Modelling the lexical complexity of homogenous texts: a time series approach

Yanhui Zhang

Accepted: 25 May 2022 / Published online: 18 June 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract
Lexical complexity of homogeneous texts, especially when produced by an institutional author over time, exhibits a generally observed increasing trend with local random fluctuations. Such an irreversible entropic process fits very cogently into the dynamical complexity system theory, where the social, economic, and cultural missions such texts set to serve constitute the underlying driving momentum for the texts to adapt themselves from low to high complexity. Structural equations have been shown effective in modeling such macroscopic behavior of the entropic process of the homogeneous texts. The current work formulates the problem from a time series modeling approach applied to a large socio-linguistic corpus in written Chinese. The findings show that such an alternative approach not only produces as valid models with strong goodness of fit as the structural equation approach, but also exhibits, by design, additional benefits in explaining the entropic process of homogeneous texts in the dynamical complexity system framework. Some technical challenges, such as phase change in model calibration, are also solved with less cost using the newly proposed approach. Further directions are pointed out to more fully compare these approaches in the setup of the current study and corpus linguistics in general.

Keywords Homogeneous texts · Dynamical complexity · Entropy · Time series · ARIMA model

1 Introduction

Homogeneity provides a natural classification criterion to otherwise highly randomized organisms in various systems. In the context of linguistic studies, particularly corpus linguistics, homogeneous texts refer to a collection of texts profiling much enough in common in one or more linguistic characteristics that an underlying homogeneity is believed to be present to account for such observed commonness. The production mechanism of the texts is often, although not always, the source of homogeneity (Lai et al. 2020; Zhang, 2016; Boyd and Fraurud, 2010). From quantitative modeling and empirical analysis
perspectives, homogeneity may help filter potential confounding factors and random noises in the observed realization of the response variables, increasing the validity and explaining the power of the model constructed. For instance, Mu and Dooley (2015) implicitly relied on the homogeneity in the first language (L1) background in both experiment and model design and ANOVA analysis in demonstrating how household background contributed to the literacy development among heritage Chinese learners in Australia. Such dichotomy of homogeneity and heterogeneity served as a critical methodological defense for many studies, explicitly or implicitly, in assessing how relatively effective the different L1 backgrounds or the different designs of learning and instructional schemes may affect the L2 literacy development (Maxwell et al. 2018; Li et al. 2019; Paap et al. 2014; Komachali and Khodareza, 2012; Zhou et al. 2019). More examples illustrate how homogeneity plays an essential role in the theoretical formulation or experiment design. Such instances include the cross corpora analysis (Kilgarriff, 2001; Kilgarriff and Grefenstette, 2003; Denoual, 2005; Smith and Kelly, 2002), topical dispersion analysis (Sahlgren and Karlgren, 2005), second language (L2) writing classification (Crossley and McNamara, 2011), and language assessment validations (Yao and Chen, 2020; Jang et al. 2015).

A critical dimension of homogeneous texts is lexical complexity, especially when the exact linguistic production mechanism generates the texts at different times. According to the dynamic complex system theory (Larsen-Freeman and Cameron, 2008; Verspoor et al. 2011), the corpus of homogenous texts can be viewed as an entropy system. In the system, the texts perform the communicative functions, constituting the propelling force to fuel the evolution of the corpus’ entropic process. The dynamically complex nature of the corpus is compellingly exhibited in several aspects, including the expansive linguistic dimensions, such as semantic, syntactic, discourse, the infinite combinatorial patterns of the constituents, the nonlinear or even chaotic interactions, and the potential self-organizational mechanisms. Such dynamic complexity is reinforced when the system allows information or energy to exchange between the system and its sustaining environment.

In the case of institutional writings such as government work reports, the external environment is broadly made up of various socio-economic and cultural resources of the society, which provide but also constrain the government’s capacity in achieving its policy agenda. In other words, these reports are both socially conditioned and socially constitutive in that the collective mentalities of the society regulate their linguistic norms and styles. An essential example is the annual China Government Work Report (CGWR), delivered in the high profile “Two Sessions” – the annual plenary meetings of the Chinese National People’s Congress and the Chinese People’s Political Consultative Conference (CPPCC). As such, the CGWR is a dominantly important public document laying out the government’s policy priorities for the following year or years to come and encapsulating the achievements and tasks accomplished in the previous year. Given the above shown interacting relationship between language and society, an analysis on the evolution of the lexical complexity of the CGWR texts, which is the focus of the current study, is highly relevant for an insightful understanding of how the socio-economic and cultural milieu are linguistically manifested from a dynamical sociolinguistic perspective.

One major challenge for dynamical complexity approaches to linguistics is the quantifiability of the theory with conceptually justified and computationally rigorous models. As a recent development, Zhang (2016) proposed a structural equation model describing the global entropic pattern of CGWR at a macroscopic level. One limitation of the structural equation modeling, however, is its time-invariance, making it intuitively less consistent with the dynamic complexity theory, where the system’s entropy is essentially assumed as irreversible in time. The current work attempts to model the lexical complexity process of
CGWR with time series models which are generically more consistent with the dynamical and ever-increasing nature, at least in an average sense, of the entropic process of homogeneous texts. Accordingly, the nonstationarity of a time series foretells the emerging state or unsaturatedness of a dynamic complexity system. The interconnectedness and interacting nature of the dynamical complexity system are also naturally explained by the autocovariance of the process at different times. What is further, but not least, is that time series modeling can easily handle jumps, phase changes, and other nonlinear phenomena in a dynamic system.

Time series methods and modeling (Shin, 2017; Box-Steffensmeier et al., 2014; Lameli, 2018) had been developed in the early 1980s, mostly for analyzing highly randomized socioeconomic data. An early example of time series approach in language studies goes back to Pawłowski (1997) for quantifying the underlying basis of rhythms in coded Latin hemameter syllables, but applications of time series analysis in humanity and social sciences, particularly in linguistic studies, are comparatively scarce. One major reason for the relatively scarce application of time series analysis in the linguistic field is the lack of datasets with high regularity and large sample size. But such a hurdle was gradually overcome with the ever-faster advancement in corpora development, automated testing, natural language processing, data mining, and machine learning technologies. Time series analysis approaches were recently applied, for instance, by research studies regarding word formation (Lameli, 2018) and the adoption and propagation of linguistic innovation (Blytheand Croft, 2021). Michel et al. (2011) conducted a time series analysis on cultural trends historically reflected in English books by drawing insights from diverse perspectives such as the adoption of technology, lexicography, and grammar evolution. Tarnish (2018) applied time series analysis to study how the use of modals and semi-modals in British and American English has evolved differently in the past 200 years. Koplenig (2017) applied the method to identify the correlated changes in chronological corpora.

The significance of the current study is two-fold. First, the present study empirically demonstrates a unique angle to understand the interconnectedness between the linguistic features such as lexical complexity of institutional writings and the physical, socio-economic environment by which such writings are conditioned. Specifically, the results show that the lexical complexity of the CGWR, to date, is not a stationary process, implying that the texts are still in their initial states from entropic production and evolution point of view. Thus, the study evinces the ongoing ever-increasing complexity of China’s social, cultural, and economic realities, despite the dramatic changes and reformations undergone in recent decades. Methodologically, the current study provides a long-wished quantitative and analytical support to the theorists in dynamic complexity research by demonstrating that time series analysis is a conceptually valid and computationally rigorous approach to model the dynamic complexity phenomena in the homogeneous texts of institutional writings. Compared to the structural equations, the currently proposed approach is less challenged with variable identification problems and more robust in capturing the structural changes or other irregular patterns that could be more of an unexceptional occurrence with more incoming data of CGGR in the future. In sum, the approach and findings posit a methodological baseline to the corpus linguistics modeling in general where the datasets are of time series nature.

In the following section Two, a more detailed literature review focuses on the pertinence to entropic modeling of the homogenous text of CGWR. Section Three focuses on the data description of the CGWR texts and the methods for successful modeling purported in the current study. Section Four provides the primary analysis, results, and discussions on
the model validity and comparison with existing models. Finally, section Five presents the concluding remarks and future directions.

2 Related literature

2.1 Homogeneity analysis

Homogeneity is a multi-field concept, the reference of which could be context-dependent. For instance, in music, homogeneity could mean how musical scores are stylistically similar from the music composition and appreciation perspective. But it could also tell how equal the opportunities in the music production industry are to the different backgrounds of musicians concerning, for instance, social class, gender, and ableism (Li et al. 2010). In ecology, species homogeneity is used, in duality with species richness, to measure how diversified an environment encompasses and sustains different species (Newbold et al. 2015). In corpus linguistics, homogeneity refers to the extent to which a group of texts shares much enough in common in one or more linguistic features, e.g., genre, prosody, lexical spectrum, and lexical distance at macroscopic level; or the choice of word, punctuation, tense, directness or indirectness, or the use of questions at local levels (Scott and Tribble, 2006; Nielbo et al., 2019; Kilgarriff, 2001; Kilgarriff and Grefenstette, 2003; Denoual, 2005; Sahlgren and Karlgren, 2005). As such, texts produced by the same author at different times could be reasonably rated as homogenous from a stylometry analysis point of view, even though the writing styles of the same author may still show variations over time. With such definition, the writings by Thomas Hardy are homogeneous when the author-ship of texts is underscored. However, the writings of Thomas Hardy, Jane Austin, Charles Dickens, Virginia Wolf, and Mary Shelley, are also collectively homogeneous when highlighting that these writings are all of the same genres of classical English fiction.

Homogeneity is essential for text classification. But perhaps more importantly, it accredits the models and analysis adopted in corpus linguistic research by filtering out the disturbing or even distorted information from the raw data and increases the probability and reliability of any proper pattern under seeking. For instance, from a statistical regression perspective, the challenges usually arise less from exploring new variables but more often from the confounding effects of many potential factors. Consequently, the non-homogeneity of factors is empirically demonstrated as one of the most frequent reasons leading to weak explaining powers of the conclusions made in linguistic and other social researches (Aguinis, 2003; Jang et al. 2015). With this rationale, homogeneity has increasingly attracted scholarly attention in linguistic practices. For instance, Kilgarriff and Grefenstette (2003) and Sahlgren and Karlgren (2005) relied on homogeneity to devise the metric of text distance for cross-corpora comparison studies. Denoual (2005) heavily highlighted the homogeneity of participants in the experiment to support the validity of the modeling results. Additionally, many research studies in frequency-based corpus studies premised their analysis on the homogeneity of the texts under review (Johansson, 2008; Voleti et al. 2019).

2.2 Level of lexical complexity

One of the core properties of homogeneous texts that may evolve is the level of complexity – lexical complexity in particular, as focused in the current study. From the information
filtration and processing perspective, quantitatively primarily by entropy, lexical complexity refers to the degree of complicatedness the lexises are organized under various linguistic constraints (Hales, 2016; Lowder et al. 2018; Zhang, 2016). However, the notion of “complexity” in linguistic studies, particularly from the language acquisition and language proficiency perspectives, is somewhat complex and much-debated in its own right. According to Crossley et al. (2011) and Jarvis (2013), lexical complexity and lexical sophistication bear the same meaning to some extent. In contrast, the latter possesses more tinge of the high usage of advanced or more words but less of the combinatorial and configurational complexity as focused by the current study. On the other hand, research studies such as Housen et al. (2012) umbrellaed lexical complexity under the concept of linguistic complexity, including syntactic complexity, mainly referring to the organization of the text at phrasal, sentential, and clausal levels. Such taxonomy of linguistic complexity has also been echoed by research from cognitive tasks and assessment designs (Robinson, 2007; Housen et al. 2012).

Another important contrast to draw is between lexical complexity and lexical richness (LR) or lexical diversity (Malvern and Richard, 2012; Daller and Xue, 2007), where LR entails a broader notion of a language user’s lexical proficiency. Such proficiency is hypothesized to symbolize how skillful the user is in maneuvering or manipulating different words (e.g., synonyms) when expressing the meaning with subtle differences. More specifically, lexical richness refers to the variability of word use profiled by a speaker or writer (Jarvis, 2013; Zhang and Wu, 2021). Under this umbrella, a variety of lexical richness measures was invented—each with pros and cons in assessing the level of LR of a text quantitatively, typically using frequency-based calculations. One of the oldest and most widely used LR measures is the Type to Token Ratio (TTR), where type refers to the number of unique words used in a text, while token refers to the total running words of the text. While simple, TTR has been demonstrated by substantial empirical studies to outperform many sophisticated LR measures, such as Sichel, Honor, or Yule (Zhang and Wu, 2021). There have been modifications and improvements over TTR, namely, LogTTR (Herdan 1960), rootTTR (Guiraud, 1960), D (Malvern and Richard, 2012), for instance. These attempts were invested mainly to overcome the length effect of TTR. In particular, several studies have shown D as more potent in describing and predicting the LR of speakers, although the computing of D, especially the simulation-based computation as required by the original definition of D, is more costly. Given the marginal benefit of the obvious more tedious and less intuitive formulas, it is a question whether highly algebraically sophisticated formulas yield a more insightful understanding of LR. Zhang and Wu (2021) recently reported a performance ranking of eighteen known LR measures for lexical differentiation of L1 and L2 speakers. Although not the best, TTR is shown as a reliable predictor for lexical proficiency. Instead of deep exploration of frequency-based formulas, some emerging LR studies are trying to leverage the configurational notion of diversity to enrich the understanding of LR, as hinted by diversity phenomena in ecology and biology (Javis, 2013).

In contrast to the lexical manipulation skills of language users, lexical complexity focuses on how complex the words are produced and aligned in the text. In other words, how unpredictable are the next word or string of words given the known lexical information up to the current word in the text. From cognitive information retrieval and transmission perspective, the more unpredictable a system is, the more randomly organized its constituents, and hence the higher level of complexity of the system. Accordingly, the higher level of complexity of a text, the more memory and information processing capacity are required to store, retrieve, and process the information contained in the text. Thus, the
major difference between a simple and a complex text is the probability of accurate prediction. Given a simple text, it is easier to predict the contents and meaning when an initial segment of the text is known, whereas, for a complex text, such prediction is more complicated and less accurate.

2.3 Entropy measure of complexity

Entropy, a total frequency spectrum accounted measure, has been shown as a natural and robust measure for the level of complexity of any close or open system with interconnected and interactive constituents. Let the system be composed of \( T \) different components out of \( N \) components in total, where each of the \( i \)th components shares \( p_i \) percentage in the total constituents, then the entropy of the system is defined as

\[
\text{Entropy}(S) = - \sum_{i=1}^{T} p_i \ln(p_i)
\]

where the negative sign is to ensure the positivity of the quantity. In the context of corpus linguistics, \( T \) corresponds to the number of types, and \( N \) corresponds to the number of tokens of a text.

The seminal introduction of the above entropy formulation goes to Shannon (1951). The entropy was initially introduced to measure the information load from the information processing perspective. For this reason, the entropy defined as such is called Shannon-entropy, in contrast to other versions of entropies that have been developed ever since. Nevertheless, the notion of entropy was widely used in the science community quite long before Shannon. Notably, entropy was originated from thermodynamics as a metric defining the level disorderedness of a system; in particular, how the degree of disorderedness of all the molecules in a closed space will change over time as the temperature of the space increases or decreases (Ramshaw, 2018). The higher the entropy of a system, the less likely to predict with precision the state of positions or velocities of all the particles in the system.

Because it includes all component information in one compact format and its easy computation, entropy since then has been adopted in many other fields, including the well-known Shannon entropy form information communications. In the list of other applications, environmental entropy refers to species richness and balance, thus the sustainability of an ecological system. Furthermore, entropy is an essential notion in cell biology, defining the degree of variability or uncertainty, and therefore the differentiation potency towards diversified biological forms in cellular dynamics (Gros, 2011; Sethna, 2006). One crucial remark is that a more generalized version of entropy has been developed, namely, Renyi entropy, which includes Shannon entropy as a particular case (Acharya et al. 2017). Specifically, Renyi entropy introduced a parameter alpha in the definition that

\[
\text{Renyi Entropy}(S) = \frac{1}{1 - \alpha} \ln \left( \sum_{i=1}^{T} p_i^\alpha \right)
\]

where \( p_i \) has the same definition as that for Shannon entropy. It is easily seen that when \( \alpha = 0 \), Renyi entropy reduces to Shannon entropy. Recent researches show that, in some scenarios, Renyi entropy of nonzero order could become desirable when the sample size becomes large.
2.4 Dynamical complexity theory for linguistics

The theory of dynamic complex system (Byrne and Callaghan, 2013; Larsen-Freeman and Cameron, 2008; Guastello et al. 2008; Gros, 2011) is one of the most ontologically important modern treatises of any physical or social system, aiming at understanding the movement or behavioral pattern of the system in both spatial and time domain, most often from a macroscopic perspective. Such treatise is fundamentally different from classical sciences, symbolized by a small number of components with limited dependent variables and free of external influence. Even facilitated with the ever-increasing computing technologies, a complete depiction of the whole microscopic structure and movement of a dynamic complexity system is yet not possible, so probabilistic modeling and data fitting are widely adopted to uncover the evolutionary patterns a statistical sense of the system at large.

The system is admittedly dynamic and complex because the number of composite entities is often enormous. Secondly, the system is complex because of the interaction between the components. Even a pair-wise interaction between two components in the system may result in an incomprehensively high number of evolution routes momentarily due to the effect of combinatorics. Additionally, the dynamical complexity is naturally associated with the non-linear changes allowed by the system. Many natural entities may behave nonlinearly and chaotically given sufficient time and external influences, starting from a simple linear rule at a local level. Last but not least, the reason for the dynamical complexity is the almost ubiquitous self-organization phenomena widely observed at both the cosmos level and the microscopic level. Putting together, dynamic complexity is a fundamental mechanism governing any system, where the components of subordinate entities are significant in number, allowing for spontaneous interactions, self-organization, nonlinear changes in time, and exchange of energy or substances or information in between and with external environment.

Dynamic complexity theory has been well elaborated in language and linguistic research, and empirically tested by numerous studies spanned from L1 and L2 acquisition to corpus linguistics. According to Larsen-Freeman and Cameron (2008), for instance, language entities such as lexical units, grammatical rules, language input from daily life, and so on, are all interactive components of a linguistic system when a proper context or environment is referred. These linguistic constituents, analogous to biological cells in life sciences, are interconnected, constantly interacting, subject to external sociolinguistic influences and functional constraints, and may also reshape their evolution through spontaneous self-organization. Such a framework has been applied to explain lexical skill development (Verspoor et al. 2011), language attrition in German (Hopp and Schmid, 2013), and English dialect variation across different regions (Clopper and Smiljanic, 2011). A distinct lineage of research motivated by dynamic complexity is framed as evolutionary linguistics, underscoring the self-adaptive nature of languages (Lee and Schumann, 2003; Croft, 2008; Steffensen and Fill, 2014).

3 Method and procedure

The corpus used for the current study is the CGWR from 1954, when the first report was delivered, to 2000, excluding the years when the report was not delivered. CGWR has always drawn wide attention from various perspectives, including academics and businesses, due to its significant socio-economic importance, as indicated in the Introduction section. The archives of CGWR texts are publicly available at the website www.gov.cn. Table 1 lists the main tags of the texts as a time series, the basic lexical statistics, and the
main characteristics from a functional discourse point of view. To have a further glimpse of the data, Fig. 1 plots the types and tokens of each CGWR text up to 2020, which corresponds to a sample size of 546,015 Chinese words in total.

A quantitative analysis of word frequency may further illustrate patterns of keywords of each CGWR text at different times. Zhang (2016) provides a diagram of the first three most frequently appeared content words to demonstrate the word frequency changes in CGWR as a dynamic complexity system. A more visual aiding plot of word cloud could be drawn for similar purposes. Figure 2 is such a plot displaying the key content words in proportion to their frequencies in the CGWR texts of 2007 and 2008, where one may sense the topical continuation and variation through the frequency changes in the top-ranking nouns such as “development” and “system” or verbs “improve” and “promote”, hence an inkling to the evolution of the lexical networks therein profiled. The dynamicity of such a complex system as visualized is more rigorously evaluated using Shannon entropy. As discussed in

| Table 1 | Descriptive statistics of the CGWR texts in the current study |
|-----------------|-------------------|-----------------|
| **Time series tags** |  |  |
| Total number of texts | 52 |  |
| Frequency of delivery | Annual |  |
| Serial correlation | Yes |  |
| Type | 2212 |  |
| **Average lexical statistics (in Chinese words)** |  |  |
| Token | 10,500 |  |
| Entropy | 6.621 |  |
| Genre | Institutional writing |  |
| Domain | Socio-economic, public, world affairs |  |
| **Discourse parameters** |  |  |
| Register | Public archive, internet |  |

![Fig. 1 Types and tokens in words for the corpus under study](image)
Literature Review, entropy is a proper and effective measure of the level of information complexity of a dynamic system, although probably not necessarily an optimal choice for lexical richness when the focus is lexical proficiency.

Zhang (2016) used the CGWR texts up to 2011 to construct an exponential structural equation model to explain the macroscopic entropic pattern CGWR. As pointed out in the concluding remark of the paper, the model fitting and forecasting accuracy will be substantially affected if the entropic process undergoes phase change phenomena. To potentially harness such drawback of the structural equation approach, among other methodological considerations, the current study uses times series analysis approach to model the entropic process of the CGWR texts using the updated data (1954–2000). The setup of the time series analysis, assuming that the text’s entropic value at the current time is a lagged function of the entropic values at previous times, does not allow for reversibility of time in the definition. To run a time series model, an initial value of the process should be specified before the following sequence of the series can be recursively determined. To highlight the major steps for time series analysis, stationary should be examined upfront. Technically there are also non-time-series approaches for handling nonstationary data, which is not the scope of the current study. Fortunately, there are methods in the time-series domain to transform a large class of nonstationary processes into stationary, notably by detrending or difference operation, for instance. Once the dataset is given, several tests are available to test the stationarity condition. They could be done through autocorrelation function (ACF) and partial autocorrelation function (PACF) plots for a gross graphical check. Stationarity could also be quantitatively assessed by statistical tests, including the augmented Dicky-Fuller (ADF) test, Phillips-Perron (PP) test, and Kwiatkowski-Phillips-Shin (KPSS) test. Some research studies also use Ljung-Box (LB) test for an additional measure convincing the stationarity, although LB is used primarily for testing autocorrelations.

Once a process is proved stationary, the next step is to choose a specific model, such as an ARMA model with an adequate number of lags and proper coefficient parameters for each lag. There could be multiple ways to determine the best model depending on the research context. The minimization of information statistics, including Akaike information criterion (AIC) and Bayesian Information Criterion (BIC), is often a widely used metric to determine the optimal model, again usually in the ARMA class. After model selection, the next major task will be model diagnostics in terms of well-established statistics such as MSE, F-statistic, p-value, and R-squared values. Lastly, the validity of a time series model usually scrutinizes an out-sample forecasting analysis. The specific procedure

Fig. 2  Word cloud example of the CGWR for the years 2007–2008
typically involves splitting the available data into two sets—a training set for model fitting and the other as a “pseudo-future” set to benchmark how much the model forecasted values would deviate from these “pseudo-future” values at hand.

Lastly, a word frequency analysis tool is needed for computing the raw frequency distributions of characters and words in each text of the CGWR corpus. The task was mainly carried out using The Computerized Language Analysis Program (CLAN) — a publicly available word frequency analyzer developed and maintained at Carnegie Mellon University (CMU)’s Child Language Data Exchange System (CHILDES). It is increasingly recognized by various research studies such as second language acquisition (SLA) and corpus linguistics. One may refer to MacWhinney (2007a, b), for instance, to get familiar with the tool. In contrast, there are similar tools capable of Chinese word segmentation and frequency analysis, including those developed at Stanford University and Tsinghua University. After the raw frequency distributions for all types and tokens of each text are obtained, model estimation and diagnostics can be conducted in programming languages, such as Python.

4 Results and analysis

As underscored in the Method and Procedure section, stationarity should be examined upfront in the data analysis and model selection. But for better contextualization, the time series of the entropy is first plotted by the following Fig. 3, shown together with which are the linear fit and the confidence interval (the light blue band) for the mean response at a significance level alpha = 0.01. In addition, the cubic-spline interpolation of the data is also plotted, showing an irregular cyclic pattern with heteroscedastic variances. The Discussion section provides a more detailed discussion from a dynamical sociolinguistic perspective. Finally, Fig. 4 gives a graphical check of stationarity using the ACF plot. The series is concluded as nonstationary because the ACF values decay rather slowly at a roughly linear rate than a geometric rate characterized by a stationary process.

Consistent with the graphic judgment, the following Tables 2 and 3 report the ADF statistics of the data for the three different significance levels and various choices of the characteristic formulation of the unit roots. The maximum number of lags included in the test is set as 12 (larger than what the ADF test conventionally requires for the current context, where such a conventional threshold is usually set as twelve multiplied by the fourth root of the one-hundredth of the number of the observations). Minimization of AIC information is set as the default criterion to choose the optimal number of lags to be included. Given the testing power of ADF may be biased by the sampling context and sampling size, a further comparison of unit root test using more influential statistics are provided by the following Table 2. As shown, there is no statistical evidence to reject that null hypothesis, at a significance level of 0.01, that unit root is present in the process under study. The test was conducted with constant not assumed for stationarity. On the other hand, the other choice of such condition, i.e., constant allowed for stationary, does not change the conclusion with ADF statistic = -1.4300, p-value = 0.5678, and corresponding critical values (CV) = -3.5689, -2.9214, -2.5987 to reject the null hypothesis, respectively for significance levels of 0.01, 0.05, and 0.10 under AIC information characterization for unit root. Test with BIC information characterization for unit root pointed to the same conclusion.

Table 2 and the above summary show that the series has unit roots unless a hidden trend, linear or nonlinear irrespectively, is allowed. Correspondingly, the process, at optimality, is
Fig. 3 Plot of the CGWR entropic process and the linear trend fitting

Fig. 4 ACF plot of the CGWR entropic process

Table 2 Unit root test statistics to the entropic process in original series (maximum number of lags included: 12; test mode: nc)

| Unit root criterion | ADF statistic | p-value | CV (significance level 0.01) | CV (significance level 0.05) | CV (significance level 0.10) |
|---------------------|--------------|---------|-----------------------------|-----------------------------|-----------------------------|
| AIC                 | 1.9491       | 0.9887  | −2.6150                     | −1.9479                     | −1.6122                     |
| BIC                 | 0.7414       | 0.8746  | −2.6119                     | −1.9475                     | −1.6124                     |
intrinsically an integrated autoregressive moving average process (ARIMA) with the minimum integration order of one. Thus the process is not possibly modeled by a stationary ARMA model of any finite orders with a reasonable fit. In other words, the entropic process of the CGWR texts is not stationary, consistent with the fundamental assumption of the dynamic complex system point of view, where the lexical complexity of institutional writings, at emerging stages, is an entropic increasing process. A better understanding of the upward entropic pattern of CGWR texts should go beyond language use in itself as, from a dynamic sociolinguistic perspective, language and society are inseparable. Specifically, the observed entropy-increasing pattern of the CGWR texts manifests, in duality, the dramatic social and economic transformations and reformations at both macro and mundane aspects of China in the past many decades. In such interconnectedness between the linguistic domain and various exogenous sociocultural factors, coupled with probabilistically infinite reformations of such connections, other evolutionary phenomena, including possibly phase changes or irregular cyclic patterns, are linguistically observed in CGWR.

Next, a difference operation was applied to the original data hoping that a stationary time series could be attained with such transformations. Let $\Delta y_t = y_t - y_{t-1}$, the first-order difference of the entropy process of $y_t$, for $t = 2, 3, \ldots, T$, be presumably modeled by an ARMA $(p,q)$ model in the form of

$$
\Delta y_t = \beta_0 + \beta_1 \Delta y_{t-1} + \cdots + \beta_p \Delta y_{t-p} + \gamma_1 \epsilon_{t-1} + \cdots + \gamma_q \epsilon_{t-q} + \epsilon_t
$$

where $\epsilon_t$ is assumed to follow a normal distribution with a mean of 0 and a variance of $\sigma^2_t$. Accordingly, a stationarity check was performed on the $\Delta y_t$ series and the numeric statistics of the test are reported in the following Table 2, in addition to the ACF and PACF plots presented in Fig. 5. The geometrically fast decay pattern in the ADF and PACF plots and the significant enough absolute values of ADF statistics ($<-3.5778$ in real values) tend to affirm the stationarity of the $\Delta y_t$ process at a significance level of 0.01. The test was conducted with constant mode for a stationary check. However, lifting the constant assumption does not change the conclusion with ADF statistic $=-12.4857$ (far lower than the CV $=-2.6119$ at significance levels of 0.01), and corresponding $p$-value $=8.8206e-23$, with AIC set as the information criterion for unit root characterization. The test statistic is significant enough to reject the null hypothesis of the existence of unit root. The conclusion remains the same when the BIC information criterion is used.

Following the rule of thumb for empirically determining the optimal orders for the ARMA process (Box-Steffensmeier et al. 2014), one may select $p=2$ and $q=3$ for the model as stated in the above Eq. (1). To double confirm the optimality of the order of the model decided firstly by the empirical rule of thumb based on ACF and PACF plots,
we run the same optimization algorithm for all the possible combinations of \((p, q)\) for \(0 \leq p \leq 15, 0 \leq q \leq 15\) against the AIC values. Indeed the minimum of the AIC function is reached at roughly \(p=2\) and \(q=3\).

A maximum likelihood procedure was carried out to fit the \(\Delta y_t\) series with the specified ARMA(2, 3) model to estimate the parameters for the model using a maximum number of iterations set at 100,000. The fitted curve of the \(\Delta y_t\) and the original series of \(\Delta y_t\) are comparatively plotted in the following Fig. 6. The estimated parameter values and the in-sample goodness of fit were tabulated in Table 4. To make it a fair comparison in terms of in-sample forecasting accuracy, we use Eq. (1) to convert the differenced series back to the original entropic values, noting that \(y_{t+1} = y_t + \Delta y_t\) and \(\tilde{y}_{t+1} = y_t + \Delta \tilde{y}_{t+1}\). The corresponding statistics and significance for the overall model and each parameter are provided in Table 4. As demonstrated, the \(p\)-values for the model and each coefficient parameter are significant, validating the choice of ARMA type of model and the order of the model. Overall, the optimized ARMA(2,3) model has a satisfactory fit with \(F\)-statistic = -7.0117, \(p\)-value = 6.8959e-10. MSE = 0.0168 and R-squared = 0.6830 further show the goodness of fit. Given the high nonlinearity of the entropic dynamics of CGWR texts over time, such R-squared value is at a reasonable and acceptable level of model fitting (see Shin, 2017; Wittink, 1988, for discussions on a reasonable benchmark of R-squared values in social sciences studies).

In addition to the above in-sample forecasting performance assessment, we also conducted the out-sample forecasting evolution based on the empirical rule of thumb of model validation using three to one split of the full sample into the training and prediction sets. Four prominent measures of out-sample forecast performance, namely, MSE, MAE, MAPE, and Theil-statistics (Theil), were calculated for the 1-step, 2-step, and 3-step forward prediction. The selected ARMA(2,3) model has demonstrated sufficient out-sample forecasting power, as demonstrated in Table 5. At the same time, it might be intuitive to conjecture that the longer steps into the future, the less accurate the prediction, there could be local inconsistencies to such notion. However, as demonstrated, the foresting accuracy for the 2-step forward prediction is higher than that for both 1-step forward and 3-step forward predictions.

Lastly, a normality check was conducted to evaluate how well the residuals between the fitted values and the sample values of the differenced entropic process fit the normal distribution as part of the module assumptions. Figure 7 presents the QQ plot of residuals, where the normal quantiles were well followed overall with very scarce discrepancies.
It is again consistent with the assumption of the potential phase changes and other nonlinear phenomena in the entropic process of CGWR texts, despite the fact highlighted by the current study that ARIMA models are comparatively more robust in fitting such dynamical processes. Normality may also be demonstrated with several pertinent test statistics, notably Shapiro–Wilk (SW) test, the Lilliefors test, and Anderson–Darling (AD) test. Specifically, the residuals’ SW, Lilliefors, and AD statistics are 0.9877, 0.0639, and 0.2636, respectively. The corresponding $p$-values for these statistics are 0.8748, 0.8702, and 0.7026, all confirming the normality of residuals at the significance level of 0.05.

![First-order difference process and the fitted curve](image)

**Fig. 6** First-order difference process and the fitted curve
Overall, the entropic process of the CGWR texts exhibits ARIMA type of nonstationarity with integration order one. An ARMA(2,3) model has been estimated and demonstrated to best fit the first-order difference of the original entropic process using the data by 2020. The validity and effectiveness of the proposed model have been consistently justified by the pertinent statistical procedure and results, including the reported unit root test, ACF and PACF plots, information test, and the in-sample and out-sample forecasting accuracies.

5 Discussion

The first discussion concerns the cyclic pattern, although irregular in a sense, of the entropic process as observed along with the linear trend. First, we would like to remark that cyclicity is not a new language phenomenon. For instance, Bouzouita et al. (2019) documented a wide range of cyclical styles such as diachronic connections with different typological probabilities regarding, for example, markers of sentential negation and tense inflection on verbs. In the context of the current study, however, we argue that one should appeal to a broader socio-economic context, instead of focusing on language use alone, to explain better the trend-surrounding cyclic pattern of the lexical complexity in the CGWR texts. From a dynamic complexity system perspective, the mechanism of the pattern is generically associated with the interdependence and interconnectedness between the texts and the external resources, where the present level complexity depends critically on the previous levels. One particularly influential factor having a cyclical feature, by design, goes to the so-called Five-Year Plan, a comprehensive blueprint outlining the government’s policy priorities, together with specific or qualitative evaluation metrics. The identified social, economic, and cultural significance initiatives are reviewed and possibly adjusted in the annual CGWR. The physical cycles of such topically significant projects, from inception to implementation and completion, likely affect the
configurations of the linguistic counterparts of the CGWR, hence the observed quasi-cyclic pattern in the evolution of the lexical complexity. Not surprisingly, apart from the prominent cyclic elements inherited from the socio-economic world, other latent variables possibly contribute to lexical complexity dynamics, making exact cycles unattainable. Unforeseeable environmental or natural events may also cause changes to the overall pattern. The most recent impactful example goes to COVID-19, the main reason that caused the postponed delivery of the CGWR as of the year 2020, and possibly also the substantial decrease in the length of the report.

The following discussion concerns comparing the current approach and the structural equation approach. As mentioned in the Introduction section, the structural equations are usually invariant in the time direction. In contrast, the ARIMA model introduced in the current study is timely irreversible by design, a property naturally consistent with the irreversible nature of the entropic evolution of dynamic complex systems. For comparison, we reran the model selection and estimation for the exponential model in Zhang (2016) using the updated CGWR data up to 2020. First, the empirical results demonstrated that the difference between the two models’ statistics in terms of goodness of fit is slight. Second, from the nuanced difference between the model statistics, it is generally observed that the structural exponential model tended to perform better for the in-sample fit. In contrast, the ARMA model provided a better out-sample fit when the number of prediction steps is not significant. Notwithstanding, we do not wish to generalize such observations regarding the relative accuracy of the exponential and the AMRA models. A resolute conclusion to such generalization is likely not realistic as it may largely depend on the research context and the data-generating mechanisms.

Another comparison is between the currently proposed ARMA model and a possible GARCH type model. Although sampling autocorrelations does not make specific the choice of GARCH models, trials are worthwhile if one is concerned by whether this may imply further heteroscedasticity and variance clustering phenomena, which was confirmed with a formal arch effect test based on Lagrange multiplier statistics. Thus Garch type models are also potentially suitable for the differenced data, although cautions should be added in terms of the data’s low frequency and small size. To this end, a series of model selections and comparisons were done for the GARCH type models as for the ARMA (p, q) model selection explained in the above section. Then the corresponding time series were generated using simulations methods, with the selection criterion being to minimize the MSE. Setting the maximum simulations being 100,000, we derived the optimal GARCH model as GARCH (3, 1), and the corresponding R-squared value is 0.403, which is lower than that of the ARMA(2,3). This result demonstrated that a GARH model focusing on modeling the variance clustering does not yield better performance in fitting the up-to-date data of the CGWR. Finally, of course, one may further explore the possibility of ARMA-GARCH combined models. However, we opt to leave it for further directions given the concerns of the overfitting and interpretability of the parameters with the limited sample size.

6 Conclusion

Homogeneous texts, particularly those of the same individual or institutional authorship produced over different times, provide uniquely helpful resources for various linguistic research, particularly from a sociolinguistic perspective, as the discourse of such institutional writings are both socially conditioned and constitutive. The dynamical complexity theory offers a well-fit framework to analyze such homogeneous texts to unveil
prominent linguistic patterns and insights with potent socio-economic implications. The current work analyzed the stylish features of the lexical profiles of the CGWR published from 1954 to 2000 using the Box-Jenkin approach. The results show that the entropic process of CGWR texts is not stationary from time series analysis perspective. Instead, an ARIMA process characterizes it, the stationarity of which requires an order one difference operation on the actual process. Further noticeable entropic features, such as the irregular cyclic pattern around the increasing linear trend, have also been characterized and explained in the dynamical complex sociolinguistic perspective. In particular, the government’s strategic socio-economic initiatives, such as education enhancement or countryside preservation, are inseparable factors to account for the stylish features exhibited in the CGWR texts. As such, the current study provides a unique angle for researchers interested in understanding the social dynamics of China from sociolinguistic text analysis.

Theoretically significant, the dynamical complexity narrative is often challenged with the lack of pertinent and quantifiable models for the diversified field research in corpus linguistics. Despite emerging research studies with elements of structural equations and time series models as surveyed in the Introduction and Related Literature sections, such quantitative treatise specific to dynamic complexity theory is still scarce. With a rigorous implementation of time series modeling, the current study has exemplified the possibility of making the dynamical complexity a methodologically well-equipped framework instead of a hypothetical fad. More concretely, a thorough ARIMA modeling, including model identification, estimation, diagnostics, and assessment, was conducted based on the updated texts of CGWR. The validity and effectiveness of the currently proposed model were positively demonstrated, together with the pros and cons analysis compared to the more classical structural equation approaches. The overall goodness of the proposed mode of ARMA (2, 3) in particular is equivalent or slightly more substantial than the structural models of exponential type depending on the context of statistics. The current approach is strongly advocated with the extended horizon model fitting performance yet to be tested with future data. It naturally fits the dynamics complexity setting of the homogeneous texts’ entropic or information evolution process. With various toolkits of time series analysis facilitated with upgraded computing technologies, the current approach appears robust enough to accommodate the potentially high complexities of the incoming data.

As noticed in the Discussion section, whether removing hidden irregular cyclic trends may produce a better model, in general, remains an open question. Similarly, the GARCH type models, designed to model variance clustering in the process, in particular, are not extensively explored by the current study. Although there are noticeable autocorrelations in the series, variance heteroscedasticity and variance clustering are not critically apparent. Hence, GARCH type modeling does not yield a better model fit. Among others, overfitting and interpretability are the main concerns.

Moreover, given the so far limited sample size of the CGWR texts, the usefulness regarding GARCH models—typically applicable to high frequency and extensive sample data—could be inconclusive. However, these value-adding directions are worthwhile for future explorations. Another recommendable research direction is to apply the models of the current study to other types of homogeneous texts. This would generate more empirical verifications to the current modeling and provide more complete mappings in terms of the universality of the dynamical complexity system theory.
Acknowledgements The author is grateful to the anonymous referees and the journal editors for the valuable comments and suggestions in improving the earlier manuscript.

Funding The authors have not disclosed any funding.

Declarations

Conflict of interest The author has no competing interests to declare related to this article.

References

Acharya, J., Orlitsky, A., Suresh, A.T., Tyagi, H.: Estimating renyi entropy of discrete distributions. IEEE Trans. Inf. Theory 63(1), 38–56 (2017)
Aguinis, H.: Regression analysis for categorical moderators. Guilford Press (2003)
Blythe, R.A., Croft, W.: S-curves and the mechanisms of propagation in language change. Language 88(2), 269–304 (2012)
Bouzouita, M., Breitbart, A., Witzenhausen, E.: Cycles in language change. Oxford University Press (2019)
Box-Steinmesier, J.M., Freeman, J.R., Hitt, M.P.: Time series analysis for the social sciences. Cambridge University Press (2014)
Boyd, S., Fraurud, K.: Challenging the homogeneity assumption in language variation analysis: Findings from a study of multilingual urban spaces. In: Auer, P., Schmidt, J.E. (eds.) language and space: an international handbook of linguistic variation, pp. 686–706. Mouton De Gruyter (2010)
Byrne, D., Callaghan, G.: Complexity theory and the social sciences : the state of the art. Taylor and Francis (2013)
Clopper, C.G., Smiljanic, R.: Effects of gender and regional dialect on prosodic patterns in American English. J. Phon. 39(2), 237–245 (2011)
Croft, W.: Evolutionary linguistics. Annu. Rev. Anthropol. 37, 219–234 (2008)
Crossley, S.A., McNamara, D.S.: Shared features of L2 writing: intergroup homogeneity and text classification. J. Second. Lang. Writ. 20(4), 271–285 (2011)
Crossley, S.A., Salsbury, T.L., McNamara, D., Jarvis, S.: Predicting lexical proficiency in language learner texts using computational indices. Lang. Test. 28(4), 561–580 (2011)
Daller, H., Xue, H.: Lexical richness and the oral proficiency of Chinese EFL students. In: Daller, H., Milton, J., Trefers-Daller, J. (eds.) Modelling and assessing vocabulary knowledge, pp. 150–164. Cambridge University Press (2007)
Denoual, E.: 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 (2005)
Gro’s, C.: Complex and adaptive dynamical systems: a primer. Springer (2011)
Guastello, S.J., Koopmans, M., Pincus, D.: Chaos and complexity in psychology: the theory of nonlinear dynamical systems. Cambridge University Press (2008)
Hale, J.: Information-theoretical complexity metrics. Language Linguistics Compass 10(9), 397–412 (2016)
Herdan, G.: Quantitative linguistics. Butterworth (1960)
Hopp, H., Schmid, M.S.: Perceived foreign accent in first language attrition and second language acquisition: the impact of age of acquisition and bilingualism. Appl. Psycholinguist. 34(2), 361–394 (2013)
Housen, A., Kuiken, F., Vedder, I.: Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA. John Benjamins Publishing (2012)
Jang, E.E., Cummins, J., Wagner, M., Stille, S., Dunlop, M.: Investigating the homogeneity and distinguishability of STEP proficiency descriptors in assessing english language learners in Ontario Schools. Lang. Assess. Q. 12(1), 87–109 (2015)
Jarvis, S.: Capturing diversity in lexical diversity. Lang. Learn. 63, 87–106 (2013)
Kilgarriff, A.: Comparing corpora, International. J. Corpus Linguistics 6(1), 1–37 (2001)
Kilgarriff, A., Grefenstette, G.: Introduction to the special issue on the web as corpus. Comput. Linguist. 29(3), 333–348 (2003)
Komachali, M.E., Khodareza, M.: The effect of using vocabulary flash card on iranian pre-university students’ vocabulary knowledge. Int. Educ. Stud. 8(3), 134–147 (2012)
Koplenig, A.: A data-driven method to identify (correlated) changes in chronological corpora. J. Quant. Linguistics 24(4), 289–318 (2017)
Lai, Y. A., Zhu, X., Zhang, Y., and Diab, M. T.: Diversity, Density, and Homogeneity: Quantitative Characteristic Metrics for Text Collections. In Proceedings of The 12th Language Resources and Evaluation Conference, 1739–1746 (2020)

Lameli, A.: The replacement of diminutive suffixes in the New High German period: a time series analysis in word formation. J. Historical Linguistics 8(2), 273–316 (2018)

Larsen-Freeman, D., Cameron, L.: Complex systems and applied linguistics. Oxford University Press (2008)

Lee, N., and Schumann, J. H.: The evolution of language and the symbolosphere as complex adaptive system, paper presented at the American Association of Applied Linguistics Conference, VA (2003)

Li, S., Ellis, R., Zhu, Y.: The associations between cognitive ability and L2 development under five different instructional conditions. Appl. Psychol. 40(3), 693–722 (2019)

Liu, Z., Wang, C., Guo, L., Bai, Y., and Wang, J.: LyDAR: A Lyrics Density based Approach to Non-homogeneous Music Resizing. In 2010 IEEE International Conference on Multimedia and Expo, 310–315 (2010)

Lowder, M.W., Choi, W., Ferreira, F., Henderson, J.M.: Lexical predictability during natural reading: effects of surprisal and entropy reduction. Cogn. Sci. 42, 1166–1183 (2018)

MacWhinney, B.: A unified model, in P. Robinson and N. Ellis, editors, Handbook of Cognitive Linguistics and Second Language Acquisition, Lawrence Erlbaum Associates, Mahwah, NJ (2007a)

MacWhinney, B.: The talkbank project. In: Beal, J.C., Corrigan, K.P., Moisil, H.L. (eds.) Creating and digitizing language corpora: synchronic databases. Palgrave-Macmillan, UK (2007)

Malvern, D.D., Richards, B.: Measures of lexical richness. In: Chapelle, C. (ed.) The encyclopedia of applied linguistics. Blackwell Wiley (2012)

Maxwell, O., Payne, E., and Billington, R.: Homogeneity vs heterogeneity in Indian English: Investigating influences of L1 on f0 range. In 19th Annual Conference of the International Speech Communication Association: Speech Research for Emerging Markets in Multilingual Societies, 2191–2195 (2018)

Michel, J.B., Kui Shen, Y., Presser Aiden, A., Veres, A., Gray, M.K., Pickett, J.P., Aiden, E.L.: Quantitative analysis of culture using millions of digitized books. Science 331(6014), 176–182 (2011)

Mu, G.M., Dooley, K.: Coming into an inheritance: family support and Chinese Heritage Language learning. Int. J. Biling. Educ. Biling. 18(4), 501–515 (2015)

Newbold, T., Hudson, L.N., Hill, S.L.L., Contu, S., Lysenko, I., Senior, R.A., Collen, B.: Global effects of land use on local terrestrial biodiversity. Nature 520(7545), 45–50 (2015)

Nielbo, K.L., Baunvig, K.F., Liu, B., Gao, J.: A curious case of entropic decay: persistent complexity in textual cultural heritage. Digital Scholarship Humanities 34(3), 542–557 (2019)

Paap, K.R., Johnson, H.A., Sawi, O.: Are bilingual advantages dependent upon specific tasks or specific bilingual experiences? J. Cogn. Psychol. 26(6), 615–639 (2014)

Pawłowski, A.: Time-Series analysis in linguistics: application of the ARIMA method to cases of spoken Polish. J. Quant. Linguistics 4(1–3), 203–221 (1997)

Ramshaw, J.D.: Statistical foundations of entropy. World Scientific Publishing (2018)

Robinson, P.: Task complexity, theory of mind, and intentional reasoning: effects on L2 speech production, interaction, uptake and perceptions of task difficulty. Int. Rev. Appl. Linguistics Language Teach. 45(3), 193–213 (2007)

Sahlgren, M., and Karlsgren, J.: 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, Argentina (2005)

Scott, M., and Tribble, C.: Textual Patterns: Key words and corpus analysis in language education. Second Language Development: Methods and Techniques, 25–38, John Benjamins, Amsterdam (2006)

Sethna, J.P.: Statistical mechanics: entropy. Oxford University Press (2006)

Shannon, C.E.: Prediction and entropy of printed English. Bell Syst. Tech. J. 30, 50–64 (1951)

Shin, Y.: Time series analysis in the social sciences: the fundamentals. University of California Press (2017)

Smith, J.A., Kelly, C.: Stylistic constancy and change across literary corpora: using measures of lexical richness to date works. Comput. Humantit. 36, 411–430 (2002)

Steffensen, S., Fill, A.: Ecolinguistics: the state of the art and future horizons. Lang. Sci. 41(6), 6–25 (2014)

Tarnish, A.H.: A corpus analysis of changes in the use of British and American English modal verbs in semi-modal clauses. Int. J. English Cul. Stud. 1(1), 1–22 (2018)

Verspoor, M.H., Bot, K., Lowie, W.: A Dynamic approach to second language development: methods and techniques. John Benjamins (2011)

Voleti, R., Woolridge, S., Liss, J. M., Milanovic, M., Bowie, C. R., and Berisha, V.: Objective assessment of social skills using automated language analysis for identification of schizophrenia and bipolar disorder. In 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language, 1433–1437 (2019)
Wittink, D.R.: The application of regression analysis. Allyn and Bacon (1988)
Yao, D., Chen, K.: Gender-related Differential Item Functioning Analysis on an ESL Test. J. Language Test. Assess. 3(1), 5–19 (2020)
Zhang, Y.: Entropic evolution of lexical richness of homogeneous texts over time: a dynamic complexity perspective. J. Language Modell. 3(2), 569–599 (2016)
Zhang, Y., Wu, W.: How effective are lexical richness measures for differentiations of vocabulary proficiency? a comprehensive examination with clustering analysis. Language Test. Asia 11(1), 1–19 (2021)
Zhou, B., Tan, Y., Pan, D.L.: Investigating the validity of different peer groupings in the assessment of English writings. Int. Educ. Stud. 12(12), 55–61 (2019)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.