Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2019, 7, 228-249
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Origin of Dynamic Correlations of Words in Written Texts

Hiroshi Ogura, Hiromi Amano, Masato Kondo

Department of Information Science, Faculty of Arts and Sciences, Showa University, Fujiyoshida, Japan
Email: ogura@cas.showa-u.ac.jp, kayanm@cas.showa-u.ac.jp, mkondo@nr.showa-u.ac.jp

How to cite this paper: Ogura, H., Amano, H. and Kondo, M. (2019) Origin of Dynamic Correlations of Words in Written Texts. *Journal of Data Analysis and Information Processing*, **7**, 228-249.
https://doi.org/10.4236/jdaip.2019.74014

Received: August 26, 2019
Accepted: October 12, 2019
Published: October 15, 2019

Abstract

In a previous study, we introduced dynamical aspects of written texts by regarding serial sentence number from the first to last sentence of a given text as discretized time. Using this definition of a textual timeline, we defined an autocorrelation function (ACF) for word occurrences and demonstrated its utility both for representing dynamic word correlations and for measuring word importance within the text. In this study, we seek a stochastic process governing occurrences of a given word having strong dynamic correlations. This is valuable because words exhibiting strong dynamic correlations play a central role in developing or organizing textual contexts. While seeking this stochastic process, we find that additive binary Markov chain theory is useful for describing strong dynamic word correlations, in the sense that it can reproduce characteristics of autocovariance functions (an unnormalized version of ACFs) observed in actual written texts. Using this theory, we propose a model for time-varying probability that describes the probability of word occurrence in each sentence in a text. The proposed model considers hierarchical document structures such as chapters, sections, subsections, paragraphs, and sentences. Because such a hierarchical structure is common to most documents, our model for occurrence probability of words has a wide range of universality for interpreting dynamic word correlations in actual written texts. The main contributions of this study are, therefore, finding usability of the additive binary Markov chain theory to analyze dynamic correlations in written texts and offering a new model of word occurrence probability in which common hierarchical structure of documents is taken into account.

Keywords

Autocorrelation Function, Autocovariance Function, Word Occurrence, Stochastic Process, Additive Binary Markov Chain

1. Introduction

Introducing the notion of time to written texts reveals dynamical aspects of word occurrences, allowing us to apply standard dynamical analyses developed and used in the fields of signal processing and time series analysis. In a previous study [1], we used a set of serial sentence numbers assigned from the first to last sentence in a given text as a discretized time. Using this time unit, we successfully defined an autocorrelation function (ACF) appropriate to words in written texts, then calculated ACFs according to this definition for words frequently appearing in twelve famous books. We found that the resulting ACFs could be classified into two groups: words showing dynamic correlations and those with no correlation type. Words showing dynamic correlations are called Type-I words, and their ACFs are well-described by a modified Kohlrausch-Williams-Watts (KWW) function. Words showing no correlation are called Type-II words, and their ACFs are modeled as a simple stepdown function. We showed that this stepdown function can be theoretically derived from the assumption that the stochastic process governing occurrences of Type-II words is a homogeneous Poisson point process.

Type-I words are known to occur multiple times in a text in a bursty and context-specific manner, and such occurrences ensure that the word has a dynamic correlation. Put another way, Type-I words are important for describing an idea or topic, and are therefore expected to be highly correlated with a duration, typically several tens of sentences in which the idea or topic is described. In contrast, Type-II words are not context-specific and their appearance is governed by chance. Type-I words are therefore more important than Type-II words, in the sense that they play a central role in explaining the author's ideas or thoughts. The author's insights and thought process should thus be discernible through modeling of the stochastic process that generates Type-I words. However, despite the importance of Type-I words, the stochastic process yielding them could not be clarified in the previous study.

The purpose of the present study was to find such a stochastic process for Type-I words. We found that additive binary Markov chain theory is suited to this purpose because this theory can capture characteristic behaviors for dynamic correlations of Type-I words in actual written texts. To our knowledge, this is the first application of the theory of additive binary Markov chain to analyze written texts, although the theory has been utilized to model natural phenomena such as wind generations [2]. Using this theory, we further calculated a time-varying probability that describes the occurrence probability of a given word as a function of time (*i.e.*, sentence number). The evaluated time-varying probability has two distinctive features: probability values in the text are discretized into several values, and sentence numbers at which the occurrence probability takes the same value among several discretized values seem to aggregate along a time (sentence number) axis. The ultimate goal of this study was construction of a recursive model for probability distributions, which is easily converted to a time-varying probability having the two features described above. We constructed this recursive model by incorporating common hierarchical document structures (chapters, sections, subsections, paragraphs, and sentences), so it can be applied to a broad range of applications for analyzing written texts. Although long-range correlations in written texts have been modeled from various point of views, the methodologies used are completely different from ours [3] [4].

The remainder of this paper is organized as follows: In Section 2, we define the autocovariance function (ACVF), an unnormalized ACF used throughout this study. In Section 3, we describe the additive binary Markov chain theory, which allows mutual conversion between memory functions and ACVFs. We also present the relation between the time-varying probability of word occurrence and the memory function, which allows us to estimate the time-varying probability of a given word. In Section 4, we present typical examples of time-varying probability for Type-I words and their two distinctive features. In Section 5, we describe how to establish a recursive model for a probability distribution that successfully reproduces the two features of the time-varying probability. Finally, in Section 6, we present our conclusions and suggest directions for future research.

2. Autocorrelation and Autocovariance Functions

2.1. Definitions of Autocorrelation and Autocovariance Functions

The autocovariance function gives the covariance of a given process with itself at pairs of time points. A standard definition of the ACVF for a weak stationary process $\{X_t\}$ is

$$K(\tau) = E\left[(X_t - \mu)(X_{t+\tau} - \mu)\right], \tag{1}$$

where $E[\,]$ is an expectation operator and $\mu = E[X_t]$ is the mean of $\{X_t\}$ [5] [6]. The ACF, which measures the similarity between signals X_t as a function of the time lag τ between them, is a normalized autocovariance defined as [5] [6]

$$\rho(\tau) = \frac{K(\tau)}{K(0)}. \tag{2}$$

As Equation (1) and Equation (2) show, these definitions for ACVF and ACF use the deviation of X_t from its mean instead of X_t itself. In our previous study [1], we used X_t itself to define the ACF so that the limit of the ACF at infinitely large τ gives important information regarding Type-II words, that is, the limit of ACF using X_t itself approaches a constant rate λ of a corresponding homogeneous Poisson point process for a given Type-II word when $\tau \to \infty$. However, because our main interest in this study is in Type-I words, there is no need to extract information from the ACFs of Type-II words. We therefore use the standard definition of ACVF, Equation (1), throughout this study.

Another difference between the previous and present studies is that we extensively use ACVFs instead of ACFs because ACVFs more directly link to additive binary Markov chain theory, as will be shown later.

2.2. Examples of ACVFs for Type-I and Type-II Words

Because we use the set of serial sentence numbers assigned from the first to the last sentence in a considered text as discretized time, and because we intend to analyze word occurrence characteristics in terms of ACVFs, we define the signal X_t representing word occurrence or non-occurrence as

$$X_t = \begin{cases} 1 & \left(\text{when a given word occurs in the } t\text{th sentence}\right) \\ 0 & \left(\text{when a given word does not occur in the } t\text{th sentence}\right) \end{cases} \tag{3}$$

Figure 1 and Figure 2 show examples of X_t and ACVFs calculated from X_t for typical Type-I and Type-II words, respectively, extracted from a set of frequent words in Charles Darwin's most famous work, *On the Origin of Species*. These words were chosen as typical Type-I and Type-II words in the previous study, and are also used here for comparison. Here, a "frequent" word is one appearing in at least 50 sentences in the text. Text preprocessing procedures performed before calculating ACVFs are the same as in the previous study.

As Figure 1 shows, ACVFs—namely $K(\tau)$ as defined in Equation (1)—for Type-I words are monotonically decreasing, indicating that dynamic correlations decrease as lag τ increases. They also show an apparent persistence of dynamic correlations with durations of several tens of sentences. In contrast, Figure 2 clearly shows that ACVFs for Type-II words show no dynamic correlations, suggesting that Type-II words are generated from a memoryless stochastic process. Figure 1 and Figure 2 show that the functional forms and characteristic behaviors of ACVFs are the same as those for the corresponding ACFs reported

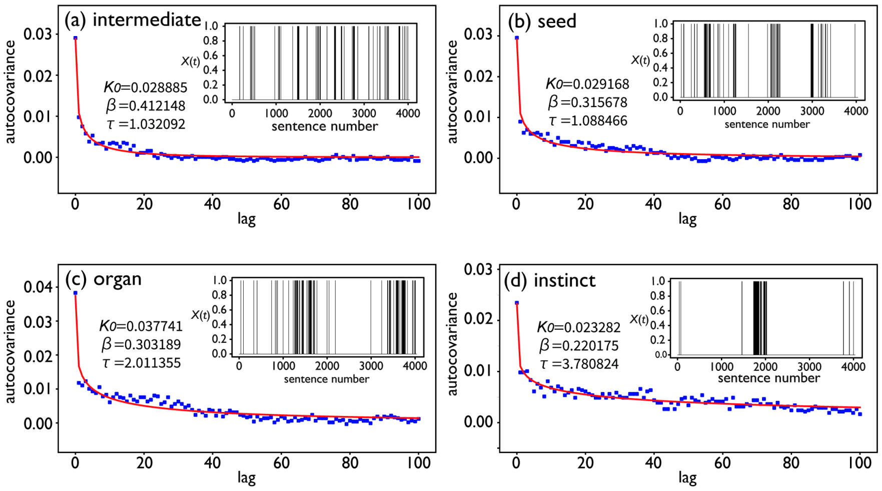

Figure 1. Examples of X_t (insets) and ACVFs (blue plots) for typical Type-I words. Red curves represent the best-fitted KWW functions (see Section 4) of which optimized parameters are shown in the plot areas. X_t and ACVFs for (a) "intermediate", (b) "seed", (c) "organ" and (d) "instinct", from a set of frequent words in Darwin text.

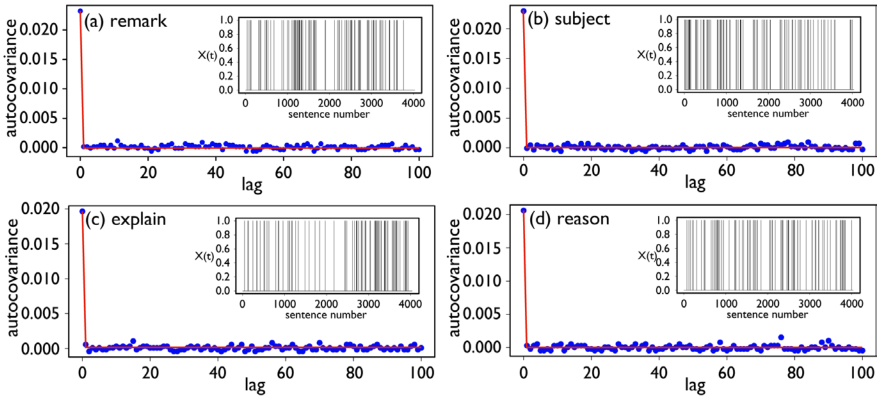

Figure 2. Examples of X_t (insets) and ACVFs (blue plots) for typical Type-II words. Red lines represent the stepdown functions. X_t and ACVFs for (a) "remark", (b) "subject", (c) "explanation", and (d) "reason", from a set of frequently used words in the Darwin text.

in the previous study. Therefore, the model functions used in the previous study to describe the ACFs of Type-I and Type-II words are still appropriate for describing ACVFs for each word type. Specifically, the Kohlrausch-Williams-Watts (KWW) function used to model the ACFs of Type-I words and the stepdown function used to describe the ACFs of Type-II words still provide full descriptive power when applied to modeling ACVFs. Figure 1 and Figure 2 show fitting results from these two model functions as red lines, indicating the validity of using KWW and the stepdown functions to model the two ACVF types. Details of the fitting results using the KWW function for the ACVFs of Type-I words will be described in Section 4.

Our previous study found that, without exception, all frequent words appearing in the twelve famous books are well classified into Type-I or Type-II words. This study also showed that the stochastic process governing occurrences of Type-II words is a homogeneous Poisson point process, which is completely memoryless. In the following section, we describe our attempts to determine a stochastic process for generating Type-I words in written text. We also investigate a mechanism for providing dynamic correlations to Type-I words.

3. Additive Binary Markov Chain

3.1. Necessity of an Additive Markov Chain

One standard approach to analyzing time series with dynamic correlations is to use a Markov chain model [7] [8] [9]. Because a first-order Markov chain is simplest, we first apply one to model Type-I word occurrences and check whether or not the model can reproduce actual ACVFs exhibiting dynamic correlations. Consider a random variable X_t where t denotes a discretized time. If the probability of moving to the next state at the next time depends only on the present state, that is, if

$$Pr\left(X_{n+1}=x \mid X_1=x_1, X_2=x_2, \cdots, X_n=x_n\right)=Pr\left(X_{n+1}=x \mid X_n=x_n\right) \quad (4)$$

holds, then $\{X_t\}$ is a first-order Markov chain. In the case of a binary Markov chain in which signal X_t takes only values 0 or 1 as in Equation (3), the stochastic properties of the first-order Markov chain can be completely determined by defining a transition matrix

$$P = \begin{pmatrix} P_{00} & P_{10} \\ P_{01} & P_{11} \end{pmatrix}, \tag{5}$$

where P_{ij} denotes a transition probability from state *i* to state *j*. To determine all values of P_{ij} in the transition matrix from signals $\{X_t\}$ observed in actual written texts, we simply used maximum likelihood estimators

$$P_{ij} = \frac{n_{ij}}{n_{i0} + n_{i1}}, \tag{6}$$

where n_{ij} is the number of transitions from state *i* to state *j* observed in signals $\{X_t\}$. **Table 1** shows transition probabilities thus obtained for the typical Type-I words shown in **Figure 1**.

After obtaining all values of P_{ij} listed in **Table 1**, we can simulate occurrences of a given word by using a simple Monte Carlo procedure as follows:

1) Arbitrarily set an initial state X_0 as 0 or 1.

2) Determine the next state X_1, by comparing P_{00} or P_{10} with a generated random number $p \in [0,1]$ following the standard uniform distribution $U(0,1)$. If the initial state was $X_0 = 0$, compare random number *p* with P_{00}. Otherwise, if $X_0 = 1$, compare *p* with P_{10}. For example, consider the case where $X_1 = 0$. Then if $p < P_{00}$ we set the next state as $X_1 = 0$, while if $p \ge P_{00}$ we set $X_1 = 1$. This procedure ensures that $Pr(X_1 = 0 \mid X_0 = 0) = P_{00}$ and $Pr(X_1 = 1 \mid X_0 = 0) = P_{01}$ because from Equation (6), $P_{00} + P_{01} = 1$ always holds. Similarly, given $X_0 = 1$, if $p < P_{10}$ we set $X_1 = 0$, otherwise we set $X_1 = 1$ so that $Pr(X_1 = 0 \mid X_0 = 1) = P_{10}$ and $Pr(X_1 = 1 \mid X_0 = 1) = P_{11}$.

3) Repeat Step 2 with replacements $X_1 \to X_i$ and $X_0 \to X_{i-1}$ $(i = 2,3,\cdots,n)$. By repeating this procedure *n* times, we can obtain simulated signals $X_1, X_2, \cdots, X_n$, the set of which is a first-order Markov chain having the transition matrix *P* given by Equation (5).

Insets in **Figure 3** show examples of simulated signals X_t for typical Type-I words with transition matrices *P*, elements of which are listed in **Table 1**. That figure also shows ACVFs calculated from those simulated signals. Obviously, these ACVFs are completely different from the actual ACVFs shown in **Figure 1**, in that they do not have long durations over several tens of sentences, although they exhibit dynamic correlations over much shorter durations. This indicates that the first-order Markov chain cannot reproduce actual dynamic correlations of Type-I words.

A direct and intuitive way to discover dynamic correlations with long durations in simulated signals X_t is to consider higher-order Markov chains [7] [8] [9]. In an *m*th order Markov chain, the probability of moving to the next state is

$$\begin{aligned} &Pr(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ &= Pr(X_{n+1} = x \mid X_{n-m+1} = x_{n-m+1}, X_{n-m+2} = x_{n-m+2}, \cdots, X_n = x_n). \end{aligned} \tag{7}$$

Table 1. Transition probabilities for typical Type-I words estimated by Equation (6).

	Intermediate	Seed	Organ	Instinct
P_{00}	0.9800766	0.9787887	0.9723871	0.9860371
P_{01}	0.0199234	0.0212114	0.0276129	0.0139629
P_{10}	0.6446281	0.6747967	0.6645963	0.5670103
P_{11}	0.3553719	0.3252033	0.3354037	0.4329897

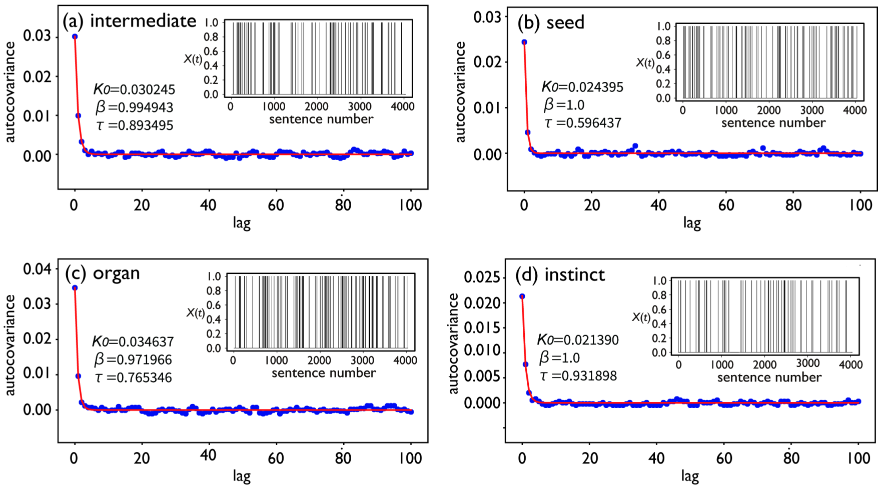

Figure 3. Simulated signals X_t (insets) for the typical Type-I words obtained as a realization of first-order Markov chains and ACVFs (blue plots) calculated from those X_t. Red curves represent best-fitted KWW functions, optimized parameters for which are shown in the plot areas. Results are for the words (a) "intermediate", (b) "seed", (c) "organ", and (d) "instinct".

However, one difficulty is that the number of transition probabilities, each of which is an element of the transition matrix, grows exponentially with the order. In the case of binary Markov chains, we must evaluate 2^{m+1} transition probabilities to model the *m*th order Markov chain because we must consider all possible transition patterns in the last m signals. If $m = 10$, we must determine 2048 transition probabilities. This is impossible because we cannot obtain sufficient samples to evaluate these probabilities when the number of signals X_t, which is equal to the number of sentences in the considered text, is the same order as the number of transition probabilities to be evaluated. For example, *The Origin of Species* has 3991 sentences, so it is impossible to determine 2048 transition probabilities with sufficient statistical reliability from 3991 signals. We must therefore consider another model that is tractable and can generate dynamic correlations with long durations. In the following subsection, we introduce additive binary Markov chains for this purpose.

3.2. Framework of Additive Binary Markov Chain Theory

Melnyk *et al.* [10] [11] proposed a theory of additive binary Markov chains, the framework of which is described below. In the following subsection, we will show that the theory is successfully applied to model occurrences of Type-I words.

An additive Markov chain of order m has the property

$$Pr\left(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \cdots, X_{n-m} = x_{n-m}\right) = \sum\nolimits_{r=1}^{m} f\left(x_n, x_{n-r}, r\right). \quad (8)$$

This means influences of previous states at different times are mutually independent on next states and thus can be expressed in additive form. In the binary case, where signal X_t is restricted to a value of 0 or 1, the theory tells us that the conditional probability of Equation (8) can be modified as

$$\begin{aligned} &Pr\left(X_n = 1 \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \cdots, X_{n-m} = x_{n-m}\right) \\ &= \bar{X} + \sum\nolimits_{r=1}^{m} F\left(r\right)\left(x_{n-r} - \bar{X}\right), \end{aligned} \quad (9)$$

where $\bar{X}$ is the mean of signals $\{X_t\}$, and $F\left(r\right)$ is a memory function representing the degree of influence of a previous signal occurring r time steps before. Thus, if $F\left(r\right)$ takes a large value, then the occurrence of a given word at $t = n - r$ positively affects the word occurrence at $t = n$. Another implication of Equation (9) is that by obtaining $F\left(r\right)$ for a given word, we can calculate the probability of signal X_t being 1 by this equation, and we can therefore generate signals X_t by use of a simple Monte Carlo procedure with that probability. Because the parameters needed to simulate generation of X_t are $F\left(r\right)$, the number of parameters to be evaluated is equal to the order m of the considered additive Markov chain. Tractability is therefore greatly improved because the number of parameters is only linearly dependent on the order m of a considered Markov chain.

Furthermore, the theory of additive binary Markov chains [10] [11] offers simple simultaneous equations that directly relate ACVFs—$K\left(r\right)$ given by Equation (1)—to the memory functions $F\left(r\right)$ appearing in Equation (9) as

$$K\left(r\right) = \sum_{s=1}^{m} K\left(r - s\right) F\left(s\right) \quad \left(r = 1, 2, \cdots, m\right) \quad (10)$$

Note that the relations

$$K\left(r\right) = K\left(-r\right), \quad (11)$$

$$K\left(0\right) = \bar{X}\left(1 - \bar{X}\right), \quad (12)$$

always hold by the definition of $K\left(r\right)$. By using Equation (11) and Equation (12), we can regard Equation (10) as m simultaneous equations relating ACVFs $K\left(1\right), K\left(2\right), \cdots, K\left(m\right)$ to memory functions $F\left(1\right), F\left(2\right), \cdots, F\left(m\right)$. We can thus calculate ACVFs, $K\left(r\right)$, from a theoretically assumed memory function $F\left(r\right)$, or we can conversely calculate memory functions $F\left(r\right)$ for a given word at $r = 1, 2, \cdots, m$ from its actual ACVFs.

Before applying additive binary Markov chain theory to analyze dynamic correlations for Type-I words, we confirm the validity of the theory as follows. First, we assume the tentative memory function

$$F\left(r\right) = \begin{cases} 0.1 - 0.05r & \left(1 \le r < 20\right) \\ 0 & \left(20 \le r\right) \end{cases} \quad (13)$$

which is shown as the solid line in Figure 4(a). This function is plausible in the sense that memory decreases as lag r increases. We set other conditions as $\bar{X} = 0.05$ and $X_0 = 0$, tentatively determined to fit actual word occurrences. We then generate signals X_t according to the conditional probability given by Equation (9) using a simple Monte Carlo procedure, the algorithm for which is basically the same as that for generating $\{X_t\}$ using the first-order Markov chain. The Monte Carlo simulation to generate X_t applies two conditions (C1) and (C2), as follows:

(C1) Signal $X_n = 1$ if a generated random number $p \in [0,1]$ following standard uniform distribution $U(0,1)$ is less than the conditional probability given by Equation (9). Otherwise, $X_n = 0$. To calculate the conditional probability, we substitute $F(r)$ as calculated from Equation (13) and the past m signal values into Equation (9). This procedure is repeated until we obtain the desired length of signals $\{X_t\}$.

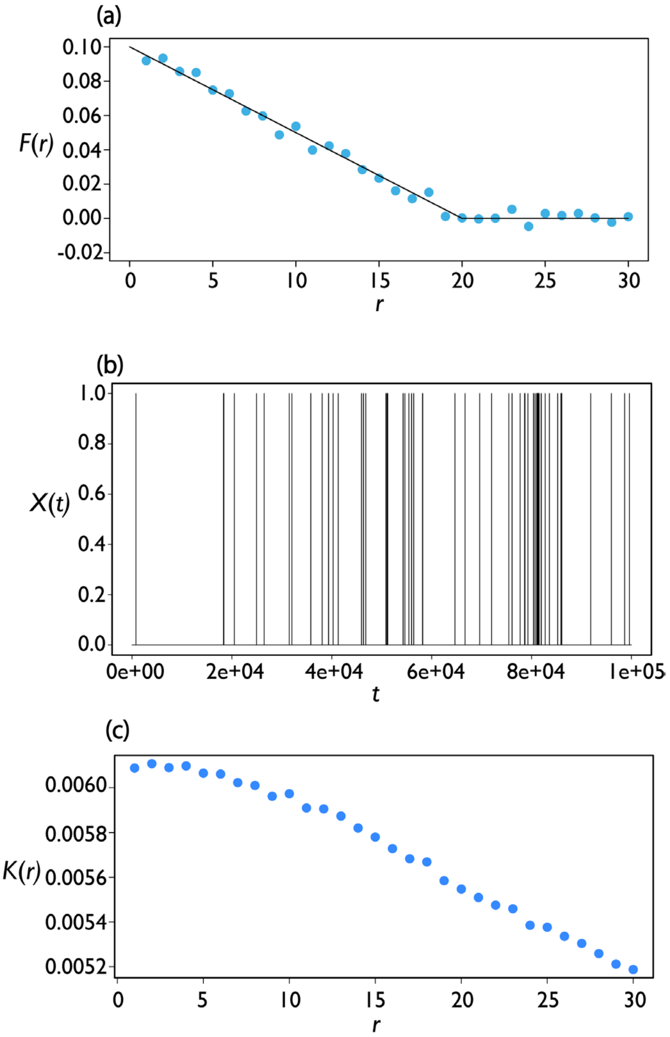

Figure 4. (a) Memory function $F(r)$ given by Equation (13) (solid line) and reproduced $F(r)$ from ACVFs with Equation (10) (blue plots); (b) Simulated signals X_t generated by a simple Monte Carlo procedure; (c) ACVFs calculated from simulated signals X_t.

(C2) The number of past signals is insufficient for generating the first $m-2$ signals $X_1, X_2, \cdots, X_{m-2}$ because Equation (9) requires past m signals $X_{n-1}, X_{n-2}, \cdots, X_{n-m}$ to calculate the conditional probability of X_n being 1. In these cases, we use all available past n signals from X_0 to X_{n-1} to calculate Equation (9) and ignore other terms that require $X_{-1}, X_{-2}, \cdots$.

Obtained X_t and estimated ACVFs from X_t with a condition $m=30$ are shown in Figure 4(b) and Figure 4(c), respectively. The functional form of ACVFs (Figure 4(c)) is very similar to the used memory function (the solid line in Figure 4(a)), which is consistent with previously reported results [10] [11]. To confirm the validity of the additive binary Markov chain theory, we calculate values for the memory function $F(r)$ from ACVFs in Figure 4(c) by inversely using the simultaneous equations, Equation (10). Figure 4(a) compares the memory function calculated from ACVFs (blue plots) and that of the original Equation (13) (solid line). As that figure shows, the original and calculated $F(r)$ from ACVFs satisfactorily agree, ensuring validity of the theory. The following section applies additive binary Markov chain theory as a basic framework for investigating characteristics of the stochastic process that generates Type-I words.

4. Memory Functions and Occurrence Probabilities of Type-I Words

In this section, we apply the theory of additive binary Markov chain to Type-I word occurrences to clarify characteristics of the stochastic process that generates dynamic correlations of Type-I words. Since ACVFs for Type-I words can be calculated from actual signals X_t by use of Equation (1) and these ACVFs can be used as observed $K(r)$ values in Equation (10), it is easy to determine $F(r)$ values at each lag r from Equation (10). Figure 5 shows examples of thus

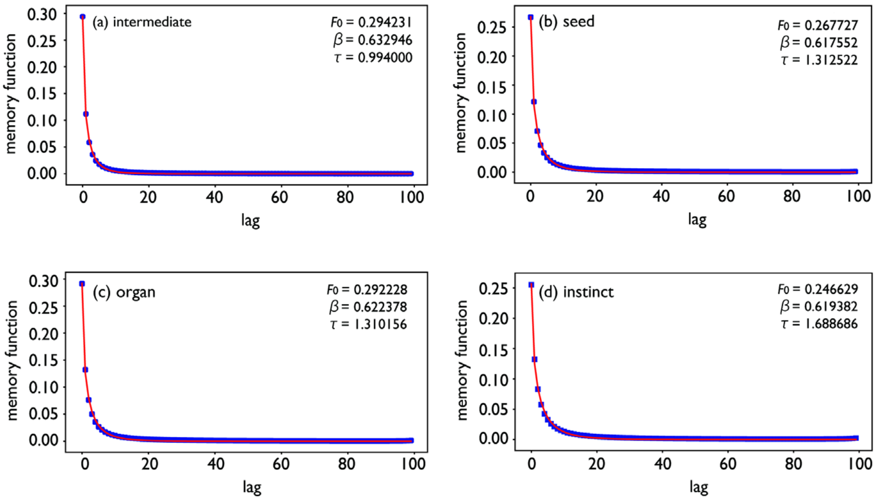

Figure 5. Examples of $F(r)$ (blue plots) for typical Type-I words calculated from ACVFs and fitted lines with the KWW function, for which optimized parameters are shown in the plot areas (red curves).

obtained $F(r)$ for typical Type-I words. When calculating $F(r)$, we have used $K(r)$ represented by best-fitted KWW functions at each lag step instead of the original ACVFs to reduce noise effects. Red lines in that figure represent results fitted to $F(r)$ by use of KWW functions, indicating that memory functions $F(r)$ for Type-I words can also be well described by KWW functions as in the case of fitting ACVFs.

Strictly speaking, because the memory function $F(r)$ is defined at $r \geq 1$, as seen in Equation (9) and Equation (10), we use a modified form of the KWW function for $F(r)$:

$$F(r) = F_0 \exp\left\{-\left(\frac{r-1}{\tau}\right)^{\beta}\right\}, \tag{14}$$

while when fitting ACVFs we use the standard form of the KWW function, namely,

$$K(r) = K_0 \exp\left\{-\left(\frac{r}{\tau}\right)^{\beta}\right\}. \tag{15}$$

Table 2 summarizes the evaluated fitting parameters of the KWW functions, Equation (14) and Equation (15), obtained through fittings for the typical Type-I words. These optimized parameters are obtained by nonlinear least-squares fitting. Fitting parameters for $K(r)$ are more scattered than those for $F(r)$, particularly in values of β and τ. This means that slight differences in functional form for $F(r)$ are amplified and become distinct in $K(r)$.

To confirm this observation, we conducted nonlinear least-squares fittings for all Type-I words from five well-known academic books, described in detail in the **Appendix**. Table 3 summarizes means and standard deviations of the fitting parameters for the Type-I words, and Table 4 lists coefficients of variation (CV), defined as σ/μ and calculated from the data in Table 3. Table 4 reconfirms that fitting parameter values are more scattered for $K(r)$ than for $F(r)$ because CVs of β and τ for $K(r)$ are much larger than those for $F(r)$.

Consistency of the additive binary Markov chain theory for modeling occurrences of Type-I words can be used to confirm whether the theory can reproduce observed ACVFs. The procedure for confirming consistency is as follows. Note that we use the symbol X_t for actual signals of word occurrence or nonoccurrence as observed in actual written text, and X_n designates simulated signals obtained through simple Monte Carlo procedures.

Table 2. Optimized fitting parameters of Equations (14) ($F(r)$) and (15) ($K(r)$) obtained by nonlinear least-squares fitting.

	$K(r)$			$F(r)$		
Word	K_0	β	τ	F_0	β	τ
Instinct	0.0233	0.220	3.78	0.257	0.619	1.69
Intermediate	0.0289	0.412	1.03	0.294	0.633	0.994
Organ	0.0374	0.303	2.01	0.292	0.622	1.31
Seed	0.0292	0.316	1.09	0.268	0.618	1.31

Table 3. Mean μ and standard deviation σ for fitting parameters in $K(r)$ (Equation (15)) and $F(r)$ (Equation (14)) calculated for all Type-I words appearing in the corresponding book and described in the form $\mu \pm \sigma$.

Book	Number of Type-I words	$K(r)$			$F(r)$		
		K_0	β	τ	F_0	β	τ
Darwin	109	0.0278 ± 0.0187	0.262 ± 0.115	0.485 ± 1.14	0.142 ± 0.0769	0.584 ± 0.0574	1.88 ± 0.545
Einstein	17	0.0732 ± 0.0199	0.261 ± 0.0420	0.224 ± 0.169	0.161 ± 0.0353	0.589 ± 0.0195	1.77 ± 0.281
Freud	14	0.0465 ± 0.0176	0.276 ± 0.144	0.658 ± 1.51	0.160 ± 0.0760	0.596 ± 0.0683	1.81 ± 0.564
Kant	142	0.0283 ± 0.0250	0.245 ± 0.106	0.237 ± 0.391	0.140 ± 0.0534	0.580 ± 0.0506	1.97 ± 0.442
Smith	382	0.0141 ± 0.0143	0.256 ± 0.106	0.305 ± 1.01	0.131 ± 0.0570	0.583 ± 0.0561	1.91 ± 0.491

Table 4. CV for the fitting parameters in $K(r)$ (Equation (15)) and $F(r)$ (Equation (14)) calculated from σ/μ values in Table 3.

Book	Number of Type-I words	$K(r)$			$F(r)$		
		K_0	β	τ	F_0	β	τ
Darwin	109	0.673	0.439	2.351	0.542	0.098	0.290
Einstein	17	0.272	0.161	0.754	0.219	0.033	0.159
Freud	14	0.378	0.522	2.295	0.475	0.115	0.312
Kant	142	0.883	0.433	1.650	0.381	0.087	0.224
Smith	382	1.014	0.414	3.311	0.435	0.096	0.257

1) ACVFs for a Type-I word are calculated using Equation (1) from actual signals X_t observed in the text.

2) Curve fitting to fit Equation (15) to ACVFs obtained in the previous step is performed to obtain optimized values for fitting parameters.

3) $K(r)$ is calculated at each lag step r using Equation (15) with the optimized fitting parameters. We set the maximum lag step as $r_{\max} = 100$, which, as will be recognized in the following steps, is equivalent to setting the order of the additive binary Markov chain to $m = 100$. This setting is sufficient to cover the longest durations of dynamic correlations for Type-I words [1].

4) $K(r)$ obtained in the previous step are substituted into the simultaneous equations, Equation (10), to obtain $F(r)$.

5) $F(r)$ is used to calculate the conditional probability, Equation (9), and the calculated conditional probability of X_n being 1 is used to generate simulated signals X_n through simple Monte Carlo procedures. Conditions (C1) and (C2) described in Subsection 3.2 are still applied in this step. Before starting the Monte Carlo procedures, we set $\bar{X}$ to the averaged value of actual signals for a given word.

6) From the simulated signals X_n, ACVFs are calculated using Equation (1) and compared with those obtained in step 1. If the ACVFs calculated from simulated signals agree well with the actual ACVFs, we can consider the additive binary Markov chain theory as consistent.

Figure 6(a) and Figure 6(d) show comparisons between ACVFs calculated from actual signals (blue plots) and those reproduced from simulated signals (red plots) for two typical Type-I words, "intermediate" and "seed". Both are in good agreement, indicating that dynamic correlations of Type-I words are well modeled by the theory of additive binary Markov chains. Figure 6 also shows simulated signals X_n obtained through the simple Monte Carlo procedures (Figure 6(b) and Figure 6(e)) and the conditional probabilities of X_n being 1 (Figure 6(c) and Figure 6(f)).

As Equation (9) shows, signals X_t are considered to be consequences of the time-varying conditional probabilities for word occurrence given by Equation (9) in the framework of additive binary Markov chain theory. In this sense, the unobserved time-varying probabilities given by Equation (9) are more essential than the observed signals X_t. The time-varying probabilities shown in Figure 6(c) and Figure 6(f) seem plausible in the sense that the probabilities take higher values in time intervals with higher word occurrence rates.

A closer look at Figure 6(c) and Figure 6(f) reveals two further characteristics of the time-varying probability: that the probability value does not continuously change, but instead seems to be discretized to form several levels in an approximate sense, and that same-level probabilities seem to aggregate within a certain period of time. That is, the sentence numbers at which the occurrence probability takes approximately the same values among several discretized values seem to be aggregated along the time (sentence number) axis.

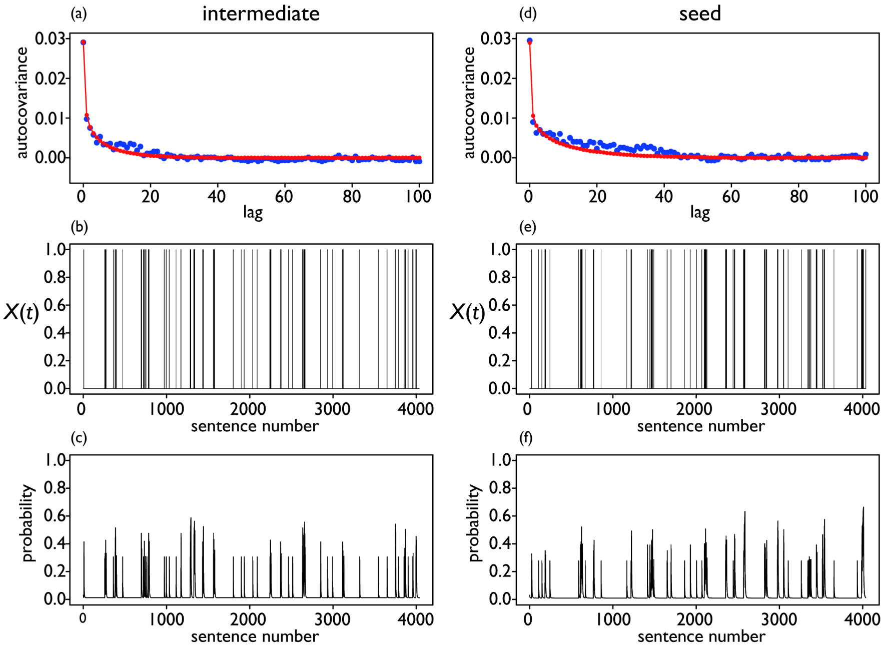

Figure 6. (a) and (d) ACVFs calculated from actual signals X_t observed in the text (blue plots) and those reproduced from simulated signals X_n (red plots); (b) and (e) Simulated signals X_n generated using the simple Monte Carlo procedures; (c) and (f) Conditional probabilities of X_n being 1 calculated using Equation (9). The left and right columns show results for the words "intermediate" and "seed" respectively.

We further calculated the time-varying probabilities of X_n being 1 for typical Type-I words selected from five academic books. Specifically, we chose the two words having the largest and second largest ΔBIC in each book because ΔBIC is a measure of dynamic correlation and thus indicates the importance of a given word [1]. Figure 7 shows the results for the words selected in this way. As that figure shows, the two features described above are common among all cases, indicating that the two features are substantial for all Type-I words having

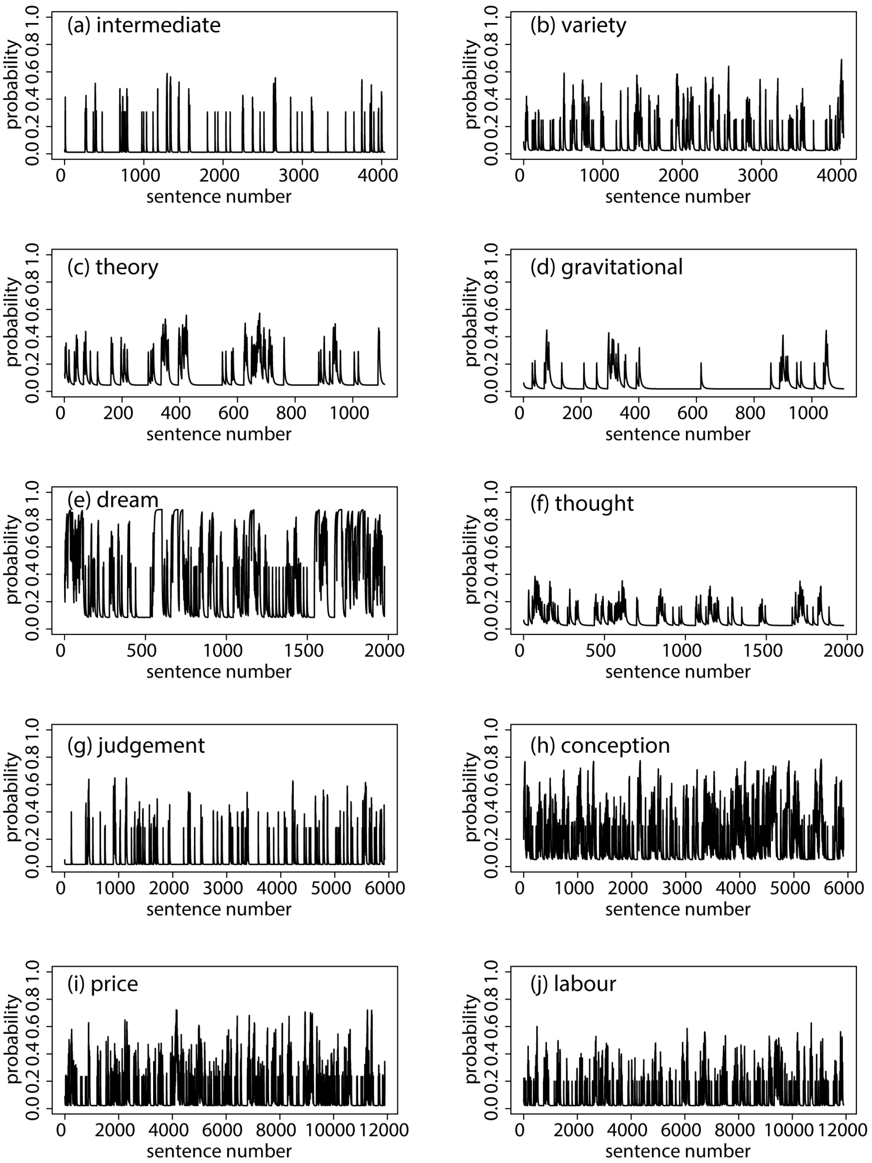

Figure 7. Conditional probabilities of X_n being 1 calculated using Equation (9). The procedures for obtaining these conditional probabilities are the same as those for Figure 6. Plots (a) and (b) show results for the words "intermediate" and "variety", respectively, in the Darwin text. Plots (c) and (d) show results for the words "theory" and "gravitational", respectively, in the Einstein text. Plots (e) and (f) show results for the words "dream" and "thought", respectively, in the Freud text. Plots (g) and (h) show results for the words "judgement" and "conception", respectively, in the Kant text. Plots (i) and (j) show results for the words "price" and "labour", respectively, in the Smith text.

strong dynamic correlations, although the second feature (aggregation along the horizontal axis) is not easy to see in **Figures 7(g)-(j)** due to compression of the horizontal axes.

The two features described above cannot be directly explained from Equation (9), so another viewpoint beyond the scope of additive binary Markov chain theory is needed to explain them. In the following section, we propose a recursive probability distribution model in which hierarchical structures of documents are considered to explain these features.

5. Hierarchical Model of Probability Distribution for Word Occurrences

Almost all documents have a hierarchical structure consisting of chapters, sections, subsections, paragraphs, and sentences. This section describes the construction of a probability distribution model that reflects such hierarchical structures. The constructed model is expected to reproduce the two features of the time-varying probability of word occurrence described in the previous section. However, because our aim is to capture dynamic correlations of Type-I words with a simple model and we do not intend to build a complex or sophisticated model, some details of the proposed model will be tentatively determined. We believe, however, that the hierarchical structures of documents induce dynamic correlations of Type-I words, and that our model essentially captures the origin of these correlations.

Our model is based on a recursive probability redistribution that constructs a hierarchical probability distribution. After obtaining the hierarchical probability distribution, we convert it to the time-varying probability of word occurrence. The probability redistribution and conversion are performed in the following manner.

1) As a starting point, we consider the standard uniform distribution $U(0,1)$ illustrated in **Figure 8(a)**, in which the density function is fixed to 1 in the unit interval [0, 1] and 0 otherwise. 0 and 1 on the horizontal axis correspond to the positions of the first and the last sentences in the text, so the unit interval [0, 1] corresponds to the entire text. No structures are introduced in the distribution at this point, so the occurrence probability of a given word is exactly the same for all sentences in the text.

2) The unit interval [0, 1] is divided into 5 subintervals of the same length, indexed as 1, 2, 3, 4, 5 (**Figure 8(a)**). These subintervals correspond to five chapters in the text. We then select two indices, a_1 and a_2, taken from a discrete uniform distribution with possible values $\{1,2,3,4,5\}$. Here, we select two other indices, b_1 and b_2, differing from a_1 and a_2 by sampling without replacement.

3) The rectangles representing probabilities at subintervals with indices a_1 and a_2 are removed and stacked on rectangles having indices b_1 and b_2. In **Figure 8(b)**, $a_1 = 1$, $a_2 = 5$, $b_1 = 2$, and $b_2 = 4$. Consequently, the occurrence probabilities for a given word within subintervals having indices 1 and 5 become

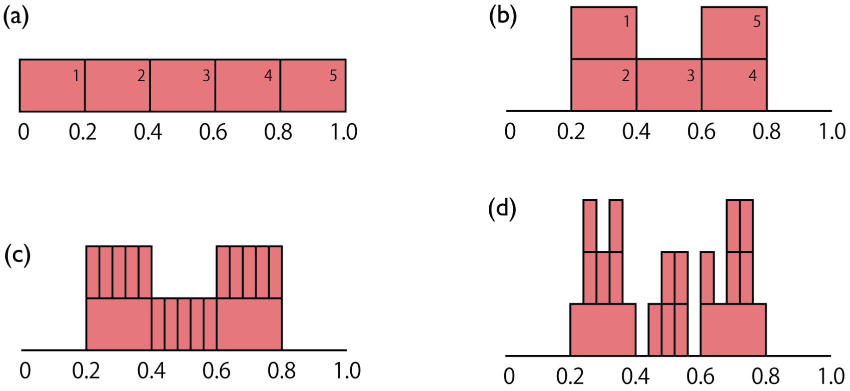

Figure 8. The procedure for probability redistribution considering hierarchical structures in the actual written text.

zero, indicating that chapters 1 and 5 become irrelevant to descriptions using the considered word, while chapters 2, 3, and 4 are considered relevant.

4) Division into five subintervals, choosing indices a_1, a_2, b_1, and b_2 and the stacking of probabilities are repeated for each portion of the top layer in the probability distribution as in Figure 8(c) and Figure 8(d). This procedure is recursively repeated to form a desired hierarchical structure for the probability distribution. For example, if we repeat this procedure twice, we obtain the probability distribution in Figure 8(d) with four different density function values (zero and three positive values 1, 2 and 3). Similarly, by repeating this three times, we obtain a probability distribution with five function values.

5) After obtaining a desired hierarchical probability distribution, that is, after repeating the probability redistribution a predefined number of times, we discretize the horizontal axis and convert it into sentence numbers. For example, if we repeat the previous step three times, that is, if we set the number of repetitions as $r = 3$, then the unit interval [0, 1] is divided into $5^3 = 125$ subintervals, considered to be 125 sentences with serial sentence number from 1 to 125. In this case, because the text length of 125 sentences is too short to calculate ACVFs with statistical reliability, we concatenate 10 different obtained hierarchical probability distributions, each having 125 sentences to form a text with 1250 sentences. Concatenation of 10 hierarchical probability distributions is applied to the cases of $r = 2,3,4$ and concatenation of 5 hierarchical probability distributions is used for $r = 5$.

6) The vertical axis of the probability distribution is also converted. Originally, a vertical axis value represents a density function value defined within the unit interval [0, 1]. However, because we intend to obtain time-varying probabilities, these vertical axis values must be converted into word occurrence probabilities at corresponding sentences. That is, we want to transform the probability density function to the time-varying probability, as in Figure 6(c) and Figure 6(f). The conversion of vertical axis values is achieved by dividing each value by the maximum value on the axis. For example, consider the case in which step 4 was repeated 3 times, causing the density function to take 5 different values. After

division, the vertical values are still limited to one of 5 values, but their maximum is now 1. Each of the five values corresponds to the probability of word occurrence in sentences with 5 different relevancies. The lowest value represents the probability of word occurrence in sentences in irrelevant chapters, which is equal to 0. The second, third, and fourth lowest values represent probabilities of word occurrence at sentences in relevant chapters, sections, and paragraphs, respectively. The highest value represents the probability of word occurrence in relevant sentences, and this probability equals 1, indicating that the word always appears in relevant sentences. As the recursive procedures described above clearly show, relevant sections must be in upper layers of relevant chapters, relevant paragraphs must be in relevant sections, and relevant sentences must be in relevant paragraphs.

Figure 9(a), Figure 9(d), Figure 9(g), and Figure 9(j) show examples of time-varying probabilities obtained by the procedures described above. These plots correspond to cases where the recursive procedure is repeated 2, 3, 4, and 5 times, respectively. Figure 9 also presents simulated signals X_t generated from the time-varying probabilities using the simple Monte Carlo procedure (Figure 9(b), Figure 9(e), Figure 9(h), and Figure 9(k)) and ACVFs calculated from simulated X_t (Figure 9(c), Figure 9(f), Figure 9(i), and Figure 9(l)).

Note that the time-varying probabilities shown in Figure 9(a), Figure 9(d), Figure 9(g), and Figure 9(j) reveal the two features observed in the time-varying probabilities of actual Type-I words. That is, the discretization of probability values observed for actual Type-I words is well reproduced in Figure 9(a), Figure 9(d), Figure 9(g), and Figure 9(j). Similarly, these plots confirm that aggregation of probabilities at the same level among several discretized levels on the horizontal axis is also reproduced.

Another important finding observed in the ACVFs in these plots is that dynamic correlations become more prominent and durations of dynamic correlations increase with the number of repetitions of the recursive procedure. This indicates that the origin of dynamic correlations observed for Type-I words is closely connected to the hierarchical structure of a considered text. Judging from the ACVFs in Figure 9, suitable repetition times to reproduce actual ACVFs observed for Type-I words seem to be $r=4$ and $r=5$ because long duration times over several tens of sentences are achieved for these cases (Figure 9(i) and Figure 9(l)).

Figure 10 shows the results of fittings using the KWW function (Equation (15)) to ACVFs displayed in Figure 9(i) and Figure 9(l). The fittings are good. In particular, obtained optimized parameters for $r=4$, namely, $\tau \cong 2.05$ and $\beta \cong 0.265$ (Figure 10(a)), are considered to be typical values for actual Type-I words. Regarding K_0, the obtained value of $K_0 \cong 0.159$ is too large compared with typical values for actual Type-I words because typical K_0 values for actual Type-I words are in the range of 0.02 - 0.04 (see Figure 1). However, K_0 is easily adjusted by multiplying all values for time-varying probabilities by some constant $c<1$, because this multiplication uniformly lowers the probability of word

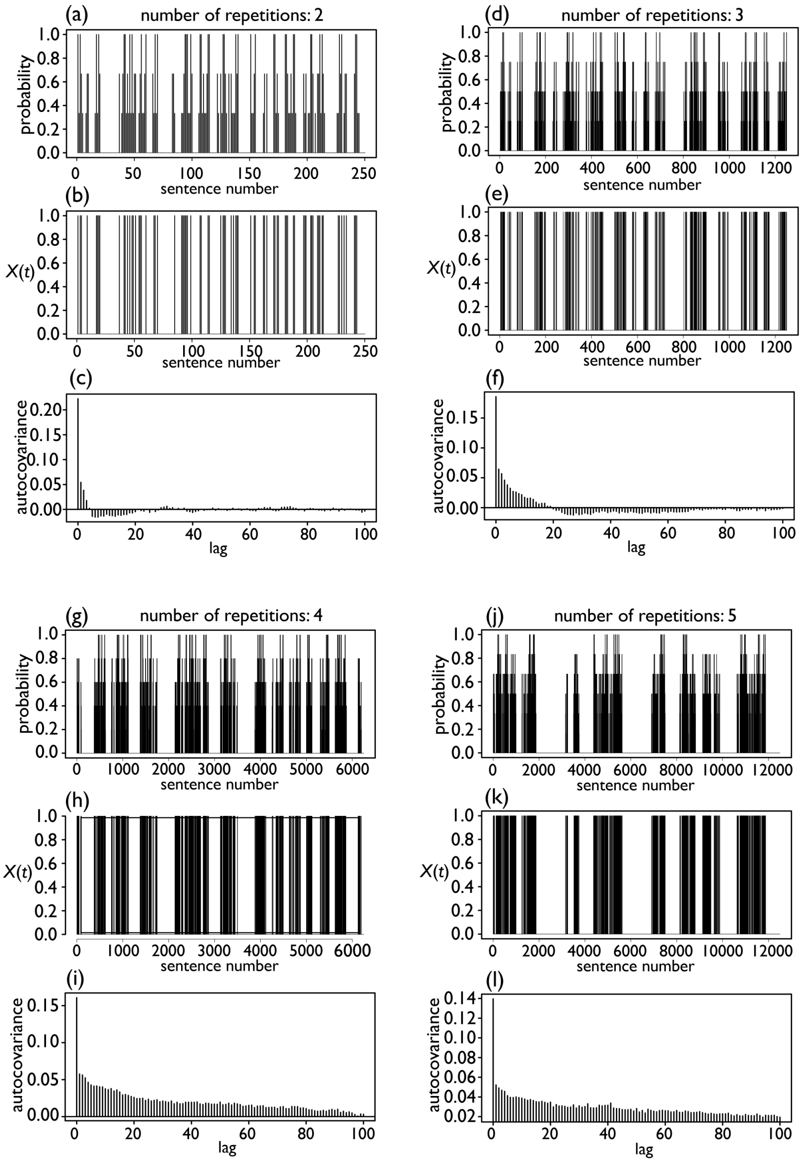

Figure 9. Time-varying probabilities of word occurrence ((a), (d), (g) and (j)), simulated signals X_t generated using a simple Monte Carlo procedure with the time-varying probabilities ((b), (e), (h) and (k)), and ACVFs calculated from simulated signals X_t ((c), (f), (i) and (l)). Plots (a)-(c) show the case of two repetitions of the recursive procedure, (d)-(f) show the case of three repetitions, (g)-(i) show the case of four repetitions, and (j)-(l) show the case of five repetitions.

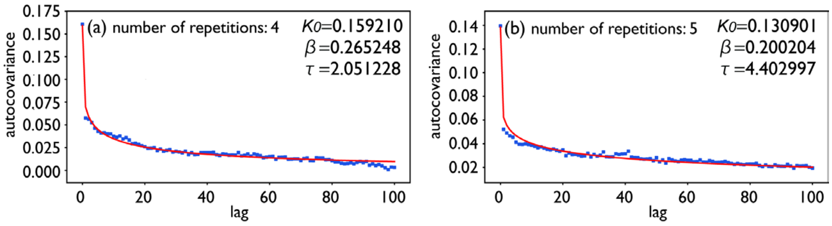

Figure 10. Fitting results by use of KWW function (red curves) to ACVFs (blue plots) obtained from simulated X_t. The recursive procedure for constructing the time-varying probability was repeated (a) 4 times and (b) 5 times.

occurrence and hence lowers the values $\bar{X}$ and $K_0 = \bar{X}\left(1-\bar{X}\right)$. The disagreement in K_0 mentioned above is thus not serious, and we can therefore conclude that the proposed model for recursive probability distribution appropriately reproduces dynamic correlations of Type-I words.

6. Conclusions

Type-I words are those used to describe notions or ideas in written texts over some duration of sentences in context-specific manners. Therefore, if we consider occurrences of a given word as time-series data by regarding serial sentence numbers as time, the words exhibit dynamic correlations that are well captured in autocovariance functions (ACVFs). In this study, we investigated a stochastic process that governs word occurrences with the hope that the origin of dynamic correlations of Type-I words is well interpreted by that process.

To identify this process, we applied additive binary Markov chain theory to observe word occurrence signals for considered Type-I words to estimate memory functions and time-varying probabilities of word occurrence. The obtained time-varying probabilities represent the probability of word occurrence as a function of time (*i.e.*, serial sentence number). These probabilities show two distinctive features: values of time-varying probabilities seem to be discretized, and similar probability values aggregate in some time-axis range.

To explain these features, we attempted to construct a recursive model of probability distribution that considers hierarchical structures of documents such as chapters, sections, subsections, paragraphs, and sentences. This construction was based on a recursive probability redistribution in a probability density function defined on the unit interval [0, 1]. After obtaining the hierarchical probability distribution in the density function, we converted the density function to time-varying probabilities of word occurrence by discretizing the horizontal axis and rescaling the vertical axis. We found that the obtained time-varying probabilities well reproduce the two distinctive features mentioned above. By using those time-varying probabilities, we generated signals X_t representing word occurrence or non-occurrence over the entire text and calculated ACFVFs from the signals. The resultant ACVF with four repetitions, which is the number of recursive procedure repetitions needed to construct the hierarchical probability distribution, was quite similar to actual ACVFs for Type-I words.

At this stage, we have not yet considered optimization of the construction procedure for the recursive model of probability distribution. For example, the use of five subintervals in the procedure was tentatively determined without statistical verifications. To construct more realistic models, we should extract the number of subintervals from some adequate probability distribution each time this number is needed. Therefore, increased sophistication of the construction procedures described in Section 5 is one area for future research. Another is to interpret the time-varying probabilities of word occurrences as a fractal time series. As the recursive procedure for constructing the hierarchical probability distribution shows, the obtained probability distribution can be regarded as a statistical fractal [12] [13], so the time-varying probabilities thus obtained can be considered as a fractal time series [14] [15]. Any relations between fractal dimension of time-varying probabilities and characteristics of dynamic correlations may provide new insights into written texts. A more detailed study along this line, through which we will try to identify any such relations, is reserved for future work.

Acknowledgements

We thank Dr. Yusuke Higuchi for useful discussion and illuminating suggestions. This work was supported in part by JSPS Grant-in-Aid (Grant No. 25589003 and 16K00160).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Ogura, H., Amano, H. and Kondo, M. (2019) Measuring Dynamic Correlations of Words in Written Texts with an Autocorrelation Function. *Journal of Data Analysis and Information Processing*, **7**, 46-73. https://doi.org/10.4236/jdaip.2019.72004

[2] Juliane, W., Christopher, Z. and Dirk, W. (2018) Modeling Long Correlation Times Using Additive Binary Markov Chains: Applications to Wind Generation Time Series. *Physical Review E*, **97**, 32138-32150. https://doi.org/10.1103/PhysRevE.97.032138

[3] Montemurro, M.A. and Pury, P.A. (2002) Long-Range Fractal Correlations in Literary Corpora. *Fractals*, **10**, 451-461. https://doi.org/10.1142/S0218348X02001257

[4] Chatzigeorgiou, M., Constantoudis, V., Diakonos, F., Karamanos, K., Papadimitriou, C., Kalimeri, M. and Papageorgiou, H. (2017) Multifractal Correlations in Natural Language Written Texts: Effects of Language Family and Long Word Statistics. *Physica A*, **469**, 173-182. https://doi.org/10.1016/j.physa.2016.11.028

[5] Pillai, S.U. and Papoulis, A. (2002) Probability, Random Variables and Stochastic Processes. McGraw-Hill Europe.

[6] Hwei, H. (1997) Probability, Random Variables, and Random Processes. McGraw-Hill, New York.

[7] Berchtold, A. and Raftery, A.E. (2002) The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series. *Statistical Science*, **17**, 328-356. https://doi.org/10.1214/ss/1042727943

[8] Ataharul Islam, M. and Chowdhury, R.I. (2006) A Higher Order Markov Model for Analyzing Covariate Dependence. *Applied Mathematical Modelling*, **30**, 477-488. https://doi.org/10.1016/j.apm.2005.05.006

[9] Ching, W.K., Huang, X., Ng, M.K. and Siu, T.K. (2013) Higher-Order Markov Chains. In: *Markov Chains. International Series in Operations Research & Man-*

agement Science, Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-6312-2

[10] Yampol'skii, V.A., Apostolov, S.S., Melnyk, S.S., Usatenko, O.V. and Maiselis, Z.A. (2006) Memory Functions and Correlations in Additive Binary Markov Chains. *Journal of Physics A*: *Mathematical and General*, **39**, 14289-14301. https://doi.org/10.1088/0305-4470/39/46/004

[11] Melnyk, S.S., Usatenko, O.V. and Yampol'skii, V.A. (2006) Memory Functions of the Additive Markov Chains: Applications to Complex Dynamic Systems. *Physica A*: *Statistical Mechanics and Its Applications*, **361**, 405-415. https://doi.org/10.1016/j.physa.2005.06.083

[12] Padua, R. and Borres, M. (2017). From Fractal Geometry to Statistical Fractal. *Recoletos Multidisciplinary Research Journal*, **1**. https://rmrj.usjr.edu.ph/rmrj/index.php/RMRJ/article/view/80

[13] Chatterjee, S. and Yilmaz, M.R. (1992) Chaos, Fractals and Statistics. *Statistical Science*, **7**, 49-68. https://doi.org/10.1214/ss/1177011443

[14] Kantelhardt, J.W. (2012) Fractal and Multifractal Time Series. In: Meyers, R., Ed., *Mathematics of Complexity and Dynamical Systems*, Springer, New York. https://doi.org/10.1007/978-1-4614-1806-1_30

[15] Li, M. (2010) Fractal Time Series—A Tutorial Review. *Mathematical Problems in Engineering*, **2010**, Article ID: 157264. https://doi.org/10.1155/2010/157264

Appendix

In this study, we selected the English edition of five famous academic books as samples of written texts and analyzed all Type-I words appearing therein to clarify the features of Type-I words. Unlike in our previous study, we omitted novels from our text samples because the features of Type-I words are more prominent in academic books [1]. The five books were downloaded as text files from Project Gutenberg (https://www.gutenberg.org). **Table A1** lists the details of the five books used.

Table A1. Summary of selected English texts.

Short name	Title	Author	Download URL
Darwin	*On the Origin of Species*	Charles Darwin	https://www.gutenberg.org/ebooks/1228
Einstein	*Relativity: The Special and General Theory*	Albert Einstein	https://www.gutenberg.org/ebooks/5001
Freud	*Dream Psychology*	Sigmund Freud	https://www.gutenberg.org/ebooks/15489
Smith	*An Inquiry into the Nature and Causes of the Wealth of Nations*	Adam Smith	https://www.gutenberg.org/ebooks/3300
Kant	*The Critique of Pure Reason*	Immanuel Kant	https://www.gutenberg.org/ebooks/4280

Figure A1 presents some basic statistics of the five books, evaluated after the pre-processing procedures.

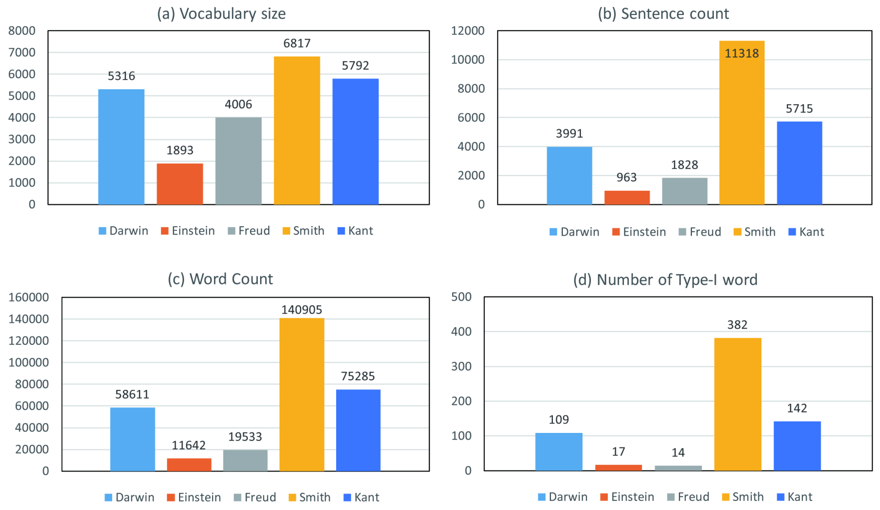

Figure A1. Basic statistics for the five texts, evaluated after pre-processing procedures.