Accepted Manuscript

Measuring information-based energy and temperature of literary texts

Mei-Chu Chang, Albert C.-C. Yang, H. Eugene Stanley, C.-K. Peng

PII: S0378-4371(16)30929-3
DOI: http://dx.doi.org/10.1016/j.physa.2016.11.106
Reference: PHYSA 17762

To appear in: *Physica A*

Received date: 1 May 2016
Revised date: 15 September 2016

Please cite this article as: M.-C. Chang, A.C.-C. Yang, H. Eugene Stanley, C.-K. Peng, Measuring information-based energy and temperature of literary texts, *Physica A* (2016), http://dx.doi.org/10.1016/j.physa.2016.11.106

Measuring Information-Based Energy and Temperature of Literary Texts

Mei-Chu Chang[1,2], Albert C.-C. Yang[3,4], H. Eugene Stanley[2], and C.-K. Peng[1,3]
[1] *Research Center for Adaptive Data Analysis, National Central University, Jhongli 32001, Taiwan*
[2] *Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA*
[3] *Center for Dynamical Biomarkers, Division of Interdisciplinary Medicine and Biotechnology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts 02215 USA*
[4] *Taipei Veterans General Hospital, Taipei 11217, Taiwan*

(Dated: September 13, 2016)

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 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] and [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