Physica A 468 (2017) 783–789

Contents lists available at ScienceDirect

Physica A

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

Measuring information-based energy and temperature of literary texts

Mei-Chu Chang [a,b,*], Albert C.-C. Yang [c,d], H. Eugene Stanley [b], C.-K. Peng [a,c]

[a] Research Center for Adaptive Data Analysis, National Central University, Jhongli 32001, Taiwan

[b] Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA

[c] Center for Dynamical Biomarkers, Division of Interdisciplinary Medicine and Biotechnology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts 02215, USA

[d] Taipei Veterans General Hospital, Taipei 11217, Taiwan

HIGHLIGHTS

- The relative temperature of text, as an author's writing capacity, illustrates that how an author uses the 100 most frequents words to write a text.
- The Shannon entropy is used to measure the text complexity to quantify the author's lexical processing.
- A statistical method – information-based energy – constituted of the relative temperature and the Shannon entropy is applied to quantify an author's writing performance.
- Two authors – Shakespeare in English and Jin Yong in Chinese – are analyzed to demonstrate the method. It is found that their well-known works are associated with higher information-based energies.

ARTICLE INFO

Article history:
Received 1 May 2016
Received in revised form 15 September 2016
Available online 27 November 2016

Keywords:
Linguistic analysis
Thermodynamics and statistical mechanics
Boltzmann distribution
Shannon entropy

ABSTRACT

We apply a statistical method, information-based energy, to quantify informative symbolic sequences. To apply this method to literary texts, it is assumed that different words with different occurrence frequencies are at different energy levels, and that the energy-occurrence frequency distribution obeys a Boltzmann distribution. The temperature within the Boltzmann distribution can be an indicator for the author's writing capacity as the repertory of thoughts. The relative temperature of a text is obtained by comparing the energy-occurrence frequency distributions of words collected from one text versus from all texts of the same author. Combining the relative temperature with the Shannon entropy as the text complexity, the information-based energy of the text is defined and can be viewed as a quantitative evaluation of an author's writing performance. We demonstrate the method by analyzing two authors, Shakespeare in English and Jin Yong in Chinese, and find that their well-known works are associated with higher information-based energies. This method can be used to measure the creativity level of a writer's work in linguistics, and can also quantify symbolic sequences in different systems.

Sequences of symbols carrying information are commonly found in nature, e.g., human language and genetic codes. How to quantify these informative symbolic sequences based on the occurrence and rank of repetitive patterns is an interesting

* Corresponding author.
E-mail address: meichu@bu.edu (M.-C. Chang).

http://dx.doi.org/10.1016/j.physa.2016.11.106
0378-4371/

and open issue. Here we focus on the words of literary texts and introduce a statistical method of quantifying authors and of quantifying hidden structures in informative sequences in such systems as genetic codes and human heart rate time series.

In linguistics, the occurrence of different words [1–4], word ranks in the table of occurrence frequency [1–3,5], vocabulary richness [6,7], and entropy-based measures [8,9] can be used to quantify writing styles of literary texts. For example, word occurrence frequency-rank order statistics and phylogenetic tree construction have been used to resolve literary authorship disputes [3]. For novels written by different authors, not only are power-law distribution differences observed, but the exponents also differ [10].

The concept of "text temperature" has been introduced to linguistic analysis [11–16] under the assumption that human language can be described as a physical system within the framework of equilibrium statistical mechanics. It can be used to measure communicative ability [13], or it can be associated with text size [14]. Recently, the authors have successfully associated words with energies (i.e., word energies) based on a general standard Maxwell–Boltzmann distribution [15,16]. It is found that, the linguistic relative temperature of a book can be determined by measuring the deviation from a standard Maxwell–Boltzmann distribution of a corpus of English words [15]. The relative temperature can also measure vocabulary complexity relative to the academic level of the text and to the target readership in different languages [16]. The word energies can be defined by using the American National Corpus in Ref. [15] or the Project Gutenberg corpuses of English in Ref. [16] as a general standard Boltzmann distribution.

In this work, an information-based energy for literary texts is applied by combining the relative temperature [15,16] and information Shannon entropy [17] of the text. This information-based energy can be viewed as a quantifier of authorial writing performance of a text. It is assumed that different words with different occurrence frequencies have different word energies, and that the word energy-occurrence frequency distribution obeys a Boltzmann distribution [15,16]. The temperature introduced by the Boltzmann distribution may be a representative of the author's writing capacity and their repertory of thoughts. Unlike the corpuses gathered from different authors in Refs. [15,16], the word occurrence frequencies in this work are determined by considering the corpus from *a single author*. Then the word energies can be observed using a Boltzmann distribution associated with the reference temperature. Note that, by considering the corpus from a single author, how the relative temperature concept plays a role in different literary writing styles or genres of the same author can be unveiled by getting rid of interferences from other authors.

When an author writes a text, he/she must change his/her writing capacity to express the specific thoughts of the text. We assume that, the change of author's writing capacity leads to the temperature change of the text associated with the change of the Boltzmann distribution of the word occurrence frequencies in the text. The relative temperature is defined as the ratio between the temperature of the text and the reference temperature of the same author's corpus. By combining the information-based energy with the information Shannon entropy [17] to measure the author's text complexity (how the author enriches the text), we can quantify the author's writing performance. We demonstrate the method by analyzing two authors, Shakespeare in English and Jin Yong in Chinese (currently the best-selling living Chinese author). We find that the famous works of the two authors display high energies, even when translated into other languages.

We first analyze William Shakespeare's 35 plays and we classify them into four genres [18]. In order to avoid the finite size effect, we exclude the plays with fewer than 5000 words (i.e., *Richard III* and *Love's Labour's Lost*). We first construct the word rank-occurrence frequency P distribution [black squares in Fig. 1(a)] by collecting the words in Shakespeare's 35 plays and find that there are 22,292 different words out of a total of 846,036 words. For example, the first and second most frequently used words (i.e., "the" and "and") are ranked first and second, respectively. We find that in the rank regime of approximately 20–2000 the occurrence frequency distribution P in Fig. 1(a) obeys an empirical Zipf law which states that the occurrence frequency of a word and its rank follow a power-law relation [1,2,19–21]. The exponent of the power law in Fig. 1(a) is approximately -1.1. The curve departs from the power-law regime at the lowest and highest rank regimes. Note that the step-like plateaux are a finite size effect associated with infrequent words related to specific descriptions, such as names or places. In this distribution different word forms, e.g., "hand" and "hands" or "write" and "wrote", are treated as different words. This provides a useful simplicity, but it also allows us examine the author's writing preferences in detail.

A Boltzmann distribution is used to describe the relation between the occurrence frequencies of words collected from Shakespeare's 35 plays and the word energies, i.e.,

$$P(i) \sim \exp(-E_i/kT), \tag{1}$$

which describes the occurrence frequency $P(i)$ of a given word i related to the word energy E_i, where k is the Boltzmann constant, and T is the constant reference temperature [15,16]. Here the reference temperature T relates to Shakespeare's entire writing capacity. For convenience, we set $1/kT = 1$, and the word energy E_i of each word i can be determined directly [$E_i = -\ln P(i)$]. The black solid line shown in Fig. 1(b) depicts the corresponding E_i vs. $P(i)$ Boltzmann distribution. Note that different authors may have different capacities; therefore, different reference temperatures T should be considered. It is the main reason why we choose the corpus only from a single author, instead of the American National Corpus [15] or the Project Gutenberg corpuses of English [16] from many authors.

To express their thoughts, an author composing a text must adjust their writing capacity (temperature). The words in the text exhibit another word occurrence frequency distribution p. For example, the blue circles in Fig. 1(a) indicate the occurrence frequency p distribution in Shakespeare's play *Hamlet* according to the rank order in the occurrence frequency P table [22]. The step-like plateaux are due to finite size effects. Thus the E_i vs. $p(i)$ energy-occurrence frequency distribution of *Hamlet* [blue circles in Fig. 1(b)] deviates from the Boltzmann distribution [black line in Fig. 1(b)], i.e., $p(i) \sim \exp(-E_i/kT')$,

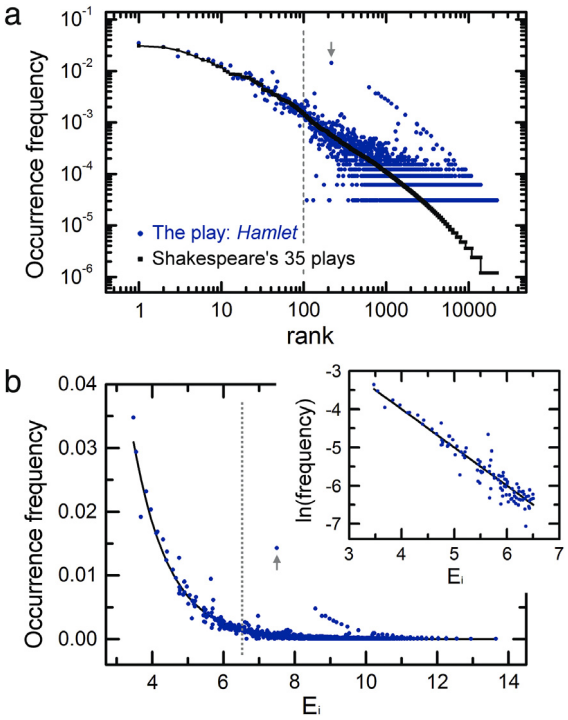

Fig. 1. (Color online) (a) and (b) The word rank-occurrence frequency distribution and word energy-occurrence frequency distribution of texts. The highest frequent word in (a) is ranked as one. The black squares in (a) and (b) are the word occurrence frequency distribution of collections of Shakespeare's 35 plays, and the blue circles in (a) and (b) are the word frequency distribution of the play *Hamlet*. The gray arrows pointed to the blue circles depict the word "*Hamlet*". The vertical gray dotted line in (b) shows the energy of the word ranked as 100. The insect is the semi-log plot of (b), where only the 100 most frequent words ranked from 1–100 in the occurrence frequency P table are used to fit.

where the temperature T' is the author's writing capacity while writing *Hamlet*. Thus the relative temperature T'/T can be derived from the inverse of the slope value by linearly fitting the $\ln p(i)$ vs. E_i plot, where $k = 1/T$ [the inset in Fig. 1(b)]. It represents Shakespeare's ability to organize words while constructing the play. The higher (or lower) relative temperature T'/T leads to a wider (or narrower) word distribution.

For the linear fittings of the $\ln p(i)$ vs. E_i plots, we use the 100 most frequent words ranked in the occurrence frequency P table. Most of these 100 words are function words that express grammatical relationships with other words. They also constitute ∼50% of the total words in both the occurrence frequency P distribution of all 35 plays and in the occurrence frequency p distribution of any single play. The words excluded from this fitting are specific to the content of individual plays, e.g., names and places. For example, the words indicated by the gray arrows in Fig. 1(a) and (b) are "Hamlet", the name of the main character in the play *Hamlet*. If we include all of the words in *Hamlet* in the fitting, specific words and the finite size effect will strongly affect the relative temperature T'/T of the slope. Hence, by focusing the 100 most frequent words only, the relative temperature T'/T as the author's writing capacity can be interpreted as how the author frames the text via the 100 most frequent words. If the author uses a more widely distribution of the 100 most frequent words in a text, the higher relative temperature of the text will be observed.

Fig. 2(a) shows the relative temperatures T'/T of Shakespeare's 35 plays classified in four genres, i.e., comedies, histories, tragedies, and romances. The average T'/T of the comedies, histories, tragedies, and romances are 0.963, 0.946, 1.011, and 1.010, respectively. Note that the tragedies have the highest averaged T'/T and the histories the lowest. Shakespeare used the 100 most frequent words more widely distributed when he wrote tragedies and more narrowly distributed when he wrote histories. Within Shakespeare's 35 plays, the T'/T of *Macbeth*, considered one of his darkest and most powerful works, is the highest. The two plays *Henry VIII* and *Pericles, Prince of Tyre* are generally assumed to be collaborations with John Fletcher [23] and George Wilkins [23], respectively, and this possibly explains why they have a higher T'/T value than the other history plays. Note that the romantic genre is a hybrid containing comic and tragic elements, and its average relative temperature is slightly lower than the tragedies. If the T'/T of the play *Pericles, Prince of Tyre* is excluded, the average T'/T of romances falls between that of the comedies and the tragedies.

After determining the T'/T of a play we can then measure how the author enriches the lexical words of a text. We introduce Shannon information entropy [17] to measure the text complexity indicating the author's ability to fill in the details of a text, i.e., $H = -\sum_{i=1}^{N} p_i \ln p_i$, where N is the number of different words and p_i is the occurrence frequency of word i in one text. When there is a higher H values more information is being included. Fig. 2(b) depicts the Shannon entropies H of Shakespeare's 35 plays, and the corresponding averaged H of plays in comedies, histories, tragedies, and

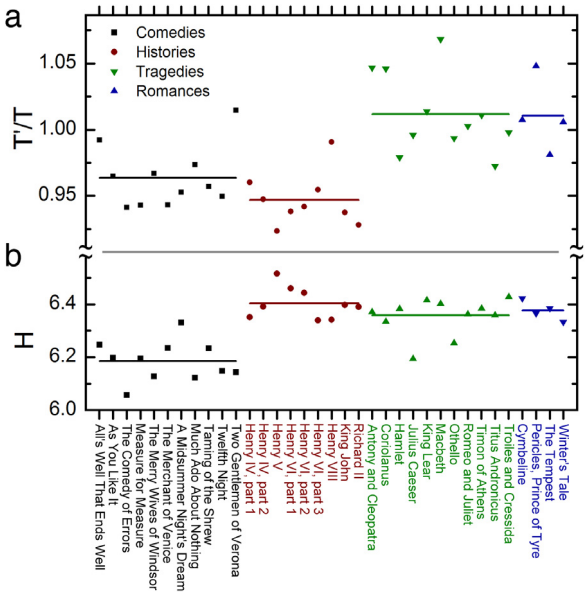

Fig. 2. (Color online) (a) and (b) The relative temperature T'/T and the Shannon entropy H of Shakespeare's 35 plays classified into 4 genres. Black squares, red circles, green upside-down triangles, and triangles are comedies, histories, tragedies, and romances. The horizontal lines of genres are the averaged T'/T in (a) and the averaged H in (b). The averaged T'/T of plays in comedies, histories, tragedies, and romances are 0.963, 0.946, 1.011, and 1.010. The averaged H of plays in comedies, histories, tragedies, and romances are 6.185, 6.403, 6.358, and 6.376.

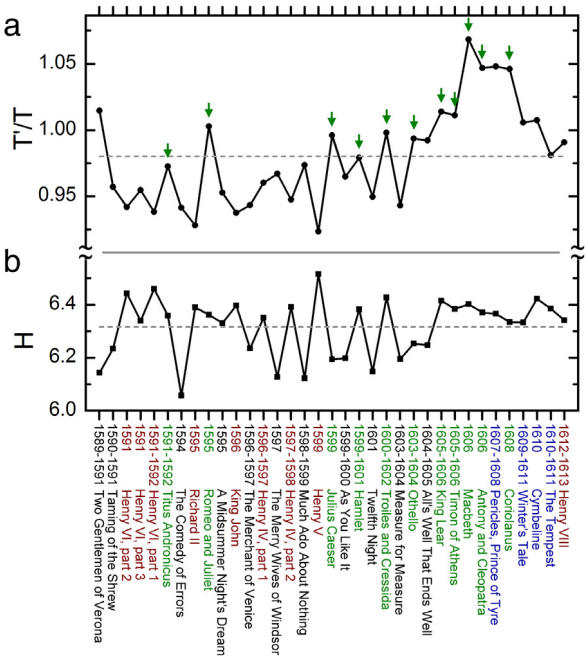

Fig. 3. (Color online) (a) and (b) The relative temperature T'/T and the Shannon entropy H of Shakespeare's 35 plays with the chronological order. Annotations of plays are classified into 4 colors. Black, red, green and blue colors represent the comedies, histories, tragedies, and romances. Green arrows in Fig. 3(a) specifically label tragedies. The gray dashed lines in (a) and (b) are the averaged relative temperature (=0.98) and the Shannon entropy (=6.316), respectively.

romances are 6.185, 6.403, 6.358, and 6.376, respectively. The histories now have the highest averaged H, possibly due to descriptive words associated with wars and politics [8].

Fig. 3 shows the relative temperature T'/T and the Shannon entropy H of Shakespeare's 35 plays with the chronological order. Annotations of plays are classified into 4 colors. Black, red, green and blue colors represent the comedies, histories, tragedies, and romances. The gray dashed lines in Fig. 3(a) and (b) are the averaged relative temperature (=0.98) and the Shannon entropy (=6.316), respectively. We find that, the T'/T trend in Fig. 3(a) increases and then drops. By simply

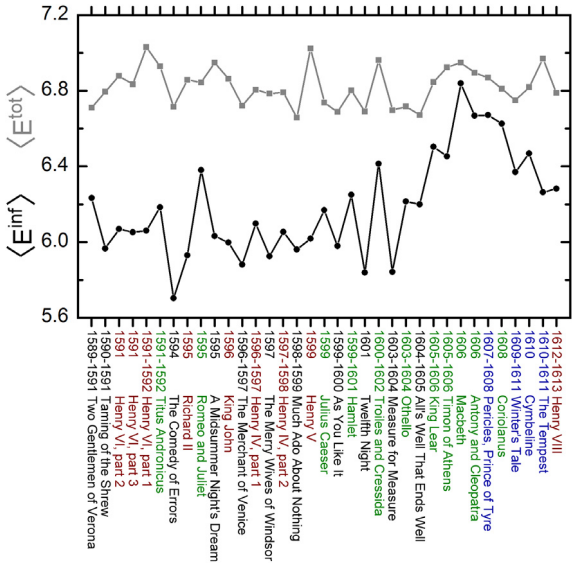

Fig. 4. (Color online) The energies of Shakespeare's 35 plays with the chronological order. The black squares are the information-based energy $\langle E^{\rm inf}\rangle$ considering the text relative temperature and the Shannon entropy. The gray circles are the normalized total energy $\langle E^{\rm tot}\rangle$ considering word energies E_i directly. Annotations of plays are classified into 4 colors. Black, red, green and blue colors represent the comedies, histories, tragedies, and romances.

grouping the plays, 19 and 16 plays are included before and after A.D. 1600. Only 3 plays (out of 19 plays) before A.D. 1600 but 13 plays (out of 16 plays) after A.D. 1600 with T'/T are higher than the averaged value. For H, 11 plays (out of 19 plays) before A.D. 1600 and 12 plays (out of 16 plays) after A.D. 1600 with H are higher than the averaged value. For the tragedies labeled by green arrows in Fig. 3(a), the trend of T'/T increases and reaches to its maximum in A.D. 1606. For the last 12 plays constituting the peak of T'/T in Fig. 3(a), 6 plays are tragedies and 4 plays are romances as a hybrid containing comic and tragic elements. By focusing tragedies [by green arrows in Fig. 3(a)], the highest relative temperature is the play *Macbeth* at the later stage (A.D. 1060). The relative temperature of the play *Titus Andronicus* as the first tragedy at the early stage (A.D. 1591–1592) is the lowest. It indicates that the author's writing capacity changes, even the plays at different stages belong to the same genre.

Here we consider the author's writing performance by putting the author's writing capacity of text and the text complexity together. This leads to the information-based energy $\langle E^{\rm inf}\rangle$ as the writing performance, introduced as the multiplication of the relative temperature of 100 most frequent words and the Shannon entropy of all words in one text,

$$\langle E^{\rm inf}\rangle \equiv \frac{T'}{T} \times \left(-\sum_{i=1}^{N} p_i \ln p_i\right). \tag{2}$$

Fig. 4 shows the information-based energy $\langle E^{\rm inf}\rangle$ of Shakespeare's 35 plays arranged in chronological order (black circles). (This chronological order is the one listed in the *Oxford Shakespeare* [23].) We find that in Shakespeare's 35 plays, the $\langle E^{\rm inf}\rangle$ trend of a play initially increases and then later drops. Because the word energies (E_i) are first estimated in the beginning, the normalized total energy can simply been tested, i.e., $\langle E^{\rm tot}\rangle = \sum_{i=1}^{N} p_i E_i = -\sum_{i=1}^{N} p_i \ln P_i$. The normalized total energies $\langle E^{\rm tot}\rangle$ of plays are shown as squares in Fig. 4. We find that the information-based energy $\langle E^{\rm inf}\rangle$ has a better resolution than the normalized total energy $\langle E^{\rm tot}\rangle$.

Can the method of using information-based energy as a quantifier for an author's performance be applied to literary texts in other languages? We examine 12 martial arts and chivalry (wuxia) novels by Jin Yong, considered one of the finest wuxia writers of all time, and find 5284 different Chinese characters out of a total of 7,086,619 [24]. Although character combinations in Chinese alter the meaning, for simplicity we only count the occurrence of single characters. We calculate the relative temperature of each novel using the 400 most frequently used characters appearing in the novel. Fig. 5 shows the information-based energy $\langle E^{\rm inf}\rangle$ of Jin Yong's 12 novels in chronological order. The annotations are the names of the novels in Chinese and their translations in English [25].

Fig. 5 shows that the highest $\langle E^{\rm inf}\rangle$ belongs to the novel *The Legend of the Condor Heroes*, the most famous novel in the Chinese language. The gray arrows indicate the novels *The Legend of the Condor Heroes*, *The Return of the Condor Heroes*, and *The Heaven Sword and Dragon Saber*, which constitute the *Condor Trilogy* and should be read in that order. *Demi-Gods and Semi-Devils* is a precursor to the *Condor Trilogy*, and that may account for its information-based energy also being higher. The trend of $\langle E^{\rm inf}\rangle$ decreases in the successive novels. Note that methods of decomposing the meaningful Chinese characters are still to be developed. This topic will be studied in a future work.

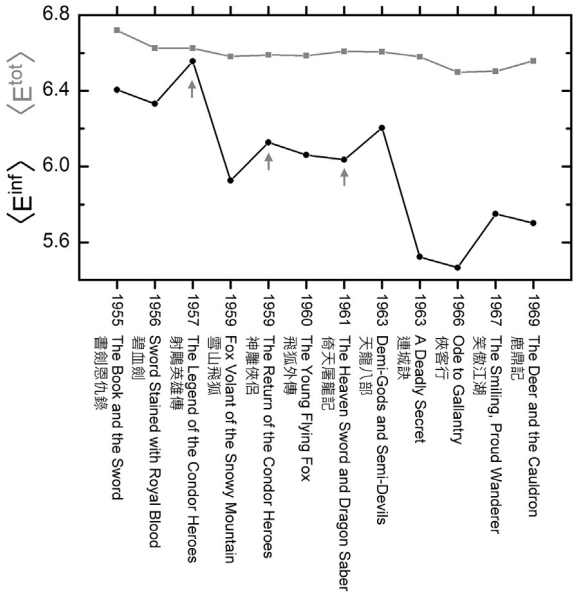

Fig. 5. The energies of Jin Yong's 12 Chinese martial arts and chivalry plays in chronological order. The black squares are the information-based energy $\langle E^{\text{inf}}\rangle$, and the gray circles are the normalized total energy $\langle E^{\text{tot}}\rangle$ considering word energies directly.

In conclusion, we apply a quantitative measure, information-based energy, that combines the relative temperature of the 100 most frequent words and the Shannon entropy of all words in a single text, to evaluate an author's writing performance. (i) By assuming that the word occurrence frequency and word energy obeys a Boltzmann distribution, the relative temperature of a text indicates the author's writing capacity. Different writing capacities lead to different uses of the 100 most frequent words. (ii) Using the Shannon entropy of a text to measure the text complexity quantifying the author's lexical processing. (iii) We therefore define information-based energy to be the relative temperature of a text multiplied by the Shannon entropy of the text. We apply our method to two authors, Shakespeare in English and Jin Yong in Chinese. Examining Shakespeare's plays, we find that different genres exhibit different relative temperatures and information-based energies. Taking the plays in chronological order, we find that the information-based energy values first increase and then decrease. In the Chinese novels of Jin Yong, the trend of the information-based energy decreases chronologically. The highest text energies are found in the most popular texts, i.e., the famous texts that have a wide readership. This phenomenon may be linked to the creativity of a writer and may have important implications in quantifying a person's thoughts, whether in English or in Chinese.

In future work we will study how this informative-based energy can be applied to genetic sequences associated with evolution and to human heart rates associated with physiological states.

Acknowledgment

This work is supported by the MOST support for the Center for Dynamical Biomarkers and Translational Medicine, National Central University, Taiwan (MOST 103-2911-I-008-001).

References

[1] G.K. Zipf, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley, Cambridge MA, 1949.
[2] G.K. Zipf, Selected Studies of the Principle of Relative Frequency in Language, Harvard University Press, Cambridge MA, 1932.
[3] C.-C. Yang, C.-K. Peng, H.-W. Yien, A.L. Goldberger, Physica A 329 (2003) 473;
C.-K. Peng, C.-C. Yang, A.L. Goldberger, Chaos 17 (2007) 015115.
[4] F. Mosteller, D.L. Wallace, J. Amer. Statist. Assoc. 58 (1963) 275;
J. Burrows, Comput. Humanit. 37 (2003) 5.
[5] S. Havlin, Physica A 216 (1995) 148.
[6] D.I. Holmes, J. Roy. Statist. Soc. Ser. A 155 (1992) 91;
R. Thisted, B. Efron, Biometrika 74 (1987) 445;
F.J. Tweedie, R.H. Baayen, Comput. Humanit. 32 (1998) 323.
[7] L.L. Goncalves, L.B. Goncalves, Physica A 360 (2006) 557.
[8] O.A. Rosso, H. Craig, P. Moscato, Physica A 388 (2009) 916.
[9] D.L. Hoover, Comput. Humanit. 37 (2003) 151;
P. Thoiron, Comput. Humanit. 20 (1986) 197;
M.A. Montemurro, D.H. Zanette, Adv. Complex Syst. 5 (2002) 7.
[10] C.K. Hu, W.C. Kuo, POLA Forever (2005) 115–139.
[11] B. Mandelbrot, in: W. Jackson (Ed.), Communication Theory, Academic Press, New York, 1953, pp. 486–502.

[12] H. de Campos, J.M. Tolman, Phys. Today 3 (1982) 177.
[13] K. Kosmidis, A. Kalampokis, P. Argyrakis, Physica A 366 (2006) 495.
[14] A. Rovenchak, S. Buk, Physica A 390 (2011) 1326.
[15] S. Miyazima, K. Yamamoto, Fractals 16 (2008) 25.
[16] H.H.A. Rego, L.A. Braunstein, G. D'Agostino, H.E. Stanley, S. Miyazima, PLoS One 9 (2014) e110213.
[17] C.E. Shannon, Bell Labs Tech. J. 27 (1948) 379.
[18] The 35 plays of William Shakespeare are accessed from.
[19] R.F. Cancho, R.V. Solé, Proc. Natl. Acad. Sci. 100 (2003) 788.
[20] A.M. Petersen, J.N. Tenenbaum, S. Havlin, H.E. Stanley, M. Perc, Sci. Rep. 2 (2012) 943.
[21] M.A. Montemurro, Physica A 300 (2001) 567.
[22] For example, the word '*Hamlet*' can have a different rank order according to occurrence frequency tables *P* composed of 35 plays or *p* composed of the play *Hamlet*. In the occurrence frequency tables *P* and *p* of *Hamlet*, the rank orders of the word '*Hamlet*' are 218 and 9, respectively.
[23] S. Wells, G. Taylor, J. Jowett, W. Montgomery, The Oxford Shakespeare: The Complete Works, second ed., Oxford University Press, Oxford, 2005.
[24] The 12 novels chosen from the 15 total novels of Jin Yong were selected because they each have more than 100,000 words.
[25] Most of the novels were initially published in daily installments in newspapers, thus the chronology extends across a range of years.