The Influence of Background Data Size on the Performance of a Score-Based Likelihood Ratio System: A Case of Forensic Text Comparison

SHUNICHI ISHIHARA[1,2]
1Forensics Stream of Speech and Language Lab
2Linguistics Program, Australian National University

shunichi.ishihara@anu.edu.au

The likelihood ratio framework

- The only way of assessing the uncertainty inherited in evidential evaluation (Aitken, 2018; Aitken and Taroni, 2004; Good, 1991)
- The logically and legally correct framework for analysing forensic evidence in court (Balding, 2005; Evett et al., 1998; Marquis et al., 2011; Morrison, 2009; Neumann et al., 2007)
- The application of the likelihood ratio framework has been described:
  - DNA (Evett and Weir 1998); voice (Morrison et al. 2018, Rose 2002)
    - fingerprint (Neumann et al. 2007); handwriting (Chen et al. 2018, Hepler et al. 2012)
    - hair (Hoffmann 1991); MDMA tablet (Bolck et al. 2009); evaporated gasoline residue (Vergeer et al. 2014)
    - earmarks (Champod et al. 2001) and more

$$LR = \frac{p(x,y|H_{SA})}{p(x,y|H_{DA})} = \frac{\text{the similarity between the offender and suspect samples}}{\text{the typicality of them in the relevant population}}$$

- x = evidence from the crime scene (source-unknown, offender sample)
- y = evidence from the suspect (source-known, suspect sample)
- H_{SA} = prosecution or same-author hypothesis
- H_{DA} = defence or different-author hypothesis

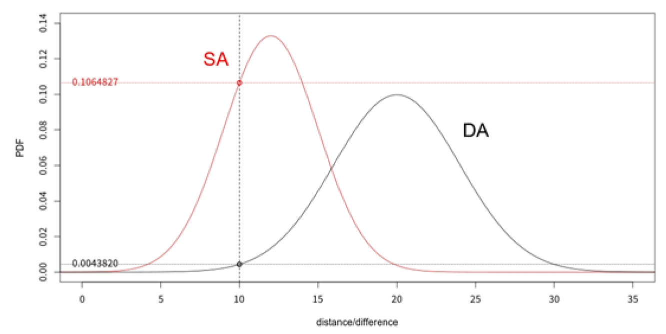
- LR > 1 => same-author hypothesis
- LR < 1 => different-author hypothesis
- A task for the forensic scientist is to estimate the weight of evidence via LR

- Background data is necessary for the relevant population
- Aim: To investigate the robustness and stability of a LR-based forensic text comparison system against the size of the background data

Score-based Likelihood ratios

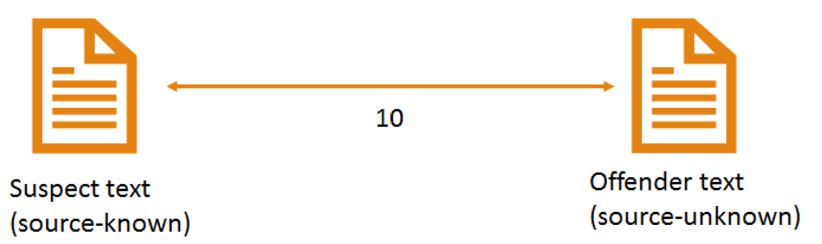
$$LR = \frac{f(\Delta(x,y)|H_{SA})}{f(\Delta(x,y)|H_{DA})} = \frac{f(\Delta(\{w_1^x, w_2^x \cdots w_N^x\}, \{w_1^y, w_2^y \cdots w_N^y\})|H_{SA})}{f(\Delta(\{w_1^x, w_2^x \cdots w_N^x\}, \{w_1^y, w_2^y \cdots w_N^y\})|H_{DA})}$$

- f = probability density function
- x = source–unknown document
- y = source-known document
- $\Delta(x,y)$ = the measured difference between the documents
- x, y = represented as vectors of relative word frequencies ($A = w_i^j, i \in \{1 \cdots N\}, j \in \{x, y\}$)



Score-to-LR conversion model

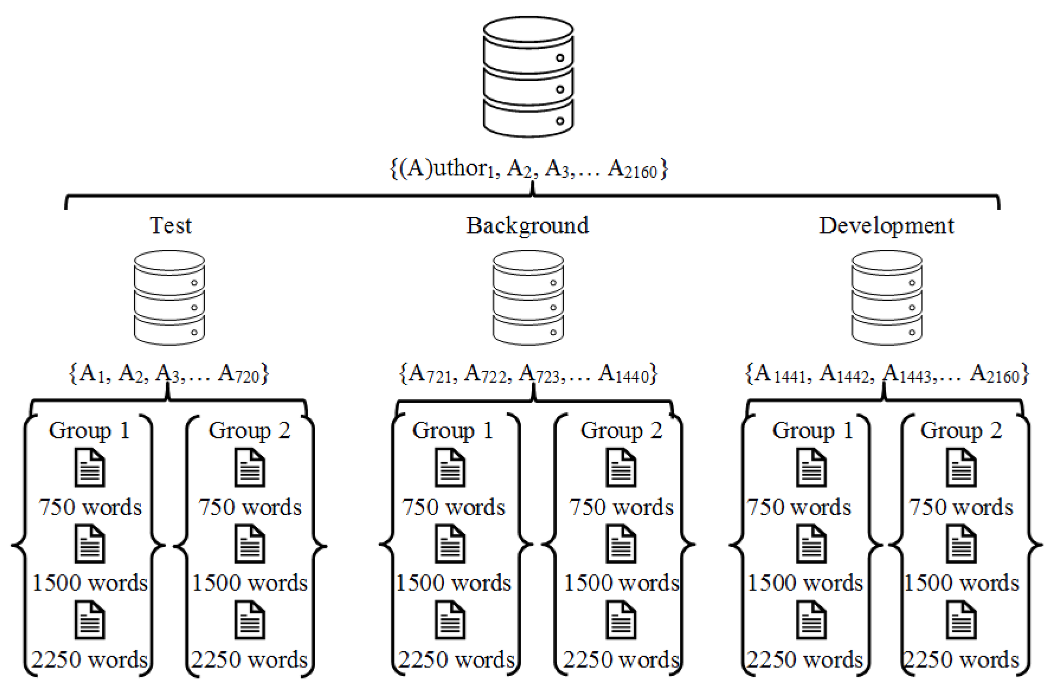
Background, relevant population data



$$LR = \frac{p(E|H_{SA})}{p(E|H_{DA})} = \frac{0.1064827}{0.0043820} = 24.2996$$

Database

- A portion of the Amazon Product Data Authorship Verification Corpus (Halvani et al., 2017)
  - The review texts were equalised to be 4kB in size (approximately 750 words in length)
  - 2,160 reviewers who contributed 6 review texts
  - Each author (reviewer) has 3 pairs of documents which are different in word length (750, 1500, 2250)



Tokenisation and bag-of-words model

- All characters were changed to lower case
- Punctuation marks were not removed; the punctuation marks were thus considered single-word tokens
- No stemming algorithm was applied

- The 420 most frequent words appearing in the entire dataset were selected as components for the bag-of-words model
- The relative frequencies of the words in the model were then calculated for each document
- The word frequencies of the bag-of-words vector were z-score normalised

Gradient assessment metric

- log-likelihood-ratio cost (C_{llr}) (Brümmer & du Preez, 2006)

$$C_{llr} = \frac{1}{2}\left(\left[\frac{1}{N_{SA}}\sum_{i}^{N_{SA}} log_2\left(1+\frac{1}{LR_i}\right)\right] + \left[\frac{1}{N_{DA}}\sum_{j}^{N_{DA}} log_2\left(1+LR_j\right)\right]\right)$$

- N_{SA} and N_{DA} are the number of SA and DA comparisons, and LR_i and LR_j are the linear LRs derived from the SA and DA comparisons, respectively
- The lower, the better
- $C_{llr} > 1$ means the evidence does not provide any useful info
- $C_{llr} = C_{llr}^{min}$ (discrimination loss) + C_{llr}^{cal} (calibration loss)

Experiment 1

- To identify under what conditions the system yields the best outcome
- With different sizes (N) of the bag-of-words vector (N={20,40,60…420})
  - Cosine distance
  - Parametric models (Weibull, Normal, Log Normal, Gamma) for the score-to-LR conversion models
  - Document lengths (750, 1500, 2250 words)

Experiment 1, Result

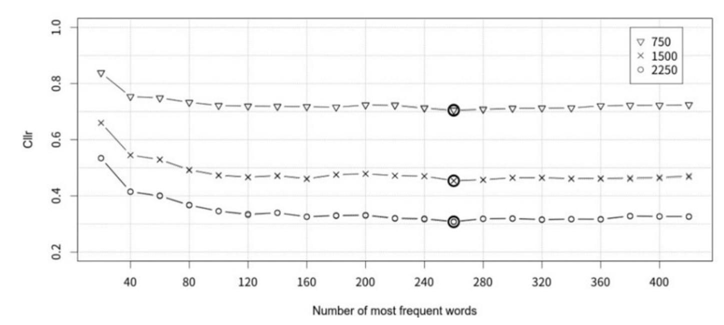


Figure 1: C_{llr} values plotted as a function of the number of features, separately for the word lengths of 750, 1,500 and 2,250. The large circles indicate the best C_{llr}

- Regardless of the word length, the system performed best with N=260
- The overall trend for the C_{llr} trajectory is similar across the word lengths, revealing a relatively large improvement in performance as the N increased from 20 to 120 and the C_{llr} values started converging towards N=260
- After N=260, the performance remained relatively unchanged, indicating that the inclusion of less-frequent words did not contribute to the improvement

Experiment 2

- Probability density models (score-to-LR conversion model) were trained with the background database which consists of texts written by 720 authors
- Using this model as the basis, the scores of X number of authors (X = {5,10,20,30,40,60,80…720}) were randomly generated 20 times to build the score-to-LR conversion model

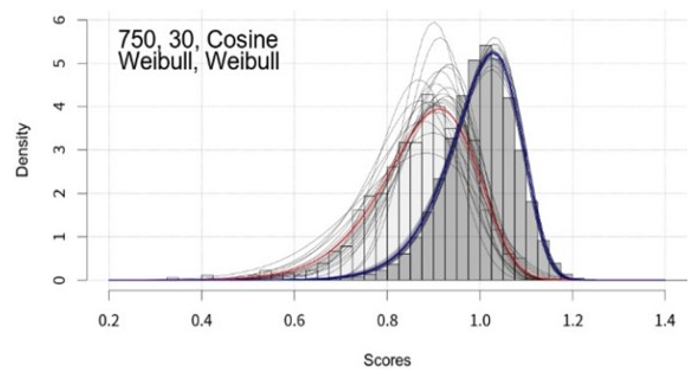


Figure 2: Illustration of a Monte-Carlo simulation with the base SA and DA scores, of which the histograms are white and grey, respectively. The red and blue curves are models of the SA and DA scores, respectively. The thin lines represent the models of the 20 sets of randomly generated scores from 30 authors

Experiment 2, Result

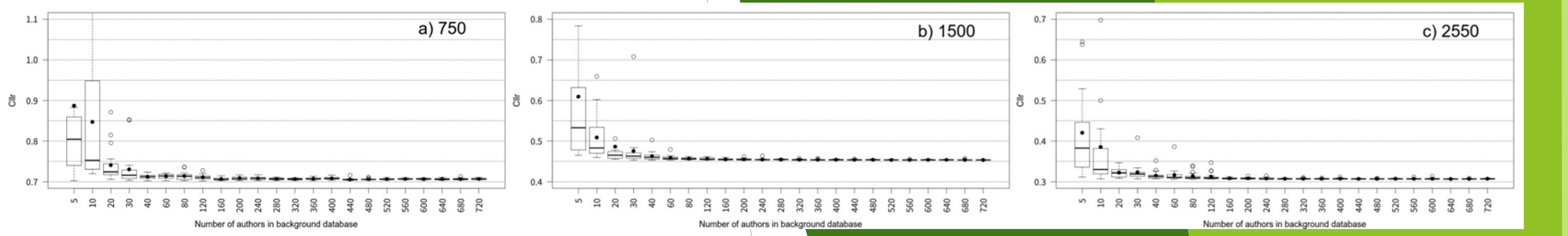


Figure 3: Boxplots displaying the degree of fluctuation in C_{llr} values as a function of the size of the background database. Black circles indicate the mean C_{llr} values for each size of the background database

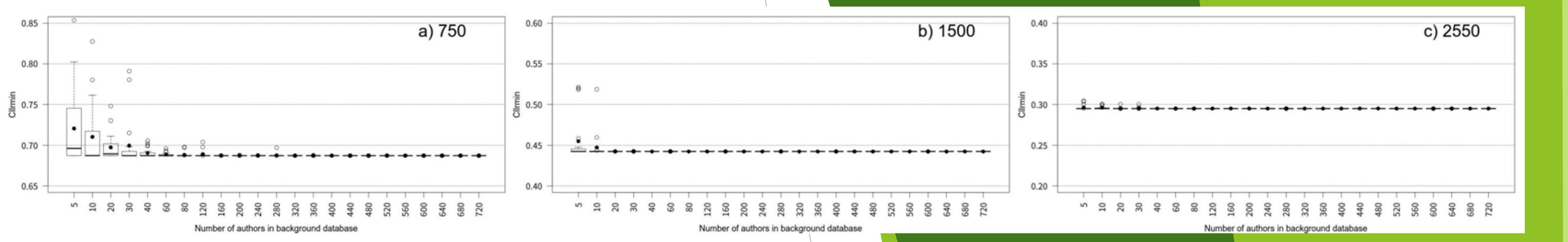


Figure 4: Boxplots showing the degree of fluctuation in C_{llr}^{min} as a function of the size of the background database. Black circles indicate the mean C_{llr}^{min} values for each size of the background database

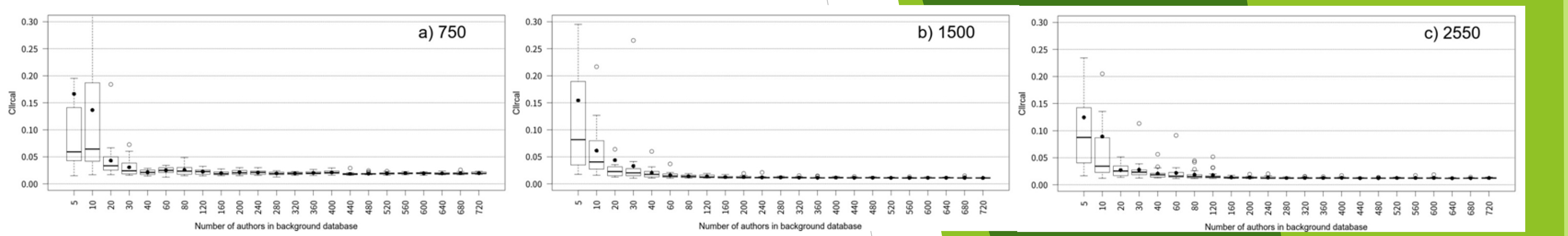


Figure 5: Boxplots displaying the degree of fluctuation in C_{llr}^{cal} values as a function of the size of the background database. Black circles indicate the mean C_{llr}^{cal} values for each size of the background database

- It is evident from Figure 3 (black circles) that the system's overall performance improves exponentially from N=5 to N=40, resulting in the outcome in which the performance with N=40 is nearly compatible with its performance with N=720
- As can be observed in Figure 4, being apart from the word length of 750, the system's discriminability is highly stable, even with small Ns. Specifically, regarding the word length of 2,250, Figure 4c reveals that the C_{llr}^{min} values are constant and fair stable, as they are not affected by the number of authors in the background database. That is, in terms of discrimination performance, when many words (e.g., 1,500 and 2,250 words) are available, the system is robust and stable against a small background population size
- In contrast, Figure 5 indicates that the C_{llr}^{cal} values exhibit a highly similar trend to that of the C_{llr} values that are plotted in Figure 3—in that, a great variability in the C_{llr}^{cal} values is observed when the number of authors is small (e.g., N=5–10); however, this variability be-gins converging rapidly with more authors. This signifies that the C_{llr}^{cal} values also demonstrate a quick recovery with more authors
- The observations drawn from Figures 4 and 5 reveal that the poor performance associated with a small number of authors (N=5–10), as indicated by the C_{llr} values from Figure 3, is not due to the system's poor discriminability, but due to poor calibration.

Conclusions

- The experiments' results revealed that
- The score-based forensic text comparison system is fairly robust and stable in performance against the limited number of background population data
  - For example, with 40–60 authors, the performance is both nearly compatible and as stable as with 720 authors
  - This is a beneficial finding for forensic text comparison practitioners
- The instability and suboptimal performance observed in terms of C_{llr} with a small number of data (e.g., 5–20 authors) were mainly attributed to poor calibration (i.e., the derived LRs were not calibrated) rather than to the poor discriminability potential

References

Aitken, C. G. G. (2018) Bayesian hierarchical random effects models in forensic science. *Frontier in Genetics 9*(Article 126): 1-14. https://doi.org/10.3389/fgene.2018.00126
Aitken, C. G. G. and Lucy, D. (2004) Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society, Series C (Applied Statistics) 53*(1): 109-122. https://dx.doi.org/10.1046/j.0035-9254.2003.05271.x
Aitken, C. G. G. and Stoney, D. A. (1991) *The Use of Statistics in Forensic Science*. New York: Ellis Horwood.
Aitken, C. G. G. and Taroni, F. (2004) *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester: John Wiley & Sons.
Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6): 716-723. https://dx.doi.org/10.1109/TAC.1974.1100705
Balding, D. J. (2005) *Weight-of-Evidence for Forensic DNA Profiles*. Hoboken: John Wiley & Sons.
Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S. and Matsuo, A. (2018) quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software 3*(30): 774-776. https://doi.org/10.21105/joss.00774
Bolck, A., Ni, H. F. and Lopatka, M. (2015) Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: Applied to forensic MDMA comparison. *Law, Probability and Risk 14*(3): 243-266. https://dx.doi.org/10.1093/lpr/mgv009
Brümmer, N. and du Preez, J. (2006) Application-independent evaluation of speaker detection. *Computer Speech and Language 20*(2-3): 230-275. https://dx.doi.org/10.1016/j.csl.2005.08.001
Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C. and Vitt, T. (2017) Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities 32*(suppl_2): ii4-ii16. https://doi.org/10.1093/llc/fqx023
Evett, I. W., Lambert, J. A. and Buckleton, J. S. (1998) A Bayesian approach to interpreting footwear marks in forensic casework. *Science & Justice 38*(4): 241-247. https://dx.doi.org/10.1016/S1355-0306(98)72118-5
Good, I. J. (1991) Weight of evidence and the Bayesian likelihood ratio. In C. G. G. Aitken and D. A. Stoney (eds.), *The Use of Statistics in Forensic Science* 85-106. Chichester: Ellis Horwood.
Halvani, O., Winter, C. and Graner, L. (2017). Authorship verification based on compression-models. *arXiv preprint arXiv:1706.00516*. Retrieved on 25 June 2020 from http://arxiv.org/abs/1706.00516
Hepler, A. B., Saunders, C. P., Davis, L. J. and Buscaglia, J. (2012) Score-based likelihood ratios for handwriting evidence. *Forensic Science International 219*(1-3): 129-140. http://dx.doi.org/10.1016/j.forsciint.2011.12.009
Ishihara, S. (2014) A likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams. *International Journal of Speech Language and the Law 21*(1): 23-50. http://dx.doi.org/10.1558/ijsll.v21i1.23
Ishihara, S. (2016) An effect of background population sample size on the performance of a likelihood ratio-based forensic text comparison system: A Monte Carlo simulation with Gaussian mixture model. In T. Cohn (ed.), *Proceedings of Proceedings of the Australasian Language Technology Association Workshop 2016*: 113-121.
Ishihara, S. (2017a) Strength of forensic text comparison evidence from stylometric features: A multivariate likelihood ratio-based analysis. *The International Journal of Speech, Language and the Law 24*(1): 67-98. https://doi.org/10.1558/ijsll.30305
Ishihara, S. (2017b) Strength of linguistic text evidence: A fused forensic text comparison system. *Forensic Science International 278*: 184-197. https://doi.org/10.1016/j.forsciint.2017.06.040
Lund, S. P. and Iyer, H. (2017) Likelihood ratio as weight of forensic evidence: A closer look. *Journal of Research of the National Institute of Standards and Technology 122*(Article 27): 1-32. https://doi.org/10.6028/jres.122.027
Marquis, R., Bozza, S., Schmittbuhl, M. and Taroni, F. (2011) Handwriting evidence evaluation based on the shape of characters: Application of multivariate likelihood ratios. *Journal of Forensic Sciences 56*(Suppl_1): S238-242. https://dx.doi.org/10.1111/j.1556-4029.2010.01602.x
Morrison, G. S. (2009) Forensic voice comparison and the paradigm shift. *Science & Justice 49*(4): 298-308. https://dx.doi.org/10.1016/j.scijus.2009.09.002
Morrison, G. S. (2013) Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences 45*(2): 173-197. https://dx.doi.org/10.1080/00450618.2012.733025
Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A. and Bromage-Griffiths, A. (2007) Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Science 52*(1): 54-64. https://dx.doi.org/10.1111/j.1556-4029.2006.00327.x
Ramos, D., Krish, R. P., Fierrez, J. and Meuwly, D. (2017) From biometric scores to forensic likelihood ratios. In M. Tistarelli and C. Champod (eds.), *Handbook of Biometrics for Forensic Science* 305-327. Cham: Springer.
Robertson, B., Vignaux, G. A. and Berger, C. E. H. (2016) *Interpreting Evidence: Evaluating Forensic Science in the Courtroom* (2nd ed.). Chichester: John Wiley and Sons, Inc.
Smith, P. W. H. and Aldridge, W. (2011) Improving authorship attribution: Optimizing Burrows' Delta method. *Journal of Quantitative Linguistics 18*(1): 63-88. https://dx.doi.org/10.1080/09296174.2011.533591
Zipf, G. K. (1932) *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge: Harvard University Press.