Florentina Armaselu

Text, Fractal Dust and Informational Granularity: A Study of Scale

Abstract: This chapter proposes a method of text analysis that combines conceptual aspects from the model of scalable or zoomable text (z-text), topic modelling and fractal geometry. It argues that this type of methodology may assist in detecting different levels of generality and specificity in texts and reveal some characteristics of the assemblage of blocks of text, above the word level, at different scales of representation. Applications of such an approach can range from hermeneutics and discourse analysis to text (and possibly z-text) generation and summarization.

Keywords: scalable text, topic modelling, fractal geometry, informational granularity, digital hermeneutics

1 Introduction

Scale in text analysis has often been considered in relation to big collections of data and the possibility offered by digital methods and tools to provide insights into patterns, trends, outliers and linguistic phenomena that are hard to detect and cover by human reading alone. Several terms, such as *distant reading*, sometimes opposed or compared to *close reading* (Moretti 2013; Underwood 2019), *scalable reading* (Mueller 2014), *macroscope* (Hitchcock 2014) and *long zoom* (Johnson 2007), have been coined to define this type of approach that allows for shifts from a bird's eye view to individual details. However, what seems to have been less studied so far is the significance of the concept of scale and its possible applications as an inherent feature of text itself. Under the magnifying glass, a text is far from being a flat conceptual structure; it may reveal a stratified organization with different layers of general and specific, abstract and concrete, simple and complex units of meaning.

This chapter will focus on scale in textual forms, starting from the assumption that a text can be conceived as a scalable construct containing different levels of detail and can be explored by zooming in and zooming out (Armaselu 2010;

Acknowledgements: The author would like to thank the C^{2}DH and Andreas Fickers for the thinkering grant that made this project possible, Dario Rodighiero, from the University of Groningen, for his review and suggestions of improvement, and Sarah Cooper, from the Language Centre of the University of Luxembourg, for English proofreading.

https://doi.org/10.1515/9783111317779-012

Armaselu and Van den Heuvel 2017). It will investigate the possibility of computer-based detection and analysis of scale-related structures in text, as well as the potential meaning attached to these scales and forms of interpretation.

The analysis will combine topic modelling (Blei 2011), for the preparation of the data, with the theory of fractals that is known for its applications in various domains, from mathematics and physics to statistical economics and linguistics. In his book *The Fractal Geometry of Nature*, Mandelbrot (1983) describes the concept of *fractal*, derived from the Latin *fractus, frangere* (to break), and its use in modelling highly irregular and complex forms from nature such as coastlines, clouds, mountains and trees, whose study goes beyond standard Euclidean geometry and dimensions. One of the fractal forms utilized as a model of a coastline at various scales is the Koch curve (Figure 1). The process of generating such forms starts with a straight interval called the *initiator*. A second approximation replaces the straight line with a broken line formed of "four intervals of equal length", called the *generator.* New details such as promontories and subpromontories appear through the iterative replacement of the generator's four intervals by a reduced generator. Although its irregularity is too systematic, the Koch curve is considered to be a "suggestive approximation" of a coastline (Mandelbrot 1983: 34–35, 42–45).

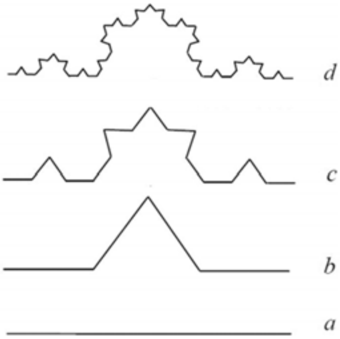


Figure 1: The first four iterations in building a Koch curve (adaptation of Sapoval 1989, 19): a – first iteration, *initiator*; b – second iteration, *generator*; c, d – third and fourth iteration.

Applications of fractal theory in text analysis have exploited various aspects of the concept of *scaling* and *self-similarity*. In scaling systems, a part is in "some way a reduced-scale version of the whole", and when "each piece of a shape is geometrically similar to a whole, both the shape and the cascade that generate it are called *self-similar*" (Mandelbrot 1983: 345, 34). Elaborating on Zipf's law (Zipf 2012: 23–27), stating the relationship between word frequencies and their ranks as evidence of vocabulary balance in a text, Mandelbrot illustrated how constructs such as "scaling lexicographical trees" provide generalized proofs of the Zipf law while exhibiting scaling properties and fractal, non-integer dimensions (Mandelbrot 1983: 346). Other studies have applied fractal-based methodologies to automatic keyword extraction by assessing the “degree of fractality” of words as a measure of their relevance and non-uniform versus uniform distribution in texts (Najafi and Darooneh 2015). Fractal indicators, computed by considering the sequence of words and the number of letters in these words, have been used to compare the style of different types of texts, such as scientific, journalistic, conversational, epistolary and poetic (Kaminskiy et al. 2021). Fractal analysis has also been employed to determine the optimal number of topics based on the detection of “self-similar fractal regions” using a “density-of-states function” for texts in different languages (Ignatenko et al. 2019). More theoretical approaches have conceptualized language as a system that displays fractal features, such as “structural autosimilarity”, “fractal dimension” and “iterative order” in generating linguistic structures (Pareyon 2007) or “self-similar patterns” of discourse through the “process of identifying recursive semantic components” (Tacenko 2016).

Although a variety of methods for the fractal-based processing or conceptualization of language have been proposed, mainly taking into account the composition of texts as compounds including letters, syllables, words and sentences, the scalable nature of text and its stratified structure from a conceptual perspective has been less studied so far. In this chapter, I propose an approach that combines topic modelling techniques and fractal theory-related measures to detect different layers of generality and specificity and analyze a text at different scales. For this purpose, I use a corpus of texts from historiography, literature and philosophy. Section 2 will describe the initial assumptions about text scalability and the data used to test and assess the methods illustrated in Sections 3 and 4. Section 5 will present the results and possible interpretations of the approach, while Section 6 will summarize the findings and propose hypotheses for future work.

2 Datasets

A particular area of research in digital history and humanities, that of global microhistory (Trivellato 2011), has presented interest for the study. The dataset used in the experiments contains books considered representative for the objective of this type of research: *1688. A Global History* (Wills 2001); *Plumes* (Stein 2008); *The Inner Life of Empires* (Rothschild 2011); *The Two Princes of Calabar* (Sparks 2004); and *Vermeer’s Hat* (Brook 2009). These books combine methods of analysis spanning various conceptual levels, from micro to macro perspectives on the investigated historical phenomena, by connecting, for example: a series of paintings and art objects with the growth of trade and exploration in the seventeenth century (Brook 2009); micro- and macro-histories through the history of a family's own connections (Rothschild 2011); micro-historical accounts with the history of enslaved Africans in the early modern Atlantic world (Sparks 2004); or the perspective of particular actors (people, commodities, one year in time) with the history of specific groups and cultures (Stein 2008; Wills 2001). For comparison purposes, one literary and one philosophical text were included in the dataset, *Gulliver's Travels* (Swift 2009) and *Beyond Good and Evil* (Nietzsche 2009), available via Project Gutenberg (Hart 2004). The size of the corpus was relatively small to allow for closer analysis of the methodology as a proof of concept. The main question to address was to what extent the applied digital methods were able to detect various conceptual levels in the studied texts. The historiography group of books was presumed to already possess such a variety given their analytical coverage ranging from broad overviews to detailed examination in their unfolding of arguments related to world history and microhistory. It was expected that the two other books, from literature and philosophy, would contain a certain type of stratification as well, as an inherent structure of text itself that would be revealed by the analysis.

The books were divided into separate text files corresponding to chapters or parts (when chapters were too short), deemed as meaningful units of analysis for the exploration of scale in text. It was assumed that chapters and parts preserve a certain coherence and similarity in terms of varying degrees of generality and specificity in disclosing the content of the book. Figure 2 illustrates the structure of a book containing topics grouped on levels: from more general, representing a larger number of units, to more particular, mostly characterizing a single unit.

Preliminary experiments consisted in reorganizing excerpts from the books as *zoomable* texts or *z-texts*[1] using a dedicated interface, *z-editor*[2] (Armaselu 2010). The z-editor allows the user to start with a sequence of *z-lexias*[3] on the surface level and to expand or explore them by zooming in and out along the Z-axis and adding or revealing details that belong to deeper levels. I referred to the corresponding processes of expansion and exploration of z-lexias by zooming in and out as *z-writing* and *z-reading*. Each level of the structure corresponds to an XML-TEI file that stores the content and relations of parent and children z-lexias. For the author or reader of a z-text, the inner XML mechanism of the interface is transparent.

[1] Accessed July 23, 2023. http://www.zoomimagine.com/AboutProject.html.

[2] Accessed July 27, 2023. http://www.zoomimagine.com/ZEditor.html.

[3] Fragments of texts as units in the writing or reading process, inspired by Barthes's (1974: 13) *lexias*, "units of reading".

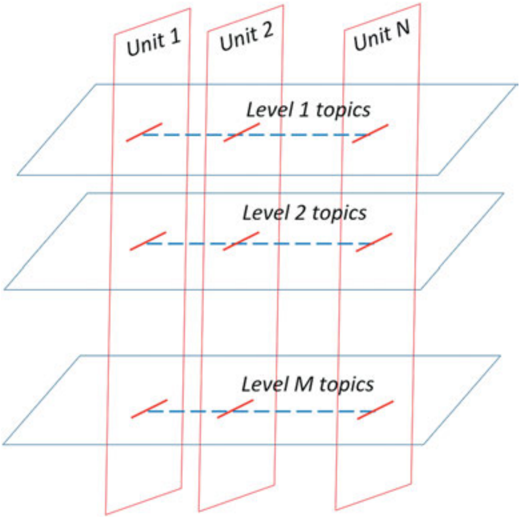
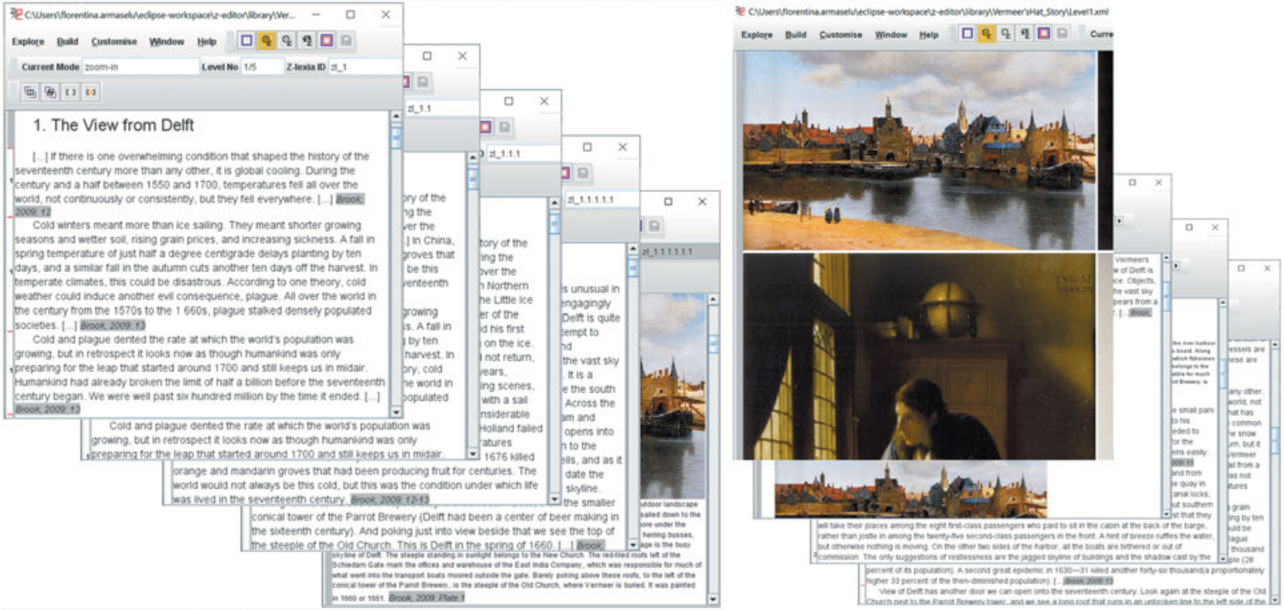


Figure 2: Structure of a book organized horizontally (left to right) by units (e.g. chapters, parts) and vertically (top-down) by conceptual level (e.g. from general to specific).

Figure 3 (left) shows a z-text constructed with fragments from the first chapter of Brook's (2009) book and the result of successive zoom-ins on a fragment from a historical standpoint. The exercise implied a preliminary interpretation of the book as a stratified representation of meaning. For instance, the top level in the *View from Delft* chapter contains fragments that describe events from a world history perspective, such as global cooling, plague and maritime trade in the sixteenth and seventeenth century. Details are added on the following levels: the focus gradually moves to more localized depictions of China's heavy frosts and the Little Ice Age in Northern Europe to the winter landscapes by Pieter Bruegel the Elder in the Low Countries and Vermeer's painting *View from Delft*. The painting is explained in more detail as containing several "doors" into the world of the seventeenth century. One door is the herring boats captured in the picture, as evidence of the herring fishery moving south under the control of Dutch fishermen due to climate change. Another door is the home of the Dutch East India Company, the VOC, also visible in the picture, which points to the network of trade that linked the Netherlands to Asia from the late sixteenth to the late eighteenth century.

Figure 3 (right) illustrates another hypothesis following the storyline more closely. More precisely, the text can be restructured starting from the other direction, i.e. the "doors" which are the paintings themselves corresponding to each chapter, then zoom in to open those doors and gradually expand the text. For instance, the chapter one z-text unfolds from the artwork and its description through


C:\Users\florentina.armaselu\eclipse-workspace\z-editor\library\Ver...
Explore Build Customise Window Help
Current Mode zoom-in Level No 1/5 Z-lexia ID zl_1
zl_1.1
zl_1.1.1
zl_1.1.1.1
zl_1.1.1.1.1
C:\Users\florentina.armaselu\eclipse-workspace\z-editor\library\Vermeer'sHat_Story\Level1.xml
Explore Build Customise Window Help Curre

1. The View from Delft

[...] If there is one overwhelming condition that shaped the history of the seventeenth century more than any other, it is global cooling. During the century and a half between 1550 and 1700, temperatures fell all over the world, not continuously or consistently, but they fell everywhere. [...] Brook, 2009: 12

Cold winters meant more than ice sailing. They meant shorter growing seasons and wetter soil, rising grain prices, and increasing sickness. A fall in spring temperature of just half a degree centigrade delays planting by ten days, and a similar fall in the autumn cuts another ten days off the harvest. In temperate climates, this could be disastrous. According to one theory, cold weather could induce another evil consequence, plague. All over the world in the century from the 1570s to the 1 660s, plague stalked densely populated societies. [...] Brook, 2009: 13

Cold and plague dented the rate at which the world's population was growing, but in retrospect it looks now as though humankind was only preparing for the leap that started around 1700 and still keeps us in midair. Humankind had already broken the limit of half a billion before the seventeenth century began. We were well past six hundred million by the time it ended. [...] Brook, 2009: 13

Cold and plague dented the rate at which the world's population was growing, but in retrospect it looks now as though humankind was only preparing for the leap that started around 1700 and still keeps us in midair.

orange and mandarin groves that had been producing fruit for centuries. The world would not always be this cold, but this was the condition under which life was lived in the seventeenth century. Brook, 2009: 12-13

conical tower of the Parrot Brewery (Delft had been a center of beer making in the sixteenth century). And poking just into view beside that we see the top of the steeple of the Old Church. This is Delft in the spring of 1660. [...] Brook,

skyline of Delft. The steeple standing in sunlight belongs to the New Church. The red-tiled roofs left of the Schiedam Gate mark the offices and warehouse of the East India Company, which was responsible for much of what went into the transport boats moored outside the gate. Barely poking above these roofs, to the left of the conical tower of the Parrot Brewery, is the steeple of the Old Church, where Vermeer is buried. It was painted in 1660 or 1661. Brook, 2009: Plate 1

will take their places among the eight first-class passengers who paid to sit in the cabin at the back of the barge, rather than jostle in among the twenty-five second-class passengers in the front. A hint of breeze ruffles the water, but otherwise nothing is moving. On the other two sides of the harbor, all the boats are tethered or out of commission. The only suggestions of restlessness are the jagged skyline of buildings and the shadow cast by the

percent of its population). A second great epidemic in 1630—31 killed another forty-six thousand(a proportionately higher 33 percent of the then-diminished population). [...] Brook, 2009: 11

View of Delft has another door we can open onto the seventeenth century. Look again at the steeple of the Old Church next to the Parrot Brewery tower, and we see a long roof that runs in an unbroken line to the left side of the

Figure 3: Brook (2009)[4] z-text: history layout (left); story layout (right).

[4] Text fragments reproduced with permission (©Timothy Brook, 2009, *Vermeer's Hat: The Seventeenth Century and the Dawn of the Global World*, Profile Books and Bloomsbury Publishing Inc.). Image credits: Johannes Vermeer, *View of Delft* (Mauritshuis, The Hague); Johannes Vermeer, *The Geographer* (Städel Museum, Frankfurt am Main).

events of local history to the large-scale view on global cooling and world trade development. The labels marked in grey in the figures indicate bibliographical notes that contain pages in the book from where the fragments were extracted, such as pages 11, 12, 13 and Plate 1 (for the painting). The restructuring of the original text as a z-text presupposes that various clusters of meaning, belonging to different levels of detail, are scattered throughout the chapters of the book, not necessarily in contiguous areas. My assumption is that grouping them together by level on the same plane and stratifying the representation on several levels would provide new insights into the text and its complex multi-layered structure. This implies various conceptual scales dispersed through fragments that link to each other horizontally and vertically over shorter and longer distances. The methodology applied to detect this type of structure has followed this intuition.

3 Topic modelling

For the preparation and first phase of analysis of the corpus, I used MALLET (McCallum 2002), a software package applying latent Dirichlet allocation (LDA) for topic modelling (Blei 2011), combined with Microsoft Excel functions and Visual Basic for Applications (VBA) procedures that I created for the project.[5] The choice of MALLET and Excel was driven by their accessibility and the possibility of creating output and diagnosis files for further analysis and processing. However, the methodology should be applicable to other types of software as well (for instance, to an integrated Java package that may implement the proof of concept described in this chapter in a second phase of the project).

3.1 Entropy

Each book from the dataset was analyzed with MALLET.[6] Each folder, corresponding to a book organized into files for chapters or parts, was imported via *import-dir* with the options *keep-sequence* and *remove-stopwords* (strings were converted to lower case by default). The topic models were built with *train-topics*

[5] Experiments with hierarchical LDA (hLDA) (Blei et al. 2009) were ongoing at the time of writing and are not described in this paper.

[6] For more details about the options used for analysis, see Graham et al. (2012) and the online MALLET documentation at https://mimno.github.io/Mallet/topics and https://mallet.cs.umass.edu/diagnostics.php, accessed July 27, 2023.

including the options *output-state*, *output-topic-keys* and *diagnostics-file* to produce a series of XML and tab/space delimited files. The resulting data were imported into Microsoft Excel for processing through built-in functions and VBA procedures that I wrote for this purpose. After a set of tests with various numbers of topics (8, 10, 15, 20, 25) and analysis of topic quality, the number of topics was empirically set at 20, with an *optimize-interval* value of 20 and the default value of 20 for *num-top-words*. The decision was based on the observation that the number of 20 topics produced a topic distribution that included at least two dominant topics appearing in almost all the chapters/parts of the books from the collection considered in the study. This observation was considered as a first indicator of a structure layered from general to specific.

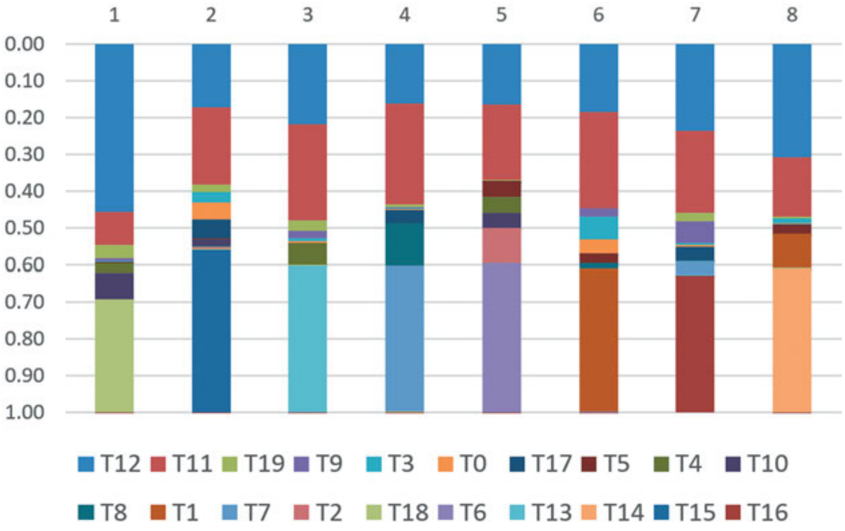


Figure 4: Topics sorted by *document_entropy* in Brook (2009), with topic probabilities (vertical axis) and their distribution by chapter (1–8, horizontal axis).

The goal of post-processing the MALLET files was to devise a methodology for detecting the levels of generality and specificity that characterize each book. For this purpose, I used the topic distribution per document (chapter or part) from the composition file, combined with the *document_entropy* measure from the diagnostics file. Topics with low entropy values are concentrated in a few documents, while topics with higher entropy values are spread evenly over many documents (MALLET documentation – *Topic model diagnostics*). This metric was considered as an indicator of generality versus specificity within the chapters/parts of the books included for analysis.

Figure 4 shows the topic distribution per chapter for Brook (2009), with topics sorted in descending order of their *document_entropy*. One can observe that topics T12 and T11 (top of the bars), and to a lesser degree T19 and T9, are the ones spread throughout the chapters, while topics such as T18, T15 and T13, at the other end of the spectrum (bottom of the bars), are mostly concentrated in a single chapter (chapters 1, 2 and 3 respectively). Intermediate topics are represented by thinner strips in the middle area of the bars.

Figure 5 presents the topic distribution by chapter (Brook 2009) for each of the 20 topics, arranged from more general to more specific (left to right and top-down). Table 1 shows excerpts of top words for the most generic and most specific topics and the chapters where these topics are prominent.

The first two topics (T12, T11) are almost evenly distributed throughout the chapters of the book. They are part of Brook's recurrent argumentation that outlines the emergence of global trade in the seventeenth century, connecting Europe with the world. Narrower descriptions of particular events, developed in relation with the eight paintings by Vermeer and other artworks, stand for articulation points chosen by the author as "doors" or "passageways" to the seventeenth-century world for each chapter (e.g. T18, T15, T13). Intermediary topics (e.g. T17) that cover fewer chapters (but more than one) appear to be less coherent[7] and are probably referring in the texts to fragments that make the transition between more general and more specific themes.

3.2 Levels

To detect the number of levels of generality and specificity inherent to a text and assign topics to such levels, I used the *document_entropy* metric and the computation of the *slope* (Excel built-in function) as a measure of the generality/specificity variation from one topic to another in the graph (Figure 6).

I considered that two adjacent topics T_i, T_j belong to the same level if the absolute value of the slope computed using their corresponding *document_entropy* is less than the value of the average interval computed as $(\max - \min)_{document\_entropy}$ divided by the number of topics.[8] The resulting mapping of topics to levels is shown in Table 2. One can observe that same level topics tend to appear together on plateau, while a change of level is marked by steeper or longer slope lines in the

[7] Also according to the MALLET *coherence* indicator computed for the topics.

[8] Except for the two most general topics that were considered by default as belonging to two separate levels, 1 and 2.

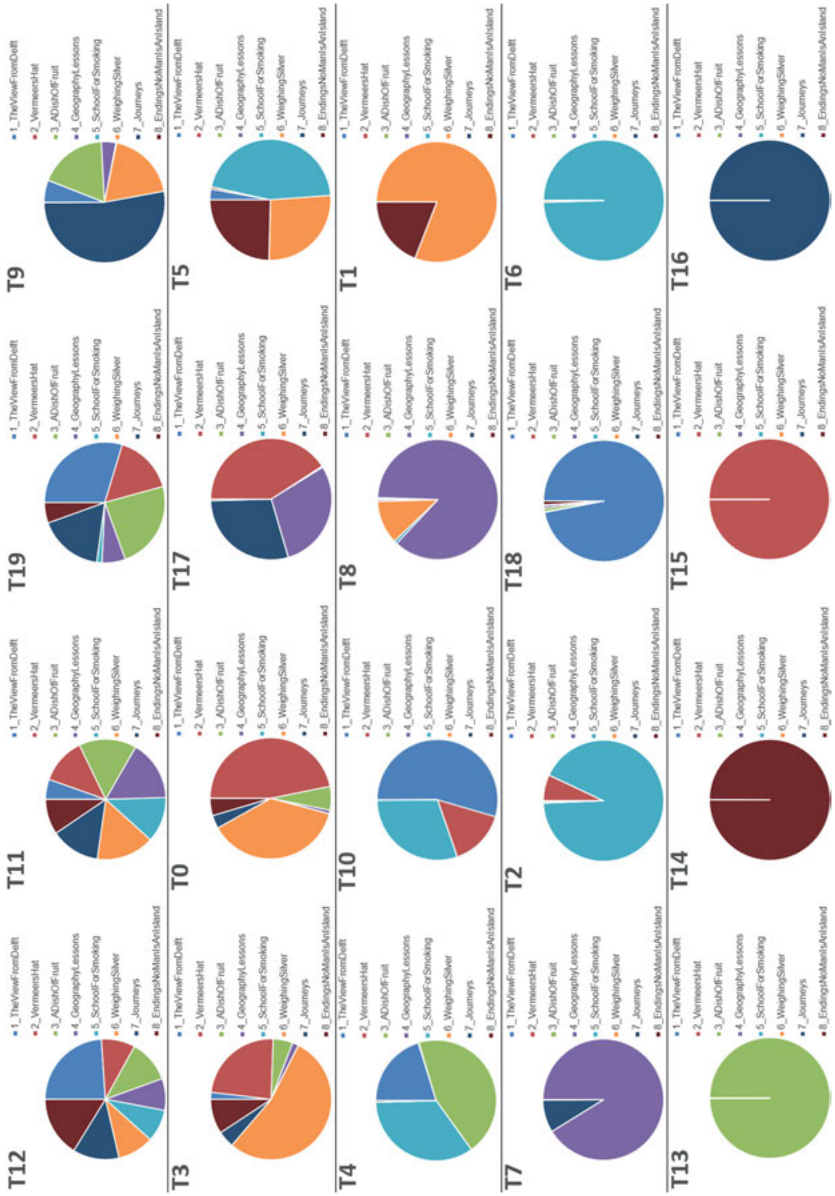


Figure 5: Topic (T0–T19) distribution by chapter (1–8)[9] in Brook (2009).

[9] 1_TheViewFromDelft, 2_VermeersHat, 3_ADishOfFruit, 4_GeographyLessons, 5_SchoolForSmoking, 6_WeighingSilver, 7_Journeys, 8_EndingsNoManIsAnIsland.

Table 1: Most generic and most specific topics. Excerpts from Brook (2009).

Topic	Top 20 words	Distribution by chapter
T12	century world dutch time vermeer seventeenth painting years people delft place trade life voc men things long great sea side	Even
T11	chinese china back made european trade europe europeans south coast end make called ship people spanish found needed portuguese japan	Even
. . .	. . .	. . .
T17	gunpowder fashion bound kill feet boundaries fishing passage role hung dutchmen rebel lack supplies semedo scattered correctly infiltrating reaching europe's	Mostly in 2_VermeersHat, 4_GeographyLessons and 7_Journeys
. . .	. . .	. . .
T18	delft paintings shanghai view canal pearl rotterdam built schouten chamber schiedam web buildings oude surface cold herring kolk contacts dong	Mostly in 1_TheViewFromDelft
T15	champlain french lake beaver native arquebus champlain's huron mohawks hurons allies felt hat montagnais lawrence hats dream iroquois chiefs map	Mostly in 2_VermeersHat
T13	porcelain white objects ships dishes lion dutch wen potters pieces amsterdam taste voc cargo dish portuguese produced ceramic lam blue	Mostly in 3_ADishOfFruit

diagram. The topic levels also seem to be correlated with the degree of generality/specificity or distribution by chapter shown in Figure 5.

Once the topics were assigned to levels, these levels were propagated to all the words in the text belonging to the topics. For this purpose, I used Excel to further process the MALLET output obtained via *output-state*, which is a file containing the words of the corpus (book), after stopword removal, with their topic assignments, index and position within each document (chapter or part of the book). In this way, each word was assigned to the level of the corresponding topic. Since the analysis of text as a scalable structure was intended for units of text larger than words, I created a set of procedures in Excel VBA to propagate levels from words to segments of a given length in number of MALLET words.[10]

[10] That is, the number of word tokens after stopword removal by MALLET (this is how *words* are also referred to for the rest of the chapter).

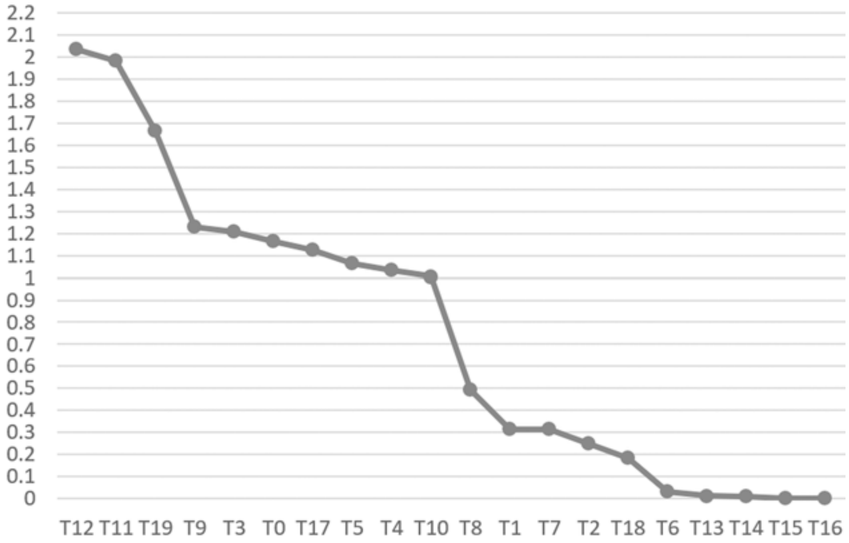


Figure 6: Topics (T0–T19) in decreasing order of their *document_entropy* (vertical axis) in Brook (2009).

Table 2: Topic to level mapping, from generic to specific (top-down), in Brook (2009).

Topic	Level
T12	1
T11	2
T19	3
T9, T3, T0, T17, T5, T4, T10	4
T8	5
T1, T7, T2, T18	6
T6, T13, T14, T15, T16	7

First, the probability of each word to belong to a topic was computed by counting the number of times a word *w* was assigned to topic *t* and dividing this value by the total number of words assigned to that topic. Then, given the length of a segment *s* defined as a number of words inside a document, a score was computed for each level according to the following formula:

$$score(s, l) = \frac{\text{count\_lvl}_{\text{l in s}}}{\text{seg\_size}_{\text{s}}} \times \frac{\text{avg\_prob\_word\_topic}_{\text{l in s}}}{\text{avg\_word\_distance}_{\text{l in s}}} \tag{1}$$

where: $\text{count\_lvl}_{\text{l in s}}$ is the number of times level *l* appears in segment *s* (i.e. the number of words assigned to level *l* in segment *s*); $\text{seg\_size}_{\text{s}}$ is the size in number of words of segment *s*; $\text{avg\_prob\_word\_topic}_{\text{l in s}}$ is the average word-topic probability for words belonging to *l* in *s*; $\text{avg\_word\_distance}_{\text{l in s}}$ is the average distance between words belonging to *l* in *s*.

A segment containing words from different levels is therefore assigned to the level that has the highest score according to (1). This is the level that appears many times in the segment and has many words assigned to it, whose average probability of words belonging to that level is higher, and which involves words that appear grouped together at smaller distances (presumed to form more compact clusters of meaning). In the case of Brook (2009), seven levels were detected (Table 2) and propagated to segments by applying this method.

Tests were run for different segment sizes, from one segment of the size of each book, then sizes iteratively divided by 4 up to 1,024 (six iterations),[11] in a process similar to the generation of the Koch curve that divides the segments by 4 at every iteration. Segment counting was reset at the beginning of each unit (chapter or part), except for the first iteration when a single segment of the length of the book was considered. Thus, segments of different sizes could result from an iteration (either for values larger than a unit size, when the actual size of the segment was the unit size, or for segments placed at the end of the unit containing the remaining words after the division corresponding to the iteration). The process was intended to simulate, by iterative reductions of the segment size, the representation of text at various scales, revealing a stratification by levels and a fragmented rather than flat structure where all the components are placed on a single line. The segment-level diagrams in Figure 7 were computed in Excel following the method for step charts without risers (Peltier 2008).[12] It was observed that for large segment size values (large scale), when one segment covers a full unit (chapter or part), the assigned level can differ from unit to unit, and it is not always a level corresponding to the most general topics, as would have been expected. Sometimes it may be a specific level or, less often, an intermediate level. Figure 8 displays the detail of the word distribution by level for the first 11 words and the first segment of size 35, 140 and 560 words in Brook's book, chapter 1. We can compare it with the three bottom diagrams from Figure 7 (read from right to left). According to the score computed by formula (1), segment 1 is assigned to level 6 when considered at a small scale (segment size: 35 words) and to

[11] With the ratio of $2^{2(k-1)}$, where k = 1, 2, . . . , 6 represents the number of the iteration.

[12] For simplification, all the segments are represented equally. Segment i spans i to i+1 (starting with 1), where i stands for the numerical labels on the horizontal axis of segments. For visibility and analogy purposes, the segments were represented as 15pt-wide bars (instead of points) in the Excel diagrams. The vertical axis represents the levels, from 1 to 7 for Brook's book.

level 1 when the scale increases further (140, 560 words). For larger scales (Figure 7, top, right to left), the first segment remains at level 1 for the next two iterations, but is assigned to level 7 when a single segment of the size of the whole book is considered. This way of looking at the text as made of building blocks of increasing size as the observation scale increases can provide insights into the mechanisms of meaning production which involve assembling words with different degrees of generality and specificity to form more complex units. The specificity or generality of these higher order units, such as sentences, groups of sentences or paragraphs, chapters, parts and whole book, could therefore be detected and mapped on different levels, revealing a stratified conceptual structure rather than a linear layout.

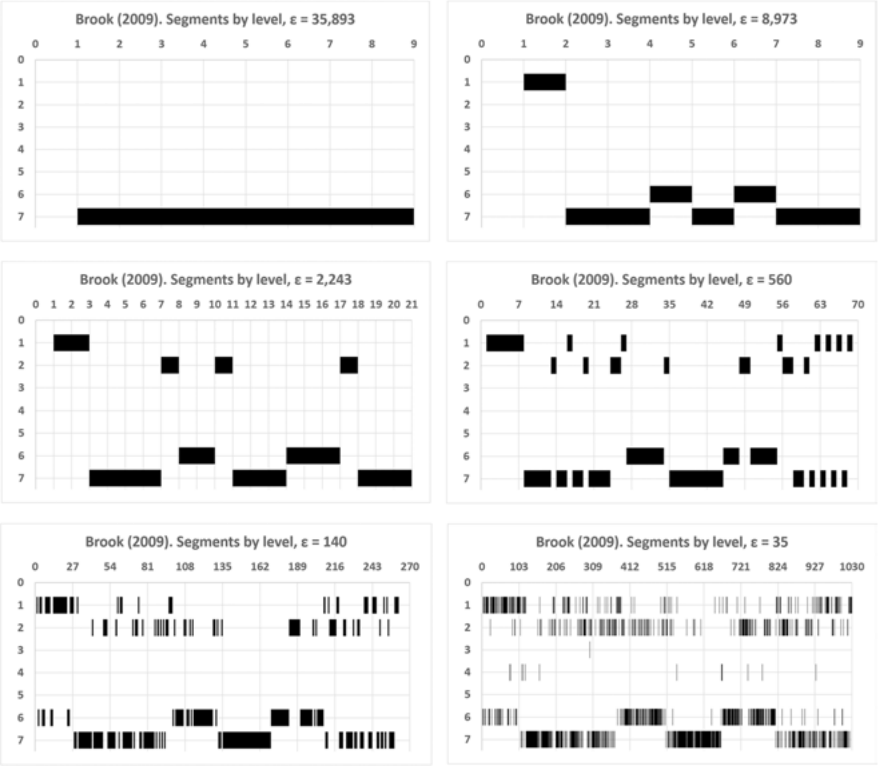


Figure 7: Segment distribution (horizontal axis) by level (vertical axis) at different scales, with ε the size of the segment in number of words (Brook 2009).

If we read Figure 7 in reverse order, we can interpret the progression left to right and top-down by analogy with a process in physics. First, the whole text-bar is assigned by the algorithm to the most specific level 7. By exposure to external fac-

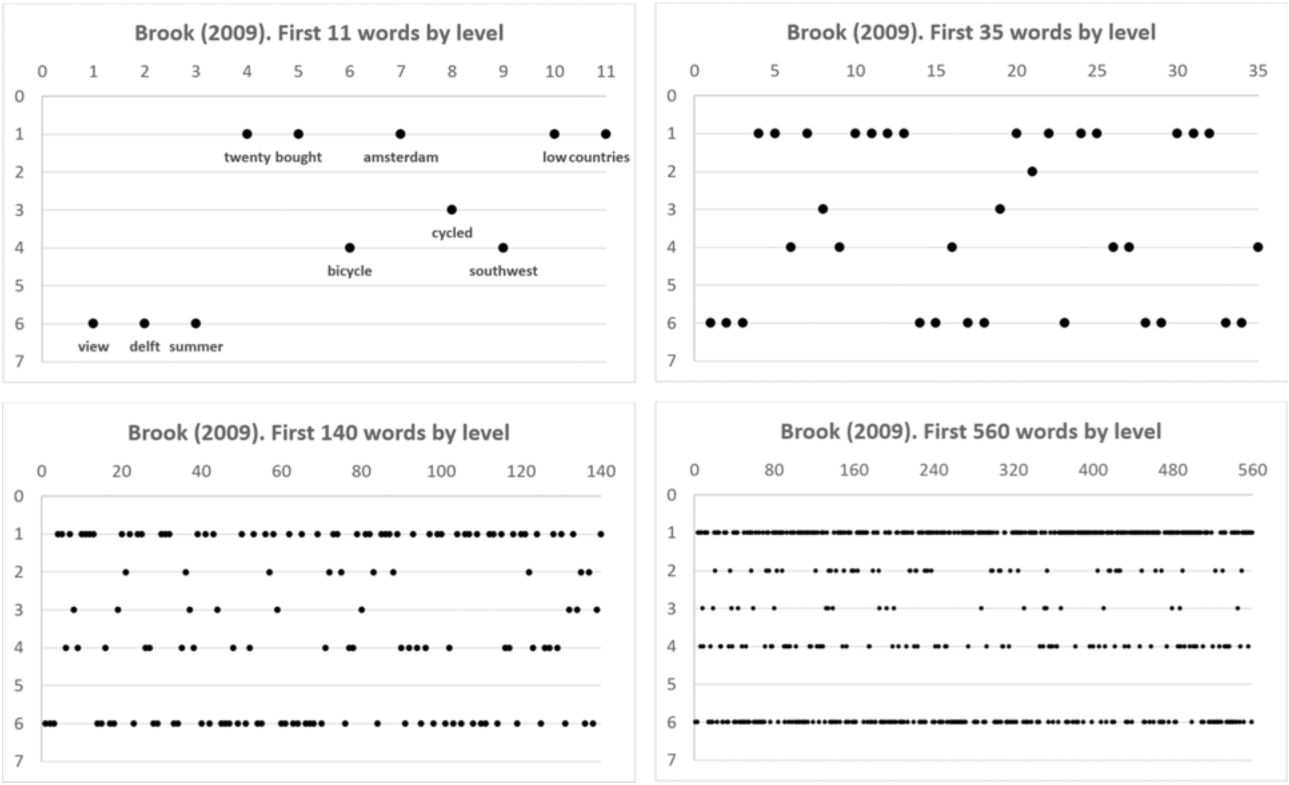


Figure 8: Word distribution by level, first 11 to 560 words from segment 1, chapter 1, at different scales (Brook 2009).

tors (in our case the analysis at different scales)[13] the text is broken into smaller and smaller units of analysis, which seems to increase the mobility of the resulting segments and their migration to more generic levels.

From a conceptual point of view, it therefore appears that the focus of Brook's book gradually moves from specific to generic (or from analysis to synthesis) with the decrease in size of the investigation unit. Following this line of thought, which seems to align with the storyline z-text layout depicted in Figure 3 (right),[14] we may infer that the strategy of argument unfolding of the book is primarily articulated, when considered at a large scale, around the concept of "doors", symbolized by the eight artworks corresponding to each chapter. This conceptual framework, corresponding to the microhistory level, is gradually enriched and contextualized within a global-history perspective that depicts the "dawn of the global world in the seventeenth century" through storytelling and synthesis inserts that become more "visible" or prominent at intermediary and smaller scales of analysis. In his study of semantic similarity in a taxonomy, Resnik (1995) proposes a measure to quantify the "information content" of a concept according to its position in a taxonomy (e.g. the top concepts being more abstract). Starting from this idea but applying it to textual fragments, I will use the term *informational granularity map* for the diagram depicted in Figure 7 that graphically describes the degrees of generality and specificity characterizing the units of analysis at different scales of representation. Granularity is understood both as a measure of the scale of observation, represented through the segment size, and as an expression of the degrees of generality and specificity of the content itself. Section 5 will provide a comparison with the informational granularity maps built for the other books in the collection (Figure 15).

4 Fractal geometry

This stratification by levels suggested that texts might possibly be interpreted as fractals; that is, as irregular, non-linear forms characterized by a certain degree of self-similarity. In this view, a text, first considered in its entirety as a single segment of words and a single conceptual unit, becomes scattered along several conceptual lines when iteratively broken into smaller segments and analyzed at decreasing scales. The question was whether this type of structure can be more

[13] In physics this may correspond to exposure to radiation or higher temperatures that produces segment fragmentation.

[14] It should be noted, however, that the z-text model operates with segments of the order of several sentences or one or two paragraphs.

formally portrayed as a fractal and if it is possible to determine its degree of irregularity as measured through a fractal dimension. This measure may then be used to compare the fragmentation of different texts by level and as a possible indicator of their complexity in terms of multi-layered rather than linear structures. To do this I applied a method called *box counting* that is used in mathematics, physics and other natural sciences to detect the fractal dimension of irregular shapes (Falconer 2014; Gouyet 1996).

4.1 Fragmentation

I considered text representations for the corpus like the one shown in Figure 7. A grid of squares of side ε can be imagined as covering the image at every scale, where ε takes iteratively different values for each of the analyzed books.[15] The algorithm consists in counting the number N(ε) of squares (or boxes) of side ε that intersect the text shape, for grids of different granularity.

Figure 9 shows an example of box counting for Brook's book and a box of side ε = 8,973 words. The number of boxes N(ε) in this case is six, that is the number of squares in the grid that contain segments of text. At this coarser granularity, only three of the seven levels detected for the book are occupied with segments, the most general level (L1) and two of the most specific ones (L6, L7). The segments and boxes were modelled in Excel as integer intervals taking account of the number of words in the segments and their succession, and the (i, j) pairs, where i represented the column and j the row corresponding to a box. The total number of boxes in a grid at a certain scale was defined based on the scale factor $s = 2^{2(k-1)}$ plus 1, where k was the number of the iteration. For instance, at the iteration k = 2, the total number of words for Brook's book (35,893) was divided by the scale factor $s = 2^2 = 4$, resulting in segments of ε = 8,973 words.[16] Since the division might not always be exact, the scale factor was increased by 1 to include the remaining words of the last unit. Thus, a grid of (s+1)*(s+1) = 5*5 = 25 boxes was devised. The position of the levels and the level interval ω on the vertical axis were determined by dividing the maximum value of the squared grid by the number of levels N_l. For the Brook example in Figure 9, the level interval $ω = [(s+1)*ε]/N_l = (5*8,973)/7 = 44,865/7 = 6,409$.

[15] E.g. 35,893; 8,973; 2,243; 560 . . . words for Brook's book corresponding to the six iterations, Section 3.2.

[16] At this scale, ε exceeded the size of each unit (chapter) and the actual segment sizes corresponded to the sizes of the eight chapters of the book, labelled $S_1 – S_8$ in the figure.

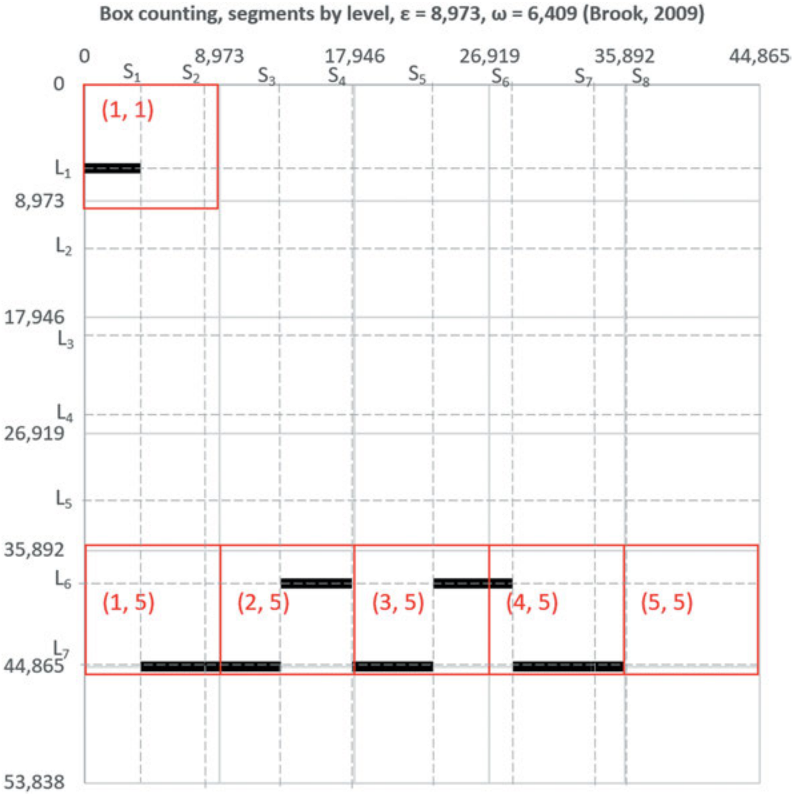


Figure 9: Example of box counting, second iteration, k = 2 (Brook, 2009).

The process was repeated and the number of boxes intersecting the text shape was calculated for six iterations (k = 1–6). I then built the diagram for log(1/ε) and log(N(ε)) (Figure 10). Usually, the fractal literature considers that if this curve exhibits an approximately linear behavior (at least for a certain subset of the plane), then N(ε) obeys a power law of the form: $N(\varepsilon) \approx c^{*}\varepsilon^{-D}$, where c is a constant. Therefore, the studied object may be assumed to possess fractal properties in the linear region, and D represents its fractal dimension.[17] Intuitively, D reflects how the number of counted boxes grows with the decrease in box side, the way in which the analyzed object fills the space, the ratio of change in detail to change in scale, or the inherent complexity of an irregular form (Falconer 2014: 27–28; Karperien and

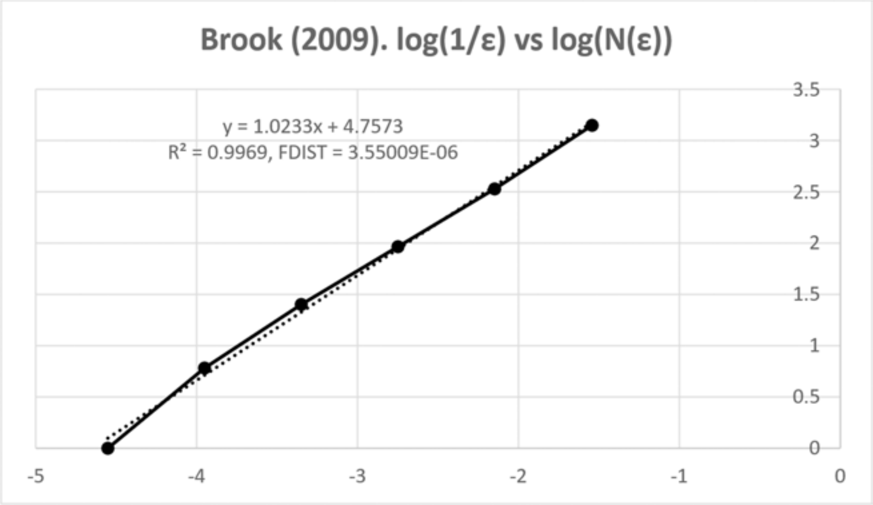
[17] Or box-counting dimension. There are different types of fractal dimension. See Mandelbrot (1983) and Falconer (2014) for a survey of these types and their degree of equivalence.

Jelinek 2016: 20). Various applications of the box-counting method to images either considered a grid of iteratively reduced size overlaid on the same image (Ostwald and Vaughan 2016) or used mathematical functions instead of pixel pictures to eliminate the distortions due to zoom-in (Wu et al. 2020). Unlike the image-based approach, I considered that the representation of text at each scale also changes with the side of the grid cell ε (according to the algorithm described in 3.2) and I worked with an interval-based modelling of the boxes and segments in Excel. This type of representation was intended to capture the disposition of segments on levels at various scales by simulating the effect of a zoom-in that makes more and more details visible as the scale decreases, and to test the application of the power law for fractal behavior on these different scale-driven configurations.

In practice, D is calculated as the slope of the linear region of the graph (Figure 10) for a certain number of iterations. Studies in a variety of research fields have shown that despite its relative simplicity, the box-counting method presents a series of drawbacks. For instance, the value of D varies with the range of box sides, the number of iterations[18] and certain characteristics of the grid or image to be analyzed (positioning, resolution) (Datseris et al. 2021; Ostwald and Vaughan 2016; Harrar and Hamami 2007; Klinkenberg 1994). In their analysis of fractal patterns of words in a text, Najafi and Darooneh (2015) observe that detecting the fitting range in the log-log plot of the number of filled boxes against box side, and the fractal dimension as the slope of the line of best fit is quite challenging to do automatically. Other studies have pointed to the need to provide other statistical measures, such as the correlation coefficient, mean and standard deviation over multiple samples used to compute the fractal dimension, to assess the accuracy and limitations of the model (Karperien and Jelinek 2016: 23–26).

To estimate the fractal dimension, I applied the method of the least squares to compute the slope of the $\log(1/\varepsilon)$ vs $\log(N(\varepsilon))$ graph (Harrar and Hamami 2007) and computed related statistical measures as first accuracy estimators. Figure 10 displays a 1.0233 slope and 4.7572 intercept (left) and the variation of the number of filled boxes with box side (right). The R^2 statistic shows a proportion of 0.9969 of variability in $Y = \log(N(\varepsilon))$ that can be explained using $X = \log(1/\varepsilon)$ and a measure of their linear relationship and correlation in the sample. A value close to 1 indicates that a large proportion of the response has been explained by the regression, while a number near 0 suggests the opposite (James et al. 2017: 69–71). The FDIST statistic estimates the probability that the observed relationship between the two variables

[18] For instance, Ostwald and Vaughan (2016: 40) recommend "at least eight and preferably ten or more comparisons" for better accuracy, to reduce the error rate to "around ±1 % or less", in their study of the fractal dimension in architecture.



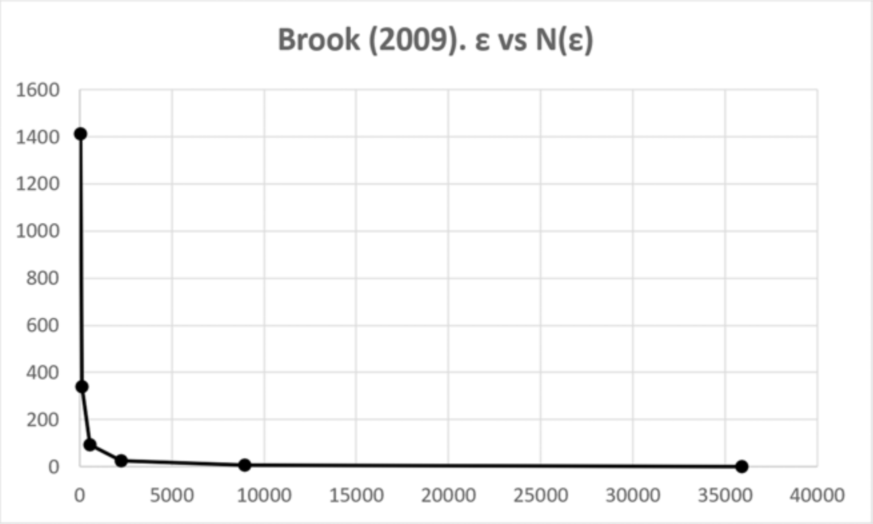


Figure 10: Fractal dimension as the slope of the double-log plot, log (1/ε) (horizontal) vs log(N(ε)) (vertical) (top); box side (horizontal) vs number of filled boxes (vertical) (bottom) for iterations k = 1 to 6 (Brook 2009).

occurs by chance, which is very low (3.55009E-06) in this case.[19] Although the influence of some factors on the box-counting dimension, such as the side of the box and the number of iterations, requires further analysis, I considered the values obtained through this approach as a rough approximation of the fractal dimension and basis of comparison for the texts included in the study (see also Table 4, Section 5.2).

Once the dimension value had been computed, another question needed to be addressed, i.e. how this dimension could be interpreted within the fractal theory framework. Mandelbrot (1983) generically called *dust* the fractals with dimensions in the interval 0 to 1. A classic example from this category is *Cantor dust*, whose generation is illustrated in Figure 11. Its construction starts with a straight line as an initiator, followed by a generator obtained by removing the middle third from the initiator. The process is repeated at smaller and smaller scales, by continuing to delete the middle third.

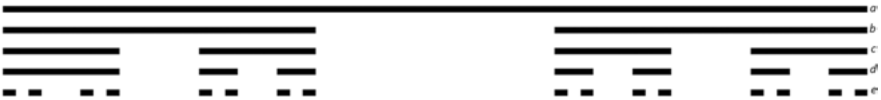


Figure 11: Cantor dust (adaptation of Mandelbrot 1983, 80): a – first iteration, *initiator*; b – second iteration, *generator*; c, d, e – third, fourth and fifth iteration.

While the Koch curve (Figure 1) is a fractal with an approximate dimension of 1.2618, Cantor dust has a fractal dimension of 0.6309 (Mandelbrot 1983: 36, 80). Two- or three- dimensional Cantor dusts or sets can also be generated, with fractal dimensions above 1, such as 1.2619 and 1.8928 respectively (Sławomirski 2013; Tolle et al. 2003). One can observe a certain similarity between Cantor dust (Figure 11) and the segments of Brook's book represented at smaller and smaller scales (Figure 7). Some differences may be noted as well. First, the appearance of the latter is not as regular as that of the former since the textual segments and their distribution were generated through a score-driven procedure with variable results rather than the application of an invariant pattern of iteration. Second, Cantor dust corresponds to a linear structure with the matter from the gaps being gradually incorporated into the dark areas of increasing density through the process of "curdling", according to Mandelbrot's terminology. In contrast, Figure 7 depicts a multilinear, two-dimensional arrangement derived from the dispersion of the textual matter on the plane over several levels of generality and specificity with the decrease in scale.

[19] See also "LINEST function", *Excel for Microsoft 365*, https://support.microsoft.com/en-us/office/linest-function-84d7d0d9-6e50-4101-977a-fa7abf772b6d, accessed July 27, 2023.

I would therefore argue that texts may exhibit fractal geometry, which reveals a stratified conceptual structure on layers of generality and specificity at various scales. We can recall the procedure described in Section 3.2 and formula (1) that assigns a segment to the level with the highest score. This score considers the word count and the probability of words belonging to that level, as well as their average distance. If we imagine the segments as a form of dust, they seem to be attracted, at different scales, to levels that correspond to various degrees of generality and specificity. This "force of attraction" may be determined by the existence of more compact and coherent clusters of meaning in the text, which are more likely to belong to a certain level as compared with the other levels.

Following this interpretation, a closer look at Figure 7 suggests that at the largest scale, when the segment size coincides with the length of the whole book in number of words, this segment is "attracted" to the most specific level 7. It therefore seems that clusters containing words that are more specific are prevalent in conceptually depicting Brook's book at a global scale. However, for smaller and smaller scales, the dust segments are gradually scattered and attracted towards more generic, surface levels, as shown by the following iterations in the figure. This raises the question of whether this attraction and the movement of "dust" fragments from one level to another as the scale changes may indicate the existence of a certain type of pattern, correlation or persistence over a longer range, or whether it is completely random.

4.2 Memory

To determine whether such a "memory" of the text exists, I performed an analysis of fluctuation and long-range correlation for the selected corpus. The method was proposed by Peng et al. (1992) to uncover the correlation between basic structural units of nucleic acids over long distances by mapping nucleotide sequences onto a so-called "DNA-walk". In this random walk model, a function u(i) describes a walker's move through two values, $u(i) = +1$ and $u(i) = -1$, if the walker moves either up or down for each step i of the walk. Based on the "net displacement", the sum of the unit steps u(i) after l steps, Peng et al. (1992: 168) compute a measure called "root mean square fluctuation", F(l), that characterizes the average of the displacement. A power law function of the form $F(l) \sim l^{\alpha}$ where $\alpha \neq \frac{1}{2}$ may indicate a long-range correlation in the considered walk, while a value of $\alpha = \frac{1}{2}$ can be the indicator of a random walk. A straight line on the plot log(l) vs log(F(l)) would confirm a power law between the two measures, with α representing the slope calculated through the method of the least squares. Studies in linguistics, such as those by Pavlov et al. (2001), have applied the method to investigate long-range correlations between letters and combinations of symbols in English novels. In her analysis of children's language, Tanaka-Ishii (2018: 8, 5) observed that "long-range correlation is due to the arrangement of frequent words and rare words" and that "rare words tend to cluster".

The aim of my analysis was to determine whether any long-range correlation can be observed in the "walk" of the word segments from one level to another at different scales of representation. I considered the iterations k = 2 to 6. The function u(i) modelled the walk through the values: +1, if a segment j was followed in the text by a segment j+1 placed on an upper level (move up); −1 for a move down to a lower level; and 0 for a segment j+1 remaining on the same level as segment j. The number of steps was l = Ns–1, where Ns represented the number of segments detected for each iteration k. The fluctuation F(l) was computed in Excel using the formula proposed by Peng et al. (1992: 168). Figure 12 shows the plots for log(l) vs log (F(l)) for the iterations 4–6 (left to right) and Brook's book.

For larger scales (k = 2 to 3), the linear regions of the curves log(l) vs log(F(l)) covered shorter portions of the graph. As illustrated in Figure 12, for decreasing scales (k = 4 to 6), the diagram started to exhibit more visible linear regions, especially in the first part of the curves (left). A selection of the linear portions (right) allowed the values of α to be computed as the slope of the plots for these zones. As ε (the size of the segment) decreased, α took values from 0.1917 to 0.4218 (Appendix, Table 6, Brook). What can therefore be inferred about the segment "walk" through levels and the degree of correlation over longer distances corresponding to this walk? The values of α below 0.5 seem to suggest an "anti-persistent" behavior or "mean reversion" (Ijasan et al. 2017; Saha et al. 2020; Hu et al. 2021) when a move in one direction is followed by a move in the opposite direction. This behavior refers to the linear regions of the graphs (Figure 12, right), i.e. to a number of steps l = 9, 30 and 509 and segment size ε = 560, 140 and 35. In the case of Brook (2009), these segment sizes roughly correspond to blocks of 7–8 paragraphs, 2-2$\frac{1}{2}$ paragraphs and 2 and $\frac{1}{2}$–3 sentences respectively. As the segment size decreases, the range in number of steps with walk correlation apparently increases. Shorter blocks at smaller scales would therefore exhibit longer memory. Table 6 (Appendix) summarizes the results of the experiments calculating the long-range correlation in the walk for the original and shuffled data.[20] One can observe that for this book, the segments of 140 words (~ 2–2$\frac{1}{2}$ paragraphs) display a smaller value of α than the segments of 560 and 35 words, which seems to be related to their relative immobility (a higher per-

[20] The levels assigned to segments at each scale were randomly shuffled. The procedure was then applied in the same way as for the original data to model the walk and to compute the fluctuation and the value of slope α in the linear region.

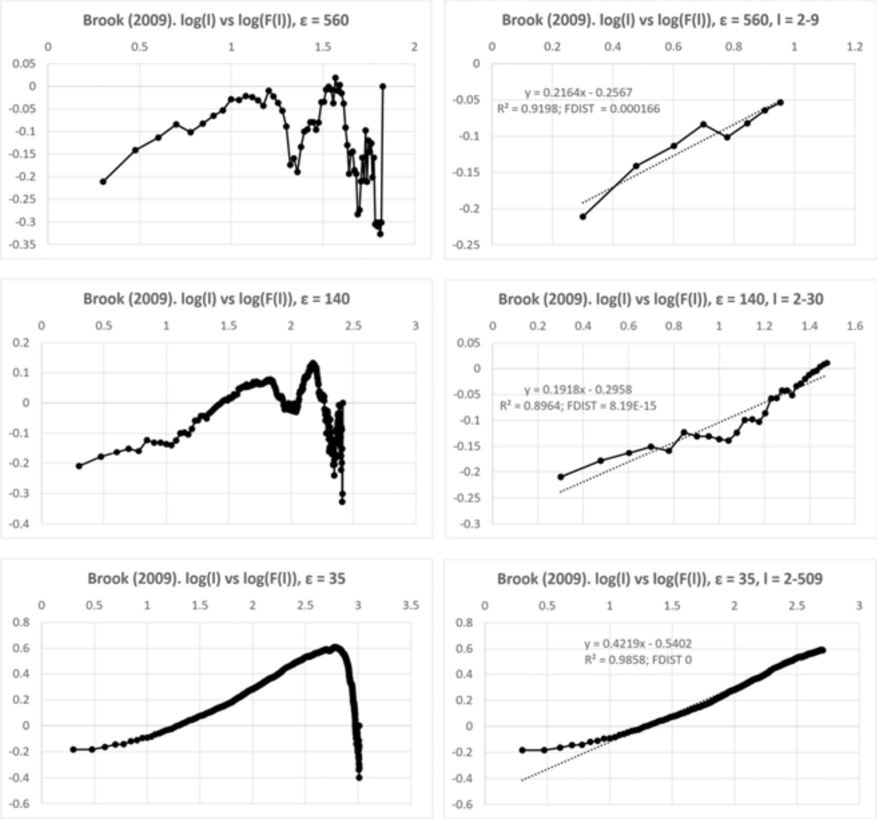


Figure 12: Fluctuation in the "walk" of word segments through levels for iterations k = 4 to 6 (Brook 2009).

centage of no moves, when the segments remain on the same level for longer periods) as compared with the other two block types. It is not yet clear why this happens, but a possible explanation may be related to the inner configuration of meaning of the book and the way this configuration is modelled by the algorithm defined by formula (1). In other words, 140-word blocks may be more dependent on the surrounding segments and thus less susceptible to independent movement across levels than the two other block types. Although there are significant differences between the values of α for initial and shuffled data, the latter still displayed values of α < 0.5, thus indicating an anti-persistent behavior, while a more random walk would have been expected. More shuffling rounds would be needed to draw a conclusion, but it is possible that the relatively high proportion of no moves (60%–68%) that characterized the walk at various scales had an impact on the results of the shuffling process. Therefore, the segments seem to fluctuate or remain still around or on some dominant levels at any examined scale. Mainly similar behavior was observed for the other books in the collection. Section 5 will provide a discussion of this aspect and its possible connections with Zipf's (2012) notion of specific-generic balance in texts.

4.3 Lacunarity

Another concept that presented interest for this study was that of lacunarity, as a "measure of the 'gappiness' or 'hole-iness' of a geometric structure" (Plotnick 1993: 202). In the domain of fractal geometry, this may be applied, for example, to distinguish objects with close or identical fractal dimensions, which display differences in the distribution and size of the gaps. The "gliding box algorithm" is one of the methods often applied to measure lacunarity (Allain and Cloitre 1991; Plotnick et al. 1993; Da Silva 2008). The method involves firstly representing an object against a grid of squares like the one shown in Figure 9. Then a box of variable side length (e.g. ε, 2*ε, 4*ε, 8*ε, etc.) is placed on the upper left corner of the grid and the number of occupied squares of side ε within the box is counted. After moving the box one column to the right, the filled squares in the box are counted again. The process is repeated over all the rows and columns of the grid, and for different sizes of the gliding box. The lacunarity is defined as a function of the side of the gliding box and the number of squares occupied by the object at different scales.

While the fractal dimension measures "*how much* the object (or data) fills the space", the "amount of space-filled, or the mass in some sense", lacunarity measures "*how* the data fill the space", the "spatial size of gaps and their structure within a set" or the "mass distribution" (Di Ieva 2016: 10; Tolle et al. 2003: 131). Lacunarity may thus indicate the "level of contagion between occupied sites at a particular scale" or the "degree of spatial clumping or aggregation" of certain populations (Plotnick 1993: 208). I considered that such a measure may be useful for a closer analysis of the distribution of gaps and the movement of segments to one level or another with the change in scale.

Figure 13 and Table 7 (Appendix) present the values and shapes obtained for the measure of lacunarity in the books in the collection. To compute this measure, I used the "gliding box algorithm" (Plotnick et al. 1993). For each scale and iteration (k = 1–6) corresponding to a certain segment and grid cell side (ε), the lacunarity was calculated using different values for the side of the gliding box, i.e. 2^0, 2^1, 2^2, . . ., as multiples of ε until a certain threshold for each iteration was

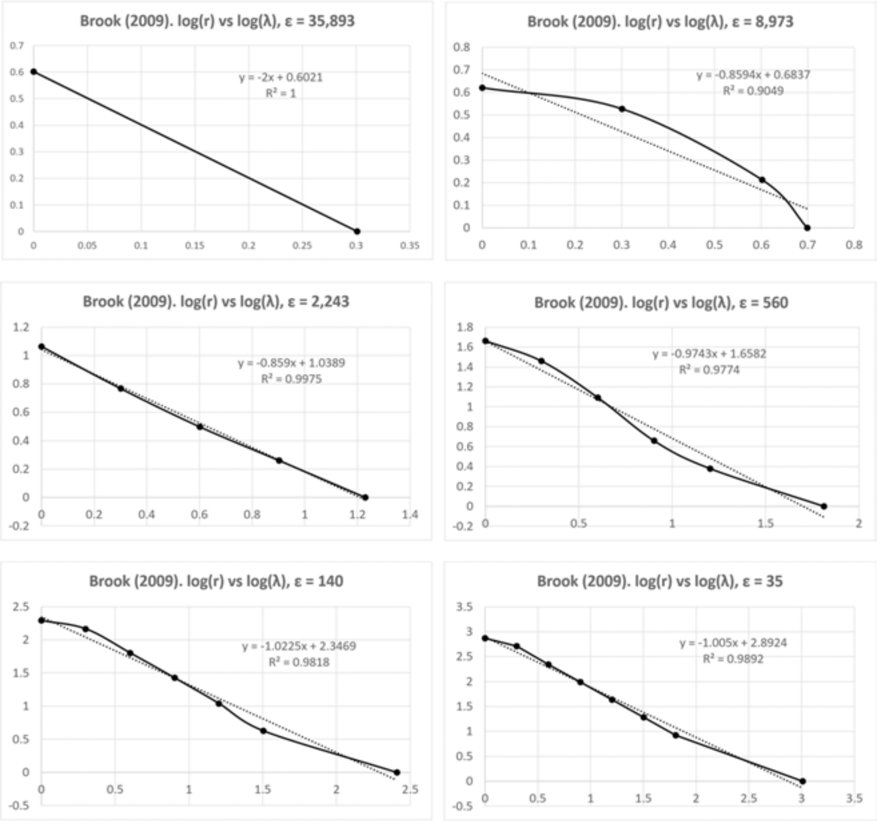


Figure 13: Log lacunarity (λ) by log side of the gliding box (r), iterations 1–6 (Brook 2009).

attained. The calculations also included the case of r corresponding to the side of the grid, M, as the maximum value (r = 2^{i-1}, 1 ≤ i < k+2; r = M, i ≥ k+2).[21]

As noted by Plotnick et al. (1993), the highest values of lacunarity were recorded at each scale (k = 1–6) for values of r = 1 when the gliding box was equal in size with the cell of the grid of side ε, while lower lacunarities were obtained as the box side increased. At different scales, the curves exhibited a certain degree of linearity (Figure 13), which according to Allain and Cloitre (1991) is a feature of self-similar fractals. In general, configurations with low variation in gap sizes display lower lacunarity, while objects with a wide range of gap sizes or larger areas of clumped sequences show higher values of lacunarity. For Brook

[21] Table 7 shows the specific values of r applied for each iteration.

(2009) the average lacunarity (λ) increased with the decrease in scale and segment size (ε) and in the fraction represented by the occupied cells in the grid (P) (Table 7, Figures 7 and 9). This suggests that the variability of gap sizes increased with the decrease in segment size corresponding to the increase in segment movability from one level to the other. A comparison of the lacunarity measure of all the books selected for the study will be presented in the following section.

5 Discussion

The experiments performed so far with the corpus of seven books, five from historiography, one from literature and one from philosophy, indicate that an analysis at various scales of the texts may reveal fractal properties and a stratified structure.[22] To detect these levels, I combined the concept of zoomable text (z-text), as a starting point, with topic modelling and elements from fractal theory and applications. Although some limitations were identified and further analysis is needed, it can be argued that the initial assumption of the text as a multi-layered conceptual construct seems to be confirmed.

5.1 Attractors

Various types of layers can characterize the internal organization of a text on levels, e.g. from simple to complex, abstract to concrete, global to local, etc. In the present study I modelled this type of layered pattern through the generic to specific spectrum. To this end, I considered that topics spread over several documents are more general than topics distributed mostly over a smaller number of documents or just one document. The existence of a specific-generic balance in semantic systems was formulated by Zipf (2012: 185) through the metaphor of the artisan involved in the task of classification by a number of *n* criteria and correlations, and the Principle of Least Effort. Specific correlations describe a small set of particular classes of events more completely, while generic correlations depict

[22] It should be noted that although theoretical fractals can go down indefinitely to smaller and smaller scales and we can imagine the scaling down of text to segments of size below 1 (word level), such as morphemes, letters, letter fragments, etc., for the present study I considered cut-offs in segment size above word level, as defined by the six iterations. See also Mandelbrot's (1983: 38) discussion on the Koch curve cascade of smaller and smaller promontories and the cut-off scales applied to real coastlines.

a larger set of classes but less completely. The balance between specific and generic correlations would therefore be maintained by the artisan in his attempt to generalize upon the basis of specific correlations, and particularize upon the basis of generic correlations, with the aim of minimizing *n* and the classification effort. Approaching the same question but from a different perspective, Lafon (1981) proposed a probabilistic model to discern between basic (non-specific) and specific forms in a corpus divided into parts.

My hypothesis was that this type of inference involves several degrees of generality and specificity that can be examined through longitudinal and transversal cuts of texts into units (e.g. chapters, parts) and levels (Figure 2), and different scales of observation. I assumed that a gradual unfolding of generic and specific arguments can be observed in the global-micro history texts built upon thematic aspects that varied from broad worldviews to minute examination of distinctive historical events, people, objects or points in time. Texts from other domains such as literature and philosophy were also presumed to exhibit a gradual relationship between generic and specific elements, from words defining the general theme and basis of communication to localized forms characteristic of certain units only. The topic modelling approach used in level detection offered a first glimpse into this type of conceptual structure. Table 3 lists the number of levels detected for each text in the collection against the number of words and units for each book. The influence of these two factors considered in isolation is not clear. A closer look at the percentage of intermediary levels and topics and the distribution of topics by level may suggest a possible explanation.

Table 3: Number of detected levels, book length (MALLET words), analysis units (chapter or parts), and intermediary levels and topics (sorted by length).

Book	Length	Units	Detected levels	Intermediary levels (%)	Intermediary topics (%)
1688. A Global History (Wills 2001)	51,846	8	8	62.50	35.00
Gulliver's Travels (Swift 2009)	37,892	5	9	66.66	75.00
The Inner Life of Empires (Rothschild 2011)	37,112	9	10	70.00	85.00
Vermeer's Hat (Brook 2009)	35,893	8	7	57.14	65.00
Plumes (Stein 2008)	29,023	7	9	66.66	65.00
Beyond Good and Evil (Nietzsche 2009)	24,196	10	7	57.14	45.00
The Two Princes of Calabar (Sparks 2004)	16,504	7	6	50.00	40.00

If we exclude the most generic (the first two) and the most specific (the lowest plateau) topics in the diagrams (Figure 14), we can infer which books exhibit a larger proportion of their topics and levels in the intermediary area. Thus, the books with a higher number of levels (≥ 8) in Table 3 are also those with a higher number of intermediary levels, such as Wills (2001), Swift (2009), Rothschild (2011) and Stein (2008). The number of detected levels may therefore be influenced by the length of the texts and the number of units, the way in which the authors shape their discourse through generic and specific classes of words and topics, and also classes of words and topics that belong to the area in between.

Studying the texts at various scales revealed that some levels act as "attractors" of segments (considered as "fractal dust"). The movement of segments from one level to another does not appear to be random. As shown in Section 4.2, this movement seems to be characterized by a particular type of "memory" of the segments and values of α situated below 0.5, which would correspond to an "anti-persistent" behavior. Table 6 presents a summary of this type of memory. One can observe a certain symmetry in moves up and down and a relatively high percentage of no moves or stationary behavior of the segments for all the books in the collection, which may indicate segment fluctuation around a dominant level at the smaller scales. Why this happens is not yet completely clear. As shown in Figures 7 and 15, the books as whole aggregates start on a more generic or specific level, and then, with the decrease in scale, a certain equilibrium between generic and specific tends to be established by the migration of segments from one level to another. The tension between generic and specific alluded to by Zipf (2012) therefore seems to operate at the smaller scales (and possibly word level) characterized by higher segment "mobility". It should be noted, however, that the generic-specific dichotomy is not binary, but multi-value and involves different degrees, or levels.

Table 6 provides insights into segment mobility at smaller scales. The lowest values of α are displayed for Stein (2008) and Wills (2001), the former with a generic level, the latter with a specific dominant level (Figure 15), and segment sizes of 453, 113 and 810 words.[23] These types of block therefore seem less mobile for these books, given the high percentage of no moves that characterizes them, or may follow a movement logic that is only feebly anti-persistent. For all the books, the memory interval (in number of steps) increases with the decrease in scale (segment size), and for more than half of the books (Rothschild, Stein, Wills and

[23] Corresponding to $5\frac{1}{2} \div 6\frac{1}{2}$, $1 \div 1\frac{1}{2}$, and respectively $9 \div 11$ paragraphs.

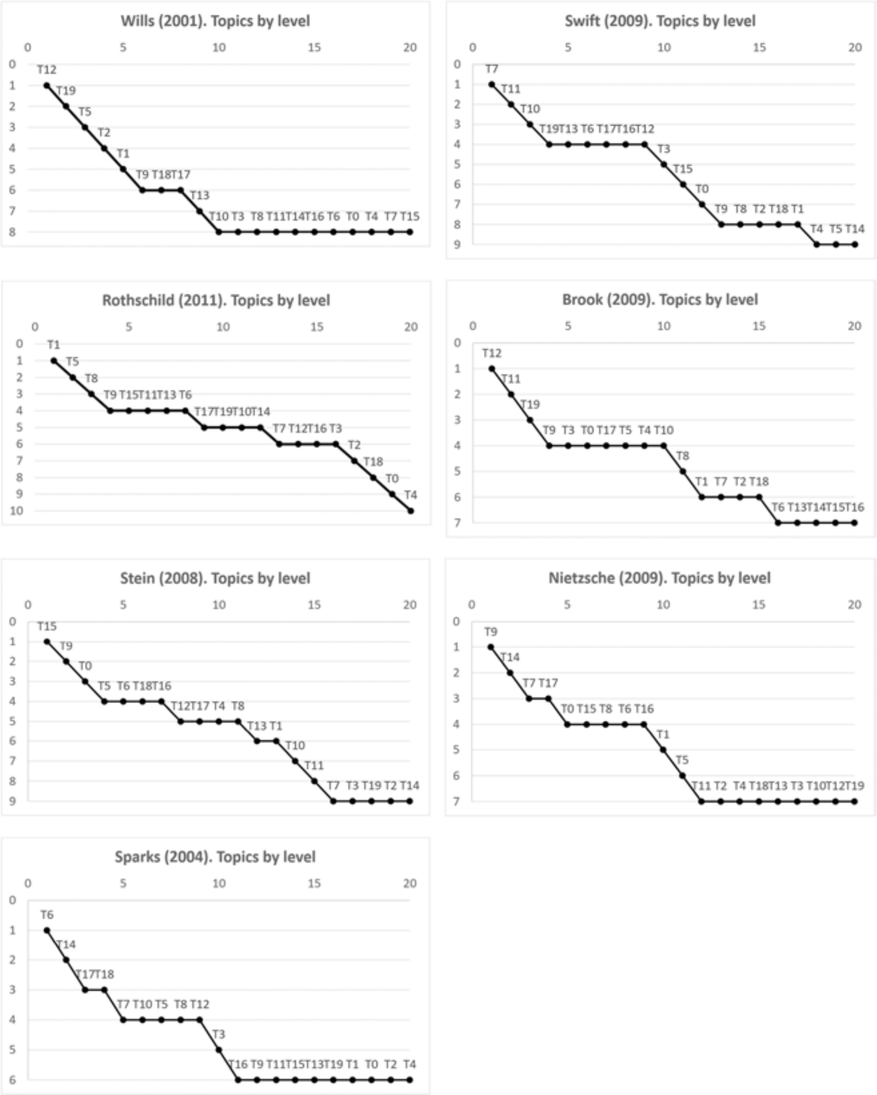


Figure 14: Topic distribution by level and book in the collection (sorted by length, left-right, top-down).

Swift), α increases with the decrease in scale.[24] These books, as can be observed in Figure 15, present a dominant level, with the highest density of segments, at all

[24] It would be interesting to investigate the value of α in the case of segments of 1 word in size, to see if it approaches 0.5, representing a random walk, or if it remains below this value, and

scales. In this case, anti-persistent behavior would consist in a tendency of most segments to remain on the dominant levels, except for those that have enough mobility (or energy) to escape their attraction. The increase in α seems to capture this phenomenon, since this form of energy appears to increase with the decrease in scale. Thus, Zipf's property would manifest itself not as a generic-specific balance, but as a tendency to maintain stability around a certain level of generality or specificity. Two books (Brook and Sparks) exhibit a different pattern through a slightly lower value of α for the middle sizes ($k = 5$). In the case of Brook (2009), this may be related to a higher value for the number of no moves, as explained in Section 4.2. It should also be noted that this book contains the longest memory interval and the highest value of α for the smallest scale studied ($k = 6$). This may be due to a certain equilibrium between the occupation of generic, specific and intermediary levels (Figure 7, bottom, right) and the strategy, discernible at this scale, of an unfolding of detail vs global view, constructed around the eight artworks chosen "not just for what they show, but for the hints of broader historical forces that lurk in their details" (Brook 2009: 7). For Sparks (2004), the lower value of α ($k = 5$) may be related to the nature of the units themselves,[25] whose fluctuation suggests a slightly lower tendency to return to the dominant level at every move than the units corresponding to the previous iteration ($k = 4$).[26] The case of Nietzsche (2009) is more intriguing, since it shows a decrease of α with the decrease in scale. This behavior could be caused by the sensitivity of smaller scales to the style of this philosophical text that alternates very long assertive sentences with very short questions, which may impose a logic of segment assignment to levels that is perhaps less anti-persistent in nature than when articulated at a larger scale.

5.2 Dispersion

The study of the books at different scales and their fractal geometry seem to offer a new standpoint on text as a conceptual object. This may reveal a certain type of dynamics in the stratification of levels and the way in which these levels attract word segments with changes in scale. Such behavior may be related to the aggre-

thus continues to show an anti-persistent behaviour. The time allocated to the writing of this chapter allowed only for partial experiments of this type, meaning that it is not possible to draw a conclusion at this stage.

[25] $\frac{1}{2} \div 1$ paragraph.

[26] $2\frac{1}{2} \div 3$ paragraphs.

gation of clusters of meaning, correlation over long distances and the modes in which conceptual building blocks of variable sizes are formed in language.

Table 4: Fractal dimensions and generated statistics by book (sorted by dimension).

Book	Fractal dimension (D)	R^2	FDIST
Gulliver's Travels (Swift 2009)	1.032885269	0.997971565	1.544E-06
Vermeer's Hat (Brook 2009)	1.023331752	0.996924745	3.55009E-06
The Two Princes of Calabar (Sparks 2004)	1.0148522	0.992653067	2.02913E-05
Plumes (Stein 2008)	1.01428358	0.997639195	2.09167E-06
1688. A Global History (Wills 2001)	1.00456998	0.997032804	3.30487E-06
The Inner Life of Empires (Rothschild 2011)	0.99646761	0.988888016	4.64762E-05
Beyond Good and Evil (Nietzsche 2009)	0.994695321	0.991815719	2.51873E-05

All the books in the collection showed a fractal dimension (D) slightly below and above 1 (Table 4) and a resemblance with the category of dusts (Mandelbrot 1983). Fractal dimension is considered an indicator of the degree of "change in detail" or complexity that becomes apparent with the "change in scale" (Karperien and Jelinek 2016: 20) or a way to describe, together with lacunarity, the "visual look" of a dataset (Tolle et al. 2003: 129). The books with the highest values of D are Swift (2009), Brook (2009) and Sparks (2004), which show a higher degree of dispersion of segments with the decrease in scale as compared with the others (Figures 7 and 15). While the books by Swift and Sparks exhibit lower values of average lacunarity (λ) (Table 7), Brook's displays a higher value that may be interpreted as a marker of higher variability in the size and structure of the gaps, and therefore a more complex pattern of detail unfolding with the decrease in scale. Sparks' and Stein's are very close in terms of fractal dimension but the texture of their segment distribution differs by a higher average occupation fraction (P) and a lower average lacunarity for the former and the reverse for the latter, which is also characterized by a simpler detail pattern with a concentration of mass on the second level and larger gaps. A simpler pattern, similar to Stein's, can also be observed for Wills', which has the highest value of average lacunarity discernible through large areas of empty space and a high density of segments on the last level at every scale. The values of D, λ and P of the last two books in Table 4 are somewhat harder to interpret: first, because both exhibit a value of D that is less than 1 (although by a very small amount), which brings them closer to the category of one-dimensional dusts, despite a relatively higher average occupation fraction of the grid (P) as compared with the others. What distinguishes the two books is a higher average lacunarity for Nietzsche, and thus a higher variability of gap configurations, while for Rothschild the average λ is the lowest of the whole collection, possibly due to a more homogeneous size and structure of the gaps. With these observations summarized, the question that arises is how these measures relate to the initial assumption of generality and specificity levels characterizing these texts from a conceptual point of view.

Differences were observed in the level of generality or specificity to which the segments corresponding to the largest scale were assigned in the first iteration (Table 5). This level may be interpreted as the *initiator* by analogy with the construction of theoretical fractals such as the Koch curve and Cantor dust (Figures 1 and 11). It is from this initial level that the dispersion of segments towards other levels begins when the reduction in scale is applied through the 6 iterations. This may suggest that from a global perspective, the words belonging to the initiator tend to group together in more compact or coherent clusters than those at other levels. An additional hypothesis may consider these levels as potential starting points in the writing process by the real or a hypothetical author, or by an automatic process of text generation.

As also illustrated in Figure 15 (read top-down), there are four books with an initiator corresponding to the more generic levels 1 and 2 (Rothschild, Swift, Stein and Nietzsche). The second row in the figure shows the distribution by level of segments with a size usually comparable to that of a unit (chapter or part). Recall that generic and specific levels are based on topics with respectively larger and smaller values of document entropy, which means broader or narrower unit coverage. Rothschild's initiator, placed on level 2, was assigned to a topic with large coverage and top words that refer to members of the Johnstone family (*john william james george betty alexander*) or to specific conditions, places and entities (*slaves scotland grenada east india company*). A closer look at the contexts of these words in the book revealed biographical details and fragments of letters and documents from various archives (via citations of primary and secondary sources), which can be associated with a microhistory perspective. For subsequent iterations, segments of smaller size spread either up, to level 1, corresponding to a topic with broad coverage (*johnstones empire information history slavery ideas*) and thus a view apparently closer to a macro-history perspective; or down, to deeper levels (i.e. 4–7) that cumulated more localized topics in terms of unit coverage but were variable in terms of micro- vs macro-historical standpoints (*henrietta illness anxious litigation; individuals historians enlightenment microhistory*). Rothschild's book therefore seems to be articulated around micro-historical characters and events, a conceptual unifying stratum discernible from a bird's-eye view and more localized macro-historical arguments that become visible in the layered representation only with a decrease in the scale of analysis. Stein's initiator was also placed on level 2, corresponding to a dominant topic (*ostrich feathers trade industry plumes jewish*). The decrease in scale resulted in the migration of segments either to the upper level

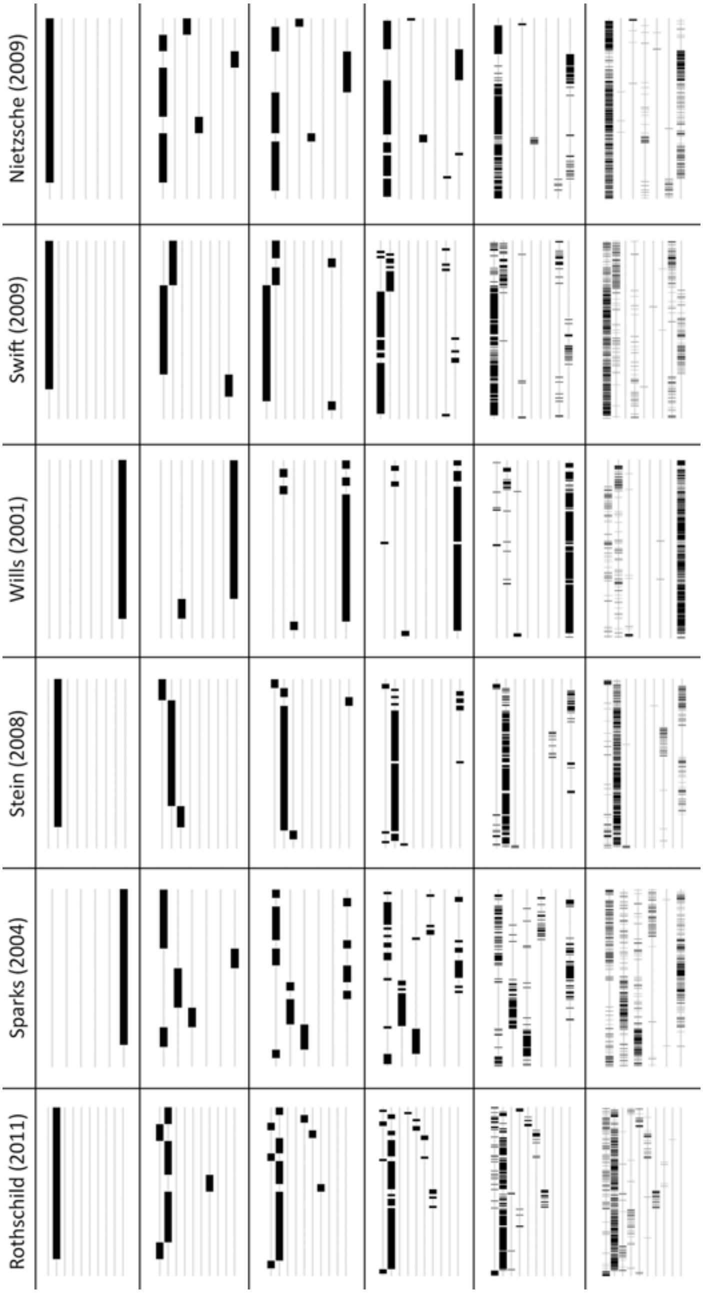


Figure 15: Segment distribution by level at different scales for six books.

Table 5: Initial level by book corresponding to the largest scale, iteration k = 1 (sorted by number of levels).

Book	Initial level	Total levels
The Inner Life of Empires (Rothschild 2011)	2	10
Gulliver's Travels (Swift 2009)	1	9
Plumes (Stein 2008)	2	9
1688. A Global History (Wills 2001)	8	8
Beyond Good and Evil (Nietzsche 2009)	1	7
Vermeer's Hat (Brook 2009)	7	7
The Two Princes of Calabar (Sparks 2004)	6	6

based on a general topic (*jews modern global commerce history commodity*) or to deeper levels (7, 9) corresponding to topics pertaining to manufacturers and the ostrich feather trade in different regions of the world, in Africa, America and Europe. Unlike Rothschild, Stein predominantly seemed to adopt a global history perspective with synthesis explanations that accompanied the microhistory accounts, noticeable at smaller scales of the visual representation (which I will call *map* from now on). The initiators in the case of Swift and Nietzsche occupied level 1, with a very generic topic (*great made time people good found hundred court*) for the first and a similarly generic one for the second (*man men good time great soul life taste world people love morality*). With the decrease in scale, the fluctuations of segments revealed some of the intermediary and deepest levels of the maps. For Swift, the dispersed segments were placed mostly on levels 2, 4, 8 and 9, attached to topics that went from more general (*country master reason nature honour*), through medium generality referring to objects, situations or institutions common to different places visited by the protagonist (*majesty majesty's left palace royal; yellow settled tolerable disposition*), to more localized aspects specific to a certain country or event (*yahoos houyhnhnms yahoo; emperor blefuscu imperial; island king luggnagg*). For smaller scale iterations in Nietzsche's text, the intermediary level 4 and the specific levels 6 and 7 were more prevalent. These levels referred to topics with medium coverage in the book (*recognized process psychological; unconditionally utility experienced; love woman vanity*) and to more localized, philosophical themes *(judgments sensation faculty impulses*; *skepticism germany greatness scientific; morals morality herd gregarious).*

The other three books (Brook, Sparks, Wills) exhibited a different pattern containing an initiator on the deepest level and segment dispersion gradually involving the upper levels with the decrease in scale (Figures 7 and 15). For Brook, also discussed in Sections 3 and 4, the second iteration produced a movement up of the segments corresponding to chapters 1, 4 and 6, i.e. from level 7 to level 1 for the first and level 6 for the two others. These chapters were the first to move since they were probably less stable with their most specific topics (T18, T7, T1) belonging to level 6 (Figures 4, 5, 14). The next iterations produced fluctuations of segments from these chapters between levels 6, 1 and 2, while the segments from the other chapters mainly fluctuated between levels 7, 1 and 2. Excerpts of top words from the most generic and specific topics of this book (Table 1) and the movements of segments (Figure 7) suggest a prevalence of the micro-historical perspective at larger scales, gradually balanced by global history arguments that become more discernible on the map with the decrease in segment size. Sparks' case was more intriguing since it displayed an initiator placed on level 6, and in the second iteration, a movement up to levels 1, 2 and 3 of most of the chapter segments, except for chapter 3 (Figure 15). A possible explanation of the relative stability of chapter 3 on the last level resides in its higher percentage of specific topics (T11, T19) (see also Figure 14). Although globally attached to the most specific level, the book exhibited a variable pattern at smaller scales. Thus, for subsequent iterations, the first third fluctuated around the top levels 1, 2 and 3 with topics mainly related to the slave trade (*robin johns slave trade; traders slaves trade calabar; town king english captains; atlantic africa numbers individuals*). The two other thirds showed an increasing density of segments on levels 4 (end) and 6 (middle), corresponding to intermediary or narrower themes and events (*wesley charles africans god christianity; roseau mortality shore believed night*). This threefold pattern suggests a certain similarity with Swift's distribution, also displaying a higher concentration of segments on the last level in the middle of the book; this may correlate with Sparks' argument unfolding style, which seems closest to that of a storyteller of all the five history books analyzed. Wills' map showed the simplest configuration, with an initiator on level 8 and fluctuations involving levels 1 and 2 with the decrease in scale (Figure 15). What is surprising about Wills' text is the higher segment density on the deepest level in contrast with sparser inserts at the upper levels at every scale, as compared with the other books. This pattern can be explained by the "baroque" composition of the book (as also suggested by its "Baroque Prelude"), intended to create the "portrait" of one year, 1688, from discrete depictions of people, places and events around the world at that particular time. The global perspective is therefore constructed by general topics from the upper levels 1, 2 and 3 (*time world year good long work; great people power made years trade; world voices voice sense human baroque*), while the bottom level 8 cumulates topics with more precise but narrower scope (*jews thy children jerusalem; spanish slaves coast portuguese; william king england james; muslim mughal ottoman hindu*). The relative sparsity of segments on the upper levels and the abundance of mass on the lowest level can therefore be attributed to Wills' method itself, based on "[s]erendipity, surprise, and letting one thing lead you to another" (2001: XI), which involves the author less and the reader more in making connections and fitting together the pieces of the global history puzzle of 1688.

5.3 Informational granularity maps

The combined approach of topic modelling and fractal geometry led me to scalable representations of the analyzed texts (Figures 7 and 15) and their generic-specific dynamics, representations that I will call *informational granularity maps*. The term *map* was inspired by Bjornson's (1981) "cognitive mapping" and its role in the "comprehension of literary texts". Bjornson distinguishes two modes of thought in the elaboration of a textual image. The first refers to the construction by readers of a "general idea" or "image" about what they are reading, i.e. "a poem, a play, a novel" and "how it can be expected to operate as they read". This general idea is then made more specific with the progression of the reading process, in the same way as "archeologists confronted by a heap of potsherds, start with a general idea about the nature of pottery and gradually refine that idea as they reconstruct a particular pot" (1981: 58). The second mode of thought is related to the hypotheses readers make about the world and the confirmation, alteration or denial of these hypotheses as they continue to read, operations through which "information is added to the textual image by assimilation and accommodation". These two modes of thought would therefore produce a flexible cognitive construct, a "schematized map of the text and its imaginary territory – a map that facilitates remembering what has been read". Bjornson also assumes that although these constructions will differ from person to person, there are invariant features that "tend to recur in different readers' mapping of the same text" (1981: 59). Ryan used Bjornson's concept of cognitive maps in a narrower sense, referring to a "mental model of spatial relations" (2003: 215). She conducted a series of experiments with high school students, who were asked to draw maps of the story world of the *Chronicle of a Death Foretold* by Gabriel García Márquez, to investigate the readers' mental construction of the narrative space. Ryan notes that text processing, in the process of reading, operates at different levels, "words, sentences, paragraphs, passages", to which one may add the level of the "global meaning or narrative macro-structure" (2003: 234). She also discerns two types of memory involved in the reading process, the "long-term memory", where the global representation of a text is stored, and the "sketch-pad of short-term, or episodic, memory" affected by "smaller textual units", where the readers form their "most detailed visualizations" or "picture-like representations" (2003: 234). Based on the results of her experiments, Ryan concludes that early in the reading process readers create a global but schematic representation of the spatial configuration of the textual world, and that they then concentrate on the plot, characters and visualization of the current scene, without the need to reorganize the whole map, which remains relatively resistant to new input. According to Ryan, this would explain the differences but also the common elements in the students' sketches that, although not completely identifiable to cognitive maps, may document the selective work of long-term memory.

My visual representations (Figures 7 and 15) did not involve experiments with readers and their cognitive constructs of the textual world or the spatial configurations expressed within it; instead they were built through automatic analysis, namely, from the perspective of the texts themselves and a particular type of *information* carried by them, independently of the readers. Thus, the term *informational* was chosen, also inspired by studies in information and communication theory (Shannon 1948; Dretske 1999; Resnik 1995). These visual representations were intended to illustrate the *informational granularity* of the analyzed texts. That is, how texts, cut into smaller and smaller units of analysis, may change their geometry according to the reconfiguration at different scales of the spectrum of generic to specific themes characteristic to each text. I considered these barcode-like representations as *maps*[27] that depicted how the initiators, and their positioning on a generic or specific level, encompassed at global scale a sort of "long-term memory" of the texts considered in their entirety. Shorter segments also exhibited a certain type of memory, identified as mainly anti-persistent, possibly indicating a tendency of the segments to fluctuate around dominant levels (or attractors) at different scales. It would be interesting to compare via dedicated experiments, e.g. inspired by cognitive map studies (Bjornson 1981; Ryan 2003), the levels assigned by the algorithm (formula 1) with the levels of generality or specificity assigned by human readers of the texts, for each of the six iterations considered in the project. The fractal particularities of these maps (dimension, lacunarity) need further analysis. However, they seem to suggest some correlations between the visual characteristics of segment dispersion and the strategy of argument or story unfolding of the books. This drew attention, for instance, to certain words and topics that synthesized and linked together the conceptual threads of the texts, as entities evenly distributed throughout the units of analysis (chapters or parts), or on the contrary, to elements that narrowed down the scope of the narrative through localized descriptions or focused analyses of detailed content. The topic modelling approach used in level detection offered a first glimpse into this type of layered conceptual structure, despite its inherent limitations related

[27] Some similarities with "genetic maps" were also observed (see for instance Fang et al. 2020: 4).

to topic instability and dependence on the choice of the number of topics. More general techniques should be investigated as potential alternatives, for instance those derived from information, entropy and energy theory (Shannon 1948; 1951; Marcus 1970; Onicescu 1966), the study of lexical cohesion and lexical chains (Morris and Hirst 1991; Barzilay and Elhadad 1997) and the analysis of rare word clustering (Tanaka-Ishii and Bunde 2016). The potential connections between thermodynamics, entropy, energy, the so-called "temperature of discourse" and the fractal dimension (Mandelbrot 1983: 347), and the stratified representations of texts and the dynamics of segment attraction to levels at various scales proposed in this study should also be further examined.

6 Conclusion and future work

The study proposed a method of text analysis that combined conceptual aspects from the model of zoomable text, topic modelling and fractal geometry. It was assumed that this type of methodology may assist in detecting different levels of generality and specificity in texts and reveal some characteristics of the assemblage of blocks of text, above the word level, at different scales of representation. Applications of such an approach can range from hermeneutics and discourse analysis to text (and possibly z-text) generation and summarization.

Further work will consist in deepening the analysis of measures such as lacunarity, fluctuation and long-range correlation in conjunction with that of fractal dimension. A closer examination of the limitations of the applied techniques (e.g. impact on the results of certain factors in box counting and topic modelling) and the applicability of alternative methods from other fields of research such as information theory, physics or genetics may also be envisaged.

Appendix

Table 6: Long-range correlation by iteration (k), number of steps in the linear region (l interval)$_{lin}$ and value of the slope (α), and corresponding R^2 and FDIST measures, for original (lin) and shuffled (lin-sh) data, and segment sizes ε with approximate number of corresponding sentences (s) or paragraphs (p).[28]

k	Measures Book	Brook (2009)	Rothschild (2011)	Sparks (2004)	Stein (2008)	Wills (2001)	Swift 2009)	Nietzsche (2009)
4	**ε**	560 ($7\frac{1}{2} \div 8$p)	579 ($7\frac{1}{2} \div 8$p)	257 ($2\frac{1}{2} \div 3p$)	453 ($5\frac{1}{2} \div 6\frac{1}{2}p$)	810 ($9 \div 11$p)	592 ($5 \div 6\frac{1}{2}p$)	378 ($4 \div 4\frac{1}{2}p$)
	(l interval)$_{lin}$	2–9	2–28	2–28	2–9	2–21	2–11	2–12
	Moves$_{all}$ (%)	no (60.29)	no (74.28)	no (67.64)	no (80.59)	no (89.55)	no (75.75)	no (86.95)
		up (20.58)	up (12.85)	up (17.64)	up (10.44)	up (4.47)	up (12.12)	up (5.79)
		down (19.11)	down (12.85)	down (14.70)	down (8.95)	down (5.97)	down (12.12)	down (7.24)
	α_{lin}	0.216384333	0.185040878	0.261332118	0.046823725	0.006867853	0.181172257	0.291666169
	R^2_{lin}	0.919798726	0.88342838	0.875057937	0.243531002	0.00451974	0.830157976	0.9693114
	FDIST	0.000166306	3.59628E-13	8.59407E-13	0.213958273	0.778237534	0.000244858	4.07101E-08
	$\alpha_{lin\text{-}sh}$	0.149986078	0.016463879	0.084144612	0.041551025	−0.050632471	0.113057825	0.170276019
	$R^2_{lin\text{-}sh}$	0.602186505	0.029770492	0.241980519	0.131817593	0.174296436	0.616295395	0.840515564
	FDIST$_{lin\text{-}sh}$	0.023587474	0.389453858	0.009155519	0.376707047	0.06702235	0.00714031	7.16962E–05
5	**ε**	140 ($2 \div 2\frac{1}{2}$p)	144 ($2 \div 2\frac{1}{2}$p)	64 ($\frac{1}{2} \div 1p$)	113 ($1 \div 1\frac{1}{2}$p)	202 ($2 \div 3$p)	148 ($\frac{1}{2} \div 1\frac{1}{2}p$)	94 ($\frac{1}{2} \div 1\frac{1}{2}$p)
	(l interval)$_{lin}$	2–30	2–50	2–97	2–36	2–31	2–20	2–40
	Moves$_{all}$ (%)	no (68.19)	no (72.24)	no (64.36)	no (77.39)	no (87.35)	no (67.82)	no (80.00)
		up (16.47)	up (13.30)	up (18.39)	up (11.49)	up (6.51)	up (15.50)	up (9.61)
		down (15.32)	down (14.44)	down (17.24)	down (11.11)	down (6.13)	down (16.66)	down (10.38)

[28] The number of sentences and paragraphs are indicative and represent rough estimations based on manual counting and the observation of one or two chapters from the beginning of the books. They are not systematic counts or average values representative of the entirety of the books.

	α_{lin}	0.191770271	0.234109443	0.250216607	0.068862996	0.138374811	0.201002756	0.222016254
	R^2_{lin}	0.89637017	0.972306688	0.921194349	0.725516927	0.893800989	0.983533176	0.954688903
	FDIST	8.19074E-15	2.91902E-38	1.17451E-53	8.70955E-11	3.65465E-15	1.33272E-16	1.83752E-26
	$\alpha_{lin\text{-}sh}$	0.234708099	0.216825884	0.133893403	0.179243835	0.055466597	0.156645701	0.07041099
	$R^2_{lin\text{-}sh}$	0.946291969	0.969363938	0.552195165	0.950569927	0.544944655	0.90812337	0.730288518
	$\text{FDIST}_{lin\text{-}sh}$	1.1192E-18	3.13713E-37	4.37272E-18	3.98671E-23	3.21971E-06	3.0635E-10	4.47479E–12
6	**ε (s or p)**	35 ($2\frac{1}{2} \div 3$s)	36 ($2\frac{1}{2} \div 3$s)	16 ($\frac{1}{2} \div 1$s)	28 ($1 \div 1\frac{1}{2}$s)	50 ($2\frac{1}{2} \div 3$s)	37 ($2 \div 2\frac{1}{2}$s)	23 ($2 \div 4\frac{1}{2}$s)
	$(\text{l interval})_{lin}$	2–509	2–344	2–197	2–221	2–201	2–153	2–240
	Moves_{all} (%)	no (62.58)	no (58.55)	no (49.27)	no (68.58)	no (72.30)	no (51.11)	no (66.82)
		up (18.95)	up (20.38)	up (25.41)	up (15.65)	up (14.03)	up (24.24)	up (16.68)
		down (18.46)	down (21.06)	down (25.31)	down (15.75)	down (13.65)	down (24.63)	down (16.49)
	α_{lin}	0.421862798	0.29979272	0.331368534	0.321025	0.386141823	0.264320874	0.150407153
	R^2_{lin}	0.985846869	0.97521342	0.979490072	0.91998802	0.965578685	0.986571656	0.9892173
	FDIST	0	7.1547E-276	1.0531E-165	1.5633E-121	8.0632E-147	2.6167E-142	3.9323E-235
	$\alpha_{lin\text{-}sh}$	0.224735648	0.189948059	0.252615074	0.23359415	0.287609244	0.374482602	0.279634208
	$R^2_{lin\text{-}sh}$	0.877148638	0.96534973	0.922715812	0.920340097	0.939106299	0.97994726	0.973250293
	$\text{FDIST}_{lin\text{-}sh}$	1.5528E-232	4.6008E-251	8.3091E-110	9.6651E-122	2.7493E-122	3.0237E-129	2.2764E-188

Table 7: Average lacunarity (λ) by iteration (k), side of the gliding box (r) and occupation fraction (P).

k	Measures Book	Brook (2009)	Rothschild (2011)	Sparks (2004)	Stein (2008)	Wills (2001)	Swift (2009)	Nietzsche (2009)
1	**avg_λ** (r=1, 2)	2.5	2.5	2.5	2.5	2.5	2.5	2.5
	P (occupied fraction)	0.25	0.25	0.25	0.25	0.25	0.25	0.25
2	**avg_λ** (r=1, 2, 4, 5)	2.538580247	1.5521262	1.789035467	2.090974625	9.726666667	2.179421769	1.588036704
	P (occupied fraction)	0.24	0.36	0.32	0.24	0.24	0.24	0.32
3	**avg_λ** (r=1, 2, 4, 8, 17)	4.671352876	3.978259635	4.460605475	5.402852824	11.95368659	5.711927248	5.190134757
	P (occupied fraction)	0.08650519	0.093425606	0.089965398	0.07266436	0.07266436	0.07266436	0.076124567
4	**avg_λ** (r=1, 2, 4, 8, 16, 65)	15.83437841	15.10998148	15.5291666	17.6682462	29.01423336	16.95334049	17.30767549
	P (occupied fraction)	0.021775148	0.020118343	0.020591716	0.018461538	0.01704142	0.019171598	0.017514793
5	**avg_λ** (r=1, 2, 4, 8, 16, 32, 257)	63.83729508	50.52015076	52.41730563	55.70682809	93.71280392	53.09860446	62.7550491
	P (occupied fraction)	0.005132553	0.005026571	0.005314236	0.004769187	0.00439068	0.005147693	0.004693485

6	**avg_λ** (r=1, 2, 4, 8, 16, 32, 64, 1025)	205.8479551	165.1325696	163.611997	193.8029156	243.8531417	161.2342388	196.8867079
	P (occupied fraction)	0.00134301	0.001390601	0.001481975	0.001298275	0.001261154	0.001453421	0.001334444
1–6	**avg_λ**	49.20492696	39.79884795	40.05135169	46.1953029	65.12675538	40.27958879	47.70460066
	avg_P (occupied fraction)	0.10079265	0.121660187	0.114558888	0.09786556	0.097559602	0.098072845	0.111611215

References

Allain, C., and M. Cloitre. “Characterizing the Lacunarity of Random and Deterministic Fractal Sets.” *Physical Review* A 44.6 (1991): 3552–3558.

Armaselu (Vasilescu), Florentina. “Le livre sous la loupe. Nouvelles formes d’écriture électronique” (PhD thesis, Université de Montréal, 2010). Accessed July 1, 2023. https://papyrus.bib.umontreal.ca/xmlui/handle/1866/3964.

Armaselu, Florentina, and Charles Van den Heuvel. “Metaphors in Digital Hermeneutics: Zooming through Literary, Didactic and Historical Representations of Imaginary and Existing Cities.” *Digital Humanities Quarterly (DHQ)* 11.3 (2017). Accessed July 1, 2023. http://www.digitalhumanities.org/dhq/vol/11/3/000337/000337.html.

Barthes, Roland. *S/Z*. New York: Hill and Wang, 1974.

Barzilay, Regina, and Michael Elhadad. “Using Lexical Chains for Text Summarization.” *Intelligent Scalable Text Summarization* (1997): 10–17.

Bjornson, Richard. “Cognitive Mapping and the Understanding of Literature.” *SubStance* Vol. 10.1 Issue 30 (1981): 51–62.

Blei, David M. “Introduction to Probabilistic Topic Models.” *Communications of the ACM* 55 (2011). Accessed July 1, 2023. https://www.researchgate.net/publication/248701790_Introduction_to_Probabilistic_Topic_Models.

Blei, David M., Thomas L. Griffiths, and Michael I. Jordan. “The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies.” arXiv:0710.0845, 27 August 2009. Accessed July 1, 2023. http://arxiv.org/abs/0710.0845.

Brook, Timothy. *Vermeer’s Hat: The Seventeenth Century and the Dawn of the Global World*. London, UK: Profile Books, 2009.

Da Silva, David. “Caractérisation de La Nature Multi-Échelles Des Plantes Par Des Outils de Géométrie Fractale, Influence Sur l’interception de La Lumière.” (PhD thesis, Université Montpellier 2, 2008). Accessed July 1, 2023. https://www.researchgate.net/publication/256093785_Evaluation_des_caracteristiques_geometriques_d’une_structure_vegetale_dans_le_cadre_de_l’analyse_fractale.

Datseris, George, Inga Kottlarz, Anton P. Braun, and Ulrich Parlitz. “Estimating the Fractal Dimension: A Comparative Review and Open Source Implementations.” arXiv, 13 September 2021. Accessed July 1, 2023. http://arxiv.org/abs/2109.05937.

Di Ieva, Antonio, ed. The Fractal Geometry of the Brain. Springer Series in Computational Neuroscience. New York: Springer, 2016.

Dretske, Fred I. *Knowledge and the Flow of Information*, The David Hume Series, Philosophy and Cognitive Science Reissues, Leland Stanford Junior University: CSLI Publications, 1999.

Falconer, Kenneth. *Fractal Geometry: Mathematical Foundations and Applications*. West Sussex, United Kingdom: John Wiley & Sons, Incorporated, 2014.

Fang, Yunxia, Xiaoqin Zhang, Xian Zhang, Tao Tong, Ziling Zhang, Gengwei Wu, Linlin Hou, et al. “A High-Density Genetic Linkage Map of SLAFs and QTL Analysis of Grain Size and Weight in Barley (Hordeum Vulgare L.).” *Frontiers in Plant Science* 11 (2020). https://doi.org/10.3389/fpls.2020.620922.

Gouyet, J.-F. *Physics and Fractal Structures*. Paris: Masson Éditeur, 1996. Accessed July 1, 2023. https://vdoc.pub/documents/physics-and-fractal-structures-p7bgb9hco6c0.

Graham, Shawn, Scott Weingart, and Ian Milligan. “Getting Started with Topic Modeling and MALLET.” *Programming Historian*, 2012. https://doi.org/10.46430/phen0017.

Harrar, K, and L Hamami. “The Box Counting Method for Evaluate the Fractal Dimension in Radiographic Images.” In *6th International Conference on Circuits, Systems, Electronics, Control & Signal Processing* (CSECS’07), 6. Cairo, Egypt, 2007. Accessed July 1, 2023. https://www.researchgate.net/publication/254455405_The_Box_Counting_Method_for_Evaluate_the_Fractal_Dimension_in_Radiographic_Images.

Hart, Michael. “The Project Gutenberg Mission Statement.” 2004. Accessed July 1, 2023. https://www.gutenberg.org/about/background/mission_statement.html.

Hitchcock, Tim. “Big Data, Small Data and Meaning.” *Historyonics*, 9 November 2014. Accessed July 1, 2023. http://historyonics.blogspot.com/2014/11/big-data-small-data-and-meaning_9.html.

Ignatenko, V, S Koltcov, S Staab, and Z Boukhers. “Fractal Approach for Determining the Optimal Number of Topics in the Field of Topic Modeling.” *Journal of Physics: Conference Series* 1163 (2019). https://doi.org/10.1088/1742-6596/1163/1/012025.

Ijasan, Kolawole, George Tweneboah, and Jones Odei Mensah. “Anti-Persistence and Long-Memory Behaviour of SAREITs.” *Journal of Property Investment & Finance* 35.4 (2017): 356–368. https://doi.org/10.1108/JPIF-09-2016-0073.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer, 2017. https://doi.org/10.1007/978-1-4614-7138-7.

Johnson, Steven. “The long zoom.” *Seminars about Long-Term Thinking*, 11 May 2007. Accessed July 1, 2023. https://longnow.org/seminars/02007/may/11/the-long-zoom/.

Kaminskiy, Roman, Nataliya Shakhovska, Jana Kajanová, and Yurii Kryvenchuk. “Method of Distinguishing Styles by Fractal and Statistical Indicators of the Text as a Sequence of the Number of Letters in Its Words.” Edited by Marcin Hernes. Mathematics 9, no. 2410 (2021). https://doi.org/10.3390/math9192410.

Karperien, Audrey L., and Herbert F. Jelinek. “Box-Counting Fractal Analysis: A Primer for the Clinician.” In *The Fractal Geometry of the Brain* edited by Antonio Di Ieva, 13–42 New York: Springer, Springer Series in Computational Neuroscience, 2016.

Klinkenberg, Brian. “A Review of Methods Used to Determine the Fractal Dimension of Linear Features.” *Mathematical Geology* 26.1 (1994): 23–46. https://doi.org/10.1007/BF02065874.

Lafon, Pierre. “Sur la variabilité de la fréquence des formes dans un corpus.” *Mots* 1.1 (1980): 127–165. https://doi.org/10.3406/mots.1980.1008.

Mandelbrot, Benoit B. *The Fractal Geometry of Nature*. New York: W.H. Freeman and Company, 1983.

Marcus, Solomon, *Poetica matematică* (*Mathematical Poetics*), Bucharest, Editura Academiei Republicii Socialiste România, 1970.

McCallum, Andrew Kachites. *MALLET: A Machine Learning for Language Toolkit*. V. 2.0.8. University of Massachusetts Amherst. 2002. Accessed July 26, 2023. https://mallet.cs.umass.edu/download.php.

Moretti, Franco. *Distant Reading*. London, UK, New York, US: Verso, 2013.

Morris, Jane, and Graeme Hirst. “Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure.” *Computational Linguistics* 17.1 (1991): 21–48. https://doi.org/10.1016/B0-08-044854-2/05234-2.

Mueller, Martin. “Shakespeare His Contemporaries: Collaborative Curation and Exploration of Early Modern Drama in a Digital Environment.” *Digital Humanities Quarterly* (*DHQ)* 8.3 (2014). Accessed July 27, 2023. http://www.digitalhumanities.org/dhq/vol/8/3/000183/000183.html.

Najafi, Elham, and Amir H. Darooneh. “The Fractal Patterns of Words in a Text: A Method for Automatic Keyword Extraction.” Edited by Francisco J. Esteban. PLOS ONE 10.6 (2015): e0130617.

Nietzsche, Friedrich. *Beyond Good and Evil*. The Project Gutenberg eBook, 2009, reprint of the Helen Zimmern translation from German into English of “Beyond Good and Evil,” as published in The

Complete Works of Friedrich Nietzsche (1909–1913). Accessed July 24, 2023. https://www.gutenberg.org/ebooks/4363.

Onicescu, Octav, “Energie informationnelle”, *Comptes Rendus Acad. Sci.*, Paris, 263, 1966, 22, 841–842, cited in Marcus (1970).

Ostwald, Michael J., and Josephine Vaughan. *The Fractal Dimension of Architecture*. Cham: Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-32426-5.

Pareyon, Gabriel. “Fractal Theory and Language: The Form of Macrolinguistics.” In *Form and Symmetry: Art and Science Buenos Aires Congress*, 2007. Accessed July 23, 2023. https://www.mi.sanu.ac.rs/vismath/BA2007/sym79.pdf.

Pavlov, Alexey N., Werner Ebeling, Lutz Molgedey, Amir R. Ziganshin, and Vadim S. Anishchenko. “Scaling Features of Texts, Images and Time Series.” *Physica A: Statistical Mechanics and Its Applications* 300.1–2 (2001): 310–324. https://doi.org/10.1016/S0378-4371(01)00341-7.

Peltier, Jon. “Step Chart Without Risers.” *Peltier Tech. Peltier Technical Services – Excel Charts and Programming* (blog), 24 May 2008. Accessed July 1, 2023. https://peltiertech.com/line-chart-without-risers/.

Peng, C.-K., S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley. “Long-Range Correlations in Nucleotide Sequences.” *Nature* 356.6365 (1992): 168–170. https://doi.org/10.1038/356168a0.

Pérez-Mercader, Juan. “Scaling Phenomena and the Emergence of Complexity in Astrobiology.” In *Astrobiology*, edited by Gerda Horneck and Christa Baumstark-Khan, 337–360. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. https://doi.org/10.1007/978-3-642-59381-9_22.

Plotnick, Roy E., Robert H. Gardner, and Robert V. O’Neill. “Lacunarity Indices as Measures of Landscape Texture.” *Landscape Ecology* 8.3 (1993): 201–211. https://doi.org/10.1007/BF00125351.

Resnik, Philip. “Using Information Content to Evaluate Semantic Similarity in a Taxonomy.” arXiv: Cmp-Lg/9511007 November, 1995. Accessed July 1, 2023. http://arxiv.org/abs/cmp-lg/9511007.

Rothschild, Emma. *The Inner Life of Empires: An Eighteenth-Century History*. Princeton, US, Oxford, UK: Princeton University Press, 2011.

Ryan, Marie-Laure. “Cognitive Maps and the Construction of Narrative Space.” In *Narrative Theory and the Cognitive Sciences*, edited by David Herman, 214–242. Stanford, California: CSLI Publications, 2003.

Saha, Kunal, Vinodh Madhavan, and Chandrashekhar G. R. “Pitfalls in Long Memory Research.” In *Cogent Economics & Finance*, edited by David McMillan 8.1 (2020): 1733280.

Sapoval, Bernard. *Les Fractales. Fractals*. Paris: Aditech, 1989.

Shannon, Claude E. “A Mathematical Theory of Communication.” *The Bell System Technical Journal* 27 (1948): 379–423, 623–656.

Shannon, Claude E. “Prediction and Entropy of Printed English.” The Bell System Technical Journal, January 1951, 12.

Sławomirski, Mariusz R. “Fractal Structures and Self-Similar Forms in the Artwork of Salvador Dalí.” *Prace Instytutu Mechaniki Górotworu PAN*, Instytut Mechaniki Górotworu PAN 15.3–4 (2013): 131–146.

Sparks, Randy J. *The Two Princes of Calabar: An Eighteenth-Century Atlantic Odyssey*. Cambridge, Massachusetts, London, England: Harvard University Press, 2004.

Stein, Sarah, Abrevaya. *Plumes: Ostrich Feathers, Jews, and a Lost World of Global Commerce*. New Haven and London: Yale University Press, 2008.

Swift, Jonathan. *Gulliver’s Travels into Several Remote Nations of the World.* The Project Gutenberg eBook, 2009, first published in 1726. Accessed July 24, 2023. https://www.gutenberg.org/ebooks/829.

Tacenko, Natalija. “Fractal Theory of Discourse Construction: Some Hypothetic Ideas.” UDC 81’111, November 2016, 1–8. Accessed July 1, 2023. https://www.researchgate.net/publication/313403670_FRACTAL_THEORY_OF_DISCOURSE_CONSTRUCTION_SOME_HYPOTHETIC_IDEAS.

Tanaka-Ishii, Kumiko. “Long-Range Correlation Underlying Childhood Language and Generative Models.” *Frontiers in Psychology* 9 (2018): 1725. https://doi.org/10.3389/fpsyg.2018.01725.

Tanaka-Ishii, Kumiko, andArmin Bunde. “Long-Range Memory in Literary Texts: On the Universal Clustering of the Rare Words.” Edited by Tobias Preis. PLOS ONE 11.11 (2016): e0164658. https://doi.org/10.1371/journal.pone.0164658.

Tolle, Charles R., Timothy R. McJunkin, David T. Rohrbaugh, and Randall A. LaViolette. “Lacunarity Definition for Ramified Data Sets Based on Optimal Cover.” *Physica D: Nonlinear Phenomena* 179.3–4 (2003): 129–152. https://doi.org/10.1016/S0167-2789(03)00029-0.

Trivellato, Francesca. “Is There a Future for Italian Microhistory in the Age of Global History?” *California Italian Studies* 2.1 (2011): 25. Accessed July 24, 2023. http://escholarship.org/uc/item/0z94n9hq.

Underwood, Ted. *Distant Horizons. Digital Evidence and Literary Change*. Chicago and London: The University of Chicago Press, 2019.

Wills, John E., Jr. *1688*: *A Global History*. New York, London: W.W. Norton & Company, 2001.

Wu, Jiaxin, Xin Jin, Shuo Mi, and Jinbo Tang. “An Effective Method to Compute the Box-Counting Dimension Based on the Mathematical Definition and Intervals.” *Results in Engineering* 6 (2020). https://doi.org/10.1016/j.rineng.2020.100106.

Zipf, George Kingsley. *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. Mansfield Centre, CT: Martino Publishing, 2012, first published by Addison-Wesley Press, 1949.