Physica A 415 (2014) 307–314

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Complex dynamics of text analysis

Xiaohua Ke [a,b,*], Yongqiang Zeng [b,c], Qinghua Ma [a], Lin Zhu [a]

[a] *Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou, China*
[b] *Social Science Key Laboratory of Language Engineering and Computing of Guangdong Province, Guangzhou, China*
[c] *Guangdong Teachers College of Foreign Language and Arts, Guangzhou, China*

HIGHLIGHTS

- We examine changes in the level of text quality using complex network measurements like degrees, clustering conferences, and dynamical networks.
- Complex network features of different text qualities can be clearly revealed.
- Decreasing amounts of nodes in complex networks indicating higher quality of writing.
- Complex network theories can be applied to potential applications of text analysis.

ARTICLE INFO

Article history:
Received 29 March 2014
Received in revised form 12 July 2014
Available online 19 August 2014

Keywords:
Complex network
Text analysis
Text quality series variation
Network feature

ABSTRACT

This paper presents a novel method for the analysis of nonlinear text quality in Chinese language. Texts produced by university students in China were represented as scale-free networks (word adjacency model), from which typical network features such as the in/outdegree, clustering coefficient and network dynamics were obtained. The method integrates the classical concepts of network feature representation and text quality series variation. The analytical and numerical scheme leads to a parameter space representation that constitutes a valid alternative to represent the network features. The results reveal that complex network features of different text qualities can be clearly revealed and applied to potential applications in other instances of text analysis.

1. Introduction

In China, achievement tests[1] are common for measuring the effects that learning and teaching have on individuals. In college, they were adopted in most courses via short writings. Text quality plays a fundamental role in assessing writing activities whereas text features of score series reveal complex dynamical phenomena. For the last few years, nonlinear dynamics of complex network have been employed as statistical or stochastic analytical methods of measurement in financial, electrical, mechanical, thermal, and other types of physical systems [1]. Because the representation and analysis of written texts in terms of complex networks offer a promising and challenging research opportunity for autoscoring writings and translations [2,3], it is worthy of involving mathematical tools usual in the analysis of nonlinear dynamics from text quality series variation.

* Correspondence to: 2# Baiyun Dadao, Guangzhou, China.
E-mail addresses: Ck0900@hotmail.com (X. Ke), 199110513@oamail.gdufs.edu.cn (Y. Zeng), gdqhma@21cn.com (Q. Ma), 1107509010@qq.com (L. Zhu).

[1] An achievement test is to measure skills and knowledge learned in a given grade level. Its scores are often used in an educational system to determine what grade-level a student belonging to.

http://dx.doi.org/10.1016/j.physa.2014.08.022
0378-4371/© 2014 Elsevier B.V. All rights reserved.

By using unilingual or bilingual corpus, some researchers managed to adopt complex networks to investigate texts' formats and content by several measurements, like the immediate associations between words, and the node degree, clustering coefficient and network dynamics [4]. Other researchers used complex network theory to retrieve information on some properties of the text, such as style and authorship [5], measure the relevance of words in a text [2], and identified the works of renowned English writers from the perspective of first-order statistics of words and other features obtained from texts [6]. A series of studies indicated that word adjacency networks [7], word association networks [8], semantic networks [1], and syntactic networks [9] are graphs that show features present in classical examples of complex networks. One of the important consequences of such studies is the presence of hubs in linguistic networks, e.g. the World Wide Web and social networks.

This paper presents a novel method for the analysis of nonlinear text quality in Chinese language. In Section 2, a method based on complex network theory is presented to analyzes text features like structure, coherence and cohesion, adherence to writing topics, and theme adequacy. Several complex networks of texts are constructed based on different score levels. In Section 2.3, we provide the discussions, conclusions, and future work.

2. Modeling approach and experiments

In the procedure for constructing complex networks of language, we firstly collect over 500 short-answer responses to a topic-discussing prompt. It was in a final exam that all the students read the prompt and write a response in Chinese language with an average article size of 173 words (or 205 characters). For better understanding, we translated several response samples, several key words into English in this paper, although they were originally written and processed in Chinese. At the same time, we provided the prompt and reference answer, and selected two samples of short-answer responses to be displayed in the Appendix.

To meet the challenge of maintaining testing reliability, each text was firstly scored by two human raters, in accordance with holistic rubrics that described six levels of performance, scored 1 through 6, where a score of "6" indicates the *highest* quality writing, and a score of "1" indicates one of the *lowest* quality, otherwise 0 was used to designate 'unresponsive' short-writings (texts that are off topic, repetitions of the prompt, or otherwise unintelligible or blank). Most large-scale tests require essays to be double-rated for reliability, with a third rater used if the first two raters disagree [9,10]. In our research, if the scores from the two human raters differed by less than two points, the average of the two scores became the final score. If the first two readers disagreed by two points or more, the essay was then rated by a third human rater and the final score was the average of the two closest scores. Human raters used 2 criteria related to text quality, namely (i) adherence to standard short-writing sample and (ii) theme adequacy/development. These are the criteria generally employed to mark short-writings for university students applying to achievement tests in China. Because of the large score dispersion for some texts, we select 80 texts within scores of 3, 4, 5, and 6, which means there are totally 320 texts in this research.

Then, we construct word adjacency/co-occurrence complex networks by considering the proximity between words. In the first step, we do the pretreatments for all the text, including: (i) Separate each texts into sentences $(S_1, S_2, \ldots S_n)$. (ii) For every sentence, a Chinese word segmentation program[2] had been launched to obtain separated words, e.g. w_1; w_2; w_3; ..., implying a directed edge from w_1 to w_3 and another directed edge from w_2 to w_3. (iii) The stopwords are removed and most of the synonym words were merged. Stopwords are also known as non-meaningful words, such as verb to be, some adverbs, and words from closed classes like articles, pronouns, prepositions and conjunctions. Then, the remaining words were merged according to a synonym corpus. Therefore, different occurrences of meaning-related words are represented uniquely by the same node in the complex network. In the second step, word adjacency networks of different score-levels were obtained. Obviously, there are 80 texts in a score-level, e.g. texts with a score of 3, and each word in a text can be represented as a node, while subsequent words are defined associations. Nodes are connected with edges when they appear in one sentence of the text. Typically, pairs of subsequent words, excluding texts and other connecting words are connected with unity length by considering couples of subsequent words.

Thirdly, three measurements of complex networks are estimated after the construction of the networks. That means the indegree and outdegree, the clustering coefficient, and the network dynamics are estimated while the number of connected components is monitored during complex networks' growing. So as to calculate topological network features considering word associations in a complex network. Keeping in mind that each complex network had been stored as one weight matrix, for example matrix *A*, all network measurements calculated here were performed based on the specific matrix.

2.1. Indegree and outdegree

Each node *i* in the complex network has its own indegree and outdegree. The in/outdegrees are defined respectively as

$$ID_i = \sum_{j=1}^{N} A_{ij}, \qquad (1)$$

[2] We use the ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System).

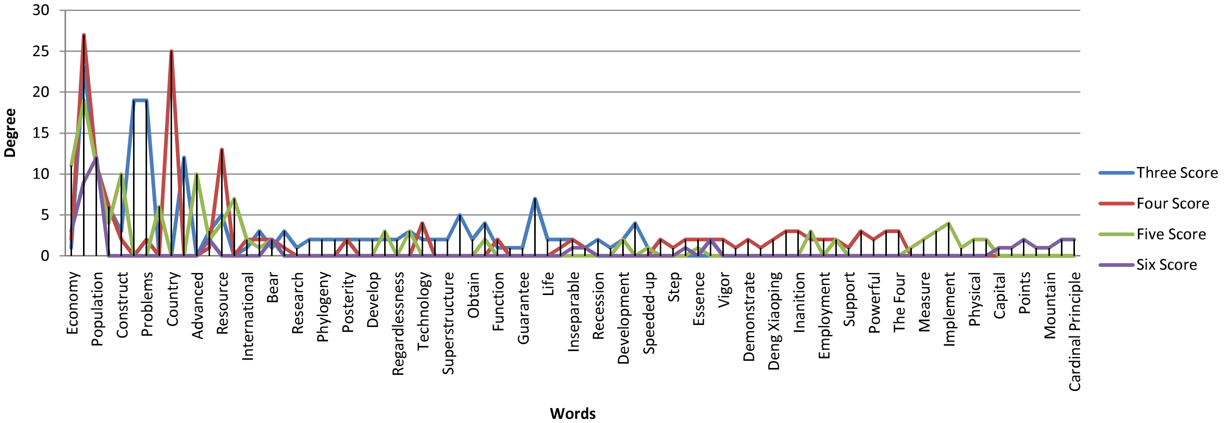

Fig. 1. Degrees of the four score level texts. The horizontal axes are words presented as network nodes. The vertical axes are the numbers of degree of each node in the specific score level text set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Pearson correlation of degrees of words in the 4 scores.

	Score_4	Score_5	Score_6
Score_3	.374**	.365**	.425**
Score_4		.477**	.510**
Score_5			.627**

** Correlation is significant at the 0.01 level (2-tailed). $N = 81$.

and

$$OD_i = \sum_{j=1}^{N} A_{ji}. \tag{2}$$

The indegree ID is obtained from the arithmetic mean of every w_i and so is the outdegree obtained similarly. Because the average value of the indegrees coincide with that obtained for the outdegrees, only the latter will be considered and presented as the degree henceforth. There are totally 81 keywords obtained from the texts. They perform a large dispersion among the degrees, as illustrated in Fig. 1, and each word will be presented as one node in complex networks.

Fig. 1 shows how the number of degrees evolves with the nodes for the four score level texts, being therefore representative of the evolution of connectivity in a given text. It is organized with the words distributed along the horizontal axes, while the degrees are positioned in the vertical axes. Different colors of plots refer to the measurements taken from texts with scores of 3, 4, 5, and 6, respectively. There are totally 81 words with in/out degrees in our research. However, only several words own more than 2 non-zero degrees, e.g. "Environment" owns degrees of (21, 26, 19, 9) when score = 3, 4, 5, and 6, respectively, while "Distribution" owns degrees of (2, 0, 0, 0) when score = 3, 4, 5, and 6, respectively. A further investigation depicts that there are only 27 nodes with more than 2 nonzero degrees in the four score levels.

Here comes a doubt from us: could such a small number of nonzero-degree-nodes make sense to complex network features of different score levels? We use correlation and dispersion of degrees in different score levels to answer this question. (i) Correlations in the values of degrees in different scores are calculated as Pearson correlation coefficient and listed in Table 1. It is obvious that the degrees of words in different score levels are significant correlated. (ii) The amount of degrees in each score level are also in some kind of law that the sum of degrees of score 3 through 6 own numbers of 172, 148, 120, and 44 respectively. In short, Fig. 1 and Table 1 indicate that the scores assigned by human raters increased with decreasing number of nodes and the sum of degree. Most significant are the texts of 5 marks and 6 marks. Intuitively, the bigger value of a node's degree means that in some way the more "important" that node is. That means we can judge how powerful a node is by comparing its degree with all the others in one complex network.

To investigate more features in complex network of text, a quantitative treatment of the data in Fig. 1 was carried out by calculating the clustering coefficient and network dynamics. This measurement will be referred to as "complex network dynamics of scores" in the remainder of this paper.

2.2. Clustering coefficient

The clustering coefficient of node v reflects the relationship between node v and its neighbors. To calculate the clustering coefficient of node v, firstly all nodes connecting directly with node v are identified and put into a given set $N(v)$, with

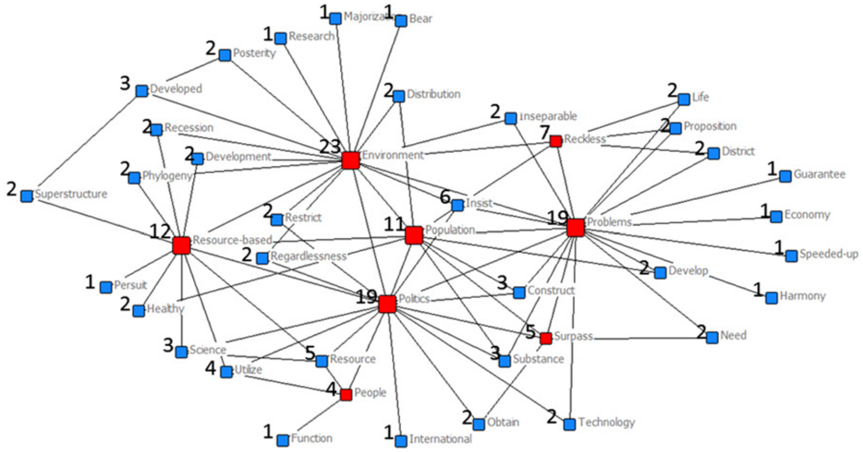

Fig. 2. The Complex Network I, score = 3, 80 texts, 44 nodes.

$E_i = |N(v)|$. Secondly, the total number of edges between all the nodes in $N(v)$ can be obtained from $k_i(k_i - 1)/2$, taking into account the edges directions, i.e., edge $i \to j$ is different from edge $j \to i$. Then, the clustering coefficient of node v is

$$C_i = \frac{2E_i}{k_i \times (k_i - 1)}. \tag{3}$$

In case k_i is smaller or equal to 1, then $C_i = 0$. The network's clustering coefficient C is the arithmetic mean of all individual clustering coefficients C_i.

2.3. Network dynamics

Measurements of a given text was calculated and taken into account for considering the dynamics of growth for the complex network in every score levels. The number of connected components (or clusters) was extracted after adding each word associations to the complex network, yielding a topological feature which is a function of the number of associations and, consequently, of the evolution of complex network construction [4]. For each text, the network was initiated with all N meaning-words, each one representing a single component, and the connections were established by each word association that occurred along the text. When a word association was recognized, a new edge was created, and the degree, known as D, of an already existing edge was increased. As a consequence of the word adjacency model, the number of connected components always converged to one after all words that had been introduced. A quantitative treatment of the plots in a complex network is carried out by calculating the extent to which the real plot deviated from other plots. The deviation in the network dynamics is calculated as

$$CF_i = \frac{\alpha C_i}{\sum_{j=1}^{N} C_j} + \frac{(1-\alpha) \times D_i}{N}, \tag{4}$$

where $0 < \alpha < 1$, and N is the total number of plots in a complex network.

Figs. 2–5 reveal the deviation from nonlinear dynamics in complex networks' growth according to score levels. Because of the marks dispersion for all the 500 texts, we perform the complex network analysis taking only texts of score 3, 4, 5, and 6 with the medium dispersion for each criterion.

As report [4] discussed, the measurements associated with complex networks could be used to distinguish between low-quality and high-quality texts. Here, the Complex Network I through IV also depict that the complex network measurements can easily capture the quality of texts. When comparing these four complex networks, we can easily find out that scores assigned by the human raters decrease with the increasing sum of nodes. The degree of nodes in a complex network can represent cohesion/coherences and adherences to standard writing conventions clearly. We examined the nodes with a degree of 4 or bigger in all the complex networks, typical nodes are the words of "development", "population", "economy", and "environment". The sum of these words' degrees counted up to 82% of all. In fact, they have a great amount of information, and other nodes are clustered around them while presenting great deviation. As a result, one may argue that the writer kept discussing several arguments via repeating those big-degree-words, leading to a high text quality. Moreover, as we illustrate in Figs. 2 through 5 for the nodes of words with their degrees, the fewer amounts of nodes in a complex network indicating higher quality of writing. For the reason that main concepts are highly interconnected, which represents

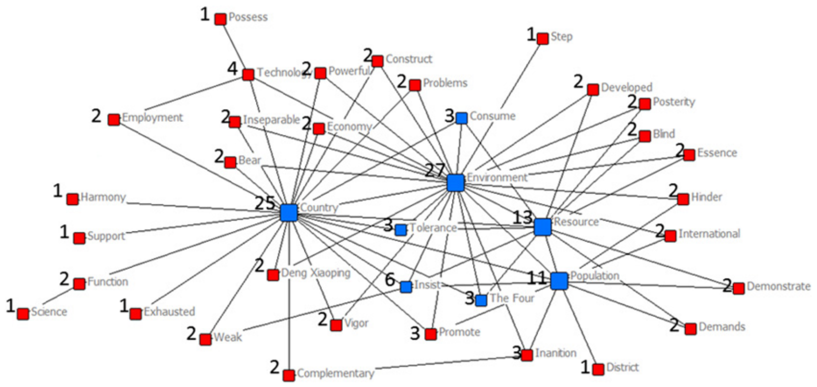

Fig. 3. The Complex Network II, score = 4, 80 texts, 40 nodes.

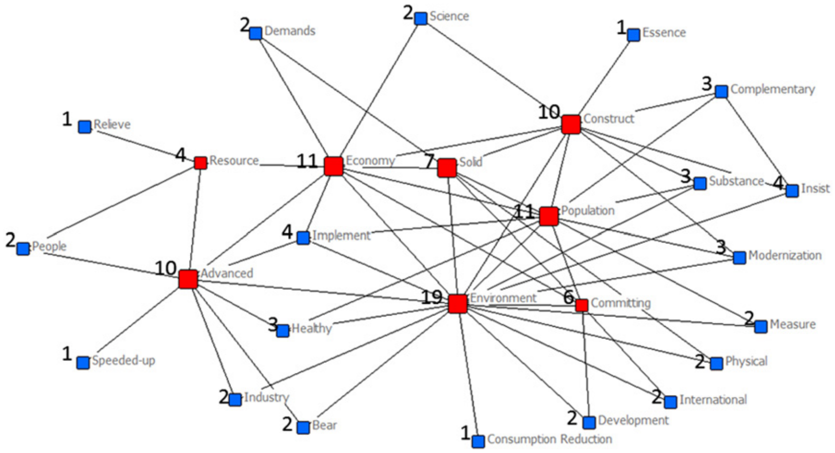

Fig. 4. The Complex Network III, score = 5, 80 texts, 28 nodes.

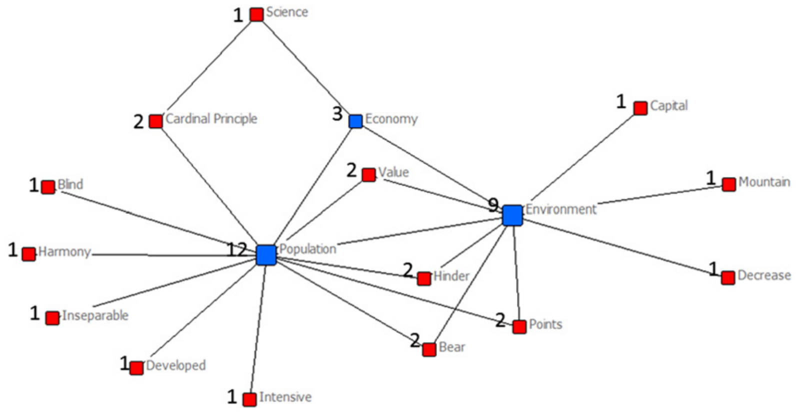

Fig. 5. The Complex Network IV, score = 6, 80 texts, 18 nodes.

that the writer had focused on the key concepts and viewpoints. Noted that in Table 1, the significant correlation between degrees of different score texts were not obtained by chance.

In addition, in subsidiary experiments we observed that essentially the same conclusions and trends apply for the complex network, which also included those with large dispersions in the scores assigned by the human judges. We arrange all the non-zero degrees of each node in different score levels, then 4 linear trend lines are drawn in Fig. 6.

In Fig. 6, a few words have a large number of degrees, these lines rapidly dropped down in a much similar trend. From a linguistic point of view, the big-degree-words appear to represent whether the flow of the prose is adequate or not, which is reflected especially in the cohesion and coherence.

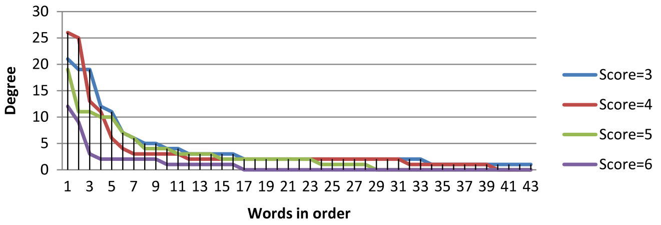

Fig. 6. The trend lines of degrees of all nodes in four score levels.

Table 2
The average deviations of the 4 complex networks.

	Complex Network I	Complex Network II	Complex Network III	Complex Network VI
Ave_N_D	3.17	2.89	2.13	1.41

With the analysis of network dynamics, we can easily figure out one node's deviation [4]. Network dynamics can be used to distinguish between low and high quality texts. Hence, it is meaningful to calculate the average network dynamics of each complex network considering all the nodes inside. The average deviations of the Complex Network I through VI are listed in Table 2.

Table 2 points to the text quality decreasing with increasing network dynamics. Similar conclusions can be drawn in accordance with what we had found in Figs. 2 through 5. For example, the average network dynamics of the Complex Network I is 3.17, while only 1.41 in the Complex Network VI. Similar conclusions can be drawn as above. From the perspective of article structure, the deviation of nodes in a complex network indicates how faster or closer a writer introduces new concepts in a text. It seems that, the intense of the big-degree-words in a complex network would course a low-quality text. As we know, if a writer simply kept repeating some words, probably copied from the prompt, in the remainder of the writing process, leading to a low-quality text. Respectively, there is apparently difference with the presentation of the Complex Network I and Complex Network IV, which practically resulting diverse scores in dissimilar context. This corroborates the earlier finding in previous report [4].

The trend of toward decreasing scores with the number of degrees and network dynamics suggests that text lose quality if the concepts are highly interconnected. Despite the many parameters and unavoidable subjectivity of human language, complex networks present potential to be used as a subsidy for a more objective and reproducible means to evaluate text quality, rapidly refining the central idea and key distribution of texts, or automatically retrieving abstracts of articles.

3. Conclusions

This study addressed the analysis of complex and nonlinear dynamics in texts analysis. Responses to an achievement test in Chinese language were characterized by means of indices with considerable degrees, clustering conferences, and dynamical networks making text features in different score levels representative. The proposed methodology reformulates the classical methods lead to a new model based in the text analysis, text classification, and autoscoring short writing applications. In future study, we would adopt more measurements of complex network theory to capture features on text sets. For example, the shortest path and entropy of texts can be used to automatically judge the quality of texts.

Acknowledgments

The authors are very grateful to the anonymous referees for their very detailed referencing and helpful suggestions to improve this paper. Also, the first author would like to express her sincere acknowledgments to the National Ministry of Education (11YJC740046) and the University innovation projects of Guangdong province (2014) as research topics and sponsored this study. The authors also thank the Centre of Translation Studies of Guangdong University of Foreign Studies for useful discussions about this research while it was in progress, as well as helpful comments from the Collaborative innovation center of foreign language studies and service. The third author would like to thank The University Scientific and Technological Innovation Project of Guangdong Province (2013KJCX0068) for sponsoring this study. A further report of this research will appear as Correlation between essay scores and complex dynamics analysis by Ke, Zeng, and colleagues (in press).

Appendix

The prompt, reference answer, and two samples of short-answer responses.

The prompt (in Chinese)	请简要分析以下观点：经济发展是实现人口、资源、环境与经济协调发展的根本出路；但同时，经济的发展也离不开人口、资源和环境的支持。
The prompt (be translated into English)	Briefly discuss *the relationships of economy, environment, and population in China society: in one way, the economic development is the fundamental way to achieve the coordinated development among population, resources, environment, and economic itself; in the other way, the economy cannot develop without the support of population, resources, and environment in China.*
Reference answer (in Chinese)	经济发展是我国社会主义发展的核心，发展是党执政兴国的第一要务。经济发展为我国政治发展、文化发展以及其他方面的发展提供了坚实的物质基础和保障，是解决我们当前社会主义初级阶段所面临所有矛盾和问题的根本途径，因此也是实现人口、资源、环境与经济协调发展的根本出路； 但是同时，经济发展也离不开人口、资源和环境的支持。因为人是经济发展的主体，资源是经济发展的要素，环境为发展提供依托。所以经济的发展要同时考虑到人口的承载力，资源的支撑力，环境的承受力，走可持续发展道路。
Reference answer (be translated into English)	Economic development is the core of China's socialist development and the CCP Party’s absolute priority for governing and rejuvenating the country. As we know, economic development provides the material foundation and guarantee for China's political development, cultural development and other aspects. It is a fundamental way to solve the contradictions and problems that we are facing in current primary stage of socialism. Therefore, economic development is the fundamental way to achieve the coordinated realization of population, resources, environment and economic development; However, the economic development also cannot do without the support of population, resources and environment. For the reason that people are the principle part of economic development, resources are main factors in the development of economy, and environment provides a solid foundation for all developments, the development of economy must take them into account. To develop economy, we should consider the supporting capacity of population, resources, and environment; take the road of sustainable development.
Short-answer sample_1 (score = 6)	The core idea of socialism is to emancipate and develop our productive forces while focusing on economic construction. Firstly, keeping the Scientific Development Concept in mind: development is the primary task for the ruling party. Only when the means of economy were fully developed would the fundamental interests of the people can be ensured, the material civilization, spiritual civilization, and political civilization can achieve their developmental foundation. Therefore, with the development of economy, people's living standards improve, cultural level increase, the efficient use of resources and environmental protection consciousnesses also increase. Secondly, along with the development of economy, there will bring some destruction of resources and environment, and the growth of population. The way to defuse them is to coordinate the development of economic, population, resources, and environment. Keep a balance of developing economy and resources, and the protection of environment and population growth. Bearing the essential idea of the Scientific Development Concept in our mind, we are proposing to build an ecological civilization as well as coordinating the expense of the environment, resources, and population. So as to build a resource-saving, environment-friendly society.
Short-answer sample_2 (score = 3)	The relations among the economic development, population, resources, and environment are closely linked. In a region, the allocation and distribution of population, environment, and resources determines its level of economic development, to a large extent. They guarantee the basic requirement of economic development, so the statement in this prompt is correct. However, to the coordinated development of population, resources, environment and economy, the fundamental way out is to build environment-friendly society.

References

[1] A.J. Holanda, I.T. Pisa, O. Kinouchi, A.S. Martinez, E.E.S. Ruiz, The saurus as a complex network, Physica A 344 (2004) 530–536.

[2] H. Zhou, G.W. Slater, A metric to search for relevant words, Physica A 329 (2003) 309–327.

[3] J.A. Tenreiro, Complex dynamics of financial indices, Nonlinear Dyn. (2013) Published online: 04 June.

[4] L. Antiqueira, M.G.V. Nunes, O.N. Oliveir Jr., L. da, F. Costa, Strong Correlations between text quality and complex networks features, Physica A 373 (2007) 811–820.

[5] M.A. Montemurro, D.H. Zanette, Entropic analysis of the role of words in literary texts, Adv. Complex. Syst. 5 (1) (2002) 7–17.

[6] L.L. Goncalves, L.B. Goncalves, Fractal power law in literary English, 2005 condmat/0501361.

[7] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, U. Alon, Super families of evolved and designed networks, Science 303 (2004) 1538–1542.

[8] L.F. Costa, What's in a name? Int. J. Mod. Phys. C 15 (2004) 371–379.

[9] R. Ferreri Cancho, R.V. Solé, R. Kohler, Patterns in syntactic dependency networks, Phys. Rev. E 69 (2004) 051915.

[10] M.D. Shermis, J. Burstein, Handbook of Automated Essay Evaluation: Current Applications and New Directions, NJ, Routledge, 2013, 38–39.