Advances in Language and Literary Studies
ISSN: 2203-4714
www.alls.aiac.org.au

A Study of the Lexical Complexity of Homogeneous Texts Using Stochastic Modeling and Analysis

Yanhui Zhang*

Faculty of Humanities and Social Sciences, University of Nottingham Ningbo China, 199 Taikang East Road, Ningbo, 315100, China

Corresponding author: Yanhui Zhang, E-mail: zhangyeonline@hotmail.com

ARTICLE INFO

Article history
Received: July 15, 2020
Accepted: September 12, 2020
Published: October 31, 2020
Volume: 11 Issue: 5
Advance access: October 2020

Conflicts of interest: None
Funding: None

Key words:
Lexical Richness,
Homogeneous Texts,
Dynamical Complexity,
Language Diffusion,
Stochastic Modeling

ABSTRACT

This paper takes a system dynamic approach to study homogeneous texts where the dynamics of the lexical richness of such texts over time are of the focal concern. It is hypothesized that the progress of the lexical complexity is driven by how far away this process is from the maximum level of complexity, while is subject to the fluctuations due to the dynamic nature of the system. It is shown that the lexical dynamics of homogeneous texts can be effectively modeled by a stochastic differential equation with proper upper bounds. The linguistic validity and the statistical goodness of the model are empirically tested with the texts of CGWR. Given the ubiquity of the diffusion phenomena in various settings of language and linguistic studies (e.g. language development), the findings of the current work should provide a useful methodological reference in comparison to classic approaches such as statistical regressions.

INTRODUCTION

Background and Research Objectives

Homogeneous texts, as hinted by the term, are a collection of texts, sharing so much in common in one or more language features, that are believed to be produced by the same linguistic production mechanism. To constitute a same production mechanism, texts can be written by a single unique author, when the writings of Thomas Hardy are referred to; can be restricted to a particular genre of writing, when the bankruptcy statement of businesses are under discussion; can be concerned with a specific syntactic characteristic when they are collected according to the usage of passive verbs or BÀ-structure (把字句) in Chinese language, for instance; or can be topically congruent when scientific papers on the motion of the solar system are referred to, and can be a combination of the aforesaid features and beyond.

Homogeneity is studied from multi-perspectives. Literatures such as Kilgarriff (2001) concerned mainly the quantitative measures of homogeneity and applications in copra comparison. Crossley and McNamara (2011) relied on the notion of inter-group homogeneity, and correspondingly cross-group heterogeneities, to study the development L2 English composition proficiencies. Argumentative texts produced by school children at a given grade were treated as homogeneous in Feike (1996) to analyze the role of syntactical component complexity and argumentative implicitness in achieving and developing discourse coherence. As an application in stylometric analysis, the idea of homogeneity was used by Gurney and Gurney (1998) for indicating that subsets of texts typically do not provide as effective gauge of vocabulary usage as the entire texts do. Homogeneity also served a critical notion in Yang and Luk (2003) for construction of cross-lingual thesaurus, where thematic homogeneity was premised for the purpose of text segmentation.

Homogeneity concerned in these literatures is mostly based on the text productions at the same time horizon, thus the profiled linguistic features of such homogeneous texts are supposedly uniform from one to another. On the other hand, homogeneous texts can be produced at different times, examples of which can be the texts of the speech of an assembly archived from the early years of a speaker to the elderly years of the same speaker, or the tax codes and publications published by government in different years. It is natural query, and thus the objective of the current work, to study whether any systematic properties, in one or more linguistic

Published by Australian International Academic Centre PTY.LTD.

http://dx.doi.org/10.7575/aiac.alls.v.11n.5-p.1

aspects, can be generalized from the series of homogeneous texts produced at different times. And accordingly, what are the mechanisms, if any, leading to such properties.

Research Problem Formulation

One of the most prominent quantifiable linguistic properties in differentiation of texts at corpus level is lexical richness, which refers to the level of verbal variation and sophistication represented by a given text. The measure of lexical richness can take many forms, including TTR, D, and entropy. TTR, defined as the ratio of the number of types divided by the number of tokens of a text, is a classic and widely known measure of lexical richness, the seminal introduction of which goes back to Herdan (1960). D, the arithmetic form of which is defined in relation to TTR as

$$TTR = \frac{D}{N}\left(\sqrt{1+2\frac{N}{D}}-1\right),$$

was more recently proposed by Malvern et al. (2004), intended to overcome the length effect of TTR. Entropy, the origin of which arose from thermodynamics (see Bailyn, 1994, e.g.), is defined by the equation $E= -\sum_{i=1}^{T} P_i \log P_i$, where P_i is the probability for the ith word to appear in an interested text. One of the apparent advantages of using entropy to quantify lexical richness is its universality in many other disciplines such as biology or information science. Indeed, ontologically speaking, the laws governing the evolution of lexical richness in language domain are fundamentally analogous to the evolution laws in other fields. But researchers have shown that none of these measures is perfect. For instance, Jarvis (2013) discussed the importance of introducing the rareness dimension in lexical richness to better reflect its linguistic intuition. Johansson (2008) showed arguments and empirical observations why D is not optimal as a lexical richness measure.

In sense, all these measures are functionally related since all their underlying constructs are based on the frequency distribution of words. As demonstrated in the above, TTR and D are actually algebraically related to each other by an identity equation. This paper will show that lexical richness, expressed in terms of all the selected forms, of homogeneous texts follows distinctively similar evolving patterns. For this purpose, the quantitative attributes of lexical richness of homogeneous texts, recorded as time-dependent series, are formulated and analyzed with stochastic differential equations. Instead of being treated as discrete non-related texts created at different times, the whole corpus is thought of as a continuous production of a single, unique, and integrated linguistic mechanism. From the dynamic complexity perspective, the corpus is viewed as a linguistic organism which continuously evolves itself to fit into the changing sociocultural environment (Zhang, 2015). Such organism is a dynamic system in itself, undergoing continuous information exchange, feedback, adaptation, and self-organization. The quantitative law discovered by this paper explains how the level of lexical richness of such a complex system evolves from low to high over time.

The law is fundamentally described by a stochastic differential equation, where the unknown variable is the level of lexical richness of an interested text at a prescribed time horizon. The equation is expressed in terms of how the level of lexical richness will change in relation to a small change of time. This change is shown to be positively proportional to the current level of lexical richness and the distance between the current level of lexical richness and the maximum level of lexical richness for the text with the given size and sociolinguistic constraints. The solution to the proposed model, when the random part is removed, verifies an empirical exponential model, recently reported in Zhang (2015).

To contrast, the currently proposed model is validated with the same corpus data of the CGWR as explained and modeled by Zhang (2015). Four types of lexical richness measures, namely, TTR, root TTR, D, and entropy are tested and compared. The estimation procedures are demonstrated as efficient and stable, and all the estimated parameters reported in the current paper are statistically significant. The next Section 2 describes the data and the corpus used for the study as well as the stochastic differential equation methodology. Section 3 provides the numerical results of implementation of the proposed approach, together with statistical analysis and model testing. Section 4 presents further discussions to the current results with comments of future directions.

METHODOLOGY

The homogenous texts are not uncorrelated and index-invariant. This paper argues, from a dynamic system standpoint, that they are sequentially related, continuously interconnected, and asymmetric in time. Instead of being static, the corpus, in its own right, undergoes inception, emergence, development and maturing, the course of which may involve nonlinear changes, adoption of new entries, removal of superannuated elements and syntactic structures, and possibly other interruptions. To model this dynamic process with random noises, the following quantitative framework is proposed.

Let $P(t)$ be the degree of lexical richness of the homogeneous text of CGWR at time t, where t=1 be the year of 1954, t=2 the year of 1955, and so on. Let dP be the change of the level of lexical richness within a time interval of dt. The evolution of the lexical dynamics is expressed as

$$dP(t) = \alpha\left(L - P(t)\right)dt + \sigma\left(L - P(t)\right)dB(t) \qquad (1)$$

where L is the asymptotic limit of $P(t)$. In another word, L is the upper bound of the level of lexical complexity for the given size of the CGWR text, which in turn may be constrained by the linguistic functions that the CGWR is set to perform. From the definition of L, $P(t) < L$ for all $t \geq 0$. L is assumed as constant in the current study; however, the scenarios where L is time-dependent are possible and are discussed as a potential future direction in the concluding remarks. Here $B(t)$ is the standard Brownian motion,

reflecting and modeling the randomness resulted from the dynamic nature of the process. Quantitatively, $B(t)$ has the property $dB(t) \sim N(0,t)$, i.e, $dB(t)$ follows a normal distribution with mean 0 and variance t or standard deviation $\sqrt{t}$. For an introduction of Brownian motion with application in the field of social science and humanity research, one may refer to Gardiner (2009).

It is seen from the equation (1) that the process is composed of two forces. First, consider the deterministic case where the volatility factor sigma is assumed 0. Then the increment of $P(t)$ is positive if alpha is positive and if the process of $P(t)$ starting from somewhere between zero and L. But due to the constraints framed by the linguistic syntactic style, function, prosody, or other sociocultural metric, the growth of $P(t)$ will be eventually flattened, unless there are emerging factors that may lead to level change which is not focused by the current study. Thus the farther away the $P(t)$ is from L, the asymptotic upper bound of the level of lexical complexity, the higher rate of increase in $P(t)$. The same scaling rule can be applied to the diffusion term of the equation (1). This is despite the fact that $dB(t)$, by definition, has the same expected value for different time t,

Now let $Q(t) = L - P(t)$, then equation (1) is equivalently written as

$$dQ(t) = -\alpha Q(t)\,dt + \sigma Q(t)\,dB(t) \tag{2}$$

In term stochastic modeling, this differential equation is solvable with

$$Q(t) = Q(0)e^{(-\alpha-\frac{\sigma^2}{2})t + \sigma B(t)} \tag{3}$$

or equivalently,

$$\ln\frac{Q(t)}{Q(0)} = \left(-\alpha - \frac{\sigma^2}{2}\right)t + \sigma B(t) \tag{4}$$

Now the increment of $Q(t)$ in time interval of $(t, t+\Delta t)$ takes the form of

$$\ln Q(t+\Delta t) - \ln Q(t) = \left(-\alpha - \frac{\sigma^2}{2}\right)\Delta t + \sigma B(\Delta t)$$

Thus to simulate the process of $Q(t)$ for $t_0 < t_1 < \cdots < t_n$, one can appeal to the following iterations:

$$P(t_{i+1}) = P(t_i)e^{\left(-\alpha-\frac{\sigma^2}{2}\right)\Delta t + \sigma\sqrt{\Delta t}Z_{i+1}}$$

where $Z_1, Z_2, \cdots Z_n$ are independently drawn from the identical standard normal distributions. Here it is assumed that the intervals between t_i and t_{i+1} for $i = 0, \cdots n$ are uniformly spaced. If not, simply replace Δt with $(t_{i+1} - t_i)$. Now let

$$x_i = l_n \frac{Q_i}{Q_{i-1}}, i = 1, 2, \cdots\cdots n,$$

then $$x_i \sim N\left[\left(-\alpha - \frac{\sigma^2}{2}\right)\Delta t, \sigma^2\Delta t\right]$$

It is well known that the mean and variance of a sample data whose distribution is described by a normal distribution can be estimated through the method of maximum likelihood estimation (MLE). The MLE procedure applied to the supposedly normally distributed sample of x_is gives the following estimations for the parameters in model (1):

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \frac{\left(\sum_{i=1}^{n} x_i\right)}{n})^2}{n.\Delta t}}$$

$$\hat{\alpha} = -\frac{1}{n.\Delta t}\left\{\sum_{i=1}^{n} x_i + \frac{1}{2}\sum_{i=1}^{n}\left(x_i - \frac{\left(\sum_{i=1}^{n} x_i\right)}{n}\right)^2\right\}.$$

Now for any given $L > \max(P_i)$, the corresponding values of x_i's, $i = 1, 2, \cdots n,$ are fixed. Then the parameter estimates $\hat{\alpha}$ and $\hat{\sigma}$ can be determined by the above two formulas. There are several statistics concerning the testing the goodness of fit, including Chi square test, KS test, and Shapiro-Wilk test. Since Chi square test can have biased conclusion for small sample data, the current study uses jointly the KS and Shapiro-Wilk tests to decide on the best model fitting. To remark, the maximum value of L is easy to comprehend since it is well known that the lexical richness of any text in any language is bounded. For instance, the maximum character based entropy of Chinese language is about 9.7, according to Yuan et al. (1987). And the letter based entropy of English language is capped at 4.03, according to Feng (1991). For reference of entropy calculation and entropies of selected languages, one can refer to Levitin and Reingold (1994). And because of the relationships between TTR, D, and P, it is straightforward to find the corresponding domain for them.

Empirical Results and Analysis

The data used for the current study are the CGWR corpus archived from the year 1954, where the first CGWR was launched, to 2014, excluding those years where the report was not delivered, namely, 1959-1961, 1966-1976. The 45 CGWR texts in total constitute the corpus for the current study. The average number of the types of each text is about 1166 in terms of the number of unique Chinese characters used, including punctuations, and the average number of tokens for these texts is 22041. The standard deviations for the types and tokens of the selected texts are about 132 and 7994 respectively. Zhang (2015) provides a more detailed description of the corpus as well as a structured equation approach for modeling the entropic information of the corpus. The following Table 1 outlines the key statistics of one sample set from the CGWR corpus.

Table 1. Descriptive statistics of the CGWR 1954 text, adapted form Zhang (2015)

Year	Types	Tokens	TTR	Entropy	Maximum entropy
1954	1205	23168	0.052	5.8601	10.0505

Table 2. MLE based parameter estimations and normality

	TTR	Root TTR	Entropy	D
L	0.2931	10.0834	6.0238	51.9544
alpha	-9.4904e-04	-0.0909	-0.3712	-0.1168
sigma	0.0680	0.4364	0.8915	0.4936
H_ Lilliefors	0	0	0	0
KS Lilliefors	0.1052	0.1147	0.0817	0.1144
CV Lilliefors	0.1338	0.1338	0.1324	0.1338
P_ SW	0.2546	0.0152	0.8558	0.0075
SW Sig	0.289	0.046	0.856	0.023

The following Table 2 reports the estimated parameter values for the diffusion processes of TTR, root TTR, entropy, and D, respectively, when they are described by the proposed stochastic model with an assumed upper bound. They are outputted from the iterative algorithms, implementing the MLE procedure associated with the model as discussed in the section of Methodology using the TTR, root TTR, entropy, and D data of the CGWR texts. The following Table 2 also presents the goodness of fit tests associated to the respective sets of parameter estimations. As it is hypothesized in the current paper that the lexical richness data follows a normal distribution model after transformations, it is important to examine whether the observed values, after logarithm transformations, are truly normal. There are a couple of popular statistical tests serving for this purpose, including Chi square test, KS test, Lilliefors test, and Shapiro-Wilk test (SW). Lilliefors test is more preferable for the current study since Chi-square test can be biased for small samples and KS test alone is not suitable for the normal distributions with unspecified means and variances. Overall, KS-Lilliefors test is chosen as the benchmark test, but at the same time, the results of SW test are also provided for comparison.

DISCUSSION AND CONCLUDING REMARKS

The current works attempts to identify and model the diffusion phenomena empirically observed in the homogeneous texts of CGWRs. Lexical complexity is of pivotal interest to language teachers, researchers and practitioners. For language teaching, lexical richness models may suffice a better understanding and assessment of learners' vocabulary development at different stages of learning so as to facilitate the designing of an optimal learning ladder (Malvern et al., 2004; Crossley et al., 2011). For sociolinguists, lexical richness models often constitute critically important linguistic references for interested sociocultural query (Yang & Luk, 2003; Zhang, 2015). The fulfillment in all such aspects entails a sound and preferably concise description the lexical complexity as observed in the homogeneously constructed corpus. A good portion of the existing works related to lexical complexity analysis either involves only comparison of different lexical richness measures or falls short in terms of quantitative rigor and model efficiency (Lu, 2013; Diekmann & Mitte, 2014).

The stochastic modeling method proposed by the current study demonstrates the desired the clarity and robustness for the task, where the model implementation as well as the pertaining parameter calibrations are implemented with computer assisted routines. Key statistical properties such as model significance and normality are well maintained as tested with the homogeneous texts of CGWR. One notable novelty of the current works is that it decomposes the process into a drift part and a random part, where the drift part is determined and measured by how far away the process is approaching the upper bound of the lexical diversity of the corpus, while the random part models the amount of uncertainty ensued from internal and external noises of the process. A decomposition of the upper bound by a linear translation plus a logarithm operation results in a transformed series shown to fit with a lognormal diffusion model.

Within the framework of the lognormal model proposed in the current paper, one interesting and possibly intriguing direction deserving future study is to compare the optimal upper bound L reported in the current paper and the theoretical maximum levels of lexical richness of Chinese texts at given sizes, measured in terms of the four metrics used in the current work. One worthy attempt in this regard can be found in Shannon (1951), where the entropies of English are further analyzed in subclasses of zero-order entropy, first order entropy, and so on, up to the infinite order entropy, and are calibrated using simulation approach. However, the search for the maximum level of lexical richness in terms of entropy or other measures of a text for any given length is rather a challenging problem and beyond the scope of the current study. As far as our knowledge goes, there do not exist a comprehensive result on the bounds of lexical complexity of Chinese language at given sizes and given genres of texts.

While the appropriateness of the approach has been confirmed by the extensive statistical tests in terms of the stability of parameter estimation, robustness of the algorithm, the goodness of fit, and normality check, it is certainly possible and worthy of future study to find more fitting models with similar diffusion properties. Choices of such model improvement, again within the domain of diffusion stochastic equations, include to allow for a broader class of functions and combinations of constants, time parameters, and $P(t)$ at proper places in the equation (1). Such possibilities include, for example, to add a power to the drift coefficient or add a power to the diffusion coefficient. Or one may try to let the parameters in the equation (2) be functions of time instead of constants as they are in the current form. All these explorations can be rewarding in terms of goodness of fit and, in the meantime, posing new challenges in terms of parameter estimation, robustness, and other issues pertaining to general concerns for model selection.

REFERENCES

Bailyn, M. (1994). A survey of thermodynamics. American Institute of Physics, New York.

Crossley, S. A., Salsbury, T., & McNamara D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2), 243-263.

Crossley,S.A.,& McNamara, D.S. (2011). Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing*. doi:10.1016/j.jslw.2011.05.007

Diekmann, A., & Mitte, P. (2014). Stochastic Modelling of Social Processes, Elsevier.

Feilke H. (1996). From syntactical to textual strategies of argumentation: Syntactical development in written argumentative texts by students aged 10 to 22. *Argumentation*, 10, 197-212.

Feng, Z. (1991). Shuxue Yu Yuyan (Mathematics and Language). Hunan Education Press.

Gardiner, C. (2009). Stochastic Methods: A Handbook for the Natural and Social Sciences, Springer.

Gurney, P. J., & Gurney, L. W. (1998). Subsets and homogeneity: Authorship attribution in the Scriptories Historiae Augustae. *Literacy & Linguistic Computing* , 13 (3), 133-140.

Herdan, G. (1960). Quantitative linguistics. Butterworth, London.

Jarvis, S. (2013). Capturing diversity in lexical diversity. *Language Learning,* 63, 87-106.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective, Lund University, Dept. of Linguistics and Phonetics, Working Papers 53 (2008), 61

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1), 1–37.

Koizumi, R., & In'nami Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40, 522-532.

Lamprier, S., Amghar, T., Levrat, B., and Sanbion, F. (2007), SegGen: A genetic algorithm for linear text segmentation. *Proceeding of the 20th International Joint Conferneces on Artificial Intelligence*. AAAI Press, Menlo Park, CA. 1647-1652.

Levitin, L. B., & Reingold, Z. (1994). Entropy of natural languages: Theory and experiment. *Chaos, Solitons, and Fractals*, 4 (5), 709-743.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal,* 96(2), 190–208.

MacWhinney, B. (2007). The TalkBank Project. In J. C. Beal, K. P. Corrigan,&H. L. Moisl (Eds.), Creating and digitizing language corpora: Synchronic databases (Vol. 1, pp. 163–180). Houndmills, UK: Palgrave-Macmillan.

Malvern D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19. 85-104.

Malvern, D., Richards, B., Chipere, N., &Duran, P. (2004). Lexical diversity and language development: Quantification and assessment. Palgrave Macmillan.

McCarthy, P. and Jarvis S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*. 42 (2), 381-392.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. Language Testing, 24(4), 459–488.

Wan, F. W. M. (2019). Stochastic Models in the Life Sciences and Their Methods of Analysis. World Scientific Publishing Co.

Yang, C. C., & Luk, J. (2003). Automatic Generation of English/Chinese Thesaurus Based on Corpus in Laws. *Journal of the American Society for Information Science and Technology*. 54 (7), 671-682.

Yuan, L., Wang, D., & Zhang, S. (1987). The probability distribution and entropy and redundancy in printed Chinese. In: Proceedings of International Conference on Chinese Information Processing, 505–509.

Zhang, Y. (2015). Entropic evolution of lexical richness of homogeneous texts over time: A dynamic complexity perspective. *Journal of Language Modeling*, 3 (2), DOI: http://dx.doi.org/10.15398/jlm.v3i2.111