International Journal of Innovative Studies in Sciences and Engineering Technology (IJISSET)

ISSN 2455-4863 (Online) www.ijisset.org Volume: 5 Issue: 9 | 2019

Autoregressive model of word occurrences in written texts

Hiroshi Ogura, Hiromi Amano, Masato Kondo

Department of Information Science, Faculty of Arts and Sciences, Showa University,
Fujiyoshida City, Yamanashi, Japan

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 occurrence 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. In this study, we find that an autoregressive (AR) model is useful for describing strong dynamic word correlations in the sense that it can reproduce characteristics of actual ACFs.*

Keywords: *Autocorrelation function, Stochastic process, Word occurrence, Autoregressive model*

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.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 is to find such a stochastic process for Type-I words.

2. AUTOCORRELATION FUNCTION FOR WORD OCCURRENCES

2.1 Definition of Autocorrelation Function

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 ACFs, we define the signal $A(t)$ representing word occurrence or non-occurrence as

$$A(t) = \begin{cases} 1 \text{ (when a given word occurs in the } t\text{th sentence)} \\ 0 \text{ (when a given word does not occur in the } t\text{th sentence).} \end{cases} \quad (1)$$

In the standard signal processing theory, the definition of ACF for a stationary system [2] and its normalized expression, $C(t)$ and $\Phi(t)$, are given by

$$C(t) = \lim_{T\to\infty} \frac{1}{T} \int_0^T A(\tau)A(\tau+t)d\tau, \quad (2)$$

$$\Phi(t) = \frac{C(t)}{C(0)} = \frac{\lim_{T\to\infty} \frac{1}{T} \int_0^T A(\tau)A(\tau+t)d\tau}{\lim_{T\to\infty} \frac{1}{T} \int_0^T A(\tau)A(\tau)d\tau}, \quad (3)$$

where $A(t)$ is a time varying signal which is interested in. As seen in the equations, the ACF measures the correlation of a signal $A(\tau)$ with itself shifted by some time delay t.

For our case in which signal $A(t)$ is restricted to a value of 0 or 1 as in eq. (1) and time t takes only positive integers, eqs. (2) and (3) are proved to be modified as [1]

$$C(t) = \frac{1}{N-t} \sum_{j=1}^{m} A(p_j+t), \quad (4)$$

$$\Phi(t) = \frac{C(t)}{C(0)} = \frac{N}{m(N-t)} \sum_{j=1}^{m} A(p_j+t), \quad (5)$$

where p_j is the ordinal sentence number at which a considered word occurs, N the number of sentences in a considered text. In eqs. (4) and (5), we have assumed that the total occurrences of the word in a text is m

times. Throughout this work, we use eq. (5) to calculate the normalized ACF of a considered word.

2.2 Examples of ACFs for Type-I and Type-II words

Figures 1 and 2 show examples of ACFs calculated from $A(t)$ 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*. As Figure 1 shows, ACFs for Type-I words are monotonically decreasing, indicating that dynamic correlations decrease as lag increases. They also show an apparent persistence of dynamic correlations with durations of several tens of sentences. In contrast, Figure 2 clearly shows that ACFs for Type-II words show no dynamic correlations, suggesting that Type-II words are generated from a memoryless stochastic process.

2.3 Curve Fitting using model Functions

To analyze the characteristic behaviors of ACFs described above, we have introduced two model functions to express ACFs and have attempted to fit these two parametrized functions to the calculated ACFs [1]. One of the model functions is $\Phi_{KWW}(t)$ which is used for ACFs of Type-I words showing dynamic correlations as in Fig. 1 and is defined by

$$\Phi_{KWW}(t) = \alpha \exp\left\{-\left(\frac{t}{\tau}\right)^{\beta}\right\} + (1-\alpha), \qquad (6)$$

where α, β and τ are fitting parameters which satisfy inequality conditions; $0 < \alpha \leq 1$, $0 < \beta \leq 1$ and $0 < \tau$. Equation (6) is a modification of the well-known Kohlrausch-Williams-Watts (KWW) function, which has been widely used to describe relaxation phenomena in complex systems [3, 4].

Another model function is $\Phi_{Poisson}(t)$ which is suitable for ACFs of Type-II words exhibiting no dynamical correlations as in Fig. 2. $\Phi_{Poisson}(t)$ is defined as a step-down function;

$$\Phi_{Poisson}(t) = \begin{cases} 1 \ (t = 0) \\ \gamma \ (t \geq 1) \end{cases}, \qquad (7)$$

where γ is a fitting parameter satisfying a condition $0 < \gamma < 1$.

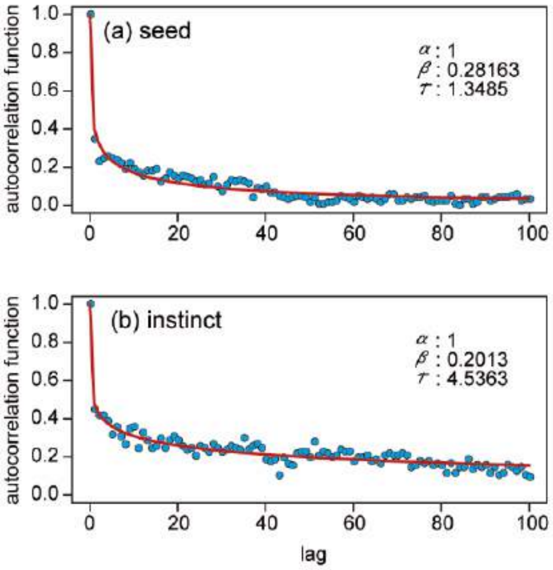

Fig -1: Examples of ACFs of Type-I words exhibiting strong dynamic correlations. They are ACFs of the words (a)"seed" and (b)"instinct", which are picked from the set of frequent words of Darwin text.

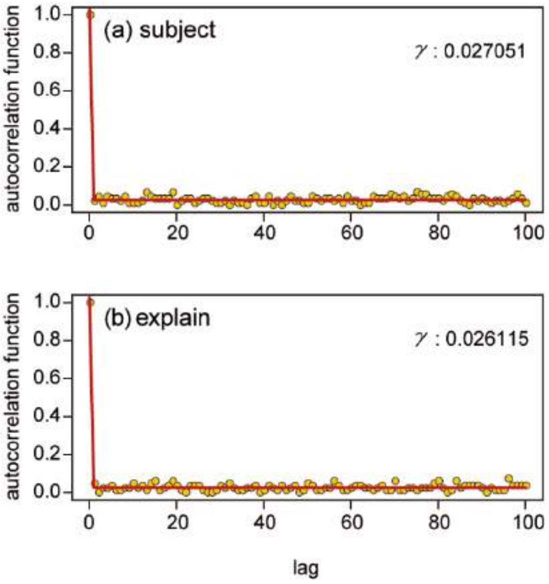

Fig -2: Examples of ACFs of Type-II words exhibiting no type of correlations. They are ACFs of the words (a)"subject" and (b) "explain", which are picked from the set of frequent words of Darwin text.

Figures 1 and 2 show fitting results using these two model functions as red lines, indicating the validity of using $\Phi_{KWW}(t)$ and $\Phi_{Poisson}(t)$ to model the two ACF types. Optimized values of fitting parameters are shown in the plot areas of these figures. Our previous study [1] found that, without exception, all frequent words appearing in 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.

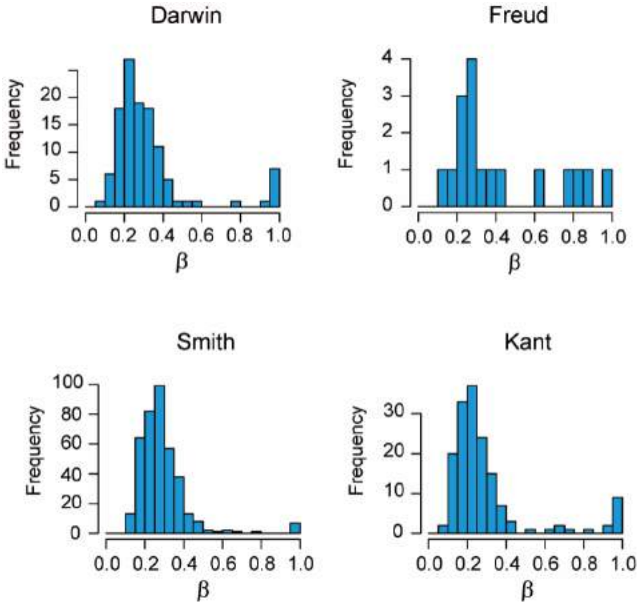

Fig -3: Histograms representing β distributions of Type-I words.

Figure 3 shows histgrams representing distributions of β values, one of the fitting parameters used in $\Phi_{KWW}(t)$. In the figure, results of fittings for all of the Type-I words appear in four well-known academic books (*On the Origin of Species* by Charles Darwin, *Dream Psychology* by Sigmund Freud, *An Inquiry into the Nature and Causes of the Wealth of Nations* by Adam Smith and *The Critique of Pure Reason* by Immanuel Kant) are employed. Some information of the used four books will be described in the Appendix. As seen in Fig. 3, typical distribution of β is characterized by a skewed unimodal distribution peaked around $\beta = 0.2$ and ranging from 0.1 to about 0.6. In the next section, we will try to construct a stochastic model which has an ability to reproduce ACFs with typical values of τ and β observed for real Type-I words.

3. MODEL OF WORD OCCURRENCES FOR TYPE-I WORDS WITH AUTOREGGRESSIVE PROCESS

In this section, we try to construct an autoregressive (AR) model to represent the random process of word occurrences exhibiting strong dynamic correlations. The reason for utilizing the AR model is that the model can describe a wide variety of random processes and the time range of correlation in the model can be arbitrarily adjusted.

First we consider AR(1) process defined by [5, 6]

$$B_t = \phi B_{t-1} + \varepsilon_t + c, \qquad (8)$$

where ϕ and c are constants, ε_t the white noise process with mean zero and variance σ_ε^2. The equation indicates that the time series B_t explicitly depends only on its firstbackshift B_{t-1}, which means that the dynamic correlation of the AR(1) is short range. Since B_t takes any real values in eq. (8), we have to convert B_t to another time series A_t which only takes binary values 0 and 1 in order to simulate occurrence/nonoccurrence of a considered word in each sentence of a document. We tentatively applied the simplest conversion

$$A_t = \begin{cases} 1 & (B_t \geq \theta) \\ 0 & (B_t < \theta) \end{cases}, \qquad (9)$$

where θ is a predefined constant. We will see below that an example of the ACF calculated from the time series B_t defined by eqs. (8) and (9) and will examine whether the obtained ACF is suitable to describe the dynamic correlations in real texts. The AR(1) process can be extended to the AR(p) process which involves the explicit dependence on its own previous values given by more than two backshift terms. The AR(p) process is defined by [5, 6]

$$B_t = \sum_{k=1}^{p} \phi_k B_{t-k} + \varepsilon_t + c. \qquad (10)$$

Obviously, the range of correlations in AR(p) process with $p \geq 2$ are wider than the AR(1) process and we can thus expect that the correlations over paragraphs can be expressed by choosing appropriate value of p in eq. (10). The conversion of eq. (9) was still adapted for the AR(p) case. Eq. (10) has an ability of describing a wide variety of stochastic processes by choosing p and adjusting ϕ_k. Unfortinately, we do not have any principles nor practical experiences to determine optimized p and ϕ_k in order to express dynamic correlations observed in real documents. Thus, the following equation of B_t with specified ϕ_k is not the only solution; it is one of the possible modelings by use of eq. (10) to describe the dynamic correlations of word occurrences observed over paragraphs. Our model equation obtained through trial and error is

$$B_t = \begin{cases} \alpha \left\{ \sum_{k=0}^{p-1} \frac{B_{t-(p-k)}}{(p-k)^\gamma} \right\} + \varepsilon_t & (t \geq p+1) \\ \alpha \left\{ \sum_{k=1}^{t-1} \frac{B_k}{(t-k)^\gamma} \right\} + \varepsilon_t & (t < p+1) \end{cases}. \qquad (11)$$

where $p > 1$, $\alpha > 0$ and $\gamma > 1$ are parameters of the model.

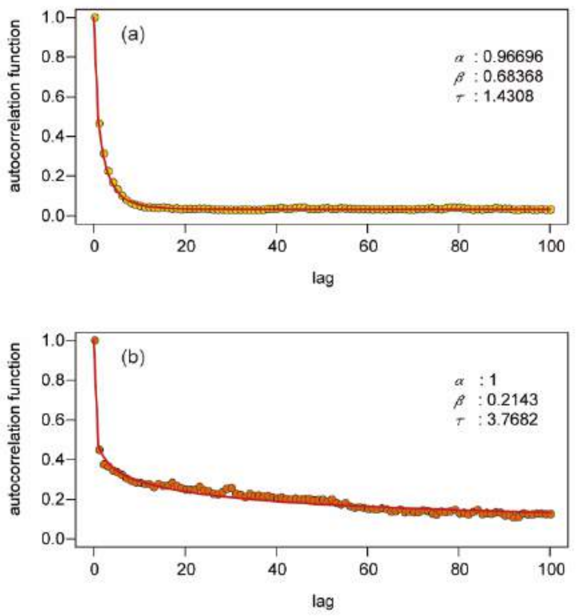

Fig -4: ACFs of the time series A_t. (a) B_t was generated from AR(1) process, i.e., eq. (8) with $\phi = 0.8$ and then converted to A_t using eq. (9) with$\theta = 3.0$. (b) B_t was generated from one of the AR(p) processes, eq. (11) with $p = 40, \alpha = 0.4, \gamma = 0.4$ and then converted to A_t using eq. (9) with $\theta = 2.0$. For both cases, A_t and B_t were generated from $t = 1$ to $t = 100{,}000$.

Figure 4 (a) shows an example of ACF calculated from time series A_t which were generated by use of eqs. (8) and (9). To obtain the ACF, we first generated B_t with settings $\phi = 0.8$ and $c = 0$ in eq. (8), converted B_t to A_t with setting $\theta = 3.0$ in eq. (9), calculated ACF of A_t by use of eq. (5) and then fitted it by using eq. (6). Figure 4(b) was obtained by similar procedures, but we used eq. (11) with $p = 40$, $\alpha = 0.4$ and $\gamma = 1.5$ insted of using eq. (8) to generate B_t. In the conversion from B_t to A_t, eq. (9) with $\theta = 2.0$ was also used. For $\varepsilon_t$in eqs. (8) and (11), we drew it from the standard normal distribution. From fig. 4(a) and (b), we can conclude that β decreases from about 0.7 to 0.2 when p in the AR(p) model increases from 1 to 40. Considering that the β distribution is heavily populated around $\beta = 0.2$ for ACFs of real texts as seen in Fig. 3, $\beta \cong 0.2$ obtained by the AR(p) process with $p = 40$ shown in Fig. 4(b) seems to be consistent with the ACFs of real Type-I words in their behaviors, indicating that the explicit dependence of the stochastic variable on more than several tens of its previous values is common in real Type-I words with low β values. On the other hand, the ACFs given by AR(1) process take βvalues larger than 0.5 for most cases as in Fig. 4(a).

The values of two other parameters, $\alpha = 1$ and $\tau \cong 3.8$ obtained in Fig. 4(b) are also consistent with those of real Type-I words, as confirmed in Fig. 1. The consistency of the ACF of fig. 4(b) with those of real Type-Iwords suggests us that the memory ranging over several tens of sentences is the key to understand the dynamic behaviors of important words in texts.

4. CONCLUSIONS

To confirm the importance of the long-range correlation, we proposed an autoregressive (AR) model of order p, AR(p), which uses linear combination of past p values of stochastic variable to produce a current value. Setting the order as $p = 40$ and adjusting the coefficients appropriately gave a ACF which is similar to the ACFs of important words in real written texts. This result indicates that the long-range memory with duration times from several to several tens of sentences is the key to reproduce the ACFs of the important words.

At present, the validity of using the AR(p) process to model word occurrences in texts is not strictly confirmed. More detailed study along this line, through which we try to identify the stochastic process suitable to describe word occurrences in real texts, is reserved for our future work.

ACKNOWLEDGEMENT

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).

REFERENCES

[1] H. Ogura, H. Amano and M. Kondo, "Measuring Dynamic Correlations of Words in Written Texts with an Autocorrelation Function". Journal of Data Analysis and Information Processing, 2019, Vol. **7**(2), pp. 46–73.

[2] P. Dunn, *Measurement, data analysis, and sensor fundamentals for engineering and science*, 2nd ed., CRC press, 2010.

[3] D. C. Johnston, "Stretched exponential relaxation arising from a continuous sum of exponential decays", Physical Review B, 2006, Vol. 74, 184430.

[4] A. V. Milovanov, A. Marcelo and K. Rypdal, "Stretched-exponential decay functions from a self-consistent model of dielectric relaxation", Physics Letters A, 2008, Vol. 372, pp. 451-461.

[5] M. Douglas, C. L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*, 2nd ed., John Wiley & Sons, 2015.

[6] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Application*, 3rd ed., Springer, 2011.

APPENDIX

In this study, we selected the English edition of four well-known academic books as samples of written texts and analyzed all Type-I words appearing therein to clarify the features of Type-I words. The four books were downloaded as text files from Project Gutenberg (https://www.gutenberg.org). Table A-1 lists the details of the four books used.

TABLE A-1. 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
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

Table A-2 presents some basic statistics of the four books, evaluated after the pre-processing procedures. Here, "frequent words" are those appearing in at least 50 sentences in the relevant text.

TABLE A-2. Summary of selected English texts.

Text	Vocabulary size	Word count	Sentence count	Number of frequent words	Number of Type-I words
Darwin	5316	58611	3991	212	109
Freud	4006	19533	1828	30	14
Smith	6817	140905	11318	537	382
Kant	5792	75285	5715	289	142