Article

Modeling Long-Range Dynamic Correlations of Words in Written Texts with Hawkes Processes

Hiroshi Ogura *, Yasutaka Hanada, Hiromi Amano and Masato Kondo

Department of Information Science, Faculty of Arts and Sciences, Showa University, Fujiyoshida 403-0005, Japan; hanada@cas.showa-u.ac.jp (Y.H.); kayanm@cas.showa-u.ac.jp (H.A.); mkondo@nr.showa-u.ac.jp (M.K.)
* Correspondence: ogura@cas.showa-u.ac.jp

Abstract: It has been clarified that words in written texts are classified into two groups called Type-I and Type-II words. The Type-I words are words that exhibit long-range dynamic correlations in written texts while the Type-II words do not show any type of dynamic correlations. Although the stochastic process of yielding Type-II words has been clarified to be a superposition of Poisson point processes with various intensities, there is no definitive model for Type-I words. In this study, we introduce a Hawkes process, which is known as a kind of self-exciting point process, as a candidate for the stochastic process that governs yielding Type-I words; i.e., the purpose of this study is to establish that the Hawkes process is useful to model occurrence patterns of Type-I words in real written texts. The relation between the Hawkes process and an existing model for Type-I words, in which hierarchical structures of written texts are considered to play a central role in yielding dynamic correlations, will also be discussed.

Keywords: Hawkes process; autocorrelation function; stochastic process; waiting time distribution; word occurrence; long-range correlation

1. Introduction

Considering written texts as time series data and analyzing occurrence patterns of any components of texts by using methods of time series analysis have been attempted for various purposes including rhythm analyses [1–4], analysis of word distributions [5], gathering language statistics for rare words [6], and measuring importance of words [7]. One of the major and actively investigated problems in time series analysis of written texts is to elucidate the origin of long-range dynamic correlations which are observed at various levels of components [8–13]. For example, although word-level long-range correlations have been clearly detected especially for words which play important roles in describing the main theme of texts [7,14,15], the stochastic process which brings the dynamic correlations to word occurrences is unknown. More specifically, there exist models of stochastic process yielding long-range dynamic correlations of important words [15,16], but no clear conclusion has yet been obtained. Since occurrences of a word in a written text can be treated as a point process as will be described in the next section, this problem is equivalent to identifying a suitable point process which can reproduce the occurrence patterns of the considered word. We treat this problem in this study, and propose a Hawkes process as a powerful and useful candidate for this point process.

In a wide range of fields, it has gradually been realized that point processes with strong long-range correlations are suitably described by using Hawkes processes [17]. For example, the Hawkes processes have been successfully adopted to model occurrences of earthquakes [18–20], neuronal spikes [21–23], transactions in financial markets [24], behavior patterns of people on social networking sites [25,26], and patterns of COVID-19 transmission [27]. We expect that the process is also effective for describing word occurrence patterns having dynamic correlations.

Before explaining why this is to be expected, here we briefly describe the definitions of Type-I and Type-II words. Words that have dynamic correlations in their occurrence patterns in a document are called Type-I words, while words that do not exhibit any type of dynamic correlations are called Type-II words. As will be described later, an autocorrelation function (ACF) is used to determine whether a considered word has dynamic correlation or not.

The reasons why we expect the Hawkes process to be applicable to the description of occurrence patterns of Type-I words are as follows.

- Important words which play central roles in describing some notions or ideas in texts are all classified into Type-I words [7,14]. This is because every important word appears in texts with a "burst" nature; once an important word appears in a considered text, then it appears again and again for the duration in which the notion or the idea is described. Thus, the "burst" nature brings dynamic correlations in the occurrence patterns and consequently the word becomes Type-I. The "burst" nature of word occurrences reminds us of the fact that once an earthquake occurs, earthquakes occur more frequently for a short period of time. The Hawkes process can adequately treat such "burst" phenomena because it has a built-in property of self-excitability.
- Type-I words often show long durations of dynamic correlations ranging from several tens to several hundreds of sentences [7,14,15]. These durations correspond to lengths of sentences in which some notion/idea that deeply related with a considered word are described. The Hawkes process is expected to be able to reproduce such long-range dynamic correlations because of its self-excitability.

Note that the Hawkes process is expected to describe only the occurrence patterns of Type-I words, not those of Type-II words. This is because Type-II words appear with almost constant probabilities of occurrence regardless of context, which is different from the self-excited pattern that can be described by the Hawkes process.

In this study, we try to describe occurrence patterns of Type-I words by use of the Hawkes process and check the validity of the description. To our knowledge, this is the first attempt to apply the Hawkes process to analyze written texts. If the description by the Hawkes process is successful, it will not only be an important application of the Hawkes process, but also be an important step in attempts to describe document generation by stochastic models.

The rest of the paper is organized as follows. In the next section, we present the method of how to find/optimize an adequate Hawkes process for a considered Type-I word, given occurrence patterns of the word. The section also presents procedures that check the validity of the optimized Hawkes process through simulation. Section 3 is devoted to describing the results of validation of the optimized Hawkes processes. In Section 4, we discuss the relation between the Hawkes process and an existing stochastic model of generating Type-I words in which a hierarchical structure of written texts (volumes, chapters, sections, subsections, paragraphs, sentences) is taken into account. In the last section, we give our conclusion and indicate a direction of future study.

2. Methodology

The main purpose of this study is to verify whether the process of generating Type-I words can be regarded as a Hawkes process or not. To achieve the purpose, we take the following 3 steps.

1. From the occurrence patterns of Type-I words in real written texts, we calculate autocorrelation functions (ACFs) to characterize dynamic correlations of these words. Details are given in Section 2.1.
2. We optimize a Hawkes process so that it can express the stochastic process of yielding observed occurrence patterns of a considered word. For the optimization, we utilize a log-likelihood function of the Hawkes process, and maximize the function by optimizing parameters of the process, given the observed word occurrence signals. Then, the Hawkes process, having the kernel function with the optimized values of

parameters, is considered to be the best way to express the observed word occurrence signal of the considered word in the sense of maximum likelihood estimation (MLE). A detailed description is given in Section 2.2.

3. We generate word occurrence signals of the considered word from the optimized Hawkes process. This is achieved by standard simulation procedures of point processes [28]. The simulated word occurrence signals are then used to calculate ACFs. The ACFs of the simulated signals are compared with the ACFs obtained in step 1 to validate the optimized Hawkes process. Section 2.3 explains the procedure in detail.

As described above, we mainly utilize ACF as a characteristic quantity of a stochastic process. In general, waiting time distributions (WTDs) are also used along with ACFs in time series analysis because they both contain equivalent information in principle. However, ACFs is more suitable in this study, because ACFs offer more precise descriptions of dynamic correlations for the case of word occurrence signals [16].

2.1. Model Functions of ACFs for Type-I and Type-II Words

One of the methods to convert occurrence patterns of a considered word in the considered text to time series data is to utilize the following definition of a binary time-dependent signal $X(t)$:

$$X(t) = \begin{cases} 0 & \text{(if a considered word does not occur in the } t\text{th sentence)} \\ 1 & \text{(if a considered word occurs in the } t\text{th sentence)} \end{cases}. \tag{1}$$

Here, t is an ordinal number of sentences that assigned from the first to the last sentences in a considered text and it plays a role of time along the text. By defining a binary time-dependent signal $X(t)$ as in Equation (1), we can utilize various results in point process theory for our investigation.

Two examples of word occurrence signals $X(t)$ are shown in Figure 1a,d. The two words, "organ" and "seem" used in the figure, are both picked from "On the Origin of Species" by Charles Darwin. The word "organ" is a typical Type-I word in the book, and thus it shows a "burst" nature in $X(t)$ (Figure 1a) and in a cumulative count of the word occurrences along the text (Figure 1b). On the other hand, since the word "seem" is a typical Type-II word in the book, its $X(t)$ (Figure 1d) and cumulative count (Figure 1e) do not exhibit a "bursty" nature but show word occurrences with a constant occurrence rate, which indicates the occurrences are purely governed by chance.

Figure 1c,f shows ACFs of word "organ" and "seem", which are calculated from $X(t)$ displayed in Figure 1a,d, respectively. Since $X(t)$ is a discrete-time signal, a general definition of ACF for a continuous-time signal $A(t)$ given by

$$\Phi(t) = \frac{\lim_{T \to \infty} \int_0^T A(\tau)A(\tau + t)d\tau}{\lim_{T \to \infty} \int_0^T A(\tau)A(\tau)d\tau}, \tag{2}$$

is extended for our discrete-time case and the extended definition is used for the calculations [7]. The ACF shown in Figure 1c indicates that the dynamic correlation gradually decreases as lag increases, which is a typical behavior of ACFs in usual linear systems. On the other hand, ACF in Figure 1f shows abrupt decrease from its initial value one at $t = 0$ to some constant value of almost zero at $t > 0$. This behavior of ACF indicates that Type-II words are generated from stochastic processes that do not have any type of dynamic correlations.

To represent characteristics of ACFs of Type-I words, we introduce an empirical model function which gives satisfactory fittings of observed ACFs for Type-I words [7,14]. The

function is called the Kohlrausch–Williams–Watts (KWW) function and has a stretched exponential form given by

$$\Phi_{KWW}(t) = \alpha \exp\left\{-\left(\frac{t}{\tau}\right)^{\beta}\right\} + (1-\alpha), \tag{3}$$

where α, β and τ are fitting parameters satisfying $0 < \alpha \le 1$, $0 < \beta \le 1$ and $0 < \tau$. The fitting result for the Type-I word "organ" by use of Equation (3) is also shown as a red curve in Figure 1c with optimized values of fitting parameters.
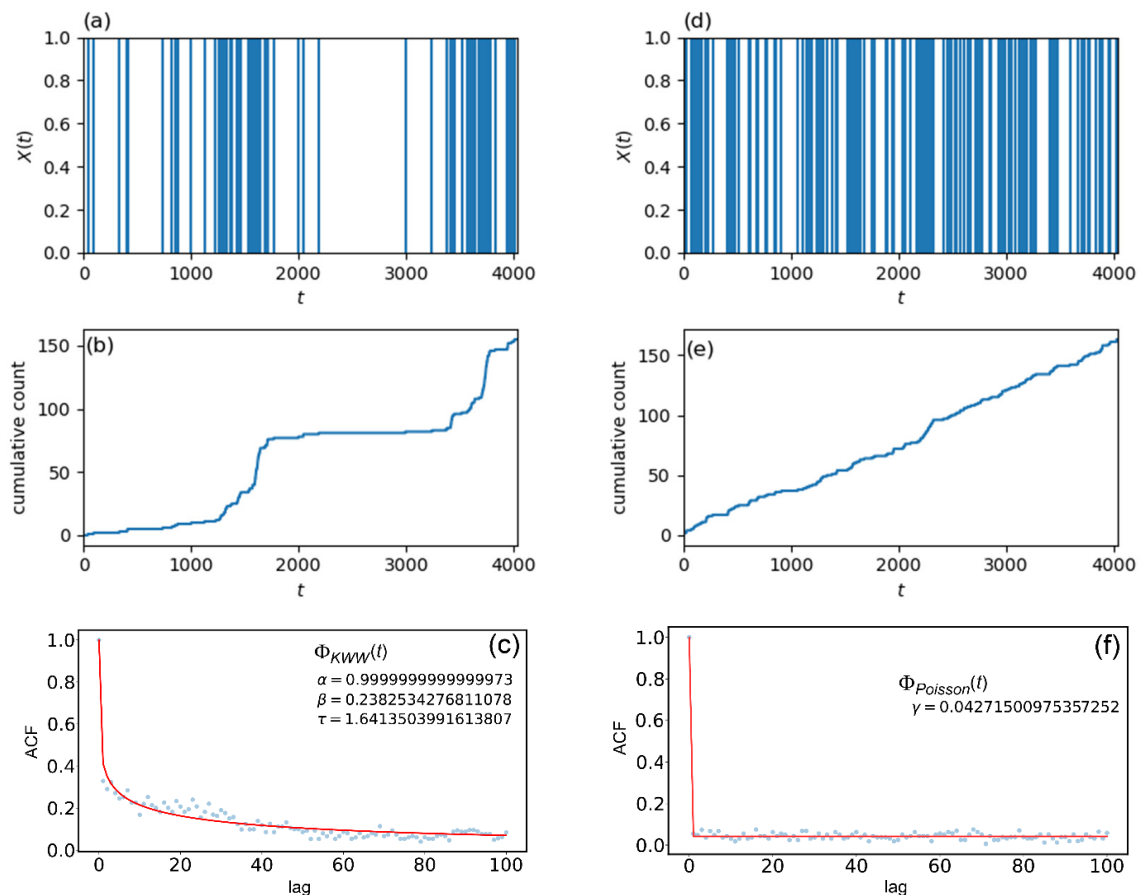

Figure 1. Word occurrence signals $X(t)$ (**a**,**d**), as defined by Equation (1); cumulative count of word occurrences (**b**,**e**), and ACFs (**c**,**f**) of the words "organ" (**a**–**c**) and "seem" (**d**–**f**). The words "organ" and "seem" are typical Type-I and Type-II words, respectively, picked from the Darwin text. Occurrences of "organ" are in a context-specific and bursty manner (**a**,**b**), and long-range dynamic correlations are seen in the ACF plot (**c**), while occurrences of "seem" show an approximately constant occurrence rate (**d**,**e**) and no type of dynamic correlation is seen in ACF (**f**). Red lines in (**c**,**f**) show best fitted results by use of Equations (3) and (4), respectively.

A model function for ACFs of Type-II words is given by

$$\Phi_{Poisson}(t) = \begin{cases} 1 & (t = 0) \\ \gamma & (t > 0) \end{cases}, \tag{4}$$

where γ is a fitting parameter satisfying $0 < \gamma < 1$ and it actually takes a value almost equal to zero. Note that Equation (4) is theoretically derived under an assumption that the stochastic process yielding Type-II words is a Poisson point process [7]. In Figure 1f, the result of fitting with Equation (4) is displayed as a red line.

Classifying whether a given word is Type-I or Type-II is performed as follows. First, we execute two non-linear least squares fittings by use of Equations (3) and (4) simultaneously on the one observed ACF of a considered word and compare two values of the Bayesian information criterion (BIC) each of which obtained at the fitting using Equation (3) and at the fitting using Equation (4). If the BIC of the fitting using Equation (3) is smaller than that using Equation (4), then the word is classified as Type-I and otherwise the word is classified as Type-II. We can classify an arbitrary word in written texts as Type-I or Type-II without any ambiguities by using this procedure with some additional criteria [7].

Seven texts employed in this study, which are famous academic books chosen so as to represent wide range of written texts, are listed in Table 1 with their short names and some information. The procedure for preparing texts is the same as in the previous paper [16]. In this study, we determined Type-I words to be analyzed through following steps. First, words that appear at least 50 sentences or more in each of the 7 texts are chosen as frequent words. All frequent words are then classified whether Type-I or Type-II by the classification method described above. Finally, stop words are removed from the set of the Type-I words, and remaining words are used for further analysis. As stop words, we used the same list of stop words that the MySQL 8.0 system uses for full-text queries (https://dev.mysql.com/doc/refman/8.0/en/fulltext-stopwords.html accessed on 2 March 2022).

Table 1. Summary of English texts employed.

Short Name	Title	Author	Vocabulary Size	Length in Sentences	Number of Type-I Words
Darwin	On the Origin of Species	Charles Darwin	5728	4036	124
Einstein	Relativity: The Special and General Theory	Albert Einstein	2222	1107	20
Freud	Dream Psychology	Sigmund Freud	4520	1977	18
Kant	The Critique of Pure Reason	Immanuel Kant	5157	5920	157
Lavoisier	Elements of Chemistry	Antoine Lavoisier	5558	3899	122
Plato	The Republic	Plato	5686	5268	49
Smith	An Inquiry into the Nature and Causes of the Wealth of Nations	Adam Smith	8399	11906	433

2.2. Log-Likelihood Function of Hawkes Process

Here we recall the basic notation of Hawkes process for later reference [25,29]. A Hawkes process is a kind of self-exciting point process and has been applied in diverse areas because of its self-exciting nature. This 'self-excite' means that each arrival/occurrence increases the rate of future arrivals/occurrences for some period. The conditional intensity function of the Hawkes process is given by

$$\lambda(t|H_t) = \mu + \sum_{t_i < t} g(t - t_i), \tag{5}$$

where H_t denotes the history of all events occurring before time t, μ is the background intensity, t_i means a time of the ith event occurrence before time t, and $g(\tau)$ is a kernel function which determines how past events will affect the future. Two frequent choices of the kernel function, which are also used in this study, are

$$g_{exp}(\tau) = ab \exp(-b\tau), \tag{6}$$

$$g_{pow}(\tau) = \frac{K}{(\tau + c)^p}, \tag{7}$$

where a and b in Equation (6) and c, p and K in Equation (7) are parameters of the kernels taking non-negative real values. Examples of an occurrence signal, a cumulative count, and a conditional intensity function for a Hawkes process with an exponentially decaying kernel, Equation (6), are shown in Figure 2. Figure 3 shows corresponding quantities to Figure 2 for a Hawkes process with a power-law decaying kernel, Equation (7). As seen in these figures, since a value of the intensity function at the current time is enhanced by the history of past generated events, once an event is generated, it tends to be generated intensively in a short period of time in Hawkes processes.
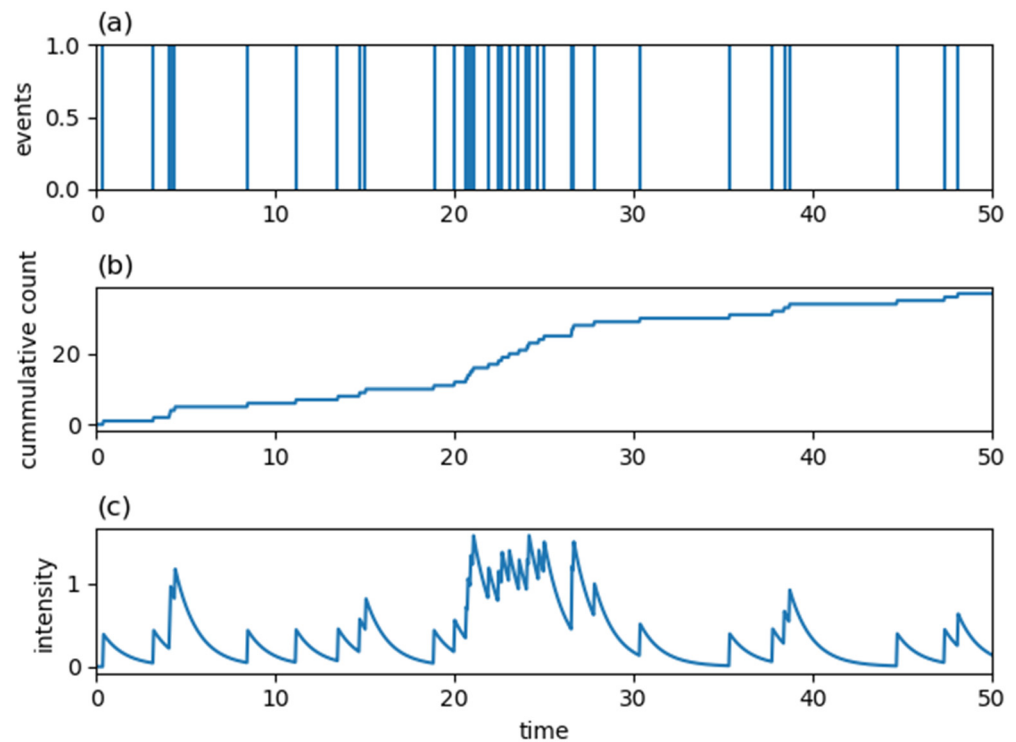

Figure 2. Examples of (**a**) the occurrence signal of events, (**b**) the cumulative count of events and (**c**) the conditional intensity function for the Hawkes process with $\lambda(t|H_t) = 0.2 + \sum_{t_i < t} 0.4 \exp\{-0.8(t - t_i)\}$.

As mentioned before, the main objective of this study is to verify whether the Hawkes process is eligible to describe stochastic processes yielding Type-I words or not. The first step of the verification is to find the Hawkes process that best approximates the stochastic process yielding the actual occurrence signal of a considered word within the descriptive power of Hawkes processes. The most suitable Hawkes process, which is expected to be able to reproduce real word occurrence signals $X(t)$ of a considered Type-I word, is searched for by maximizing the log-likelihood function of the Hawkes process. Therefore, a searching method to find the optimized Hawkes process described below is in the manner of standard maximum likelihood estimation (MLE).

Given the history of all events in the time interval $[0, T]$, i.e., given the record of all occurrence times

$$\boldsymbol{D}^{[0,T]} = \{t_i\}_{i=1}^n, \tag{8}$$

the log-likelihood function of the Hawkes process having conditional intensity function of Equation (5) is given by [30]

$$l\left(\boldsymbol{\theta}|\boldsymbol{D}^{[0,T]}\right) = \log L(\boldsymbol{\theta}|\boldsymbol{D}^{[0,T]}) = \sum_{i=1}^n \log \lambda(t_i|H_{t_i}) - \int_0^T \lambda(t|H_t)dt, \tag{9}$$

where θ denotes a set of parameters of the Hawkes process. If we combine Equations (5) and (6), then Equation (9) becomes

$$
\begin{aligned}
l_{exp}&\left(\mu,\, a,\, b\,\middle|\,\{t_i\}_{i=1}^n\right)\\
&= \sum_{i=1}^n \log\left[\mu + \sum_{j<i} ab\exp\left\{-b(t_i - t_j)\right\}\right]\\
&\quad - \left[\mu T + \sum_{i=1}^n a\{1 - \exp(-b(T - t_i))\}\right].
\end{aligned}
\tag{10}
$$

In the same way, combining Equations (5) and (7) makes Equation (9) to

$$
\begin{aligned}
l_{pow}&\left(\mu,\, c,\, K,\, p\,\middle|\,\{t_i\}_{i=1}^n\right)\\
&= \sum_{i=1}^n \log\left[\mu + \sum_{j<i} \frac{K}{(t_i - t_j + c)^p}\right]\\
&\quad - \left[\mu T + \sum_{i=1}^n \frac{K}{p-1}\left\{\frac{1}{c^{p-1}} - \frac{1}{(T - t_i + c)^{p-1}}\right\}\right].
\end{aligned}
\tag{11}
$$

In our case, the set of occurrence times of events, $\{t_i\}_{i=1}^n$ in Equations (10) and (11), is equivalent to the set of times (sentence numbers) at which word occurrence signal $X(t)$ takes value one. MLE of $\theta_{exp} = (\mu, a, b)$ for exponentially decaying kernel, Equation (6), and that of $\theta_{pow} = (\mu, c, K, p)$ for power-low decaying kernel, Equation (7), are thus obtained by substituting $\{t_i\}_{i=1}^n$ to Equations (10) and (11), respectively, and then maximizing these functions. To maximize Equation (10) as a function of $\theta_{exp} = (\mu, a, b)$, and to maximize Equation (11) as a function of $\theta_{pow} = (\mu, c, K, p)$, we use a quasi-Newton method with the BFGS algorithm [31].
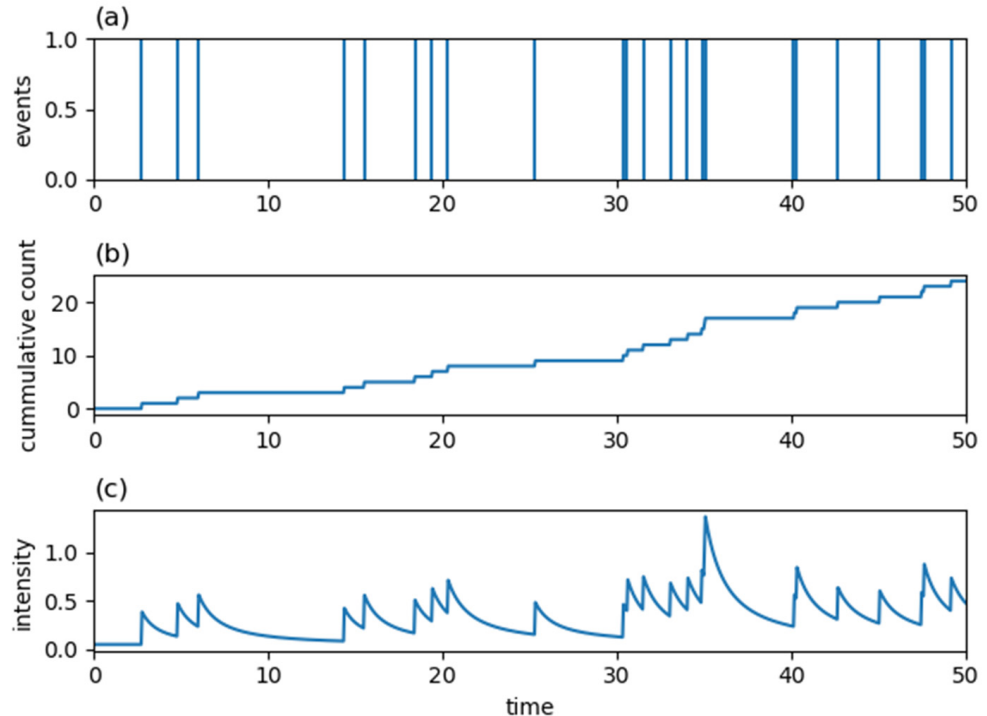

Figure 3. Examples of (**a**) the occurrence signal of events, (**b**) the cumulative count of events and (**c**) the conditional intensity function for the Hawkes process with $\lambda(t|H_t) = 0.05 + \sum_{t_i < t} \frac{0.9}{(t - t_i + 1.7)^{1.8}}$.

In actual procedures of the quasi-Newton method, we introduce new parameters m_0, a_0, b_0, c_0, K_0 and p_0 instead of using original parameters μ, a, b, c, K and p in order to stabilize convergence calculations. The original and new parameters are related as follows.

$$\mu = 0.5 + 0.5\tanh(m_0), \tag{12}$$

$$a = 0.5 + 0.5\tanh(a_0), \tag{13}$$

$$b = \exp(b_0), \tag{14}$$

$$c = \exp(c_0), \tag{15}$$

$$K = 0.5 + 0.5\tanh(K_0), \tag{16}$$

$$p = 3.2 + 2.0\tanh(p_0). \tag{17}$$

Note that all of the original parameters have a restriction that they should be non-negative, but newly introduced parameters can take any real values. Equations (12)–(17) also indicate that the conditions of the original parameters expressed by the following inequalities are automatically satisfied; $0 < \mu < 1$, $0 < a < 1$, $0 < b$, $0 < c$, $0 < K < 1$, $1.2 < p < 5.2$. The last condition for p, $1.2 < p < 5.2$, is needed in a practical sense because the Hawkes process becomes unstable when $p < 1$. To determine which of the two kernel functions, Equation (6) and Equation (7), is more appropriate given actual word occurrence data $\{t_i\}_{i=1}^n$, we use the Akaike's Information Criterion (AIC) for the judgement; i.e., we select Equation (6) as a better kernel function when the AIC of MLE calculation with Equation (10) is smaller than that with Equation (11), and otherwise we select Equation (7). The AIC is an estimator of prediction error and therefore it provides a mean for model selection in the same way as BIC [32]. The reason for using AIC instead of using BIC is that AIC has a proven track record of being applied in model selection involving Hawkes processes [30].

2.3. Simulating Word Occurrence Events from Hawkes Process

The conditional intensity function, Equation (5), can be determined for each of the Type-I words by using optimized kernel parameters $\theta_{exp} = (\mu, a, b)$ in Equation (6) or $\theta_{pow} = (\mu, c, K, p)$ in Equation (7), which are obtained by MLE procedures with the word occurrence signal $\{t_i\}_{i=1}^n$ of a considered word. Once the conditional intensity function, Equation (5), is fixed, then we can simulate word occurrence signal $X(t)$ from the Hawkes process having that fixed intensity function. The simulation period T was set to be equal to the actual text length (length in sentences of a considered book) to check the validity of the Hawkes process; i.e., if the Hawkes process employed in the simulation is valid, then the number of occurrences in simulated $X(t)$ is almost equal to the number of occurrences of a considered word in real written text. For simulating $X(t)$, we use a standard thinning algorithm for simulation of point processes [28]. Since the simulated word occurrence times t_i can take any real values between 0 and T, we convert these values of t_i to the nearest integer values in order to meet the condition that real t_i take only integer values in time along the text.

Then, we calculate ACFs from the simulated signal $X(t)$ and we further perform the curve fitting with Equation (3), as described in Section 2.1. We then evaluate the degree of agreement between the ACFs calculated from actual word occurrence signals $X(t)$ and the ACFs calculated from simulated $X(t)$ from the Hawkes process. The evaluation is made through comparisons of two sets of optimized fitting parameters; one is the set of parameters in Equation (3) obtained for real ACFs and the other is the set of parameters obtained for simulated ACFs calculated from simulated $X(t)$.

3. Results

3.1. MLE of Hawkes Process

Table 2 summarizes the results of MLE performed given the real word occurrence signal $X(t)$ of the word "organ" which is shown in Figure 1a. The parameter values listed in the table are optimized ones to maximize Equations (10) and (11). Comparing the two AIC values in the table, we can judge that $g_{exp}(\tau)$ is more suitable than $g_{pow}(\tau)$ to simulate the real word occurrence signal $X(t)$. Thus, we employ the Hawkes process with

$g_{exp}(\tau)$ having the parameter values of (μ, a, b) shown in Table 2 to simulate $X(t)$. The simulated $X(t)$, cumulative count obtained from the simulated $X(t)$ and ACF calculated from the simulated $X(t)$ are shown in Figure 4a–c, respectively. Figure 4c also shows the results of the curve fitting using Equation (3). Although the "burst" periods are different in Figures 1 and 4, overall behaviors of ACFs are almost coincide in Figures 1c and 4c.

Table 2. Results of MLE of the Hawkes process with the word occurrence signal of "organ" in Darwin text.

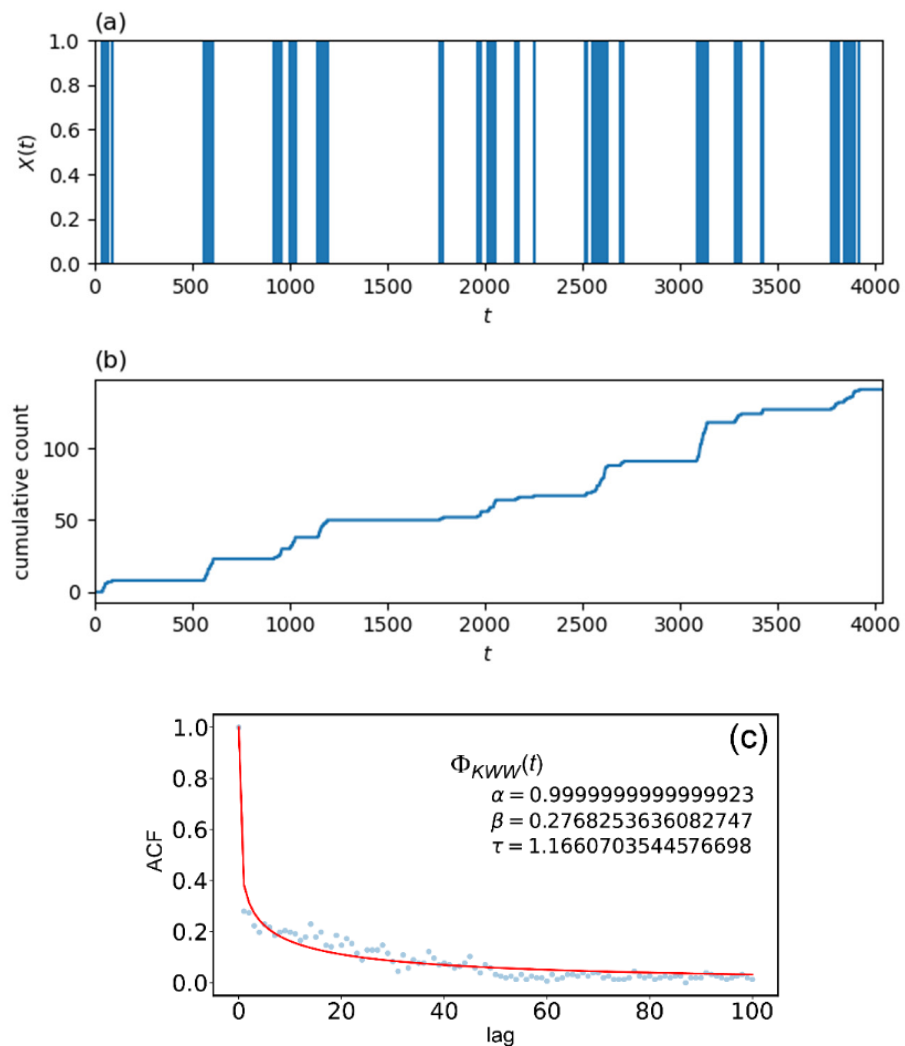| Kernel Type	Parameters Vector	AIC	BIC
$g_{exp}(\tau)$	$(\mu, a, b) = (0.00893, 0.23944, 0.25068)$	540.9253578	546.7790890
$g_{pow}(\tau)$	$(\mu, K, c, p) = (0.00743, 0.62804, 5.05592, 1.56729)$	541.3719225	549.1768974

Figure 4. Results of the simulation for generating word occurrence signal of "organ" in Darwin text. (**a**) Simulated word occurrence signal $X(t)$ obtained from the optimized Hawkes process for "organ", (**b**) cumulative count of word occurrences obtained from simulated $X(t)$, and (**c**) ACFs calculated from simulated $X(t)$ and the best fitted curve by use of Equation (3).

Table 3 shows comparison of the curve fitting results of two different ACFs; one is the ACFs calculated from the real $X(t)$ of the word "organ", and the other is the ACF obtained from simulated $X(t)$ for the same word. From Table 3, it can be seen that the number of word occurrences are almost the same for real and simulated $X(t)$, and the values of other parameters are also similar. This result indicates that the Hawkes process with optimized

parameters can reproduce a word occurrence signal having the same statistical properties as the actual signal of the word "organ".

Table 3. Fitting results for two ACFs of the actual and and the simulated signals.

Type of X(*t*)	Number of Word Occurrences	Fitting Parameters in Equation (3)			BIC
		α	β	τ	
observed signal in real written text	155	1.00000	0.23825	1.64135	−691.30057
simulated signal from Hawkes process	157	1.00000	0.27683	1.16607	−680.82002

3.2. Results of Simulations for All Type-I Words

Figures 5–11 show comparisons of the curve fitting results of two different ACFs for all Type-I words which appear in 7 texts listed in Table 1. In each plot of (a) in these figures, the horizontal axis represents occurrence numbers of words in real written text, while the vertical axis represents occurrence numbers in the simulated $X(t)$. In each plot of (b), (c), and (d), the horizontal axis represents one of the fitting parameters of ACFs obtained from the real signal $X(t)$, while the vertical axis represents that of ACFs obtained from simulated $X(t)$. In each plot of (e), the horizontal axis shows values of fitting parameter γ, which appears in Equation (4), for ACFs obtained from the real signal $X(t)$, while the vertical axis shows values of γ for ACFs obtained from simulated $X(t)$. One may wonder why we use here the parameter γ which appears in the model function for Type-II words, Equation (4). The value of γ is equal to the intensity rate λ in a simple Poisson point process [7] and can therefore be regarded as the averaged value of the conditional intensity function $\lambda(t|H_t)$ of a Hawkes process. Thus, if the horizontal and vertical values in plot (e) are approximately equal, then overall behaviors of real and simulated $X(t)$ are considered to be similar. In each plot of (f), the horizontal axis shows BICs of fitting ACFs obtained from the real signal $X(t)$, while the vertical axis shows BICs of fitting ACFs obtained from simulated $X(t)$. If the horizontal and vertical values in plot (f) are approximately equal, then overall behaviors of ACFs obtained from real and simulated $X(t)$ are considered to be similar. In all plots, one plot point corresponds to one Type-I word; for example, horizontal and vertical values of one plot point in a plot (a) represent the real occurrence number of a considered Type-I word (the horizontal value) and occurrence number in the simulated signal for that word (the vertical value). The red line in each plot shows the relation $y = x$, on which plot points should be located when Hawkes processes are sufficient to represent original word yielding processes. The correlation coefficients for the vertical and horizontal quantities are also displayed in the title of each plot.

The degree of agreement between the vertical and horizontal quantities of each plot in Figures 5–11 indicates that the actual signals of Type-I words can be reproduced somewhat accurately using the optimized Hawkes processes. In particular, for the reasons listed below, we can conclude that the Hawkes processes have sufficient descriptive power to express original stochastic processes yielding Type-I words.

- The correlation coefficients of the number of occurrences (plot (a)) show strong positive correlations in most texts. This fact indicates that actual and simulated $X(t)$ share same statistical properties.
- The correlation coefficients of γ (plot (e)) and those of BIC (plot (f)) also show strong positive correlations. This indicates that ACFs of real $X(t)$ and those of simulated $X(t)$ are very similar in overall behaviors.
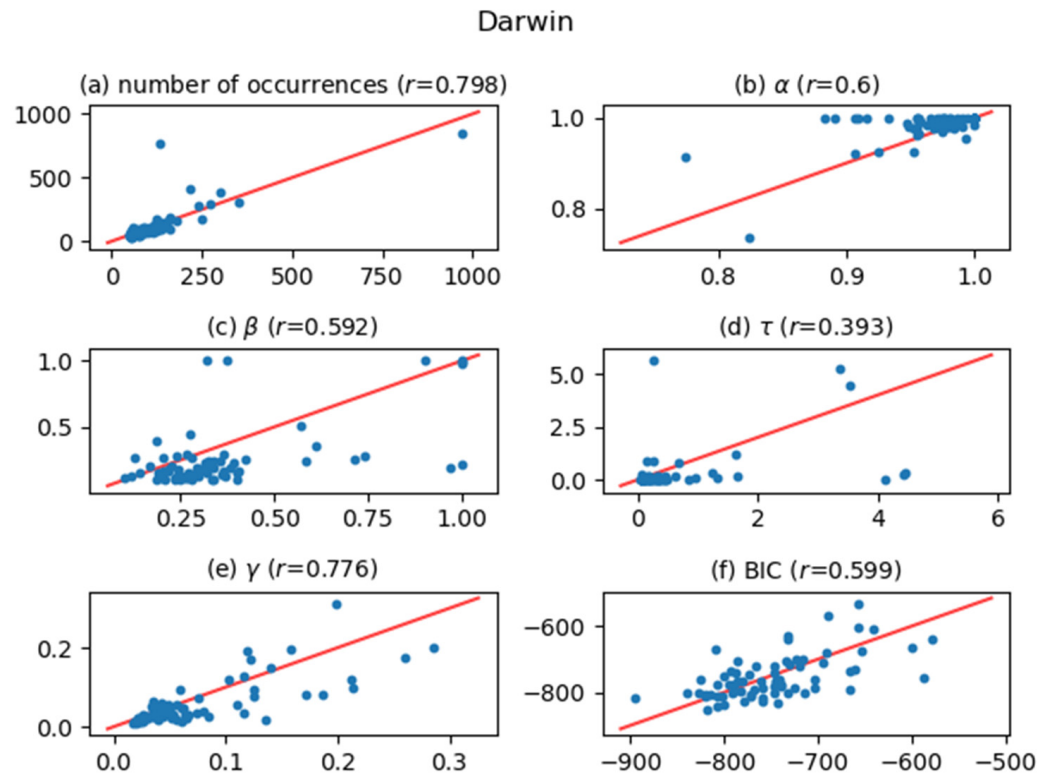
Darwin

Figure 5. Comparisons between values evaluated from real $X(t)$ (horizontal) and those evaluated from simulated $X(t)$ (vertical) across 6 characteristic quantities. All quantities were calculated for each of Type-I words that appears in Darwin text. The red line in each plot represents the relation $y = x$. The title of each plot includes a value of correlation coefficient, r, which indicates strong positive correlation when r is larger than about 0.6.
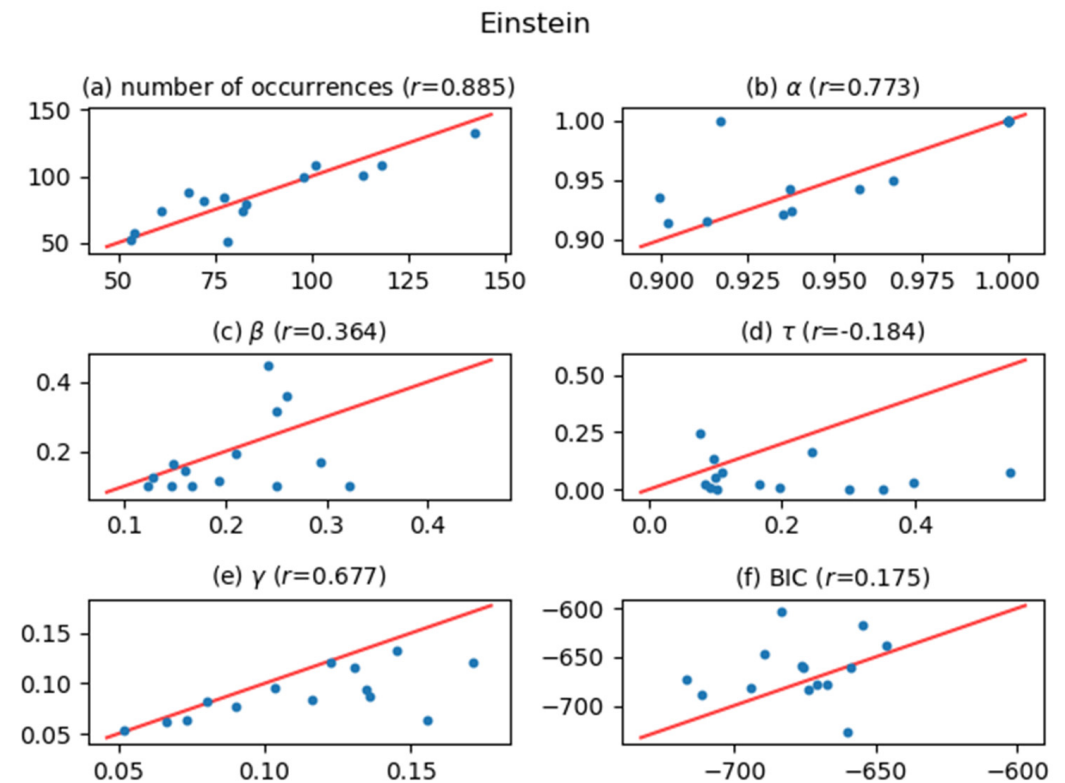
Einstein

Figure 6. The same meaning plots as in Figure 5 for the case of the Einstein text.
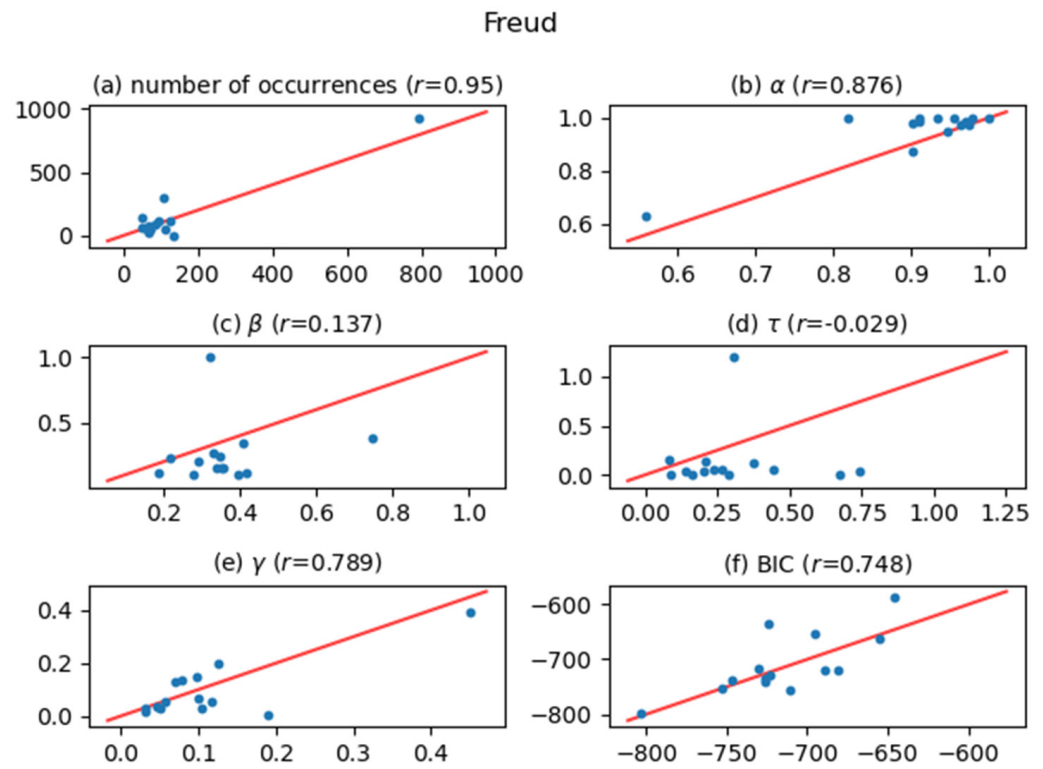
Freud

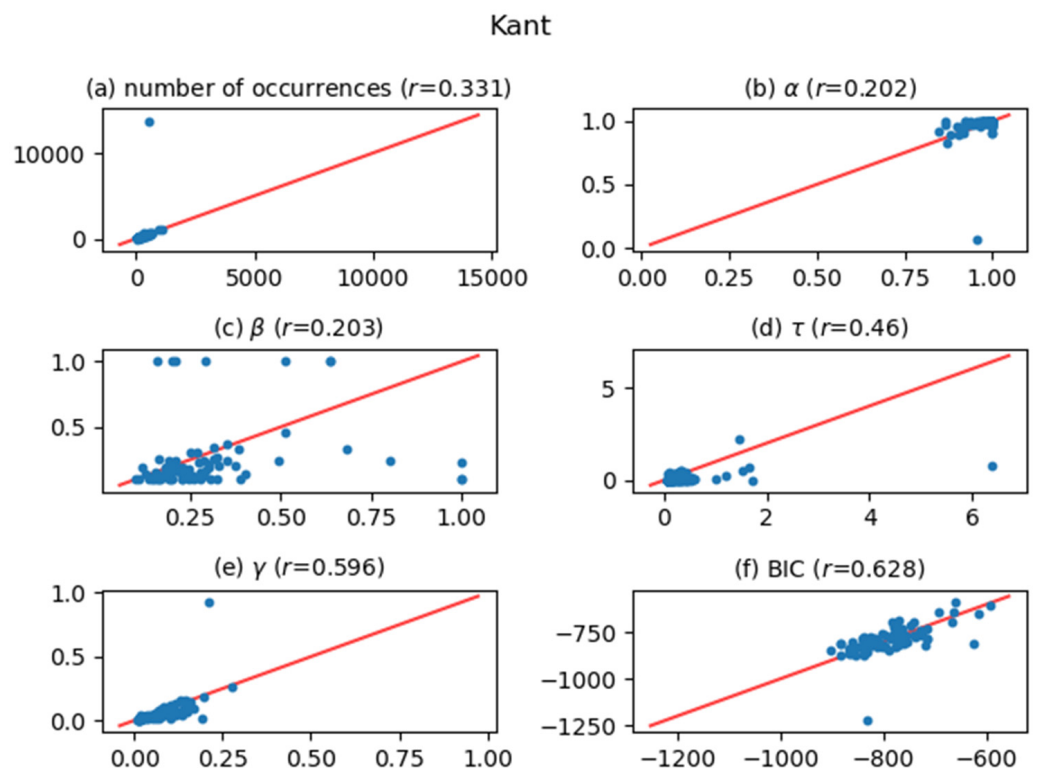**Figure 7.** The same meaning plots as in Figure 5 for the case of the Freud text.

Kant

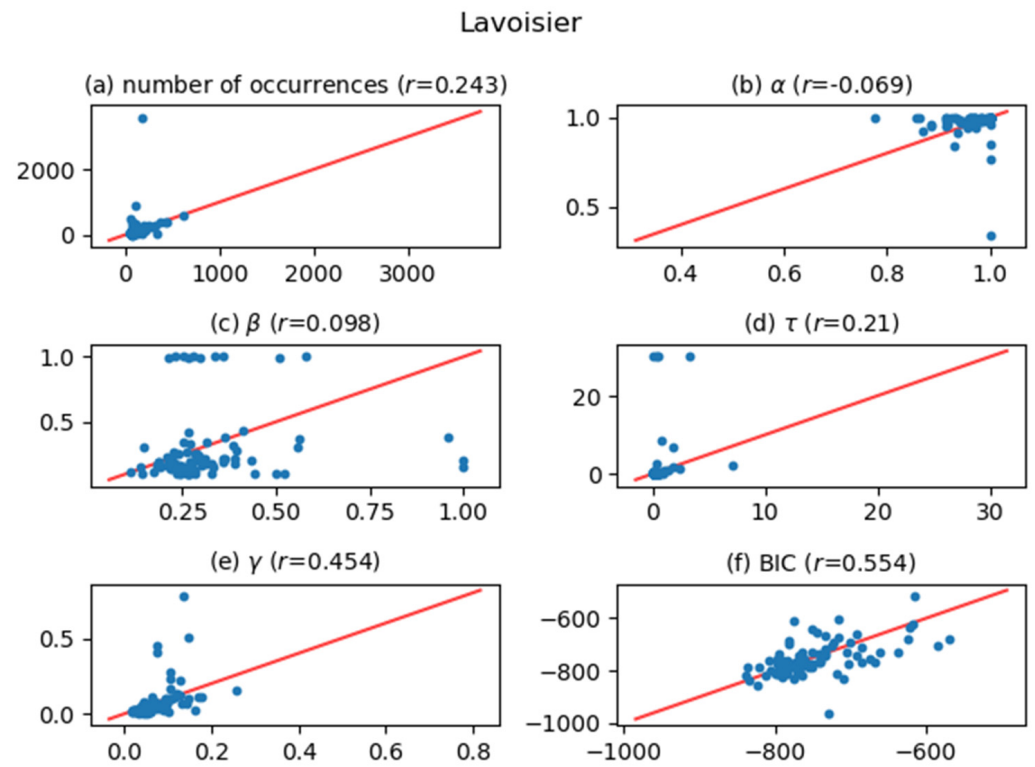**Figure 8.** The same meaning plots as in Figure 5 for the case of the Kant text.

Lavoisier

Figure 9. The same meaning plots as in Figure 5 for the case of the Lavoisier text.
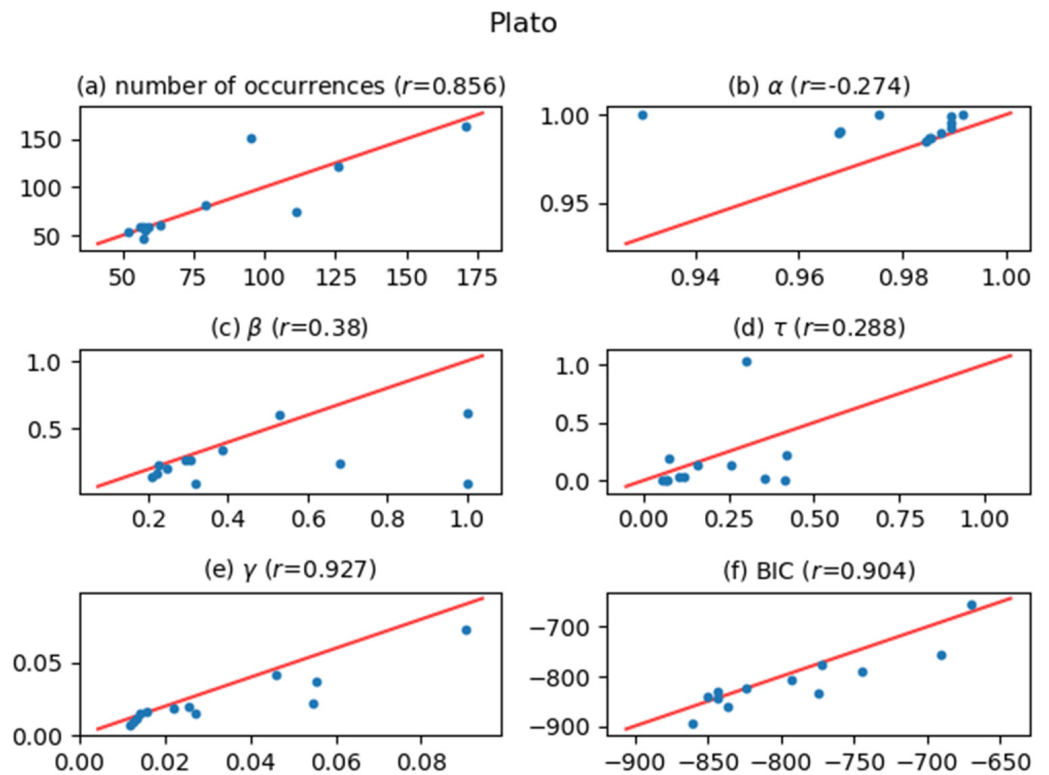
Plato

Figure 10. The same meaning plots as in Figure 5 for the case of the Plato text.
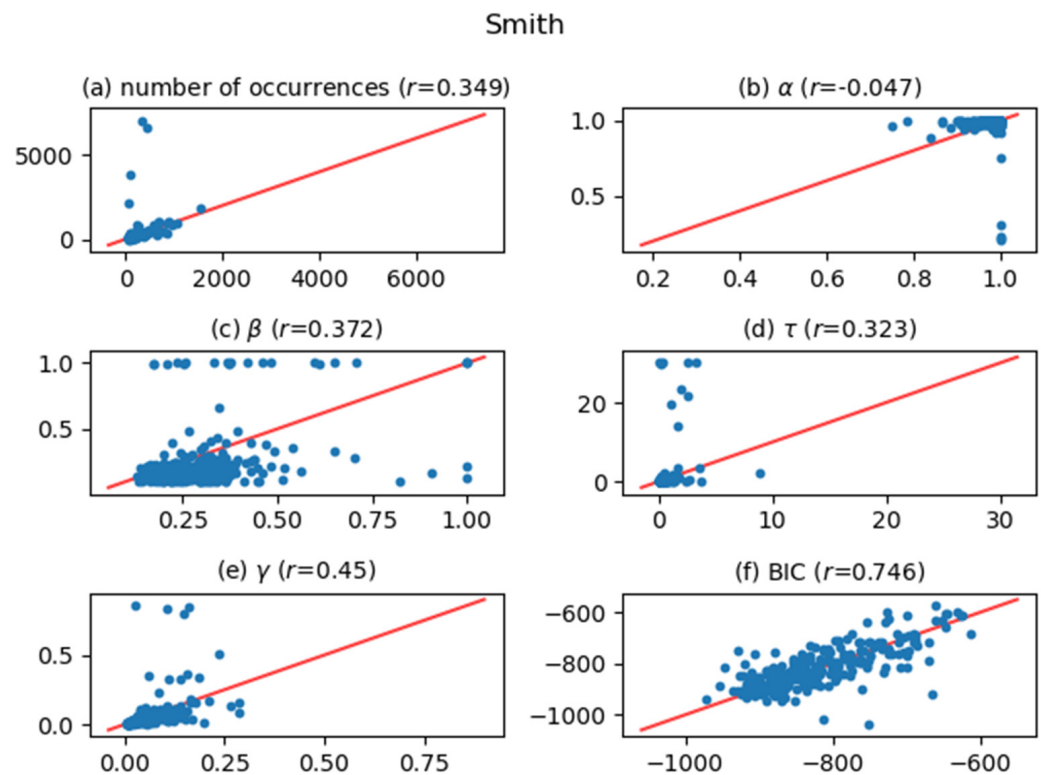
Smith

Figure 11. The same meaning plots as in Figure 5 for the case of the Smith text.

The most significant reason why the horizontal and vertical quantities do not perfectly match in these figures is that the simulated $X(t)$ generated by a Hawkes process with optimized parameters is only one sample of the time series data among infinite realizations generated by the optimized Hawkes process. If we prepare sufficient samples, i.e., large number of simulated signals $X(t)$ from the optimized Hawkes process for one Type-I word, and use averaged values for all simulated $X(t)$ to obtain one vertical value, then the vertical value tends to approach to the corresponding horizontal value of the considered word. This means that each of all runs generating $X(t)$ offers one vertical value, and averaging all vertical values justifies their convergence toward the real value. However, this requires high computational costs, and is out of the scope of this study.

4. Discussion

In our previous study [16], a model of stochastic process that yields Type-I words has been proposed. The characteristic of the model is that a waiting time distribution (WTD) of word occurrences has a fractal structure, which is naturally introduced from the hierarchical structure of written texts, i.e., volumes, chapters, sections, subsections, paragraphs, and sentences. More specifically, the fractal nature of WTDs has been clarified through the following procedures [16].

1. First, we construct an intensity function of word occurrence along a text, $P(t)$, which describes the occurrence probability of a considered word at time t. The construction is done in a recursive way in which the hierarchical structure of written texts is considered. Note that $P(t)$ represents word occurrence probability per unit time at time t, and thus it corresponds to $\lambda(t|H_t)$ of Hawkes processes.
2. A Monte Carlo simulation is performed by use of $P(t)$ to generate a word occurrence signal $X(t)$.
3. Waiting times t_w, which denote a time between two successive word occurrences, and their distribution $P(t_w)$ are calculated from the $X(t)$.
4. The resultant log-log plot of t_w vs. $P(t_w)$ shows a linear relationship, indicating that the WTD has a fractal structure.

Since the Hawkes process is defined by Equations (5)–(7), and since the structure of written texts is not taken into account among these equations, the above model does not seem to be related to the Hawkes process. However, the conditional intensity function $\lambda(t|H_t)$ of Hawkes processes employed in this study and $P(t)$ described in the first procedure above, are very similar to each other. To illustrate this fact, and to follow procedures 1 to 4 mentioned above by use of simulated signals from the Hawkes process, we present another simulation result in which the simulated period is set to be a longer value of $T = 10,000$ to make result clearer. Figure 12 shows the simulated signal $X(t)$ from a Hawkes process, cumulative count of word occurrences, and $\lambda(t|H_t)$ of the process over the period of $[0, 10,000]$. The conditional intensity function $\lambda(t|H_t)$ shown in Figure 12c is very similar to the previously reported $P(t)$ [14,16] in the point that it seems to be restricted to take several approximately discretized values.
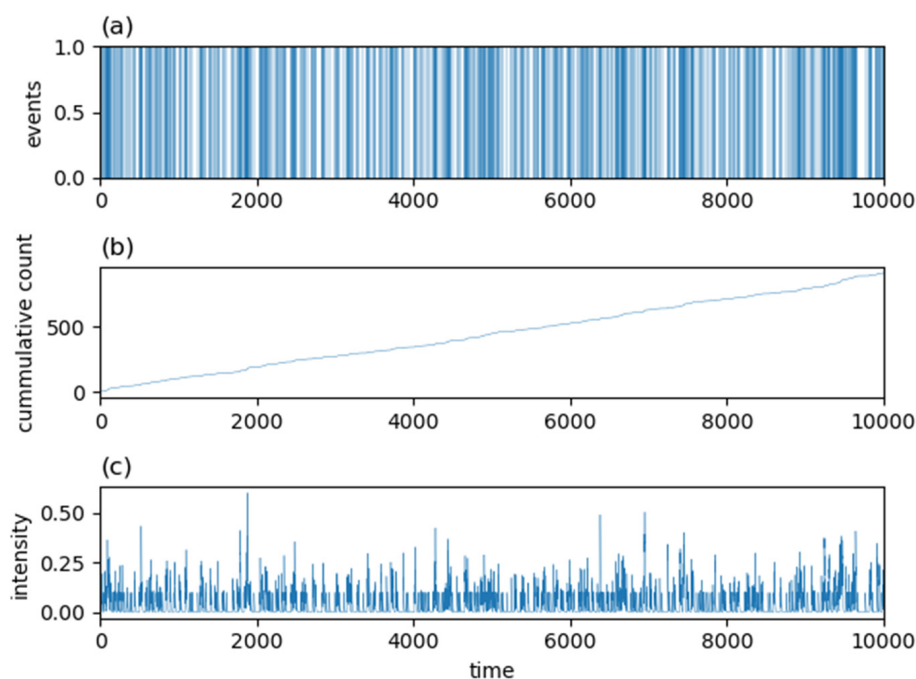

Figure 12. Simulated signal $X(t)$ (**a**), cumulative count of word occurrences (**b**), and conditional intensity function $\lambda(t|H_t)$ (**c**), of the Hawkes process defined by $\lambda(t|H_t) = 0.05 + \sum_{t_i < t} 0.1 \exp\{-0.25(t - t_i)\}$.

Figure 13 shows ACF calculated from simulated $X(t)$ shown in Figure 12a and its fitting results by use of Equation (3). Note that the fitting parameter τ takes a very small value of about $\tau \cong 0.03$ in Figure 13. In general, when τ in Equation (3) becomes smaller, then the resultant $\Phi_{KWW}(t)$ defined by Equation (3) approaches to $\Phi_{Poisson}(t)$ given by Equation (4), and at the limit of $\tau \to 0$, Equation (3) becomes Equation (4). Thus, the ACF shown in Figure 13 has intermediate properties between Figure 1c,f. In the same context, Figure 12a has intermediate properties between Figure 1a,d, and Figure 12b is in between Figure 1b,e. Therefore, the simulated signal shown in Figure 12a has a dynamic correlation with "intermediate" strength.

Figure 14 shows two examples of the relationship between waiting time t_w versus their distribution $P(t_w)$ in double logarithmic plots. The values of t_w and $P(t_w)$ used in the figure are obtained from two simulated $X(t)$ from Hawkes processes; one is the $X(t)$ simulated for the word "organ" (Figure 4a) and the other is $X(t)$ shown in Figure 12a. Note that a linear relationship seems to hold between $\log t_w$ and $\log P(t_w)$ for both cases, indicating that a fractal structure exists in the WTDs. It is confirmed that, therefore, the Hawkes processes bring some fractal structure in WTDs even in the case at which the Hawkes process does not have strong dynamic correlations but an "intermediate" correlation level. From the results shown in Figure 14, it seems almost certain that the Hawkes processes optimized to

describe the generations of Type-I words have some fractal structures within their WTDs. However, in order to clearly demonstrate this conclusion, large-scale simulations with longer periods T are needed.
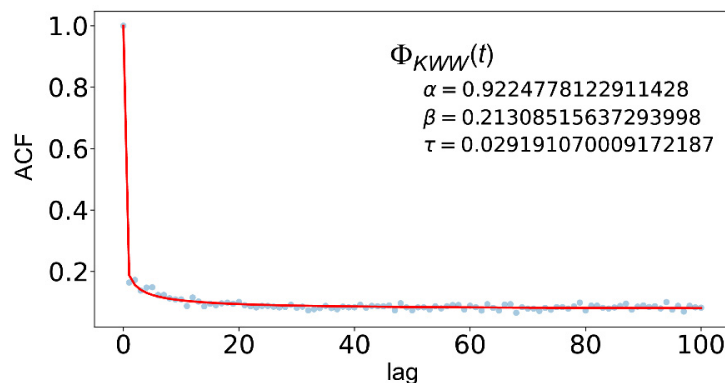

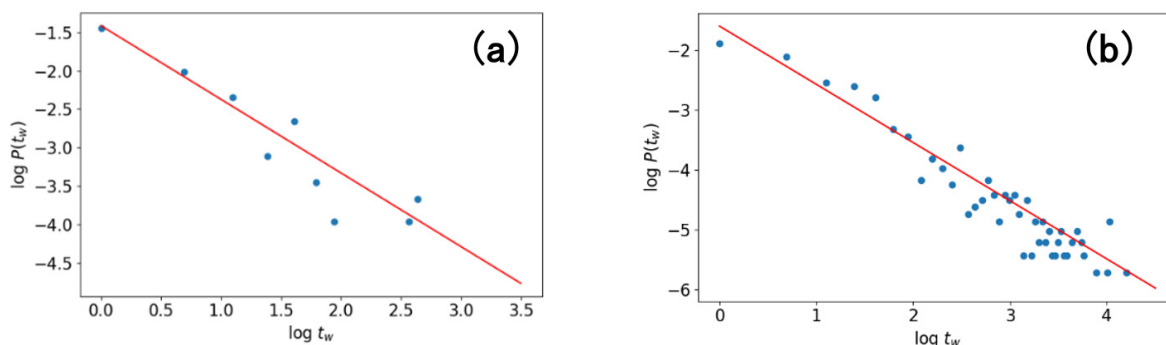**Figure 13.** ACF calculated from $X(t)$ shown in Figure 12a and its fitting result with Equation (3).

Figure 14. Double logarithmic plot of waiting time t_w and its distribution $P(t_w)$ which are calculated from simulated $X(t)$. The plot (**a**) uses $X(t)$ shown in Figure 4a and the plot (**b**) uses $X(t)$ shown in Figure 12a. Note that sample size is smaller in (**a**) than in (**b**) because the length of the used text is also shorter for (**a**) ($l = 4036$) than for (**b**) ($l = 10,000$). The red lines shown in the figure indicate the relations of $\log P(t_w) = -0.95661 \log t_w - 1.41603$ in plot (**a**) and $\log P(t_w) = -0.97394 \log t_w - 1.60345$ in plot (**b**) which were obtained by the method of weighted least square. To avoid noise affecting the fittings, we omit t_w having an occurrence count less than or equal to 2.

5. Conclusions

Occurrence patterns of Type-I words that appear in seven famous academic books were simulated by use of Hawkes processes. To seek an optimized Hawkes process for each of the Type-I words, we performed maximum likelihood estimation of the parameters of Hawkes process with the log-likelihood function given the actual occurrence pattern of a considered word observed in real written text. With that optimized Hawkes process, the occurrence signal of the word is generated and compared to the actual occurrence pattern of the word. The validity of the optimized Hawkes process was confirmed through comparisons between ACFs obtained from real word occurrence signals $X(t)$ and those obtained from the simulated $X(t)$. Degrees of agreement in various characteristic quantities of ACF show that Hawkes processes have a satisfactory ability to reproduce actual word occurrence signals in real written texts. Therefore, Hawkes processes can be utilized to express or simulate real word occurrence patterns of Type-I words. One of the advantages of using the Hawkes process in this manner is that it allows us to infer how a considered Type-I word works in a document. More specifically, we can determine the characters of the dynamic correlation of the word from the parameters vector, θ_{exp} or θ_{pow}, although the accuracy of parameter estimation needs to be further improved for this purpose.

We further found that simulated word occurrence signals from Hawkes processes have a property that waiting time t_w and its distribution $P(t_w)$ show linear relationship in double-logarithmic plots, indicating that the employed Hawkes processes have some fractal structure in their WTDs. The generalization of this finding through large scale simulations is an interesting theme for our future research.

Another possible direction of a future study is to establish a link between the stochastic model of Type-I words and some kind of diffusion model. Indeed, in our previous study [16], we utilized a Weierstrass random walk model [33–35] and modified it to derive the linear relationship between $\log t_w$ and $\log P(t_w)$ which was observed in WTDs of Type-I words. More generally, a methodology that directly relates point processes to diffusion processes has already been proposed [36,37]. This may allow us to apply various findings on fractional Brownian motion [38–40] to the analysis of the generation process of Type-I words.

This study confirms that the yielding processes of Type-I words in seven famous academic books can be described somewhat accurately by the Hawkes processes, which was established through the curve fittings in which ACFs of simulated signals $X(t)$ generated from Hawkes processes are well fitted by the KWW function. This result leads to a new research question: can the ACFs of signals generated from Hawkes processes always be described by the KWW functions? Solving this problem requires either running large-scale simulations of a new design or deductive arguments by developing the relevant point process theory. This issue is also an interesting research direction for the future.

Author Contributions: Conceptualization, H.O.; methodology, H.O.; software, H.O.; validation, H.O., Y.H., H.A. and M.K.; writing—original draft preparation, H.O.; writing—review and editing, Y.H., H.A. and M.K.; visualization, H.O.; supervision, H.O.; project administration, H.O. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pawlowski, A. Time-Series Analysis in Linguistics. Application of the Arima Method to Some Cases of Spoken Polish. *J. Quant. Linguist.* **1997**, *4*, 203–221. [CrossRef]
2. Pawlowski, A. Language in the Line vs. Language in the Mass: On the Efficiency of Sequential Modelling in the Analysis of Rhythm. *J. Quant. Linguist.* **1999**, *6*, 70–77. [CrossRef]
3. Pawlowski, A. Modelling of Sequential Structures in Text. In *Handbooks of Linguistics and Communication Science*; Walter de Gruyter: Berlin, Germany, 2005; pp. 738–750.
4. Pawlowski, A.; Eder, M. *Sequential Structures in "Dalimil's Chronicle"*; Mikros, G.K., Macutek, J., Eds.; Walter de Gruyter: Berlin, Germany, 2015; pp. 147–170. [CrossRef]
5. Altmann, E.G.; Pierrehumbert, J.B.; Motter, A.E. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE* **2009**, *4*, e7678. [CrossRef] [PubMed]
6. Tanaka-Ishii, K.; Bunde, A. Long-range memory in literary texts: On the universal clustering of the rare words. *PLoS ONE* **2016**, *11*, e0164658. [CrossRef]
7. Ogura, H.; Amano, H.; Kondo, M. Measuring Dynamic Correlations of Words in Written Texts with an Autocorrelation Function. *J. Data Anal. Inf. Process* **2019**, *7*, 46–73. [CrossRef]
8. Schenkel, A.; Zhang, J.; Zhang, Y. Long range correlation in human writings. *Fractals* **1993**, *1*, 47–57. [CrossRef]
9. Ebeling, W.; Pöschel, T. Entropy and long-range correlations in literary english. *Europhys. Lett.* **1994**, *26*, 241. [CrossRef]
10. Montemurro, M.A.; Pury, P.A. Long-range fractal correlations in literary corpora. *Fractals* **2002**, *10*, 451–461. [CrossRef]
11. Alvarez-Lacalle, E.; Dorow, B.; Eckmann, J.P.; Moses, E. Hierarchical structures induce long-range dynamic correlations in written texts. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 7956–7961. [CrossRef]
12. Altmann, E.G.; Cristadoro, G.; Esposti, M.D. On the origin of long-range correlations in texts. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 11582–11587. [CrossRef]

13. Chatzigeorgiou, M.; Constantoudis, V.; Diakonos, F.; Karamanos, K.; Papadimitriou, C.; Kalimeri, M.; Papageorgiou, H. Multifractal correlations in natural language written texts: Effects of language family and long word statistics. *Physica. A* **2017**, *469*, 173–182. [CrossRef]
14. Ogura, H.; Amano, H.; Kondo, M. Origin of Dynamic Correlations of Words in Written Texts. *J. Data Anal. Inf. Process.* **2019**, *7*, 228–249. [CrossRef]
15. Ogura, H.; Amano, H.; Kondo, M. Simulation of pseudo-text synthesis for generating words with long-range dynamic correlations. *SN Appl. Sci.* **2020**, *2*, 1387. [CrossRef]
16. Ogura, H.; Hanada, Y.; Amano, H.; Kondo, M. A stochastic model of word occurrences in hierarchically structured written texts. *SN Appl. Sci.* **2022**, *4*, 77. [CrossRef]
17. Hawkes, A.G. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika* **1971**, *58*, 83–90. [CrossRef]
18. Ogata, Y. Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **1988**, *83*, 9–27. [CrossRef]
19. Ogata, Y. Seismicity analysis through point-process modeling: A review. *Pure Appl. Geophys.* **1999**, *155*, 471–507. [CrossRef]
20. Zhuang, J.; Ogata, Y.; Vere-Jones, D. Stochastic declustering of space-time earthquake occurrences. *J. Amer. Statist. Soc.* **2002**, *97*, 369–380. [CrossRef]
21. Truccolo, W.; Eden, U.T.; Fellows, M.R.; Donoghue, J.P.; Brown, E.N. A Point Process Framework for Relating Neural Spiking Activity to SpikingHistory, Neural Ensemble, and Extrinsic Covariate Effects. *J. Neurophysiol.* **2005**, *93*, 1074–1089. [CrossRef]
22. Reynaud-Bouret, P.; Rivoirard, V.; Tuleau-Malot, C. Inference of functional connectivity in Neurosciences via Hawkes processes. In Proceedings of the 1st IEEE Global Conference on Signal and Information Processing, Austin, TX, USA, 3–5 December 2013. [CrossRef]
23. Gerhard, F.; Deger, M.; Truccolo, W. On the stability and dynamics of stochastic spiking neuron models: Nonlinear Hawkes process and point process GLMs. *PLoS Comput. Biol.* **2017**, *13*, e1005390. [CrossRef]
24. Bacry, E.; Mastromatteo, I.; Muzy, J. Hawkes Processes. *Market. Microstruct. Liq.* **2015**, *1*, 1550005. [CrossRef]
25. Rizoiu, M.A.; Lee, Y.; Mishra, S.; Xie, L. A tutorial on hawkes processes for events in social media. *arXiv* **2017**, arXiv:1708.06401. [CrossRef]
26. Palmowski, Z.; Puchalska, D. Modeling social media contagion using Hawkes processes. *J. Pol. Math. Soc.* **2021**, *49*, 65–83. [CrossRef]
27. Chiang, W.H.; Liu, X.; Mohler, G. Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *Int. J. Forecast.* **2022**, *38*, 505–520. [CrossRef] [PubMed]
28. Ogata, Y. On Lewis' simulation method for point processes. *IEEE Trans. Inf. Theory* **1981**, *27*, 23–31. [CrossRef]
29. Laub, P.J.; Taimre, T.; Pollett, P.K. Hawkes Processes. *arXiv* **2015**, arXiv:1507.02822. [CrossRef]
30. Omi, T.; Hirata, Y.; Aihara, K. Hawkes process model with a time-dependent background rate and its application to high-frequency financial data. *Phys. Rev. E* **2017**, *96*, 012303. [CrossRef]
31. Bonnans, J.F.; Gilbert, J.G.; Lemaréchal, C.; Sagastizábal, C.A. *Numerical Optimization—Theoretical and Practical Aspects*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2003; Chapter 4. [CrossRef]
32. Ogura, H.; Amano, H.; Kondo, M. Classifying Documents with Poisson Mixtures. *Trans. Mach. Learn. Arti. Intell.* **2014**, *2*, 48–76. [CrossRef]
33. Shlesinger, M.F. Fractal time and 1/f noise in complex systems. *Ann. N. Y. Acad. Sci.* **1987**, *504*, 214–228. [CrossRef]
34. Klafter, J.; Shlesinger, M.F.; Zumofen, G. Beyond Brownian motion. *Phys. Today* **1996**, *49*, 33–39. [CrossRef]
35. Wolfgang, P.; Baschnagel, J. *Stochastic Processes*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2013; Chapter 4.
36. Scafetta, N.; Hamilton, P.; Grigolini, P. The thermodynamics of social processes: The Teen Birth Phenomenon. *Fractals* **2001**, *9*, 193–208. [CrossRef]
37. Mega, M.S.; Allegrini, P.; Grigolini, P.; Latora, V.; Palatella, L.; Rapisarda, A.; Vinciguerra, S. Power-law time distribution of large earthquakes. *Phys. Rev. Lett.* **2003**, *90*, 188501. [CrossRef] [PubMed]
38. Samorodnitsky, G. Long range dependence. *Found. Trends Stoch. Syst.* **2006**, *1*, 163–257. [CrossRef]
39. Decreusefond, L.; Üstünel, A.S. Fractional Brownian motion: Theory $X(t)$ and applications. *ESAIM Proc.* **1998**, *5*, 75–86. [CrossRef]
40. Shevchenko, G. Fractional Brownian motion in a nutshell. *arXiv* **2014**, arXiv:1406.1956. [CrossRef]