b)** **(c)**

Figure 4.3. Examples of unweighted networks with different values of assortativity coefficient. The network in (a) is assortative ($r = 0.74$), the network in (b) is dissasortative ($r = -0.82$), and in the network in (c), the degrees of directly linked vertices are not correlated ($|r| < 0.01$). In each network, the vertices are colored according to their degree - blue, purple and red colors correspond respectively to low, medium and high degree.

Modularity

Modularity is a global characteristic of a network, measuring the extent to which the set of network's vertices can be divided into disjunctive subsets which maximize the density of edges within them, and minimize the number of edges connecting one with another .

Let $G = (V, E)$ be an unweighted network. A *partition of the network* G is some division of V into disjoint subsets (called modules, clusters or communities). Let a_{uv} denote the adjacency matrix element (defined in Eq. 4.1). Let c_v denote the module to which the vertex v is assigned by some given partition. The *modularity of the partition* is defined as:

$$q_u = \frac{1}{2M} \sum_{u,v \in V} \left(\left[a_{uv} - \frac{\deg(u) \deg(v)}{2M} \right] \delta(c_u, c_v) \right), \tag{4.10}$$

where M is the number network's edges, $\deg(u), \deg(v)$ are degrees of vertices u and v, and function $\delta(c_u, c_v)$ has value 1 if $c_u = c_v$ and 0 otherwise.

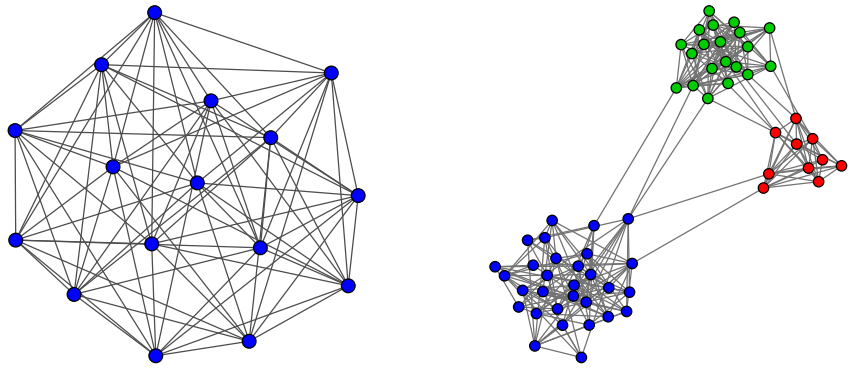
Modularity of a partition has value between -1 and 1, and indicates whether the density of edges within the given modules is higher or lower than it would be if edges were distributed at random. The random network that serves as a reference in this definition is constructed using the *configuration model* (presented in section 4.1.2).

The *modularity of a network*, denoted by Q_u, is the maximum value among modularities q_u of all possible partitions. Determining the network's modularity precisely is computationally intractable, hence a number of heuristic algorithms have been proposed. In this work, modularity is estimated using the Louvain algorithm [304].

The generalization of modularity onto weighted networks can be done by replacing the quantities appearing in Eq. 4.10 by their weighted counterparts. If w_{uv} denotes the weight of an edge connecting vertices u and v, W is the sum of all edge weights, and $\mathrm{str}(u)$, $\mathrm{str}(v)$ are the strengths of vertices u, v, then the modularity of a given partition is equal to:

$$q_w = \frac{1}{2W} \sum_{u,v \in V} \left(\left[w_{uv} - \frac{\mathrm{str}(u)\,\mathrm{str}(v)}{2W} \right] \delta(c_u, c_v) \right). \tag{4.11}$$

Again, the weighted modularity of the network, Q_w, is the greatest of modularities obtained in all possible partitions of the network. Examples of networks with different values of modularity are presented in Fig. 4.4.



(a) **(b)**

Figure 4.4. Examples of unweighted networks with (a) low modularity ($Q_u = 0.07$) and (b) high modularity ($Q_u = 0.58$). In figure b) the colors represent the partition which leads to the given value of modularity.

4.1.2 Random network models

A number of complex network properties can be considered universal to some extent, since they are shared among networks representing many different systems. The existence of such properties drives the development of various random network models [305] - which can be understood as numerical procedures designed to generate networks having some properties predefined, but being random in terms of the remaining characteristics. The term "random network" is sometimes used to refer to one particular network model - Erdős–Rényi model (discussed below) - but here it is used in a more general sense - to refer to networks constructed with the use of procedures involving some random processes. Random network models allow to investigate the origin of certain phenomena and organization patterns observed in networks representing diverse natural systems. Moreover, random network models

are useful in situations when there is a need to compare the properties of some studied network with the properties of a random network keeping some features of the original network (like the size), but lacks other traits implied by original network's structure. Such a comparison can be used to assess the significance of various results - when some characteristic of the original network is not observed in a random network, then it can be considered as resulting from network's specific organization rather than a coincidence.

Erdős–Rényi networks

The most "basic" model of random networks is the *Erdős–Rényi model* [306, 307]. The name comes from the names of the authors, P. Erdős and A. Rényi, but it is worth noting that the model was independently studied by E. Gilbert [308]. The model generates graphs which are unweighted and, in the version presented here, also undirected (however, generalization onto directed networks is straightforward). The model exists in two slightly different variants. In the first one, here denoted by $G(N, M)$, the network is chosen uniformly at random from the set of all graphs with N vertices and M edges. Constructing a $G(N, M)$ network consists of defining an N-element vertex set, randomly choosing M pairs of distinct vertices from the set of all possible pairs of distinct vertices (the number of all possible pairs is $M_{\max} = N(N-1)/2$), and introducing edges constituted by the chosen pairs into the network. In the second variant of the model, denoted by $G(N, p)$, each pair of distinct vertices is connected by and edge with a fixed probability p. Constructing a $G(N, p)$ network consists of defining an N-element vertex set, iterating over all pairs of distinct vertices and connecting each pair with an edge with probability p or leaving it unconnected with probability $1 - p$, independently from other pairs. Since deciding the presence or the absence of each edge can be treated as a Bernoulli trial with success probability equal to p, and the number of pairs of distinct vertices is equal to $M_{\max} = N(N-1)/2$, the number of edges M in a network generated by the $G(N, p)$ network is a random variable having a binomial distribution with parameters $M_{\max}$ and p. Hence, the probability $P(M)$ that the generated network has M edges can be expressed as:

$$P(M) = \binom{M_{\max}}{M} p^M (1-p)^{M_{\max}-M}, \tag{4.12}$$

where $\binom{x}{y}$ denotes the value of the binomial coefficient for x and y.

The expected value of the number of edges $\langle M \rangle$ is equal to $pM_{\max}$. For a fixed p (different from 0 and 1) and $N \to \infty$, the value of $M_{\max}$ also goes to infinity; then, by virtue of de Moivre-Laplace theorem, the binomial distribution with parameters $M_{\max}$ and p can be approximated by the normal distribution with mean $pM_{\max}$ and standard deviation $\sqrt{p(1-p)M_{\max}}$. This allows to show that relative dispersion of the number of edges M, which might be expressed by the standard deviation of M divided by the average value of M, goes to zero in the considered limit ($N\infty$, fixed p). So, for large enough networks, one can approximate the number of edges in a $G(N, p)$ network by the average value $\langle M \rangle = pM_{\max}$ (as the relative error of such an approximation becomes negligible for large networks). This approach allows to establish a correspondence between $G(N, M)$ and $G(N, p)$ models - it can be stated that for large enough N, networks generated by the $G(N, p)$ model behave similarly to networks generated by the $G(N, M)$ model with $M = pN(N-1)/2$. However, the correspondence is only an approximation and holds only in terms of some properties. For example, if the property of interest is having an even number of edges, then $G(N, M)$ and $G(N, p)$ might behave differently irrespective of how

large are the networks ($G(N, M)$ generates a fixed number of edges, while $G(N, p)$ can generate both odd and even numbers of edges for a single combination N and p).

Erdős–Rényi networks are a very useful model of "purely random" networks - networks which have no specific patterns of organization beyond the ones arising from distributing edges randomly over the graph. Their capability of modeling real-world networks (networks representing the organization of various real-world systems) is severely limited. One of fundamental reasons for this is the distribution of node degrees. Node degrees in Erdős–Rényi networks are binomially distributed; if k denotes the degree of a node and $P(k)$ denotes the probability mass function, then for a large enough $G(N, p)$ network ($N \to \infty$) one can write:

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{4.13}$$

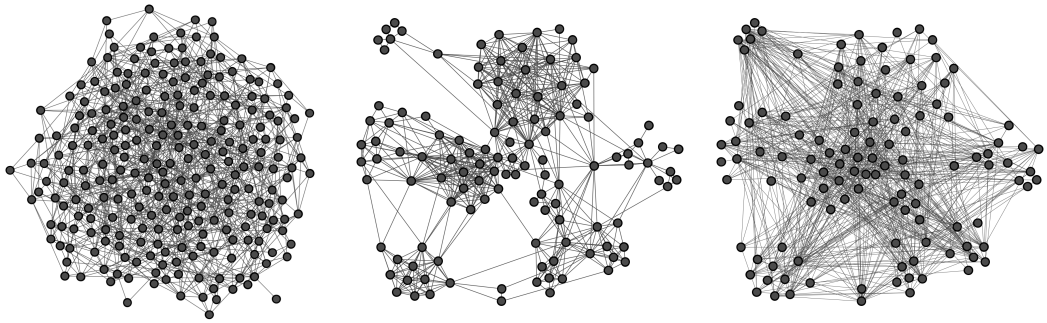
For the reasons mentioned above, the relative dispersion of the this binomial distribution decreases with growing N. Therefore, it can be stated that in an Erdős–Rényi network, node degrees are concentrated around the average value. Node degrees in real-world complex networks, on the other hand, usually span many orders of magnitude, which is often a consequence of being distributed according to a power-law distribution. The fact that certain properties of Erdős–Rényi networks (like the shape of node degree distribution) seem to be unrealistic in many situations, led to the development of other random network models, attempting to mimic at least some of the characteristics of real-world complex networks.

Configuration model

Configuration model [291] is a model which generates networks with an explicitly prescribed node degree distribution; the distribution is specified by a sequence of numbers, in which each number is the degree of one node. If the given sequence of numbers $k_1, k_2, k_3, ..., k_N$ satisfies conditions required to constitute a valid degree sequence (the sum of all the numbers in the sequence has to be an even number, for instance), then, in the simplest variant of the model, the network (undirected and unweighted) is generated as follows. The set of N nodes is created, and each node is given the number of edge stubs equal to its target degree, in other words, the i-th node gets k_i edge stubs. Then two stubs are chosen uniformly at random and connected to form an edge. Connecting pairs of randomly chosen stubs is repeated until there are no unconnected stubs left. The resulting network has exactly the degrees specified by the sequence $k_1, k_2, k_3, ..., k_N$.

The configuration model is often used in situations when there is a need to determine whether certain properties of some network are directly related to node degree distribution. The model allows to construct a randomized version of the studied network (using the degree sequence taken from that network) which can be expected to preserve properties resulting from node degree distribution and to be random in other regards. Just as the Erdős–Rényi model can be considered a model of networks whose properties are a result of a random arrangement of edges not subject to any specific restrictions (in a $G(N, M)$ model all arrangements of M edges on among N vertices are equally probable), the configuration model can be considered a model of a network whose structure is random, but has a condition of preserving prescribed node degrees imposed on it. This, together with the fact that node degrees are of fundamental importance in network analysis, makes configuration model particularly useful in studies on complex networks.

It is important to note that the procedure of generating networks presented above is the simplest algorithmic approach to configuration model, which has certain undesirable properties. As the pairing of edge stubs to connect is unrestricted with



(a) **(b)** **(c)**

Figure 4.5. Examples of random network models usage. An Erdős–Rényi network $G(N, p)$ with parameters $N = 250$ and $p = 0.03$ is shown in (a). In panel (b), an example of a network with moderately modular organization is presented; the randomization of that network based on the configuration model (that is, a network with the same number of nodes and edges and having the same degrees as in the original network) is presented in (c).

regard to the allowed choices of pairs, a network generated by the presented method might contain loops and parallel edges, which make it a multigraph. Also, when the number of edges is relatively small, the generated network might not be connected. This is a potential problem when configuration model is used as a method of generating a randomized version of a connected network, as it might be preferable for the randomized network to be connected as well. However, there exist methods [309] which overcome these problems and generate networks with the desired properties.

The configuration model can serve as a starting point for more complicated network randomization procedures. For example, when a weighted network needs to be randomized in such a way that the (unweighted) node degrees are kept unchanged, then one of the possibilities is generating a random network according to the configuration model (using the unweighted degrees of the considered weighted network as the input degree sequence), and then randomly assigning the edge weights from the original network to the edges of the generated network. The distributions of node degrees and of edge weights in the obtained network are identical to the distributions describing the original network. However, other properties, even the ones related to node degrees and edge weights (node strengths, for instance), might not be preserved. In case when such properties are to be kept unchanged, some other method of randomization needs to be applied.

The two presented types of random network models (Erdős–Rényi model and configuration model) are examples of well-known, widely studied models having a significant influence on research on complex networks; however, they certainly do not constitute an exhaustive list. There exist many more models of random networks, designed for various applications, like testing computational methods of network analysis or explaining phenomena observed in systems having a network representation. An example worth mentioning in that context is the *Barabási-Albert model* [125], which uses a mechanism based on a Yule process to generate networks with power-law distributions of node degrees; the model constitutes an important contribution to the study of power laws' emergence in complex networks.

4.1.3 Minimum spanning trees

One of the problems sometimes encountered in research on complex networks is related to the fact that the analysis of highly complicated structures might suffer from the presence of a large number of details which can make it difficult to get the general understanding of the essential properties. Investigating the characteristics

of large, densely connected networks sometimes benefits from removing certain elements which can be considered redundant from the viewpoint of the analysis. One of the concepts particularly useful in this context is the so-called *minimum spanning tree* (MST). Minimum spanning tree is defined as a subnetwork of a given weighted network satisfying certain conditions; constructing the minimum spanning tree of a network can be thought of as a procedure which removes all the edges except for the most important ones. Therefore, in certain cases, it can be considered a particular kind of a "filter" removing information of little importance.

To define a minimum spanning tree, it is convenient to define the notions of a *tree* and a *spanning tree* first. A tree is a graph which is connected, undirected and acyclic (not containing any cycles). In a tree, exactly one path exists between each pair of vertices. A spanning tree of some connected graph $G = (V, E)$ is a tree with the vertex set V and the edge set being a subset of E. In other words, the minimum spanning tree of a connected graph G is a tree containing all of the vertices of G and some subset of the edges of G. Let G be an undirected connected graph in which every edge has a real number, called *edge cost*, assigned to it. The minimum spanning tree (MST) of G can be defined as the spanning tree of G having the minimum possible sum of edge costs. The fact that costs are numbers assigned to each edge of G makes it possible to consider G a weighted graph and to treat costs as edge weights. In fact, definitions of MST often do not utilize the notion of edge cost, and define MST as the spanning tree of a weighted graph minimizing the sum of edge weights. However, in some situations it is desirable to introduce costs distinct from weights and to minimize the sum of costs instead of the sum of weights. This happens, for example, when the studied network is a weighted network in which edge weights can be interpreted in terms of connection intensity - the greater the weight, the stronger the relationship between the connected vertices. In such a case, introducing edge costs equal to, for example, the reciprocals of edge weights, and using those costs to construct an MST allows to treat the obtained MST as a subnetwork keeping only the most important edges of the original network. There are several algorithms finding the minimum spanning tree of a graph; the algorithm used in this work is the Prim's algorithm [310,311]. An example of constructing the MST of a network is presented in Figure 4.6.

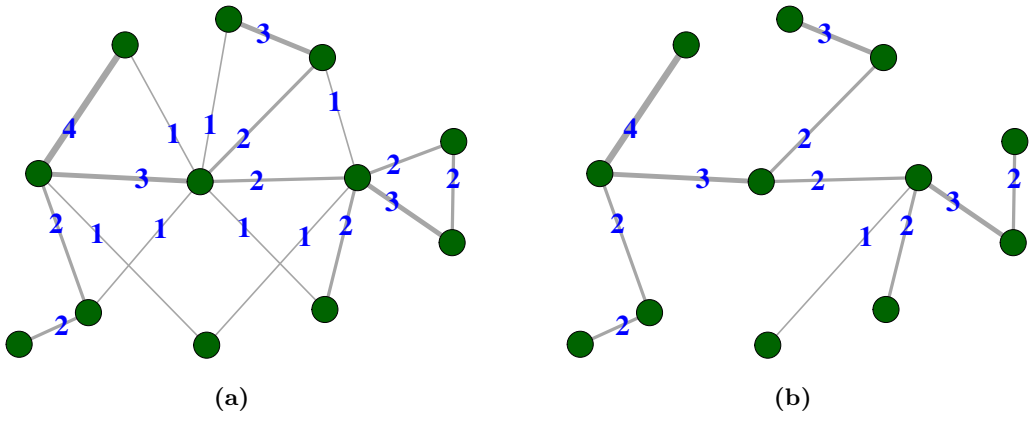


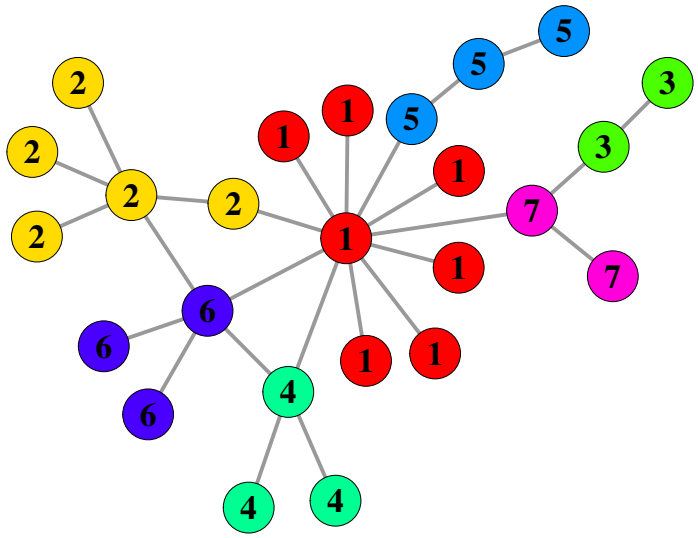
Figure 4.6. A weighted network (a) and a minimum spanning tree of that network (b). The numbers labeling edges are edge weights; edge costs (minimized by the MST) are equal to the reciprocals of weights.

Minimum spanning trees have a number of applications in various fields, like the study of financial markets [142, 144], image processing [312], or the analysis of brain networks (networks representing certain aspects of brain organization, constructed from neuroimaging data) [313]. It is worth mentioning that although the

concept of MST is inherently related to weighted networks, there exist methods of applying MST to the analysis of unweighted networks. This is done by transforming an unweighted network into a weighted one and constructing the MST of the latter. For example, edges in an unweighted network can be assigned weights based on the number of shortest paths they are part of - the greater the number of network's shortest paths containing the considered edge, the greater the weight assigned to that edge; this approach comes down to the analysis of the so-called *betweenness* centrality [314].

4.1.4 Fractal analysis of networks

Scale-free networks owe their name to the particular form of their node degree distributions; those distributions, being power laws, lack characteristic scale. But the specific shape a of node degree distribution is not the only aspect in which organization of the network can be scale-free. A network can be organized into a hierarchical, statistically self-similar structure, which can be identified with the use of fractal analysis methods, like the estimation of box-counting dimension. Box counting method applied to a network relies on partitioning the set of network's nodes V into possibly small number of boxes, that is, disjoint subsets V_i ($\bigcup_i V_i = V$) such that for every pair of nodes v_1, v_2 belonging to the same V_i, the distance $d(v_1, v_2)$ between v_1 and v_2 is less or equal to some fixed value, s (box size). The number of boxes obtained in such a partition is denoted by $N(s)$. Making partitions with many different values of s gives a collection of different values of $N(s)$; if a power-law describes $N(s)$ behavior: $N(s) \propto s^{-d_C}$, then d_C can be interpreted as the box-counting dimension of the network [315].



(a)

Figure 4.7. Exemplary covering of a network with boxes of size $s = 2$. Nodes having the same label belong to the same box.

Figure 4.7 shows an example of covering a network with boxes. In general, the minimum number of boxes of given size required to cover a network cannot be computed exactly except for very small networks. For that reason, a number of algorithms finding an approximate solution have been proposed; the algorithm used in this work is based on the idea of *greedy coloring* [316]. To cover a network G with boxes of size s, the algorithm transforms G into another network G' with an identical set of nodes and having the following property: for any nodes u, v, an edge between

u and v exists if and only if the distance between u and v in the original network G is greater than s. Then a greedy algorithm is used on the resulting network G' to solve the problem of *vertex coloring*, that is, labeling the nodes in such a way that no two nodes connected by an edge have the same label. The obtained labels constitute the partition of the original network G - nodes having the same label are assigned to the same box. The construction procedure of G' along with the coloring rule ensure that nodes separated by distances larger than s never end up in the same box.

Fractality, meaning the presence of scale-free, statistically self-similar organization of a network, is related to several other properties and patterns of network's behavior under certain circumstances. For example, fractality has been shown to be related to a tendency of high-degree nodes to be separated instead of being directly connected (this effect, occasionally called hub repulsion, is expressed by network's disassortativity); also, fractality decreases network's vulnerability to attacks (which can be interpreted as the number of high-degree nodes which need to be removed - together with the incident edges - to make large pieces of the network disconnect entirely from the remaining part) [317].

It is worth mentioning that fractal analysis is sometimes applied to some transformed or filtered versions of a network instead of the original one; one of the examples of transformation used for this purpose is constructing a minimum spanning tree (which in this context is sometimes referred to as *network skeleton*) [318]. In some networks, fractal properties of the original structure and of the network skeleton are approximately the same; however, there exist networks whose self-similarity can be detected only after applying a transformation which removes edges of little importance [319].

4.2 Word-adjacency networks

A number of problems related to natural language can be studied with the use of network theory. Networks allow to represent language on various levels of its structure - they can represent word co-occurrences, semantic similarities or grammatical relationships, for instance. Such networks, collectively called linguistic networks, often consist of a large number of nodes and edges and exhibit complex patterns of organization, but graphs with relatively simple structure (for example consisting of over a dozen of vertices), also have their applications in language-related areas of research (examples of such graphs are parse trees, presented in Chapter 1). Graphs and networks have been used to approach various practical problems related to natural language processing, like keyword selection, document summarization, word-sense disambiguation or machine translation [320–323]. Also, network formalism is used in research areas at the interface between linguistics and other scientific fields, for example in sociolinguistics, which, by studying social networks (networks representing the organization of human communities), investigates human language usage and evolution [324, 325].

An example of a linguistic network with a very simple construction procedure is the word-adjacency network, sometimes also called word co-occurrence network. A word-adjacency network is constructed from a text or a corpus in the following way: each (unique) word in a corpus becomes a vertex of the network. If two words appear next to each other in the text at least once, then the nodes corresponding to those words are connected by an edge. A network defined in such a way is an unweighted word-adjacency network. If the numbers of co-occurrences of the words

are assigned to edges as weights, then a weighted word-adjacency network is created. Word-adjacency network can be directed or undirected - depending on whether the ordering of word pairs is taken into consideration. Words can be lemmatized or kept in their original forms - each of the choices leads to slightly different variant of the resulting network. All word-adjacency networks discussed in this chapter are undirected, weighted networks. When their unweighted characteristics are of interest, edge weights are neglected. The words used to construct the networks are not lemmatized. Examples of word-adjacency networks are shown in Figure 4.8.

Despite their simplicity, word-adjacency networks have a number of interesting properties and provide a useful language representation, as they are able to capture a number characteristics of the underlying text. Since each occurrence of a particular word adds 1 to the weight of the edge between that word and the previous word, the strength $\mathrm{str}(v)$ of a vertex v in a weighted word-adjacency network is equal to $2\,\omega(v)$, where $\omega(v)$ is the frequency of the word represented by v (the number of times it appears in the text). The exception from this rule happens in cases in which two or more consecutive occurrences of the same word appear next to each other in the text; such cases are ignored in the process of network construction (no edges connecting a vertex with itself are created). Since such cases are relatively rare, it can be stated that $\mathrm{str}(v) \approx 2\,\omega(v)$. Vertex degree (which is strongly correlated but in general not equal to vertex strength - Fig. 4.13) gives information about how many different co-occurrence pairs a word forms with other words in the corpus. Clustering coefficient describes the structure of the node's neighborhood; it reveals how often the words being direct neighbors of some word v are also direct neighbors of each other. Measuring network's assortativity provides information about the correlations between the quantities describing words occurring next to each other (degrees and strengths), and modularity gives insight into the extent to which the vocabulary of a text can be divided into clusters of words frequently appearing together.

4.3 Comparing networks of different sizes

Word-adjacency networks constructed from texts of different lengths in general have different sizes - they differ in the numbers of nodes and edges and in magnitudes of edge weights. To compare the properties of word-adjacency networks representing different texts, it is sometimes useful to perform some type of characteristics' normalization. One of possible ways of doing that is referring to the characteristics of a network randomized in a specific way. The randomization procedure adopted in this chapter consists of shuffling the order of the words in a text in a random fashion, and constructing a word-adjacency network from the so-obtained random text (Figure 4.9). It is worth mentioning that this procedure preserves node strengths, as word frequencies remain unchanged. After constructing the randomized network, the characteristic of interest - either global (pertaining to the whole network) or local (describing specific words) is determined for the randomized network. Randomizing the network and computing the desired characteristic is repeated multiple times and the results are averaged; the average value of the studied quantity in a randomized network serves as a reference for the value obtained for original network. Consequently, the investigated quantities can be of the form $g - g^{rand}$ or g/g^{rand}, where g is some network characteristic like assortativity coefficient, clustering coefficient, or modularity, and g^{rand} is the average value of the same characteristic in a randomized network (network constructed form a randomized text). When g is a local

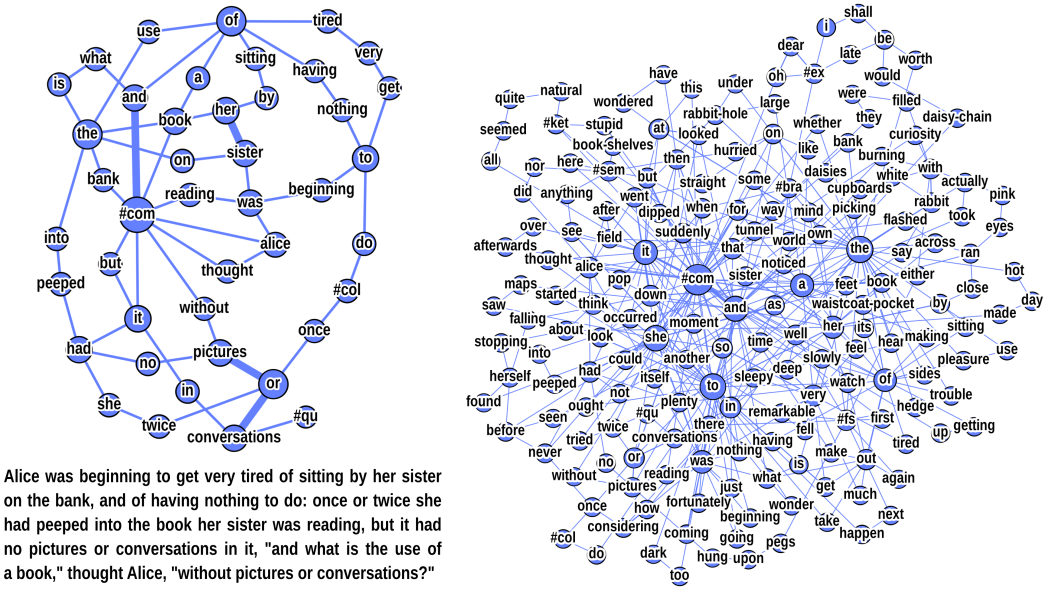


Figure 4.8. Word-adjacency networks constructed from text samples of different lengths excerpted from *Alice's Adventures in Wonderland* by Lewis Carroll. The text samples used to construct the networks are: the first sentence of the book (a), the first 10 sentences (b), the first 5000 words (c), the whole book (d). In (c) and (d) node labels are not shown because of networks' size. In all the presented networks, punctuation marks are treated as words; their labels start from "#" symbol. The sentence used to construct the network in (a) is shown below the resulting network.

characteristic, like the clustering coefficient of a node representing a specific word in a network, then g^{rand} is the average value of that characteristic in a randomized network, computed for the node corresponding to the same word. The choice of $g - g^{rand}$ or g/g^{rand} as a quantity to investigate is to some degree arbitrary; it may be specific to a particular problem and depend on a particular characteristic.

Normalizing network parameters is aimed to allow for a meaningful comparison between networks of different sizes, and also to compensate for the effects being a direct result of words' different frequencies. Since a randomized text has the same length and word frequency distribution as its original source text, it can be anticipated that normalized characteristics neglect the purely frequency-based effects and capture the properties of network's specific organization, as they express how the

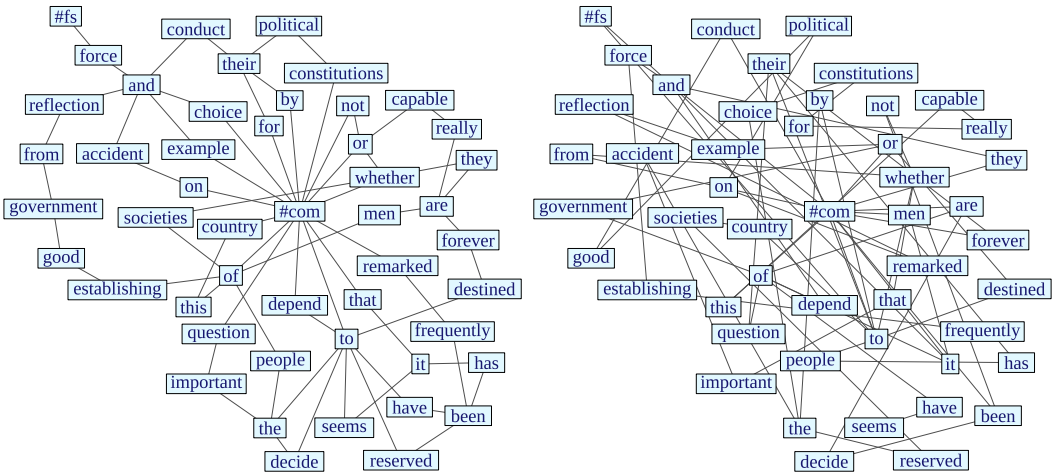
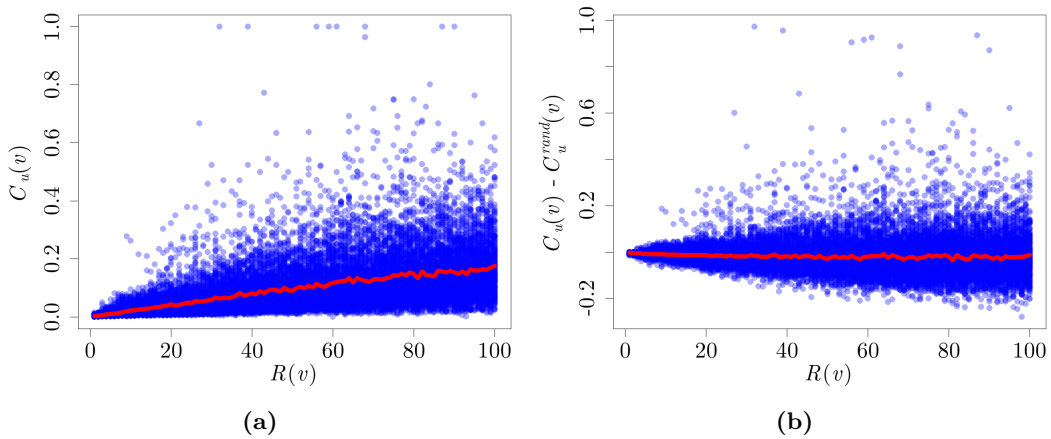


Figure 4.9. An unweighted word-adjacency network and its randomization. Figure (a) presents the network created from one sentence excerpted from *The Federalist Paper No. 1* by Alexander Hamilton. Figure (b) shows the network created from the same piece of text, but with randomly shuffled words. Punctuation is taken into consideration in the construction of network; comma and full stop are denoted by "#com" and "#fs", respectively.

network structure differs from the structure that would be observed if words were placed randomly in the text. An example of normalization's impact on network characteristics can be seen in Figure 4.10, where unweighted clustering coefficients $C_u(v)$ of 100 most frequent words in each of the texts listed in Appendix B.1 are presented, along with their normalized counterparts. The tendency of $C_u(v)$ to increase with increasing word rank $R(v)$, is not observed for $C_u(v) - C_u^{rand}(v)$, expressing the differences between the clustering coefficients in the original and in the randomized network.

4.4 Punctuation in word-adjacency networks

The word-adjacency networks' construction procedure can be extended to include objects other than words into the analysis. An extension which seems natural in this context is taking punctuation marks into consideration - when treated as words, punctuation marks become vertices of word-adjacency networks (the set of studied punctuation marks is the same as in section 3.7 - it consists of period, question mark, exclamation mark, ellipsis, comma, dash, semicolon, colon, left parenthesis and right parenthesis; in the figures presented in this work these marks are denoted by #fs, #qu, #ex, #ell, #com, #pau, #sem, #col, #bra and #ket, respectively). It turns out that in terms of certain parameters expressing their role in a network, nodes representing punctuation marks behave in the same way as nodes corresponding to words of comparable frequency in the text [254]. An example of such an effect can be seen in Figure 4.11. The results presented there are obtained by transforming



(a) **(b)**

Figure 4.10. An example of the influence of word frequencies on word-adjacency network's local characteristics. For each of the networks representing individual texts from dataset specified in Appendix B.1, the unweighted local clustering coefficient in its unnormalized (a) and normalized (b) form ($C_u(v)$ and $C_u(v) - C_u^{rand}(v)$, respectively) is plotted against the rank of a word (the position on the list of most frequent words); this is done for 100 most frequent words. The red line represents the average values, that is the values of clustering coefficients obtained for consecutive ranks, averaged over all the considered networks. It can be observed that while the unnormalized characteristic depends on word frequency (as evidenced by the increase of average $C_u(v)$ with increasing $R(v)$), the normalized quantity does not seem to exhibit such a dependence. Therefore, it can be anticipated that the variability of the normalized characteristic can be attributed to the specific traits of a word in a particular network rather than to word frequency.

the corpora consisting of books from Appendix B.1 into word-adjacency networks (with punctuation marks included), and investigating selected words' properties, expressed by local clustering coefficients and average shortest path lengths. For each language, the set of studied words is composed of two parts: one being a set of a few most frequent words (and punctuation marks) in the analyzed corpus, and the other consisting of words listed in Table 4.1. While the role of the most frequent words in language is often mainly grammatical, the words in the table are words carrying certain meanings - they refer to specific objects and concepts. Their frequencies are significantly lower than the frequencies of a few most frequent words, but high enough to allow for statistically reliable analysis of their properties. It can be seen in Figure 4.11 that in terms of normalized unweighted local average shortest path lengths $\ell_u - \ell_u^{rand}$ and normalized unweighted local clustering coefficients $C_u - C_u^{rand}$, the two groups of words exhibit different patterns of variability. While most frequent words tend to be confined along the horizontal axis on the $(\ell_u - \ell_u^{rand}, C_u - C_u^{rand})$ plane, the words listed in Table 4.1 are more scattered with respect to their normalized clustering coefficients. In all the studied languages, punctuation marks seem to belong to the regime determined by most frequent words. This is in accordance with the results of frequency analysis, supporting the idea that from certain points of view punctuation marks can be treated as words; not only they have frequencies comparable to the frequencies of most frequent words and fit into the power-law regime of rank-frequency distributions, but also some of their properties in word-adjacency networks resemble the properties of high-frequency words. As a consequence of the presented line of reasoning, all the word-adjacency networks studied in this chapter involve punctuation marks.

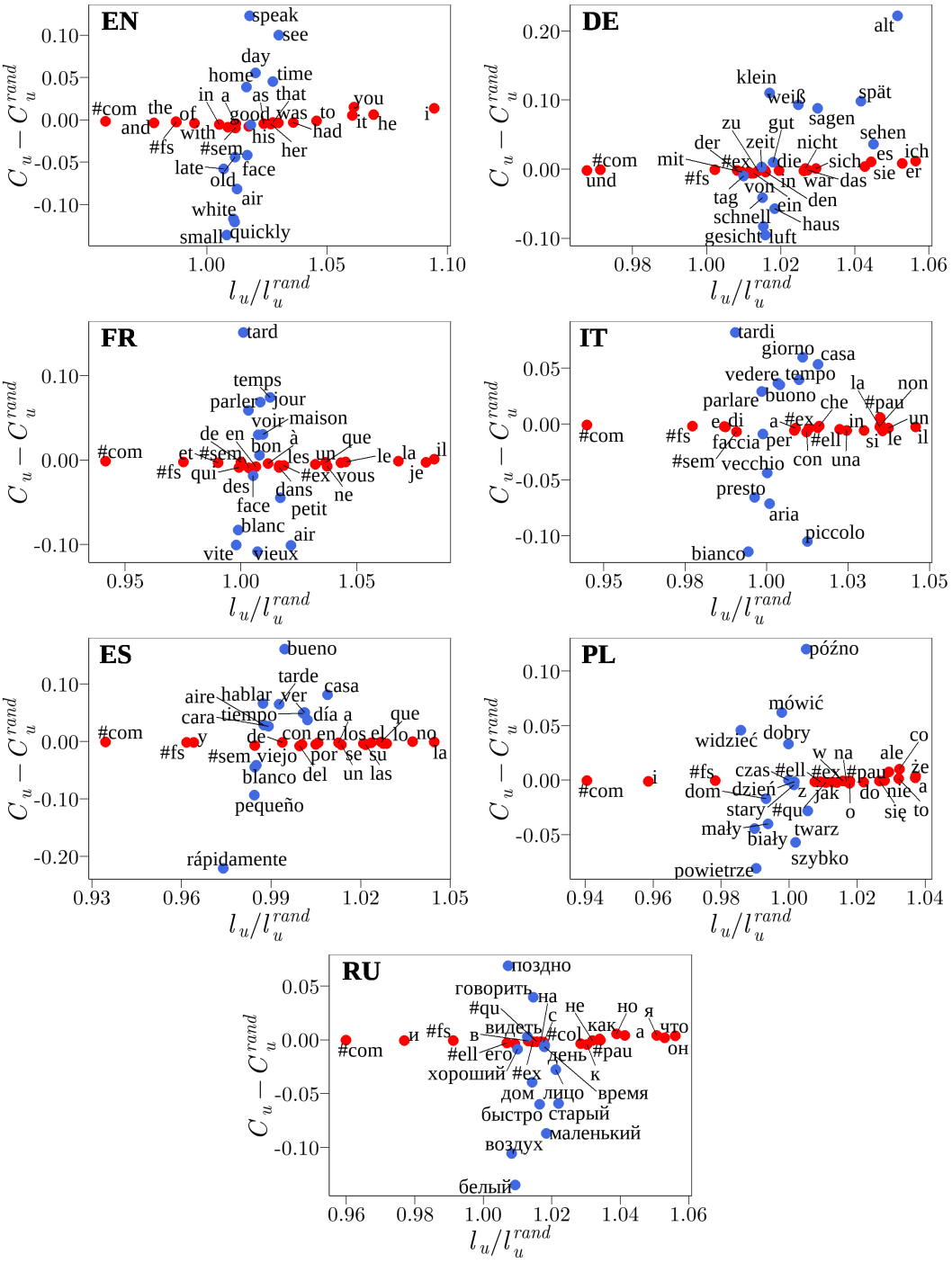


Figure 4.11. Average shortest path lenghts and clustering coefficients (both unweighted and normalized) of selected words and punctuation marks in word-adjacency networks constructed from corpora in 7 languages: English (EN), German (DE), French (FR), Italian (IT), Spanish (ES), Polish (PL) and Russian (RU). The corpora consist of books listed in Appendix B.1. The set of words considered in each language is composed of two word groups. The first group is constituted by 20 words having the highest frequencies in the corpus (punctuation marks are treated as words); those words are marked with red dots in the figures. The second group consists of the words collected in Table 4.1. In contrast to most frequent words, whose role is usually mostly grammatical, the words from the table have specific meanings - they are references to specific objects and concepts. They are marked with blue dots in the figures. It can be observed that on the plane $(\ell_u/\ell_u^{rand}, C_u - C_u^{rand})$ the most frequent words tend to be aligned along the horizontal axis, while the second group of words has more variability in the vertical direction (but some dispersion in the perpendicular direction is also visible). In terms of the presented characteristics, punctuation marks behave similarly to ordinary words with high frequencies.

Table 4.1. Words used in the analysis presented in Fig. 4.11 (along with the set of most frequent words). The table is constructed by translating a set of arbitrarily selected, predefined meanings into all of the 7 studied languages (it should be noted that the accuracy of the translations might vary, as a word in one language might not have a counterpart with exactly the same meaning in other languages). One table row corresponds to one meaning. The words in the table are typically words with moderate frequencies; their ranks (positions on the list of most frequent words) in the studied corpora (corpora constructed from books listed in Appendix B.1) are given in parentheses.

English	German	French	Italian	Spanish	Polish	Russian
time	Zeit	temps	tempo	tiempo	czas	время
(82)	(110)	(107)	(103)	(101)	(194)	(89)
face	Gesicht	face	faccia	cara	twarz	лицо
(149)	(230)	(376)	(248)	(253)	(148)	(135)
home	Haus	maison	casa	casa	dom	дом
(195)	(295)	(195)	(79)	(63)	(421)	(323)
day	Tag	jour	giorno	día	dzień	день
(138)	(279)	(128)	(111)	(93)	(179)	(140)
air	Luft	air	aria	aire	powietrze	воздух
(313)	(468)	(330)	(622)	(326)	(715)	(806)
old	alt	vieux	vecchio	viejo	stary	старый
(86)	(894)	(324)	(273)	(375)	(167)	(561)
good	gut	bon	buono	bueno	dobry	хороший
(95)	(145)	(178)	(682)	(273)	(532)	(1871)
white	weiß	blanc	bianco	blanco	biały	белый
(283)	(174)	(697)	(668)	(654)	(1235)	(1687)
small	klein	petit	piccolo	pequeño	mały	маленький
(292)	(1154)	(160)	(445)	(1292)	(690)	(1024)
late	spät	tard	tardi	tarde	późno	поздно
(539)	(1039)	(507)	(567)	(215)	(1359)	(966)
quickly	schnell	vite	presto	rápidamente	szybko	быстро
(836)	(401)	(541)	(324)	(2876)	(391)	(275)
see	sehen	voir	vedere	ver	widzieć	видеть
(94)	(166)	(117)	(225)	(105)	(739)	(432)
speak	sagen	parler	parlare	hablar	mówić	говорить
(311)	(128)	(254)	(290)	(305)	(303)	(228)

4.5 Word-adjacency networks in various languages

Figure 4.12 shows the log-log plots of degree distributions and edge weight distributions of word-adjacency networks constructed from corpora consisting of books listed in Appendix B.1. The form of degree distributions indicates that the networks can be considered approximately scale-free, with degrees described by power laws with exponents of the survival function slightly above 1. This result is to a degree an expected one: Zipf's law ensures that word frequency distributions are described by a power law; word frequencies are approximately equal to node strengths, and since degrees are strongly correlated with strengths, one can anticipate that degree distributions in word-adjacency networks are significantly influenced by word frequency distributions. Edge weights in a word-adjacency network correspond to frequencies of 2-grams (pairs of words) in the underlying text. It can be seen that edge weight distributions in the studied networks can be approximated by power laws; this can be associated with the fact that frequencies of certain linguistic constructs larger than words also seem to be conforming to power-law distributions [326, 327].

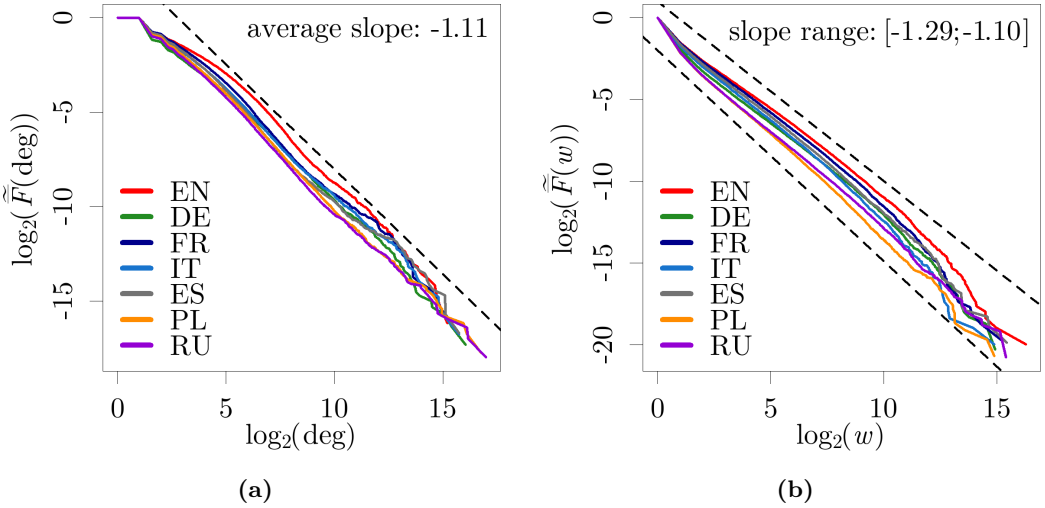


Figure 4.12. Log-log plots of empirical survival functions $\tilde{\tilde{F}}$ representing node degree distributions (a) and edge weight distributions (b) of word-adjacency networks constructed from corpora consisting of books listed in Appendix B.1 (each corpus is constructed from books in one language). The slope of the black dashed line in (a) is equal to the average of slopes of the lines fitted to each node degree distribution (-1.11). The slopes of black dashed lines in (b) represent the minimal and the maximal slope of the lines fitted to edge weight distributions (-1.29 and -1.10, respectively).

It should be noted, however, that the effect is not as evident as in case of individual words, described by Zipf's law [97, 328, 329].

While a number of properties of word-adjacency networks can be considered general and possibly universal, some traits seem to be specific to particular languages, at least to some degree. Figure 4.14 shows how the values of selected (normalized) global network characteristics are distributed in texts in different languages; the texts used are the books from the dataset specified in Appendix B.1. The differences between the distributions of the some of the considered characteristics in different languages are evident. This results in a tendency of texts in the same language to group together in the space of quantities describing network structure. One of the ways of detecting such clustering is using a hierarchical clustering algorithm [330]. The algorithm applied here can be described as follows: given a set of m points in some space, hierarchical clustering aims to link together the points that are close to each other, according to a certain metric (a function specifying the distance between points). It starts from creating m clusters and assigning one point to each cluster. Then the clusters that are closest to each other are merged together, and this is repeated until there is only one cluster. The distance between clusters can be defined in multiple ways; in the approach utilized here, the distance between two clusters is the greatest Euclidean distance between pairs of elements of these clusters (each element in a pair belongs to different cluster). The result of clustering can be presented in the form of a dendrogram - a tree-like diagram representing the consecutive merges along with the height at which they take place (the height of a merge is the distance between the merged clusters). Fig. 4.15 shows a dendrogram of the clustering of the books from Appendix B.1 in the space of the characteristics presented in Fig. 4.14. It can be seen that such a clustering tends to link texts with other texts in the same language. However, the separation between texts in different languages is far from perfect.

The differences between languages can also be illustrated with the use of another method of data analysis - the so-called Linear Discriminant Analysis (LDA) [331]. Given a set of points in some space, each belonging to some predefined class and labeled by that class, LDA sequentially finds vectors (orthogonal to each other) such

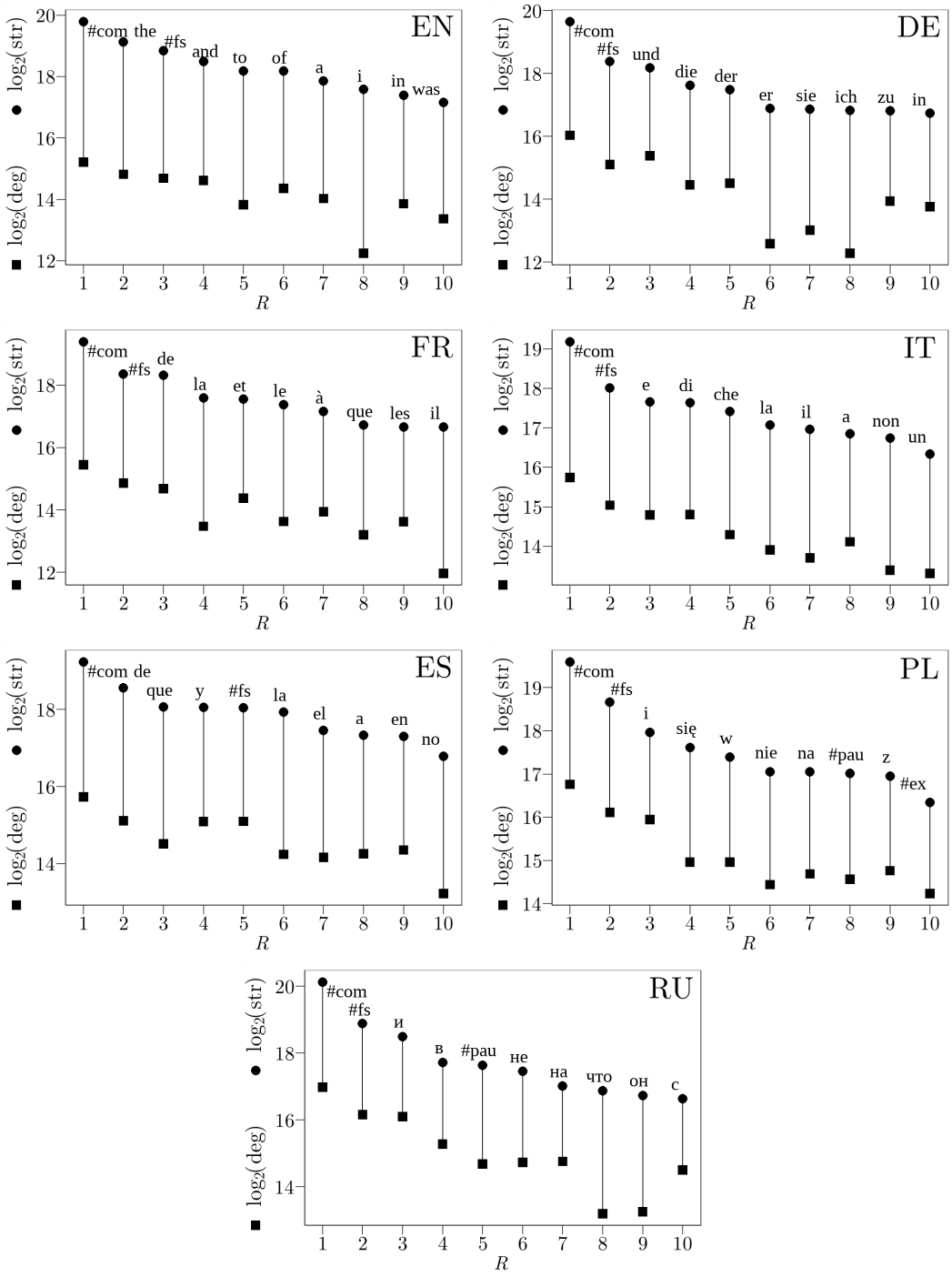
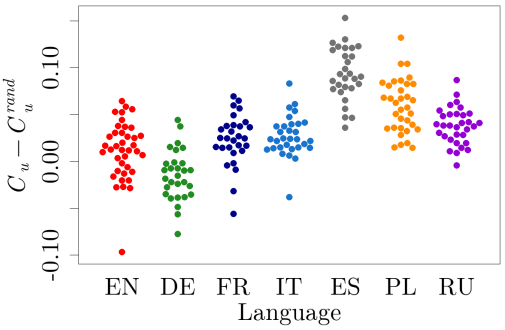
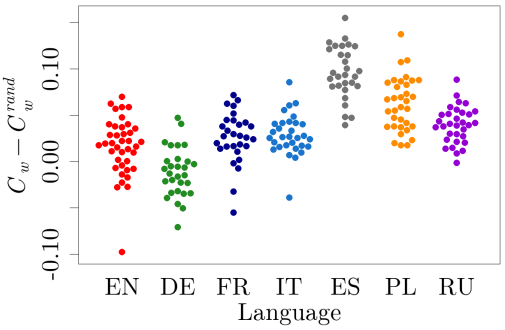


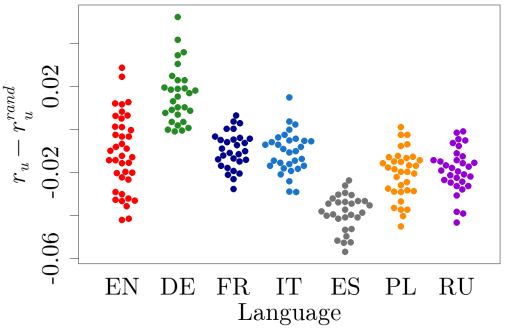
Figure 4.13. The highest node strengths and degrees in word-adjacency networks constructed from corpora consisting of books listed in Appendix B.1. Each of the figures pertains to one language (one corpus) - English (EN), German (DE), French (FR), Italian (IT), Spanish (ES), Polish (PL) or Russian (RU). The hoizontal axis specifies consecutive ranks R of words (their positions on the list of most frequent words), while vertical axis is used to specify both the degree (squares) and the strength (dots) of the nodes representing those nodes in the network; both quantities are under logarithm. It can be observed that although degree and strength are strongly related, their values might be significantly different from each other. Punctuation marks are treated as words and included in the analysis.



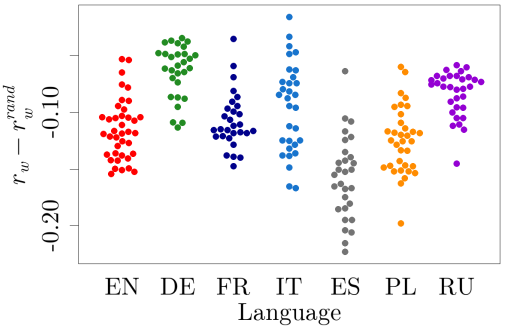
(a) Clustering coefficient, unweighted



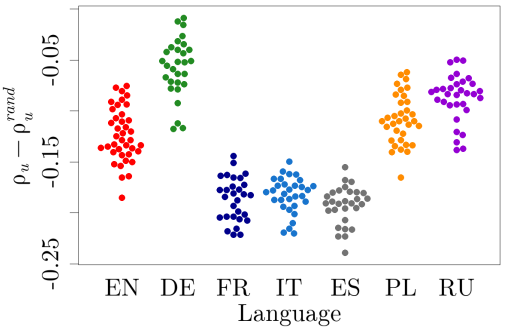
(b) Clustering coefficient, weighted



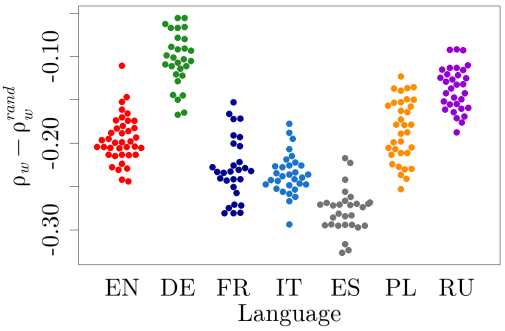
(c) Assortativity coefficient, unweighted



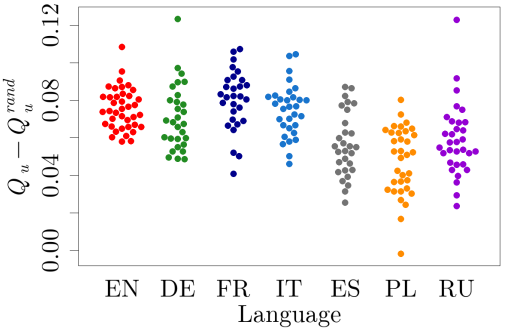
(d) Assortativity coefficient, weighted



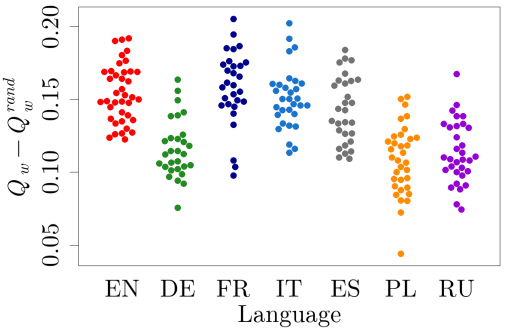
(e) Rank assortativity coefficient, unweighted



(f) Rank assortativity coefficient, weighted



(g) Modularity, unweighted



(h) Modularity, weighted

Figure 4.14. Global characteristics of word-adjacency networks constructed from texts in different languages. Each figure pertains to a single (normalized) characteristic; each dot in the figure represents one text from the dataset specified in Appendix B.1. It can be seen that the distributions of individual characteristics can be substantially different for different languages, but the presence and the strength of that effect varies among languages.

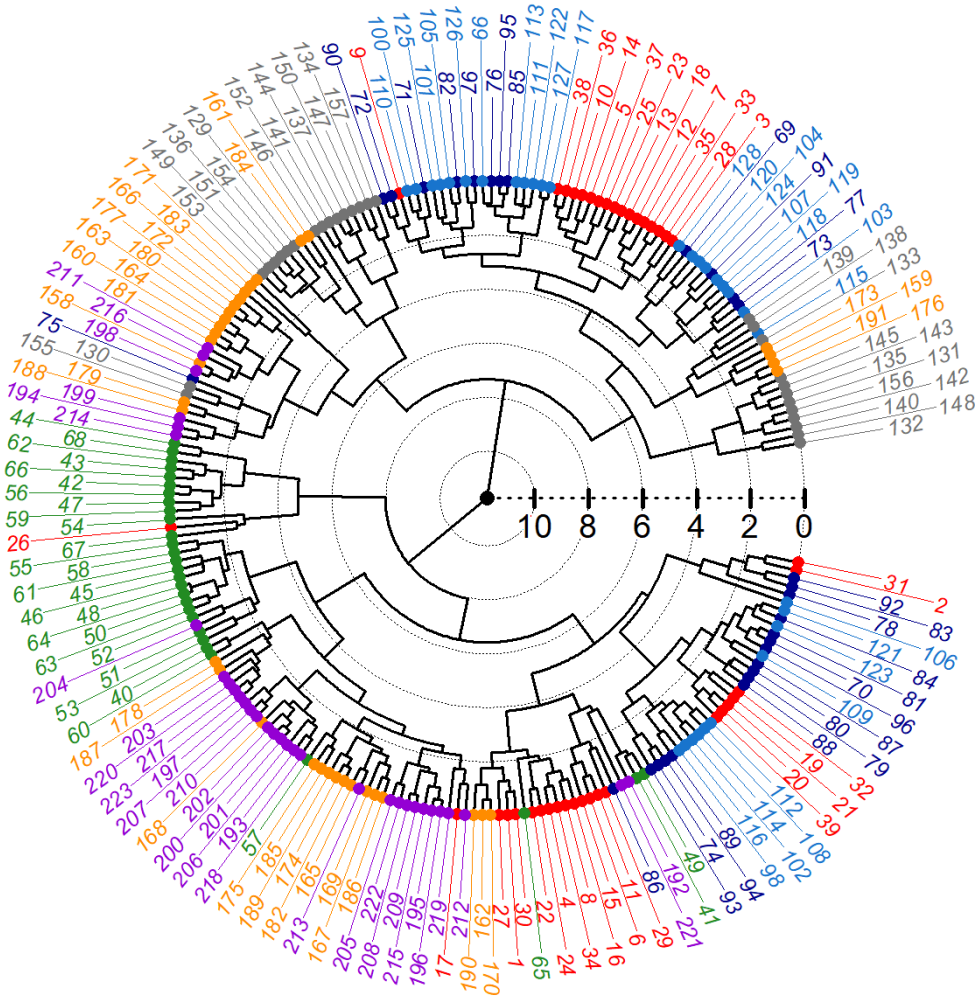


Figure 4.15. The dendrogram of hierarchical clustering of texts listed in Appendix B.1, in the attribute space consisting of (normalized) global characteristics of word-adjacency networks, presented in Fig. 4.14. The scale on dendrogram's radius gives the height of merges (the distance between merged clusters). Numeric labels correspond to the numbers assigned to books in the studied dataset; colors correspond to languages: red - English, green - German, dark blue - French, light blue - Italian, gray - Spanish, orange - Polish, purple - Russian. The presence of clusters consisting of texts in the same language indicates the presence of word-adjacency network's traits specific to individual languages.

that the projection of the data points on the subspace spanned by those vectors results in the maximum possible separation between classes. The first vector maximizes the class separation, and each subsequent vector maximizes the separation under the condition that it is orthogonal to all the preceding vectors. Since the ability of each vector to discriminate between classes is weaker than for the preceding vectors, projecting the data onto the subspace spanned by the first few such vectors can be sufficient to detect the patterns of variability between classes. Hence, LDA is often used as a dimensionality reduction technique. Therefore it can be treated as a method of finding a linear subspace in which the overlap between the clouds of points belonging to different classes is minimal. Figure 4.16 presents the projection of the dataset onto a 2-dimensional space spanned by vectors obtained by performing LDA on the books from Appendix B.1, in the space of four network characteristics: $C_u - C_u^{rand}$, $C_w - C_w^{rand}$, $\rho_u - \rho_u^{rand}$, $\rho_w - \rho_w^{rand}$, being the unweighted and weighted normalized clustering coefficients and the unweighted and weighted normalized rank assortativity coefficients, respectively. The results indicate the presence of clusters

of points representing texts in the same language, but there is still some overlap; among the studied languages, distinguishing between French and Italian and between Polish and Russian seems particularly difficult in terms of the network parameters considered. It can be concluded that using word-adjacency network representation allows to investigate certain statistical properties of texts, expressed by quantities characterizing the structure of such networks, and to describe quantitatively how languages differ in terms of those properties. In that context, it is worth noting that other types of linguistic networks, for example networks based on syntactic relationships between words, can also have the property of displaying different patterns of organization for different languages [332].

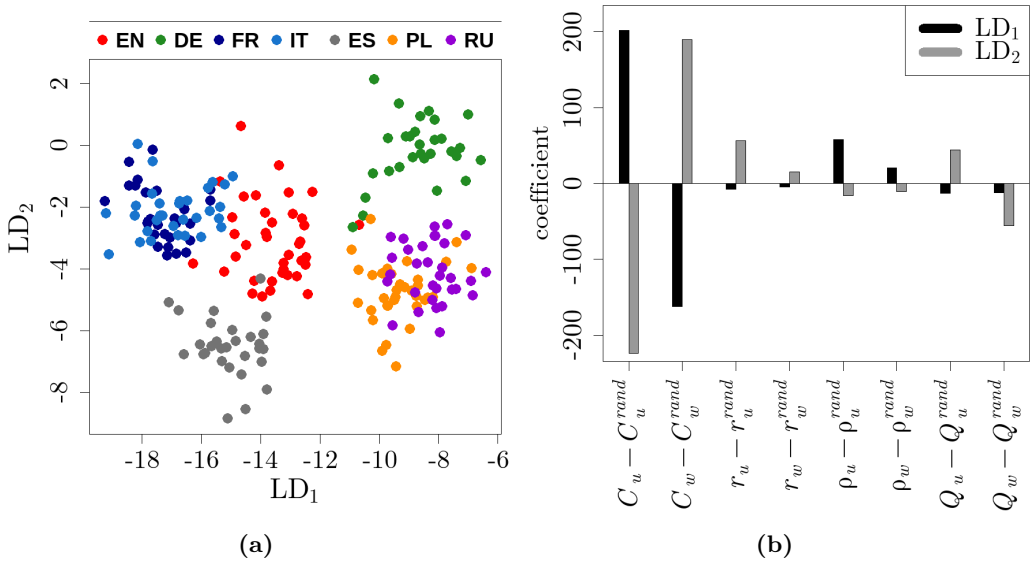


Figure 4.16. Texts in different languages in a subspace of the space constructed from word-adjacency networks' global characteristics. The original space consists of the (normalized) characteristics presented in Fig. 4.14. The 2-dimensional subspace presented in (a) is spanned by the first two linear discriminants LD_1, LD_2 - vectors computed with the use of LDA, determining the directions along which the texts are best separated with respect to language. Each dot in (a) represents one text from the dataset specified in Appendix B.1; colors represent languages. LD_1 and LD_2 are linear combinations of the basis vectors (each basis vector represents one of the studied network characteristics); the coefficients of those linear combinations are shown in (b). The absolute values of the coefficients can be treated as quantities measuring the extent to which a particular characteristic allows to separate texts in different languages; in that regard, clustering coefficient (both in its weighted and unweighted form) seems to be most significant.

4.6 Word-adjacency networks and text authorship

Since word-adjacency networks, apart from having a number of properties being common for various corpora, exhibit a tendency to distinguish between texts in different languages, a question arises whether network representation allows to classify texts by attributes other than language. An example of what can be studied in that context is writer's individual style of writing [333]. Stylometry (the analysis of writing style) relies on the fact that the specific way in which a particular author uses language is, at least to some degree, reflected by certain statistical properties of texts written by that author. It can be anticipated that the structure and organization of word-adjacency networks are also influenced by individual writing style. A trace

of such an effect can be seen in Fig. 4.17, where selected network characteristics of English and Polish 5000-word text samples drawn randomly from books written by different authors are shown. It can be observed that although derived from different books, the networks corresponding to texts of the same author tend to be similar in terms of the computed characteristics. The number of authors considered in each of the plots in Fig. 4.17 is only three; with a larger number of authors, the overlap between regions occupied by points representing text samples of different authors overlap to a larger degree and the differences between authors are much less evident.

A more detailed analysis of the utility of network representation in recognizing the authorship of texts is presented below. The used dataset consists of English and Polish books listed in Appendix B.3; in each of the languages there are 8 different authors and 6 books of each of the authors, giving 48 books per language and 96 books in total. Since it is the authorship, not the properties of a particular language, that is of interest, the analysis is performed separately for English and for Polish books.

Figure 4.18 shows the dendrograms of hierarchical clustering performed in space of four normalized, global, unweighted characteristics of networks. The studied characteristics are: average shortest path length, clustering coefficient, assortativity coefficient and modularity. The characteristics are normalized by dividing their values by the values obtained in network constructed from a text with words shuffled randomly, so the quantities constituting the clustering space are ℓ_u/ℓ_u^{rand}, C_u/C_u^{rand}, r_u/r_u^{rand}, Q_u/Q_u^{rand} where ℓ_u, C_u, r_u, Q_u denote, respectively: average shortest path length, clustering coefficient, assortativity coefficient and modularity of the original network, and ℓ_u^{rand}, C_u^{rand}, r_u^{rand}, Q_u^{rand} are the same characteristics computed for a randomized network. From the results presented in Fig. 4.18 one can conclude that hierarchical clustering is not sufficient to reliably distinguish between the authors present in the dataset; however, one can observe a number of small clusters (consisting of 2-3 books) belonging to the same author.

To further examine the correspondence between the authorship and the parameters of word-adjacency networks, one can use a method of supervised machine learning. The reasoning behind such an approach can be explained as follows. The sharper the differences between the structure of networks representing texts of different authors, the easier it should be to train an algorithm of statistical classification to recognize authorship in the space of network characteristics. So the accuracy of the classifier can be considered a measure of how well texts of different authors are separated in network parameters' space. The classifier used here is an ensemble of decision trees; it is based on the so-called *bagging* (*bootstrap aggregating*) of decision trees [334]. A short characterization of decision tree ensembles is given in Appendix A. The set of texts in each of the studied languages (English and Polish), represented by a set of points in the space of network characteristics, is randomly divided into two disjoint sets: the training set and the test set. These sets are constructed in such a way that each author is equally represented; more precisely, from six books of each author four are randomly selected to the training set, and the remaining two are assigned to the test set. Then an ensemble of 100 trees is trained on the training set to classify books with respect to the author - each book constitutes one observation whose attributes are the appropriate network parameters. Then the classifier classifies observations in the training set. The partition of the data into training and test sets and training the decision tree ensemble is repeated 10000 times. The average accuracy of the classifier (the fraction of observations with correctly recognized authorship) in the test set is treated as a measure of how distinct the structures of word-adjacency networks constructed from texts of different authors are. if the authors were not distinguishable at all, then the classification

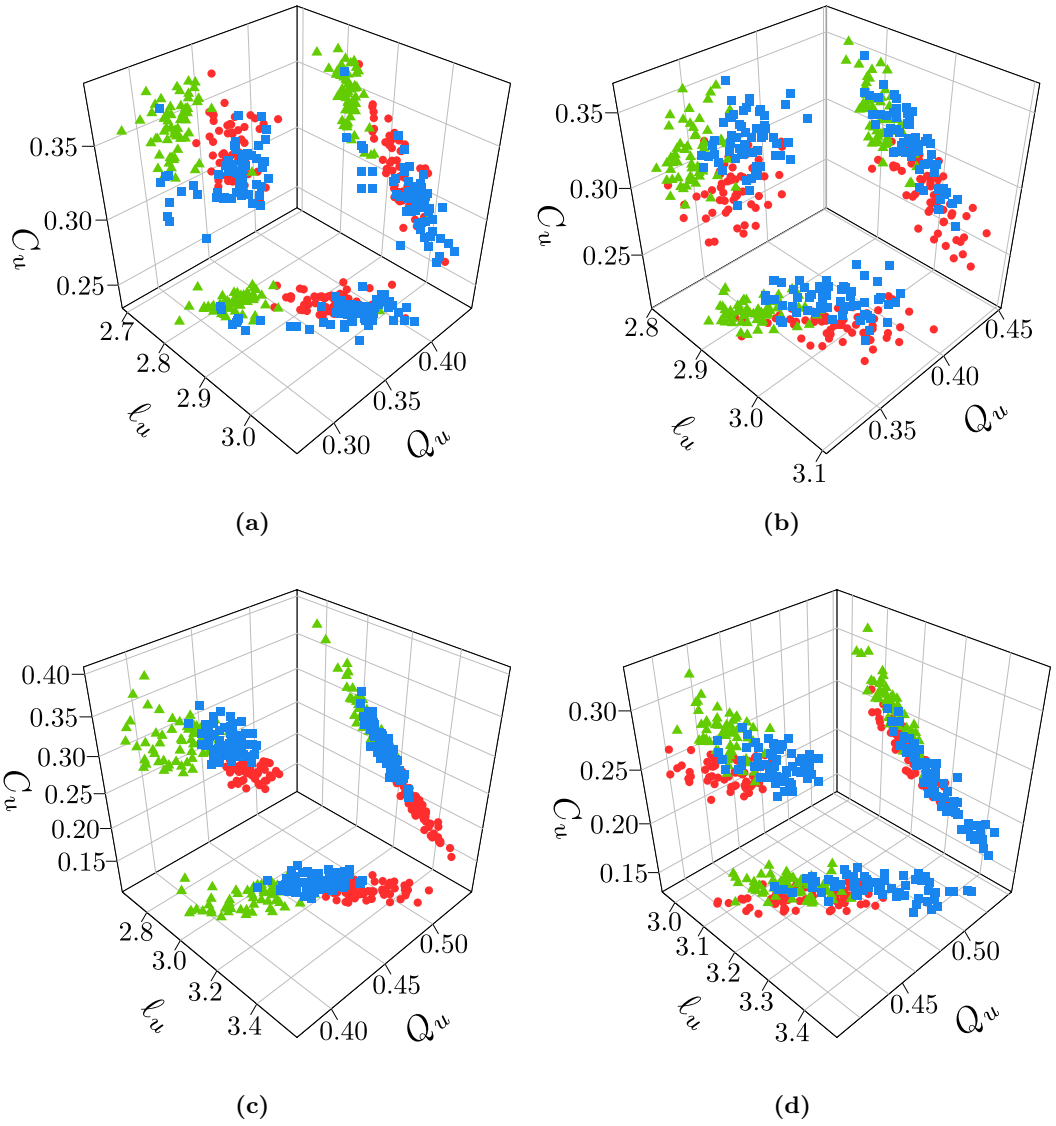


Figure 4.17. The projections of the triplets of word-adjacency network characteristics (ℓ_u, Q_u, C_u) onto planes (ℓ_u, Q_u), (ℓ_u, C_u), (Q_u, C_u), for texts in English (a, b) and Polish (c, d). Each triplet of characteristics pertains to one chunk of text of length 5000 words. Texts samples were randomly chosen from all of the studied works of considered authors (contained in the dataset specified in Appendix B.3). Different markers denote different authors - red dots, green triangles and blue squares denote respectively: Charles Dickens, Daniel Defoe and Mark Twain in (a), George Eliot, Jane Austen, Joseph Conrad in (b), Władysław Reymont, Janusz Korczak, Jan Lam in (c) and Henryk Sienkiewicz, Józef Ignacy Kraszewski, Stefan Żeromski in (d). It can be seen that points representing texts of different authors tend to occupy different regions of space. The characteristics are not normalized, as all the samples have the same length.

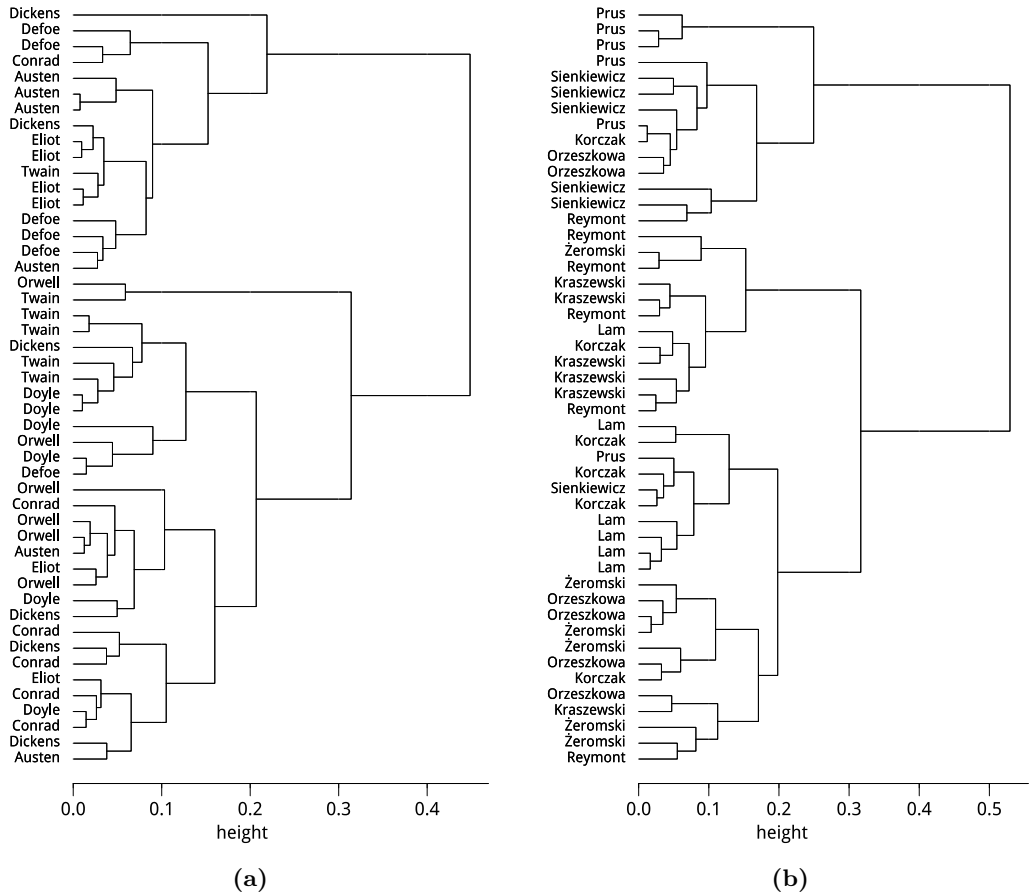


Figure 4.18. The dendrograms of the hierarchical clustering of (a) English and (b) Polish books from the dataset specified in Appendix B.3, in the space of (normalized) **unweighted global characteristics of word-adjacency networks**. Each text is labeled by the surname of its author.

would not be significantly different from a random choice, which has the expected accuracy of $1/n$, where n is the number of authors.

The results of classification in the same space that the one considered in Fig. 4.18, that is, a 4-dimensional space constructed from ℓ_u/ℓ_u^{rand}, C_u/C_u^{rand}, r_u/r_u^{rand}, and Q_u/Q_u^{rand}, are presented in Table 4.2. The table is organized as follows. In each part (one part corresponds to one language), the number in the i-th row and the j-th column is the probability of classifying a text of the i-th author as a text of the j-th author, obtained by counting such classifications in the test set and dividing the number of counts by the number of performed repetitions of the test set selections (10000). The probabilities of correct classifications reside on the diagonal. The sum of values in each row is equal to 1, as it is the probability of assigning a text to any of the authors. The average overall probabilities of correct classification in the test set are: 35% with standard deviation of 10% for English, and 41% with standard deviation of 10% for Polish. This indicates that the some information about the authorship of texts is indeed encoded in the parameters of the networks, as the results are clearly better than a random classification; however, there is certainly room for improvement.

One of possible improvements is taking into account the fact that networks are weighted, and including weighted characteristics into the analysis. Unweighted characteristics, typically being conceptually simpler and easier to compute, can be combined with their weighted counterparts form a space in which the algorithms of hierarchical clustering or statistical classification can be applied. Table 4.3 presents the results of classification performed by decision tree ensembles in an 8-dimensional

Table 4.2. The results of the classification of (a) English (b) Polish books (from the dataset specified in Appendix B.3) with respect to the authorship, in the space of (normalized) **unweighted global characteristics of word-adjacency networks**. A number in the i-th row and j-th column is the probability of classifying a text of i-th author as a text of j-th author.

(a) The classification of English books. The authors are denoted by the two first letters of their surnames: Au - Austen, Co - Conrad, De - Defoe, Di - Dickens, Do - Doyle, El - Eliot, Or - Orwell, Tw - Twain.

	Au	Co	De	Di	Do	El	Or	Tw
Au	**.33**	.22	.17	.06	.01	.11	.10	.00
Co	.08	**.34**	.01	.21	.02	.11	.14	.09
De	.21	.00	**.54**	.00	.09	.04	.12	.00
Di	.18	.21	.00	**.05**	.23	.18	.08	.07
Do	.00	.12	.06	.09	**.28**	.03	.24	.18
El	.10	.22	.16	.05	.04	**.39**	.04	.00
Or	.10	.04	.09	.07	.23	.12	**.18**	.17
Tw	.00	.00	.00	.03	.12	.16	.03	**.66**

(b) The classification of Polish books. The authors are denoted by the two first letters of their surnames: Ko - Korczak, Kr - Kraszewski, La - Lam, Or - Orzeszkowa, Pr - Prus, Re - Reymont, Si - Sienkiewicz, Że - Żeromski.

	Ko	Kr	La	Or	Pr	Re	Si	Że
Ko	**.18**	.23	.14	.18	.04	.00	.09	.14
Kr	.16	**.57**	.00	.02	.00	.18	.03	.04
La	.21	.06	**.54**	.02	.05	.00	.08	.04
Or	.14	.00	.02	**.26**	.11	.00	.22	.25
Pr	.02	.00	.09	.12	**.65**	.00	.10	.02
Re	.05	.21	.00	.01	.00	**.44**	.07	.22
Si	.04	.00	.18	.21	.09	.00	**.46**	.02
Że	.19	.10	.01	.32	.00	.17	.00	**.21**

space of normalized global network characteristics - average shortest path length, clustering coefficient, assortativity coefficient and modularity, in unweighted and weighted variant. The average classification accuracy obtained in the test set is 42% with standard deviation of 11% for texts in English, and 44% with standard deviation of 11% for Polish texts. The increase of classification accuracy is rather negligible, indicating that the information regarding the individual style of particular authors encoded in networks' weighted characteristics overlaps to a large degree with the information carried by the characteristics in unweighted variant. in terms of the In term of the ability to distinguish between authors, the results of hierarchical clustering utilizing both unweighted and weighted characteristics are similar to the results obtained with unweighted characteristics only; therefore the dendrograms are not presented here. Also, using only the weighted variants of the studied quantities (without the unweighted ones) leads to the distinguishability between authors at a level similar to the one obtained in the analysis of unweighted characteristics only, both in clustering and classification.

Table 4.3. The results of the classification of (a) English (b) Polish books (from the dataset specified in Appendix B.3) with respect to the authorship, in the space of (normalized) **unweighted and weighted global characteristics of word-adjacency networks**. A number in the i-th row and j-th column is the probability of classifying a text of i-th author as a text of j-th author.

(a) The classification of English books. The authors are denoted by the two first letters of their surnames: Au - Austen, Co - Conrad, De - Defoe, Di - Dickens, Do - Doyle, El - Eliot, Or - Orwell, Tw - Twain.

	Au	Co	De	Di	Do	El	Or	Tw
Au	**.20**	.17	.18	.12	.00	.22	.11	.00
Co	.12	**.62**	.00	.06	.02	.07	.03	.08
De	.11	.00	**.52**	.11	.09	.14	.01	.02
Di	.16	.03	.03	**.21**	.22	.18	.09	.08
Do	.02	.10	.00	.22	**.33**	.01	.14	.18
El	.17	.16	.13	.04	.00	**.46**	.00	.04
Or	.06	.03	.05	.04	.05	.01	**.55**	.21
Tw	.00	.01	.03	.02	.10	.13	.22	**.49**

(b) The classification of Polish books. The authors are denoted by the two first letters of their surnames: Ko - Korczak, Kr - Kraszewski, La - Lam, Or - Orzeszkowa, Pr - Prus, Re - Reymont, Si - Sienkiewicz, Że - Żeromski.

	Ko	Kr	La	Or	Pr	Re	Si	Że
Ko	**.22**	.22	.16	.09	.01	.01	.15	.14
Kr	.21	**.35**	.03	.07	.01	.29	.00	.04
La	.06	.04	**.55**	.05	.12	.00	.03	.15
Or	.16	.04	.02	**.34**	.13	.00	.12	.19
Pr	.01	.00	.08	.18	**.63**	.00	.10	.00
Re	.01	.12	.00	.00	.00	**.50**	.18	.19
Si	.10	.00	.07	.06	.15	.01	**.61**	.00
Że	.22	.03	.08	.08	.00	.28	.00	**.31**

A slightly different approach to identifying how various texts differ in their network representations is based on studying local characteristics - quantities describing individual nodes in a word-adjacency network. In each text, nodes corresponding to n most frequent words in the studied language are identified, their characteristics are computed and supplied as components of the space in which the classification of clustering takes place. Punctuation marks are included into the analysis, and they are treated in the same way as words; this means that they are present in the frequency ranking from which n most frequent words are chosen. Identifying the most frequent words in language requires a large enough corpus; here these words are extracted from the corpus consisting of all studied texts. The characteristics investigated in the analysis are: vertex degree, local clustering coefficient, and average shortest path length, all in both unweighted and weighted variants. The characteristics are normalized by dividing their value by the average value in a randomized network. Node strength (the weighted counterpart of node degree) is an exception here - since it is roughly equal to the doubled frequency of the corresponding word in the underlying text, and word frequencies remain unchanged during randomization, dividing node strength by its randomized counterpart always gives a value equal to or very close to 1. Therefore the normalization of node strength is performed in another way: the normalized strength of a node v in a network is the strength of v divided by the sum of strengths of all nodes in that network, $\mathrm{str}(v)/\sum_{u\in V}\mathrm{str}(u)$. This is equivalent to word's relative frequency, that is the number of its occurrences divided by the length of the considered text.

A dendrogram of the clustering in 12-dimensional space of the weighted clustering coefficients of $n = 12$ most frequent words is presented in Fig. 4.19. Results of classification with decision tree ensembles performed in the same space are collected in Table 4.4. The obtained overall classification accuracy is 90% with standard deviation of 8% for English texts, and 86% with standard deviation of 8% for the Polish ones. The choice of the network parameter (clustering coefficient) and the number of words to study ($n = 12$) is a consequence of the results presented in Fig. 4.20, which pertain to the effectiveness of particular network characteristics in distinguishing between texts of different authors. From what is presented in Fig. 4.20 one can conclude that weighted clustering coefficient gives the best classification results in English and one of the best in Polish for a wide range of the most frequent words studied. In both languages, it is sufficient to analyse the 11-12 most frequent words to obtain the accuracy of 85-90%; further increase in the number of words does not improve the classifier's performance.

The results indicate that the quantities describing individual words in a word-adjacency network are much more effective in capturing the writing styles of individual authors compared to the characteristics pertaining to the whole network. The number of attributes associated with each text can be significantly larger when analyzing local characteristics instead of the global ones, as the number of words taken into consideration can be chosen arbitrarily. However, even for comparable dimensions of the classification space, local characteristics (like nodes' clustering coefficients) provide classification accuracy significantly higher than the accuracy obtained with global characteristics.

From Fig. 4.20 it can be seen that a characteristic which works comparably well to normalized weighted clustering coefficient in recognizing the authorship of texts is normalized node strength, which, due to reasons mentioned above, expresses the relative frequency of a word in the text. The analysis of word frequencies is a basic yet effective method of authorship attribution, widely applied and discussed in the relevant literature [335–339]. The fact that word frequency analysis is often able to provide a decent result leads to a question whether it is viable to introduce

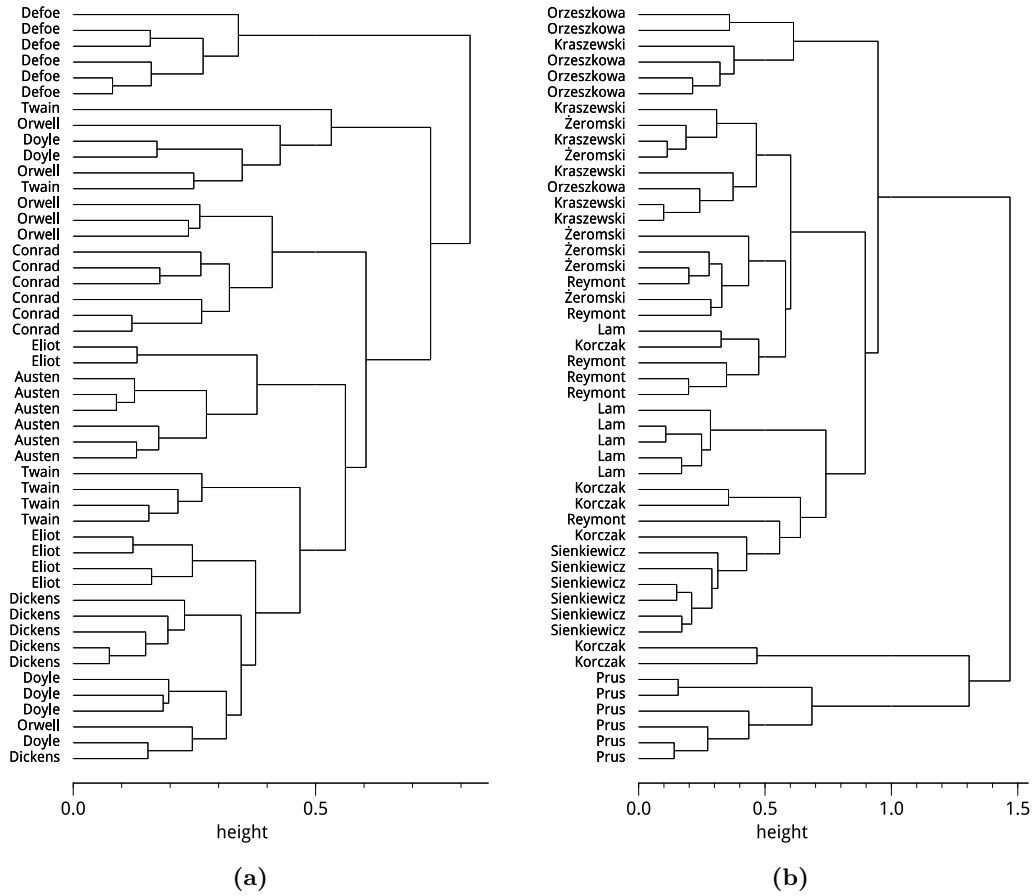


Figure 4.19. The dendrograms of the hierarchical clustering of (a) English and (b) Polish books from the dataset specified in Appendix B.3, in the space of the **weighted clustering coefficients of word-adjacency networks' nodes, corresponding to 12 most frequent words**. Each text is labeled by the surname of its author.

Table 4.4. The results of the classification of (a) English (b) Polish books (from the dataset specified in Appendix B.3) with respect to the authorship, in the space of the (normalized) **weighted clustering coefficient of word-adjacency networks' nodes, corresponding to 12 most frequent words**. A number in the i-th row and j-th column is the probability of classifying a text of i-th author as a text of j-th author.

(a) The classification of English books. The authors are denoted by the two first letters of their surnames: Au - Austen, Co - Conrad, De - Defoe, Di - Dickens, Do - Doyle, El - Eliot, Or - Orwell, Tw - Twain.

	Au	Co	De	Di	Do	El	Or	Tw
Au	**.96**	.00	.00	.01	.01	.02	.00	.00
Co	.00	**.92**	.00	.00	.00	.04	.04	.00
De	.00	.00	**1.0**	.00	.00	.00	.00	.00
Di	.01	.00	.00	**.88**	.04	.00	.00	.07
Do	.01	.00	.00	.05	**.93**	.01	.00	.00
El	.02	.02	.00	.03	.01	**.87**	.02	.03
Or	.00	.13	.00	.01	.01	.07	**.78**	.00
Tw	.00	.01	.00	.00	.01	.02	.07	**.89**

(b) The classification of Polish books. The authors are denoted by the two first letters of their surnames: Ko - Korczak, Kr - Kraszewski, La - Lam, Or - Orzeszkowa, Pr - Prus, Re - Reymont, Si - Sienkiewicz, Że - Żeromski.

	Ko	Kr	La	Or	Pr	Re	Si	Że
Ko	**.73**	.00	.00	.00	.00	.06	.03	.18
Kr	.00	**.80**	.00	.12	.00	.00	.05	.03
La	.00	.00	**1.0**	.00	.00	.00	.00	.00
Or	.00	.15	.00	**.83**	.00	.00	.01	.01
Pr	.00	.00	.00	.00	**1.0**	.00	.00	.00
Re	.04	.00	.00	.00	.00	**.77**	.02	.17
Si	.00	.05	.00	.00	.04	.01	**.90**	.00
Że	.01	.15	.00	.00	.00	.03	.02	**.79**

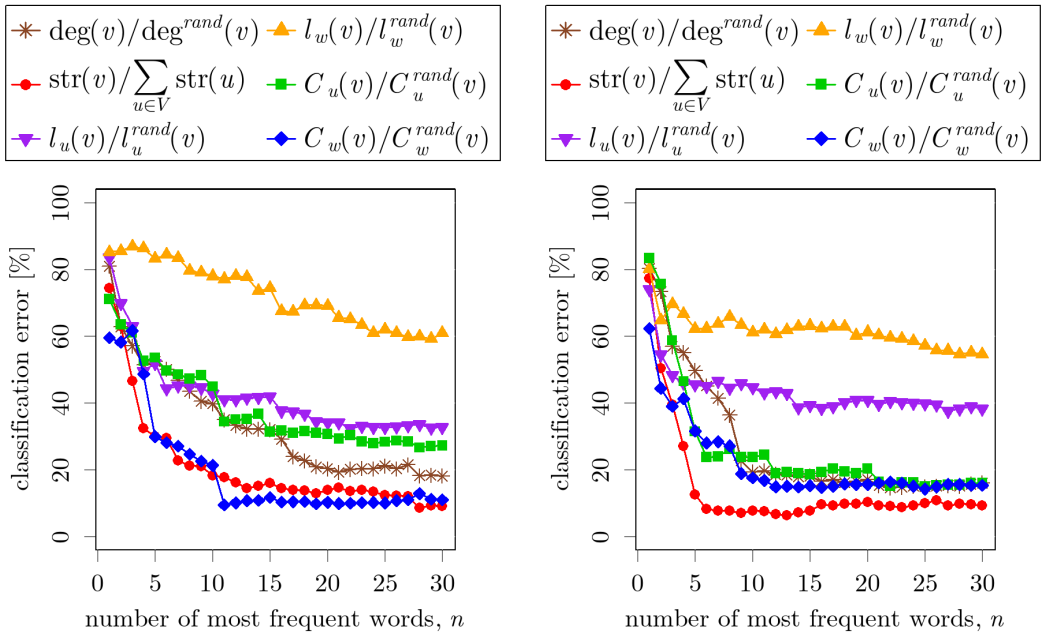


Figure 4.20. The classification of books from the dataset specified in Appendix B.3, in the feature spaces constructed from a single word-adjacency network parameter determined for a set of n words occurring most frequently in the whole book collection. Charts (a) and (b) present the average classification error as a function of n, for English and Polish books, respectively. Each point on a chart represents the average classification error in the test set, obtained in one experiment. One experiment consists of selecting the n most frequent words, computing one network parameter for each of these words in each text, and performing the cross-validation of classification in the so obtained n-dimensional space, 10000 times. All the considered network characteristics are normalized. The studied characteristics are: vertex degree and strength ($\deg(v)/\deg^{rand}(v)$, $\mathrm{str}(v)/\sum_{u\in V}\mathrm{str}(u)$), unweighted and weighted average shortest path length ($\ell_u(v)/\ell_u^{rand}(v)$, $\ell_w(v)/\ell_w^{rand}(v)$), unweighted and weighted clustering coefficient ($C_u(v)/C_u^{rand}(v)$, $C_w(v)/C_w^{rand}(v)$). The legend specifying the symbols used to represent the characteristics is given above each plot.

the network formalism when a simpler, frequency-based method allows to recognize authorship with satisfying accuracy. An answer to such a question can be given based on the results presented in Fig. 4.21: it turns out that combining frequencies with purely network-based characteristics (clustering coefficients) can constitute a beneficial approach to author identification; classification using both frequencies and clustering coefficients leads to accuracy better than the accuracy obtained when studying either frequencies or clustering coefficients alone. This indicates that the information encoded in word frequencies is to a certain degree distinct from the information contained in the set of word's clustering coefficients. A practical application of the presented results can therefore be based on combining the already known features used in text classification with the ones pertaining to word-adjacency network structure. It can be anticipated that classification aimed at identifying traits other than authorship can also benefit from incorporating network characteristics into the analysis.

Recognizing authorship with the use of word frequencies (or equivalently, node strengths) focuses on studying the tendency of individual authors to use particular words more often than others. Clustering coefficient of a node representing word in a word-adjacency network describes the structure of the neighborhood of that word in the network (in its normalized form, it describes how that structure is different from the one observed in a randomized network). So the classification based on clustering coefficients relies on investigating the way in which words are used more

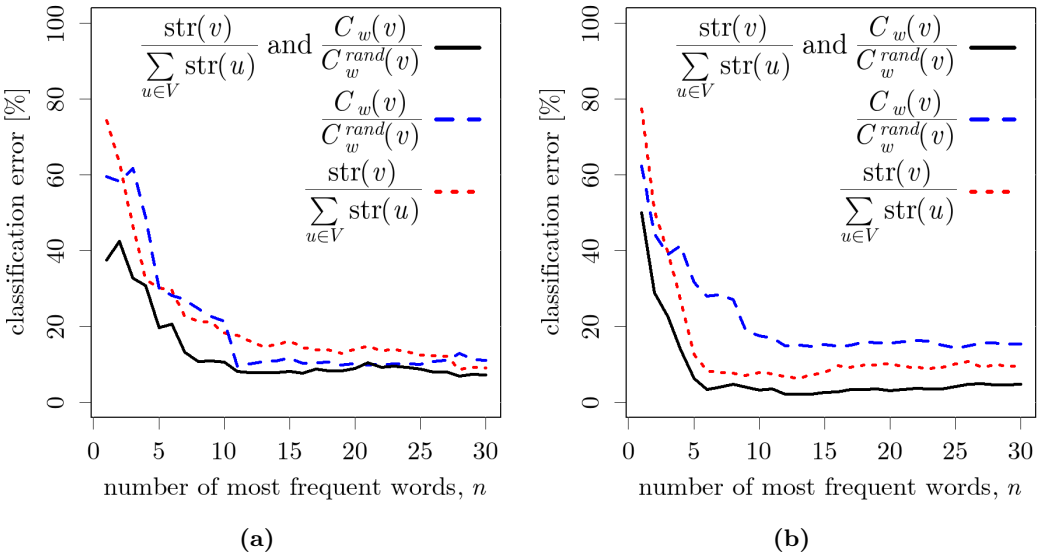


Figure 4.21. The classification of books from the dataset specified in Appendix B.3, in the feature spaces constructed from the sets of word-adjacency network parameters, determined for a set of n words occurring most frequently in the whole collection of books. Charts (a) and (b) present the average (obtained from 10000 repetitions of cross-validation) classification error in the test set, as a function of n, for English and Polish books, respectively. 3 sets of quantities describing words in texts were investigated, namely: (1) normalized vertex strength $\mathrm{str}(v)/\sum_{u\in V}\mathrm{str}(u)$, (2) normalized weighted clustering coefficient $C_w(v)/C_w^{rand}(v)$, (3) normalized vertex strength $\mathrm{str}(v)/\sum_{u\in V}\mathrm{str}(u)$ together with normalized weighted clustering coefficient $C_w(v)/C_w^{rand}(v)$.

than their frequencies. The fact that the best results are obtained when both types of information are included seems to be in agreement with a quite natural expectation that author's individual style is expressed by both the preference to use particular words or linguistic constructs and the specific way of using them.

Among the studied characteristics, average shortest path length (especially in its weighted variant) seems to be the weakest in authorship attribution (Fig. 4.20); this suggests that average shortest path lengths of nodes representing most frequent words behave in roughly the same way in texts of various authors. Indeed, when a text is sufficiently long, the construction procedure of word-adjacency network leads to the formation of a densely connected cluster of most frequent words, in which all nodes are connected by a relatively short path. The distance between an arbitrarily chosen node v_1 in the network and a node v_2 belonging to the cluster is similar for all choices of v_2. To describe that effect quantitatively, one can consider a subnetwork consisting of some arbitrary number of nodes representing most frequent words, compute the (global) average shortest path length in that subnetwork, ℓ^{sub}, and compare it with the global average shortest path length in the whole network, ℓ. Figure 4.22 presents the strip charts of the distributions of the quotient ℓ^{sub}/ℓ, for all the studied networks, with the subnetworks consisting of the 50 most frequent words. For the unweighted networks, the value of ℓ^{sub}/ℓ is usually around 0.4-0.5, and for weighted ones it is typically smaller than 0.1. The difference between the two is a consequence of the fact that while for unweighted path lengths the smallest possible distance between two distinct vertices is equal to 1, the distance taking edge weights into account can be made arbitrarily small by increasing those weights. Hence, for any v being one of 50 most frequent words, the weighted local average shortest path length $\ell_w(v)$ has roughly the same value. This effect is present also for networks constructed from randomized texts (Fig. 4.22), which confirms that it is related to how word-adjacency networks are constructed more than to specific properties of texts.

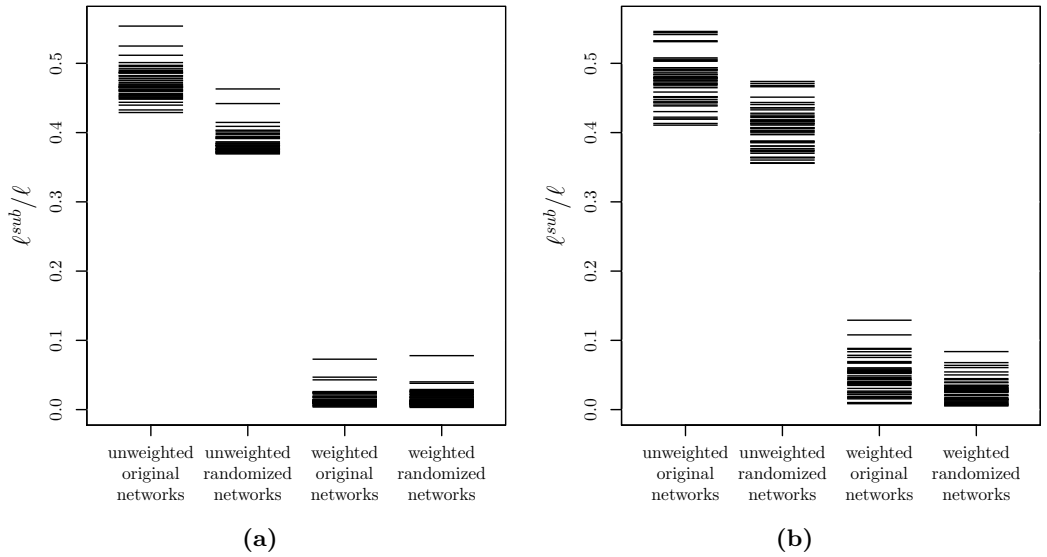
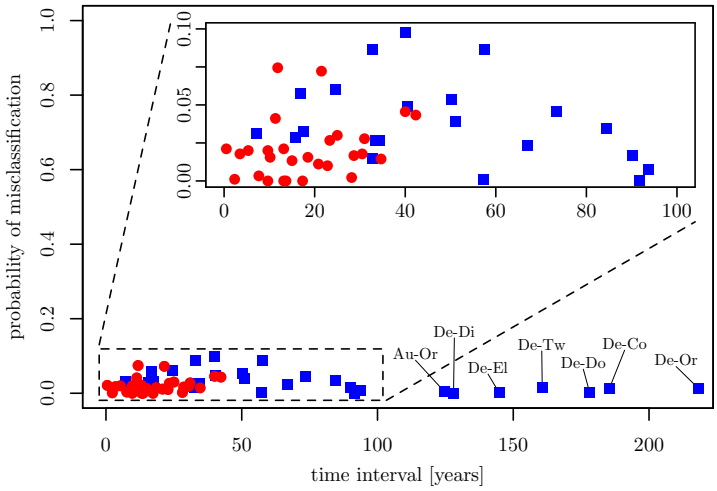


Figure 4.22. The distributions of the quantity ℓ^{sub}/ℓ, for word-adjacency networks constructed from (a) English and (b) Polish books from the dataset specified in Appendix B.3. ℓ^{sub} denotes the global average shortest path length in a subnetwork consisting of 50 most frequent words, and ℓ denotes the average shortest path length in the whole network. Both original and randomized networks are considered, in both weighted and unweighted version.

The publication dates of the studied books vary from 1719 to 1949. Since the evolution of language certainly has an influence on some of its statistical properties, it can be suspected that the fact that books come from different centuries might affect the results of authorship attribution. One can anticipate that literary works written around the same time are more likely to be confused than these separated by a long time interval. The approach to this issue adopted here is based on investigating how the distinguishability between authors varies with time interval separating their works. This is done as follows. For each author, a "time centroid" is determined as an arithmetic mean of publication dates of that author's works considered in the analysis. Then a distance between the centroids (in years) is assigned to each pair of authors and treated as a time interval separating them. Next, for each two authors a classification considering only their works is performed multiple times (in feature space consisting of frequencies and clustering coefficients of the 12 most frequent words; the set of texts of each author is split into training and test set with ratio 4:2). The error rates in such pairwise classifications (for each pair of authors) are then plotted versus time intervals to assess whether the authors separated by greater time intervals are easier to distinguish between or not. The results are presented in Fig. 4.23. It can be seen that in case of English, for the writers who are separated by more than 100 years the probability of misclassification is lower than 2%, while for others it exhibits more variability. It is worth noting however, that in all except one pair of authors separated by time intervals longer than 100 years one of the authors is Daniel Defoe, who lived and wrote much earlier than the rest of the studied writers. When only time intervals shorter than 100 years are considered (Figure 4.23, inset), no clear relationship between the classification accuracy and the length of the time interval separating authors can be observed. So for the analyzed set of texts the effects of language change over time seem to have little to no effect. However, this does not mean that language evolution does not have an effect on text classification in general - this might depend on the type of classification and especially on the studied timespan.



(a)

Figure 4.23. The influence of differences in publication dates on text authorship classification's accuracy, for the books listed in Appendix B.3. The scatterplot presents the error rates of pairwise authorship classifications and the time intervals separating the studied writers. Each marker denotes one pair of authors; squares pertain to English, and dots pertain to Polish language. All points with time interval greater than 100 years, except one, correspond to pairs in which one author is Daniel Defoe. These points are labeled with writers' names, abbreviated in the same manner as in Tables 4.2, 4.3, 4.4. For the rest of pairs, the scatterplot is presented in more detail in the inset.

The data presented in Fig 4.20 and Fig. 4.21 allows to observe some differences between classification of English and of Polish texts. It can be noticed, for example, that reaching the limit of the classification accuracy in Polish requires analyzing less words than for the books in English (it is most clearly visible in the case of classification based on word frequencies and on both frequencies and weighted clustering coefficients). This effect might be an example of consequences of structural differences between languages. In Polish, being an inflected language, the syntactic role of a word in a sentence is determined mainly by inflection. The syntax of English relies more on the word order and on utilizing function words (articles, auxiliary verbs, etc), which are among the most frequent words in the language. It can be anticipated that in a language whose grammar imposes stronger conditions on the usage and order of some of the most frequent words, there is less room for the diversity of usage patterns related to individual writing styles. In such a case, more words need to be included into the analysis to obtain a reliable classification accuracy.

An issue worth pointing out is the role that punctuation has in the classification. Quantifying the influence of including punctuation marks into the analysis can be done by removing punctuation marks from the set of studied words (in which punctuation marks are included), replacing them with subsequent words from the list of most frequent words - to keep the dimension of the attribute space unchanged, and comparing the accuracy of classification with and without punctuation marks. It turns out that the accuracy substantially decreases when punctuation is removed; For example, a classification in the space of the normalized weighted clustering coefficients $C_w(v)/C_w^{rand}(v)$ of $n = 12$ most frequent words without punctuation marks gives the average accuracy of 75% for English and 76% for Polish texts. These values being significantly lower than the ones obtained in classification including punctuation marks indicate that specific patterns of punctuation usage constitute a non-negligible contribution to the writing style and that taking punctuation into account on the same terms as words has a potential to considerably improve the effectiveness of stylometric analysis.

4.7 Semantic networks and word-association networks

As mentioned before, there are multiple aspects of language structure that can be described with the use of complex networks. The formalism discussed in the context of word-adjacency networks can be applied also to relationships other than word co-occurrences. An example of a class of networks used to represent the aspects of language organization different from the ones described by word-adjacency networks are the so-called semantic networks. A semantic network is a network in which nodes represent concepts and edges express semantic relationships between these concepts. Among the examples of semantic networks are networks representing the structure of certain linguistic databases like WordNet [340,341]. WordNet is a lexical database consisting of words grouped into collections of synonyms, called synsets, which express certain concepts. Synsets can be connected with each other with various semantic relations, like hyponymy and hypernymy (if Y is a subtype of X, then Y is a hyponym of X and X is a hypernym of Y; examples of hypernym-hyponym pairs are *plant* - *tree* or *astronomical object* - *star*), or meronymy and holonymy (if Y is a part of X, then Y is a meronym of X and X is a holonym of Y; *tree* - *leaf* and *building* - *window* are examples of holonym-meronym pairs). An example of a network representing hypernymy-holonymy relations in an excerpt from WordNet database is shown in Fig. 4.24.

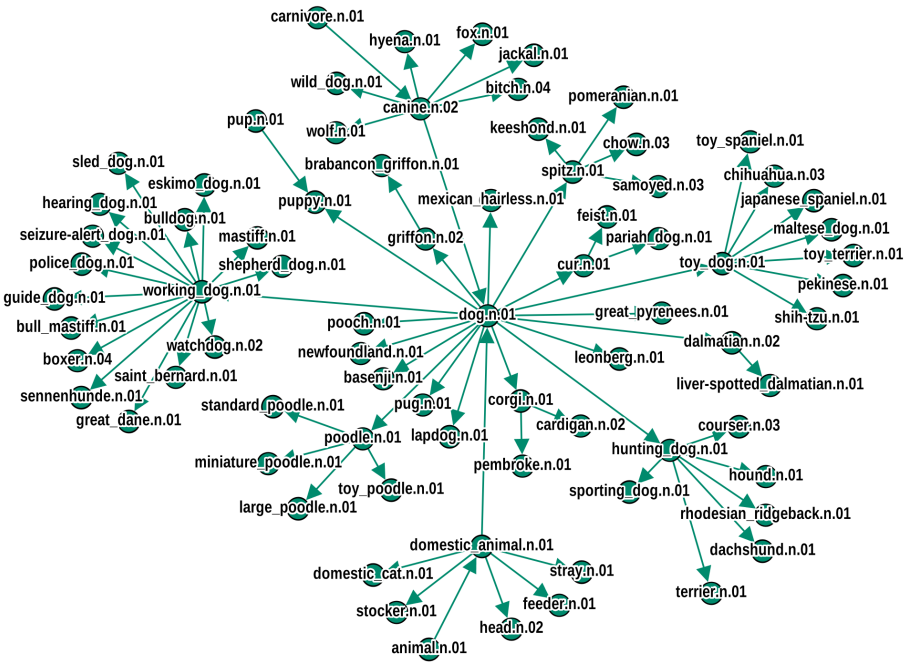


Figure 4.24. An excerpt from the network representing hypernym-hyponym relations between nouns in the WordNet database. In such a network, synsets are represented by nodes, and hypernym-hyponym pairs are connected by (unweighted) directed edges (from hypernyms to hyponyms). The presented piece of the original WordNet-based network consists of synsets that are at most 2 steps from the synset "*dog.n.01*" (the letter "n" denotes nouns; the numbers appended to synsets' names are used to distinguish between synsets which might be confused - for example, "bank" meaning a financial institution and "bank" meaning a sloping raised land). It is constructed from the original network of hypernym-hyponym releations by removing all the nodes whose distance (ignoring the direction of edges) from the synset "*dog.n.01*" is greater than 2.

Apart from having a number of applications in the field of automatic natural language processing [342,343], semantic networks are studied in psycholinguistics - a field of research focused on cognitive mechanisms responsible for representing and

processing language in human brain. For example WordNet, now serving as a lexical resource in a wide range of natural language processing solutions and language-related research, has been initially developed as a lexical database consistent with certain hypotheses regarding how semantic memory (the knowledge of words) is organized in human mind. Theories developed in 1960s and 1970s suggested that memory is organized in a hierarchical fashion, with concepts on deeper levels of hierarchy inheriting the properties assigned to relevant higher-level concepts [344]. Some aspects of such a view have been found to be oversimplified [345]; however, the idea of using network formalism to study the organization of words and concepts in human mind (often referred to as *mental lexicon*) is more general than the mentioned theories and remains highly influential in relevant research. A class of networks important in that context are word-association networks (or *associative networks*). A word-association network can be described as a network whose nodes represent words and edges correspond to associations between words. Edges often have weights, representing strengths of individual associations. An usual way of constructing a network representing word associations typical for a population of users of some particular language is conducting an experiment in which participants are presented words - one word at a time - and are asked to write down the first word that comes to their mind. Collecting the data from many participants and for many different words allows to build a network in which words become nodes and associations manifest themselves as connections between nodes, with weights proportional to the number of participants giving a particular response.

The significance of word associations has been investigated in a number of psycholinguistic experiments, involving tasks like word memorization or recognition (an example of word recognition task is deciding whether a given sequence of letters constitutes a word or not); in one of possible variants of such an experiment, the actual task - recognizing some word or recalling a previously memorized word - is preceded by showing some other word to a participant. It has been established that the performance in tasks of that kind depends on the strength of relationship that the word of interest and the preceding word have in a word-association network (this strength can be measured by the number of associates shared by the words) [346, 347].

The fact that the characteristics of word-association networks allow to make predictions regarding the performance in tasks involving word processing and usage [348,349] supports the claim that the structure of a word-association network can be in some contexts considered a rough approximation of the structure of the lexicon in mind [348, 350]. This allows to describe certain activities involving language processing in terms of network theory, for example, a task consisting of finding a word that matches semantically to a set of given words can be represented by an appropriate walk on a word-association network [348]. Therefore, studying the structure of word-association networks with the use of tools and methods applied to complex networks modeling various systems in nature [345, 349, 351] has a potential to give an insight into some aspects of how language is organized and processed in human brain.

Selected properties of word-association networks are presented here with the use of data coming from two datasets: *University of South Florida Free Association Norms* [352] (here abbreviated as *USFFA*) and *Edinburgh Associative Thesaurus* [353–355] (abbreviated as *EAT*); both of them are available at [356] in the form allowing for their easy transformation into networks. The data in USFFA and EAT datasets was collected in experiments involving a large group of people and conducted according to the scheme mentioned above: participants were asked to write down the first words coming to their minds in response to some presented word. Although there are some differences between USFFA and EAT regarding the

details of the data collection procedure, the general idea of the experiment is shared by both datasets. The networks constructed from raw data are directed weighted networks; nodes correspond to words and edges correspond to associations: the presence of an edge from node v_1 to node v_2 means that word v_2 was given as a response to word v_1. The weight of an edge from v_1 to v_2 represents the number of participants who responded with v_2 when being presented v_1. Networks in that form were preprocessed before being subject to the analysis.

Preprocessing consisted of removing nodes not representing "typical" words (sequences of digits, for instance), and transforming directed networks into undirected ones, by ignoring edge directions and retaining their weights; in cases where two oppositely directed edges were present between a pair of nodes, these edges were replaced with an undirected edge with weight equal to the sum of the weights of the original edges. After that, the final step of preprocessing took place: removing the nodes with strengths equal to 1, as they represent the words which appeared only once in the whole experiment, and hence are treated as providing information less reliable than the others. The resulting undirected weighted networks are the basis of the analysis discussed here; keeping the names of the datasets from which they are derived, they are referred to as the USFFA network and the EAT network. USFFA has 9958 nodes and 62491 edges; EAT has 15184 nodes and 90236 edges.

Apart from the two discussed networks in their original form, their randomizations are studied. The randomization consists of two steps: in the first step, the network is randomized according to the configuration model, treating the network as if it was unweighted. Then, in the second step, edge weights from the original network are randomly assigned to the edges of the randomized network. Hence the obtained network has a randomized structure, but preserves the distributions of node degrees and of edge weights. The randomized versions of the USFFA network and the EAT network are here referred to as USFFA-RAND and EAT-RAND networks, respectively. Since different realizations of randomization are different from each other, studying the characteristics of randomized networks involves averaging the results obtained in multiple realizations. In addition to USFFA and EAT networks and their randomizations, the minimum spanning trees (MSTs) of each of them are considered in the analysis. For a word-association network, determining the minimum spanning tree (with edge costs inversely proportional to weights) can be treated as a procedure which tends to remove all the associations except the strongest ones. Both the MST of a randomized network and the randomized network itself serve as reference networks allowing to decide whether the studied properties of the original networks are due to the presence of specific network organization, or whether they can be attributed solely to the distributions of node degrees and of edge weights. Figure 4.25 presents the visualizations of the USFFA network and the USFFA-RAND network (one particular realization), along with the visualizations of the MSTs of those networks. Figure 4.26 demonstrates an example of how words are organized in a word-association network; it shows a subnetwork of the USFFA network and a subnetwork of the MST of the USFFA network.

Figure 4.27 shows node degree distributions in USFFA network, EAT network, and in their MSTs. The tails of the distributions in USFFA and in EAT networks can be approximated by power laws. Since USFFA-RAND and EAT-RAND networks are constructed with the use of the configuration model, their node degree distributions are identical to the distributions in USFFA and EAT, respectively, and therefore they are not shown in Fig. 4.27. The MSTs of both the original networks (USFFA and EAT) and their randomizations also have node degree distributions approximately described by power laws. In terms of node degree distributions, USFFA and EAT are quite similar; this applies also to their MSTs. Moreover, node degree distributions

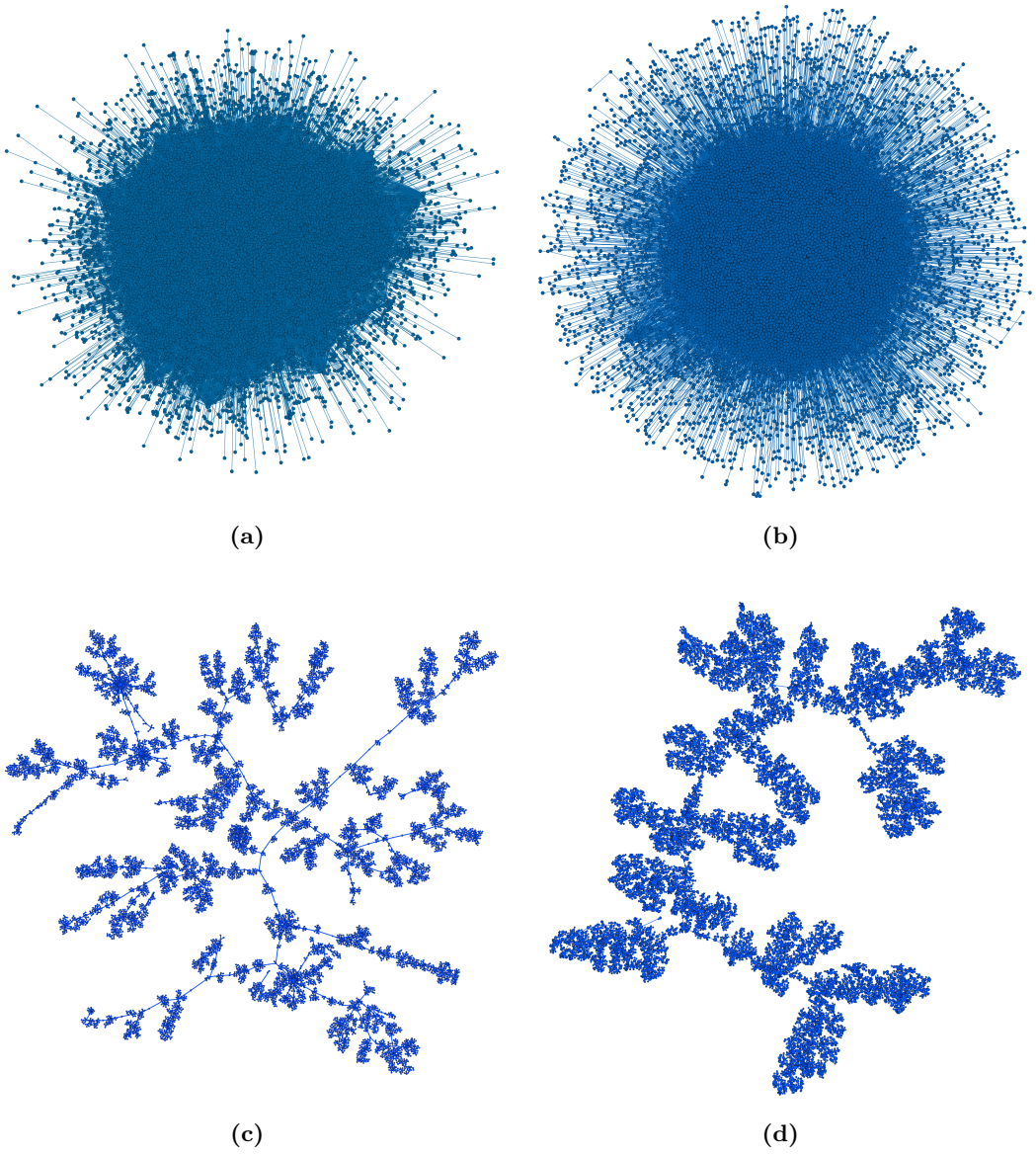
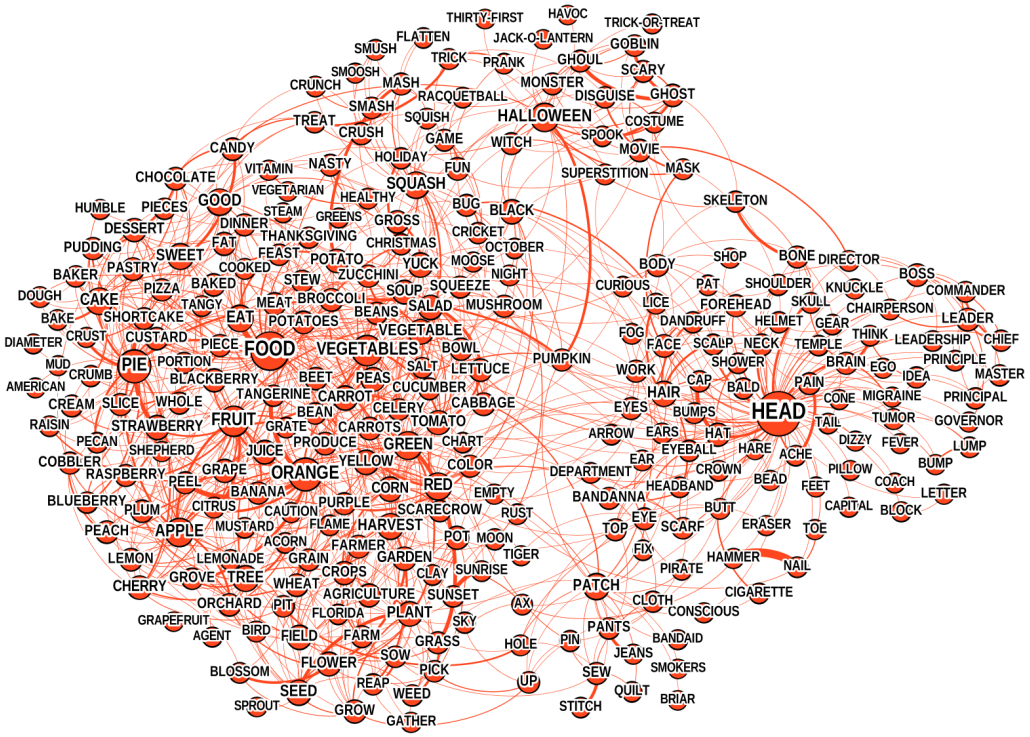
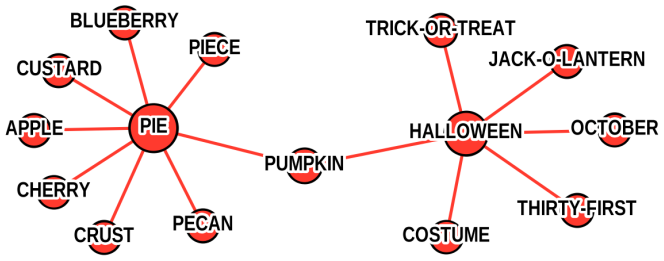


Figure 4.25. Visualizations of selected examples of networks studied in relation to word associations: (a) USFFA network, (b) USFFA-RAND network (one particular realization), (c) minimum spanning tree of the USFFA network, (d) minimum spanning tree of the USFFA-RAND network (one particular realization).



(a)



(b)

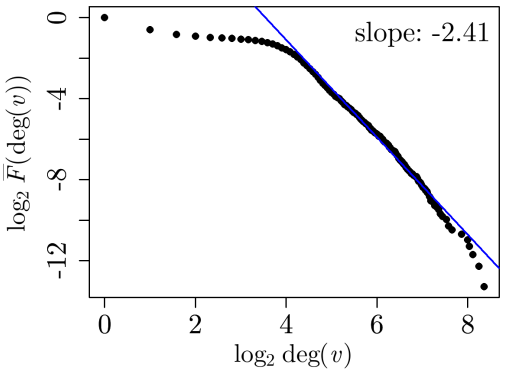
Figure 4.26. Examples of word-association networks. The network in (a) is a subnetwork of the USFFA network consisting of nodes which can be reached in at most 2 steps starting from the word "pumpkin". Node size and edge thickness represent node degree and edge weight, respectively. The network is constructed from the original USFFA network by removing all the nodes whose unweighted distance (the distance computed as in an unweighted network) from the word "pumpkin" greater than 2. The network in (b) is a subnetwork of the minimum spanning tree of the USFFA network, also restricted to nodes at most 2 steps away from the word "pumpkin".

in MSTs of USFFA and of EAT are not much different from the ones observed in MSTs of USFFA-RAND and EAT-RAND, in terms of distribution shape and the value of power law exponent.

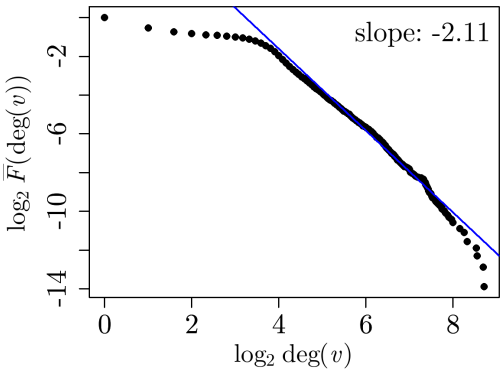
Table 4.5 presents selected (unweighted) characteristics of the studied networks. While some of the results can be considered as expected - for example, clustering coefficient C_u close to 0 in randomized networks and in MSTs, or average shortest path length ℓ_u being much longer in MSTs than in the networks from which MSTs are constructed - some of the presented quantities provide more specific information about the considered networks. The value of ℓ_u in USFFA and EAT being only slightly higher than in, respectively, USFFA-RAND and EAT-RAND, indicates that in each of these networks, at least some edges connect the parts of the network which would be distant from each other if those edges were absent; such edges serve as "shortcuts" responsible for keeping the average shortest path length relatively low, similar to the one observed in a random network. The relatively high values of modularity Q_u (0.44 and 0.43; significantly higher than in randomized networks) reflect the fact that to some extent, words can be grouped into clusters such that the associations connecting words are more dense inside the clusters than outside them. The presence of such clusters can also be related to the nonzero values of clustering coefficient C_u in both USFFA and EAT. In case of MSTs, modularity does not provide much information, since a network having the structure of a tree cannot contain densely connected clusters. Negative values of assortativity coefficients, r_u and ρ_u, express the preference of edges to connect high-degree nodes with low-degree nodes. This is a common situation in networks having many nodes with low degrees and a certain number of hubs with high degrees; this is also the case here. However, in USFFA and EAT networks, the effect cannot be attributed solely to degree distributions, as both r_u and ρ_u are close to 0 in USFFA-RAND and EAT-RAND. For MSTs, the observed disassortativity seems to be related more to how an MST is constructed than to any particular property of the studied networks, as MSTs of USFFA and of EAT have assortativity coefficients similar to MSTs of USFFA-RAND and EAT-RAND, respectively.

Table 4.5. Selected characteristics of the studied word-association networks - USFFA network, EAT network, and networks derived from these two. Each row corresponds to one network; in the first column network names are given, the remaining columns contain unweighted, global characteristics of networks: clustering coefficient C_u, average shortest path length ℓ_u, modularity Q_u, assortativity coefficient r_u and rank assortativity coefficient ρ_u.

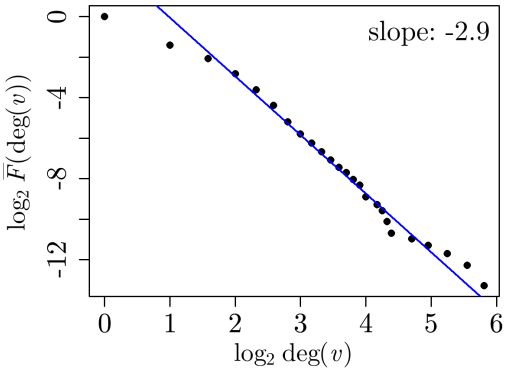
network	C_u	ℓ_u	Q_u	r_u	ρ_u
EAT	0.10	4.06	0.44	-0.09	-0.07
EAT-RAND	0.01	3.83	0.24	-0.02	-0.01
MST of EAT	0.00	27.78	0.98	-0.10	-0.36
MST of EAT-RAND	0.00	35.78	0.98	-0.13	-0.40
USFFA	0.12	3.95	0.43	-0.08	-0.07
USFFA-RAND	0.01	3.77	0.23	-0.01	-0.01
MST of USFFA	0.00	31.07	0.98	-0.12	-0.41
MST of USFFA-RAND	0.00	32.22	0.98	-0.16	-0.40



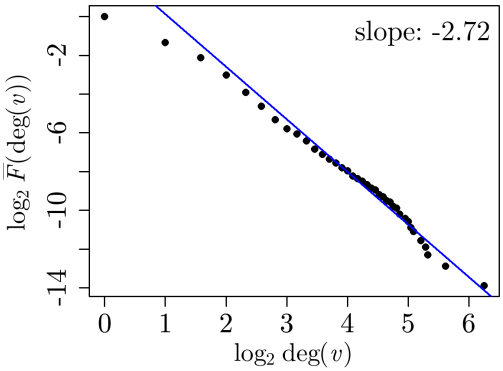
(a) USFFA network



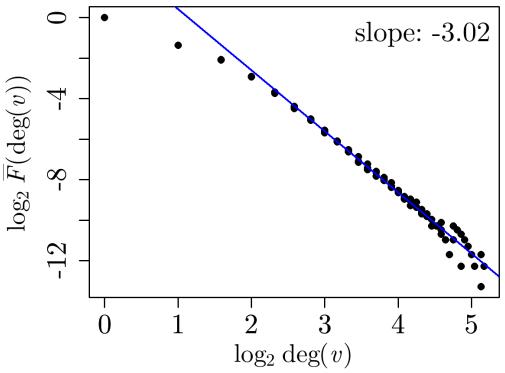
(b) EAT network



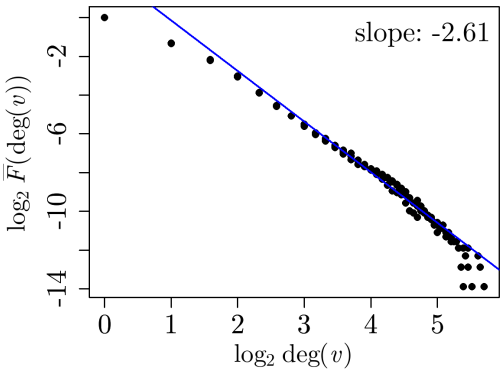
(c) MST of USFFA network



(d) MST of EAT network



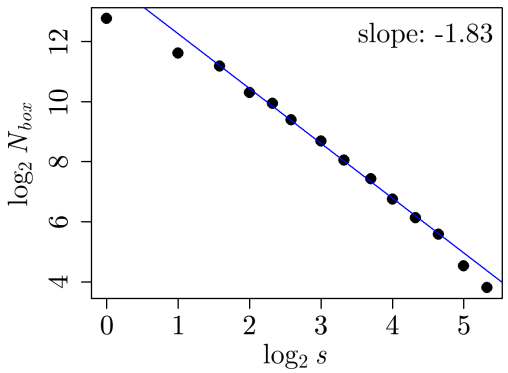
(e) MST of USFFA-RAND network (multiple realizations)



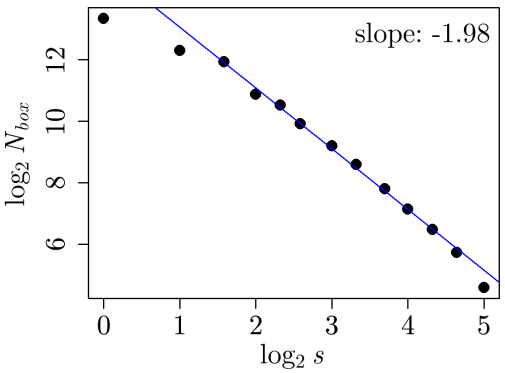
(f) MST of EAT-RAND network (multiple realizations)

Figure 4.27. Log-log plots of survival functions of node degree distributions $\bar{F}(\deg(v))$, for the USFFA network (a), EAT network (b), MST of the USFFA network (c), MST of the EAT network (d), MST of the USFFA-RAND network (e), MST of the EAT-RAND network (f). The slopes of the blue lines are given in the top-right corner of each figure. The distributions in the USFFA-RAND network and in the EAT-RAND network are not shown, as they are identical to tge distributions in USFFA and EAT networks, respectively. The figures pertaining to MSTs of randomized networks - (e) and (f) - present the superimposed results for three independent realizations of randomization.

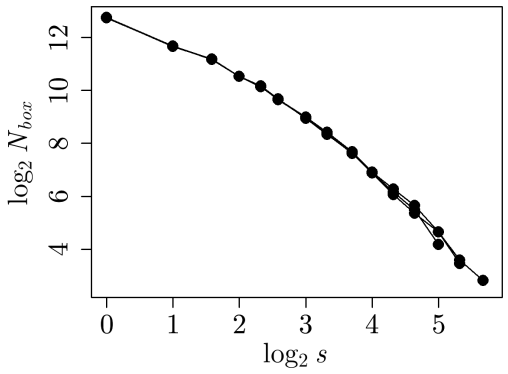
An interesting characteristic shared by USFFA and EAT networks is the fact that their MSTs are organized in a hierarchical fashion, exhibiting statistical self-similarity. In that regard, both USFFA and EAT networks differ from their randomized counterparts - as the MSTs of USFFA-RAND and of EAT-RAND do not have a fractal MST. Figure 4.28 presents the results of box-counting analysis performed for the discussed networks (MSTs of USFFA, EAT, and their randomizations)- it shows how the number of boxes N_{box} of given size s needed to cover a network depends on box size; a straight line on a log-log plot of $N_{box}(s)$ can be interpreted as the presence of fractality. Although the presented analysis might suffer from relatively small sizes of the available networks (resulting in restricted range of box sizes, which can affect the numerical stability of certain results, like fractal dimensions), it allows to detect qualitative differences between the structure of the MSTs of word-association networks and the structure of their randomizations. The existence of such differences suggests that the "skeletons" of the studied word-association networks, representing the strongest associations between words, are organized into a self-similar structure, which would not be observed if associations were connecting words in a random fashion. The estimated fractal dimensions of both networks have comparable values, 1.83 for the MST of USFFA and 1.98 for the MST of EAT. A question arises if such a result is characteristic only for the studied networks, or if it constitutes a more general property of word-association networks. However, a larger set of data, possibly in multiple languages, would be required to decide whether having an MST exhibiting self-similarity in the form discussed above is a common property of word-association networks, and to identify the possible implications of the existence of such structure on the understanding of how language is organized in human mind.



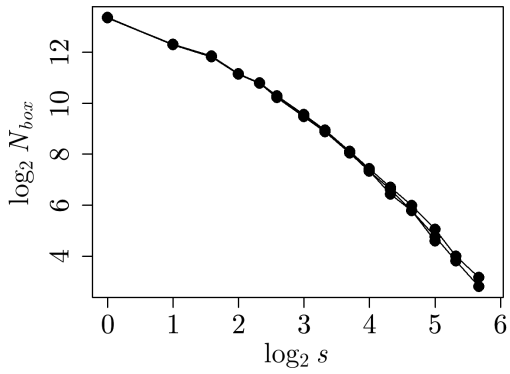
(a) Greedy coloring - MST of USFFA network



(b) Greedy coloring - MST of EAT network



(c) Greedy coloring - MST of USFFA-RAND network (multiple realizations)



(d) Greedy coloring - MST of EAT-RAND network (multiple realizations)

Figure 4.28. Identifying fractality in minimum spanning trees of word-association networks. The figures show the log-log plots of $N_{box}(s)$; s denotes the size of a box, and $N_{box}(s)$ is the number of boxes of size s needed to cover the network (using greedy coloring). The studied networks are: the MST of the USFFA network (a), the MST of the EAT network (b), the MST of the USFFA-RAND network (c) and the MST of the EAT-RAND network (d). In (a) and (b), a range in which the relationship $N_{box}(s)$ can be approximated by a power law can be identified; the slopes of the lines marked in blue are given in the top-right corner. In (c) and (d), which pertain to randomized networks, the data for three independent realizations of randomization is presented collectively.

Summary

A system so multifaceted and complicated as natural language is naturally a subject to research in diverse scientific fields. Since language is a complex system, various aspects of its organization and structure can be investigated using mathematical tools designed to study systems exhibiting complexity, ranging from basic methods of statistics and time series analysis to fractal geometry and network theory. With the use of such tools a number of natural language characteristics can be described in a quantitative way.

The analysis of word frequency distributions in literary texts confirms the validity of a well-known statistical law of natural language - Zipf's law. Studying word frequencies in more detail - for example considering words of different types (like different parts of speech) - reveals the differences between these types in terms of their statistical properties. An interesting and novel result is obtained for word rank-frequency distributions approximated by Zipf-Mandelbrot law when punctuation marks are included into the analysis - it turns out that treating punctuation marks in the same way as words decreases the value of the constant c responsible for the flattening of the rank-frequency distribution $\omega(R)$ in Zipf-Mandelbrot law (Eq. 2.53). In other words, with punctuation marks included, word frequency distribution is better approximated by a power law. The effect is present in all the languages studied in this work (English, German, French, Italian, Spanish, Polish and Russian); however, its strength depends on particular language. This can be an argument for including punctuation marks into statistical analysis of written language (especially from the standpoint of models attempting to explain power laws in word frequencies).

A valuable insight into the organization of natural language is obtained with the use of tools designed for time series analysis. Representing quantities like sentence lengths in the form of time series allows to reveal certain signs of complexity, like the presence of long-range correlations and fractal or multifractal structure. Not only the mentioned properties can be identified and quantified, but also they can be related to other specific characteristics of a text - for example, the presence of strong multiscaling in time series representing sentence lengths, evidenced by a wide multifractal spectrum $f(\alpha)$, is characteristic for texts using a narrative technique known as the stream of consciousness. An interesting perspective on certain aspects of the organization of written language is provided by analyzing the partition of a text determined by consecutive punctuation marks; a reason for which such a partition can be considered meaningful is the general purpose of punctuation in written language - that is, splitting text into parts determined by grammatical or logical consistency. From a text divided into pieces separated by punctuation marks one can construct a time series consisting of the lengths of consecutive pieces, measured by the number of words between consecutive punctuation marks. Such series of "punctuation waiting times" also have properties indicating the presence of complex patterns of organization, like long-range correlations or mutlifractality. Typically, the strength of both of the mentioned effects for punctuation waiting times is weaker than in case of time series representing sentence lengths. This is evidenced by the Hurst exponents of punctuation waiting times being lower than the Hurst exponents of sentence lengths

(but still above 0.5) and by the width of multifractal spectra $f(\alpha)$ - a text having a wide singularity spectrum of sentence lengths usually has a singularity spectrum of punctuation waiting times of significantly smaller width.

The correlation between the Hurst exponents of sentence lengths H_S and the Hurst exponents of punctuation waiting times H_{AP} reflects the fact that the partition of a text into sentences and the partition into parts determined by punctuation marks are related, since the ends of sentences are marked by a certain subset of all punctuation marks. To approach the problem of measuring the strength of that relationship, one can study the Hurst exponents H_S and H_{AP} of texts randomized in a specific way; two types of randomization procedures seem to be useful in that context. The first type randomizes the series of sentence lengths, keeping the series of punctuation waiting times unchanged; the second type randomizes punctuation waiting times, retaining the original sentence lengths. Both types of randomization result in a decrease of the Hurst exponent of the randomized series, but its value usually remains higher than 0.5. This indicates that the presence of correlations in time series representing sentence lengths and the presence of correlations in time series representing punctuation waiting times are indeed related to each other, but one cannot be fully explained by the other.

Investigating the probability distributions characterizing the values of the mentioned linguistic time series (representing sentence lengths or punctuation waiting times) leads to a conclusion that the distribution of punctuation waiting times can be successfully approximated by discrete Weibull distribution. This allows to view the arrangement of punctuation marks in written language in terms of a process whose statistical properties can be quantitatively expressed - by appropriate hazard functions. Different languages have different typical values of the distributions' parameters p and β, and therefore texts in different languages occupy slightly different regions on a (p, β) plane. It is worth noting that the approach to text analysis which leads to the discussed results can be considered a novelty - as the structures emerging from the partition of a text determined by consecutive punctuation marks have not been studied in the literature.

The analysis of time series representing sentence lengths and punctuation waiting times emphasizes the significance of punctuation in written language. Since both types of series are determined by the arrangement of punctuation marks (either all of them or the ones being of a specific type), it can be stated that punctuation in general is responsible for organizing written language in a specific way which results in the presence of complex patterns and structures. Moreover, the fact that the range of the observed Hurst exponents and the width of the multifractal spectra are smaller for punctuation waiting times than for sentence lengths indicates that while the placement of the ends of sentences leaves some freedom to the writer, the arrangement of punctuation as a whole seems to be less diverse and more uniform across texts in a given language. The same conclusion can be drawn from the analysis of probability distributions - while punctuation waiting times in texts seem to be quite universally described by discrete Weibull distribution, regularities of such kind are much harder to observe for sentence lengths. The presented results suggest that studying the structure determined by all punctuation marks can provide more information about the universal (not specific to a particular text) characteristics of written language, compared to the structure determined by sentences.

Studying linguistic networks - networks used to represent certain aspects of language structure - also leads to a number of interesting results. Word-adjacency networks - networks representing the co-occurrence of words in texts - are a tool which allows to investigate a number of statistical properties of a given language sample and to express them in terms of quantities used to describe complex net-

works. Some basic characteristics of a text, which are usually studied with the use of a representation simpler than word-adjacency networks (word frequency distribution is an example of such a characteristic), are incorporated into word-adjacency networks and can be easily retrieved. Word-adjacency networks allow to observe some effects which seem to be universal across languages, like the specific behavior of the most frequent words and the words with moderate frequencies in terms of network's local characteristics, namely local clustering coefficients and average shortest path lengths. At the same time, word-adjacency networks are able to grasp certain differences between texts. For example, global characteristics (clustering coefficient, assortativity and modularity, in several specific variants) of networks constructed from texts in different languages have slightly different ranges of variability; this allows to observe that in the space of the mentioned characteristics, different languages reside in different regions, and they are separated to some degree.

The structure of a word-adjacency network can be characteristic not only to a particular language, but also to a specific style of writing. This gives an opportunity to use word-adjacency networks in stylometry. The effectiveness of stylometric analysis utilizing word-adjacency networks is demonstrated on an example of authorship attribution task. The analysis of about 50 texts of 8 different authors in English and in Polish language shows that networks constructed from texts of different authors differ in some of their structural properties. While in terms of networks' global properties this effect can be identified only to some degree, local characteristics of selected words in word-adjacency networks allow to distinguish between texts of different authors with much better accuracy. Since local characteristics of a word-adjacency network describe certain statistical properties of word usage, and the considered words are the ones with the highest frequencies, it can be stated that structural differences between networks representing texts of different authors are a result of authors' individual patterns of using the most frequent words. A characteristic that seems to be particularly useful in grasping the information needed to recognize text authorship is the weighted variant of the clustering coefficient. Authorship attribution task, having the form of the statistical classification of texts with respect to their authorship, performed in the studied set of English and Polish books with the use of a general-purpose machine learning method - decision tree ensemble - achieves accuracy of about 80-90% when the clustering coefficients of only about 10-15 most frequent words are taken into account. It is important to note that network-based approach to text classification can be combined with other methods to improve the quality of the results.

As is the case with other representations of language samples, punctuation plays an important role also in the analysis based on word-adjacency networks. In all of the discussed word-adjacency networks punctuation marks are involved in the process of network construction, and they are treated in the same way as words. The characteristics of network nodes representing punctuation marks exhibit the behavior similar to the one exhibited by nodes representing most frequent words. This, being in accordance with the results of word frequency analysis taking punctuation into account, supports the claim that punctuation marks can be treated as words in some quantitative aspects. Moreover, patterns of punctuation usage are an important factor in the presented authorship attribution procedure: the accuracy of the performed text classification is significantly lower when punctuation marks are not included in the set of studied words. This shows that punctuation carries valuable information regarding the details of text structure dependent on factors like author's individual style of writing, and that with appropriate methods of statistical analysis the information of this kind can be utilized in practical applications.

Applying methods of network analysis to linguistic networks other than word-adjacency networks, namely word-association networks (which are designed to represent associations between words in human mind) identifies them as also having complex organization. A number of characteristics describing networks' basic statistical properties indicate that the two studied networks are similar to each other, although the data used to construct them comes from slightly different, independent experiments. An interesting fact about these networks is that their minimum spanning trees, which might be considered subnetworks consisting of only the strongest associations, have a statistically self-similar, fractal structure. The estimated fractal dimensions of both minimum spanning trees have similar values. This raises a question whether such an observation is valid only for the discussed networks, or whether it represents a more general property of word-association networks. Investigating the generality of the presented results, as well as establishing their relationship with other results regarding the organization of language in human mind, constitutes a possible direction of research in the future.

Appendix A

Decision tree bagging

In Chapter 4, the possibility of recognizing the authorship of texts is investigated with the use of a method of statistical classification known under the name of decision tree bagging. Here a short description of the key ideas of the method are discussed.

Classification with decision trees [334] (also called classification trees) is a method of supervised learning, which means that it aims at learning the patterns present in the data with the use of a set of already classified examples, called the training set. The information about the detected patterns is then used to classify other samples.

Creating an ensemble of decision trees requires constructing individual trees in the first place. Let A denote the set of n-dimensional vectors of real numbers; elements of A are called observations, and their coordinates are called attributes. Each observation is labeled with one of K categories, also called classes. The training of a single classification tree consists of: considering all possible *one-dimensional splits* of A, selecting and executing the *best split*, and repeating these steps recursively in the resulting subsets; splitting stops when A is partitioned in such a way that each subset contains observations of only one category. The *one-dimensional split* related to some attribute x_i is the choice of a constant number S and grouping the observations according to whether their coordinate x_i is smaller or greater than S. The *best split* is the one that maximizes the decrease of the diversity of distribution of classes in the considered set. The diversity can be measured, for example, by information entropy: $H = -\sum_{k=1}^{K} p_k \log p_k$ (p_k denotes the fraction of the observations in the set that belong to category k). In such case, the maximization of diversity decrease is equivalent to the maximization of the quantity $H_0 - H_{split}$, where H_0 is the initial entropy, and H_{split} is the weighted sum of entropies in the resulting subsets, with weights proportional to the numbers of elements in these subsets and adding up to 1.

The scheme of the consecutive splits of A is equivalent to a system of conditions imposed on the observations' attributes; such a system is a classification tree. A trained tree can be used to categorize observations with unknown class membership, by assigning them to appropriate subsets of A, according to the conditions satisfied by their components.

Classification with a single decision tree might suffer form instability, which means that the classifier may produce significantly different results for only slightly different training sets. Also, decision trees are prone to overfitting, which leads to decrease of classification accuracy of unknown observations.

Decision tree bagging (*bootstrap aggregating*) [334] is a method of enhancing the performance of classification based on decision tress. Given a training set with m observations, one can create N new training sets of size m, by sampling with replacement from the original set. Obtained sets, called bootstrap samples, for large m are expected to have the fraction $1 - 1/e$ (which is roughly 63.2%) of the unique observations from the original set, the rest being duplicates. A decision tree

is trained on each of the bootstrap samples, and the ensemble of N trees becomes a new classifier. When such an ensemble is given an observation to classify, each tree being its part classifies the observation on its own, and then the class that was chosen by most of trees becomes the final result of classification.

A typical method of verifying the performance of classification is cross-validation [357]. Its general idea is dividing the set A of observations with known class membership into two disjoint sets: the training set A_{train} and the test set A_{test}. The classifier is trained on A_{train} and then it classifies the observations in A_{test}, treating them as if their class memberships were unknown. Then the results are compared with the true class memberships of elements of A_{test}, and the number of correct matches indicates the classifier's performance. Partitioning A (using fixed sizes of A_{train} and A_{test}), training the classifier, and testing its performance is repeated a certain number of times, and the average result becomes the final assessment of classification's accuracy. The methods and rules of partitioning A may vary; the approach utilized in this work is the repeated random sub-sampling cross-validation (also called Monte Carlo cross-validation). In such an approach, each partition of A into A_{train} and A_{test} is random and independent of other partitions. The procedure can be modified by stratification, that is a condition imposing fixed proportions upon the numbers of elements of each class; stratification is often used to ensure that all classes are equally represented in the training set, but choices other than equal class contributions are also possible.

Appendix B

The books used in the study

B.1 Dataset B.1

Listed below are books comprising the main dataset used in this work. There are 223 books in 7 languages: English, German, French, Italian, Spanish, Polish and Russian. Each of the books is at least 1000 sentences long.

English

1. *A Room with a View* - E. M. Forster
2. *Alice's Adventures in Wonderland* - C. L. Dodgson
3. *Animal Farm* - G. Orwell
4. *Anne of Green Gables* - L. M. Montgomery
5. *Brave New World* - A. Huxley
6. *David Copperfield* - C. Dickens
7. *Frankenstein; or, The Modern Prometheus* - M. Shelley
8. *Gone with the Wind* - M. Mitchell
9. *Gulliver's Travels* - J. Swift
10. *Heart of Darkness* - J. Conrad
11. *The History of Tom Jones, a Foundling* - H. Fielding
12. *Ivanhoe* - W. Scott
13. *Jane Eyre: An Autobiography* - C. Brontë
14. *Kim* - J. R. Kipling
15. *Little Women* - L. M. Alcott
16. *Middlemarch* - M. A. Evans
17. *Moby-Dick; or, The Whale* - H. Melville
18. *Old Mortality* - S. Walter
19. *Olivier Twist* - C. Dickens
20. *Pride and Prejudice* - J. Austen
21. *Sense and Sensibility* - J. Austen
22. *Sons and Lovers* - D. H. Lawrence
23. *Tess of the d'Urbervilles* - T. Hardy
24. *The Adventures of Sherlock Holmes* - A. C. Doyle
25. *The Adventures of Tom Sawyer* - S. L. Clemens
26. *The Catcher in the Rye* - J. D. Salinger
27. *The Great Gatsby* - F. S. K. Fitzgerald
28. *The Last of the Mohicans* - J. F. Cooper
29. *Robinson Crusoe* - D. Defoe
30. *The Lion, the Witch and the Wardrobe* - C. S. Lewis
31. *The Old Man and the Sea* - E. Hemingway
32. *The Pickwick Papers* - C. Dickens
33. *The Prince and the Pauper* - S. L. Clemens
34. *The Rotters' Club* - J. Coe
35. *The Scarlet Letter* - N. Hawthorne
36. *The Thirty-Nine Steps* - J. Buchan
37. *The Time Machine* - H. G. Wells
38. *Treasure Island* - R. L. Stevenson
39. *Vanity Fair* - W. M. Thackeray

German

40. *Also sprach Zarathustra* - F. Nietzsche
41. *Buddenbrooks: Verfall einer Familie* - T. Mann
42. *Das Schloss* - F. Kafka
43. *Der grüne Heinrich* - G. Keller
44. *Der Mann ohne Eigenschaften* - R. Musil
45. *Der Process* - F. Kafka
46. *Der Schimmelreiter* - T. Storm
47. *Der Stechlin* - T. Fontane
48. *Der Untertan* - H. Mann
49. *Der Zauberberg* - T. Mann
50. *Die Blechtrommel* - G. Grass
51. *Die Chronik der Sperlingsgasse* - W. Raabe
52. *Die Judenbuche – Ein Sittengemälde aus dem gebirgichten Westfalen* - A. von Droste-Hülshoff
53. *Die Leiden des jungen Werthers* - J. W. von Goethe
54. *Die Leute von Seldwyla* - G. Keller
55. *Doktor Faustus* - T. Mann
56. *Effi Briest* - T. Fontane
57. *Emil und die Detektive* - E. Kästner
58. *Eulenpfingsten* - W. Raabe
59. *Frau Jenny Treibel* - T. Fontane
60. *Hiob* - J. Roth
61. *Klein Zaches genannt Zinnober* - E. T. A. Hoffmann
62. *Lebens-Ansichten des Katers Murr* - E. T. A. Hoffmann
63. *Pfisters Mühle* - W. Raabe
64. *Professor Unrat oder Das Ende eines Tyrannen* - H. Mann
65. *Radetzkymarsch* - J. Roth
66. *Wilhelm Meisters Lehrjahre* - J. W. von Goethe
67. *Winnetou* - K. May
68. *Wir Kinder vom Bahnhof Zoo* - C. Felscherinow

French

69. *Adolphe* - B. Constant
70. *Bel-Ami* - G. de Maupassant
71. *Candide* - F. M. Arouet
72. *Germinie Lacerteux* - J. & E. de Goncourt
73. *Jacques* - A. A. L. Dupin
74. *L'Éducation sentimentale* - G. Flaubert
75. *La Chute* - A. Camus

76. *La Peste* - A. Camus
77. *La Petite Fadette* - A. A. L. Dupin
78. *La Princesse de Clèves* - M. M. de La Fayette
79. *La Reine Margot* - A. Dumas
80. *Le Comte de Monte-Cristo* - A. Dumas
81. *Le Fantôme de l'Opéra* - G. Leroux
82. *Le Grand Meaulnes* - H. A. Fournier
83. *Le Petit Prince* - A. de Saint-Exupéry
84. *Le Rouge et le Noir, Chronique du XIXe siècle* - H. Beyle
85. *Le Tour du monde en quatre-vingts jours* - J. Verne
86. *Les Liaisons dangereuses* - P. C. de Laclos
87. *Les Misérables* - V. Hugo
88. *Les Trois Mousquetaires* - A. Dumas
89. *Lourdes* - É. Zola
90. *Madame Bovary* - G. Flaubert
91. *Mademoiselle de Maupin* - T. Gautier
92. *Histoire du chevalier Des Grieux et de Manon Lescaut* - A. Prévost
93. *Nana* - É. Zola
94. *Notre-Dame de Paris* - V. Hugo
95. *Valentine* - A. A. L. Dupin
96. *Vingt Mille Lieues sous les mers* - J. Verne
97. *Vol de nuit* - A. de Saint-Exupéry

Italian

98. *Addio, amore!* - M. Serao
99. *Canne al vento* - G. Deledda
100. *Cenere* - G. Deledda
101. *Con gli occhi chiusi* - F. Tozzi
102. *Cuore* - E. De Amicis
103. *Dell'arte della guerra* - N. Machiavelli
104. *Ettore Fieramosca* - M. d'Azeglio
105. *Fosca* - I. U. Tarchetti
106. *I Malavoglia* - G. Verga
107. *I promessi sposi* - A. Manzoni
108. *I Viceré* - F. De Roberto
109. *Il Corsaro Nero* - E. Salgari
110. *Il fu Mattia Pascal* - L. Pirandello
111. *Il marchese di Roccaverdina* - L. Capuana
112. *Il nome della rosa* - U. Eco
113. *Il piacere* - G. D'Annunzio
114. *Il romanzo di un maestro* - E. De Amicis
115. *Il romanzo della fanciulla* - M. Serao
116. *L'Illusione* - F. De Roberto
117. *La coscienza di Zeno* - I. Svevo
118. *Le avventure di Pinocchio. Storia di un burattino* - C. Collodi
119. *Le confessioni d'un italiano* - I. Nievo
120. *Le mie prigioni* - S. Pellico
121. *Le tigri di Mompracem* - E. Salgari
122. *Malombra* - A. Fogazzaro
123. *Mastro Don Gesualdo* - G. Verga
124. *Niccolò dei Lapi* - M. d'Azeglio
125. *Piccolo mondo antico* - A. Fogazzaro
126. *Tre croci* - F. Tozzi
127. *Una vita* - I. Svevo
128. *Uno, nessuno e centomila* - L. Pirandello

Spanish

129. *A la costa* - L. A. Martínez
130. *Abel Sánchez* - M. de Unamuno
131. *Amaya o los vascos en el siglo VIII* - F. Navarro Villoslada
132. *Cien años de soledad* - G. García Márquez
133. *Don Quijote de la Mancha* - M. de Cervantes
134. *Don Segundo Sombra* - R. Güiraldes
135. *Doña Bárbara* - R. Gallegos
136. *Doña Luz* - J. Valera
137. *Doña Perfecta* - B. Pérez Galdós
138. *El Criticón* - B. Gracián
139. *El Periquillo Sarniento* - J. J. F. de Lizardi
140. *El Señor de Bembibre* - E. Gil y Carrasco
141. *El sombrero de tres picos* - P. A. de Alarcón
142. *Facundo o civilización y barbarie en las pampas argentinas* - D. F. Sarmiento
143. *Fortunata y Jacinta* - B. Pérez Galdós
144. *La barraca* - V. B. Ibáñez
145. *La Regenta* - L. Alas
146. *La vida del Buscón* - F. de Quevedo
147. *La vorágine* - J. E. Rivera
148. *Los cuatro jinetes del Apocalipsis* - V. B. Ibáñez
149. *Los pazos de Ulloa* - E. Pardo Bazán
150. *María* - J. Isaacs
151. *Misericordia* - B. Pérez Galdós
152. *Pedro Páramo* - J. Rulfo
153. *Peñas arriba* - J. M. de Pereda
154. *Pepita Jiménez* - J. Valera
155. *Sab* - G. Gómez de Avellaneda
156. *Sotileza* - J. M. de Pereda
157. *Tirano Banderas* - R. M. del Valle-Inclán

Polish

158. *As* - A. Dygański
159. *Chłopi* - W. Reymont
160. *Cudzoziemka* - M. Kuncewiczowa
161. *Dewajtis* - M. Rodziewiczówny
162. *Faraon* - A. Głowacki
163. *Ferdydurke* - W. Gombrowicz
164. *Imperium* - R. Kapuściński
165. *Inny świat* - G. Herling-Grudziński
166. *Kamienie na szaniec* - A. Kamiński
167. *Kariera Nikodema Dyzmy* - T. Dołęga-Mostowicz
168. *Koroniarz w Galicji* - J. Lam
169. *Król Maciuś Pierwszy* - J. Korczak
170. *Lalka* - B. Prus
171. *Lato leśnych ludzi* - M. Rodziewiczówna
172. *Ludzie bezdomni* - S. Żeromski
173. *Na srebrnym globie. Rękopis z Księżyca* - J. Żuławski
174. *Nad Niemnem* - E. Orzeszkowa
175. *Nienasycenie* - S. I. Witkiewicz
176. *Ogniem i mieczem* - H. Sienkiewicz
177. *Ozimina* - W. Berent
178. *Poganka* - N. Żmichowska
179. *Popioły* - S. Żeromski
180. *Próchno* - W. Berent
181. *Przedwiośnie* - S. Żeromski
182. *Quo vadis* - H. Sienkiewicz
183. *Sklepy cynamonowe* - B. Schulz
184. *Stara baśń* - J. I. Kraszewski
185. *Szaleństwa panny Ewy* - K. Makuszyński
186. *Szatan z siódmej klasy* - K. Makuszyński
187. *Trans-Atlantyk* - W. Gombrowicz
188. *Trędowata* - H. Mniszkówna
189. *W pustyni i w puszczy* - H. Sienkiewicz
190. *Zaklęty dwór* - W. Łoziński
191. *Ziemia obiecana* - W. Reymont

Russian

192. *Анна Каренина (Anna Karenina)* - Л. Н. Толстой (L. N. Tolstoy)
193. *Архипелаг ГУЛАГ (Arkhipelag GULAG)* - А. И. Солженицын (A. I. Solzhenitsyn)
194. *Белая гвардия (Belaya gvardiya)* - М. А. Булгаков (M. A. Bulgakov)
195. *Бесы (Besy)* - Ф. М. Достоевский (F. M. Dostoyevskiy)
196. *Братья Карамазовы (Brat'ya Karamazovy)* - Ф. М. Достоевский (F. M. Dostoyevskiy)
197. *Чевенгур (Chevengur)* - А. П. Платонов (A. P. Platonov)
198. *Деревня (Derévnya)* - И. А. Бунин (I. A. Bunin)

199. *Доктор Живаго (Doktor Zhivago)* - Б. Л. Пастернак (B. L. Pasternak)
200. *Дворянское гнездо (Dvoryanskoye gnezdo)* - И. С. Тургенев (I. S. Turgenev)
201. *Дым (Dym)* - И. С. Тургенев (I. S. Turgenev)
202. *Герой нашего времени (Geroy nashego vremeni)* - М. Ю. Лермонтов (M. Y. Lermontov)
203. *Капитанская дочка (Kapitanskaya dochka)* - А. С. Пушкин (A. S. Pushkin)
204. *Котлован (Kotlovan)* - А. П. Платонов (A. P. Platonov)
205. *Мастер и Маргарита (Master i Margarita)* - М. А. Булгаков (M. A. Bulgakov)
206. *Мёртвые души (Mortvyye du'shi)* - Н. В. Гоголь (N. V. Gogol')
207. *Новъ (Nov')* - И. С. Тургенев (I. S. Turgenev)
208. *Обломов (Oblomov)* - И. А. Гончаров (I. A. Goncharov)
209. *Обрыв (Obryv)* - И. А. Гончаров (I. A. Goncharov)
210. *Отцы и дети (Ottsy i deti)* - И. С. Тургенев (I. S. Turgenev)
211. *Палата № 6 (Palata № 6)* - А. П. Чехов (A. P. Chekhov)
212. *Петербург (Peterburg)* - Б. Н. Бугаев (B. N. Bugaev)
213. *Пикник на обочине (Piknik na obochine)* - А. Н. & Б. Н. Стругацкий (A. N. & B. N. Strugatsky)
214. *Поединок (Poyedinok)* - А. И. Куприн (A. I. Kuprin)
215. *Преступление и наказание (Prestupleniye i nakazaniye)* - Ф. М. Достоевский (F. M. Dostoyevskiy)
216. *Путешествие из Петербурга в Москву (Puteshestviye iz Peterburga v Moskvu)* - А. Н. Радищев (A. N. Radishchev)
217. *Тарас Бульба (Taras bul'ba)* - Н. В. Гоголь (N. V. Gogol')
218. *Театральный роман (Teatral'nyy roman)* - М. А. Булгаков (M. A. Bulgakov)
219. *Тихий Дон (Tikhiy Don)* - М. А. Шолохов (M. A. Sholokhov)
220. *Три года (Tri goda)* - А. П. Чехов (A. P. Chekhov)
221. *Война и мир (Voyna i mir)* - Л. Н. Толстой (L. N. Tolstoy)
222. *Воскресение (Voskreseniye)* - Л. Н. Толстой (L. N. Tolstoy)
223. *Жизнь Арсеньева (Zhizn' Arsen'yeva)* - И. А. Бунин (I. A. Bunin)

B.2 Dataset B.2 - extension of Dataset B.1

Dataset B.2 consists of all texts from Dataset B.1 and additional books listed below. Since the dataset can be considered an extension of Dataset B.1 used for certain parts of the study, the numbering of the additional books extends the numbering of Dataset B.1.

224. *A Heartbreaking Work of Staggering Genius* - D. Eggers [English]
225. *Absalom, Absalom!* - W. Faulkner [English]
226. *As I Lay Dying* - W. Faulkner [English]
227. *Finnegans Wake* - J. Joyce [English]
228. *Pointed Roofs* - D. Richardson [English]
229. *The Ambassadors* - H. James [English]
230. *The Portrait of a Lady* - H. James [English]
231. *The Waves* - A. V. Woolf [English]
232. *Tristram Shandy* - L. Sterne [English]
233. *U.S.A.* - J. Dos Passos [English]
234. *Berlin Alexanderplatz* - A. Döblin [German]
235. *Mort à crédit* - L. F. Céline [French]
236. *2666* - R. Bolaño [Spanish]
237. *Rayuela* - J. Cortázar [Spanish]

B.3 Dataset B.3

The dataset presented in the tables below is used in the analysis of authorship attribution using the characteristics of word-adjacency networks. It consists of 96 books in English and Polish. For each language, there are 8 different authors, and 6 books of each author. Each of the books is at least 1500 sentences long.

English language

Author	Title	Year of publishing	Number of words (in thousands)	Number of punctuation marks (in thousands)	Number of sentences (in thousands)
Arthur Conan Doyle (1859-1930)	*Micah Clarke*	1888	178.1	23.8	9.2
	The Adventures of Sherlock Holmes	1892	104.6	15.0	6.9
	The Exploits of Brigadier Gerard	1896	74.7	9.8	4.0
	The Lost World	1912	75.8	10.2	4.5
	The Refugees	1893	122.9	17.3	7.8
	The Valley of Fear	1915	57.7	8.3	4.3
Charles Dickens (1812-1870)	*A Tale of Two Cities*	1859	136.2	23.2	7.8
	Barnaby Rudge	1841	254.0	44.3	12.7
	David Copperfield	1850	356.0	63.0	19.5
	Oliver Twist	1838	157.7	29.4	9.2
	The Mystery of Edwin Drood	1870	96.0	16.6	5.7
	The Pickwick Papers	1837	300.3	55.6	16.4
Daniel Defoe (1660-1731)	*Colonel Jack*	1722	141.4	20.6	4.2
	Memoirs of a Cavalier	1720	101.2	13.8	2.6
	Roxana: The Fortunate Mistress	1724	160.9	23.4	3.8
	Moll Flanders	1722	136.2	18.9	3.2
	Robinson Crusoe	1719	232.3	34.0	4.1
	Captain Singleton	1720	110.8	16.1	2.4
George Eliot (1819-1880)	*Adam Bede*	1859	215.1	28.0	9.0
	Daniel Deronda	1876	311.1	39.2	14.3
	Felix Holt, the Radical	1866	182.2	24.1	8.1
	Middlemarch	1872	318.1	41.2	14.9
	Romola	1863	227.9	29.5	9.1
	The Mill on the Floss	1860	207.3	30.5	9.0
George Orwell (1903-1950)	*Animal Farm*	1945	30.1	3.8	1.6
	Burmese Days	1934	97.7	15.1	7.4
	Coming up for Air	1939	82.8	10.4	5.3
	Down and Out in Paris and London	1933	66.8	9.9	4.0
	Keep the Aspidistra Flying	1936	87.1	14.3	7.8
	Nineteen Eighty-Four	1949	103.7	14.2	6.7
Jane Austen (1775-1817)	*Emma*	1815	160.3	26.5	8.6
	Mansfield Park	1814	159.8	22.5	6.9
	Northanger Abbey	1818	77.3	11.4	3.6
	Persuasion	1818	83.3	12.4	3.7
	Pride and Prejudice	1813	121.8	17.2	6.0
	Sense and Sensibility	1811	119.5	18.0	5.2
Joseph Conrad (1857-1924)	*An Outcast of the Islands*	1896	104.4	18.0	8.8
	Chance: A Tale in Two Parts	1913	137.4	18.3	9.6
	Lord Jim	1900	129.3	20.4	8.9
	Nostromo: A Tale of the Seaboard	1904	168.3	24.3	10.1
	Under Western Eyes	1911	112.8	16.9	8.6
	Victory: An Island Tale	1915	114.8	18.3	8.6
Mark Twain (1835-1910)	*Following the Equator*	1897	186.7	26.1	9.2
	Life on the Mississippi	1883	143.9	21.1	6.8
	The Adventures of Huckleberry Finn	1884	110.8	16.9	5.9
	The Adventures of Tom Sawyer	1876	70.5	11.6	4.9
	The Innocents Abroad	1869	192.4	25.4	8.8
	The Prince and the Pauper	1882	67.2	10.8	3.3

Polish language

Author	Title	Year of publishing	Number of words (in thousands)	Number of punctuation marks (in thousands)	Number of sentences (in thousands)
Bolesław Prus (1847-1912)	*Anielka*	1885	46.6	11.6	5.7
	Dzieci	1909	66.1	19.1	9.1
	Emancypantki	1894	249.2	63.5	30.2
	Faraon	1897	193.3	45.8	19.7
	Lalka	1890	246.4	61.2	28.2
	Placówka	1886	68.6	17.5	7.7
Eliza Orzeszkowa (1841-1910)	*Cham*	1888	59.4	14.2	4.4
	Dziurdziowie	1885	50.6	10.8	3.7
	Jędza	1891	40.1	9.9	3.7
	Marta	1873	62.1	12.5	4.5
	Meir Ezofowicz	1878	97.3	18.4	6.9
	Nad Niemnem	1888	164.4	32.6	11.8
Henryk Sienkiewicz (1846-1916)	*Ogniem i mieczem*	1884	239.0	51.9	19.4
	Potop	1886	380.4	84.7	32.8
	Quo Vadis	1896	168.3	34.1	11.4
	Rodzina Połanieckich	1894	210.2	47.3	15.1
	W pustyni i w puszczy	1912	99.5	18.1	6.5
	Wiry	1910	92.7	20.0	6.6
Jan Lam (1838-1886)	*Dziwne karyery*	1881	70.9	14.8	4.7
	Humoreski	1883	37.6	7.5	2.3
	Idealiści	1876	100.9	21.3	7.0
	Koroniarz w Galicyi	1870	62.0	11.6	3.3
	Rozmaitości i powiastki	1878	52.4	10.3	2.6
	Wielki świat Capowic	1869	37.7	6.7	1.6
Janusz Korczak (1879-1942)	*Bankructwo małego Dżeka*	1924	41.7	10.3	4.3
	Dzieci ulicy	1901	50.2	13.2	5.9
	Dziecko salonu	1906	54.1	16.7	6.3
	Kajtuś czarodziej	1934	46.0	17.9	9.8
	Król Maciuś na wyspie bezludnej	1923	45.9	12.1	5.2
	Król Maciuś Pierwszy	1923	67.0	15.1	6.2
Józef Ignacy Kraszewski (1812-1887)	*Barani Kożuszek*	1898	60.3	14.8	5.5
	Boża opieka	1873	51.3	11.9	4.9
	Boży gniew	1886	112.0	24.9	7.8
	Bracia rywale	1877	42.0	10.5	3.8
	Infantka	1884	111.3	23.2	8.2
	Złote jabłko	1853	107.5	22.6	9.7
Stefan Żeromski (1846-1925)	*Dzieje grzechu*	1908	148.5	35.0	15.9
	Ludzie bezdomni	1900	97.2	20.2	8.6
	Popioły	1902	222.4	46.5	21.0
	Przedwiośnie	1924	81.3	17.5	7.3
	Syzyfowe prace	1897	61.9	11.6	4.5
	Wierna rzeka	1912	39.5	8.6	4.0
Władysław Reymont (1867-1925)	*Bunt*	1924	40.2	9.0	3.8
	Fermenty	1897	117.9	30.9	9.8
	Lili	1899	25.0	6.3	2.2
	Marzyciel	1910	47.7	11.9	4.4
	Wampir	1911	53.0	13.0	4.5
	Ziemia obiecana	1899	170.9	38.0	13.3

Bibliography

[1] C. K. Catchpole, P. J. B. Slater, *Bird Song*, Cambridge University Press, 2012.

[2] K. von Frisch, *Decoding the Language of the Bee*, Science 185 (4152) (1974). doi:10.1126/science.185.4152.663.

[3] W. H. Kirchner, M. Lindauer, A. Michelsen, *Honeybee dance communication*, Naturwissenschaften 75 (12) (1988). doi:10.1007/bf00366482.

[4] V. M. Janik, *Whistle Matching in Wild Bottlenose Dolphins (Tursiops truncatus)*, Science 289 (5483) (2000). doi:10.1126/science.289.5483.1355.

[5] V. M. Janik, L. S. Sayigh, R. S. Wells, *Signature whistle shape conveys identity information to bottlenose dolphins*, Proceedings of the National Academy of Sciences 103 (21) (2006). doi:10.1073/pnas.0509918103.

[6] B. D. López, *Whistle characteristics in free-ranging bottlenose dolphins (Tursiops truncatus) in the Mediterranean Sea: Influence of behaviour*, Mammalian Biology 76 (2) (2011). doi:10.1016/j.mambio.2010.06.006.

[7] L. G. Domb, M. Pagel, *Sexual swellings advertise female quality in wild baboons*, Nature 410 (6825) (2001). doi:10.1038/35065597.

[8] L. Gosling, S. Roberts, *Scent-marking by male mammals: Cheat-proof signals to competitors and mates*, in: Advances in the Study of Behavior, Elsevier, 2001. doi:10.1016/s0065-3454(01)80007-3.

[9] H. Slabbekoorn, T. B. Smith, *Bird song, ecology and speciation*, Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 357 (1420) (2002). doi:10.1098/rstb.2001.1056.

[10] J. Bräuer, J. Call, M. Tomasello, *All Great Ape Species Follow Gaze to Distant Locations and Around Barriers.*, Journal of Comparative Psychology 119 (2) (2005). doi:10.1037/0735-7036.119.2.145.

[11] C. Schloegl, K. Kotrschal, T. Bugnyar, *Gaze following in common ravens, Corvus corax: ontogeny and habituation*, Animal Behaviour 74 (4) (2007). doi:10.1016/j.anbehav.2006.08.017.

[12] R. D. Paulos, K. M. Dudzinski, S. A. Kuczaj, *The role of touch in select social interactions of Atlantic spotted dolphin (Stenella frontalis) and Indo-Pacific bottlenose dolphin (Tursiops aduncus)*, Journal of Ethology 26 (1) (2007). doi:10.1007/s10164-007-0047-y.

[13] V. L. Salazar, P. K. Stoddard, *Sex differences in energetic costs explain sexual dimorphism in the circadian rhythm modulation of the electrocommunication signal of the gymnotiform fish Brachyhypopomus pinnicaudatus*, Journal of Experimental Biology 211 (6) (2008). doi:10.1242/jeb.014795.

[14] C. N. Slobodchikoff, A. Paseka, J. L. Verdolin, *Prairie dog alarm calls encode labels about predator colors*, Animal Cognition 12 (3) (2008). doi:10.1007/s10071-008-0203-y.

[15] C. Hockett, *The origin of speech*, Scientific American 203 (1960).

[16] S. Wacewicz, P. Żywiczyński, *Language Evolution: Why Hockett's Design Features are a Non-Starter*, Biosemiotics 8 (1) (2014). doi:10.1007/s12304-014-9203-2.

[17] W. O'Grady, M. Dobrovolsky, F. Katamba, *Contemporary linguistics: An introduction*, Longman, London New York, 1997.

[18] G. Yule, *The study of language*, Cambridge University Press, Cambridge, UK New York, 2010.

[19] K. R. Gibson, M. Tallerman (Eds.), *The Oxford Handbook of Language Evolution*, Oxford University Press, 2011. doi:10.1093/oxfordhb/9780199541119.001.0001.

[20] M. D. Hauser, C. Yang, R. C. Berwick, I. Tattersall, M. J. Ryan, J. Watumull, N. Chomsky, R. C. Lewontin, *The mystery of language evolution*, Frontiers in Psychology 5 (2014). doi:10.3389/fpsyg.2014.00401.

[21] A. J. Jean Aitchison, *The Seeds of Speech*, Cambridge University Press, 2012.

[22] S. Johansson, *Origins of language: constraints on hypotheses*, John Benjamins Pub, Amsterdam Philadelphia, 2005.

[23] S. Mcbrearty, A. S. Brooks, *The revolution that wasn't: a new interpretation of the origin of modern human behavior*, Journal of Human Evolution 39 (5) (2000). doi:10.1006/jhev.2000.0435.

[24] S. Pääbo, *The mosaic that is our genome*, Nature 421 (6921) (2003). doi:10.1038/nature01400.

[25] S. E. Fisher, *Evolution of language: Lessons from the genome*, Psychonomic Bulletin & Review 24 (1) (2016). doi:10.3758/s13423-016-1112-8.

[26] H. S. Mountford, D. F. Newbury, *The genomic landscape of language: Insights into evolution*, Journal of Language Evolution 3 (1) (2017). doi:10.1093/jole/lzx019.

[27] R. DeSalle, I. Tattersall, *What aDNA can (and cannot) tell us about the emergence of language and speech*, Journal of Language Evolution 3 (1) (2017). doi:10.1093/jole/lzx018.

[28] W. T. Fitch, *The Evolution of Language*, Cambridge University Press, 2015.

[29] M. Ruvolo, *Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets*, Molecular Biology and Evolution 14 (3) (1997). doi:10.1093/oxfordjournals.molbev.a025761.

[30] B. Wood, B. G. Richmond, *Human evolution: taxonomy and paleobiology*, Journal of Anatomy 197 (1) (2000). doi:10.1046/j.1469-7580.2000.19710019.x.

[31] F.-C. Chen, W.-H. Li, *Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees*, The American Journal of Human Genetics 68 (2) (2001). doi:10.1086/318206.

[32] B. J. Bradley, *Reconstructing phylogenies and phenotypes: a molecular view of human evolution*, Journal of Anatomy 212 (4) (2008). doi:10.1111/j.1469-7580.2007.00840.x.

[33] B. Wood, T. Harrison, *The evolutionary context of the first hominins*, Nature 470 (7334) (2011). doi:10.1038/nature09709.

[34] L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, J. B. Reece, *Campbell biology*, Pearson Education, Inc, New York, NY, 2017.

[35] E. D. Jarvis, S. Ribeiro, M. L. da Silva, D. Ventura, J. Vielliard, C. V. Mello, *Behaviourally driven gene expression reveals song nuclei in hummingbird brain*, Nature 406 (6796) (2000). doi:10.1038/35020570.

[36] T. C. Scott-Phillips, *Meaning in animal and human communication*, Animal Cognition 18 (3) (2015). doi:10.1007/s10071-015-0845-5.

[37] R. Moore, *Meaning and ostension in great ape gestural communication*, Animal Cognition 19 (1) (2015). doi:10.1007/s10071-015-0905-x.

[38] T. C. Scott-Phillips, *Meaning in great ape communication: summarising the debate*, Animal Cognition 19 (1) (2015). doi:10.1007/s10071-015-0936-3.

[39] M. Tomasello, J. Call, *Thirty years of great ape gestures*, Animal Cognition 22 (4) (2018). doi:10.1007/s10071-018-1167-1.

[40] 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) (2002). doi:10.1126/science.298.5598.1569.

[41] H. H. Hock, B. D. Joseph, *Language History, Language Change, and Language Relationship*, De Gruyter Mouton, 2009.

[42] T. Lansdall-Welfare, S. Sudhahar, J. Thompson, J. Lewis, N. C. and, *Content analysis of 150 years of British periodicals*, Proceedings of the National Academy of Sciences 114 (4) (2017). doi:10.1073/pnas.1606380114.

[43] T. Lansdall-Welfare, S. Sudhahar, G. A. Veltri, N. Cristianini, *On the coverage of science in the media: A big data study on the impact of the Fukushima disaster*, in: 2014 IEEE International Conference on Big Data (Big Data), IEEE, 2014. doi:10.1109/bigdata.2014.7004454.

[44] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, E. L. A. and, *Quantitative Analysis of Culture Using Millions of Digitized Books*, Science 331 (6014) (2010). doi:10.1126/science.1199644.

[45] H. Hammarström, R. Forkel, M. Haspelmath, S. Bank, *Glottolog database 4.2.1*, https://glottolog.org/glottolog/glottologinformation (2020). doi:10.5281/ZENODO.3754591.

[46] D. M. Eberhard, G. F. Simons, C. D. Fennig, *Ethnologue: Languages of the World. 23rd edition (online)*, https://www.ethnologue.com/ (2020).

[47] M. Swadesh, *The origin and diversification of language*, Aldine, Atherton, Chicago, 1971.

[48] R. Mcmahon, *Language Classification by Numbers*, Oxford University Press, 2006.

[49] R. D. Gray, Q. D. Atkinson, *Language-tree divergence times support the Anatolian theory of Indo-European origin*, Nature 426 (6965) (2003). doi:10.1038/nature02029.

[50] F. Petroni, M. Serva, *Language distance and tree reconstruction*, Journal of Statistical Mechanics: Theory and Experiment 2008 (08) (2008). doi:10.1088/1742-5468/2008/08/p08012.

[51] I. Dyen, J. B. Kruskal, P. Black, *An Indoeuropean Classification: A Lexicostatistical Experiment*, Transactions of the American Philosophical Society 82 (5) (1992). doi:10.2307/1006517.

[52] B. F. Skinner, *Verbal behavior*, Copley, Acton, Mass, 1992.

[53] M. Tomasello, *Constructing a language: a usage-based theory of language acquisition*, Harvard University Press, Cambridge, Mass, 2003.

[54] N. Chomsky, *Aspects of the theory of syntax*, MIT Press, Cambridge, Massachusetts, 1965.

[55] S. Pinker, *The language instinct*, William Morrow and Company, New York, 1994.

[56] G. K. Pullum, B. C. Scholz, *Empirical assessment of stimulus poverty arguments*, The Linguistic Review 18 (1-2) (2002). doi:10.1515/tlir.19.1-2.9.

[57] J. A. Legate, C. D. Yang, *Empirical re-assessment of stimulus poverty arguments*, The Linguistic Review 18 (1-2) (2002). doi:10.1515/tlir.19.1-2.151.

[58] A. Fernald, V. A. Marchman, *Language learning in infancy*, in: M. J.Traxler, M. A. Gernsbacher (Eds.), Handbook of psycholinguistics, Elsevier/Academic Press, Amsterdam Boston, 2006, Ch. 27.

[59] K. D. Bot, *A History of Applied Linguistics*, Taylor & Francis Ltd, 2015.

[60] P. K. Kuhl, *Brain Mechanisms in Early Language Acquisition*, Neuron 67 (5) (2010). doi:10.1016/j.neuron.2010.08.038.

[61] C. J. Price, *A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading*, NeuroImage 62 (2) (2012). doi:10.1016/j.neuroimage.2012.04.062.

[62] G. I. de Zubicaray, N. O. Schiller (Eds.), *The Oxford Handbook of Neurolinguistics*, Oxford University Press, 2019. doi:10.1093/oxfordhb/9780190672027.001.0001.

[63] N. Geschwind, *The Organization of Language and the Brain*, Science 170 (3961) (1970).

[64] P. Tremblay, A. S. Dick, *Broca and Wernicke are dead, or moving past the classic model of language neurobiology*, Brain and Language 162 (2016). doi:10.1016/j.bandl.2016.08.004.

[65] G. Hickok, D. Poeppel, *The cortical organization of speech processing*, Nature Reviews Neuroscience 8 (5) (2007). doi:10.1038/nrn2113.

[66] D. Saur, B. W. Kreher, S. Schnell, D. Kümmerer, P. Kellmeyer, M.-S. Vry, R. Umarova, M. Musso, V. Glauche, S. Abel, W. Huber, M. Rijntjes, J. Hennig, C. Weiller, *Ventral and dorsal pathways for language*, Proceedings of the National Academy of Sciences 105 (46) (2008). doi:10.1073/pnas.0805234105.

[67] G. Hickok, *The functional neuroanatomy of language*, Physics of Life Reviews 6 (3) (2009). doi:10.1016/j.plrev.2009.06.001.

[68] D. Kemmerer, *Cognitive neuroscience of language*, Psychology Press, New York, NY Hove, East Sussex, 2015.

[69] G. Nasios, E. Dardiotis, L. Messinis, *From Broca and Wernicke to the Neuromodulation Era: Insights of Brain Language Networks for Neurorehabilitation*, Behavioural Neurology 2019 (2019). doi:10.1155/2019/9894571.

[70] J. Fodor, *The language of thought*, Harvard University Press, Cambridge, Massachusetts, 1975.

[71] A. Tillas, *Language as grist to the mill of cognition*, Cognitive Processing 16 (3) (2015). doi:10.1007/s10339-015-0656-2.

[72] L. J. Kaye, *The Languages of Thought*, Philosophy of Science 62 (1) (1995). doi:10.1086/289841.

[73] J. L. Garfield, *Mentalese not spoken here: Computation, cognition and causation*, Philosophical Psychology 10 (4) (1997). doi:10.1080/09515089708573231.

[74] P. Carruthers, *Language, Thought and Consciousness: An Essay In Philosophical Psychology*, Cambridge University Press, 1998.

[75] C. Viger, *Learning to Think: A Response to the Language of Thought Argument for Innateness*, Mind and Language 20 (3) (2005). doi:10.1111/j.0268-1064.2005.00287.x.

[76] J. H. Hill, B. Mannheim, *Language and World View*, Annual Review of Anthropology 21 (1) (1992). doi:10.1146/annurev.an.21.100192.002121.

[77] J. Lucy, *Sapir–Whorf Hypothesis*, in: International Encyclopedia of the Social & Behavioral Sciences, Elsevier, 2001. doi:10.1016/b0-08-043076-7/03042-4.

[78] B. L. Whorf, *An American Indian Model of The Universe*, in: J. B. Carrol (Ed.), Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf, Technology Press of Massachusetts Institute of Technology John Wiley & Sons, Cambridge, Mass. New York, 1956.

[79] B. L. Whorf, *Science and Linguistics*, in: J. B. Carrol (Ed.), Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf, Technology Press of Massachusetts Institute of Technology John Wiley & Sons, Cambridge, Mass. New York, 1956.

[80] E. Malotki, *Hopi Time*, De Gruyter Mouton, 1983.

[81] D. L. Everett, *Cultural Constraints on Grammar and Cognition in Pirahã*, Current Anthropology 46 (4) (2005). doi:10.1086/431525.

[82] M. C. Frank, D. L. Everett, E. Fedorenko, E. Gibson, *Number as a cognitive technology: Evidence from Pirahã language and cognition*, Cognition 108 (3) (2008). doi:10.1016/j.cognition.2008.04.007.

[83] L. Boroditsky, L. A. Schmidt, W. Phillips, *Sex, Syntax, and Semantics*, in: D. Getner, S. Goldin-Meadow (Eds.), Language in Mind: Advances in the Study of Language and Thought, MIT Press, 2003.

[84] E. Ünal, A. Papafragou, *Interactions Between Language and Mental Representations*, Language Learning 66 (3) (2016). doi:10.1111/lang.12188.

[85] K. S. Jones, *Natural Language Processing: A Historical Review*, in: Current Issues in Computational Linguistics: In Honour of Don Walker, Springer Netherlands, 1994. doi:10.1007/978-0-585-35958-8_1.

[86] T. Young, D. Hazarika, S. Poria, E. Cambria, *Recent Trends in Deep Learning Based Natural Language Processing [Review Article]*, IEEE Computational Intelligence Magazine 13 (3) (2018). doi:10.1109/mci.2018.2840738.

[87] L. Deng, Y. Liu (Eds.), *Deep Learning in Natural Language Processing*, Springer Singapore, 2018. doi:10.1007/978-981-10-5209-5.

[88] N. Chomsky, *Syntactic Structures*, De Gruyter, 1957. doi:10.1515/9783112316009.

[89] N. Chomsky, *On certain formal properties of grammars*, Information and Control 2 (2) (1959). doi:10.1016/s0019-9958(59)90362-6.

[90] W. J. M. Levelt, *Formal grammars in linguistics and psycholinguistics*, John Benjamins Pub, Amsterdam Philadelphia, 2008.

[91] D. Jurafsky, J. H. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, Pearson Prentice Hall, Upper Saddle River, N.J, 2009.

[92] A. Lindenmayer, *Mathematical models for cellular interactions in development I. Filaments with one-sided inputs*, Journal of Theoretical Biology 18 (3) (1968). doi:10.1016/0022-5193(68)90079-9.

[93] A. Lindenmayer, *Mathematical models for cellular interactions in development II. Simple and branching filaments with two-sided inputs*, Journal of Theoretical Biology 18 (3) (1968). doi:10.1016/0022-5193(68)90080-5.

[94] P. Prusinkiewicz, A. Lindenmayer, *The Algorithmic Beauty of Plants*, Springer New York, 1990. doi:10.1007/978-1-4613-8476-2.

[95] G. Zipf, *The Psychobiology of Language*, Routledge, 1936. doi:10.2307/408930.

[96] G. Zipf, *Human behavior and the principle of least effort: an introduction to human ecology*, Addison-Wesley Press, 1949. doi:10.2307/409735.

[97] S. T. Piantadosi, *Zipf's word frequency law in natural language: A critical review and future directions*, Psychonomic Bulletin & Review 21 (5) (2014). doi:10.3758/s13423-014-0585-6.

[98] H. S. Heaps, *Information retrieval, computational and theoretical aspects*, Academic Press, New York, 1978.

[99] 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) (2007). doi:10.1002/asi.20524.

[100] A. Chacoma, D. H. Zanette, *Heaps' Law and Heaps functions in tagged texts: evidences of their linguistic relevance*, Royal Society Open Science 7 (3) (2020). doi:10.1098/rsos.200008.

[101] G. Altmann, *Prolegomena to Menzerath's law*, Glottometrika 2 (2) (1980).

[102] J. Milička, *Menzerath's Law: The Whole is Greater than the Sum of its Parts*, Journal of Quantitative Linguistics 21 (2) (2014). doi:10.1080/09296174.2014.882187.

[103] I. G. Torre, B. Luque, L. Lacasa, J. Luque, A. Hernández-Fernández, *Emergence of linguistic laws in human voice*, Scientific Reports 7 (1) (2017). doi:10.1038/srep43862.

[104] Á. Corral, I. Serra, *The Brevity Law as a Scaling Law, and a Possible Origin of Zipf's Law for Word Frequencies*, Entropy 22 (2) (2020). doi:10.3390/e22020224.

[105] E. G. Altmann, M. Gerlach, *Statistical Laws in Linguistics*, in: Lecture Notes in Morphogenesis, Springer International Publishing, 2016. doi:10.1007/978-3-319-24403-7_2.

[106] J. Kwapień, S. Drożdż, *Physical approach to complex systems*, Physics Reports 515 (3-4) (2012). doi:10.1016/j.physrep.2012.01.007.

[107] P. W. Anderson, *More Is Different*, Science 177 (4047) (1972). doi:10.1126/science.177.4047.393.

[108] Aristotle, *Metaphysics*, NuVision Publications, Sioux Falls, SD, 2005, translated by William David Ross.

[109] A. Roli, M. Villani, A. Filisetti, R. Serra, *Dynamical Criticality: Overview and Open Questions*, Journal of Systems Science and Complexity 31 (3) (2017). doi:10.1007/s11424-017-6117-5.

[110] R. V. Solé, S. C. Manrubia, B. Luque, J. Delgado, J. Bascompte, *Phase transitions and complex systems:Simple, nonlinear models capture complex systems at the edge of chaos*, Complexity 1 (4) (1996). doi:10.1002/cplx.6130010405.

[111] C. G. Langton, *Studying artificial life with cellular automata*, Physica D: Nonlinear Phenomena 22 (1-3) (1986). doi:10.1016/0167-2789(86)90237-x.

[112] C. G. Langton, *Computation at the edge of chaos: Phase transitions and emergent computation*, Physica D: Nonlinear Phenomena 42 (1-3) (1990). doi:10.1016/0167-2789(90)90064-v.

[113] P. Melby, J. Kaidel, N. Weber, A. Hübler, *Adaptation to the Edge of Chaos in the Self-Adjusting Logistic Map*, Physical Review Letters 84 (26) (2000). doi:10.1103/physrevlett.84.5991.

[114] M. Mitchell, *Complexity: A Guided Tour*, Oxford University Press, 2009.

[115] M. Baym, A. W. Hübler, *Conserved quantities and adaptation to the edge of chaos*, Physical Review E 73 (5) (2006). doi:10.1103/physreve.73.056210.

[116] E. Landa, I. O. Morales, R. Fossion, P. Stránský, V. Velázquez, J. C. L. Vieyra, A. Frank, *Criticality and long-range correlations in time series in classical and quantum systems*, Physical Review E 84 (1) (2011). doi:10.1103/physreve.84.016224.

[117] A. V. Getling, *Rayleigh-Bénard convection: structures and dynamics*, World Scientific, Singapore River Edge, NJ, 1998.

[118] H. Stanley, *Introduction to phase transitions and critical phenomena*, Oxford University Press, New York, 1987.

[119] B. Mandelbrot, *How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension*, Science 156 (3775) (1967). doi:10.1126/science.156.3775.636.

[120] B. T. Werner, *Complexity in Natural Landform Patterns*, Science 284 (5411) (1999). doi:10.1126/science.284.5411.102.

[121] G. F. Wiggs, *Desert dune processes and dynamics*, Progress in Physical Geography: Earth and Environment 25 (1) (2001). doi:10.1177/030913330102500103.

[122] R. Rak, J. Kwapień, P. Oświęcimka, P. Zięba, S. Drożdż, *Universal features of mountain ridge networks on Earth*, Journal of Complex Networks (2019). doi:10.1093/comnet/cnz017.

[123] K. Park, *The Internet as a Complex System*, in: K. Park, W. Willinger (Eds.), The Internet as a Large-Scale Complex System, Oxford University Press, 2005.

[124] W. Willinger, R. Govindan, S. Jamin, V. Paxson, S. Shenker, *Scaling phenomena in the Internet: Critically examining criticality*, Proceedings of the National Academy of Sciences 99 (Supplement 1) (2002). doi:10.1073/pnas.012583099.

[125] A.-L. Barabási, R. Albert, *Emergence of Scaling in Random Networks*, Science 286 (5439) (1999). doi:10.1126/science.286.5439.509.

[126] R. Albert, A.-L. Barabási, *Statistical mechanics of complex networks*, Reviews of Modern Physics 74 (1) (2002). doi:10.1103/revmodphys.74.47.

[127] P. Turchin, *Complex population dynamics: a theoretical/empirical synthesis*, Princeton University Press, Princeton, N.J, 2003.

[128] J. Hofbauer, K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, 1998. doi:10.1017/cbo9781139173179.

[129] D. S. Bassett, M. S. Gazzaniga, *Understanding complexity in the human brain*, Trends in Cognitive Sciences 15 (5) (2011). doi:10.1016/j.tics.2011.03.006.

[130] J. M. Beggs, D. Plenz, *Neuronal Avalanches in Neocortical Circuits*, The Journal of Neuroscience 23 (35) (2003). doi:10.1523/jneurosci.23-35-11167.2003.

[131] J. Beggs, D. Plenz, *Neuronal Avalanches Are Diverse and Precise Activity Patterns That Are Stable for Many Hours in Cortical Slice Cultures*, Journal of Neuroscience 24 (22) (2004). doi:10.1523/jneurosci.0540-04.2004.

[132] D. R. Chialvo, *Critical brain networks*, Physica A: Statistical Mechanics and its Applications 340 (4) (2004). doi:10.1016/j.physa.2004.05.064.

[133] R. Friedrich, A. Fuchs, H. Haken, *Spatio-Temporal EEG Patterns*, in: Springer Series in Synergetics, Springer Berlin Heidelberg, 1991. doi:10.1007/978-3-642-76877-4_22.

[134] O. Sporns, *Network analysis, complexity, and brain function*, Complexity 8 (1) (2002). doi:10.1002/cplx.10047.

[135] Q. K. Telesford, S. L. Simpson, J. H. Burdette, S. Hayasaka, P. J. Laurienti, *The Brain as a Complex System: Using Network Science as a Tool for Understanding the Brain*, Brain Connectivity 1 (4) (2011). doi:10.1089/brain.2011.0055.

[136] E. Bullmore, O. Sporns, *Complex brain networks: graph theoretical analysis of structural and functional systems*, Nature Reviews Neuroscience 10 (3) (2009). doi:10.1038/nrn2575.

[137] D. R. Chialvo, *Emergent complex neural dynamics*, Nature Physics 6 (10) (2010). doi:10.1038/nphys1803.

[138] D. Chialvo, *Life at the Edge: Complexity and Criticality in Biological Function*, Acta Physica Polonica B 49 (12) (2018). doi:10.5506/aphyspolb.49.1955.

[139] D. Sornette, *Critical market crashes*, Physics Reports 378 (1) (2003). doi:10.1016/s0370-1573(02)00634-8.

[140] J. C. Gerlach, G. Demos, D. Sornette, *Dissection of Bitcoin's multiscale bubble history from January 2012 to February 2018*, Royal Society Open Science 6 (7) (2019). doi:10.1098/rsos.180643.

[141] J.-P. Bouchaud, *Theory of financial risk and derivative pricing: from statistical physics to risk management*, Cambridge University Press, Cambridge, 2003.

[142] M. Wątorek, S. Drożdż, J. Kwapień, L. Minati, P. Oświęcimka, M. Stanuszek, *Multiscale characteristics of the emerging global cryptocurrency market*, Physics Reports 901 (2021). doi:10.1016/j.physrep.2020.10.005.

[143] I. Giardina, J.-P. Bouchaud, *Bubbles, crashes and intermittency in agent based market models*, The European Physical Journal B - Condensed Matter 31 (3) (2003). doi:10.1140/epjb/e2003-00050-6.

[144] S. Drożdż, J. Kwapień, P. Oświęcimka, T. Stanisz, M. Wątorek, *Complexity in Economic and Social Systems: Cryptocurrency Market at around COVID-19*, Entropy 22 (9) (2020). doi:10.3390/e22091043.

[145] R. Mantegna, *An introduction to econophysics: correlations and complexity in finance*, Cambridge University Press, Cambridge, UK New York, 2000.

[146] N. F. Johnson, P. Jefferies, P. M. Hui, *Financial Market Complexity: What Physics Can Tell Us about Market Behaviour*, Oxford University Press, 2003.

[147] P. Oświęcimka, J. Kwapień, S. Drożdż, *Multifractality in the stock market: price increments versus waiting times*, Physica A: Statistical Mechanics and its Applications 347 (2005). doi:10.1016/j.physa.2004.08.025.

[148] J. Kwapień, P. Oświęcimka, S. Drożdż, *Components of multifractality in high-frequency stock returns*, Physica A: Statistical Mechanics and its Applications 350 (2-4) (2005). doi:10.1016/j.physa.2004.11.019.

[149] J. Fan, J. Meng, J. Ludescher, X. Chen, Y. Ashkenazy, J. Kurths, S. Havlin, H. J. Schellnhuber, *Statistical physics approaches to the complex Earth system*, Physics Reports 896 (2021). doi:10.1016/j.physrep.2020.09.005.

[150] D. Rind, *Complexity and Climate*, Science 284 (5411) (1999). doi:10.1126/science.284.5411.105.

[151] S. Lovejoy, D. Schertzer, *The Weather and Climate: Emergent Laws and Multifractal Cascades*, Cambridge University Press, 2017.

[152] R. O. Weber, P. Talkner, *Spectra and correlations of climate data from days to decades*, Journal of Geophysical Research: Atmospheres 106 (D17) (2001). doi:10.1029/2001jd000548.

[153] N. Boers, J. Kurths, N. Marwan, *Complex systems approaches for Earth system data analysis*, Journal of Physics: Complexity 2 (1) (2021). doi:10.1088/2632-072x/abd8db.

[154] D. Zhou, A. Gozolchiani, Y. Ashkenazy, S. Havlin, *Teleconnection Paths via Climate Network Direct Link Detection*, Physical Review Letters 115 (26) (2015). doi:10.1103/physrevlett.115.268501.

[155] Z. Wang, M. A. Andrews, Z.-X. Wu, L. Wang, C. T. Bauch, *Coupled disease–behavior dynamics on complex networks: A review*, Physics of Life Reviews 15 (2015). doi:10.1016/j.plrev.2015.07.006.

[156] C. J. Rhodes, H. J. Jensen, R. M. Anderson, *On the critical behaviour of simple epidemics*, Proceedings of the Royal Society of London. Series B: Biological Sciences 264 (1388) (1997). doi:10.1098/rspb.1997.0228.

[157] M. J. Ferrari, S. Bansal, L. A. Meyers, O. N. Bjørnstad, *Network frailty and the geometry of herd immunity*, Proceedings of the Royal Society B: Biological Sciences 273 (1602) (2006). doi:10.1098/rspb.2006.3636.

[158] W. Cai, L. Chen, F. Ghanbarnejad, P. Grassberger, *Avalanche outbreaks emerging in cooperative contagions*, Nature Physics 11 (11) (2015). doi:10.1038/nphys3457.

[159] C. J. Rhodes, R. M. Anderson, *Power laws governing epidemics in isolated populations*, Nature 381 (6583) (1996). doi:10.1038/381600a0.

[160] R. Pastor-Satorras, C. Castellano, P. V. Mieghem, A. Vespignani, *Epidemic processes in complex networks*, Reviews of Modern Physics 87 (3) (2015). doi:10.1103/revmodphys.87.925.

[161] M. E. J. Newman, *Spread of epidemic disease on networks*, Physical Review E 66 (1) (2002). doi:10.1103/physreve.66.016128.

[162] R. Pastor-Satorras, A. Vespignani, *Epidemic Spreading in Scale-Free Networks*, Physical Review Letters 86 (14) (2001). doi:10.1103/physrevlett.86.3200.

[163] G. Palla, A.-L. Barabási, T. Vicsek, *Quantifying social group evolution*, Nature 446 (7136) (2007). doi:10.1038/nature05670.

[164] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, A. Arenas, *Self-similar community structure in a network of human interactions*, Physical Review E 68 (6) (2003). doi:10.1103/physreve.68.065103.

[165] R. M. Raafat, N. Chater, C. Frith, *Herding in humans*, Trends in Cognitive Sciences 13 (10) (2009). doi:10.1016/j.tics.2009.08.002.

[166] L. Zhao, G. Yang, W. Wang, Y. Chen, J. P. Huang, H. Ohashi, H. E. Stanley, *Herd behavior in a complex adaptive system*, Proceedings of the National Academy of Sciences 108 (37) (2011). doi:10.1073/pnas.1105239108.

[167] S. Drożdż, A. Kulig, J. Kwapień, A. Niewiarowski, M. Stanuszek, *Hierarchical organization of H. Eugene Stanley scientific collaboration community in weighted network representation*, Journal of Informetrics 11 (4) (2017). doi:10.1016/j.joi.2017.09.009.

[168] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, Y. Åberg, *The web of human sexual contacts*, Nature 411 (6840) (2001). doi:10.1038/35082140.

[169] T. C. Schelling, *Dynamic models of segregation*, The Journal of Mathematical Sociology 1 (2) (1971). doi:10.1080/0022250x.1971.9989794.

[170] A. Kolmogorov, *On tables of random numbers*, Sankhyā: The Indian Journal of Statistics, Series A (1961-2002) 25 (4) (1963).

[171] A. N. Kolmogorov, *Three approaches to the quantitative definition of information*, International Journal of Computer Mathematics 2 (1-4) (1968). doi:10.1080/00207166808803030.

[172] R. Solomonoff, *A formal theory of inductive inference. Part I*, Information and Control 7 (1) (1964). doi:10.1016/s0019-9958(64)90223-2.

[173] R. Solomonoff, *A formal theory of inductive inference. Part II*, Information and Control 7 (2) (1964). doi:10.1016/s0019-9958(64)90131-7.

[174] G. J. Chaitin, *On the Simplicity and Speed of Programs for Computing Infinite Sets of Natural Numbers*, Journal of the ACM 16 (3) (1969). doi:10.1145/321526.321530.

[175] P. M. Vitányi, *How Incomputable Is Kolmogorov Complexity?*, Entropy 22 (4) (2020). doi:10.3390/e22040408.

[176] M. Gell-Mann, S. Lloyd, *Information measures, effective complexity, and total information*, Complexity 2 (1) (1996). doi:10.1002/(sici)1099-0526(199609/10)2:1<44::aid-cplx10>3.0.co;2-x.

[177] M. Gell-Mann, S. Lloyd, *Effective Complexity*, in: Nonextensive Entropy, Oxford University Press, 2004. doi:10.1093/oso/9780195159769.003.0028.

[178] J. W. McAllister, *Effective Complexity as a Measure of Information Content*, Philosophy of Science 70 (2) (2003). doi:10.1086/375469.

[179] N. Ay, M. Muller, A. Szkola, *Effective Complexity and Its Relation to Logical Depth*, IEEE Transactions on Information Theory 56 (9) (2010). doi:10.1109/tit.2010.2053892.

[180] C. H. Bennett, *Logical depth and physical complexity*, in: The Universal Turing Machine: A Half-Century Survey, Oxford University Press, 1988.

[181] C. E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal 27 (3) (1948). doi:10.1002/j.1538-7305.1948.tb01338.x.

[182] C. E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal 27 (4) (1948). doi:10.1002/j.1538-7305.1948.tb00917.x.

[183] S. Lloyd, H. Pagels, *Complexity as thermodynamic depth*, Annals of Physics 188 (1) (1988). doi:10.1016/0003-4916(88)90094-2.

[184] J. P. Crutchfield, C. R. Shalizi, *Thermodynamic depth of causal states: Objective complexity via minimal representations*, Physical Review E 59 (1) (1999). doi:10.1103/physreve.59.275.

[185] A. Lempel, J. Ziv, *On the Complexity of Finite Sequences*, IEEE Transactions on Information Theory 22 (1) (1976). doi:10.1109/tit.1976.1055501.

[186] J. Ziv, A. Lempel, *A universal algorithm for sequential data compression*, IEEE Transactions on Information Theory 23 (3) (1977). doi:10.1109/tit.1977.1055714.

[187] J. Ziv, A. Lempel, *Compression of individual sequences via variable-rate coding*, IEEE Transactions on Information Theory 24 (5) (1978). doi:10.1109/tit.1978.1055934.

[188] Welch, *A Technique for High-Performance Data Compression*, Computer 17 (6) (1984). doi:10.1109/mc.1984.1659158.

[189] *Milestones: Lempel-Ziv Data Compression Algorithm, 1977. IEEE Global History Network*, http://www.ieeeghn.org/wiki/index.php/Milestones:Lempel-Ziv_Data_Compression_Algorithm,_1977, retrieved on 2021-09-20.

[190] I. Kontoyiannis, P. Algoet, Y. Suhov, A. Wyner, *Nonparametric entropy estimation for stationary processes and random fields, with applications to English text*, IEEE Transactions on Information Theory 44 (3) (1998). doi:10.1109/18.669425.

[191] Y. Gao, I. Kontoyiannis, E. Bienenstock, *Estimating the Entropy of Binary Time Series: Methodology, Some Theory and a Simulation Study*, Entropy 10 (2) (2008). doi:10.3390/entropy-e10020071.

[192] C. Bennett, *How to define complexity in physics, and why*, in: W. H. Zurek (Ed.), Complexity, Entropy And The Physics Of Information, CRC Press, 1990.

[193] B. Mandelbrot, *The fractal geometry of nature*, W.H. Freeman, San Francisco, 1982.

[194] J. Feder, *Fractals*, Springer US, 1988. doi:10.1007/978-1-4899-2124-6.

[195] T. Tél, *Fractals, Multifractals, and Thermodynamics*, Zeitschrift für Naturforschung A 43 (12) (1988). doi:10.1515/zna-1988-1221.

[196] H. E. Stanley, P. Meakin, *Multifractal phenomena in physics and chemistry*, Nature 335 (6189) (1988). doi:10.1038/335405a0.

[197] W. Kinsner, *System Complexity and Its Measures: How Complex Is Complex*, in: Studies in Computational Intelligence, Springer Berlin Heidelberg, 2010. doi:10.1007/978-3-642-16083-7_14.

[198] B. de Boer, *Self-organization and language evolution*, Oxford University Press, 2011. doi:10.1093/oxfordhb/9780199541119.013.0063.

[199] J. Liljencrants, B. Lindblom, *Numerical Simulation of Vowel Quality Systems: The Role of Perceptual Contrast*, Language 48 (4) (1972). doi:10.2307/411991.

[200] B. Lindblom, P. Macneilage, M. Studdert-Kennedy, *Self-organizing processes and the explanation of phonological universals*, in: Explanations for Language Universals, DE GRUYTER, 2014. doi:10.1515/9783110868555.181.

[201] J. Ke, M. Ogura, W. S.-Y. Wang, *Optimization Models of Sound Systems Using Genetic Algorithms*, Computational Linguistics 29 (1) (2003). doi:10.1162/089120103321337412.

[202] V. Kvasnicka, J. Pospichal, *An Emergence of Coordinated Communication in Populations of Agents*, Artificial Life 5 (4) (1999). doi:10.1162/106454699568809.

[203] A.-R. Berrah, R. Laboissière, *SPECIES: An evolutionary model for the emergence of phonetic structures in an artificial society of speech agents*, in: Advances in Artificial Life, Springer Berlin Heidelberg.

[204] L. Steels, *The Emergence of Grammar in Communicating Autonomous Robotic Agents*, in: Proceedings of the 14th European Conference on Artificial Intelligence, ECAI'00, IOS Press, Amsterdam, The Netherlands, The Netherlands, 2000.

[205] V. Schwämmle, P. M. C. de Oliveira, *A simple branching model that reproduces language family and language population distributions*, Physica A: Statistical Mechanics and its Applications 388 (14) (2009). doi:10.1016/j.physa.2009.03.038.

[206] M. Newman, *Power laws, Pareto distributions and Zipf's law*, Contemporary Physics 46 (5) (2005). doi:10.1080/00107510500052444.

[207] D. Marković, C. Gros, *Power laws and self-organized criticality in theory and nature*, Physics Reports 536 (2) (2014). doi:10.1016/j.physrep.2013.11.002.

[208] D. Sornette, *Probability Distributions in Complex Systems*, in: Computational Complexity, Springer New York, 2012. doi:10.1007/978-1-4614-1800-9_142.

[209] T. Apostol, *Mathematical analysis*, Addison-Wesley, Reading, Mass, 1974.

[210] C. Stutz, *On the Validity of Converting Sums to Integrals in Quantum Statistical Mechanics*, American Journal of Physics 36 (9) (1968). doi:10.1119/1.1975156.

[211] S. Foss, D. Korshunov, S. Zachary, *An Introduction to Heavy-Tailed and Subexponential Distributions*, Springer New York, 2013. doi:10.1007/978-1-4614-7101-1.

[212] P. Embrechts, *Modelling Extremal Events: for Insurance and Finance*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1997.

[213] A. Vezzani, E. Barkai, R. Burioni, *Single-big-jump principle in physical modeling*, Physical Review E 100 (1) (2019). doi:10.1103/physreve.100.012108.

[214] W. Wang, A. Vezzani, R. Burioni, E. Barkai, *Transport in disordered systems: The single big jump approach*, Physical Review Research 1 (3) (2019). doi:10.1103/physrevresearch.1.033172.

[215] V. Pareto, *Cours d'Économie Politique: Nouvelle édition*, Librairie Droz, 1964.

[216] M. Hardy, *Pareto's Law*, The Mathematical Intelligencer 32 (3) (2010). doi:10.1007/s00283-010-9159-2.

[217] A. Clauset, C. R. Shalizi, M. E. J. Newman, *Power-Law Distributions in Empirical Data*, SIAM Review 51 (4) (2009). doi:10.1137/070710111.

[218] J. C. Willis, G. U. Yule, *Some Statistics of Evolution and Geographical Distribution in Plants and Animals, and their Significance*, Nature 109 (2728) (1922). doi:10.1038/109177a0.

[219] G. U. Yule, *A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S*, Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character 213 (402-410) (1925). doi:10.1098/rstb.1925.0002.

[220] H. A. Simon, *On a class of skew distribution functions*, Biometrika 42 (3-4) (1955). doi:10.1093/biomet/42.3-4.425.

[221] D. D. S. Price, *A general theory of bibliometric and other cumulative advantage processes*, Journal of the American Society for Information Science 27 (5) (1976). doi:10.1002/asi.4630270505.

[222] R. K. Merton, *The Matthew Effect in Science: The reward and communication systems of science are considered*, Science 159 (3810) (1968). doi:10.1126/science.159.3810.56.

[223] M. Perc, *The Matthew effect in empirical data*, Journal of The Royal Society Interface 11 (98) (2014). doi:10.1098/rsif.2014.0378.

[224] P. Bak, C. Tang, K. Wiesenfeld, *Self-organized criticality: An explanation of the 1/fnoise*, Physical Review Letters 59 (4) (1987). doi:10.1103/physrevlett.59.381.

[225] P. Bak, C. Tang, K. Wiesenfeld, *Self-organized criticality*, Physical Review A 38 (1) (1988). doi:10.1103/physreva.38.364.

[226] H. J. Jensen, *Self-Organized Criticality*, Cambridge University Press, 2000.

[227] D. L. Turcotte, *Self-organized criticality*, Reports on Progress in Physics 62 (10) (1999). doi:10.1088/0034-4885/62/10/201.

[228] N. W. Watkins, G. Pruessner, S. C. Chapman, N. B. Crosby, H. J. Jensen, *25 Years of Self-organized Criticality: Concepts and Controversies*, Space Science Reviews 198 (1-4) (2015). doi:10.1007/s11214-015-0155-x.

[229] D. Dhar, *Self-organized critical state of sandpile automaton models*, Physical Review Letters 64 (14) (1990). doi:10.1103/physrevlett.64.1613.

[230] D. Dhar, *Theoretical studies of self-organized criticality*, Physica A: Statistical Mechanics and its Applications 369 (1) (2006). doi:10.1016/j.physa.2006.04.004.

[231] H. J. Jensen, K. Christensen, H. C. Fogedby, *1/f noise, distribution of lifetimes, and a pile of sand*, Physical Review B 40 (10) (1989). doi:10.1103/physrevb.40.7425.

[232] J. Kertesz, L. B. Kiss, *The noise spectrum in the model of self-organised criticality*, Journal of Physics A: Mathematical and General 23 (9) (1990). doi:10.1088/0305-4470/23/9/006.

[233] K. Christensen, H. C. Fogedby, H. J. Jensen, *Dynamical and spatial aspects of sandpile cellular automata*, Journal of Statistical Physics 63 (3-4) (1991). doi:10.1007/bf01029204.

[234] H. J. Jensen, *1/f noise from the linear diffusion equation*, Physica Scripta 43 (6) (1991). doi:10.1088/0031-8949/43/6/009.

[235] P. D. L. Rios, Y.-C. Zhang, *Universal 1/f Noise from Dissipative Self-Organized Criticality Models*, Physical Review Letters 82 (3) (1999). doi:10.1103/physrevlett.82.472.

[236] C. D. Manning, H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, USA, 1999.

[237] R. Ferrer-i Cancho, R. V. Solé, *Zipf's law and random texts*, Advances in Complex Systems 05 (01) (2002). doi:10.1142/s0219525902000468.

[238] B. Manaris, L. Pellicoro, G. Pothering, H. Hodges, *Investigating Esperanto's Statistical Proportions Relative to Other Languages Using Neural Networks and Zipf's Law*, in: Proceedings of the 24th IASTED International Conference on Artificial Intelligence and Applications, AIA'06, ACTA Press, Anaheim, CA, USA, 2006.

[239] R. D. Smith, *Investigation of the Zipf-plot of the extinct Meroitic language*, Glottometrics 15 (2007).

[240] J. Clark, K. Lua, J. McCallum, *Conformance of Chinese text to Zipf's law*, in: Proceedings. PARBASE-90: International Conference on Databases, Parallel Architectures, and Their Applications, IEEE Comput. Soc. Press, 1990. doi:10.1109/parbse.1990.77200.

[241] S. Shtrikman, *Some comments on Zipf's law for the Chinese language*, Journal of Information Science 20 (2) (1994). doi:10.1177/016555159402000208.

[242] R. Ferrer-i Cancho, *The variation of Zipf's law in human language*, The European Physical Journal B 44 (2) (2005). doi:10.1140/epjb/e2005-00121-8.

[243] S. Havlin, *The distance between Zipf plots*, Physica A: Statistical Mechanics and its Applications 216 (1-2) (1995). doi:10.1016/0378-4371(95)00069-j.

[244] W. Deng, R. Xie, S. Deng, A. E. Allahverdyan, *Two halves of a meaningful text are statistically different*, Journal of Statistical Mechanics: Theory and Experiment 2021 (3) (2021). doi:10.1088/1742-5468/abe947.

[245] L. Lü, Z.-K. Zhang, T. Zhou, *Zipf's Law Leads to Heaps' Law: Analyzing Their Relation in Finite-Size Systems*, PLoS ONE 5 (12) (2010). doi:10.1371/journal.pone.0014139.

[246] D. van Leijenhorst, T. van der Weide, *A formal derivation of Heaps' Law*, Information Sciences 170 (2-4) (2005). doi:10.1016/j.ins.2004.03.006.

[247] A. Kornai, *Zipf's law outside the middle range*, in: Proc. Sixth Meeting on mathematics of language. (MOL 6), 1999.

[248] 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) (2009). doi:10.1080/09296170902850358.

[249] G. A. Miller, *Some Effects of Intermittent Silence*, The American Journal of Psychology 70 (2) (1957). doi:10.2307/1419346.

[250] M. Simkin, V. Roychowdhury, *Re-inventing Willis*, Physics Reports (2010). doi:10.1016/j.physrep.2010.12.004.

[251] 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) (2001). doi:10.1076/jqul.8.3.165.4101.

[252] M. A. Montemurro, *Beyond the Zipf–Mandelbrot law in quantitative linguistics*, Physica A: Statistical Mechanics and its Applications 300 (3-4) (2001). doi:10.1016/s0378-4371(01)00355-7.

[253] Á. Corral, G. Boleda, R. F. i Cancho, *Zipf's Law for Word Frequencies: Word Forms versus Lemmas in Long Texts*, PLOS ONE 10 (7) (2015). doi:10.1371/journal.pone.0129031.

[254] A. Kulig, J. Kwapień, T. Stanisz, S. Drożdż, *In narrative texts punctuation marks obey the same statistics as words*, Information Sciences 375 (2017). doi:10.1016/j.ins.2016.09.051.

[255] P. Stoica, R. Moses, *Spectral analysis of signals*, Pearson/Prentice Hall, Upper Saddle River, N.J, 2005.

[256] J. Beran, *Statistics for long-memory processes*, Chapman & Hall, New York, 1994.

[257] J. Gao, J. Hu, W.-W. Tung, Y. Cao, N. Sarshar, V. P. Roychowdhury, *Assessment of long-range correlation in time series: How to avoid pitfalls*, Physical Review E 73 (1) (2006). doi:10.1103/physreve.73.016117.

[258] G. Rangarajan, M. Ding, *Integrated approach to the assessment of long range correlation in time series data*, Physical Review E 61 (5) (2000). doi:10.1103/physreve.61.4991.

[259] B. B. Mandelbrot, J. W. V. Ness, *Fractional Brownian Motions, Fractional Noises and Applications*, SIAM Review 10 (4) (1968). doi:10.1137/1010093.

[260] F. J. Molz, H. H. Liu, J. Szulga, *Fractional Brownian motion and fractional Gaussian noise in subsurface hydrology: A review, presentation of fundamental properties, and extensions*, Water Resources Research 33 (10) (1997). doi:10.1029/97wr01982.

[261] C. Heneghan, G. McDarby, *Establishing the relation between detrended fluctuation analysis and power spectral density analysis for stochastic processes*, Physical Review E 62 (5) (2000). doi:10.1103/physreve.62.6103.

[262] D. Delignières, *Correlation Properties of (Discrete) Fractional Gaussian Noise and Fractional Brownian Motion*, Mathematical Problems in Engineering 2015 (2015). doi:10.1155/2015/485623.

[263] R. Fossion, E. Landa, P. Stránský, V. Velázquez, J. C. L. Vieyra, I. Garduño, D. García, A. Frank, *Scale invariance as a symmetry in physical and biological systems: listening to photons, bubbles and heartbeats*, AIP Conference Proceedings 1323 (1) (2010). doi:10.1063/1.3537868.

[264] R. F. Voss, *Random fractals: Self-affinity in noise, music, mountains, and clouds*, Physica D: Nonlinear Phenomena 38 (1-3) (1989). doi:10.1016/0167-2789(89)90220-0.

[265] K. Falconer, *Fractal geometry: mathematical foundations and applications*, John Wiley & Sons Inc, Hoboken, 2014.

[266] J. F. Muzy, E. Bacry, A. Arneodo, *The multifracal formalism revisited with wavelets*, International Journal of Bifurcation and Chaos 04 (02) (1994). doi:10.1142/s0218127494000204.

[267] T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, B. I. Shraiman, *Fractal measures and their singularities: The characterization of strange sets*, Physical Review A 33 (2) (1986). doi:10.1103/physreva.33.1141.

[268] B. B. Mandelbrot, *An Introduction to Multifractal Distribution Functions*, in: Random Fluctuations and Pattern Growth: Experiments and Models, Springer Netherlands, 1988. doi:10.1007/978-94-009-2653-0_40.

[269] B. B. Mandelbrot, C. J. G. Evertsz, Y. Hayakawa, *Exactly self-similar left-sided multifractal measures*, Physical Review A 42 (8) (1990). doi:10.1103/physreva.42.4528.

[270] B. B. Mandelbrot, *New anomalous multiplicative multifractals: Left sided $f(\alpha)$ and the modelling of DLA*, Physica A: Statistical Mechanics and its Applications 168 (1) (1990). doi:10.1016/0378-4371(90)90361-u.

[271] R. Riedi, B. Mandelbrot, *Multifractal Formalism for Infinite Multinomial Measures*, Advances in Applied Mathematics 16 (2) (1995). doi:10.1006/aama.1995.1007.

[272] R. H. Riedi, *An introduction to multifractals*, in: Rice University ECE Technical Report, 1997.

[273] S. Drożdż, J. Kwapień, P. Oświecimka, R. Rak, *Quantitative features of multifractal subtleties in time series*, EPL (Europhysics Letters) 88 (6) (2009). doi:10.1209/0295-5075/88/60003.

[274] K. Matia, Y. Ashkenazy, H. E. Stanley, *Multifractal properties of price fluctuations of stocks and commodities*, Europhysics Letters (EPL) 61 (3) (2003). doi:10.1209/epl/i2003-00194-y.

[275] P. Norouzzadeh, B. Rahmani, *A multifractal detrended fluctuation description of Iranian rial–US dollar exchange rate*, Physica A: Statistical Mechanics and its Applications 367 (2006). doi:10.1016/j.physa.2005.11.019.

[276] B. B. Mandelbrot, *Self-Affine Fractals and Fractal Dimension*, Physica Scripta 32 (4) (1985). doi:10.1088/0031-8949/32/4/001.

[277] J. W. Kantelhardt, S. A. Zschiegner, E. Koscielny-Bunde, S. Havlin, A. Bunde, H. Stanley, *Multifractal detrended fluctuation analysis of nonstationary time series*, Physica A: Statistical Mechanics and its Applications 316 (1-4) (2002). doi:10.1016/s0378-4371(02)01383-3.

[278] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberger, *Mosaic organization of DNA nucleotides*, Physical Review E 49 (2) (1994). doi:10.1103/physreve.49.1685.

[279] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. Rego, S. Havlin, A. Bunde, *Detecting long-range correlations with detrended fluctuation analysis*, Physica A: Statistical Mechanics and its Applications 295 (3-4) (2001). doi:10.1016/s0378-4371(01)00144-3.

[280] M. A. Montemurro, D. H. Zanette, *Universal Entropy of Word Ordering Across Linguistic Families*, PLoS ONE 6 (5) (2011). doi:10.1371/journal.pone.0019875.

[281] M. A. Montemurro, *Quantifying the information in the long-range order of words: Semantic structures and universal linguistic constraints*, Cortex 55 (2014). doi:10.1016/j.cortex.2013.08.008.

[282] M. A. Montemurro, P. A. Pury, *Long-range Fractal Correlations in Literary Corpora*, Fractals 10 (04) (2002). doi:10.1142/s0218348x02001257.

[283] 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) (2012). doi:10.1103/physreve.86.031108.

[284] E. G. Altmann, J. B. Pierrehumbert, A. E. Motter, *Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words*, PLoS ONE 4 (11) (2009). doi:10.1371/journal.pone.0007678.

[285] S. Drożdż, P. Oświęcimka, A. Kulig, J. Kwapień, K. Bazarnik, I. Grabska-Gradzińska, J. Rybicki, M. Stanuszek, *Quantifying origin and character of long-range correlations in narrative texts*, Information Sciences 331 (2016). doi:10.1016/j.ins.2015.10.023.

[286] R. Futrell, K. Mahowald, E. Gibson, *Large-scale evidence of dependency length minimization in 37 languages*, Proceedings of the National Academy of Sciences 112 (33) (2015). doi:10.1073/pnas.1502134112.

[287] H. Liu, C. Xu, J. Liang, *Dependency distance: A new perspective on syntactic patterns in natural languages*, Physics of Life Reviews 21 (2017). doi:10.1016/j.plrev.2017.03.002.

[288] M. B. Parkes, *Pause and effect: an introduction to the history of punctuation in the West*, University of California Press, Berkeley, 1993.

[289] T. Nakagawa, S. Osaki, *The Discrete Weibull Distribution*, IEEE Transactions on Reliability R-24 (5) (1975). doi:10.1109/tr.1975.5214915.

[290] N. L. Johnson, S. Kotz, N. Balakrishnan, *Continuous Univariate Distributions*, Wiley-Interscience, 1994.

[291] M. E. J. Newman, *Networks: an introduction*, Oxford University Press, Oxford New York, 2010.

[292] S. N. Dorogovtsev, J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW*, Oxford University Press, 2003.

[293] E. Estrada, *The Structure of Complex Networks: Theory and Applications*, Oxford University Press, 2011.

[294] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D. Hwang, *Complex networks: Structure and dynamics*, Physics Reports 424 (4-5) (2006). doi:10.1016/j.physrep.2005.10.009.

[295] M. E. J. Newman, *The Structure and Function of Complex Networks*, SIAM Review 45 (2) (2003). doi:10.1137/s003614450342480.

[296] A. Fornito, A. Zalesky, E. Bullmore, *Fundamentals of brain network analysis*, Elsevier, Amsterdam, 2016.

[297] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, D. B. Chklovskii, *Structural Properties of the Caenorhabditis elegans Neuronal Network*, PLoS Computational Biology 7 (2) (2011). doi:10.1371/journal.pcbi.1001066.

[298] M. Valencia, M. A. Pastor, M. A. Fernández-Seara, J. Artieda, J. Martinerie, M. Chavez, *Complex modular structure of large-scale brain networks*, Chaos: An Interdisciplinary Journal of Nonlinear Science 19 (2) (2009). doi:10.1063/1.3129783.

[299] J.-P. Onnela, J. Saramäki, K. Kaski, J. Kertész, *Financial Market - A Network Perspective*, in: Practical Fruits of Econophysics, Springer-Verlag, 2006. doi:10.1007/4-431-28915-1_55.

[300] M. Faloutsos, P. Faloutsos, C. Faloutsos, *On power-law relationships of the Internet topology*, in: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication - SIGCOMM'99, ACM Press, 1999. doi:10.1145/316188.316229.

[301] J. Sienkiewicz, J. A. Hołyst, *Statistical analysis of 22 public transport networks in Poland*, Physical Review E 72 (4) (2005). doi:10.1103/physreve.72.046127.

[302] A. Barrat, M. Barthélemy, R. Pastor-Satorras, A. Vespignani, *The architecture of complex weighted networks*, Proceedings of the National Academy of Sciences 101 (11) (2004). doi:10.1073/pnas.0400087101.

[303] C. Leung, H. Chau, *Weighted assortative and disassortative networks model*, Physica A: Statistical Mechanics and its Applications 378 (2) (2007). doi:10.1016/j.physa.2006.12.022.

[304] V. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics: Theory and Experiment 2008 (10) (2008). doi:10.1088/1742-5468/2008/10/p10008.

[305] R. van der Hofstad, *Random Graphs and Complex Networks*, Cambridge University Press, 2016. doi:10.1017/9781316779422.

[306] P. Erdős, A. Rényi, *On random graphs*, Publicationes Mathematicae 6 (1959).

[307] P. Erdős, A. Rényi, *On the Evolution of Random Graphs*, in: Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 1960.

[308] E. N. Gilbert, *Random Graphs*, The Annals of Mathematical Statistics 30 (4) (1959). doi:10.1214/aoms/1177706098.

[309] F. Viger, M. Latapy, *Efficient and Simple Generation of Random Simple Connected Graphs with Prescribed Degree Sequence*, in: Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2005. doi:10.1007/11533719_45.

[310] R. C. Prim, *Shortest Connection Networks And Some Generalizations*, Bell Syst. Tech. J. 36 (6) (1957). doi:10.1002/j.1538-7305.1957.tb01515.x.

[311] R. E. Tarjan, *Data Structures and Network Algorithms*, Cambridge, 1987.

[312] B. Peng, L. Zhang, D. Zhang, *A survey of graph theoretical approaches to image segmentation*, Pattern Recognition 46 (3) (2013). doi:10.1016/j.patcog.2012.09.015.

[313] P. Tewarie, E. van Dellen, A. Hillebrand, C. Stam, *The minimum spanning tree: An unbiased method for brain network analysis*, NeuroImage 104 (2015). doi:10.1016/j.neuroimage.2014.10.015.

[314] D.-H. Kim, J. D. Noh, H. Jeong, *Scale-free trees: The skeletons of complex networks*, Physical Review E 70 (4) (2004). doi:10.1103/physreve.70.046126.

[315] C. Song, S. Havlin, H. A. Makse, *Self-similarity of complex networks*, Nature 433 (7024) (2005). doi:10.1038/nature03248.

[316] C. Song, L. K. Gallos, S. Havlin, H. A. Makse, *How to calculate the fractal dimension of a complex network: the box covering algorithm*, Journal of Statistical Mechanics: Theory and Experiment 2007 (03) (2007). doi:10.1088/1742-5468/2007/03/p03006.

[317] C. Song, S. Havlin, H. A. Makse, *Origins of fractality in the growth of complex networks*, Nature Physics 2 (4) (2006). doi:10.1038/nphys266.

[318] K.-I. Goh, G. Salvi, B. Kahng, D. Kim, *Skeleton and Fractal Scaling in Complex Networks*, Physical Review Letters 96 (1) (2006). doi:10.1103/physrevlett.96.018701.

[319] X. Zhang, J. Zhu, *Skeleton of weighted social network*, Physica A: Statistical Mechanics and its Applications 392 (6) (2013). doi:10.1016/j.physa.2012.12.001.

[320] L. Antiqueira, O. N. Oliveira, L. da Fontoura Costa, M. das Graças Volpe Nunes, *A complex network approach to text summarization*, Information Sciences 179 (5) (2009). doi:10.1016/j.ins.2008.10.032.

[321] R. Navigli, M. Lapata, *An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation*, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (4) (2010). doi:10.1109/tpami.2009.36.

[322] D. Amancio, M. Nunes, O. Oliveira, T. Pardo, L. Antiqueira, L. da F. Costa, *Using metrics from complex networks to evaluate machine translation*, Physica A: Statistical Mechanics and its Applications 390 (1) (2011). doi:10.1016/j.physa.2010.08.052.

[323] R. F. Mihalcea, D. R. Radev, *Graph-based Natural Language Processing and Information Retrieval*, 1st Edition, Cambridge University Press, New York, NY, USA, 2011. doi:10.1017/cbo9780511976247.

[324] L. Dall'Asta, A. Baronchelli, *Microscopic activity patterns in the naming game*, Journal of Physics A: Mathematical and General 39 (48) (2006). doi:10.1088/0305-4470/39/48/002.

[325] A. Kalampokis, K. Kosmidis, P. Argyrakis, *Evolution of vocabulary on scale-free and random networks*, Physica A: Statistical Mechanics and its Applications 379 (2) (2007). doi:10.1016/j.physa.2006.12.048.

[326] L. Q. Ha, P. Hanna, J. Ming, F. J. Smith, *Extending Zipf's law to n-grams for large corpora*, Artificial Intelligence Review 32 (1-4) (2009). doi:10.1007/s10462-009-9135-4.

[327] J. R. Williams, P. R. Lessard, S. Desu, E. M. Clark, J. P. Bagrow, C. M. Danforth, P. S. Dodds, *Zipf's law holds for phrases, not words*, Scientific Reports 5 (1) (2015). doi:10.1038/srep12209.

[328] L. Egghe, *On the law of Zipf-Mandelbrot for multi-word phrases*, Journal of the American Society for Information Science 50 (3) (1999). doi:10.1002/(sici)1097-4571(1999)50:3<233::aid-asi6>3.0.co;2-8.

[329] L. Egghe, *The Distribution of N-Grams*, Scientometrics 47 (2) (2000). doi:10.1023/a:1005634925734.

[330] P.-N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Pearson, 2005.

[331] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, Wiley John + Sons, 2000.

[332] H. Liu, W. Li, *Language clusters based on linguistic complex networks*, Chinese Science Bulletin 55 (30) (2010). doi:10.1007/s11434-010-4114-3.

[333] T. Stanisz, J. Kwapień, S. Drożdż, *Linguistic data mining with complex networks: A stylometric-oriented approach*, Information Sciences 482 (2019). doi:10.1016/j.ins.2019.01.040.

[334] C. D. Sutton, *Classification and Regression Trees, Bagging, and Boosting*, in: Handbook of Statistics, Elsevier, 2005. doi:10.1016/s0169-7161(04)24011-1.

[335] M. Koppel, J. Schler, S. Argamon, *Computational methods in authorship attribution*, Journal of the American Society for Information Science and Technology 60 (1) (2009). doi:10.1002/asi.20961.

[336] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, D. Woodard, *Surveying Stylometry Techniques and Applications*, ACM Computing Surveys 50 (6) (2017). doi:10.1145/3132039.

[337] O. de Vel, A. Anderson, M. Corney, G. Mohay, *Mining e-mail content for author identification forensics*, ACM SIGMOD Record 30 (4) (2001). doi:10.1145/604264.604272.

[338] D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, L. Ye, *Author Identification on the Large Scale*, in: In Proc. of the Meeting of the Classification Society of North America, 2005.

[339] R. Zheng, J. Li, H. Chen, Z. Huang, *A framework for authorship identification of online messages: Writing-style features and classification techniques*, Journal of the American Society for Information Science and Technology 57 (3) (2006). doi:10.1002/asi.20316.

[340] G. A. Miller, *WordNet*, Communications of the ACM 38 (11) (1995). doi:10.1145/219717.219748.

[341] C. Fellbaum, *WordNet: an electronic lexical database*, MIT Press, Cambridge, Mass, 1998.

[342] C. Fellbaum, *WordNet: An Electronic Lexical Resource*, in: S. E. F. Chipman (Ed.), The Oxford Handbook of Cognitive Science, Oxford University Press, 2015. doi:10.1093/oxfordhb/9780199842193.013.001.

[343] J. Morato, M. A. Marzal, J. Lloréns, J. Moreiro, *Wordnet applications*, in: Proceedings of GWC, 2004.

[344] C. Fellbaum, *WordNet*, in: Theory and Applications of Ontology: Computer Applications, Springer Netherlands, 2010. doi:10.1007/978-90-481-8847-5_10.

[345] M. Steyvers, J. B. Tenenbaum, *The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth*, Cognitive Science 29 (1) (2005). doi:10.1207/s15516709cog2901_3.

[346] S. D. Deyne, G. Storms, *Word Associations*, in: J. R. Taylor (Ed.), The Oxford Handbook of the Word, Oxford University Press, 2014. doi:10.1093/oxfordhb/9780199641604.013.018.

[347] D. L. Nelson, N. Zhang, V. M. McKinney, *The ties that bind what is known to the recognition of what is new.*, Journal of Experimental Psychology: Learning, Memory, and Cognition 27 (5) (2001). doi:10.1037/0278-7393.27.5.1147.

[348] O. Valba, A. Gorsky, S. Nechaev, M. Tamm, *Analysis of English free association network reveals mechanisms of efficient solution of Remote Association Tests*, PLOS ONE 16 (4) (2021). doi:10.1371/journal.pone.0248986.

[349] S. De Deyne, Y. Kenett, D. Anaki, M. Faust, D. Navarro, *Large-scale network representations of semantics in the mental lexicon*, in: M. N. Jones (Ed.), Big Data in Cognitive Science: From Methods to Insights, Routledge/Taylor & Francis Group, 2016.

[350] S. De Deyne, D. Navarro, G. Storms, *Associative strength and semantic activation in the mental lexicon: evidence from continued word associations*, in: Proceedings of the Annual Meeting of the Cognitive Science Society, 2013.

[351] S. D. Deyne, G. Storms, *Word associations: Network and semantic properties*, Behavior Research Methods 40 (1) (2008). doi:10.3758/brm.40.1.213.

[352] D. L. Nelson, C. L. McEvoy, T. A. Schreiber, *The University of South Florida free association, rhyme, and word fragment norms*, Behavior Research Methods, Instruments, & Computers 36 (3) (2004). doi:10.3758/bf03195588.

[353] G. R. Kiss, C. Armstrong, R. Milroy, J. Piper, *An associative thesaurus of English and its computer analysis*, in: A. J. Aitken, R. W. Bailey, N. Hamilton-Smith (Eds.), The Computer and Literary Studies, Edinburgh University Press, 1973.

[354] M. Coltheart, *The MRC Psycholinguistic Database*, The Quarterly Journal of Experimental Psychology Section A 33 (4) (1981). doi:10.1080/14640748108400805.

[355] M. Wilson, G. Kiss, C. Armstrong, *EAT: the Edinburgh associative corpus*, oxford Text Archive; http://hdl.handle.net/20.500.12024/1251.

[356] V. Batagelj, A. Mrvar, *Pajek datasets*, http://vlado.fmf.uni-lj.si/pub/networks/data/ (retrieved on 2022-02-13); data available under CC BY-NC-SA 2.5 license (https://creativecommons.org/licenses/by-nc-sa/2.5/) (2006).

[357] S. Arlot, A. Celisse, *A survey of cross-validation procedures for model selection*, Statistics Surveys 4 (2010). doi:10.1214/09-ss054.