Fluctuation and Noise Letters
Vol. 3, No. 1 (2003) L1–L10

World Scientific
www.worldscientific.com

STATISTICAL TECHNIQUES FOR TEXT CLASSIFICATION BASED ON WORD RECURRENCE INTERVALS

M. J. BERRYMAN*, A. ALLISON and D. ABBOTT
Centre for Biomedical Engineering (CBME) and
School of Electrical & Electronic Engineering
The University of Adelaide, SA 5005, Australia
**mattjb@eleceng.adelaide.edu.au*

Received 27 July 2002
Revised 27 November 2002
Accepted 29 December 2002

We present a method for characterizing text based on a statistical analysis of word recurrence interval. This method can be used for extracting keywords from text, and also for comparing texts by an unknown author against a set of known authors. We also use these methods to comment on the controversial question of who wrote the letter to the *Hebrews* in the New Testament.

Keywords: Stylography; text authorship; statistics of text; keyword extraction.

1. Background

The decision as to whether two texts were written by the same author is usually a difficult one. Can an analysis of how the words in a text statistically cluster shed some light on authorship? In this paper we examine both English texts and the Greek source texts of the New Testament.

The mathematical techniqes developed by Shannon [1,2] and Markov have been used for a number of years to analyse sequences of data, whether this be computer code, text, or DNA. These techniques and other probability-based techniques have enjoyed a large amount of usage in analysing DNA sequences [3] well as both written and spoken text [4,5]. Applications of linguistic methods to DNA sequence analysis have been explored by Dong and Searls [6] and others, and this is our motivation for exploring linguistic techniques for authorship (the corresponding problem in the field of DNA research is the phylogeny of organisms based on their DNA sequences). A seminal work in the area of authorship is Mosteller [7], a good overview of other work can be found in Oakes [8]. Durbin *et al.* [9] is a good reference of work done in analysing DNA sequences.

Ortuño *et al.* [10] suggest using standard deviation of the 'inter-word spacing'

to characterize word distributions and extract keywords, as opposed to using a frequency count of each word. Since inter-word spacing is commonly used to refer to the physical spacing between two words on a page or screen [11], we use the term 'word recurrence interval' (or WRI) instead. By word recurrence interval, we mean the number of words in between successive occurrences of a keyword (non-inclusive), for example if the keyword is "text", then the spacing in "text text" is zero, and the spacing in "The text is an interesting text" is three as there are three words between the two occurrences of the word "text". Initial results of plotting the scaled standard deviation of word spacing for (almost) all the words in a given text against the plot of those from other texts by the same author reveals that works by the same author have a similar distribution of word recurrence intervals. Given the work by Kac [12], in which he derives the result that the recurrence interval of a sequence w is proportional to $1/P(w)$ for $P(w)$ the probabability of the sequence, it is perhaps not surprising that a distinct usage of words (the sequences w) gives a distinct distribution of spacing. The work by Kac [12] has also been used in exploring the entropy of English text [13]. In this paper, we show a striking result obtained when we plot the WRI curves for the gospels of *Matthew* and *Luke*, and the book of *Acts* from the *Koine* Greek New Testament. This suggests a statistical approach to stylography could be taken, and we demonstrate the use of this for both texts by known authors, and for texts where the historical authorship is unclear.

In this paper we explore a number of statistical measures for determining authorship of documents and extracting keywords. We show graphical and numerical methods for comparing authorship, and present two statistical methods for extracting keywords.

2. Mathematical Analysis

2.1. *Standard deviation graphs*

Here we detail a method of displaying the WRI distribution for texts and conjecture that this can give a valuable insight into the authorship of texts.

Given a set of word spacings $\{x_1, \ldots, x_n\}$ for a given word, we compute the scaled standard deviation of WRIs,

$$\hat{\sigma} = \frac{1}{\bar{x}} \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} \,. \tag{1}$$

The reason for the scaling is to eliminate the dependence on word frequency, so $\hat{\sigma}$ values are directly comparable for words in the same text and between different texts. We repeat this for all the words in a text, giving us a set of scaled standard deviations $\{\hat{\sigma}_1, \ldots, \hat{\sigma}_m\}$. In order to generate the graphs we then rank these $\hat{\sigma}_j$ in order and plot scaled standard deviation vs. $\log_{10}(\text{rank})$. We omit those words occurring five or fewer times as being statistically insignificant. Note this method is similar to the methods examined by Zipf [14] and Mosteller [7], however we are examining the scaled standard devation of WRI and not word frequency. The seminal paper on the scaled standard deviation of WRI was by Ortuño *et al.* [10]. Although the idea of taking the scaled standard deviation, in (1), and then using

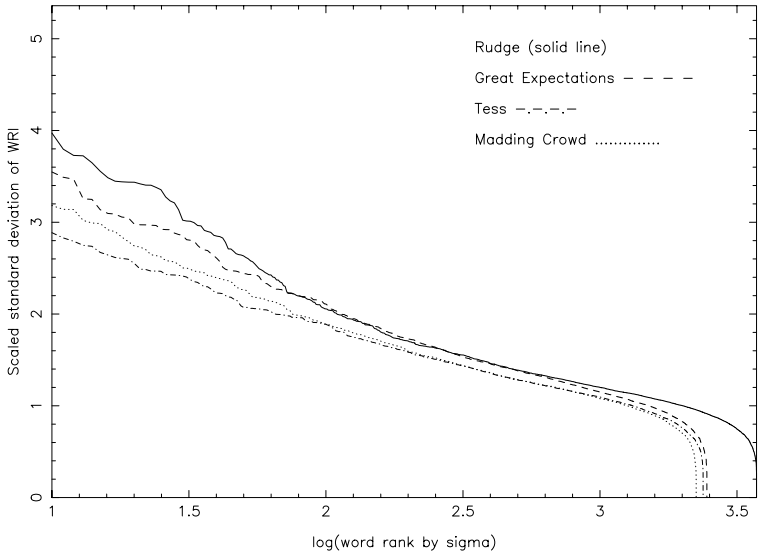

Fig. 1. Scaled standard deviation of WRI vs. log(rank) for books by Charles Dickens and Thomas Hardy. This is a Zipf-like plot based on σ_{WRI} rather than word frequency. For much of the length of the plots the graphs for the Dickens texts are nearly coincident, and likewise for the pair of Hardy texts. Although the plots are all quite different for the region 1 to 1.7, note that this apparently large region is quite small due the logarithmic scale on the x-axis. The extra length on the plot for *Barnarby Rudge* is simply due to the larger number of different words in this text when compared with the other three texts.

it as the y-variable in a Zipf-like plot is a trivial step we note it was first suggested by Carpena *et al.* [15].

Figure 1 shows the similarity between works by Charles Dickens (*Great Expectations* and *Barnaby Rudge*) and works by Thomas Hardy (*Tess of the d'Urbervilles* and *Far From the Madding Crowd*). Figure 2 shows the result of applying the method to the gospels of *Matthew* and *Luke*, and the book of *Acts*. Note that we have used the *Koine* Greek sources for the New Testament [16] to eliminate any changes in style due to translation.

In order to quantify differences in WRI between different authors' usage of keywords we introduce a chi-squared metric in Sec. 2.2. We apply the scaled standard deviation of WRI and an F-statistic to the problem of keyword extraction in Sec. 2.3.

2.2. *Chi-squared tables*

This method compares texts by focusing on words common to all the texts that come from the region of interest in the graphical method (those with a high scaled standard deviation of WRI). In order to compare several texts, we examine those words common to all texts in the selection. We rank words in descending order according to the maximum of the products $f\hat{\sigma}$ for a given word across all texts, where f is the number of times the word occurs in a text. We do this in order to try and pick those words that are statistically significant and have a high scaled standard deviation (and are thus useful keywords). We pick a selection of 30 of those words which are

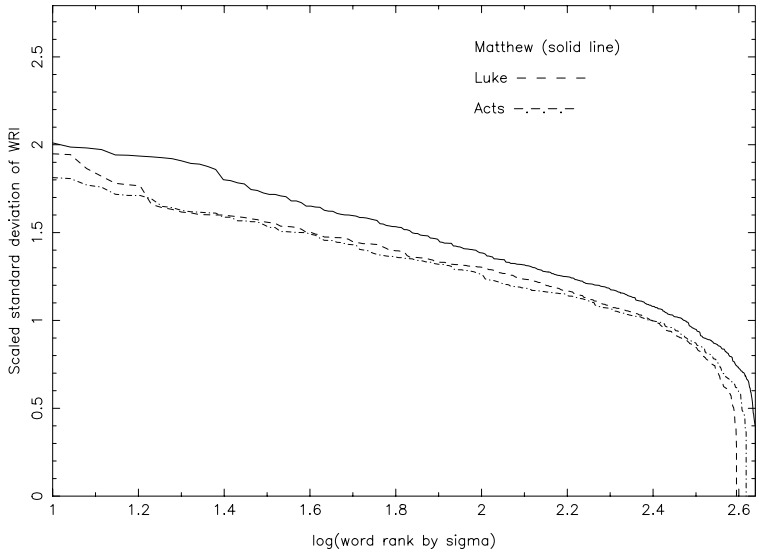

Fig. 2. A Zipf-like plot showing the scaled standard deviation of the WRI (y-axis) for each word is ranked in descending order on a logarithmic scale (x-axis). Using the original *Koine* Greek text, a remarkably close match is obtained between the gospel of *Luke* and the book of *Acts* in the New Testament, which were written by the same author. For reference, a curve of a different author is shown (the book of *Matthew*) illustrating a distinct difference (this is the upper curve). Although the match between *Luke* and *Acts* deviates for a log rank < 1.2, this represents less than four per cent of the total curve (due to the base-ten logarithmic scale). Note that uncommon words occurring less than 5 times in each text are not included in the ranking, as their scaled standard deviation values are not statistically significant.

common (for many of the texts we are considering, especially the shorter bibilical texts, there are often very few words in common to the set of texts with both a high scaled standard devaition that are statistically significant). We thus obtain sets of variances of word spacings for all the texts, $\{\hat{\sigma}^2_{11},\ldots,\hat{\sigma}^2_{I1}\},\ldots,\{\hat{\sigma}^2_{1J},\ldots,\hat{\sigma}^2_{IJ}\}$, for words $i = 1,\ldots,M$ and texts $j = 1,\ldots,J$. Then we use a formula for χ^2 [17] for a pair of texts $(k,l) \in \{1,\ldots,J\} \times \{1,\ldots,J\}$.

$$\chi^2_{kl} = \frac{1}{N_k N_l} \sum_{i=1}^{I} \frac{\left(N_l \hat{\sigma}^2_{ik} - N_k \hat{\sigma}_{il}\right)^2}{\hat{\sigma}^2_{ik} + \hat{\sigma}^2_{il}}\,, \tag{2}$$

$$N_k = \sum_{i=1}^{I} \hat{\sigma}^2_{ik}\,, \tag{3}$$

$$N_l = \sum_{i=1}^{I} \hat{\sigma}^2_{il}\,. \tag{4}$$

We thus generate a table of χ^2 values for each pair of texts. A lower χ^2 score indicates a close match between texts. To normalize the values between different tables, we multiply each χ^2 by a factor of $30/I$, 30 being the maximum number

of words we analyze. When we analyze texts where there are not enough words in common to find 30, then the values are scaled up, since we know less about how well the texts match.

To see how well the chi-squared tables match up English texts, we generated the following tables of Dickens' & Hardy's texts (see Table 1), and Oscar Wilde's *The Picture of Dorian Gray* and *Lord Arthur Savile's Crime and Other Stories* and Lewis Carroll's *Alice in Wonderland* and *Alice Through the Looking Glass* (Table 2) – as expected low scores occur for known author matches. Table 3 shows the results for the gospels plus the book of *Acts*.

Table 1. Chi-squared table for Dickens and Hardy. Here the pairs of texts by the same author get distinctly lower scores than the scores for the pairs by different authors.

	Great Expectations	**Rudge**	**Tess**	**Madding Crowd**
Great Expectations	0	3.13	6.34	8.40
Rudge	3.13	0	7.31	6.48
Tess	6.34	7.31	0	1.96
Madding Crowd	8.40	6.48	1.96	0

Table 2. Chi-squared table for Wilde and Carroll. Note the pairs of text pair up in terms of lowest scores, however Carroll's *Alice in Wonderland* seems quite similar to Wilde's *Lord Arthur Savile's Crime and Other Stories*. This suggests that the fact the pairing of *Dorian Gray* with *Lord Arthur Savile's Crime and Other Stories* is not statistically significant in determining authorship.

	Dorian Gray	**Savile's Crime**	**Alice**	**Looking Glass**
Dorian Gray	0	7.44	14.96	19.98
Savile's Crime	7.44	0	5.25	10.19
Alice	14.96	5.25	0	3.71
Looking Glass	19.98	10.19	3.71	0

Table 3. Chi-squared table for the four gospels and *Acts*. To check the significance of the match we have used our statistical method to compute the chi-squared values of variances of words in common between all the texts. The method gives the lowest score for *Acts* and *Luke* which are known to be by the same author.

	Matthew	**Mark**	**Luke**	**John**	**Acts**
Matthew	0	3.91	2.20	6.05	3.95
Mark	3.91	0	3.21	5.53	4.90
Luke	2.20	3.21	0	2.42	2.02
John	6.05	5.53	2.42	0	3.17
Acts	3.95	4.90	2.02	3.17	0

In an attempt to answer the controversial historical question of who wrote the letter to the *Hebrews* [18], we have used the chi-squared method to analyze the

Table 4. Chi-squared table of selected New Testament books. We have computed chi-squared values for these pairs of books after narrowing down the selection from a larger set of books from the New Testament. Note the very low score between *Luke* and *1 Peter* which are known to be by different authors. Here we are considering the authorship of the letter to the *Hebrews*.

	Luke	**1 Peter**	**2 Peter**	**Jude**	**Hebrews**
Luke	0	1.13	7.86	2.77	7.15
1 Peter	1.13	0	5.15	3.57	4.33
2 Peter	7.86	5.15	0	9.21	3.29
Jude	2.77	3.57	9.21	0	8.40
Hebrews	7.15	4.33	3.29	8.40	0

books of the *Koine* Greek New Testament [16]. We first ran it on a wide selection of books, in order to narrow down the list of books for closer comparison. This was done in order that we get a more representative set of words to use and the result for the smaller list are shown in Table 4.

The graphs of books (see Fig. 3) by various New Testament authors indicates books by Paul as also being close in style to the letter to the *Hebrews*. Note that in the table, the closest match for the letter to the *Hebrews* is the letter *2 Peter*. However the score for the gospel of *Luke* and the letter *1 Peter*, which were written by different authors, is even smaller. Furthermore, notice both *Jude* and *1 Peter* appear close to *Luke*, but *Jude* is not close to *1 Peter*. This type of apparent inconsistency can also be seen on examination of Tables 1 and 2. On closer inspection of the raw data we found that the cause is a Simpson reversal, also known as Simpson's paradox [19], [20]. Simpson reversals tend to occur when data from sub-populations are averaged. This is due to the limited set of words chosen and not features of the texts, such as text in common between gospels. The quite distinct texts in Tables 1 and 2 also show a Simpson reversal because of this problem. In the case of our chi-squared method we conclude that interpretation of the results is problematic and must be exercised with care. Further investigation is required to find a more transparent method of extracting the key features of our graphs. We propose examining standard statistical methods for calculating differences and similarity between graphs [21].

Finally we note that the chi-squared method could possibly be used in tracking changes to a particular text, since one could generate a minimal-spanning tree [22] using the chi-squared metric we have given in (2).

2.3. *Keyword extraction*

Here we detail the method used by Ortuño *et al.* [10] and introduce a new method for extracting keywords.

As per the scaled standard deviation graphs, we determine the set of scaled standard deviations of word spacings for all the words in a text, $\{\hat{\sigma}_1, \ldots, \hat{\sigma}_m\}$. Again, we rank the words from highest scaled standard deviation to lowest, but keeping all the words. We thus obtain a list of words ranked from high relevance to low relevance.

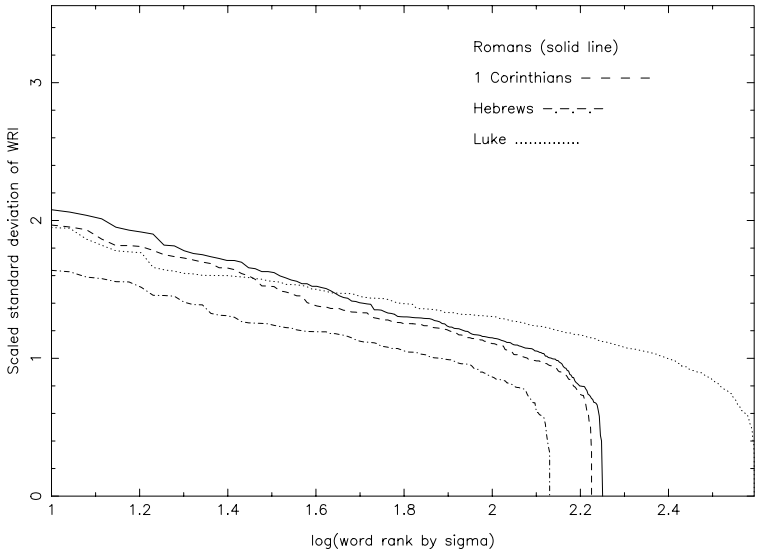

Fig. 3. This is the plot obtained for Paul's letters *1 Corinthians* and *Romans*, the letter to the *Hebrews*, and the gospel of *Luke*. Note the high level of similarity between the two letters of Paul. The similarity of the letter to the *Hebrews* is somewhat obscured by the vertical difference and the end of the plots, both are due to the end effects of the different books' sizes. It is an area of ongoing investigation to minimize these effects.

A new method we have examined for keyword extraction uses the F-statistic on the word spacings, assuming a geometric distribution. The F-statistic detects word-spacing with excess variance (relative to a maximal-entropy or "geometric" distribution). The F-statistic behaves asymptotically like a Gaussian random variable (when the number of WRI samples is large) with mean of 0 and variance of 1 so the statistical tests for relevant keywords are very easy. Given the set of word spacings $\{x_1, \ldots, x_n\}$ we use the F-statistic

$$\frac{1}{2}\ln(n)\left(\frac{s^2}{\bar{x}(1+\bar{x})} - 1\right), \tag{5}$$

where s is the normal sample standard deviation. Note the similarity between the F-statistic and the square of the scaled standard deviation. Assuming the null hypothesis then we get a maximum likelihood estimate [23] of the parameter a in the geometric pdf $p(x) = (1-a)(a^x)$, and can hence estimate the variance of the distribution and we compare this with the standard unbiased estimator for variance. The $\log(n)$ term is for scaling (to deal with the accuracy of the F-statistic for different sample sizes). Other terms are corrections as detailed in Abramowitz and Stegun [24].

Table 5 shows words from Alexander Pope's translation of Homer's *Odyssey*. Table 5 highlights the usefulness of the scaled standard deviation and F-statistic methods over an information measure as detailed in Belew [25] as obvious keywords relating to the *Odyssey* such as the title and names of main characters (Telemachus,

Table 5. Rankings according to information content (Belew [25]), scaled standard deviation and F-statistic for word in the *Odyssey*.

Rank	Info word	Info val.	$\hat{\sigma}$ word	$\hat{\sigma}$ val.	F-stat. word	F-stat. val.
1	requires	9.92	odyssey	3.77	odyssey	19.42
2	manners	9.92	alcinous	3.18	telemachus	17.63
3	scream	9.91	antinous	3.13	antinous	15.86
4	real	9.91	character	3.10	suitors	15.54
5	portico	9.91	telemachus	2.94	alcinous	15.50
6	viewless	9.91	suitors	2.91	been	14.31
7	amorous	9.91	been	2.85	character	11.33
8	tiresias	9.90	melesigenes	2.83	however	9.48
9	accents	9.90	however	2.80	thy	9.45
10	boreas	9.90	scylla	2.73	her	9.30
11	sort	9.90	menelaus	2.62	melesigenes	9.20
12	hideous	9.90	sparta	2.56	son	8.71
13	launch'd	9.90	pisistratus	2.55	bow	8.59
14	imperious	9.89	mere	2.53	suitor	8.51
15	wither'd	9.89	vulcan	2.52	any	8.44

Alcinous and Antinous) are chosen by the scaled standard deviation and F-statistic methods but not the information content method.

As a further test, we found the F-statistic gives the following top ten words as keywords of this paper: texts, the, authorship, statistic, spacing, word, deviation, letter. This can be compared with the list of words obtained by the information method: question, similarity, our, can, extracting, due, match, recurrence, or, data. Note we have run these tests before inserting the previous two sentences to avoid a self-referential loop. Qualitative tests need to be carried out to establish whether the standard deviation or F-statistic method performs best at extracting relevant keywords.

3. Conclusions

The scaled standard deviation and F-statistic methods provide useful tools for keyword extraction. Keyword extraction is important in the area of searching databases for useful information, such as searching the Internet. It may also prove useful in analyzing coding regions of DNA, which is an important open question.

Our results add weight to the generally accepted hypothesis of a common author between the gospel of *Luke* and the book of *Acts*. There is no agreement amongst scholars regarding the authorship of *Hebrews* – our results based on the graphical WRI method add weight to the idea that *Hebrews* shares authorship with the works Paul. Although the chi-squared method shows some promise, the limited sample populations give rise to Simpson reversals, thus making interpretation complex. Thus further work is required to produce a method that automatically extracts

the key features of the graphs. Future developments in this area may shed some light on a number of historical debates surrounding the question of authorship. These methods also have possible applications to both language trees (showing the evolution of languages) and with some modification to study of phylogenetic relationships by comparing DNA sequences.

The graphical and chi-squared methods we have presented for analysis of authorship need further work and testing on larger databases of texts of known authorship. A key question is to determine which methods are best suited to adding weight *for* common authorship and which are best suited for adding weight *against* common authorship between arbitrary texts. Clearly we need to apply standard statistical techniques to the sets of ranked WRI data, to better quantify the differences in style we are observing. There are some challenges in doing this. One of these is dealing with different vocabulary sizes (and hence different data set sizes). Another is in determining which parts of the graph are of interest in distinguishing authorship. Is it just a vertical separation, or other features such as slope, variance in slope, etc.?

The use of a revised chi-squared method or other statistical measures for tracking document changes is proposed for future investigation. Finally, an important open question is to determine a model of the underlying processes that govern word recurrence intervals and to characterize the distributions that underlie these word intervals in natural text.

Acknowledgements

Funding from the University of Adelaide is greatly acknowledged. We thank Pedro Carpena for useful discussions.

References

[1] R. Calderbank and N. J. A. Sloane, *Claude Shannon (1916-2001)*, *Nature* **410** (2001) 786.

[2] C. E. Shannon, *Claude Shannon: Collected Papers*, ed. N. J. A Sloane and A. D. Wyner, IEEE Press, New York (1993).

[3] A. Krogh, *Hidden Markov models for labeled sequences*, *Proceedings of the 12th IAPR International Conference on Pattern Recognition* (1994) 140–144.

[4] N. Chomsky, *Three models for the description of language*, *IRE Transactions on Information Theory* **2** (1956) 113–124.

[5] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, *Proceedings of the IEEE* **77** (1989) 257–286.

[6] S. Dong and D. B. Searls, *Gene structure prediction by linguistic methods*, *Genomics* **23** (1994) 540–551.

[7] F. Mosteller and D. L. Wallace, *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*, second edition, Springer-Verlag, New York (1984).

[8] M. P. Oakes, *Statistics for Corpus Linguistics*, Edinburgh University Press, Edinburgh (1998).

[9] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological Sequence Analysis*, Cambridge University Press, Cambridge (1998).

[10] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz and A. M. Somoza, *Keyword detection in nautral languages and DNA*, *Europhysics Letters* **57** (2002) 759–764.

[11] D. E. Knuth, *The TeXbook*, Addison-Wesley, Mass. (1984).

[12] M. Kac, *On the notion of recurrence in discrete stochastic processes*, *Bulletin of the American Mathematical Society* **53** (1947) 1002–1010.

[13] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov and A. J. Wyner, *Nonparametric entropy estimation for stationary processes and random fields, with applications to English text*, *IEEE Transactions on Information Theory* **44** (1998) 1319–1327.

[14] G. K. Zipf, *The Psycho-biology of Language: An Introduction to Dynamic Philology*, MIT Press, Mass. (1965).

[15] P. Carpena and J. O'Vari, *Private communication* (2001).

[16] E. Nestle, K. Aland, M. Black, C. Martini, B. Metzger and A. Wikgren *et al.*, *Novum Testamentum Graece*, 26th edition, Deutsche Bibelgesellschaft, Stuttgart (1979).

[17] S. Kullback, *Information Theory and Statistics*, Dover Publications, New York (1968).

[18] J. W. McGarvey, *Short Essays in Biblical Criticism*, Standard Publishing Company, Cincinnati (1910).

[19] E. H. Simpson, *The interpretation of interaction in contingency tables*, *Journal of the Royal Statistical Society* **B13** (1951) 238–241.

[20] J. Pearl, *Causality*, Cambridge University Press, Cambridge (2000).

[21] T. Hastie, R. Tibshirani and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York (2001).

[22] D. B. West, *Introduction to Graph Theory*, Prentice Hall, New Jersey (1996).

[23] R. A. Fisher, *Statistical Methods for Research Workers*, 14th edition, Oliver and Boyd, Edinburgh (1970).

[24] M. Abramowitz and A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, U.S. Govt. Printing Office, Washington DC (1964).

[25] R. K. Belew, *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*, Cambridge University Press, Cambridge (2000).