entropy

MDPI

Letter

Discovery of Kolmogorov Scaling in the Natural Language

Maurice H. P. M. van Putten

Department of Physics and Astronomy, Sejong University, Seoul 143-747, Korea; mvp@sejong.ac.kr; Tel.: +82-20-3408-3940

Academic Editors: Maxim Raginsky and Raúl Alcaraz Martínez
Received: 16 February 2017; Accepted: 26 April 2017; Published: 2 May 2017

Abstract: We consider the rate R and variance σ^2 of Shannon information in snippets of text based on word frequencies in the natural language. We empirically identify Kolmogorov's scaling law in $\sigma^2 \propto k^{-1.66\pm0.12}$ (95% c.l.) as a function of $k = 1/N$ measured by word count N. This result highlights a potential association of information flow in snippets, analogous to energy cascade in turbulent eddies in fluids at high Reynolds numbers. We propose R and σ^2 as robust utility functions for objective ranking of concordances in efficient search for maximal information seamlessly across different languages and as a starting point for artificial attention.

Keywords: Shannon information; concordances; ranking; search; attention

1. Introduction

Exponential growth of Internet usage [1] is driving the development of new algorithms to efficiently search text for potentially relevant information. As smartphones will overtake personal computers in Internet traffic by 2020 [2], identifying maximal information in concise text is increasingly important. Objective search may be approached by utility functions based on word statistics in the natural language (e.g., [3]), viewing text as linguistic networks (e.g., [4]).

Herein we discuss utility functions based on Shannon information theory [5,6] which are applicable to snippets of text obtained by key words search. These concordances ([7] and references therein) are extracted from documents obtained from the Internet. Real-world searches require robust utility functions that are well-defined in the face of words that fall outside common dictionaries or in mixed language documents. In a computerized extraction of concordances from large numbers of documents, such utility functions enable the ranking of text, facilitating in-depth document search in any given language.

Shannon quantifies information by surprise factors of symbols used in discrete encoding of messages by probability of occurrence. In digital computer communications, message size is typically much greater than the size of the dictionary of symbols $D = \{0, 1\}$. However, conveying information by snippets generally comprises very few words from a natural language dictionary. Miller [8] refers to such snippets as "chunks". Shannon and Miller hereby discuss principally distinct limits of *large* and *small* message size in data transmission, measured by size relative to the dictionary. Miller posits (without proof) that the Shannon information and variance fulfill similar roles in human communication.

In this Letter, we report on empirical power law behavior in information rate and variance in concordances as a function of size N measured by total word count. A distinguishing feature of the present study is emphasis on relatively short concordances of $N = 20$–200 words, distinct from entire books (e.g., [9]).

To illustrate and set notation, consider a dictionary $D = \{\text{Yes, No, Perfect}\}$. Messaging by a source S_0 with uniform probability distribution $p_i = 1/3$ $(i = 1, 2, 3)$ features an information rate

$$R_0 = -\sum_{w_i \epsilon D} p_i \log p_i = \log 3; \tag{1}$$

that is, $R_0 = \log_2 3$ bits per word, defined by the average surprise factor $-\log p_i$. Messaging over a proper subset $D' = \{\text{Yes, No}\}$ by a source S' with probability distribution $p_i = 1/2$ $(i = 1, 2;\ p_3 = 0)$ features a reduced information rate

$$R = \log 2 \simeq -\sum_{w_i \epsilon D'} p_i \log p_i = \frac{2}{3} R_0, \tag{2}$$

where the right-hand-side provides an approximation based on the probability distribution of S_0. By probabilities, S_0 is preferred over S as a source of messages. In the present study, we aim to rank messages accordingly, in the form of concordances extracted from online text by key word search.

In real-world applications, our symbols of encoding are words with typical frequencies determined by the natural language. Sources may display variations in word probabilities reflecting individual word preferences. For computational analysis of messages across a broad range of sources, we consider a truncated dictionary D of the most common words, whose probability distribution approximates that of all words in a more comprehensive dictionary, dropping words that are exceedingly rare, not traditionally included in dictionaries, or of foreign origin. Such D is readily extracted from a large number of documents, randomly selected over a broad range of subjects. Fixing D, p_i for each $w_i \epsilon D$ is established by normalizing inferred word frequencies,

$$\sum_{w_i \epsilon D} p_i = 1. \tag{3}$$

For instance, {Yes, No, Perfect} have relative probabilities

$$p_{yes} : p_{no} : p_{perfect} = 1 : 2.49 : 0.0559 \tag{4}$$

based on a truncated dictionary D defined by a top list of 10,000 words (Table 1).

Table 1. Probabilities of a truncated dictionary D of the 10,000 most common words in the English language. Words not listed in D are assigned probability zero. Probabilities refer to words regardless of case.

Index	Word	Probability $\times 10^{-3}$	Comment
386	apple	0.0256	Upper and lower case
6481	perfect	0.1525	"
6034	no	6.7556	"
9946	yes	2.7188	"
-	Woolsthorpe	0	Newton's city of birth, *not* in Merriam-Webster

In text, words contribute to the information rate per word according to

$$r_i = -p_i \log p_i. \tag{5}$$

A sum over all w_i in D obtains the mean rate

$$R_D = \sum_{w_i \epsilon D} r_i = 9.11 \text{ bits word}^{-1}. \tag{6}$$

Relative to the maximum $R_0 = 13.30$ bits word^{-1} for a uniform probability distribution ($p_i = 10^{-4}$ for all i), we have

$$R_D = 0.6854\, R_0, \tag{7}$$

illustrated by (2) in our example above.

The result (6) illustrates Miller's classic result on approximately 8–10 mostly binary features identified in speech analysis [8,10]. However, (7) under-estimates information rates in short expressions due to the non-unform probability distribution of words in the natural language (Table 1). It becomes meaningful only in the large N limit, for text approaching the size of a dictionary. While this limit may apply to large bodies of text, it is not representative for messages in direct human-to-human communication or human–machine interactions.

To begin, we consider the above for communications in a data base of 90,094 concordances of size $N = 20$–200 comprising a total of 8,842,720 words, extracted from the Internet by various key word searches and evaluated for Shannon information rates and associated variances (Section 2). Statistical properties are observed to satisfy power law behavior. They are analyzed for their dependence on $k = 1/N$, motivated by a heuristic analogy between information flow across messages of various size and inertial energy cascade over turbulent eddies. The latter serves as a model for energy flow over complex nonlinear dynamics involving a large number of degrees of freedom, satisfying Kolmogorov scaling as a function of wave number k [11–13] (Section 3). Results for our data base of concordances are given in terms of power law indices and compared with Kolmogorov scaling (Section 4). In Section 5, we summarize our findings.

2. A Data-Base of Concordances

We compiled a data-base of 90,094 concordances C with size $N = 20$–200 from thousands of documents on the Word Wide Web (Figure 1). They are extracted by key word search covering a broad range of generic topics in sports, culture, science, and politics [14,15]. For each search, concordances are extracted from about $M = 80$ online source pages identified with the highest document rank by existing Internet search, and downloaded for analysis as described below by parallel computing on a cluster of personal computers. Experimentally, we determined top lists of concordances ranked by information rate (8) (below), which remain essentially unchanged when M reaches 80; M less than 50 occasionally fails to capture concordances of highest text rank as defined below. On this basis, our results are also expected to be reasonably independent of the choice of Internet document search engine.

As snippets, concordances always comprise a small subset of a dictionary, including D containing most common words mentioned above. For N on the order of tens to hundreds of words, R_D in (7) is not directly meaningful for estimating information in concordances due to selection effects: snippets contain few words, many of which are relatively common by non-uniformity of word frequencies in D (Table 1) with relatively significant contributions to information flow rate r_i in (5).

Information rates in concordances are obtained by summing (5) over all distinct words therein,

$$R = -\sum_{w_i \epsilon C'} p_i \log p_i. \tag{8}$$

Our focus is to quantify statistical properties of R as a function of N given D. Here, C' shall denote its list of q distinct words in C that are in D. This reduction avoids over-counting of words, and implicitly assigns probability zero to words that are not in D. The first is important for accuracy, the second renders (8) robust in the face of words that are not included in D, because they are rarely used (Table 1) or of foreign origin. Assigned $p_i = 0$, these exceptional words are not included in (3). Since (8) has a well-defined limit as p_i approaches zero, it applies to real-world text extracted from

documents online by generic search engines. Distinct words C' in C that are in D comprise $q \leq N$ words, typically about one-half of N (Figure 1).

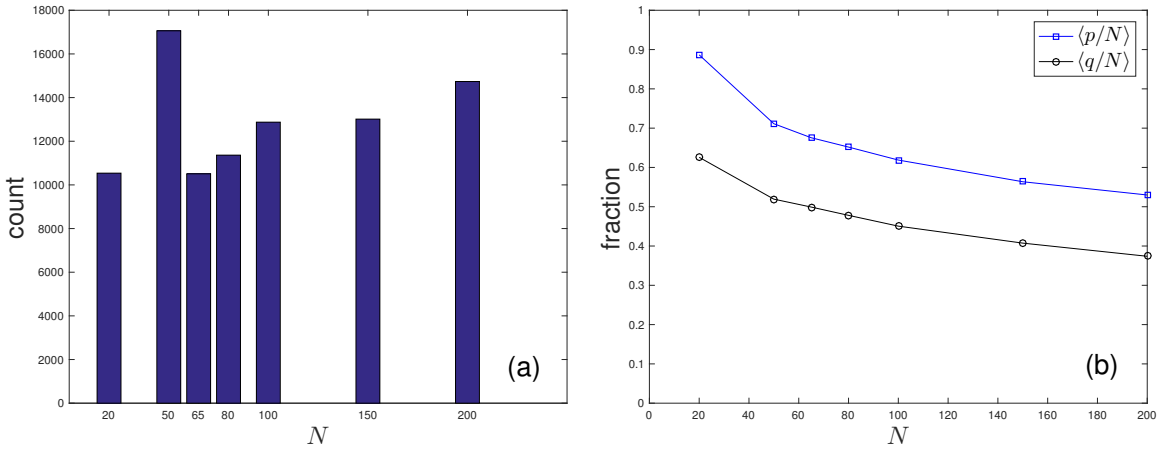

Figure 1. (**a**) Histogram of 90,094 concordances C in the data-base, shown by various sizes N measured by total word count; (**b**) Fractions p/N and q/N of counts of distinct words and, respectively, distinct words in a truncated dictionary D of the 10,000 most common words.

Table 2 illustrates a top list of concordances, obtained by the key word search "apple pie" and text ranked by R. A search with relatively few key and generic words readily obtains well over ten concordances per document (i.e., on the order of one thousand in $M = 80$ documents extracted from the Internet by existing document search). Identification of those potentially most relevant to the user necessitates computerized ranking, here by the utility function R.

Table 2. Sample concordances of $N = 50$ words selected by key words "apple pie" (bold) ranked by information rates R. Included are the standard deviations σ (obtained by [14,15]).

Rank	Concordance	R	σ
1	.. splash of brandy. My homemade **apple pie** is like a siren call to my family. All I have to do is pick up the phone and say "**pie**" to my father and he's here in less time than it takes to clear a place at the table. You know when people ..	1.1202	1.6275
2	.. of pumpkin and **apple** together just make my heart happy. The photos are gorgeous and your lattice is freaking perfect! :) I have never been to an apple orchard either but I always envision it as a marvelous occasion. Maybe one day I will go! Pinning this **pie** for future reference ;) Reply ..	1.0987	1.5144
...	...	...	...
10	.. that I've baked **apple pie**, this recipes was easy to follow AND most importantly it came out delicious. Received lots of compliments on this so Thank You!!!! Curious to know if there are any supplements for the sugar though, trying to make a version for my parents who are trying to cut back ..	1.0263	1.5126

In studying statistical properties of R as a function of concordance size N, we further consider the unnormalized variance

$$\sigma^2 = \sum_{C'} \left(r_i - q^{-1} R \right)^2 \tag{9}$$

suggested by Miller's conjecture on information measured by variance, $N^{-1}\sigma^2$. Dependence on N in the statistical properties of σ^2 will be found to satisfy a power law with index α,

$$\sigma^2 \propto N^{\alpha}. \tag{10}$$

The index α in (10) will be determined by detailed partitioning of information flow in large amounts of text into multiple messages, i.e., in expressions, statements, snippets and the like across different sizes. In the absence of detailed modeling, σ^2 in (10) is expected to be tightly correlated to R in (8) in an intuitive analogy of information and energy flow. In fluid dynamics, representing a nonlinear system with a large number of degrees of freedom, energy flow satisfies Kolmogorov scaling in conservative cascade to small scales in fully developed turbulence [16].

3. Kolmogorov Scaling in Energy Flow

Turbulent motion in high Reynolds number (Re) fluid flow demonstrates power law behavior in energy cascade by nonlinear dynamics that includes period doubling. Its inertial range comprises a large number of degrees of freedom $\propto \mathrm{Re}^{\frac{9}{4}}$, over which energy flow cascades over eddies of size λ, breaking up *conservatively* into increasingly smaller eddies across wave numbers

$$k_{min} \leq k < k_{max}. \tag{11}$$

Here, $k = 2\pi/\lambda$, k_{min} refers to eddies set by the linear size of the system, and k_{max} refers to the wave number at which viscous dissipation sets in. This cascade persists by power input ϵ_0 at k_{min}. In the inertial range (11), the ϵ_0 is conserved across k, posing a constraint on the spectral energy density $E(k)$, satisfying Plancherel's formula

$$e = \int_0^{\infty} E(k)dk \tag{12}$$

with

$$[E(k)] = [\lambda][e]. \tag{13}$$

Since $[\epsilon_0] = [e]\,\mathrm{s}^{-1}$, $[e] = \lambda^2\,\mathrm{s}^{-2}$, dimensional analysis obtains the Kolmogorov scaling

$$E(k) = C\epsilon_0^{\frac{2}{3}}k^{-\frac{5}{3}} \propto \lambda^{\frac{5}{3}}. \tag{14}$$

The Kolmogorov index 5/3 has been found to be remarkably universal in fully developed turbulence, from fluid dynamics [17] to broadband fluctuations in gamma-ray light curves [18].

4. Power Law and Kolmogorov Scaling in Information Flow

By (5), information rates of concordances

$$R = \sum_{C'} r_i < R_D \tag{15}$$

are increasing as a function of size N, formally satisfying

$$\lim_{N\to\infty} \sum_{C'} r_i = \lim_{N\to\infty} -\frac{1}{N}\sum_{C} n_i \log p_i = R_D, \tag{16}$$

where n_i denotes the number of times a word w_i appears in C. The latter is a consequence of the fact that C' approaches D in the limit as N becomes arbitrarily large.

For our data base, Figure 1 shows concordance averages of relative counts p/N of distinct words (avoiding multiplicities of recurring words) and the relative counts q/N of distinct words in D. The average value of p/N is found to be somewhat similar to R/R_0 in (7), with a minor dependence on the choice of N.

Figure 2 shows σ^2 and its correlation to R. Expressed in terms of the normalized standard deviation

$$Y = \sigma R^{-1/2} \propto N^{\frac{1}{2}}, \tag{17}$$

which points to

$$\sigma \simeq 0.2125\, N^{\frac{1}{2}} R^{\frac{1}{2}}. \tag{18}$$

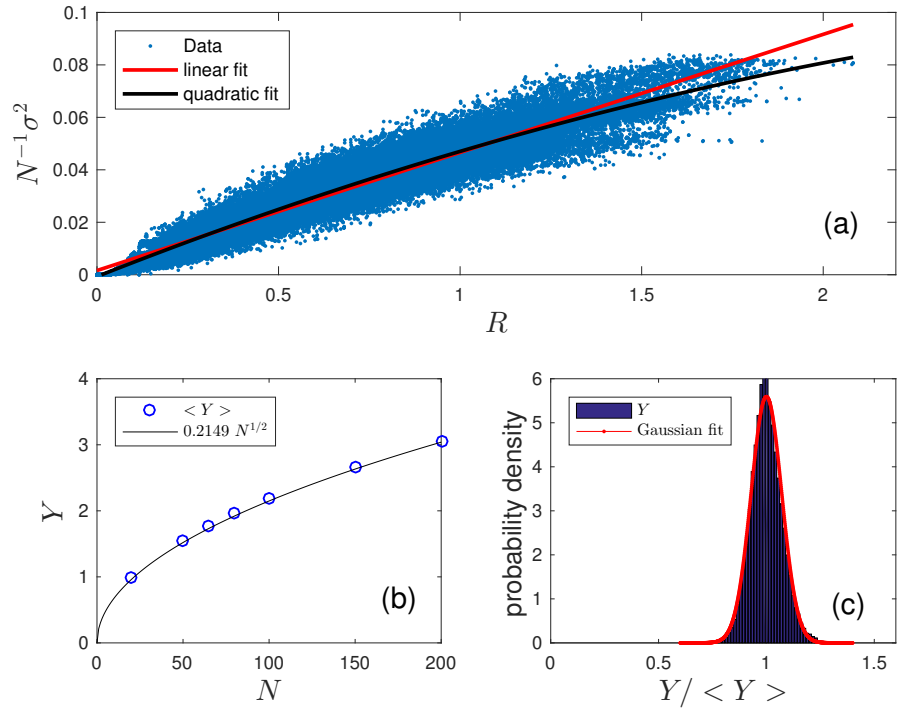

Figure 2. (**a**) $N^{-1}\sigma^2$ correlation with R (all concordances, blue dots) show scaling with R. This correlation is effectively linear for most of R (red curve), evidencing Miller's conjecture on the equivalence of variance with information. However, nonlinearity sets in when R is large (black curve); (**b**,**c**) $Y = \sigma R^{-\frac{1}{2}}$ scales with $N^{1/2}$ with an essentially Gaussian distribution in fluctuations.

Figure 3 shows the power law behavior as a function of $k = 1/N$ obtained by weighted nonlinear model regression computed by the MatLab function *fitnlm* [19] with weights according to concordance counts (Figure 1). The results with 95% confidence levels are

$$\sigma^2 \simeq 0.0013\, k^{-1.66\pm0.12}, \quad R \simeq 0.0381\, k^{-0.62\pm0.08}. \tag{19}$$

Scaling of σ^2 in (19) is consistent with (18) combined with scaling of R in (19).

Figure 3 shows a slight concave curvature in the residuals to the linear fit to the data. While this is within the 95% confidence level shown, this feature may be a real deviation from power law behavior, perhaps associated with nonlinear scaling at large R (Figure 2). A detailed consideration is beyond the present scope, however.

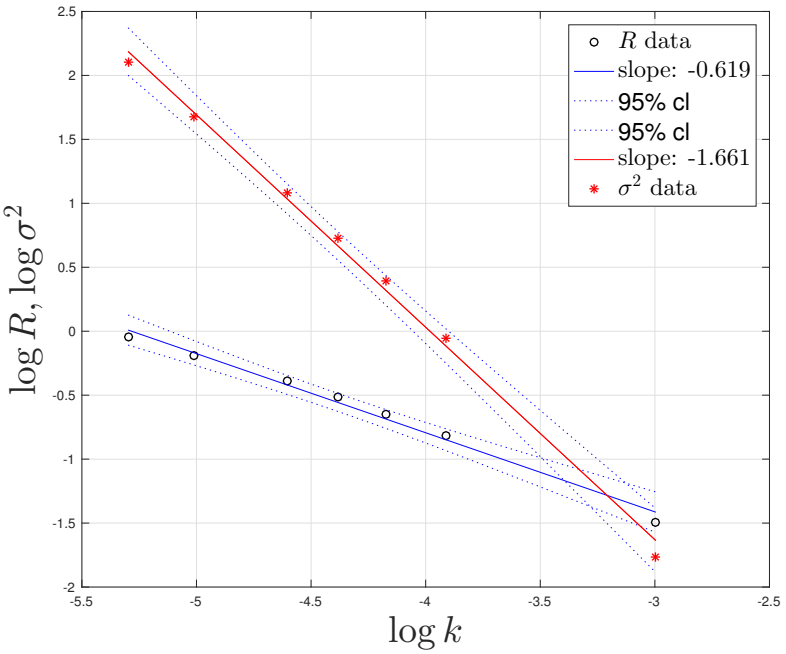

Figure 3. Power law behavior (19) in the mean R and variance σ^2 in a data set of concordances (Figure 1) as a function of $k = 1/N$.

5. Conclusions

We identify Kolmogorov scaling (19) in the variance of concordance information with power law scaling in association with information rate as a function of size $N = 20$–200.

We present Kolmogorov scaling in variance as an empirical result, which suggests that information flow over snippets is analogous to energy cascade over eddies. Since Kolmogorov scaling in the inertial range (11) critically depends on conservation of energy, (19) suggests that perhaps there is a similar conservation law at work in information flow by concordances.

Our observed power law scalings (19) give a succinct statistical summary on communication by concordances in the natural language. At sizes much smaller than the size of a dictionary, the results are fundamentally different from what is obtained in Shannon's large N limit of binary strings in computer-to-computer communications, arising from the strongly non-uniform probability distribution in words in the natural language.

Figure 4 further shows a generally increasing efficiency R/U, normalized to the upper bound U defined by word probabilities sorted in descending order ($p_{i'+1} \leq p_{i'}$, $i = 1, 2, \cdots, |D|$),

$$U = -\sum_{i \leq N} p_{i'} \log p_{i'}. \tag{20}$$

Relatively long concordances show an increase to about 40% efficiency, beyond about 25% in short concordances. *Twitter*'s tweets of 140 characters—corresponding to about 24 words on average at a mean of about five letters and one space per word in English—hereby appears sub-optimal by a factor of about two. Reasonably efficient social networking communication obtains with tweets on the order of 1000 characters.

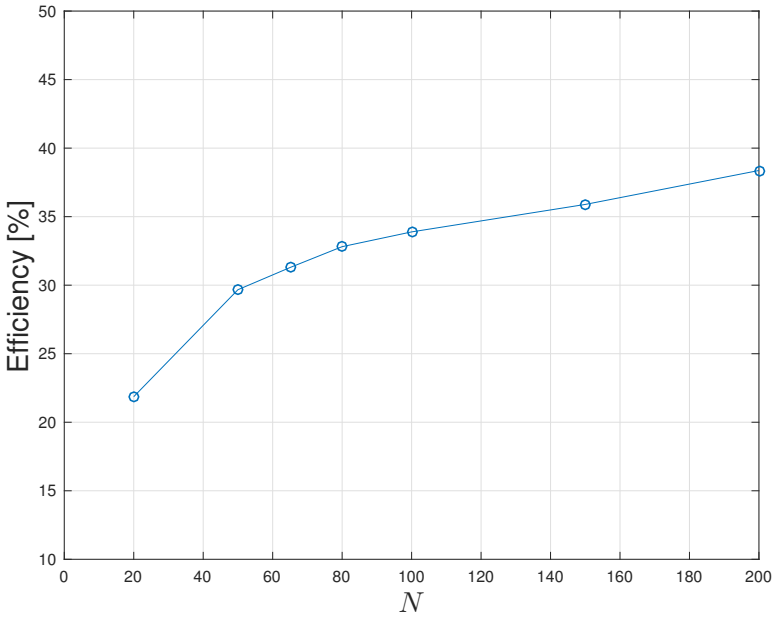

Figure 4. Information rate efficiency R/U as a function of concordance size N.

Based strictly on word frequencies in natural language, R and/or σ provide robust utility functions for objective ranking of snippets by potential relevance. This can be used for efficient search through a large body of documents from a variety of sources by limiting output to a top list of ranked concordances [14,15]. A suitable sample of source documents is readily extracted from the Internet by existing search engines.

A generalization to search seamlessly across different languages is obtained by first translating key words to a second language in which to obtain a body of source documents and concordances therein. Ranking by R and/or σ based on word frequencies in this second language produces a top list that can be translated back to the first language. This process is highly efficient, by limiting translations to a moderate number of concordances, circumventing the need for any document translation in full [20].

Power law behavior (19) in snippets of text also points to novel directions to machine learning. For instance, ranking by R and/or σ may be a first step to artificial attention—with concordances larger in size than tweets—to select snippets as input to further processing (e.g., generating new queries by artificial intelligence).

Acknowledgments: The author thanks the anonymous reviewers for constructive comments. This work was partially supported by the National Research Foundation of Korea under grants 2015R1D1A1A01059793 and 2016R1A5A1013277.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Cisco. The Zettabyte Era: Trends and Analysis, 2014. Available online: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI$_-$Hyperconnectivity$_-$WP.pdf (accessed on 27 April 2017).
2. Cisco Visual Networking Index: Forecast and Methodology, 2015–2020. Available online: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf (accessed on 27 April 2017).
3. British National Corpus, Oxford Text Archive, University of Oxford. Available online: http://www.natcorp.ox.ac.uk/ (accessed on 27 April 2017).
4. Kulig, A.; Drozdz, S.; Kwapien, J.; Oswiecimka, P. Modelling subtle growth of linguistic networks. *Phys. Rev. E* **2015**, *91*, 032810.
5. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.

6. Shannon, C.E. Communication in the presence of noise. *Proc. IRE* **1949**, *37*, 10–21.
7. Wisbey, R. Concordance Making by Electronic Computer: Some Experiences with the "Wiener Genesis". *Mod. Lang. Rev.* **1962**, *57*, 161–172.
8. Miller, G.A. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **1956**, *63*, 81–97.
9. Mehri, A.; Lashkari, S.M. Power-law regularities in human language. *Eur. Phys. J. B* **2016**, *89*, 241.
10. Jakobson, R.; Frant, C.G.M.; Halle, M. *Preliminaries to Speech Analysis: Features and Their Correlates*; MIT Press: Cambridge, UK, 1961.
11. Batchelor, G.K. *The Theory of Homogeneous Turbulence*; Cambridge University Press: Cambridge, UK, 1953.
12. Kolmogorov, A.N. The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. *Proc. R. Soc. Lond. A* **1991**, *434*, 9–11.
13. Orszag, S.A. Analytical theories of turbulence. *J. Fluid Mech.* **1970**, *41*, 363.
14. Van Putten, M.H.P.M. Method to Search Objectively for Maximal Information. U.S. Patent 20130191365A1, 25 July 2013.
15. Van Putten, M.H.P.M. Available online: www.iTopSearch.com (accessed on 27 April 2017).
16. Mathieu, J.; Scott, J. *An Introduction to Turbulent Flow*; Cambridge University Press: Cambridge, UK, 2000.
17. Nieuwstadt, F.T.M.; Boersma, B.J.; Westerweel, J. *Turbulence—Introduction to Theory and Applications of Turbulent Flows*; Springer: New York, NY, USA, 2016.
18. Van Putten, M.H.P.M.; Guidorzi, C.; Frontera, F. Broadband turbulent spectra in gamma-ray burst light curves. *Astrophys. J.* **2014**, *786*, 146.
19. Statistics and Machine Learning Toolbox, MathWorks Inc. Available online: https://www.mathworks.com/stats/index.html (accessed on 27 April 2017).
20. Van Putten, M.H.P.M. Bilingual Search Engine for Mobile Devices. U.S. Patent 20160004697A1, 7 January 2016.