Finding keywords amongst noise: Automatic text classification without parsing

Andrew G. Allison[a], Charles E. M. Pearce[b] and Derek Abbott[a]

[a]Centre for Biomedical Engineering (CBME) and School of Electrical & Electronic Engineering, The University of Adelaide, SA 5005, Australia.

[b] School of Mathematical Sciences, The University of Adelaide, SA 5005, Australia.

ABSTRACT

The amount of text stored on the Internet, and in our libraries, continues to expand at an exponential rate. There is a great practical need to locate *relevant* content. This requires quick automated methods for classifying textual information, according to subject. We propose a quick statistical approach, which can distinguish between 'keywords' and 'noisewords', like 'the' and 'a', without the need to parse the text into its parts of speech. Our classification is based on an F-statistic, which compares the observed Word Recurrence Interval (WRI) with a simple null hypothesis. We also propose a model to account for the observed distribution of WRI statistics and we subject this model to a number of tests.

Keywords: keywords, word recurrence interval, finite mixture distributions, mixed Poisson process, maximum likelihood, Kolmogorov-Smirnov

Introduction

We build on the work of M. Ortuño et al.[1] and our own previous work[2,3] to develop an F-statistic, which allows us to test whether the null hypothesis of a Poisson process with constant rate is likely. We find that keywords generally do have excessive F-statistics and noisewords do not. We examine the way in which the F-statistic scales with sample size and varies with rank. There is the possibility that the upper asymptote of the distribution of the F-statistic, with rank, follows a power law similar to Zipf's law. There is the possibility that the Hurst exponent could be used to classify different types of text.

We have studied the asymptotic behaviour of WRI distributions and found that they do not have fat-tails. The *asymptotic* behaviour of the distributions is completely consistent with the null hypothesis. This has implications for the F-statistic. When the F-statistic is excessive, it is not due to excessive second moment. It is because the first moment is smaller than predicted by the null hypothesis. We postulate that this is because the underlying process is not a homogeneous Poisson process but has two components. The resulting distribution for the WRI is therefore a finite mixture of two geometric distributions. This two-state model accurately predicts the shapes of cumulative distributions of keywords. We test this using the Kolmogorov-Smirnov statistic. The two-state model is further supported by the fact that, if we postulate more than two components in the mixture, for known keywords, then our measure of divergence, the Kolmogorov-Smirnov statistic, does not improve by very much, or may even get worse. We do not need more than two components in our model.

The two-state model predicts a non-zero correlation between successive WRI values. This is consistent with the actual observed autocorrelation in real text. We satisfy this requirement without the need to add any further hypotheses or components to our basic two-state model. The autocorrelation produced by our model is non-causal. At first glance, this seems to be paradoxical. We discuss some of the issues relating to non-causal correlation.

Finally, we find the approach used by M. Ortuño et al. to be very suitable as a first pass for other processes that can be used to classify texts.

Corresponding author: A. Allison: e-mail: aallison@eleceng.adelaide.edu.au, Tel.:+61 8 8303 5283

Noise and Stochastics in Complex Systems and Finance,
edited by János Kertész, Stefan Bornholdt, Rosario N. Mantegna,
Proc. of SPIE Vol. 6601, 660113, (2007) · 0277-786X/07/$18 · doi: 10.1117/12.724655

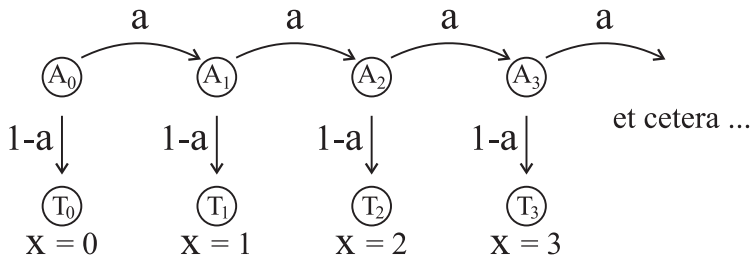

Figure 1. The discrete Poisson process, as a Markov chain: The initial state is A_0, which means 'active with $x = 0$.' This corresponds to the state where no words have yet been examined. In general, the state A_k refers to an active state with $x = k$. Physically, this corresponds to the case where we have examined k words and none of them were the proposed keyword. All states marked 'A' are active, meaning that we are still actively searching for the next instance of the proposed keyword. The state T_k refers to the terminal state with $x = k$. Physically, this corresponds to the case where we have examined $k+1$ words and the first k words were not the proposed keyword and the last word was the desired keyword. All states labelled 'T' are terminal, meaning that the Markov chain has now terminated. The labels on the arrows refer to the conditional probabilities of transition from one state to another. The conditional probability that a word is not used, in any active state, is a and the conditional probability that the word is used is $1 - a$. All events are independent, so if the probability of arriving at the initial state, A_0, is 1 then the probability of arriving at state T_k, with $x = k$, is $P(x) = (1 - a) \cdot a^x$. After the Markov chain has terminated, it immediately restarts again at state A_0, with probability of one, and the whole procedure is repeated until the end of the text is reached. This Markov chain can also be used to generate pseudo-random numbers in order to simulate the null hypothesis.

A null hypothesis for Word Recurrence Interval

For our present purpose, we consider text to be an long sequence of words, separated by white-space, including punctuation and line breaks. For any given word, it is possible to count the number of other words in between successive occurrences of the word. For example, consider the following piece of text, from Lang's translation of the Iliad of Homer:[4] "Of all these, even fifty ships, Achilles was captain. But these took no thought of noisy war; for there was no man to array them in line of battle. For fleet-footed goodly Achilles lay idle amid the ships." In this example, the Word Recurrence Interval (WRI) between the two instances of the word 'Achilles' is 26. According to this definition, WRI values have to be nonnegative integers, in $\mathbf{Z}^*$. We argue that by counting all WRI statistics, for all the words in a text, it is possible to use statistics to determine which words are more likely to be keywords.

The simplest model for WRI statistics is a Poisson process with constant rate, illustrated in Figure 1. Each words used with a certain probability and rates of use, for each individual word, are constant and do not depend on the words that have been written previously. This very simple model forms the null hypothesis that we will test. The output from the process is a random variable x. Since the WRI values are not real numbers, but nonnegative integers, the probabilities are described by a discrete probability mass function. Assume that the probability that the word is not used, in a particular location in the sequence of words, is a. Under the null hypothesis, a is constant throughout the text. The probability that the word is used, is $1 - a$. For a WRI value of x to occur we need x instances of non-use followed by one instance of use. The resulting probability mass distribution is geometric:

$$P(x) = (1 - a) \cdot a^x, \tag{1}$$

for nonnegative integers, x. The moment generating function for this distribution is: $\Omega(y) = E\left[e^{yk}\right] = (1 - a) / (1 - ae^y)$, which can be differentiated, at $y = 0$, to calculate the moments predicted by the null hypothesis. The first central moment is given by

$$\mu = \mu_1 = \frac{a}{1 - a} \tag{2}$$

and the second central moment is given by

$$\sigma^2 = \mu_2 = \frac{a}{(1-a)^2}. \tag{3}$$

These results are stated with slight variations of notation to the literature.[5–9]

An F-statistic and a test for keywords

The WRI values for a particular word in a text is a finite sequence of N values of x, which we can denote by $\{x_1, x_2 \cdots x_n \cdots x_N\}$. The maximum-likelihood estimate for the parameter, a, requires that

$$\frac{a}{1-a} = \frac{1}{N} \cdot \sum_{n=1}^{N} x_n = \overline{X}, \tag{4}$$

where $\overline{X}$ is the sample mean, which can be calculated purely in terms of observable quantities. This gives us the maximum-likelihood estimate for a in terms of observable quantities: $a = \overline{X}/\left(\overline{X}+1\right)$. This has immediate implications for Equations 2 and 3. Once we have used Equation 4 to calculate $\overline{X}$, then we can calculate estimates for the first two central moments: $\hat{\mu}_1 = \overline{X}$ and

$$\hat{\mu}_2 = \overline{X} \cdot \left(\overline{X}+1\right). \tag{5}$$

These equations will hold true, provided that the null hypothesis is true. Of course, it is also possible for us to directly estimate the second moment, based on the empirical data, $\{x_1, x_2 \cdots x_n \cdots x_N\}$. The best unbiased estimate of the sample variance is:

$$S^2 = \frac{1}{N-1}\left(\sum_{n=1}^{N} {x_n}^2 - \frac{1}{N}\left(\sum_{n=1}^{N} x_n\right)^2\right). \tag{6}$$

This is a well established result, which applies to all random variables, regardless of the distribution. This is discussed in Huntsberger,[10] Ross[11] or Wackerly[12] for example.

The two estimates for the variance, in Equations 5 and 6, should always be very similar, when the null hypothesis is true. This leads naturally to a ratio of variances or F-statistic: $F = S^2/\hat{\mu}_2$. Computer simulations reveal that F does have a mean value very close to unity. When the null hypothesis is true, large deviations of the F-statistic are very improbable. Large deviations indicate violations of the null hypothesis. This can be used to form the basis for a formal statistical test. The F-statistic turns out to be only very slightly different from the measure used by Ortuño et al.,* but we have based our work on a more formal and rigorous approach. The difference between our statistic and that of Ortuño et al. would only be significant if $\overline{X}$ were small compared to unity.

Scaling of the F-statistic with sample size

In general, an F-statistic is a scaled ratio of two independent χ^2 random variables, $F = ({\chi_1}^2/\nu_1)/({\chi_2}^2/\nu_2)$, where ν_1 and ν_2 are the numbers of degrees of freedom in ${\chi_1}^2$ and ${\chi_2}^2$ respectively †. The number of degrees of freedom for $\hat{\mu}_2$ is $\nu_2 = N-1$, since only one degree of freedom is used up in the estimation of $\overline{X}$. It would seem that the number of degrees of freedom for S^2 would be $\nu_1 = N-2$, since there are two parameters to be estimated, but this is not completely correct because the variables S^2 and $\hat{\mu}_2$ are not completely independent, because they both rely implicitly on the estimation of $\overline{X}$. These values can be regarded as approximately correct when N large, since S^2 and $\hat{\mu}_2$ will be nearly independent in this case.

*Ortuño et al. used a statistic which was equivalent to $S^2/\overline{X}^2$. This is effectively the square of the 'coefficient of variation' proposed by Karl Pearson and described by Groebner,[13] Wackerly[12] and Upton.[14]

†This is discussed in Upton and Cook[14] and in Abramowitz and Stegun.[8]

Abramowitz and Stegun[8] give an expression for a scaled F-statistic when ν_1 and ν_2 are large:

$$F' = \frac{F - \frac{\nu_2}{\nu_2 - 2}}{\frac{\nu_2}{\nu_2 - 2}\sqrt{\frac{2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)}}}. \tag{7}$$

This reduces to

$$F' \approx \frac{\sqrt{N}}{2} \cdot (F - 1) \tag{8}$$

as $N \to \infty$. The new statistic F' has a normal distribution $F' \sim N(0, 1)$. This tendency towards the normal Gaussian distribution is an example of the central limit theorem. There is a possible issue concerning speed of convergence. To guard against this we tested Equation 8 using extensive computer simulation and the result us shown in Figure 2.

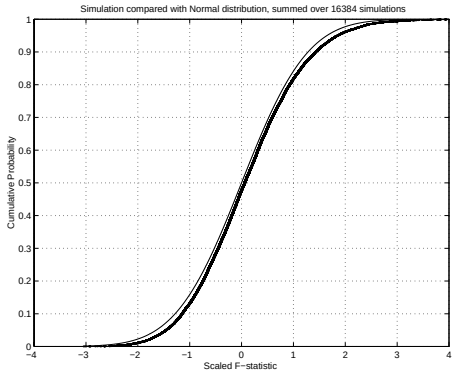

Figure 2. Partial convergence of the F-statistic to a normal distribution. The heavy line is based on data points from a large simulation based on the null hypothesis. The fine line is calculated using the scaling law, stated in Abramowitz and Stegun. The convergence is good, even for small sample sizes of the type found in the *Iliad*.

It should be possible to calculate F' using Equation 8 and then to apply a standard hypothesis test, based on the assumption that $F' \sim N(0, 1)$. The boundaries of the confidence interval for F' can be calculated using the percentage points of the normal distribution. Experience shows that words with excessively high values of F' are more likely to be keywords. Parts of speech which are definitely not keywords typically have smaller values of F', with $|F'| \leq 1$. A list of high-ranking words, with large F-statistic, from Lang's translation of *The Iliad of Homer*[4] is shown in Table 1. Of the top ten ranked words, seven of these are actually keywords in the opinion of the authors. This sorting was achieved without any parsing or attempt to understand the content of the narrative.

Variation of the F-statistic with Rank

If all the words in a text are sorted into rank according to their F-statistic, then it is possible to create a graph of the resulting distribution. An example is shown in Figure 3. It is possible, but not proven, that the F-statistic follows a power law, with the same general form as Zipf's law. The empirically fitted line has the equation $\log_{10}(F + 7.2208) = -0.1545 \cdot \log(R) + 0.9903$ where F is the F-statistic and R is the ranking. This is equivalent to a power law: $F + 7.2208 = 9.7793 \cdot R^{-0.1545}$. The slope, $\gamma = -0.1545$, can be regarded a Hurst exponent. There is some evidence that works of non-fiction (Adam Smith, Michael Faraday, Charles Darwin) have higher Hurst exponents than works of fiction (Homer, Virgil, Dante). This is a topic for further investigation. It may be possible to distinguish between fiction and non-fiction, using the Hurst exponent.

Table 1. The top ten ranked words in the Iliad, as translated by Lang.[4] The column labels refer to: The ranking (in order of F-statistic), The proposed keyword, The scaled F-Statistic F', The frequency of use in the source document, whether the word is a keyword (according to the authors) and some common alternative English spellings. The 'Keyword' column contains a logic '1' if the authors of this paper consider the word to be a keyword (based on an understanding of the text) and the column contains a logic '0' if the word is a noiseword or not a relevant word. The authors' opinion is based on a normal human reading of the text, together with discursive thought and conversation about the meaning and significance of the text.

Rank	Word	F-Statistic	Frequency	Keyword?	Comments
1	patroklos	92.50	140	1	Patroclus
2	thou	67.80	927	0	
3	achilles	59.85	361	1	
4	her	59.45	465	0	
5	hector	56.57	394	1	
6	menelaos	55.21	116	1	Menelaus
7	aias	42.55	119	1	Ajax
8	i	39.02	1011	0	
9	odysseus	37.15	108	1	
10	possessed	35.73	42	1	

The anomalous shape of the cumulative distribution for the Word Recurrence Intervals of keywords

When we compare the cumulative distribution of the WRI statistics for a known keyword, such as 'Achilles' with the cumulative distribution predicted by the null hypothesis, we observe significant divergence between the two curves. This is shown in Figure 4. The maximum divergence between the two curves, in Figure 4, is the Kolmogorov-Smirnov statistic. The maximum divergence in this case is 0.3094. Extensive simulation of the null hypothesis, shown in Figure 5, implies that any deviation greater than 0.1 is highly improbable, assuming the null hypothesis is true. The divergence between the two curves is so great that the null hypothesis is most unlikely and must be rejected. We conclude that the keyword 'Achilles' clearly occurs in a pattern that is not consistent with a pure Poisson process at constant rate. More detailed analysis of the upper asymptote of the empirical curve, in Figure 4, reveals that the upper asymptote is exponential, which is consistent with a Poisson process. The tails are not actually fat at all. The F-statistics, shown in Table 1, are not excessive because of excessive second moment but because the sample mean $\overline{X}$ is smaller than predicted by the null hypothesis. This suggests the presence of some process that affects the first moment more than the second moment.

A two-state model

We postulate that the underlying process is an inhomogeneous Poisson process with two components. The resulting distribution for the WRI statistics is a finite mixture of two geometric distributions.

The concept of a mixture distribution, being made up of two or more component distributions, is well established in the literature and extends at least as far back as Pearson[15] 1894. There is a model, used by Albert,[16] in which a mixture distribution is the output from a two-state Markov chain. There is a different Poisson process associated with each state and the result is a mixture of two Poisson processes. Our model was developed independently of Albert.

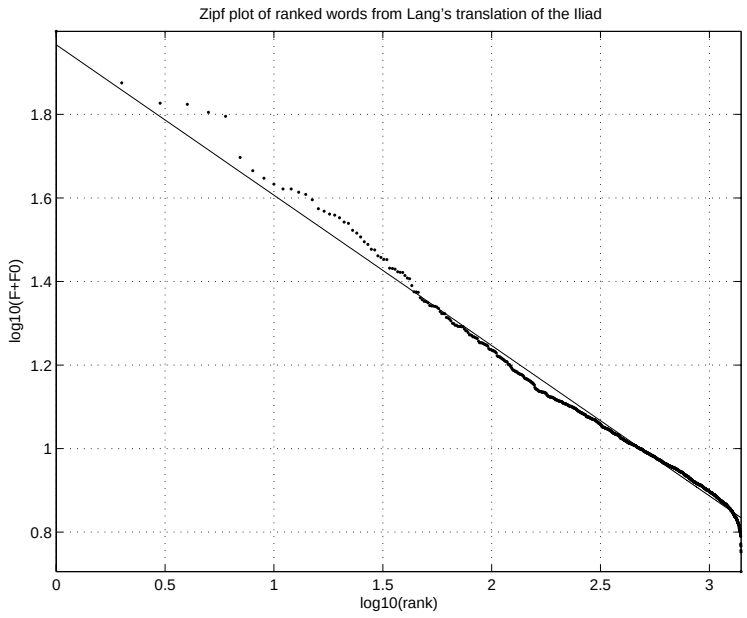

Figure 3. Zipf plot of the data from Lang's translation of the *Iliad*. The fitted line has the equation $\log(F + F_0) = \gamma \cdot \log(R) + \beta$, where $F_0 \approx 7.2208$ and $\gamma \approx -0.1545$ and $\beta \approx 0.9903$. The curve fits the upper asymptote and the bulk of the distribution but generally diverges for the lower asymptote. The offset, F_0 was introduced in order to handle negative values of F. The actual value was chosen in order to minimise the total mean square error. The use of a constant offset in F is easier to accept when we realise that an offset was added in Equation 8 and that the origin was only chosen in order to simplify statistical testing.

There are four parameters to estimate, a_1, a_2, w_1 and w_2. The first two parameters represent the probabilities of the two Poisson processes. These are analogous to the single probability, a, in Equation 1. The second two parameters relate to the proportions in which the Poisson processes act. These are weighting factors. The probability mass function for the mixture distribution is:

$$P_2(x) = \sum_{N=1}^{2} w_N \cdot (1 - a_N) \cdot (a_N)^x. \tag{9}$$

There is also one constraint on total probability: $\sum_{N=1}^{2} w_N = 1$. The process must be in one of the two defined states and produce statistics from one, or other, of the two specified Poisson processes. The generalization to more than two states, or components in the mixture, is possible but not needed.

Estimation of parameters using the method of Maximum-Likelihood

If we have a set of empirical measurements of WRI statistics, $\{X_1, \cdots X_M \cdots X_{M\max}\}$ then we can write down the log-likelihood function:

$$L_2 = \sum_{M=1}^{M\max} \log \left(\sum_{N=1}^{2} w_N \cdot (1 - a_N) \cdot (a_N)^{X_M} \right). \tag{10}$$

We can form the gradients of L_2 with respect to the parameters:

$$\frac{\partial L_2}{\partial w_k} = \sum_{M=1}^{M\max} \frac{(1 - a_k) \cdot (a_k)^{X_M}}{\sum_{N=1}^{2} w_N \cdot (1 - a_N) \cdot (a_N)^{X_M}} \tag{11}$$

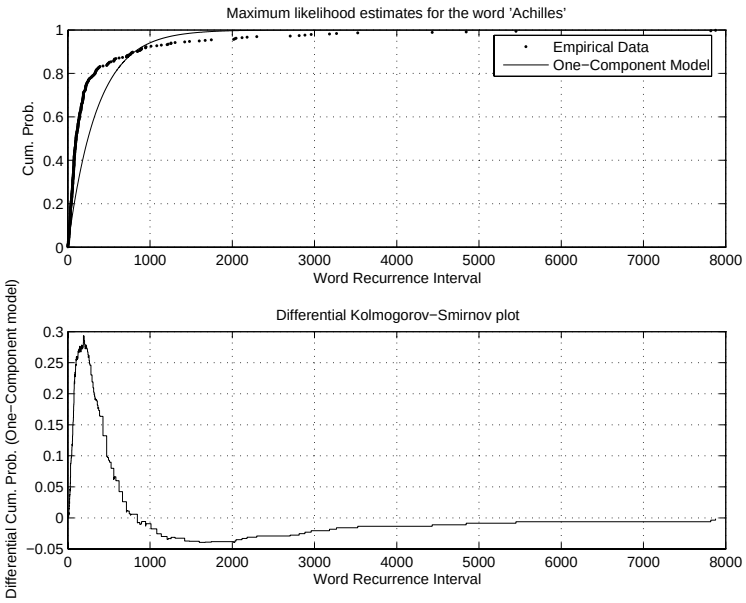

Figure 4. A Kolmogorov-Smirnov plot of empirical data for the word 'Achilles', from the translation of the *Iliad* by Lang.[4] The bottom half shows a plot of the difference between the distribution implied by the null hypothesis and the distribution of the data. The maximum difference between the two curves is the Kolmogorov-Smirnov (KS) statistic, which has the value 0.3094.

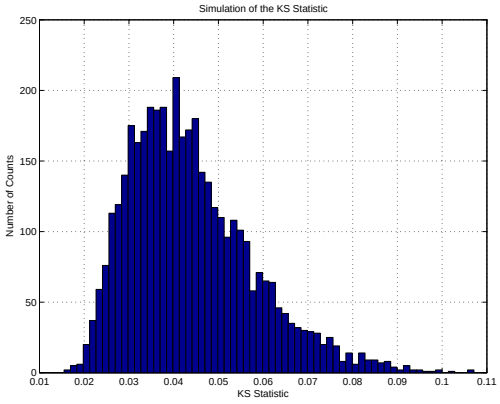

Figure 5. Extensive simulation of the null hypothesis shows that a KS statistic of greater than 0.1 is very improbable. The simulations were carried out in such a way that the scale of the process and the sample sizes were consistent with the data for Achilles. We simulated 4096 instances of the discrete Poisson process in Matlab. We calculated the associated KS statistics, which are represented as a histogram. Using this simulation, we can show that extreme deviations of the KS statistic are unusual, when the null hypothesis is true.

and

$$\frac{\partial L_2}{a_k} = \sum_{M=1}^{M\max} \frac{w_k \cdot (a_k)^{X_M} \left(\frac{(1-a_k)}{a_k} \cdot X_M - 1 \right)}{\sum_{N=1}^{2} w_N \cdot (1-a_N) \cdot (a_N)^{X_M}} \tag{12}$$

This is a constrained maximization problem and it is possible to use the principle of the Lagrange multiplier to arrive at necessary conditions for a maximum:

$$\frac{\partial L_2}{w_k} = \lambda \tag{13}$$

for some constant Lagrange multiplier, $\lambda \in \Re$, and

$$\frac{\partial L_2}{a_k} = 0. \tag{14}$$

Equations 13 and 14 can be solved using standard numerical methods, such as Newton's method. This is described with great clarity by Press et al.[17] and in Conte and de Boor.[18] The same data, for the keyword 'Achilles', in Figure 4, was processed in this way to yield: $[a_1\ ,\ a_2] = [0.9996\ ,\ 0.9943]$ and $[w_1\ ,\ w_2] = [0.0999\ ,\ 0.9001]$. The resulting Kolmogorov-Smirnov plot is shown in Figure 6. The Kolmogorov-Smirnov

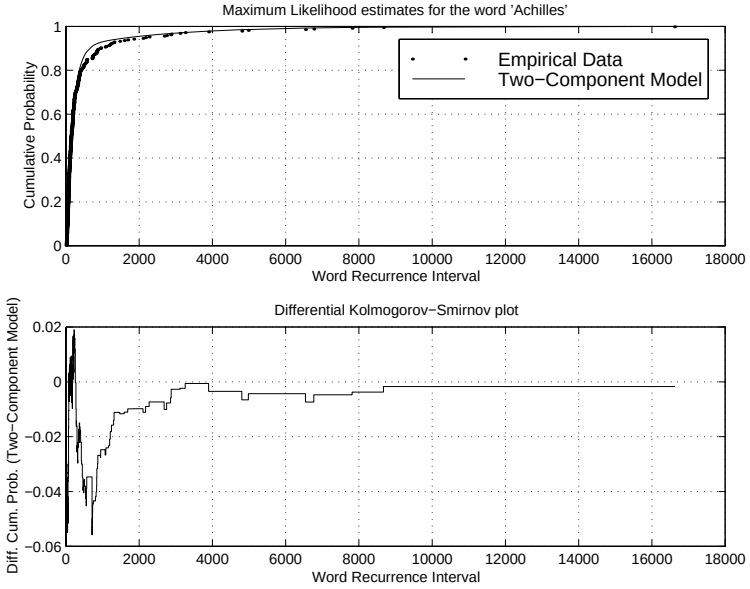

Figure 6. A Kolmogorov-Smirnov plot of empirical data for the word 'Achilles', from the translation of the *Iliad* by Lang.[4] The bottom half shows a plot of the difference between the distribution implied by a two-component mixture distribution and the empirical distribution of the data. The Kolmogorov-Smirnov statistic is now 0.0558, which is quite probable, given the new model. The empirical data is identical with that used in Figure 4, only the model has changed.

statistic is now 0.0558 which is well within the range of probable values suggested by the simulation in Figure 5. We cannot exclude the two-component mixture model described by Equation 9 in the same way that we rejected the null hypothesis, described by Equation 1. This is not a conclusive proof that the resulting distribution is a mixture of two components, but it is a sufficient condition and is suggestive.

Autocorrelation of adjacent WRI values x_t and x_{t-1}

We postulate a discrete Markov process with two states. This is shown in Figure 7. Systems of this type are described in the literature, see Norris,[19] for example. The two state model is simple, generic and well understood. It is intended to describe whether or not a particular keyword has a close relationship to the topic being discussed at any given point in the text. If the keyword is related, then it will have a high rate

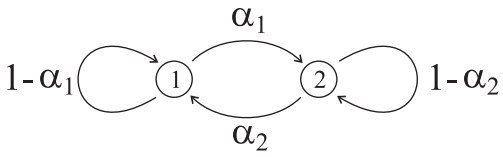

Figure 7. A two-state Markov process: This is one possible mechanism for producing the mixture of two discrete Poisson processes. When the process is in 'state 1' the first Poisson process is active and has the parameter value $a = a_1$ and when the process is in 'state 2' the second Poisson process is active and has the parameter value $a = a_2$. This gives rise to frequent changes in the value of the parameter, a, which generates a random variable with the long-term time-average density function described in Equation 9.

of use; and we regard the WRI value as being generated by a discrete Poisson process with a low mean WRI value. Conversely, if the keyword is not related, then the keyword may occur but only infrequently and we regard the WRI value as being generated by a discrete Poisson process with a high mean WRI value. The mean values are given by $\mu_1 = a_1/(1 - a_1)$ and $\mu_2 = a_2/(1 - a_2)$, which gives $[\mu_1, \mu_2] = [2499, 174]$.

We equate the weights, w_1 and w_2 in 9, with the steady state equilibrium rates of occupancy, so $w_1 = \alpha_2/(\alpha_1 + \alpha_2)$ and $w_2 = \alpha_1/(\alpha_1 + \alpha_2)$. We have estimates of the weights but we cannot estimate the transition probabilities $\{\alpha_1, \alpha_2\}$ directly, but only their ratio: $\eta = w_1/w_2 = \alpha_2/\alpha_1$. To calculate the actual *values* of the transition probabilities, we will need an additional equation. A logical approach is to match the autocorrelation predicted by the two-state model with the autocorrelation observed in the sequence of empirical WRI values.

We wish to calculate the autocorrelation between successive WRI values, $x_t\ x_{t-1}$, predicted by the two-state model. This requires an estimate of $E[x_t \cdot x_{t-1}]$. Fortunately, we can use the fact that the individual discrete Poisson processes are independent. They have no causal influence on each other in any way, which means that the expected value of the product is the product of the expected values: $E[x_t \cdot x_{t-1}] = E[x_t] \cdot E[x_{t-1}] = \mu_t \cdot \mu_{t+1}$. We can sum these expected values over the rates of transition between the recurrent states of the Markov chain to get:

$$E[x_t \cdot x_{t+1}] = \left(\left(\eta {\mu_1}^2 + {\mu_2}^2 \right) - \alpha_1 \eta \left(\mu_1 - \mu_2 \right)^2 \right) \frac{1}{1+\eta}. \tag{15}$$

We can use the definition of the Nth non-central moment, $M_N = E\left[x^N\right]$, and of the mixture distribution, in Equation 9, to show that the non-central moments obey the scaling law $M_N(X_{1\&2}) = w_1 \cdot M_N(X_1) + w_2 \cdot M_N(X_2)$, which leads to :

$$r = \frac{\eta}{1+\eta} \cdot \left(\mu_1 - \mu_2 \right)^2 \cdot \left(\frac{1}{1+\eta} - \alpha_1 \right). \tag{16}$$

Equations 15 and 16 are completely general and do not depend on the details of the base distributions. The equations do hold for geometric distributions but will also hold for more general distributions[‡]. The variance is given by

$$\sigma^2 = \left(\eta(2+\eta) {\mu_1}^2 - 2\eta \mu_1 \mu_2 + (1+2\eta) {\mu_2}^2 \right) \cdot \frac{1}{(1+\eta)^2} + (\eta \mu_1 + \mu_2) \cdot \frac{1}{(1+\eta)}. \tag{17}$$

The equation for the variance does depend on the specific distribution in Equation 9. We can use Equations 16 and 17 to calculate the correlation coefficient:

$$\rho = \frac{r}{\sigma^2}, \tag{18}$$

[‡]The derivations for Equations 15 and 16 are not difficult but they are too long for the present paper. The basic approach is to sum expected values, over all transitions between recurrent states, in proportion to their rates of occurrence. We have checked these equations numerically and we will publish the full derivation shortly.

but we have empirical measurements of ρ from the empirical WRI statistics; so we must have $\rho = 0.2959$ and we can use Equations 18, 17 and 16 to calculate the value for α_1 and hence obtain $[\alpha_1, \alpha_2] = [0.2762, 0.0309]$. Using the Markov model, shown in Figure 7, we can explain the proportions of the mixing process and the correlation observed between successive WRI values.

Non-causal correlation

We have found that embedding completely uncorrelated random-number generators inside a discrete Markov chain introduces correlation, even though the random number generators are not related in any causal manner.

The fact that causally un-related processes can be correlated is very counter-intuitive. We might suspect a mistake but the predictions of Equation 18 agree very well with simulations. The effect is a real effect that requires some explanation.

In our opinion, this counter-intuitive behaviour is easiest to understand as an artefact of the way in which the independent discrete Poisson processes are sampled. If we sample either Poisson process on its own then the results are not correlated. The independent Poisson processes have no correlation between each other. The correlation only appears when we choose to sample Poisson process number one, or Poisson process number two, using a two-state Markov chain. The Markov chain has some memory of whether the keyword is being used at a high rate, or at a low rate. This memory effect causes the balance of first and second moments, in the correlation coefficient, to be altered. We expect that this type of non-causal correlation will be quite general since Equations 15 and 16 are completely general and do not depend on the details of the base distributions.

We note that this form of non-causal correlation has a formal similarity to Parrondo's games,[20–25] where detailed balance is affected by switching the transition probabilities between two different Markov chains.

Conclusions and open questions

Physical explanation for an apparent paradox

It appears strikingly paradoxical that we have shown that random switching between two uncorrelated random sequences produces a sequence with non-zero correlation. How are we to physically understand the origin of this effect? What is the physical picture?

Firstly, it should be noted that correlation does not necessarily imply a causal relationship. In this case the original processes were random, and so the resulting paradoxical correlation is certainly not causal.

Secondly, note that the process is not invariant to time reversal. In other words, if we take a movie of the final sequence being produced from the two original uncorrelated sequences, then run the movie backwards, we *can* see a difference.

This irreversibility can be viewed as a mixing process in a thermodynamic sense. We cannot easily unravel the final sequence to reconstruct the two originating sequences. Such thermodynamic irreversibility always implies an increase of entropy or *loss* of information. Now, the originating sequences had no autocorrelation (i.e. were random) and thus from an information-theoretic viewpoint they carry *maximal* information as they are incompressible. By switching between these information-rich sequences, in an irreversible manner, we increase entropy and thus lose information. This is equivalent to increasing redundancy in the final sequence, which makes it compressible and thus it must have non-zero autocorrelation, as we verified in Equation 18. So in summary, whilst this result is initially very surprising, once we realize that all we are doing is irreversibly mixing two information-rich sequences, it is now easy to see that all we are really doing is simply losing information and thus the 'sting' is removed from the tail of the paradox.

The open question for future work is to now rigorously recast the above argumentation into a mathematical information-theoretic framework.

Application as a first-pass of a sequential process

The method of Ortuño et al. clearly does generate preferential short-lists of keywords, such as the one in Table 1. These lists are not perfect; we must remember that they are only statistical. In this respect, the method is similar to the Bayesian methods[26] that are used for the detection of spam e-mail.[27] We view the use of computers to classify texts as a tool that can be used by human experts, rather than a replacement for human experts. It is clear that we are not going to replace librarians, editors or classical scholars in the near future.

The method does not work well for fiction, especially poetry, but then poetry is not written for the purpose of conveying accurate factual information about a definite topic. Human experts can have very different opinions, regarding what a work of fiction is actually about. Possibly that the best that can be achieved is to produce a list of the key characters in the narrative, like the names in Table 1. In addition, analysis of the Hurst exponents in the Zipf plots, such as Figure 3, may tell us whether or not it makes sense to attempt to automatically allocate keywords to a given narrative.

We believe that the method will find its widest application as part of a sequential process of evaluation. For example certain words, like 'thou' 'her,' and 'I' could be excluded, or have reduced weighting, on other grounds. The Bayesian[26] framework would seem to be the best paradigm for this.

Summary of results

We make the following claims for our model:

- We have verified the work of Ortuño et al. and provided an explanation for their results. The ratios of the first and second moments become anomalous, because keywords are generated by a process that has more than one rate of word-use. Two rates, or states, or transition probabilities, a_1 and a_2 do seem to be needed in order to adequately model the observed WRI statistics.
- Ortuño's scaled-variance statistic was a useful innovation but our F-statistic provides a more rigorous basis for the work, since it compares the observed WRI statistics against a definite null hypothesis. The divergence, between the null hypotheses and the observed WRI statistics can be compared against a well understood statistical process so it is easy to construct formal tests of significance for the F-statistic.
- Our model is a natural extension of a simple Poisson process. We postulate two discrete Poisson processes, embedded within a two-state Markov chain. This essentially models the way in which an author might use a keyword with different rates in different sections of a text.
- Our model has good predictive ability and only uses four parameters: $[a_1, a_2, w_1, w_2]$.
- Our model provides an accurate description of the entire distribution of the WRI statistics, as shown in Figure 6.
- Our model accounts for the observed correlation between adjacent WRI values, $\rho(x_t, x_{t-1})$.
- Our investigation of autocorrelation has revealed that we can use the correlation coefficient to determine whether, or not, a word is a keyword. Our model imposes a relationship between the F-statistic and the correlation coefficient, which is testable. This makes our model testable, and therefore scientific in the sense used by Popper.[28]

Acknowledgements

The authors thank Matthew J. Berryman for the use of his word-counting Perl script and for his earlier work, verifying the use of the F-statistic. We also thank Mei Sheong Wong for her suggestions during the proofreading stage of this paper.

REFERENCES

1. M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz, and A. M. Somoza, "Keyword detection in natural languages and DNA," *Europhys. Lett.* **57**(5), pp. 759–764, 2002.
2. M. J. Berryman, A. G. Allison, P. Carpena, and D. Abbott, "Signal processing and statistical methods in analysis of text and DNA," in *Biomedical Applications of Micro- and Nanoengineering*, D. V. Nicolau, ed., **4937**, pp. 231–240, SPIE, November 2002.
3. M. J. Berryman, A. Allison, and D. Abbott, "Statistical techniques for text classification based on word recurrence intervals," *Fluctuation and Noise Letters* **3**, pp. L1–L10, March 2003.
4. A. Lang, W. Leaf, and E. Myers, *The Iliad of Homer, Lang, Leaf, Myers trans.*, Random House, 1950.
5. A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, Inc., 1991.
6. R. D. Yates and D. J. Goodman, *Probability and Stochastic Processes*, John Wiley & Sons, Inc., 1998.
7. P. Z. Peebles, *Probability, Random Variables and Random Signal Principles*, McGraw-Hill, Inc., 2001.
8. M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, Dover Publications Inc., New York, 9th ed., 1970.
9. E. W. Weisstein, *CRC Concise Encyclopedia of Mathematics*, CRC Press, New York, 1999.
10. D. V. Huntsberger and P. Billingsley, *Elements of Statistical Inference*, Allyn and Bacon, Inc., 4th ed., 1977.
11. S. M. Ross, *Introduction to probability and statiustics for Engineers and scientists*, John Wiley & Sons, New York, 1987.
12. D. W. Wackerly, W. Mendenhall, and R. L. Scheaffer, *Mathematical Statistics with Applications*, Duxbury Press, 1996.
13. D. Groebner and P. Shannon, *Business Statistics*, Merrill Publishing Company, 1985.
14. G. Upton and I. Cook, *The Oxford Dictionary of Statistics*, Oxford University Press, 2002.
15. K. Pearson, "Contribution to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society* **185**(A), pp. 71–110, 1894.
16. P. Albert, "A two-state Markov mixture model for a time series of epileptic seizure counts," *Biometrics* **47**, pp. 1371–1381, Dec. 1991.
17. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, Cambridge University Press, 1992.
18. S. D. Conte and C. de Boor, *Elementary numerical analysis*, McGraw-Hill book company, Aukland, third ed.
19. J. R. Norris, *Markov chains*, Cambridge University Press, 1997.
20. G. P. Harmer, D. Abbott, P. G. Taylor, and J. M. R. Parrondo, "Parrondo's paradoxical games and the discrete Brownian ratchet," in Abbott,[29] pp. 149–160.
21. G. P. Harmer, "Parrondo's paradox," *Statistical Science* **14**, pp. 206–213, 1999. Parrondo's Games.
22. G. P. Harmer and D. Abbott, "Losing strategies can win by Parrondo's paradox," *Nature* **402**, p. 864, 1999.
23. G. P. Harmer and D. Abbott, "The paradox of Parrondo's games," *Proc. Royal Soc., Series A, (Math. Phys. and Eng. Science)* **1994**(99), pp. 247–259, 2000.
24. C. E. M. Pearce, "Parrondo's Paradoxical Games," in Abbott,[29] pp. 420–425. Parrondo's Games.
25. C. E. M. Pearce, "Entropy, Markov Information Sources and Parrondo's Games," in Abbott,[29] pp. 426–431.
26. E. T. Jaynes, *Probability theory the logic of science*, Cambridge University Press, Cambridge, 2003.
27. G. Robinson, "A statistical approach to the spam problem," *Linux Journal* , pp. 58–64, March 2003.
28. A. F. Chalmers, *What is This Thing called Science?*, Queensland University Press, St. Lucia, Qld., 3rd ed., 1999.
29. D. Abbott, ed., *Proc. 2nd. Int. Conf. Unsolved Problems of Noise and Fluctuations (UPoN'99)*, **511**, American Inst. Phys., 2000.