Physica A 686 (2026) 131361

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Statistical properties of the rooted-tree encoding of $\mathbb{N}$ ☆

Pierluigi Contucci [a], Claudio Giberti [b], Godwin Osabutey [c,*], Cecilia Vernia [c]

[a] Department of Mathematics, University of Bologna, Via Zamboni 33, 40126, Bologna, Italy
[b] Department of Sciences and Methods for Engineering, University of Modena and Reggio Emilia, Via G. Amendola 2, 42122, Reggio Emilia, Italy
[c] Department of Physics, Informatics and Mathematics, University of Modena and Reggio Emilia, Via G. Campi 213/b, 41125, Modena, Italy

ARTICLE INFO

Keywords:
Number theory
Arithmetic structure
Planar rooted trees
Dyck words
Entropy
Zipf function
Correlation
Self-organisation
Deterministic language
Complexity

ABSTRACT

We prime-encode the natural numbers via recursive factorisation, iterated to the exponents, generating a corpus of planar rooted trees equivalently represented as Dyck words. This forms a deterministic text endowed with internal rules. Statistical analysis of the corpus reveals that the dictionary and the entropy grow sublinearly, compression shows non-monotonic trend, and the rank-frequency curves assume a stable parabolic form deviating from Zipf's law. Correlation analysis using mean-squared displacement reveals a transition from normal diffusion to superdiffusion in the associated walk. These findings characterise the tree-encoded sequence as a statistically structured text with long-range correlations grounded in its generative arithmetic law, providing an empirical basis for subsequent theoretical investigations and empirical ones with large language models.

1. Introduction

The sequence of natural numbers, when expressed through their iterated prime factorisation, gives rise to an ordered chain of rooted trees, a purely arithmetic structure that can be read as a symbolic text. In this representation, each number is translated into a rooted tree whose branches encode the Euclidean recursive decomposition [1,2], and the resulting corpus, once the prime labels are discarded, becomes a deterministic language written in Dyck words.

In this work, we treat that sequence of Dyck words as an empirical object. Rather than imposing a generative model, a stochastic hypothesis, or any notion of randomness, we record its observables directly: the growth of the dictionary, the rank-frequency distribution of tree types, the symmetry properties of the text, its entropy and compressibility, and the correlation structure extracted via associated walks. The approach is descriptive in the strict sense: the analysis is restricted to what is measured, without interpretation beyond the arithmetic process that generates the data.

The findings are consistent across scales. The dictionary of distinct Dyck words grows sublinearly, indicating a systematic reuse of structures and suggesting an implicit combinatorial grammar. The entropy increases with corpus size, yet remains well below the non-informative maximum, and the compression ratio exhibits a non-monotonic behaviour indicative of structural organisation. The rank-frequency curve stabilises into a parabolic form in log–log scale, reminiscent of hierarchical self-similar regimes found in other correlated corpora [3,4]. The analysis of walks and their associated mean-squared displacements, as introduced in text analysis [5], reveals a two-regime behaviour, from normal diffusion at short time scales to superdiffusion or quasi-ballistic at longer ranges, suggesting the presence of long-range deterministic correlations.

☆ This article is part of a Special issue entitled: 'Statistical mechanics for Artificial Learning' published in Physica A.

* Corresponding author.
E-mail addresses: pierluigi.contucci@unibo.it (P. Contucci), claudio.giberti@unimore.it (C. Giberti), gosabutey@unimore.it (G. Osabutey), cecilia.vernia@unimore.it (C. Vernia).

https://doi.org/10.1016/j.physa.2026.131361

Available online 4 February 2026

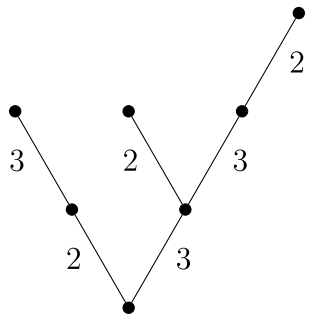


Fig. 1. Decorated tree representation of the integer $3099363912 = 2^3 \cdot 3^{18} = 2^3 \cdot 3^{2\cdot 3^2}$.

Together, these results provide an empirical characterisation of the arithmetic text as a structured, statistically organised language emerging from a fully deterministic process. The present study establishes the descriptive foundation for subsequent theoretical work and for investigations of learnability, in particular by methods of artificial intelligence like large language models trained on this corpus.

This paper is structured as follows. Section 2 introduces the tree encoding and the associated Dyck word representation. Section 3 reports the main statistical observations: dictionary growth and lexical reuse, directionality, entropy and compressibility, rank-frequency structure, and dynamical features captured through the mean-square displacement and cross-correlation of Dyck words. Section 4 summarises the conclusions and suggests future research avenues that emerge naturally from the results.

2. Tree encoding and dictionary analysis

The database studied in this paper is obtained by prime factorisation of positive integers, according to the Fundamental Theorem of Arithmetic. More precisely, every

$$n \in \mathbb{N} = \{1, 2, 3, \dots\}, \quad n > 1, \tag{1}$$

can be uniquely written as

$$n = p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}, \tag{2}$$

where $p_j \in \mathbb{P} = \{2, 3, 5, \dots\}$ are prime numbers and $n_j \in \mathbb{N}$, with

$$p_1 < p_2 < \cdots < p_k\,. \tag{3}$$

By convention,

$$n = 1 \tag{4}$$

corresponds to the empty product of primes.

In order to obtain a *pure-prime* representation of a natural number, the factorisation process may be iterated by decomposing the exponents in (2) into their own prime factorisations, and then proceed recursively [6]. In this way, every natural number is represented through primes only.

Then, for instance,

$$12 = 2^2 \cdot 3, \qquad 72 = 2^3 \cdot 3^2, \qquad 320 = 2^6 \cdot 5 = 2^{2\cdot 3} \cdot 5, \tag{5}$$

or, in a more elaborate case,

$$3099363912 = 2^3 \cdot 3^{18} = 2^3 \cdot 3^{2\cdot 3^2}. \tag{6}$$

In this way, we can *prime-encode* every natural number, expressing it entirely in terms of primes at all levels, using only products and exponentiation. It is a simple, yet remarkable observation that every natural number corresponds uniquely to a prime-decorated planar rooted tree (see Fig. 1). This correspondence can be regarded as a tree-theoretic refinement of the Fundamental Theorem of Arithmetic.

In this framework, primes correspond to the simplest planar rooted tree consisting of a single edge, and the study of the distribution of prime numbers becomes the study of the occurrence of such *undecorated* trees within the natural sequence. We then consider the sequence of all undecorated trees associated with successive natural numbers. The first ten examples are shown below:

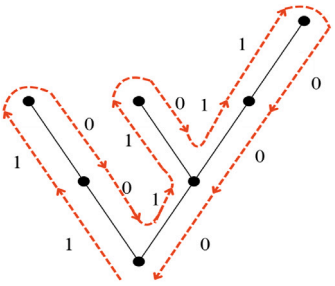
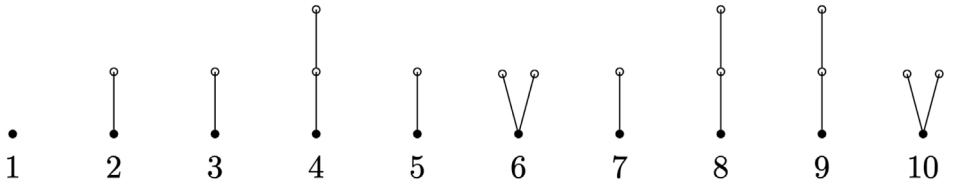


Fig. 2. A tree and the corresponding Dyck word 110011011000.



Apart from the empty tree, only three distinct types appear among the first ten natural numbers: those corresponding to primes p, to products of two primes $p \cdot q$, and to natural numbers of the form p^q.

For analytical purposes it is convenient to replace trees by their *Dyck word* encodings, namely balanced binary sequences of 1's and 0's. The one-to-one map associating a planar rooted tree to with Dyck sequence is obtained by circumnavigating the tree clockwise writing a 1 whenever the path moves up and a 0 whenever it moves down. Here is an explicit example (see Fig. 2):

This generates a natural sequence of Dyck words, and the first ten of them are

$$\emptyset, 10, 10, 1100, 10, 1010, 10, 1100, 1100, 1010 \tag{7}$$

In [7,8] it was observed that the infinite text of Dyck words exhibits a number of striking structural properties. Since there are infinitely many, each Dyck word appears infinitely many times throughout the sequence, because each undecorated tree can be decorated in infinitely many ways. Nevertheless, certain short *phrases* (finite sequences of Dyck words) occur only once, such as 10 10 or 1100 10, while others are expected to recur extremely often. For instance, the infinite recurrence of 10 1100 is equivalent to the still open Mersenne conjecture, the infinity of 1010 10 relates to the Sophie Germain conjecture, see, e.g. [8]. On the other hand, 1100 1100 is known to occur once, a fact established in 2004 with the proof of Catalan's conjecture [9].

The text is moreover oriented: if one takes a valid phrase and flips its order, the resulting string is often invalid, i.e. it never occurs. For example, 111000 10 occurs multiple times, whereas 10 111000 never appears. Overall, the text displays a complex structure and, in principle, an infinite number of rules that depend only on the underlying tree structure and not on its decoration. This naturally raises the question whether the sequence can be regarded as a *text* and, if so, whether such a text can be *learned* with the help of transformers used in Large Language Models of Artificial Intelligence.

We define $\mathbb{N}\mathcal{T}$ (Natural Text) the sequence of Dyck words and $\mathbb{N}D$ (Natural Dictionary) the ordered sequence of all the different Dyck words. This paper is devoted to the statistical properties of the two sets up to a maximal database of factorised natural numbers of 6.5 billions.

3. Text and dictionary statistical analysis

Given the Natural Text $\mathbb{N}\mathcal{T}$, we denote $\mathbb{N}\mathcal{T}_\ell^n$ the part of the text found from the position ℓ to n, and $\mathbb{N}D_\ell^n$ the Natural Dictionary found in the same part of the Natural Text. Then, for instance, $\mathbb{N}\mathcal{T}_1^{10}$ is the set of Dyck words given in (7) and $\mathbb{N}D_1^{10} = (10,\ 1100,\ 1010)$ (the Dyck word $\emptyset$ is not included in the Natural Dictionary since it appears only once in the sequence). Obviously, the Natural Dictionary size, i.e. its cardinality, satisfies $|\mathbb{N}D_1^n| \le |\mathbb{N}\mathcal{T}_1^n| = n$. In the sequel, we denote by d_n the size of $\mathbb{N}D_1^n$. We are now going to analyse $\mathbb{N}\mathcal{T}_1^{6.5\cdot 10^9}$, i.e. the dataset of Dyck words that has been obtained by the complete prime-encoding of the natural numbers up to $\overline{N} = 6.5 \cdot 10^9$.

3.1. Dictionary density

Our analysis of $\mathbb{N}\mathcal{T}_1^{6.5\cdot 10^9}$ starts with the computation of the dictionary density, i.e. $\delta_n = d_n/n$, as n increases. In Table 1 we report the size and density of the Natural Dictionary for some values of n, while in Fig. 3 $|\mathbb{N}D_1^n|$ is represented versus n. The power-law fit in the same figure highlights a sub-linear behaviour (which in linguistics is a well known phenomenon called Heaps' law [10–12]). More precisely, within the available data window, i.e. $\mathbb{N}\mathcal{T}_1^{6.5\cdot 10^9}$, the best power-law fit for the dictionary amplitude is given by $d_n \sim K n^\beta$, with $\beta = 0.234 \pm 0.005$, $K = 9.231 \pm 1.009$ and a coefficient of determination $R^2 = 0.9997$. Within the window $n \le 6.5 \cdot 10^9$ of the Natural Text, the best power-law fit of the dictionary density is $\delta_n \sim n^{-0.7656}$. However, we do not expect the power-law to be the correct asymptotic behaviour.

Table 1
Natural Text length n, the associated dictionary size $d_n = |\mathbb{N}D_1^n|$ and its density δ_n.

n	d_n	$\delta_n = d_n/n$
1	1	1
10^1	3	$3.0 \cdot 10^{-1}$
10^2	12	$1.2 \cdot 10^{-1}$
10^3	29	$2.9 \cdot 10^{-2}$
10^4	63	$6.3 \cdot 10^{-3}$
10^5	123	$1.23 \cdot 10^{-3}$
10^6	230	$2.3 \cdot 10^{-4}$
10^7	412	$4.12 \cdot 10^{-5}$
10^8	708	$7.08 \cdot 10^{-6}$
10^9	1195	$1.195 \cdot 10^{-6}$

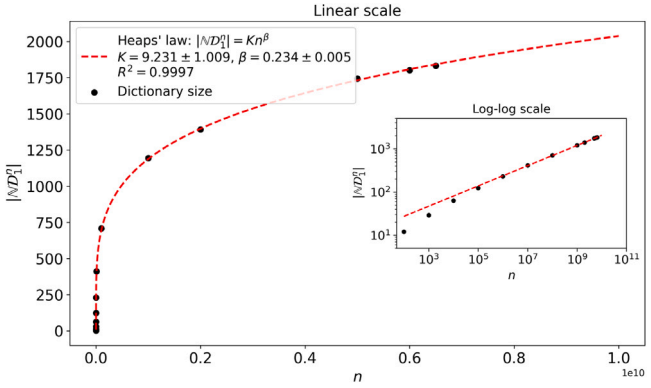


Fig. 3. The number of different Dyck words $d_n = |\mathbb{N}D_1^n|$ in the Natural Text from position 1 to n is represented versus n (black dots). The dashed line represents the power-law fit to the data. In the inset the same data are represented in a log–log plot.

3.2. Text orientation

As already highlighted in Section 2, the Natural Text presents a specific orientation, as shown by some examples. In this subsection, our aim is to provide statistical quantitative evidence of that fact by showing how it manifests itself across multiple scales in the Natural Text. To do this, we compute the fraction of ordered sequences of consecutive Dyck words that are present in $\mathbb{N}\mathcal{T}_1^n$ together with their mirror-symmetric version.

To clarify the definition, let us first illustrate the core idea with an example. Consider the following sequence:

$$\mathbb{N}\mathcal{T}_2^{10} = (10, 10, 1100, 10, 1010, 10, 1100, 1100, 1010). \tag{8}$$

For such a sequence, we list all distinct contiguous subsequences τ (tuples) of a fixed length k and check whether their mirror-reversed versions τ^{rev} also appear in the same window $\mathbb{N}\mathcal{T}_2^{10}$ of the Natural Text. For instance, considering the distinct 4-tuples in $\mathbb{N}\mathcal{T}_2^{10}$, one has that

$$\tau = (10, 10, 1100, 10) \text{ is a sub-sequence of } \mathbb{N}\mathcal{T}_2^{10}, \tag{9}$$

but its mirror-reversed version

$$\tau^{\text{rev}} = (10, 1100, 10, 10) \text{ is not a sub-sequence of } \mathbb{N}\mathcal{T}_2^{10}. \tag{10}$$

Enumerating all distinct contiguous 4-tuples of Dyck words in $\mathbb{N}\mathcal{T}_2^{10}$, we find 6 of them in total, of which 2 have their mirror-reversed versions also present. The ratio $\frac{2}{6}$ quantifies the *Statistical Symmetry* of the 4-tuples in the sequence $\mathbb{N}\mathcal{T}_2^{10}$.

We now generalise this concept. Writing

$$\mathbb{N}\mathcal{T}_1^n = (x_1, \dots, x_n) \tag{11}$$

where $x_i \in \mathbb{N}D_1^n$, we can introduce, for $1 \le k \le n$, the set formed by the *different* contiguous subsequences of length k appearing in $\mathbb{N}\mathcal{T}_1^n$, that is:

$$\Pi_k(\mathbb{N}\mathcal{T}_1^n) = \{(z_1, \dots, z_k) \in (\mathbb{N}D_1^n)^k \,|\, (z_1, \dots, z_k) = (x_i, \dots, x_{i+k-1})|$$

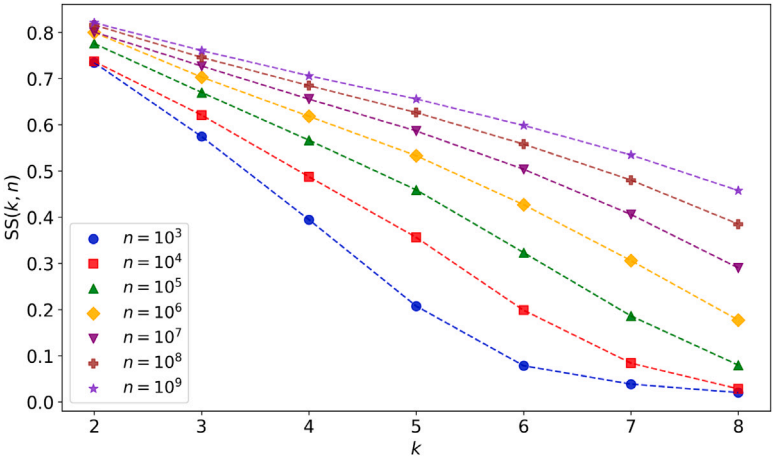


Fig. 4. Statistical Symmetry $\mathrm{SS}(k,n)$, defined in (14), as a function of the tuple size k for several values of the length of $\mathbb{N}\mathcal{T}_1^n$.

for some $1 \le i \le n-k+1\}$. (12)

Then, for each observed k-tuple $\tau = (z_1, \dots, z_k) \in \Pi_k(\mathbb{N}\mathcal{T}_1^n)$ we denote its reverse by

$$\tau^{\mathrm{rev}} = (z_k, \dots, z_1), \tag{13}$$

and define the *Statistical Symmetry* of order k at length n as the fraction of observed k-tuples in $\mathbb{N}\mathcal{T}_1^n$ whose reverse also appears in the sequence:

$$\mathrm{SS}(k,n) = \frac{1}{|\Pi_k(\mathbb{N}\mathcal{T}_1^n)|} \sum_{\tau \in \Pi_k(\mathbb{N}\mathcal{T}_1^n)} \mathbf{1}\Big(\tau^{\mathrm{rev}} \in \Pi_k(\mathbb{N}\mathcal{T}_1^n)\Big), \quad 1 \le k \le n, \tag{14}$$

where $\mathbf{1}(X)$ is the indicator of the event X. In a strictly oriented text, i.e. a text where no k-tuple has a reverse, this quantity is zero, while in the opposite case in which each Dyck word sequence appears in the text with its reverse, the Statistical Symmetry is one. In all other cases the quantity is between these two values, 0 and 1.

We have computed (14) for the length of the Natural Text $\mathbb{N}\mathcal{T}_1^n$ ranging from $n = 10^3$ to $n = 10^9$ and for k-tuples with $k = 2, \dots, 8$, see Fig. 4. The results show that, for fixed n, $\mathrm{SS}(k,n)$ is decreasing in k, which means that the longer the subsequence, the smaller the chance of finding its mirror-reversed version within the given window of the Natural Text. On the other hand, within the range of n values we have considered, the Statistical Symmetry $\mathrm{SS}(k,n)$ increases with n for fixed k. Nevertheless, the Statistical Symmetry is well below 1, since $\mathrm{SS}(k,n) \le \mathrm{SS}(k,10^9) < 0.821$, showing that the text $\mathbb{N}\mathcal{T}_1^{10^9}$ is oriented, at least with regard to subsequences that are not too long ($k \le 8$). For example, more than 29% of the sequences of length 4 and more than 54% of those of length 8 are non-invertible in $\mathbb{N}\mathcal{T}_1^{10^9}$. This suggests the existence of correlations between Dyck words, which will be examined in Section 3.5.

3.3. Complexity via entropy and compression

The fact that the dictionary density is decreasing approximately as $\delta_n \sim n^{-0.7656}$ for $n \le \overline{N}$ (see Section 3.1) implies that the number of distinct Dyck words, i.e. the dictionary size $|\mathbb{N}D_1^n|$, is significantly smaller than the total number of Dyck words in the Natural Text $\mathbb{N}\mathcal{T}_1^n$ of length $n \le \overline{N}$. As a consequence, some Dyck words are bound to appear more than once in the Text, and we are interested in computing the multiplicities of these occurrences. For this purpose, we introduce the *empirical frequency* of the Dyck word $\rho \in \mathbb{N}D_1^n$ which is defined as

$$p(\rho; \mathbb{N}\mathcal{T}_1^n) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\big(\rho = x_i\big), \tag{15}$$

where x_i is the ith Dyck word of the Natural Text $\mathbb{N}\mathcal{T}_1^n$. Thus, for instance, with reference to the text $\mathbb{N}\mathcal{T}_1^{10}$ given in Eq. (7), if $\rho = 10$, we have that $p(10; \mathbb{N}\mathcal{T}_1^{10}) = \frac{4}{10}$, while for $\rho = 1010$ we have $p(1010; \mathbb{N}\mathcal{T}_1^{10}) = \frac{2}{10}$.

Given the frequencies of all Dyck words in the dictionary $\mathbb{N}D_1^n$, we treat them as a probability distribution $\mathcal{P}_n = (p(\rho; \mathbb{N}\mathcal{T}_1^n); \rho \in \mathbb{N}D_1^n)$ from which we can compute the Shannon entropy of the Natural Text $\mathbb{N}\mathcal{T}_1^n$:

$$h(\mathbb{N}\mathcal{T}_1^n) = -\sum_{\rho \in \mathbb{N}D_1^n} p(\rho; \mathbb{N}\mathcal{T}_1^n) \log_2 p(\rho; \mathbb{N}\mathcal{T}_1^n), \tag{16}$$

which quantifies the amount of uncertainty (or information) contained in $\mathbb{N}\mathcal{T}_1^n$. Since the entropy of a m-component probability vector is not greater than $\log_2 m$, using the estimate $|\mathbb{N}D_1^n| \sim K n^{\beta}$ given in Section 3.1, we can compute an *entropic bound* for our

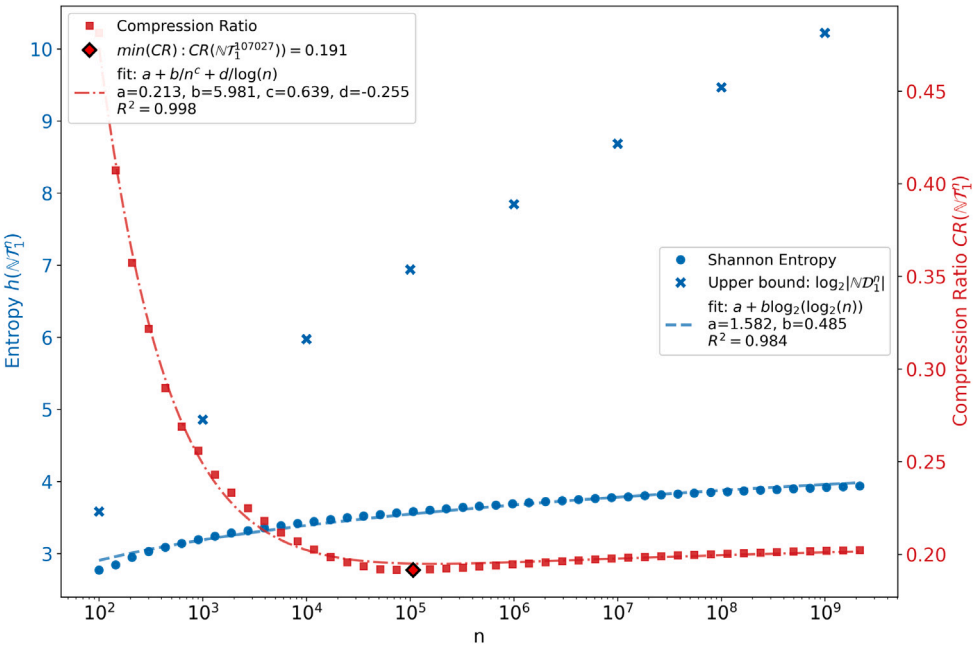


Fig. 5. Empirical entropy (blue dots) and compression ratio (red squares), as functions of sequence length n, together with model fits. Blue crosses represent the theoretical upper bound of entropy, given by $\log_2 |\mathbb{N}D_1^n|$. The rotated square marks the point where the compression ratio reaches its minimum.

ℕatural Text, as follows:

$$h(\mathbb{N}\mathcal{T}_1^n) \leq 3.206 + 0.234 \log_2 n \tag{17}$$

for $n \leq 6.5 \cdot 10^9$.

Fig. 5, where $h(\mathbb{N}\mathcal{T}_1^n)$ is represented as a function of n (blue bullets), shows that the entropy is closer to its maximum possible value for small n, but significantly lower for large n. In fact, the growth of entropy in the ℕatural Text $\mathbb{N}\mathcal{T}_1^n$ reveals a distinct scaling behaviour relative to its theoretical maximum. Although the empirical entropy increases from approximately 2.77 bits (at $n = 10^2$) to 3.94 bits (at $n \approx 2.16 \cdot 10^9$), it remains substantially below the corresponding upper bounds (see Eq. (17)), which range from 3.58 to 10.22 bits on the same scale. This persistent gap demonstrates that $\mathbb{N}\mathcal{T}_1^n$ possesses strong redundancy and a highly non-uniform distribution of Dyck words, indicating significant inherent structure rather than randomness.

It is appropriate to make some considerations regarding the computation of entropy and its validity as n varies. While the evaluation of entropy of $\mathcal{P}_n$ is virtually exact for fixed n since the empirical frequencies can be computed precisely (up to rounding errors), we clearly cannot extrapolate our results beyond the maximum length $\overline{N}$ of the window of the ℕatural Text available to us. It is well known [13] that the estimation of the entropy can be problematic even in 'standard' contexts — namely, for sequences with a finite number of distinct symbols occurring at well-defined and stationary frequencies. Indeed, the presence of complex and long-range correlations makes the evaluation of entropy nontrivial, since (15) and (16) generally underestimate this quantity due to the large fluctuations that may occur in (15) as n varies. In our case, the estimation is even more difficult, as we are dealing with a non-standard situation as evidenced, for instance, by the fact that Dyck word frequencies vanish as n becomes larger and larger.

To quantify the observed redundancy, we compute the compression ratio of $\mathbb{N}\mathcal{T}_1^n$ using the `gzip` algorithm [14]:

$$CR(\mathbb{N}\mathcal{T}_1^n) = \frac{C(\mathbb{N}\mathcal{T}_1^n)}{S(\mathbb{N}\mathcal{T}_1^n)} \tag{18}$$

where $C(\mathbb{N}\mathcal{T}_1^n)$ denotes the file size of the compressed representation of $\mathbb{N}\mathcal{T}_1^n$ and $S(\mathbb{N}\mathcal{T}_1^n)$ the original file size. The values of $CR(\mathbb{N}\mathcal{T}_1^n)$ closer to 0 indicate higher compressibility and redundancy, whereas those closer to 1 correspond to incompressible or random-like data.

The compression ratio (red squares in Fig. 5) decreases rapidly for small n, possibly reflecting the progressive discovery of some underlying rules and regularities associated with the integer factorisation. This phase corresponds to the identification of local redundancy and short-range correlations. A minimum is reached around $n = 10^5$, where the representation achieves maximal compactness, indicating that syntactic regularities are most efficiently captured. Beyond this point, the mild increase in $CR(\mathbb{N}\mathcal{T}_1^n)$ suggests the emergence of higher-level variability in which new, less frequent tree configurations and long-range dependencies reduce redundancy. A fit of the function $a + b/n^c + d/\log n$ to the data (which should not considered as asymptotic) highlights the slow growth of $CR(\mathbb{N}\mathcal{T}_1^n)$, see Fig. 5.

Table 2
Dyck words of rank 1, $\rho_n(1)$, to rank 5, $\rho_n(5)$, in $\mathbb{N}\mathcal{T}_1^n$ for several values of n.

n	$\rho_n(1)$	$\rho_n(2)$	$\rho_n(3)$	$\rho_n(4)$	$\rho_n(5)$
10	10	1100	1010	–	–
10^2	1010	10	110010	1100	101100
10^3	1010	10	110010	101010	11001010
10^4	1010	101010	10	110010	11001010
10^5	1010	101010	11001010	10	110010
10^6	1010	101010	11001010	10101010	10
10^7	101010	1010	10101010	11001010	10
10^8	101010	1010	10101010	11001010	1100101010
10^9	101010	1010	10101010	11001010	1100101010
$2 \cdot 10^9$	101010	1010	10101010	11001010	1100101010
$6.5 \cdot 10^9$	101010	1010	10101010	11001010	1100101010

The entropy trend corroborates this interpretation. Although $h(\mathbb{N}\mathcal{T}_1^n)$ increases, it remains significantly below its theoretical upper bound (blue crosses), confirming that the sequence is far from random. The growth of entropy is extremely slow, as can be appreciated from the fit with $a + b\log_2(\log_2(n))$ shown in Fig. 5 (once again, we emphasise that the fit is not intended to represent an asymptotic estimate). This sublinear growth implies that additional Dyck words contribute diminishing information per symbol, consistent with a constrained but generative structure.

Taken together, the entropy and compression analyses indicate a change in the statistical structure of the sequence as the Natural Text grows. While the system preserves a low-entropy organisation, the progressive loss of full compressibility signals the appearance of new, non-redundant configurations. This combination points to a regime in which structured correlations and long-range dependencies emerge, beyond simple repetitive order but without introducing randomness. The resulting behaviour is fully deterministic and reflects an increasing structural richness of the sequence as its length increases. For comparison, literary texts are commonly treated as approximately stationary, so that their statistical properties remain stable along the sequence and the compression rate is expected to decrease and converge toward the entropy rate of the source as the text grows. The ℕatural Text considered here does not satisfy these assumptions: its statistical properties evolve with the sequence length, and no convergence to a stationary entropy rate is observed.

3.4. The rank-frequency distribution of Dyck words: the Zipf function

In this section, to further characterise the statistical properties of the ℕatural Text, we examine the *rank-frequency distribution* $\mathcal{F}_n = (f_n(r),\, r = 1, \dots, d_n)$ of the empirical vector $\mathcal{P}_n$ whose components are given in (15) and whose dimension is d_n. The components $f_n(r)$ of $\mathcal{F}_n$ are obtained by rearranging those of $\mathcal{P}_n$ in decreasing order. More explicitly: $f_n(1) \geq f_n(2) \geq \cdots \geq f_n(r) \geq \cdots$ are the frequencies appearing in $\mathcal{P}_n$ and $r = 1, 2, \dots$ are the corresponding *ranks*. Fig. 6 displays the function $f_n(r)$, called *Zipf function*, for values of n ranging from 10 to $\overline{N}$. The figure highlights the remarkable fact that *the Zipf function $f_n(r)$ is, to a good approximation, independent of n*. Thus, for example, in every text $\mathbb{N}\mathcal{T}_1^n$ of length $n \leq \overline{N}$ the most frequent Dyck word, i.e. the one of rank 1, has an approximate frequency of 20%, while the frequency of that of rank 2 is approximately 15%.

We stress that this analysis shows that *while the frequencies (15) of individual Dyck words depend on n (see Section 3.3), the associated rank-frequency distributions remain approximately invariant. This invariance does not imply stability of the ranks of individual words, which may vary with n. Rather, it suggests that rank frequencies may admit a well-defined limit as $n \to \infty$, even when individual frequencies and ranks do not.*

Denoting $\rho_n(r)$ the Dyck word of rank r in $\mathbb{N}\mathcal{T}_1^n$, one can check that this Dyck word can change as n varies. As an example, Table 2 shows how the five most frequent Dyck words in $\mathbb{N}\mathcal{T}_1^n$, $\rho_n(1), \rho_n(2), \dots, \rho_n(5)$, change as n changes. Despite the fact that the table further suggests that $\rho_n(1)$ and $\rho_n(2)$ are constant for $10^7 \leq n \leq 6.5 \cdot 10^9$, as previously noted in other sections of this article, no definitive conclusions can be drawn regarding the asymptotic behaviour of the maps $\rho_n(r)$ in the limit as n tends to infinity.

We now examine in greater detail the nature of the Zipf function. A very common behaviour for the rank-frequency distribution encoded in the Zipf function $f(r)$, is described by the so called *Zipf's law*, in which the frequency $f(r)$ is inversely proportional to the rank r. Fig. 6 clearly shows that this law does not apply to our dataset. Instead, the decay exhibits a persistent curvature in log–log plot, forming a scale-independent parabolic envelope that extends over more than three orders of magnitude in rank. This behaviour is well approximated by another distribution, the *Parabolic Fractal Distribution* [15,16], according to which the logarithm of the frequency is a quadratic function of the logarithm of the rank.

We have fitted the rank-frequency distributions $\mathcal{F}_n$ to

$$\log f_n(r) = a_n (\log r)^2 + b_n \log r + c_n \tag{19}$$

for n ranging from 10^4 to $6.5 \cdot 10^9$. Fig. 7, in which the coefficients of the fits are reported versus n, shows that the quadratic coefficient a_n remains remarkably stable around -1 across all scales, corroborating the observation that the distributions $\mathcal{F}_n$ are almost independent of n, at least with respect to the dominant term $(\log r)^2$ in (19).

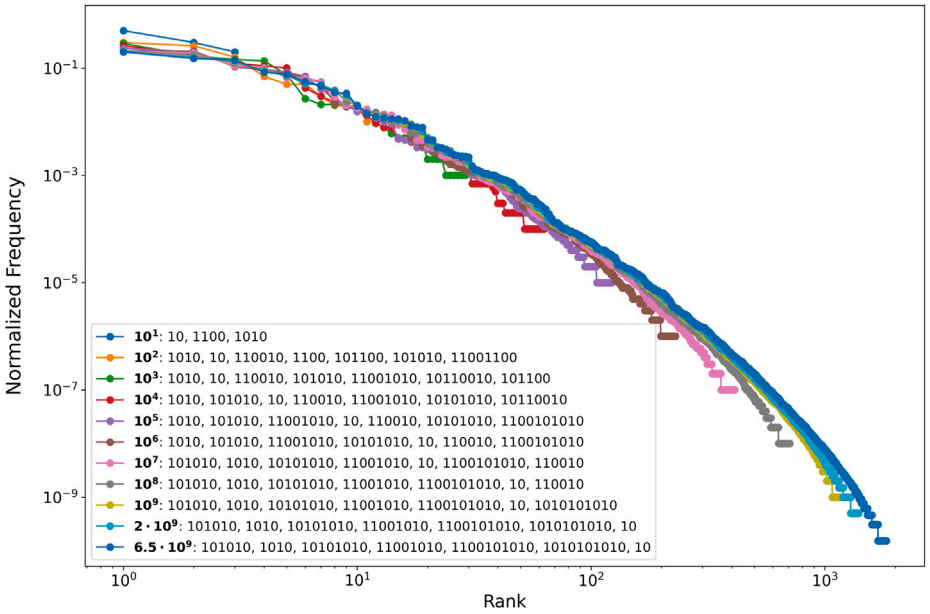


Fig. 6. Log–log plot of the rank-frequency distributions $\mathcal{F}_n$ (i.e. Zipf functions $f_n(r)$) versus the rank r, for n varying from 10 up to $6.5 \cdot 10^9$. In the legend, for each n, the most frequent Dyck words are reported.

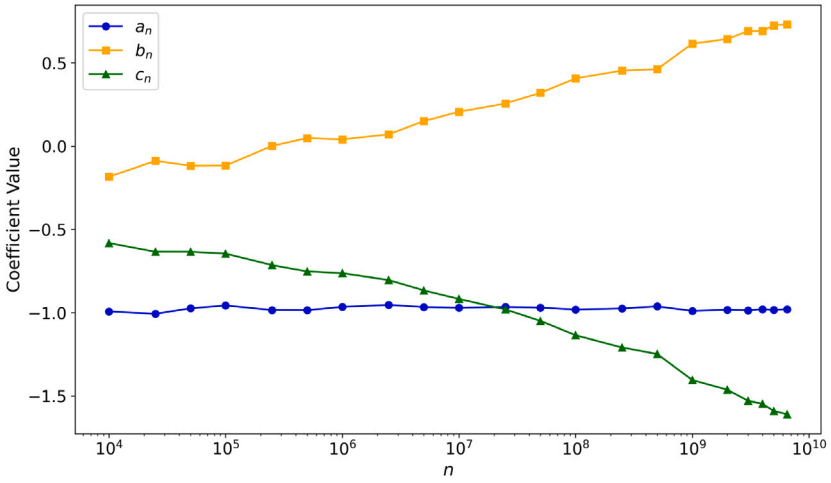


Fig. 7. Coefficients of the quadratic fits (19) for the rank-frequency distributions $\mathcal{F}_n$ (Zipf functions) versus n varying from 10^4 to $6.5 \cdot 10^9$. The coefficients a_n, b_n and c_n are represented by blue dots, orange squares and green triangles, respectively.

3.5. Correlation via mean squared displacement

In this section, we continue to study the distribution of occurrences of the Dyck words of the $\mathbb{N}$atural Text by using quantities such as the mean square displacement and the correlation functions, which are commonly employed in signals or stochastic processes analysis [17] or in the study of natural languages, see, e.g. [18,19].

Given a Dyck word $\rho \in \mathbb{N}D$, we construct the sequence $\mathbf{z} = (z_i)_{i=1}^n$ (the "signal") whose components are the indicator function:

$$z_i = \begin{cases} 1 & \text{if } x_i = \rho \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

where x_i is the ith Dyck word of the text $\mathbb{N}\mathcal{T}_1^n$, see (11). In order to construct a walk based on $\mathbf{z}$, we introduce the centred sequence $\mathbf{y}$ with components $y_i = z_i - \mu_n$, where $\mu_n = n^{-1} \sum_{i=1}^n z_i$, and define the position of a detrended walker $\mathbf{s}$ at "time" m as $s_m = \sum_{i=1}^m y_i$.

For the "temporal" dispersion of this walk from step m onward, we consider the displacement over a lag t:

$$\Delta s_{m,t} = s_{m+t} - s_m = \sum_{i=m+1}^{m+t} y_i \,. \tag{21}$$

Introducing the average $\langle\cdot\rangle$ over the $\mathbb{N}$atural Text, we compute the *mean square displacement* at time t of the walk as[1]:

$$\mathrm{MSD}_\rho(t) = \langle(\Delta s_{m,t})^2\rangle = \frac{1}{n-t}\sum_{m=1}^{n-t}\left(\sum_{i=m+1}^{m+t} y_i\right)^2 . \tag{22}$$

The mean square displacement typically exhibits a power-law behaviour,

$$\mathrm{MSD}_\rho(t) \sim t^\gamma \,, \tag{23}$$

where the exponent $\gamma > 0$ identifies the possible behaviours of the walker. In particular, according to the terminology of transport theory, we have that *normal diffusion* is characterised by $\gamma = 1$, while in anomalous cases we can recognise *subdiffusion* if $0 < \gamma < 1$, *superdiffusion* if $1 < \gamma < 2$, and *ballistic* behaviour if $\gamma = 2$. On the other hand, there are contexts in which transport is not characterised by a single exponent [20,21]. This may occur in the presence of multiple temporal regimes as, for instance, in ageing continuous time random walks (CTRW), see e.g. [22], in polymer dynamics [23] and in the behaviour of stock returns [24]. In such scenario, the mean square displacement shows a *crossover* at some time t_c:

$$\mathrm{MSD}_\rho(t) \sim \begin{cases} t^{\gamma_1} & \text{if} \quad t \ll t_c, \\ t^{\gamma_2} & \text{if} \quad t \gg t_c \,, \end{cases} \tag{24}$$

with $\gamma_1 \neq \gamma_2$.

We have computed the mean square displacement for several Dyck words, obtaining some evidence of the existence of two distinct qualitative behaviours related to the nature of the words themselves.

- Let us consider the Dyck words corresponding to *square-free numbers*, that is, natural numbers whose factorisation (2) contains each prime factor no more than once (i.e. $n_1 = \cdots = n_k = 1$). The trees corresponding to such numbers are "bushes" and their Dyck words are sequences of 10 blocks. We consider the following examples:

  $$\rho = 10,\ 1010,\ 101010,\ 10101010,\ 1010101010.$$

  The $\mathrm{MSD}_\rho(t)$ for these Dyck words are reported in Fig. 8. The log–log plots present piecewise linear behaviours, consistent with that described by Eq. (24), with parameters γ_1, γ_2, t_c depending on the considered Dyck word. In all examples, γ_1 is slightly greater than 1, but with confidence intervals of the fit that include it, while the exponent γ_2 is slightly smaller than 2. On the other hand, the crossover point t_c seems quite sensitive to the Dyck word ρ. Then, by adhering to the analogy with transport phenomena, we may say that walks associated with square-free numbers have a crossover from approximately normal or slightly superdiffusive to quasi-ballistic behaviours as the lag t crosses some critical value t_c.
- Some Dyck words associated with *non-square-free numbers*, i.e.

  $$\rho = 1100, 110010, 11001010, 1100101010, 110010101010,$$

  are considered in Fig. 9. Here too, the piecewise linear behaviour (in log–log plot) is evident, although with some differences compared to the previous case. In particular, while γ_1 is very close to 1 — slightly higher in some cases and slightly lower in others — γ_2 is consistently less than 2, even when the fit confidence interval is taken into account. In these cases, an almost normal behaviour is followed by a superdiffusive one (although not ballistic) as t increases through t_c.

The value of the exponents γ_1, γ_2, and of the crossover point t_c depend on the length n of the text $\mathbb{N}\mathcal{T}_1^n$ on which the computation is performed. The results for several values of n are shown in Fig. 10, for specific Dyck words corresponding to square-free and non-square-free numbers. More precisely, we considered 9 of the 10 most frequent Dyck words in the whole text $\mathbb{N}\mathcal{T}_1^{\overline{N}}$ (of which 4 are non-square-free), and the Dyck word 1100, which is the shortest non-square-free Dyck word (its rank is $r < 10$ only up to $\mathbb{N}\mathcal{T}_1^{100}$).

A behaviour similar to that described by Eq. (24) can be observed when considering a text of sufficiently large length n. The upper-left panel of Fig. 10 shows that the estimated value of the crossover point t_c varies significantly with n, and in most cases increases as n grows. While for some Dyck words t_c appears to approach an asymptotic value, in other cases — such as for the Dyck words 1100 and 1100101010 — the value of t_c continues to increase. Despite the variability of t_c, the values of γ_1 and γ_2 appear

[1] In practice, we evaluate $\mathrm{MSD}_\rho(t)$ for a logarithmically spaced set of lags. We set $t_{\min} = 10$ and $t_{\max} = \max\left(t_{\min}+1, \lfloor 0.1\cdot n\rfloor\right)$ then generate 80 log-spaced values between $t_{\min}$ and $t_{\max}$, round them to the nearest integer, and keep only unique values. Denoting this set by $\{t_j\}_{j=1}^J$, we compute

$$\mathrm{MSD}_\rho(t_j) = \frac{1}{n-t_j}\sum_{m=1}^{n-t_j}\left(s_{m+t_j} - s_m\right)^2, \quad j = 1,\ldots,J.$$

The choice $t_{\max} = 0.1\cdot n$ ensures that the average in (22) is taken over a sufficiently large number of terms while avoiding finite-length artifacts at very large lags.

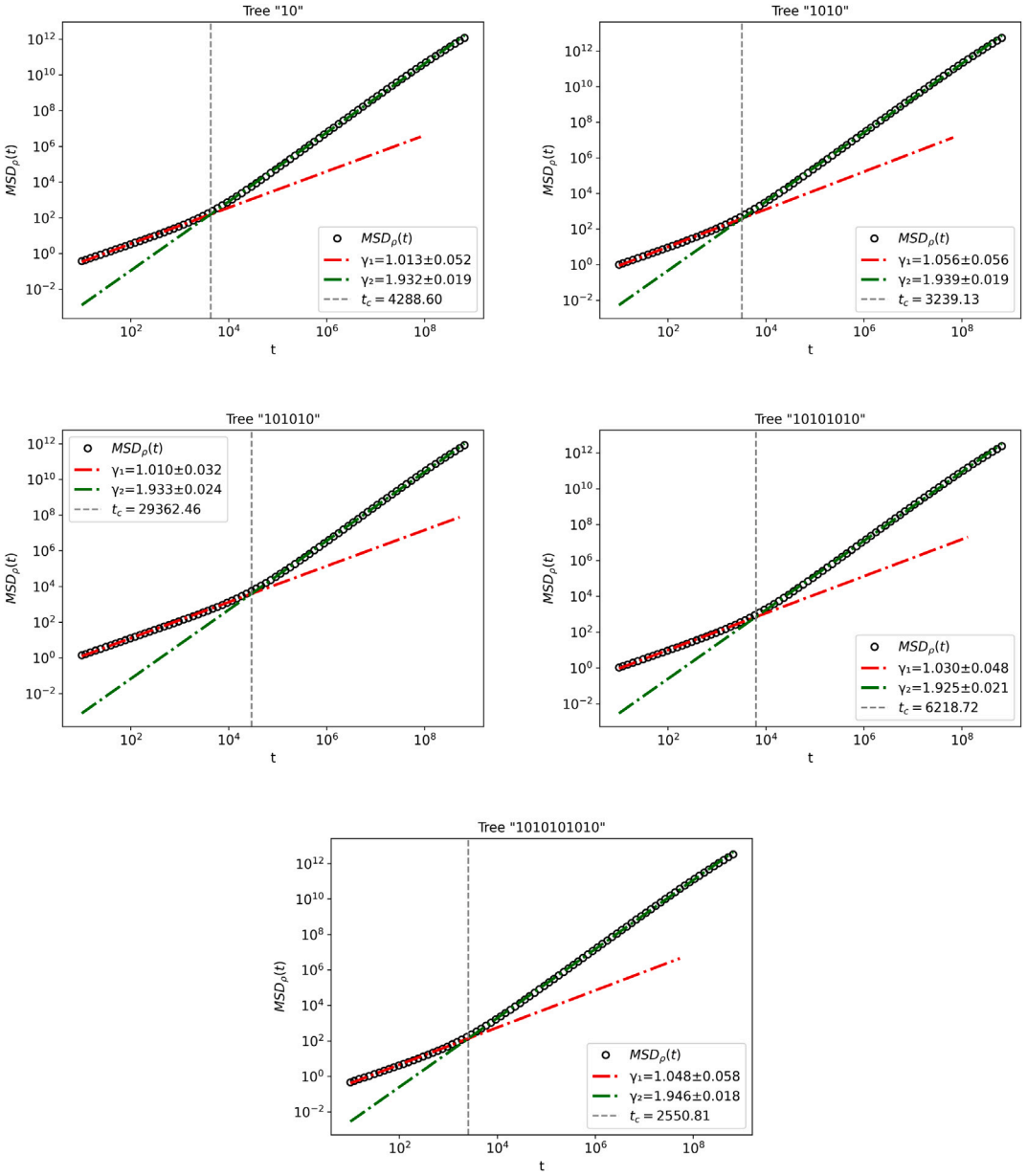


Fig. 8. Log–log plots of mean square displacement $MSD_\rho(t)$ for Dyck word (Tree) ρ associated with some square-free numbers versus the time lag t. The piecewise linear fits (dashed–dotted lines) illustrate the phase changes from an approximately normal to a superdiffusive (quasi ballistic) behaviour. The vertical dashed line represents the crossover point t_c. $MSD_\rho(t)$ is computed on the complete dataset, i.e., on the text $\mathbb{N}\mathcal{T}_1^n$ of length $n = 6.5 \cdot 10^9$.

overall closer to having reached their asymptotic regime, even in cases where t_c is still far from exhibiting limiting behaviour. The upper-right panel in Fig. 10 illustrates more explicitly that at short range (i.e. for $t < t_c$), the behaviour varies from one Dyck word to another: in some cases $\gamma_1 < 1$, while in others $\gamma_1 > 1$, thus revealing a spectrum of dynamics ranging from subdiffusion to weak superdiffusion. At long range ($t > t_c$), see the lower-left panel, the behaviour is clearly superdiffusive, as γ_2 is significantly greater than 1 — even in the case of the Dyck word 1100, where the exponent is markedly smaller than those of the other Dyck words.

Based on the present analysis, we are not in a position to formulate general statements regarding the potential monotonicity properties of the exponents γ_1 and γ_2 as functions of n. However, we observe that for the most frequent Dyck word (among the large values of n considered), namely 101010, the exponent γ_2 exhibits oscillatory behaviour, whereas γ_1 increases monotonically toward 1. We conclude by noting that γ_1 displays a variety of behaviours across the different Dyck words considered.

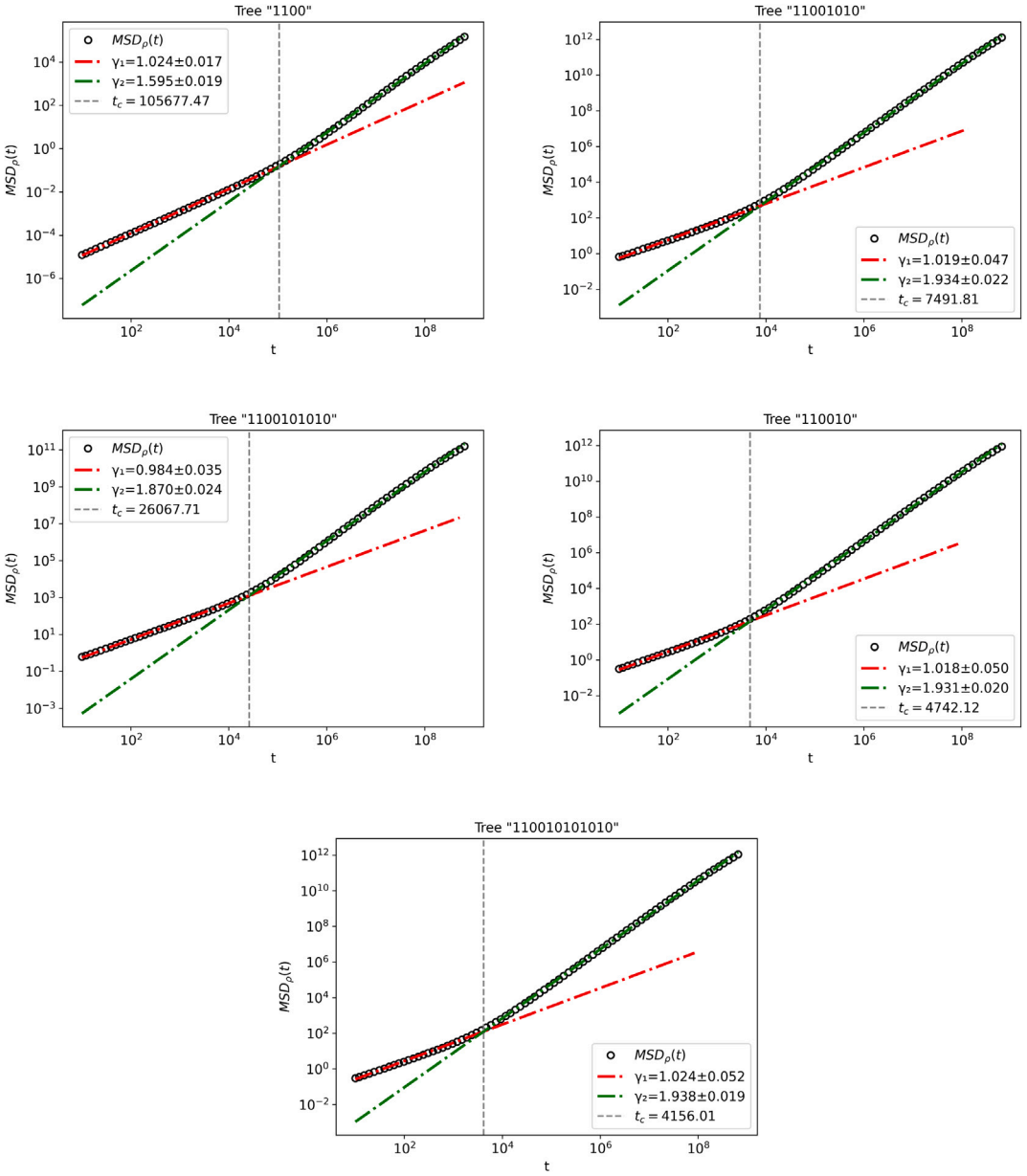


Fig. 9. Log–log plots of mean square displacement $MSD_\rho(t)$ for Dyck word (Tree) ρ associated with some non-square-free numbers versus the time lag t. The piecewise linear fitting (dashed–dotted lines) illustrate the phase changes from an approximately normal to a superdiffusive behaviour. The vertical dashed line represents the crossover point t_c. $MSD_\rho(t)$ is computed on the complete dataset, i.e., on the text $\mathbb{N}\mathcal{T}_1^n$ of length $n = 6.5 \cdot 10^9$.

3.6. Cross-correlation of Dyck words

An analysis analogous to that presented in Section 3.5 can be carried out by examining the correlation between the walks associated with two distinct Dyck words ρ_1 and ρ_2. Here we adopt the same notation introduced in Section 3.5, using the superscript j to denote quantities associated with the Dyck word ρ_j. Then, given the displacements of the walks $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ over a lag t:

$$\Delta s_{m,t}^{(j)} = s_{m+t}^{(j)} - s_m^{(j)} = \sum_{i=m+1}^{m+t} y_i^{(j)},$$

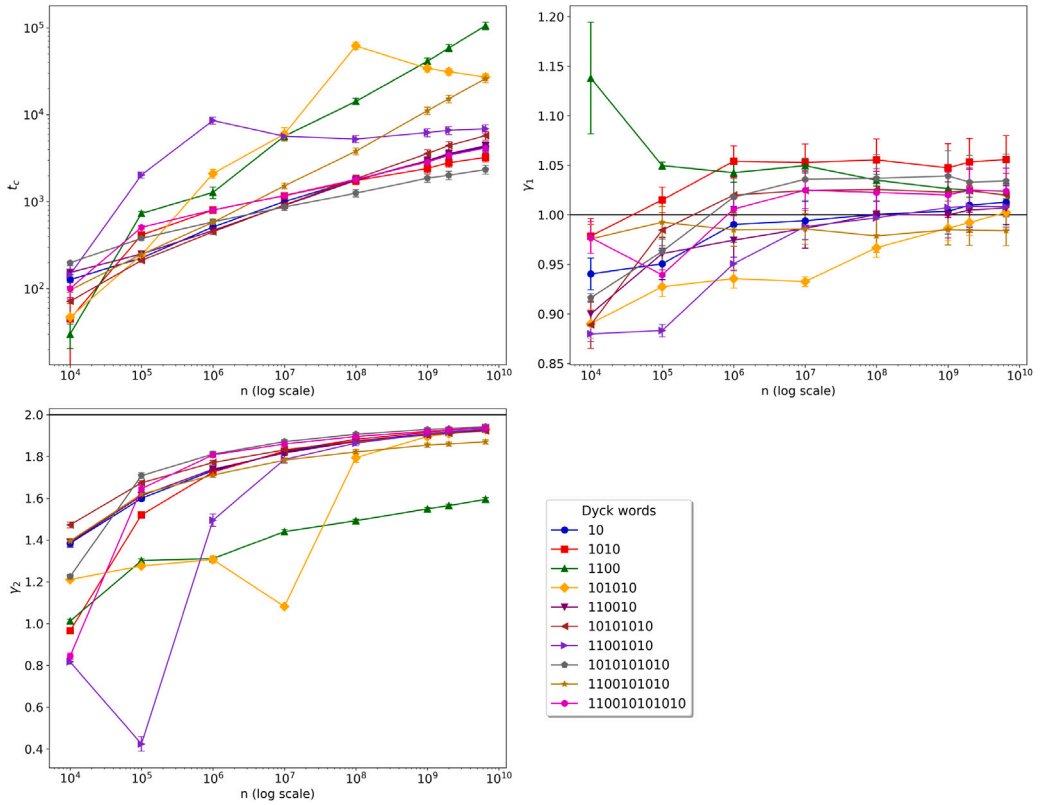


Fig. 10. Parameters t_c (left upper panel), γ_1 (right upper panel) and γ_2 (left lower panel) for several Dyck words (see the legend) computed on texts $\mathbb{N}\mathcal{T}_1^n$ of increasing length n. The horizontal axis represents the logarithm of n.

Table 3
Scaling exponents γ for square-free Dyck words.

Dyck word	Short-time scaling (γ_1)				
	10	1010	101010	10101010	1010101010
10	1.013	1.041	0.946	1.007	1.049
1010	1.041	1.056	1.028	1.041	1.053
101010	0.946	1.028	1.005	1.049	1.026
10101010	1.007	1.041	1.049	1.030	1.001
1010101010	1.049	1.053	1.026	1.001	1.049
Dyck word	Long-time scaling (γ_2)				
	10	1010	101010	10101010	1010101010
10	1.932	1.921	1.744	1.934	1.946
1010	1.921	1.939	1.899	1.939	1.949
101010	1.744	1.899	1.933	1.929	1.945
10101010	1.934	1.939	1.929	1.925	1.910
1010101010	1.946	1.949	1.945	1.910	1.946

their *cross correlation* is defined as:

$$\mathrm{MSD}_{\rho_1,\rho_2}(t) = \left\langle \left(\Delta s^{(1)}_{m,t} - \Delta s^{(2)}_{m,t}\right)^2 \right\rangle = \frac{1}{n-t}\sum_{m=1}^{n-t}\left(\sum_{i=m+1}^{m+t}\left(y_i^{(1)} - y_i^{(2)}\right)\right)^2. \tag{25}$$

As for the mean square displacement, in standard situations one may expect the cross-correlation to exhibit a power-law behaviour. However, computations performed for different choices of ρ_1 and ρ_2 show a double-regime power-law with a crossover, similar to what is observed in the mean square displacement, see Eq. (24).

Figs. 11 and 12 show the cross-correlations computed for some of the most frequent Dyck words in $\mathbb{N}D$. Specifically, the former considers pairs of square-free numbers, while the latter focuses on pairs consisting of one square-free and one non-square-free

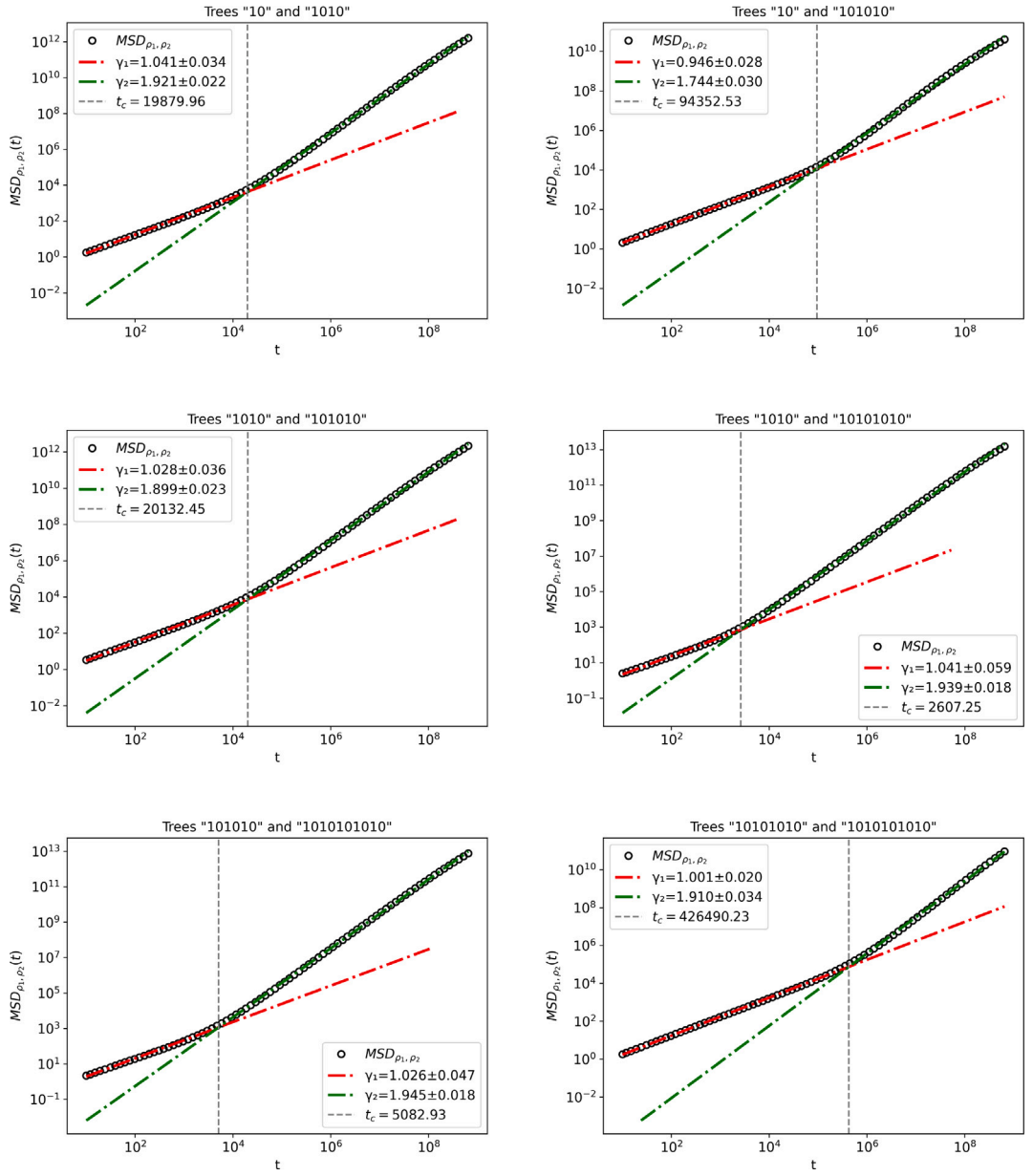


Fig. 11. Cross-correlation $MSD_{\rho_1\rho_2}(t)$ for Dyck words (Trees) ρ_1 and ρ_2 associated with some square-free numbers versus the time lag t on the log–log scale. The computation is performed on the complete dataset, i.e., on the text $\mathbb{N}\mathcal{T}_1^n$ of length $n = 6.5 \cdot 10^9$.

number, or two non-square-free numbers. The figures suggest that, at short range, the behaviour is weakly superdiffusive (though subdiffusive in one case), while at long range it becomes superdiffusive.

To provide insight into the variability of the exponents, we report some of their values in the tables below (see Table 3), which refer to the five Dyck words corresponding to square-free numbers: $\rho_1 = 10$, $\rho_2 = 1010$, $\rho_3 = 101010$, $\rho_4 = 10101010$ and $\rho_5 = 1010101010$. At the intersection of the row i and the column j, one finds the exponent γ of $MSD_{\rho_i,\rho_j}(t)$.

4. Conclusion and outlook

The analysis presented in this work examines the arithmetic sequence through its planar rooted-tree representation and, equivalently, through the associated symbolic text of Dyck words. The text is generated by iterating the Euclidean decomposition of natural numbers down to a prime-only representation, and is therefore a fully deterministic corpus. Its composition naturally includes

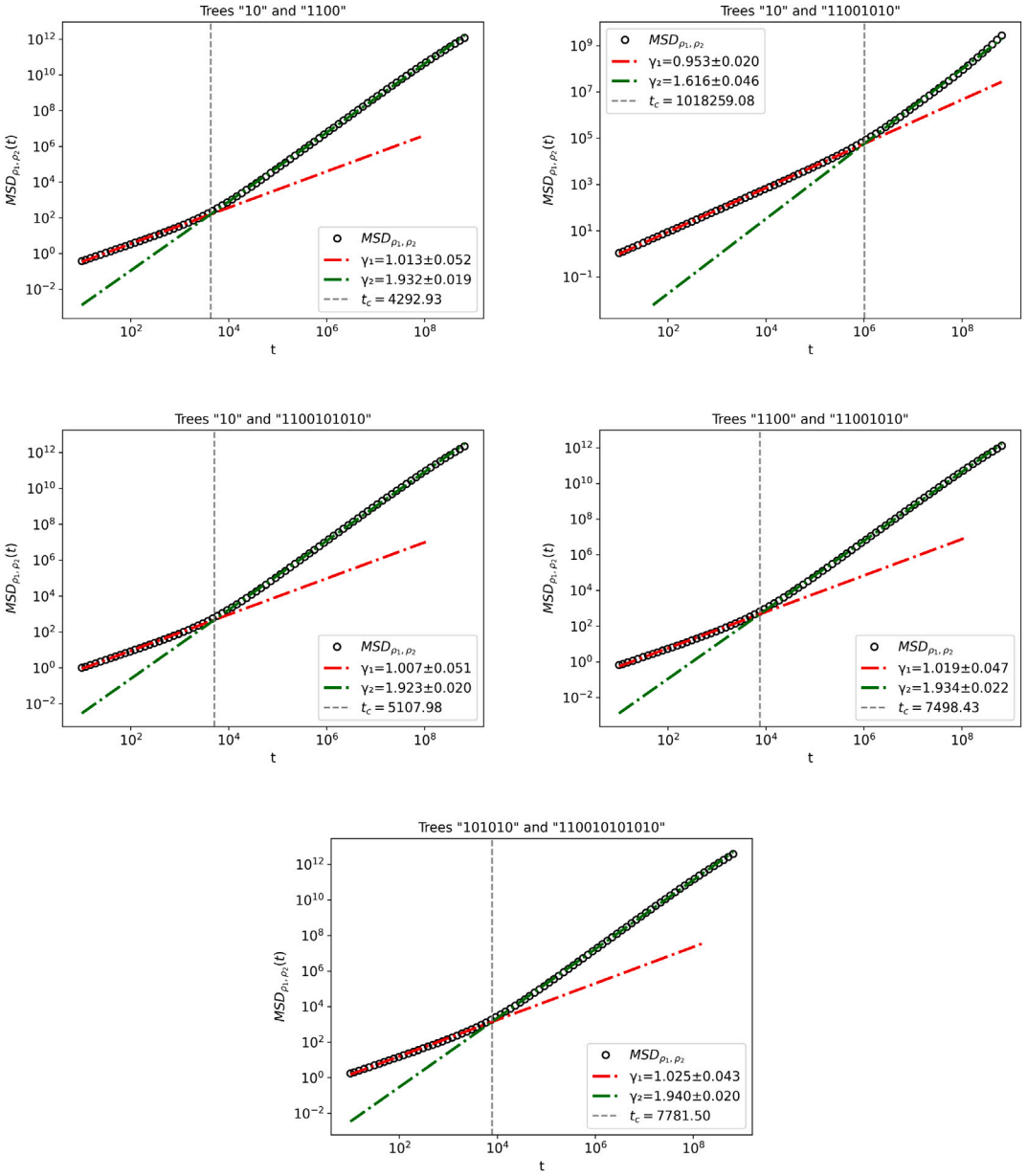


Fig. 12. Cross-correlation $MSD_{\rho_1\rho_2}(t)$ for Dyck words (Trees) ρ_1 and ρ_2 associated with some non-square-free numbers versus the time lag t on the log–log scale. The computation is performed on the complete dataset, i.e., on the text $\mathbb{N}T_1^n$ of length $n = 6.5 \cdot 10^9$.

primes and all other trees, and the resulting distribution displays structural organisation across multiple scales. Remarkably, these regularities emerge from a purely experimental analysis of the sequence. We found that the dictionary grows sublinearly, indicating sustained reuse of a limited set of combinatorial patterns. The entropy increases slowly and remains well below the theoretical bound, reflecting a high degree of redundancy. The compression ratio shows a non-monotonic behaviour, signalling a transition from local regularity to a broader form of organised complexity. The rank-frequency distribution is stable in shape and departs markedly from a Zipf law, being well approximated instead by a parabolic profile, in log–log plot, over several orders of magnitude in rank. The observed curvature, consistent with self-similarity between ranks [3,4], is analogous to departures from the classical Zipf behaviour documented in natural languages [25–28].

Transitions of the MSD from short-lag diffusion to enhanced scaling have been reported in correlated systems. Analogous behaviours were observed in dense active Brownian suspensions [29] and in disordered porous media [30,31]. A similar phenomenology was also identified in symbolic sequences, where long-range correlations in texts yield superdiffusive MSD exponents [19]. These examples indicate that MSD scaling beyond normal diffusion can arise in structured or correlated settings.

A different type of crossover occurs in processes where the transport exponent decreases with scale. In models with memory kernels, [32] reported a transition from ballistic to fractional-diffusive motion; a similar direction was found in active turbulence [33], and in ageing CTRW models where long-time behaviour becomes subdiffusive [34]. Here, the analogy lies not only in the presence of scale-dependent exponents but also in the fact that the short-time behaviour is approximately normal in both cases, rather than in the asymptotic transport regime.

Several directions follow naturally. First, the empirical regularities observed here call for theoretical explanations, particularly concerning the origin of the parabolic rank-frequency law and the mechanisms that determine the two-regime correlation structure. Second, the planar rooted-tree text offers a fully controlled benchmark for studying learnability in transformer models of artificial intelligence [35], where the role of determinism, hierarchy, and redundancy can be investigated. Finally, connecting the combinatorial depth of natural numbers with linguistic-type observables suggests a broader programme: to understand how the arithmetic structure expresses itself when viewed as a symbolic language. This study provides the empirical baseline for such developments.

CRediT authorship contribution statement

Pierluigi Contucci: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Funding acquisition, Conceptualization. **Claudio Giberti:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Godwin Osabutey:** Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Cecilia Vernia:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are grateful to Gabriele Sicuro for providing the tree-encoded database of the natural numbers. This research was performed under the auspices of Italian National Group of Mathematical Physics (GNFM) of the National Institute for Advanced Mathematics — INdAM. The authors acknowledge the financial support from the European Union — Next Generation EU - Grant PRIN 2022B5LF52. This project received support from the EU H2020 ICT48 project Humane AI Net (grant no. 952026), the Italian Ministry of University and Research PRIN 2022 (code J53D23003690006), and the Italian Extended Partnership PE01—FAIR (Future Artificial Intelligence Research, proposal code PE00000013) under the National Recovery and Resilience Plan. Claudio Giberti is a member of the Interdepartmental Centers En&Tech and InterMech at the University of Modena and Reggio Emilia, Italy.

References

[1] R.L. Childress, Recursive prime factorizations: Dyck words as numbers, 2021, arXiv preprint arXiv:2102.02777. URL: http://arxiv.org/abs/2102.02777.
[2] R.P. Stanley, Enumerative Combinatorics, second ed., in: Cambridge Studies in Advanced Mathematics, Cambridge University Press, 2011.
[3] D. Sornette, Discrete-scale invariance and complex dimensions, Phys. Rep. 297 (5) (1998) 239–270, http://dx.doi.org/10.1016/S0370-1573(97)00076-8.
[4] M.A. Montemurro, Beyond the Zipf–Mandelbrot law in quantitative linguistics, Phys. A 300 (3–4) (2001) 567–578, http://dx.doi.org/10.1016/s0378-4371(01)00355-7.
[5] A. Schenkel, J. Zhang, Yi-C. Zhang, Long range correlation in human writings, Fractals 1 (1) (1993) 47–57, http://dx.doi.org/10.1142/s0218348x93000083.
[6] V.V. Iudelevich, On the "tree" structure of natural numbers, Discrete Math. Appl. 32 (5) (2022) 325–340.
[7] R. Conti, P. Contucci, V. Iudelevich, Bounds on tree distribution in number theory, Ann. Univ. Ferrara (2024) http://dx.doi.org/10.1007/s11565-024-00535-3.
[8] R. Conti, P. Contucci, A ℕatural avenue, Exp. Math. (2025) 1–7, http://dx.doi.org/10.1080/10586458.2025.2471939.
[9] P. Mihăilescu, Primary cyclotomic units and a proof of Catalan's conjecture, J. Reine Angew. Math. 2004 (572) (2004) http://dx.doi.org/10.1515/crll.2004.048.
[10] H.S. Heaps, Information Retrieval: Computational and Theoretical Aspects, Academic Press, Inc., 1978.
[11] M. Gerlach, E.G. Altmann, Stochastic model for the vocabulary growth in natural languages, Phys. Rev. X 3 (2013) 021006, http://dx.doi.org/10.1103/PhysRevX.3.021006, URL: https://link.aps.org/doi/10.1103/PhysRevX.3.021006.
[12] P. Rosillo-Rodes, M.S. Miguel, D. Sánchez, Entropy and type-token ratio in gigaword corpora, Phys. Rev. Res. 7 (2025) 033054, http://dx.doi.org/10.1103/rxxz-lk3n, URL: https://link.aps.org/doi/10.1103/rxxz-lk3n.
[13] T. Schürmann, P. Grassberger, Entropy estimation of symbol sequences, Chaos 6 (3) (1996) 414–427.
[14] J. Ziv, A. Lempel, A universal algorithm for sequential data compression, IEEE Trans. Inform. Theory 23 (3) (1977) 337–343, http://dx.doi.org/10.1109/tit.1977.1055714.
[15] J. Laherrère, Parabolic fractal distributions in nature, C. R. Acad. Sci. - Ser. IIA - Earth Planet. Sci. 322 (7) (1996) 535–541.
[16] J. Laherrère, D. Sornette, Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales, Eur. Phys. J. B 2 (4) (1998) 525–539, http://dx.doi.org/10.1007/s100510050276.
[17] X. Michalet, Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium, Phys. Rev. E 82 (4) (2010) 041914, http://dx.doi.org/10.1103/physreve.82.041914.
[18] D.Y. Manin, On the nature of long-range letter correlations in texts, 2008, http://dx.doi.org/10.48550/ARXIV.0809.0103, URL: http://arxiv.org/abs/0809.0103.

[19] E.G. Altmann, G. Cristadoro, Mirko Degli Esposti, On the origin of long-range correlations in texts, Proc. Natl. Acad. Sci. 109 (29) (2012) 11582–11587, http://dx.doi.org/10.1073/pnas.1117723109, URL: https://www.pnas.org/doi/abs/10.1073/pnas.1117723109.
[20] M.K. Riahi, I.A. Qattan, J. Hassan, D. Homouz, Identifying short- and long-time modes of the mean-square displacement, AIP Adv. 9 (5) (2019) 055112, http://dx.doi.org/10.1063/1.5093628, URL: https://pubs.aip.org/aip/adv/article/9/5/055112/1070817/Identifying-short-and-long-time-modes-of-the-mean.
[21] E. Awad, R. Metzler, Crossover dynamics from superdiffusion to subdiffusion: Models and solutions, Fract. Calc. Appl. Anal. 23 (2020) 55–102, http://dx.doi.org/10.1515/fca-2020-0003, URL: https://link.springer.com/article/10.1515/fca-2020-0003.
[22] R. Metzler, J.-H. Jeon, A.G. Cherstvy, E. Barkai, Anomalous diffusion models and their properties: non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking, Phys. Chem. Chem. Phys. 16 (2014) 24128–24164, http://dx.doi.org/10.1039/C4CP03465A.
[23] G.D.J. Phillies, Quantitative interpretation of simulated polymer mean-square displacements, Polymers 17 (4) (2025) 516, http://dx.doi.org/10.3390/polym17040516, URL: https://www.mdpi.com/2073-4360/17/4/516.
[24] W.-J. Ma, S.-C. Wang, C.-N. Chen, C.-K. Hu, Crossover behavior of stock returns and mean square displacements of particles governed by the Langevin equation, EPL (Europhys. Lett.) 102 (2013) 66003, http://dx.doi.org/10.1209/0295-5075/102/66003, URL: https://ah.lib.nccu.edu.tw/bitstream/140.119/62211/1/66003.pdf.
[25] C.-L. Liu, S. Zhang, Y. Geng, H. l. Lai, H. Wang, Character distributions of classical Chinese literary texts: Zipf's law, genres, and epochs, 2017, arXiv preprint arXiv:1709.05587.
[26] W.B. Deng, A.E. Allahverdyan, B. Li, Q.A. Wang, Rank-frequency relation for Chinese characters, 2013, arXiv preprint arXiv:1309.1536.
[27] H. Liu, M. Nuo, J. Wu, Zipf's law and statistical data on modern tibetan, in: International Conference on Computational Linguistics, 2014, pp. 322–333, URL: https://aclanthology.org/C14-1032.pdf.
[28] L.Q. Ha, P. Hanna, J. Ming, F.J. Smith, Extending Zipf's law to n-grams for large corpora, Artif. Intell. Rev. 32 (1–4) (2009) 101–113, http://dx.doi.org/10.1007/s10462-009-9135-4.
[29] J. Reichert, L.F. Granz, T. Voigtmann, Transport coefficients in dense active Brownian particle systems: mode-coupling theory and simulation results, Eur. Phys. J. E 44 (27) (2021) http://dx.doi.org/10.1140/epje/s10189-021-00039-4.
[30] Y. Li, G. Farrher, R. Kimmich, Sub- and superdiffusive molecular displacement laws in disordered porous media probed by nuclear magnetic resonance, Phys. Rev. E 74 (2006) 066309, http://dx.doi.org/10.1103/PhysRevE.74.066309.
[31] L. Gmachowski, Fractal model of anomalous diffusion, Eur. Biophys. J. 44 (8) (2015) 613–621, http://dx.doi.org/10.1007/s00249-015-1054-5.
[32] V. Ilyin, I. Procaccia, A. Zagorodny, Stochastic processes crossing from ballistic to fractional diffusion with memory: exact results, Phys. Rev. E 81 (2010) 030105, http://dx.doi.org/10.1103/PhysRevE.81.030105.
[33] C. Singh, A. Chaudhuri, Anomalous dynamics of a passive droplet in active turbulence, Nat. Commun. 15 (2024) 3704.
[34] I.M. Sokolov, Models of anomalous diffusion in crowded environments, Soft Matter 8 (2012) 9043–9052, http://dx.doi.org/10.1039/C2SM25701G.
[35] A. Breccia, P. Contucci, F. Gerace, M. Lippi, G. Sicuro, Testing transformer learnability on the arithmetic sequence of rooted trees, 2025, arXiv preprint arXiv:2512.01870. URL: https://doi.org/10.48550/arXiv.2512.01870.