Entropic evolution of lexical richness of homogeneous texts over time: A dynamic complexity perspective

Yanhui Zhang
The Chinese University of Hong Kong

ABSTRACT

This work concerns the evolving pattern of the lexical richness of the corpus text of China Government Work Report measured by entropy, based on a fundamental assumption that these texts are linguistically homogeneous. The corpus is interpreted and studied as a dynamic system, the components of which maintain spontaneous variations, adjustment, self-organizations, and adaptations to fit into the semantic, discourse, and sociolinguistic functions that the text is set to perform. Both the macroscopic structural trend and the microscopic fluctuations of the time series of the interested entropic process are meticulously investigated from the dynamic complexity theoretical perspective. Rigorous nonlinear regression analysis is provided throughout the study for empirical justifications to the theoretical postulations. An overall concave model with modulated fluctuations incorporated is proposed and statistically tested to represent the key quantitative findings. Possible extensions of the current study are discussed.

Keywords: dynamic complexity, lexical richness, entropy, homogenous texts, language modeling

1 INTRODUCTION

Corpus linguists and experts in related fields have shown increasing interest in homogeneous texts, largely because homogenization is often an effective and statistically trustful way to filter out the unnecessary or, even worse, the distorted information from the raw meta-corpus data, thus helping to uncover the principal linguistic variables as well as the governing laws that a researcher is keen to

find. Study surrounding homogeneous texts can be undertaken from many perspectives, including homogeneity measurement, corpus selection, and applications in language acquisition and sociolinguistic analysis. For instance, the cross-corpora studies of Kilgarriff (2001), Kilgarriff and Grefenstette (2003), and Denoual (2005) relied heavily upon the notion of homogeneity. Kornai *et al.* (2006) focused on texts' homogeneity characterized by their stylistic features, particularly those discernable through author tags. Crossley and McNamara (2011) used word-based indices such as hypernymy and stem overlap to test the intergroup homogeneities among L2 English learners and cross-group heterogeneities between L2 and L1 writers so as to facilitate the understanding of the development of L2 writing. Sahlgren and Karlgren (2005) confined homogeneity to the extent of topical dispersion with empirical applications. The primary interest of the current study is to understand how the complexity of a given set of homogeneous texts progresses over time. For this purpose, the corpus is treated as an interacting, adaptive, and constantly evolving system, the evolution of which is regulated by the internal linguistic laws as well as external sociocultural conditions at large.

Lexical richness, a primary indicator of verbal variation and sophistication and hence the degree of complexity profiled by an interested text, is a particularly useful tool for quantitative and computational linguistics, the application of which can be found in Smith and Kelly (2002) for author attribution and in Johansson (2008) for language proficiency assessment. Existing literature on lexical richness is mostly concerned with the impact of spatial factors, such as how lexical richness is influenced by different writing styles or how lexical richness will vary as text length increases. This includes the above-mentioned references in this paragraph and the classic work of Shannon (1951), where maximum entropy of English was analyzed from an information science perspective, as well as the more recent works of Brown *et al.* (1992) and Genzel and Charniak (2002) with a similar focus. For all such examples, the data used and the core questions under investigation are cross-sectional, i.e., they are concerned with linguistic features at a fixed time, even though the dimension and contributing factors can be complicated.

The current study is fundamentally different in that it focuses on the evolving structure of the lexical richness over a large span of

time. In other words, it is dealing with large-scale longitude data instead of static data at a fixed time. The study investigates the lexical richness properties of a sequence of homogeneous texts, namely, the texts of China Government Work Report (CGWR) spanning from 1954 to 2011. The entropies of these texts are calculated and treated as a time series data. Under the framework of the dynamical complexity theory, the study analyzes and accurately depicts how the entropy of the CGWR texts progresses in a time span of over fifty years. Adequate probes into the data and the regression results allow us to trust on a concave and upper bounded exponential model to describe the observed entropy evolving process. Further diagnostication of the model approbates the differentiation of the whole process into two phases, namely, an initial phase where the entropy grows sharply with vehement fluctuations and a maturing phase where the process approaches a stationary baseline with small, minuscule fluctuations, where the fluctuations can be modeled by wavelet trigonometric functions. Interpolation of the initial concave exponential growth and the modulated fluctuations at the maturing phase yields a unified model that captures both the long-term trend and the local variations.

The rest of the paper unfolds as follows. Section 2 explains the CGWR corpus used for the study, followed by a preliminary analysis of the raw entropy data of the corpus. Section 3 briefly describes the dynamic complexity theory and its applications in related areas, on the basis of which postulations are drawn regarding the evolving pattern of the entropic process under review. Section 4 presents a mathematical model for capturing the global structure of the time series of the entropy data, followed by a rigorous assessment of the validity of the model. Section 5 is set out to improve the model's accuracy and predictive power by incorporating the local microscopic fluctuations of the process. Concluding remarks and possible future directions are discussed in Section 6.

2 CORPUS, MEASUREMENT, AND DESCRIPTIVE STATISTICS

2.1 *Corpus of CGWR*

The corpus used for the study consists of the CGWR written texts archived from 1954, when the first CGWR was published, to 2011,

excluding the years that the CGWR was not issued: 1961–1963, 1965–1974, and 1976–1977. Each text contains on average 22,373 Chinese characters with a standard deviation of 8256.5, making the size of the corpus approximately 962,000 characters in total. The archives of the CGWR corpus are publicly accessible at the webpage of the central government of China *www.gov.cn*. The CGWR, as one of the most important public documents in China, is drafted in accordance to a stable and formatted style, covering various major aspects of the sociocultural, political, and economic life at national level, as well as the events and projects of significance of the corresponding year.

While the sociopolitical importance of the CGWR texts is self-evident, it is their linguistically homogeneity feature that most concerns the current study. Although there exist studies such as Gries (2006) suggesting using complex techniques to quantify homogeneity, the notion of homogeneity in corpus linguistics appears rather wide and informal, as felt by Kilgarriff (2001), for instance. As to the CGWR texts in the current study, they are topically homogeneous from year to year although the emphasis may vary. They are drafted by the same institutional author whose writing style seems to be even more consistent than texts by individual authors. Moreover, the production of CGWR is periodic and subject to a strict scrutiny and modification process set by both linguistic norms and political operations.

2.2 *Entropy measure for lexical richness*

Lexical richness refers to the size of the vocabulary that is employed in language generation and how diversely the words are used. Intuitively, it reflects the degree of variations and sophistications of a spoken or written text, the production of which must of course adhere to the constraints and rules imposed by the language being used. While lexical richness is something that can be either clearly or vaguely perceived in daily conversations, assigning a numeric value to it becomes indispensable when scientific research of corpus linguistics is being conducted on a massive scale. The numeric measure adopted in the current study to quantify the lexical richness of the CGWR texts is entropy, the concept of which originates from physical sciences, particularly thermodynamics.

Consider a Chinese corpus text, denoted as T, which has n different characters indexed with 1, 2, ..., n. Assume that the relative

frequencies of each of the n characters appearing in the corpus are p_1, p_2, ..., p_n, then the entropy of the Chinese text is defined as

$$\textit{Entropy}(T) = -\sum_{i=1}^{n} p_i \ln(p_i).$$

Originally introduced in thermodynamics for quantifying the unpredictability of the microscopic state of a physical system at any given time, entropy has now become a widely accepted concept and a tested measure of uncertainty and/or complexity in many disciplines and interdisciplinary fields such as communication science, ecology, biology, and cosmology, to name a few. As useful as it is, what entropy really measures can be dependent on the context of use and field knowledge of specific disciplines. In particular, it could be naïve to treat the entropy in classical thermodynamics as equivalent to the Shannon entropy, despite that they take the same form in calculation. An insightful ontological discussion on entropy can be found in Wicken (1987), for instance. On the other hand, when used for quantifying lexical richness as in the current study, entropy should be best understood as a measure of the degree of complexity that the original system, usually composed of finite components and limited number of laws governing the interactions between the components, has developed as of today. For a fixed time horizon, what is emphasized here is the compositional complexity of the linguistic construct of a text (Jarvis 2013).

It is a simple calculation, using the above formula, to show that the maximum possible value of entropy for T is achieved when all the characters in it are different from one another, in which case $\textit{Entropy}(T) = \ln(n)$, where n is both the total number of characters (tokens) and the number of unique characters (types) appearing in the text. Nevertheless, the entropy of any meaningful text is in reality far below this number because, first, the total number of unique Chinese characters (or the total number of types of any language in general) is capped; and second, the distribution of all the unique characters (or the types of any language) is far from, not even close to, uniform distribution. As a matter of fact, the second rationale of the above partly echoes the well-known Zipf's law. Take the CGWR of 1954 as example, Table 1 provides a summary of key statistics relevant to the current study. And Figure 1 pro-

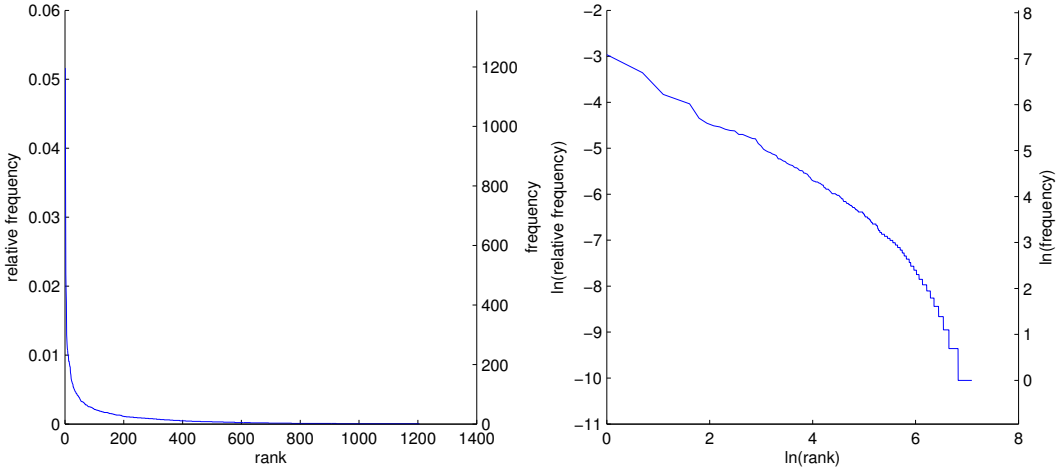


Figure 1: Frequency distribution plots of CGWR 1954 text

vides the corresponding frequency plots, where the left plot is in the original scale, and the right plot is scaled by the natural logarithm.

Table 1: Descriptive statistics of the CGWR 1954 text

Year	Total unique characters	Total characters	TTR	Entropy	Maximum entropy
1954	1205	23168	0.052	5.8601	10.0505

3 THEORETICAL FRAMEWORK AND RELATED RESEARCH

3.1 *Overview of dynamic complexity theory*

The core theoretical foundation that forms the basis for the assumptions of the current paper, and according to which the statistical models are constructed, is the theory of the dynamic complexity system. The theory, despite its diverse origins and applied fields, is formulated and commonly accepted nowadays insofar as it *corrects* the tendency in classical approaches in physical sciences to explain both natural and human phenomena with over-simplified assumptions and static mechanisms. Given its multidisciplinary and interdisciplinary nature,

it is not easy to portray a full genealogy of dynamic complexity (some antecedents of complexity theories from linguists' perspective can be found in Larsen-Freeman and Cameron 2008, pp. 2–4). Early mathematical usage of complexity using the concept of entropy is usually traced back to classic thermodynamics (Bailyn 1994), the focus concern of which is how heat is transferred in a physical system and how the system evolves in the irreversible time direction. Dynamic complexity is, in a sense, a general postulate of the second law of thermodynamics in broader disciplines beyond physics and chemistry.

It is important to keep in mind that the complexity system contextualized in contemporary scholarly research is far more "complicated" and multifaceted than its counterpart in thermodynamics. Among others, one notable difference is that traditional thermodynamics only deals with an isolated physical system, allowing no matter or energy exchange across the boundaries. Hence, the law governing the entropy process therein, as complex as it can be, is deterministic. Fundamentally different from classic sciences, the dynamic complexity theory used in this study views any examined entity as a complex and constantly evolving system, the members or components of which are interacting with each other, each evolving as a sub-system under the constraints imposed by the system as a whole. Exchange of matter, energy, and information is allowed not only among the interacting make-ups, but also between the system and the external environment in which the system is sustained. Almost as a consequence, it allows for self-organization, chaos behavior, nonlinear progression, and phase changes (Larsen-Freeman and Cameron 2008).

3.2 *Application in related research*

Nowadays, dynamic complexity theory has proven a useful framework for many applied fields in physical and social sciences. Direct or indirect introduction of dynamic complexity into studies of linguistic phenomena has led to fruitful results on a number of frontiers, particularly in the past two decades. For example, a dynamic language development approach was taken by Verspoor and Behrens (2011) to explain the role of frequency in L1 learning and the role of L1 in L2 learning. Spivey (2007) asserted the continuity of mind, emphasizing the dynamic and complex characteristic of human's cognitive, hence linguistic function. Meara (2006) adopted a similar approach for model-

ing vocabulary learning. A thorough treatment of linguistic complexity theory is presented in Larsen-Freeman and Cameron (2008), where the core rationales and properties defining "complexity systems" in language study are meticulously laid out. Many studies, such as Blevins (2004), Croft (2008), and Lee and Schumann (2003), fall within the framework of evolutionary linguistics, which partly overlaps the idea of dynamic complexity theory, particularly when the self-adaptive nature of languages is underscored. A similar approach was taken by Wang (1979) in accommodating the diffusions and randomness observed in language changes. Dynamic complexity is also presented in the competition model developed by MacWhinney (2007) in accounting for the spectrum of interrelated phenomena arising from FLA and SLAs. Useful as they are, the applications of the dynamic complexity theory in most of the existing studies are lacking a unified measure, and the analysis to date has been mostly qualitative in nature. Our statistical modeling, in part, exemplifies an attempt to bridge this gap in the focused area of corpus linguistics.

3.3 *Pertinence to CGWR*

According to Larsen-Freeman and Cameron (2008), a complex system is "a system with different types of elements, usually in large numbers, which connect and interact in different and changing ways" (p. 26). While others such as Verspoor *et al.* (2011) have summarized in different ways, virtually all the theorists agree that dynamicity and spontaneous changes between both interconnected elements as well as the system as a whole are the central property for a system to be complex. For the CGWR to be characterized as such, the constituent agents, from a complex system perspective, are the Chinese characters, words, phrases, idioms, and proper nouns commonly related to the sociocultural, political, and economic life of contemporary China. Not only are these components completely interconnected and interacting with each other spontaneously, but the discourse structure and rhetoric strategies pertaining to them are also constantly changing to fit the linguistic functions that the CGWR text is supposed to perform. When an entropic metric is imposed macroscopically, the system is unsurprisingly manifested as a self-adaptive process, evolving from simple primitive forms to more complicated ones under regulations of both

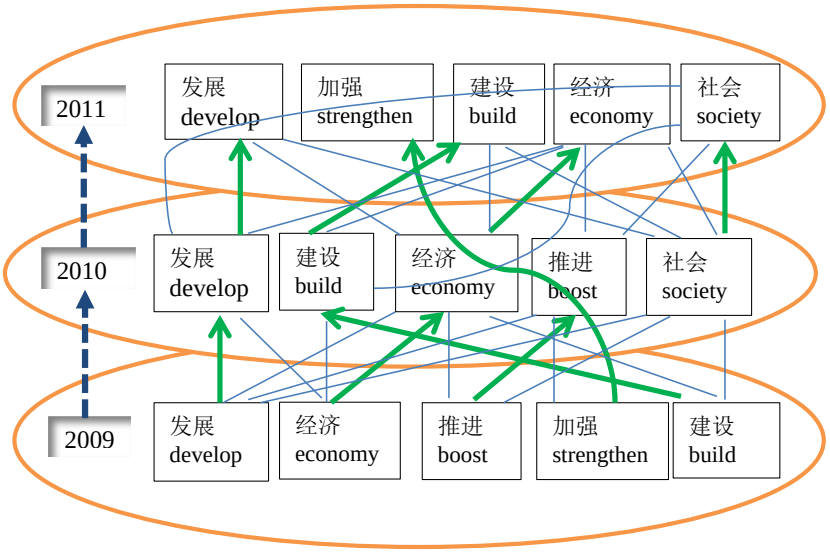


Figure 2: Most frequent content-words in the CGWR texts of 2009–11

formal linguistic rules and peripheral sociolinguistic norms of the society.

Figure 2 provides a diagram of the changes in the content-words appearing most frequently in the CGWR texts from the years 2009, 2010, and 2011, which aptly reveals the dynamic quality of the CGWR. At least three major factors contributing to the relative changes in the ranking and frequency of these content-words can be identified as follows. First is the topic continuation of CGWR over time, represented by the thick solid arrows (in green) in the diagram. For example, “to develop” or “development” played a central topical role in the CGWR in the three years under analysis; it was consistently the most frequently occurring content word across the CGWR texts during all three years (117 for 2009; 123 for 2010; 139 for 2011). Other notable topical words include “economy”, “to build”, and “to strengthen”, the relative usage of which saw more fluctuations. Second is the dynamics of lexical networks over time, denoted by the thin curve segments (in light steel blue) in the diagram, where an edge in the network can be defined by synonyms such as *tui1jin4* (to boost) and *fa1zhan3* (to develop), for instance; or a syntactic dependency as in the concurrence of *fa1zhan3* (to develop) and *jing1ji4* (economy), for instance. The third factor figuring in the dynamic quality of the

CGWR consists of the complexity explicated by the social, cultural, political, and economic contexts in which the CGWRs were drafted. This type of complexity, conceptualized by the ellipses as well as the thick dashed arrows (in dark steel blue) between such ellipses in the diagram, reflects the co-adaptive nature of the CGWR, where it allows for the exchange of energy and matter across the boundaries and draws on resources and influences from the external sociocultural environment in general. As such, a full understanding of the linguistic dynamism of the CGWR texts is not probable without reference to the parallel social, cultural, political, and economic realities of the society.

Of course Figure 2 is far from complete in depicting the infinite microscopic complexities belonging to the system under study. It only provides a glimpse, from a rather limited angle, of the vast lexical dynamics present in the CGWR texts from year to year. Many subtle changes caused by lexical inertia or a variety of cohesions are not easy to describe accurately, neither can the emergence of new words driven by technology advancement or socioeconomic shifts, for instance, be fully accounted for. Nevertheless, despite the lack of a complete microscopic description, the dynamic nature of the CGWR texts is sufficiently evident from this illustrative diagram. After all, the macro evolving pattern instead of the micro and local cause is the focus concern of the current investigation. Moreover, the goal of a dynamic approach, according to Verspoor *et al.* (2011), "is not to list possible causes for change and development but to describe the process of change and development itself by means of tracing the iterative change over time". Table 2 identifies the key properties of the CGWR serving to define its dynamic complexity nature. The items in the Field column of the table were pointed out by Larsen-Freeman and Cameron (2008) as the defining features of a system being complex. The second and third columns of the table are adapted from the same reference (p. 37).

4 GLOBAL ENTROPIC MODEL FOR CGWR TEXTS

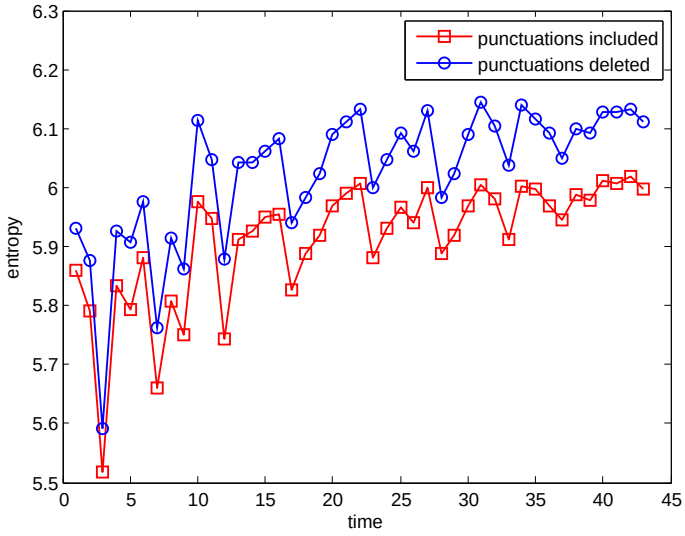
To properly envision a mathematical model that appeals to the dynamic nature of CGWR explained in the previous section and simultaneously captures its general entropic evolution pattern, it is reason-

Table 2: Defining features of CGWR and other complex systems

Field	Ecology	Classroom language learning	CGWR
Agent	individual animals	students, teachers, languages	characters, words, proper nouns
Heterogeneity	eating, nesting, breeding, habits	abilities, personalities, learning demands	meaning, lexical relationships
Organization	schools, herds, food chains	class, groups, curricula, grammars	content vs. function, part of speech, thematic group
Adaption	hunting, mating, security	imitation, memorizing, classroom behaviors	derivation, metaphor, situational context
Dynamics	predator-prey interactions, competition	classroom discourse, tasks, participation patterns	rhetorical force, styles, sociocultural influence
Emergent behavior	extinction, niches	language learning, class/group behavior, linguae francae	internet language, word fashion, popularity

able to start with a qualitative exploration of the empirically observed data. Figure 3 presents the scatterplots of the calculated entropies pertaining to the CGWR texts, where time denotes the number of years since 1953, skipping those years in which the CGWR was not issued, as pointed out in section two (the same definition applies to all the subsequent models and plots). The upper plot in Figure 3 corresponds to the data set with all punctuation deleted, and the lower plot to the data set with all punctuation included. These two series show very similar tendencies, but the entropy values for the data containing all punctuation are systematically lower than those with all punctuation deleted. The reason for such a difference is that punctuation constitutes extra linguistic constraints imposed on the text; and according to the dynamic complexity theory, the more imposed constraints, the lower

Figure 3: Scatterplot of the entropic processes



diversity of a system, with other conditions fixed. Whether punctuation should be included or not depends on the purpose of study. There exist examples where punctuation spaces are ignored (Shannon 1951, for instance) and also examples where they are included (Brown *et al.* 1992). For the subsequent analysis, all models are constructed with punctuation included, but they are equally valid for the scenario with punctuation deleted.

4.1 *Some observations*

It is a palpable observation that CGWR, as a dynamic complex system, is generally increasing in entropy. The ascending trend of the entropic process is first a manifestation of the increasing complexity of CGWR in terms of lexical choice, syntactic structuring, and discourse planning. It reflects the many and changing ways that all such constituents can interact, mutate, and concatenate with each other. To be able to appreciate this overall pattern, it helps to realize that a third-party reader will more likely to encounter new words, advanced semantic constructs, sophisticated cohesions, unprepared concepts, etc. when reading the CGWR texts in chronological order. On the other hand, CGWR is inseparably connected into the social and societal dynamics it purports to describe. This sociocultural-ecological perspective of languages (see Steffensen and Fill 2014; also Larsen-Freeman and

Cameron 2008) allows us to view the CGWR as a linguistic vessel of the society's events and histories. Ideally, the entropic process of the CGWR text shall behave in the same way as the entropic process of the societal focus it depicts, although it is quite unlikely, in reality, for such dual processes to be exactly parallel to each other. Consequently, as the complexity of human society increases (technologically, culturally, and economically), so does that of the associated linguistic agents such as the CGWR under study.

On the other hand, because the interacting linguistic components such as characters or punctuation must maintain certain lexemic, etymological and grammatical structures so as to sustain the linguistic functions of the system, the rate of increase of entropy of a homogeneous text with a roughly constant size will eventually decline, constrained by the linguistic and sociocultural conditions. Analogous arguments apply to the dual process of human society. Although interactions between parts, self-organization, randomness, nonlinear behaviors, even chaos and bifurcations are allowed in human organization, the level of possible complexities must be capped due to the constraints of, for instance, laws, cultural norms, limited capacity of production, and ethnic bonds. These conditions and constraints are necessary to conserve the defining properties of the system and prevent it from malfunctioning.

Lastly, one should expect fluctuations in the entropic process of the CGWR texts. This is different from the classic statement of the second law of thermodynamics, in which the entropy is asserted to be monotonically increasing. The difference is that a government publication is not an isolated system. Instead, it needs to accommodate the addition or deletion of lexicons, and must also allow for an inflow of foreign discourse styles, among other elements. Furthermore, such a text is subject to artificial modifications in terms of topic, theme, or size of text, in reaction to the changes in the human society system it endeavors to depict. Additionally, the fluctuations of the entropy process of a homogeneous text should be vehement in the initial stage and moderate in the maturing stage. The rationale for this, from a self-organization perspective, is that the initial stage of a system retains much less structural inertia than the maturing stage. Thus, random and nonlinear mechanics can cause dramatic changes to an emerging system with much less cost.

To summarize, the overall trend of the entropic process of the homogeneous CGWR texts is an upper bounded increasing function in time, where the trend undergoes a fast increasing initial phase before flattens to a saturated phase in the long term. In addition, fluctuations are accompanied throughout the whole process, where the magnitude of the fluctuation is large for the initial phase and small for the saturated phase.

4.2 *A basic model for the principal trend*

For quantitative modeling purpose, the study embraces a mathematical function having the following characteristics: 1) it increases rapidly in the beginning, plateauing as time goes on; 2) it is upper bounded and eventually flattens to a horizontal line, which is the upper limit of the expected entropy for the given length of the text. This leads to the following choice of equation (1), a generalized exponential type function with such postulated growth patterns:

$$E = b_1 - b_2 e^{-b_3 t}, \tag{1}$$

where t denotes time (measured in years) since the beginning of the practice of CGWR, and e is the exponential function. The same notation and definition apply to following equations and discussions. The choice of model (1) is not for the convenience of data analysis, although the exponential function, the second term of (1), is indeed a built-in class in many statistical packages, such as SPSS. It is selected because many natural phenomena, including those in linguistic processes, have been shown to develop in that way. For instance, Szmrecsanyi (2005) showed how the percentage of persistent pairs in a text, as a function of the textual distance of the pair, is decreasing exponentially. Learning effectiveness of repetition priming was reported to decay exponentially as a function of the length of the lag time (McKone 1995). Beeferman *et al.* (1997) provided an empirical study on why a model of exponential type can be used to describe the attractive and repulsive distances between word pairs with high mutual information. Despite these almost ubiquitous exponential phenomena being observed, a potential criticism might still be raised that all such examples are modeling a decaying process rather than an increasing one. But one

should note that the usual decay model taking, for example, the form $y = ae^{-bt}$, is actually a special case of (1) when the signs of the parameters are not restricted. There is a bound parameter for the usual decay model also, in the sense that zero is its lower bound.

The procedure to find the best estimates of the parameters appearing in model (1) as well as a procedure for model evaluation, under the normality assumption of errors, can be facilitated by statistical packages such as Matlab or SPSS. Precautions still need to be taken, though, in terms of choosing reasonable initial guesses of the target parameters and avoiding common pitfalls associated with nonlinear regression, e.g., over-fitting. The least square optimization procedure yields the following estimates for the model defined by equation (1):

$$b_1 = 6.0222$$
$$b_2 = 0.3089$$
$$b_3 = 0.0580$$

The corresponding standard errors for the parameter estimates are 0.0558, 0.0483, and 0.0301. The R-squared and adjusted R-squared statistics for the model are 0.5406 and 0.5177, respectively. Although the R-squared value does not seem impressively high, one should keep in mind that the validity of a nonlinear model is not solely, and not even largely, determined by the magnitude of the R-squared value when the general trend of a process is the main concern of a study. For a more rigorous explanation of why the R-squared value should not be a main concern in trend analysis, one can refer to Wittink (1988). On the other hand, the R-squared value of the basic model (1) can indeed be improved, as discussed in the next section. The t-statistics for parameters are 107.8670, 6.3982, and 1.9262, with corresponding p-values of 0.0000, 0.0000, and 0.0305 (accurate to four decimal places), respectively. Clearly each t-statistic is large enough and each p-value is small enough, which strongly justifies the statistical significance of each individual parameter in model (1). The calculated F-statistics for the model is 24.1258, with the corresponding p-value of 1.1870×10^{-7}. The overall explaining power of the model is strong. Figure 4 plots the fitted curve of the global model, together with a 95% confidence band of the regression.

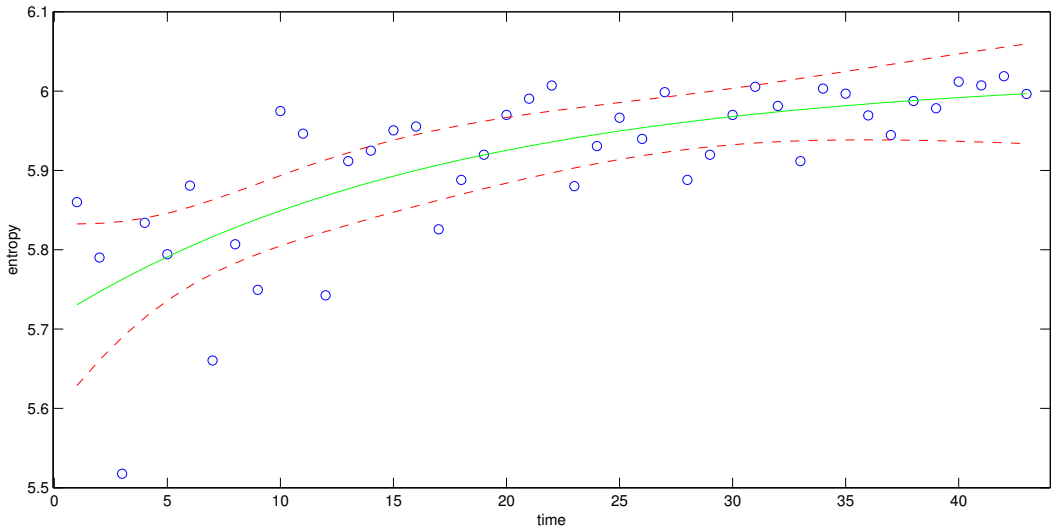


Figure 4: Plot of the original data (the circular points), basic global model (the connected curve), and 95% confidence band of the regression (the region between the dashed curves)

4.3 *Model the local fluctuations*

The backbone structure of the dynamic evolution of the entropy process is implied by the concave exponential model defined by equation (1), yielding not only a quick growth feature in the beginning of the process, but also a quick plateau effect when time is large. This said, the model does not capture the microscopic structure of the process, which exhibits large or small fluctuations at all times. There are standard statistical methods that might help to improve the model accuracy, such as smoothing and autoregression. But a relatively simple and more direct approach is to introduce the wavelike functions to the model, namely, trigonometric sine or cosine functions.

Notice the fluctuation of entropies is initially more vehement and becomes moderate later on as the process approaches steady state. It is therefore plausible to separately analyze the process in two stages: an initial quick growth stage where the process is more volatile, and the steady stage where the growth momentum is mild and the fluctuation is moderate. A clue to this can be obtained by an expository check of the scatterplot of the entropy, where the 12th data point appears to be the borderline after which the series becomes relatively stationary. To

validate statistically, one can appeal to the unit root test, a standard procedure in time series analysis for testing whether a given series is stationary or not at a prescribed level of confidence. The procedure applied to the 31 observations, i.e., the suggested steady stage of the original entropic time series, rejects the null hypothesis that the series under testing has a unit root, or equivalently affirms the hypothesized stationary nature of it, and does so at a 90% confidence level. To be specific, the augmented Dickey-Fuller statistic is −2.7603. The critical values of the test are −2.6210, −2.9640, and −3.6702 at, respectively, the 90%, 95%, and 99% confidence levels.

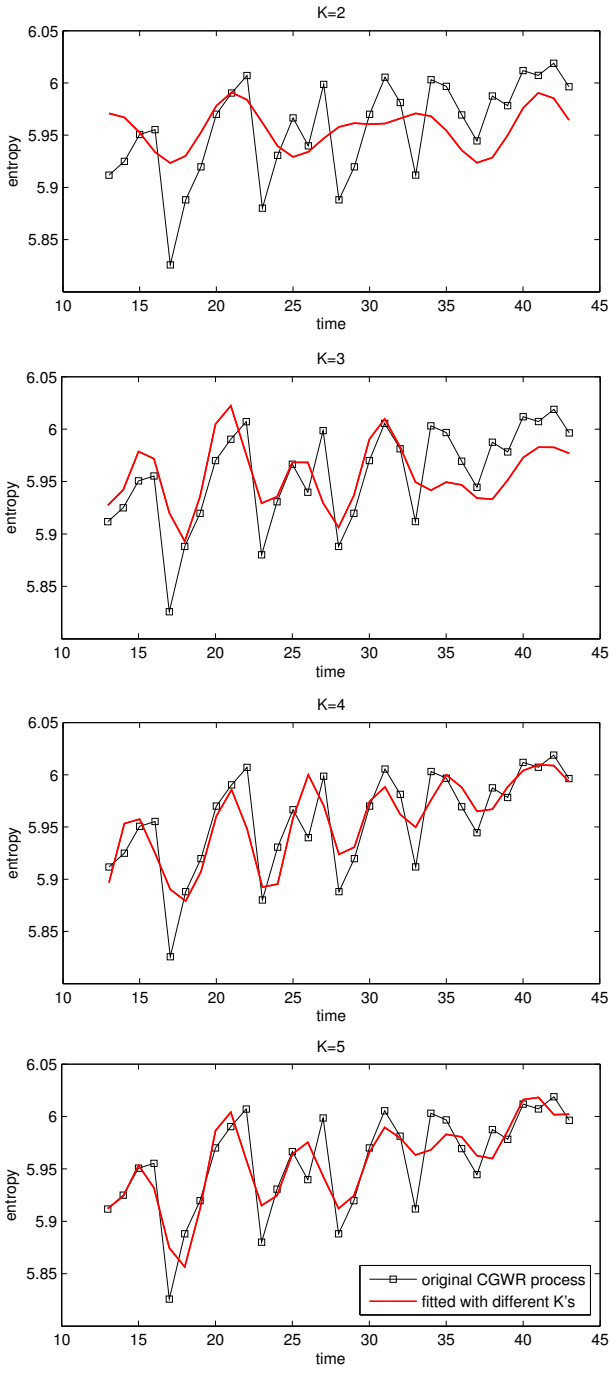
Now we turn to the modeling of the steady state with a cutting off point set at the 12th point. To model this truncated series of data with the fluctuation characteristics being the core concern, the following trigonometric functions are chosen:

$$E = k_0 + \sum_{j=1}^{\mathrm{K}} k_1(j)\sin(k_2(j)t + k_3(j)). \tag{2}$$

When K is specified, the parameters can be determined using the same regressing procedure carried out for the model defined by equation (1), and the regression evaluation can be performed accordingly. The only new issue that may complicate the procedure is the choice of K, which defines how many trigonometric functions are to be used in the model. With homogeneity of the data and validity of the model (2) in mind, one can apply an iterative scheme to find such an optimal K numerically. For the current analysis, the following rules of thumb were followed in selecting the optimal K, namely, i) the increase in adjusted R-squared divided by the increase in R-squared value is approaching maximum; ii) the majority, if not all, estimates are significant enough, judging by the corresponding t-statistics or p-values; iii) the overall F-statistics for the model is significant enough. By these rules of thumb, K = 4 is found to be optimal for the model under review. Figure 5 presents the comparison plots for cases K = 2, 3, 4, 5. Table 3 summarizes the key regression statistics, where K = 4 is observed as the choice of how many sine functions to include for best fitting the data.

One comment to add is that there appears to be a relatively large gap between the R-squared value and the adjusted R-squared value.

Figure 5:
Plot of the steady state series and the fitted curves using trigonometric functions with K = 2, 3, 4, 5



	K=2	K=3	K=4	K=5
R-squared	0.1801	0.4197	0.6761	0.7131
Adjusted R-squared	0.0248	0.1710	0.4602	0.4262
F-statistics	0.9155	1.7682	3.3054	2.6515
p-value for F-test	0.5002	0.1325	0.0099	0.0309

Table 3:
Regression statistics for the steady state series with different K

This is mainly a consequence of the small size of the data set. When the sample size is large enough compared to the number of independent variables, the adjusted R-squared value should be virtually the same as the R-squared value itself. This also implies that the model will work better as the CGWR corpus increases in size.

5 IMPROVED GLOBAL MODEL

The improvement of the global model can be achieved by a consolidation of the overall concave exponential structure and the trigonometric microscopic structure of fluctuations. In other words, the exponential component and the trigonometric component jointly depict the evolution of entropy with high resolution at local and global levels. Specifically, the following model is proposed for this purpose:

$$E = b_1 - b_2 e^{-b_3 t} + b_4 \sin(b_5 t + b_6) e^{-b_7 t}. \tag{3}$$

The product term of the exponential factor and the trigonometric factor corresponds to the interactions between the general trend and local fluctuations. The magnitude of the wavelets yielded by the product term is high when t is small, and low when t is large, making the term a suitable choice to describe the fluctuations observed in the CGWR process. The parameter estimation procedure is same as that applied to model (1). Actually the values of the estimated parameters for model (1) can be used as part of the initial guesses for the parameter vector for model (3). Going through the nonlinear regression procedure leads to the following parameter estimations:

$$\begin{aligned} b_1 &= 6.0169 \quad & b_2 &= 0.3105 \\ b_3 &= 0.0620 \quad & b_4 &= 0.1721 \\ b_5 &= 1.2703 \quad & b_6 &= 0.9612 \\ b_7 &= 0.0740 \end{aligned}$$

The model is statistically significant, with a sound explaining power, as shown in all critical aspects of observed statistics under scrutiny via various standard tests. The t-statistics for all the parameters $b_1 - b_7$ are sufficiently large and the p-values for all the parameters $b_1 - b_7$ are sufficiently small, where the observed smallest t-statistic is 2.4950 (for the parameter b_3), corresponding to a p-value of 0.0086. The significance of each parameter can also be assessed by how far away the confidence interval of the estimate is from zero, at the prescribed confidence level. Computation shows that the 90% confidence intervals for all the parameter estimates are far enough from zero. For instance, the ratio between the estimate of b_3 and the corresponding half confidence interval is about 1.4797; and it is the lowest one among the seven such ratios. The R-squared and adjusted R-squared values are 0.7432 and 0.7004 respectively, both at acceptable levels for such a highly nonlinear model with frequent fluctuations. The overall validity of the model is particularly shown in the significance of the F-statistics, which is 17.8450, and the corresponding p-value, which is 1.3692×10^{-9}. In addition to the significance of each individual parameter $b_1 - b_7$, none of the paired correlations between the estimated parameters is higher than 0.8 in absolute value except for parameters b_3 and b_1, the correlation between which is about -0.93, implying that the model basically does not have the problem of parameter redundancy and over-fitting. The full correlation matrix of the estimated parameters for model (3) is provided in Table 4.

Table 4: The correlation matrix of the estimated parameters in the global model (3)

	b_1	b_2	b_3	b_4	b_5	b_6	b_7
b_1	1						
b_2	0.4527	1					
b_3	−0.9285	−0.1677	1				
b_4	−0.0462	−0.0229	0.0443	1			
b_5	−0.1310	0.1510	0.1864	−0.1057	1		
b_6	0.1885	−0.1980	−0.2620	0.1398	−0.7675	1	
b_7	−0.0576	−0.0212	0.0551	0.7454	−0.0727	0.0945	1

It can be verified, however, that one or more of the above conclusions will be violated or weakened when one or more trigonometric terms are added, which shows that the model in the current formulation is optimal in terms of how many corrective terms need to be in-

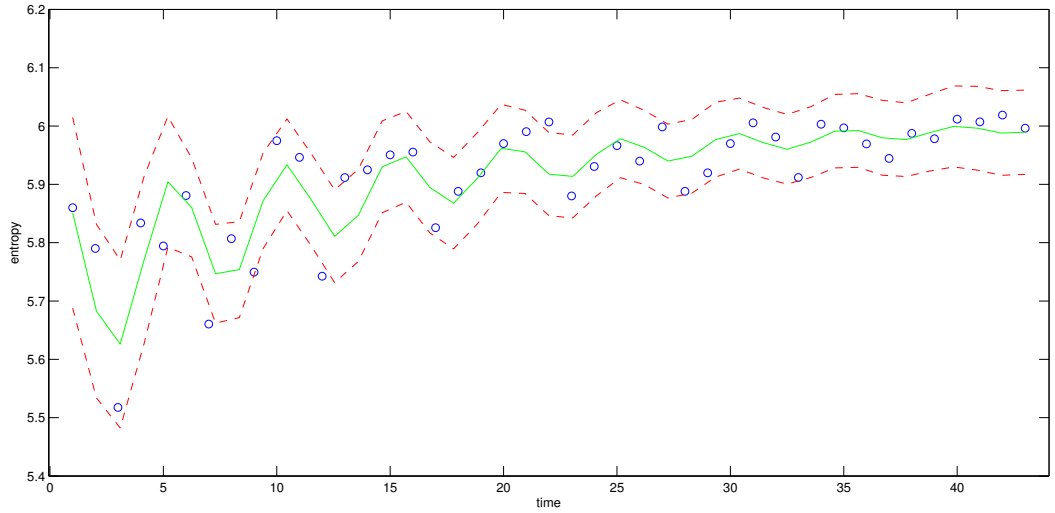


Figure 6: Plot of the original data (the circular points), improved model (the connected curve), and 95% confidence band of the regression (the region between the dashed curves)

corporated, given the current component functions and format of the model. For example, adding another trigonometric term can slightly increase the R-squared value to 0.7484, but the adjusted R-squared value will drop to 0.6214. Figure 6 plots the fitted curve and the corresponding 95% confidence band of the regression.

Although the above internal regression procedure appears to favor model (3), it is a reasonable concern that it does not overfit the data, especially because the available CGWR texts are relatively scant. The key issue here is whether adding more parameters into model (1) is a worthy effort when extrapolative prediction is also taken into consideration. Since we are mainly concerned with the evolution pattern of the CGWR text in the irreversible time direction, the appropriate numeric overfitting test to apply will be the out-of-sample prediction test, i.e., assessing which model can better forecast the future movement of the entropic process of the CGWR based on the past information. Many overfitting statistics have been developed and used for time series model selection. Here, I choose the three most widely used statistics, namely, mean squared error of out-of-sample prediction (MSE), mean absolute error of prediction (MAE), and mean absolute percentage error of prediction (MAPE) to compare the model (1) and (3). In

addition, I present two more statistics for comparison: one is prediction error variance (PEV), measuring how consistent the errors are; and the other is the Theil statistic (Theil) measuring how relatively effective the model is compared to a naïve model, where the future value is simply predicted as the current value. For detailed discussion of these statistics as well as their relevance in assessing the overfitting of time series modeling, one can, for instance, refer to Bisgaard and Kulahci (2004) and Fildes (1992).

To carry out the out-of-sample cross validation, one needs to decide on a cutting point on the time direction. Thereafter the sequential data are used as the pseudo future cases against which the predicted values are compared. While the choice of the size of this test set is not completely rigid, this study follows the fourth quarter holdout rule, i.e., the rounding point of the 25% of the time series from the end as the starting point of the out-of-sample prediction test, which has been agreed upon by most of the theorists and practices for time series modeling (Hastie *et al.* 2009). Table 5 provides the 1-step-ahead and 3-step-ahead forecast statistics of the model (1) and model (3), where it is evident that the one-year forward movement of the entropic process is consistently more predictable with model (3) than with model (1) under all the chosen testing criteria.

Table 5: Out-of-sample test statistics for model (1) and (3)

		MSE	MAE	MAPE	PEV (10^{-5})	Theil
1-step-ahead forecast	model (1)	0.0009	0.0232	0.0039	0.9390	0.5255
	model (3)	0.0007	0.0222	0.0037	0.5535	0.4169
3-step-ahead forecast	model (1)	0.0004	0.017	0.0028	0.4450	0.7879
	model (3)	0.0004	0.0182	0.0030	0.3086	0.7778

It is worth noting that the Theil statistics for both models are at an acceptable level, affirming the usefulness of both the models, regardless of the difference in the out-of-sample prediction tests. On the other hand, the improvement in prediction accuracy of the model (3) does not seem to compensate for its increased complicatedness, when judged by the multi-step-ahead forecast statistics. To interpret these statistics, it is worthwhile to refresh the idea that “overfitting is not

an absolute but involves a comparison" (Hawkins 2004). Similar precautions have been expressed by Bisgaard and Kulahci (2004) – that numerical and statistical tests of overfitting should not be applied mechanically without reference to the research contexts and purposes. Because the CGWR is viewed in the current study as a dynamic complexity system which intrinsically allows for fluctuations, model (3) appears a plausible choice when a near future prediction – such as one year in advance – is the main concern. One should nevertheless be aware that a more complete picture of the progressive pattern will only be visible as more real data are accumulated over time.

Lastly, the above models (1) and (3) are based on the assumption that the fitted residuals are normally distributed, which needs to be justified. While an apparent violation of normality can often be detected by simple graphical methods such as probability plot or QQ plot, numerical tests are necessary for subtle cases. Here four widely used procedures, namely, Kolmogorov-Smirnov (KS) test, Lilliefors (Lillie) test, Shapiro-Wilk (SW) test, and Anderson-Darling (AD) test were run and the corresponding results are presented in Table 6. For relevance and comparative powers of these tests for normality testing, one can refer to Razali and Wah (2011). For model (3), the null hypothesis (H_0) that the residuals are normally distributed is solidly affirmed as it passed all normality tests with sufficiently high p-values. On the other hand, the normality assumption is only marginally satisfied for model (1). When the significance level of the test, defined as the probability of the Type I error, is set at 0.01, model (1) can pass all the normality tests; but when the significance level is set at 0.05, it only passes KS test and fails all the rest (see Table 6 for detailed statistics).

This is to some extent understandable, as the CGWR process undergoes large fluctuations in the initial stage. Model (1) was created to capture only the main trend of the process in the first place, where the local fluctuations were not accounted until the trigonometric terms were introduced as in model (3). Because the variability of error induced by model (1) systematically decreases from large to small, the slight non-normality of it can be easily rectified. One convenient method for this purpose is to appeal to what is called the linearization transformation of random variables (Schabenberger and Pierce 2002; Chatterjee and Hadi 2012). In our case, the desired transformation is

$$Y = \ln(b_1 - E),$$

where E is entropy, the response variable of model (1), and $b_1 = 6.0222$ is the parameter appearing in model (1) defining the asymptotic upper bound of the entropic process of CGRW. A simple algebraic operation based on the above transformation yields a linear model of the following form:

$$Y = c_1 + c_2 t. \tag{4}$$

The rest of the original parameters of model (1) can be recovered from the parameters of the linearized model by $b_2 = e^{c_1}$ and $b_3 = c_2$. A Simple least square regression gives the best estimation of the transformed parameters as $c_1 = -1.2620$ and $c_2 = -0.0633$. Indeed, the transformed model defined by the above linear equation neatly satisfies the normality requirement, where the statistics of the corresponding normality tests are also tabulated in Table 6.

Table 6: Statistical tests for the normality of the models

		KS	Lillie	SW	AD
Model (1)	H_0 (0.01 significance)	accepted	accepted	accepted	accepted
	H_0 (0.05 significance)	accepted	rejected	rejected	rejected
	p-value	0.3474	0.0363	0.0137	0.0313
	Statistic	0.1387	0.1387	0.9321	0.8182
Model (1) Log-Transformed	H_0 (0.01 significance)	accepted	accepted	accepted	accepted
	H_0 (0.05 significance)	accepted	accepted	accepted	accepted
	p-value	0.9687	0.8230	0.3572	0.5978
	Statistic	0.0717	0.0717	0.9716	0.2987
Model (3)	H_0 (0.01 significance)	accepted	accepted	accepted	accepted
	H_0 (0.05 significance)	accepted	accepted	accepted	accepted
	p-value	0.9175	0.6750	0.4016	0.4216
	Statistic	0.0812	0.0812	0.9731	0.3672

One comment I would like to add is that the linearization procedure via logarithm transform only leads to the significance of nor-

mality for the modeling; it does not improve the model accuracy. In particular, model (1) is still relatively inferior compared to model (3), judged by the out-of-sample predicting errors, regardless it is expressed in terms of the original upper-bounded exponential function or the logarithm transformed linear function.

6 CONCLUDING REMARKS

Taking the CGWR as the test corpus, the current work investigates how lexical richness of a series of homogeneous texts evolves over a large time horizon. The dynamic complexity theory is shown to be a pertinent and valid foundation in the current study in the context of the existence conditions of homogeneous texts. It provides a relevant method for looking at the CGWR corpus as an open, dynamic, heterogeneous, and self-adaptive system. Additionally, the strong, distinctive, and mathematically describable properties exhibited in the associated time series of the entropic data are naturally hinted by the key theoretical implications of a dynamic complexity system. Although the base functions used in our modeling – a class of concave down exponential functions and a class of periodic trigonometric functions – seem rather simple, it takes a novel combination of them together with their interactions to produce an effective quantitative model. The models and results of the current work demonstrate that the dynamic complexity approach is not only metaphorically plausible, but is also conducive to rigorous quantitative conclusions.

A major limitation of the current study is the small size of the available data, resulting from the relatively short history of the CGWR practice. Given that the CGWR is an institutional writing process which is subject to the influence of the sociocultural environment in which it is embedded, it may be hasty to assume that the CGWR will reach its peak of lexical richness within fifty years since its inception. Because of this limitation, the models contained in the current study could have only provided a partial picture of an even larger evolving pattern which may not stand out until sufficient time has passed. For instance, it is possible that the periodicity of local fluctuations, described by the parameter b_5 in model (3), will not keep constant for an arbitrarily long time. In addition, it could be the case that the saturation state observed in the current paper is not the ultimate one

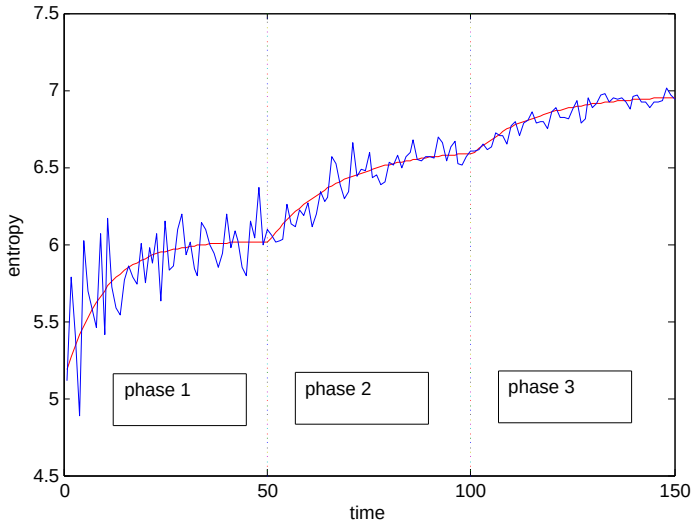


Figure 7: Illustration of multi-phase entropic growth

when the future evolution of the CGWR process is taken into consideration. Among other reasons, phase change is known to be a common characteristic of linguistic dynamics (Larsen-Freeman and Cameron 2008), implying the probableness that the saturation currently observed is only one of the multiple local saturations to come when the data is large enough. Figure 7 presents a simulated illustrative example where an entropic process undergoes three growth phases; each can be roughly estimated by model (3). This is in spite of the observation that all such three phases are again subordinated to a large-scale exponential decay model when interpolated together. More examples of multi-phase linguistic dynamics can be found in Larsen-Freeman and Cameron (2008), Verspoor *et al.* (2011), and Stachowski (2013), for instance.

The models provided in the current paper, when extrapolated backwards, can also provide useful hints to the pattern of language changes in historical linguistic studies. In particular, model (3), where it is appropriate to apply, tends to hint that language changed more dramatically in the farther past than in more recent times. To cite one example, the occurrence of paratactic constructions in written English such as left-dislocated NPs had undergone a roughly exponential decay during the years 950–1910 from Old English to Early Modern English and to Modern English, where the changes in the first 500

years were more vehement before being stabilized since about 1450 (van Kemenade and Los 2014). As an example in Chinese, the relative frequency of *ye3*, a sentence-final interjective marker which is often an emblematic of a Chinese text being classic, had seen gradual decrease from pre-10th century to 20th century, where the changes were more volatile and dramatic before 17th century (Shi 1989). This said, much depends here on the overall trending of the underlying process, there are cases where an exponential model is not suitable. The occurrence of unique Turkic glosses in Polish texts from 1388 to 1791 reported by Stachowski (2013) and the increase of the use of the English auxiliary *do* as a negative declarative demonstrated by Ellegård (1953) are such examples. Logistic functions, instead of exponential functions, should be used to best describe the respective linguistic phenomena, where a slow initial growth period is present before a more dramatic growth period emerges. In addition to model selection, the comparability and representativeness of the historical texts are also critically important when backward extrapolation is applied to infer language changes in the past. Caveats and pitfalls may arise because language data, when drawn from different historical periods, can be very inhomogeneous in dialect, genre, register, and sociolinguistic environment. For further technical precautions in using limited historical texts to extrapolate general pattern in the past one can refer to van Kemenade and Los (2014).

For future work, it is important to enrich the current research with similar empirical tests using other types of homogeneous texts in Chinese, so as to generalize the conclusions made in this study. Systematic differences in terms of the concavity of curve or parameter values or sharpness of the initial increasing phase of the curve might be detected when homogeneity changes across different corpora. Admittedly, however, the more challenging task will be how to account for such cross-corpora differences from pertinent theories, some of which can be more innately rooted in the mechanisms of language development. Dynamic complexity theory, generally concerned with the structural distribution from macro and inferential perspective, does serve as a substitute for causal effect analysis in specific linguistic fields.

In addition, testing of the models against homogeneous texts in languages other than Chinese might generate insightful comparisons. Given that Chinese and English are very different in many aspects,

including orthographic form, syntactic rules, and semantic structure (Ku and Anderson 2003; Perfetti and Tan 1998), whether or not the lexical richness measures that were developed historically for alphabetical languages are readily applicable to Chinese as an orthographic language is a reasonable concern. As shown by Figure 1 of the current paper, the frequency distribution of the CGWR text (in log-log scale) can be very different than one might expect for English. Specifically, the frequency distribution of Chinese tends to exhibit a larger concavity after a certain rank of unique characters (typically in thousands) is reached, whereas that of English tends to progress with a more stable slope. This rank-frequency distributional difference between the two languages has been verified by recent empirical studies such as Chen *et al.* (2012). How this difference will affect the entropic process of English homogeneous texts and whether the pattern uncovered in the current study will equally hold for the counterpart in English will be a worthwhile future direction.

Another aspect desiring more fine-tuned investigation is the mechanism leading to the periodic, although modulated, fluctuations manifested in the entropic process of the CGWR texts. Possible approaches may include a careful examination of the recurrent sociolinguistic themes to which the CGWR sporadically refers. For example, strategic planning is a central characteristic of China's economy, where a top-down Five-Year Plan is developed by the government every five years to mobilize resources for identified priorities. It is then a legitimate question to ask whether a sort of correlation exists between such a recurring socioeconomic initiative and the observed periodic pattern in the entropic process of the CGWR text.

ACKNOWLEDGEMENT

The author is very grateful to three anonymous referees as well as the editors of JLM, whose comments and suggestions have been of great help in improving the earlier versions of this paper.

REFERENCES

Martin BAILYN (1994), *A Survey of Thermodynamics*, American Institute of Physics, New York.

Doug BEEFERMAN, Adam BERGER, and John LAFFERTY (1997), A model of lexical attraction and repulsion, in *Proceedings of the ACL*, pp. 373–380, Madrid, Spain.

Soren BISGAARD and Murat KULAHCI (2004), *Time Series Analysis and Forecasting by Example*, John Wiley & Sons, Hoboken, New Jersey.

Juliette BLEVINS (2004), *Evolutionary Phonology: The Emergence of Sound Patterns*, Cambridge University Press, Cambridge, MA.

Peter F. BROWN, Steven A. Della PIETRA, Vincent J. Della PIETRA, Jennifer C. LAI, and Robert L. MERCER (1992), An estimate of an upper bound for the entropy of English, *Computational Linguistics*, 18(1):31–40.

Samprit CHATTERJEE and Ali S. HADI (2012), *Regression Analysis by Example*, John Wiley & Sons, New York.

Qinghua CHEN, Jinzhong GUO, and Yufan LIU (2012), A statistical study on Chinese word and character usage in literatures from the Tang Dynasty to the present, *Journal of Quantitative Linguistics*, 19:232–248.

William CROFT (2008), Evolutionary linguistics, *Annual Review of Anthropology*, 37:219–234.

Scott A. CROSSLEY and Danielle S. MCNAMARA (2011), Shared features of L2 writing: Intergroup homogeneity and text classification, *Journal of Second Language Writing*, 20(4):271–285.

Etienne DENOUAL (2005), The influence of example-data homogeneity on EBMT quality, in *Proceedings of the Second Workshop on Example-Based Machine Translation*, pp. 35–42, Phuket, Thailand.

Alvar ELLEGÅRD (1953), *The Auxiliary Do: the Establishment and Regulation of Its Use in English*, Almqvist and Wiksell, Stockholm.

Robert FILDES (1992), The evaluation of extrapolative forecasting methods, *International Journal of Forecasting*, 8:81–98.

Dmitriy GENZEL and Eugene CHARNIAK (2002), Entropy rate constancy in text, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 199–206, Philadelphia.

Stefen Th. GRIES (2006), Exploring variability within and between corpora: some methodological considerations, *Corpora*, 1(2):109–151.

Trevor HASTIE, Robert TIBSHIRANI, and Jerome FRIEDMAN (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Douglas M. HAWKINS (2004), The problem of overfitting, *Journal of Chemical Information and Computer Sciences*, 44:1–12.

Scott JARVIS (2013), Capturing the diversity in lexical diversity, *Language Learning*, 63:87–106.

Victoria JOHANSSON (2008), Lexical diversity and lexical density in speech and writing: a developmental perspective, *Lund Working Papers in Linguistics*, 53:61–79.

Adam KILGARRIFF (2001), Comparing corpora, *International Journal of Corpus Linguistics*, 6(1):1–37.

Adam KILGARRIFF and Gregory GREFENSTETTE (2003), Introduction to the special issue on the web as corpus, *Computational Linguistics*, 29(3):333–348.

Andras KORNAI, Peter HALACSY, Viktor NAGY, Csaba ORZVECZ, Viktor TRON, and Daniel VARGA (2006), Web-based frequency dictionaries for medium density languages, in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1–8, Trento, Italy.

Yu-Min KU and Richard C. ANDERSON (2003), Development of morphological awareness in Chinese and English, *Reading and Writing: An Interdisciplinary Journal*, 16(1):399–422.

Diane LARSEN-FREEMAN and Lynne CAMERON (2008), *Complex Systems and Applied Linguistics*, Oxford University Press, Oxford.

Namhee LEE and John H. SCHUMANN (2003), The evolution of language and the symbolosphere as complex adaptive system, paper presented at *the American Association of Applied Linguistics Conference*, Arlington, VA.

Brain MACWHINNEY (2007), A unified model, in P. ROBINSON and N. ELLIS, editors, *Handbook of Cognitive Linguistics and Second Language Acquisition*, Lawrence Erlbaum Associates, Mahwah, NJ.

Elinor MCKONE (1995), Short-term implicit memory for words and non-words, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21:1108–1126.

Paul MEARA (2006), Emergent properties of multilingual lexicons, *Applied Linguistics*, 27(4):620–644.

Charles A. PERFETTI and Lihai TAN (1998), The time-course of graphic, phonological, and semantic activation in Chinese character identification, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24:1–18.

Nornadiah M. RAZALI and Yap B. WAH (2011), Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests, *Journal of Statistical Modeling and Analytics*, 2(1):21–33.

Magnus SAHLGREN and Jussi KARLGREN (2005), Counting lumps in word space: density as a measure of corpus homogeneity, in *Proceedings of 12th Symposium on String Processing and Information Retrieval*, pp. 124–132, Buenos Aires, Argentina.

Oliver SCHABENBERGER and Francis J. PIERCE (2002), *Contemporary Statistical Models for the Plant and Soil Sciences*, CRC Press, New York.

Claude E. SHANNON (1951), Prediction and entropy of printed English, *Bell System Technical Journal*, 30:50–64.

Ziqiang SHI (1989), The grammaticalization of the particle le in Mandarin Chinese, *Language Variation and Change*, 1:99–114.

Joseph. A. SMITH and Colleen KELLY (2002), Stylistic constancy and change across literary corpora: Using measures of lexical richness to date works, *Computers and the Humanities*, 36:411–430.

Michael SPIVEY (2007), *The Continuity of Mind*, Oxford University Press, Oxford.

Kamil STACHOWSKI (2013), The influx rate of Turkic glosses in Hungarian and Polish post-mediaeval texts, in R. KÖHLER and G. ALTMANN, editors, *Issues in Quantitative Linguistics*, pp. 100–116, RAM-Verlag, Lüdenscheid.

Sune V. STEFFENSEN and Alwin FILL (2014), Ecolinguistics: the state of the art and future horizons, *Language Sciences*, 41(6):6–25.

Benedikt SZMRECSANYI (2005), Language users as creatures of habit: A corpus-based analysis of persistence in spoken English, *Corpus Linguistics and Linguistic Theory*, 11:113–150.

Ans VAN KEMENADE and Bettelou LOS (2014), Using historical texts, in D. SHARMA and R. PODESVA, editors, *Research Methods in Linguistics*, pp. 216–231, Cambridge University Press, Cambridge.

Marjolijn H. VERSPOOR and Heike BEHRENS (2011), Dynamic systems theory and a usage-based approach to second language development, in M. VERSPOOR, K. DE BOT, and W. LOWIE, editors, *A Dynamic Approach to Second Language Development: Methods and Techniques*, pp. 25–38, John Benjamins, Amsterdam.

Marjolijn H. VERSPOOR, Kees DE BOT, and Wander LOWIE, editors (2011), *A Dynamic Approach to Second Language Development: Methods and Techniques*, John Benjamins, Amsterdam.

William S-Y. WANG (1979), Language change: a lexical perspective, *Annual Review of Anthropology*, 8:353–371.

Jeffrey S. WICKEN (1987), Entropy and information: suggestions for common language, *Philosophy of Science*, 54:176–193.

Dick R. WITTINK (1988), *The Application of Regression Analysis*, Allyn and Bacon, Boston, MA.