Title: Strength of linguistic text evidence: A fused forensic text comparison system

Author: Shunichi Ishihara

PII: S0379-0738(17)30247-5
DOI: http://dx.doi.org/doi:10.1016/j.forsciint.2017.06.040
Reference: FSI 8903

To appear in: *FSI*

Received date: 27-3-2017
Revised date: 5-6-2017
Accepted date: 30-6-2017

Please cite this article as: Shunichi Ishihara, Strength of linguistic text evidence: A fused forensic text comparison system, Forensic Science Internationalhttp://dx.doi.org/10.1016/j.forsciint.2017.06.040

Strength of Linguistic Text Evidence: A Fused Forensic Text Comparison System

xxx
yyy

Shunichi Ishihara[1]
Department of Linguistics, the Australian National University, Canberra, Australia

Department of Linguistics, School of Culture, History and Language
College of Asia and the Pacific, Building #110
ACTON, Canberra ACT 2601, Australia

Email: shunichi.ishihara@anu.edu.au
Phone: +61 2 6125 4656
[1] Present address: Chuo University, Faculty of Letters, 742-1 Higashi Nakano, Hachioji, Tokyo 192-0393, Japan

Highlights:

- *A description of estimating strength of authorship attribution evidence within the likelihood ratio framework.*
- *Efficacy of the likelihood ratio framework for authorship attribution text evidence.*
- *Efficacy of logistic-regression-fusion for authorship attribution text evidence.*
- *Effect of data sample size on the performance of the likelihood ratio-based forensic text comparison system.*

Abstract: Compared to other forensic comparative sciences, studies of the efficacy of the likelihood ratio (LR) framework in forensic authorship analysis are lagging. An experiment is described concerning the estimation of strength of linguistic text evidence within that framework. The LRs were estimated by trialling three different procedures: one is based on the multivariate kernel density (MVKD) formula, with each group of messages being modelled as a vector of authorship attribution features; the other two involve *N*-grams based on word tokens and characters, respectively. The LRs that were separately estimated from the three different procedures are logistic-regression-fused to obtain a single LR for each author comparison. This study used predatory chatlog messages sampled from 115 authors. To see how the number of word tokens affects the performance of a forensic text comparison (FTC) system, token numbers used for modelling each group of messages were progressively increased: 500, 1000, 1500 and 2500 tokens. The performance of the FTC system is assessed using the log-likelihood-ratio cost (C_{llr}), which is a gradient metric for the quality of LRs, and the strength of the derived LRs is charted as Tippett plots. It is demonstrated in this study that i) out of the three procedures, the MVKD procedure with authorship attribution features performed best in terms of C_{llr}, and that ii) the fused system outperformed all three of the single procedures. When the token length is 1500, for example, the fused system achieved a C_{llr} value of 0.15. Some unrealistically strong LRs were observed in the results. Reasons for these are discussed, and a possible solution to the problem, namely the empirical lower and upper bound LR (ELUB) method is trialled and applied to the LRs of the best-achieving fusion system.

Keywords: forensic text comparison; likelihood ratio; logistic-regression fusion; multivariate kernel density; *N*-grams; authorship attribution features

1 Introduction

[1] Present address: Chuo University, Faculty of Letters, 742-1 Higashi Nakano, Hachioji, Tokyo 192-0393, Japan

The history of authorship attribution study is long. Mendenhall's (1887) study on Shakespeare's plays is often quoted as the first authorship attribution study based on a statistical/computational method. It was followed by many influential studies in the first half of the 20th century (Mosteller and Wallace 1964, Yule 1939, 1944, Zipf 1932). Since the end of the 20th century, due to the change in communication medium, the focus of authorship attribution has started shifting from literary texts to electronically-generated texts (e.g. emails, chatlogs, SMS), with some studies focusing on the domain of forensics (Abbasi and Chen 2005, Cohen 2009, Corney et al. 2001, Fuhrman 2008, Gao and Zhao 2005, Khan et al. 2012, Kucukyilmaz et al. 2006, 2008, Pillay and Solorio 2011, Son et al. 2008, Stolfo et al. 2003, Wei et al. 2008).

However, forensic authorship attribution has considerably fallen behind in comparison to other forensic comparative sciences in that the above forensic authorship attribution studies were not conducted in the likelihood ratio (LR) framework, which is increasingly held to be the logically and legally correct framework of evaluating forensic evidence (cf. Ishihara 2011, 2012a). In many branches of forensic sciences, including fingerprint (Neumann et al. 2007), handwriting (Bozza et al. 2008), voice (Morrison 2009), DNA (Evett et al. 1993), glass fragments (Curran 2003), earmarks (Kuchler et al. 2001) and footwear marks (Evett et al. 1998) analysis, the LR framework has been or has started being accepted as the standard framework for the evaluation of forensic evidence. The spotlight on the LR framework is, needless to say, largely attributable to the success of DNA profiling (Balding and Steele 2015, Foreman et al. 2003), as well as to some rulings (*Daubert v. Merrell Dow Pharmaceuticals Inc, 1993*; *Kuhmo Tire Co. v. Carmichael, 1999*) and reports (*Strengthening Forensic Science in the United States: A Path Forward* (2009)) regarding the rules of evidence in the United States. As a matter of fact, the use of the LR framework has been advocated for quite some time in the main textbooks on the evaluation of forensic evidence (Robertson and Vignaux 1995) and by forensic statisticians (Aitken and Stoney 1991, Aitken and Taroni 2004).

In this study, therefore, the LR framework is implemented for authorship attribution. First of all, three different procedures are trialled to estimate LRs for predatory chatlog messages – one based on authorship attribution features with the multivariate kernel density (MVKD) LR formula (the MVKD procedure); one with word token-based *N*-grams (the token *N*-grams procedure) and one with character-based *N*-grams (the character *N*-grams procedure). In the MVKD procedure, each message group (e.g. a set of messages written by a suspect or an offender) is modelled as a vector of authorship attribution features, such as the vocabulary richness feature, the average token number per message line, upper case character ratio, etc. (refer to §3.3.1 for further details on authorship attribution features). In the token and character *N*-grams procedures, each message group is modelled by token- and character-based *N*-grams, respectively. The performances of the three different procedures are compared.

In a second step, the LRs that were separately derived by the three different procedures are fused into a single LR for each comparison, representing the combined evidence. This allows us to investigate the extent to which fusion improves (or deteriorates) the performance of the forensic text comparison (FTC) system. The current study employs logistic-regression fusion (Brümmer and du Preez 2006) as it is a robust technique and has been applied to some LR-based forensic comparison systems (some examples are given in Morrison (2013)). The performance of each FTC system is assessed by means of the log likelihood ratio cost (C_{llr}) (Brümmer and du Preez 2006, van Leeuwen and Brümmer 2007), which is a gradient metric for assessing the quality of LRs. The strength of the derived LRs is visually displayed as Tippett plots. Detailed explanations of logistic-regression fusion, C_{llr} and Tippett plots are given in §3.4, §3.5 and §4, respectively.

2 Likelihood Ratio and Bayesian Theorem

Many forensic scientists and statisticians (Aitken and Stoney 1991, Aitken and Taroni 2004, Robertson and Vignaux 1995) explicitly state that the role of the forensic scientist is to estimate the strength of evidence, which can be quantified by the LR. The LR is the ratio of the probability that

the evidence (E) would occur if one hypothesis (e.g. the prosecution hypothesis – H_p) is true and the probability that the same evidence would occur if the alternative hypothesis (e.g. the defence hypothesis – H_d) is true (Robertson and Vignaux 1995). Thus, the LR can be expressed as in (1).

$$LR = \frac{p(E|H_p)}{p(E|H_d)} \quad (1)$$

Consider a typical FTC scenario where the forensic scientist is required to compare a set of messages written by the offender and another set of messages written by the suspect, and the following hypotheses are of interest:

H_p: the two sets of messages were written by the same author.
H_d: the two sets of messages were written by different authors.

In FTC, the evidence (E) consists of the measured properties of the messages written by the suspect and the messages written by the offender. The numerator of (1) determines the probability of the evidence assuming the same author hypothesis (H_p). Likewise, the denominator of (1) determines the probability of the same evidence assuming the different author hypothesis (H_d). The LR is the ratio of those two probabilities under the competing hypotheses. If the evidence is more likely to be observed under the same author hypothesis than under the different author hypothesis, then the LR will be greater than one. On the contrary, if the evidence is more likely to be observed under the different author hypothesis than under the same author hypothesis, then the LR will be smaller than one. That is to say, the relative strength of the given evidence with respect to the competing hypotheses (H_p vs. H_d) is reflected in the magnitude of the LR. The more the LR deviates from one, the greater support there is considered to be for either the prosecution hypothesis or the defence hypothesis.

In an LR estimation, the equation given in (1) is interpreted in terms of similarity (numerator) and typicality (denominator). The numerator quantifies the degree of similarity between the two samples (e.g. messages) in the comparison, and the denominator quantifies the significance of such similarity. Even if the samples of the offender and the suspect are found to be very similar, the similarity is less significant if the measured properties of the samples are very typical against the relevant population, as there would be many other individuals in the population who could present the same measured properties. Therefore, at least three different sets of data are required for estimating LRs: offender samples, suspect samples and samples from the population relevant to the case (background reference data).

An LR is a relative strength of evidence. It *indicates* whether the evidence supports the prosecution or defence hypothesis. To quantify the amount of support or obtain a probability score for the offender and suspect being the same person or otherwise, given the evidence (i.e. the probability of the hypotheses in light of the evidence; namely posterior odds or strength of hypothesis), the LR needs to be combined with the prior odds of the hypotheses via Bayes' theorem. The prior odds is the trier-of-fact's belief in relative favour of the two competing hypotheses, which is a result of initial assumptions and changes in belief after the presentation of all the relevant evidence. Such trier-of-fact's belief is not knowledgeable to the forensic scientist; thus the latter cannot logically calculate the posterior odds (Champod and Meuwly 2000). In addition, they must not calculate the posterior odds for legal reasons: referring to the posterior odds is equivalent to referring to the suspect as being guilty or not guilty, which is not the role of the forensic expert but of the fact finder; the forensic expert should not be usurping the role of the trier-of-fact (Aitken and Taroni 2004:4, Evett 1998).

3 Experiments

3.1 Database

Real pieces of chatlog communication between later-sentenced paedophiles and undercover police officers in the US, drawn from an archive of chatlog messages (http://pjfi.org/) were used for the research reported on in this paper. However, as the archive had not been designed as a database for authorship analysis studies, the messages written by each author had to be manually checked and transformed to a computer-readable format prior to the commencement of the current study. In total, the messages written by 383 authors between 2007 and 2011 were processed as described. Out of the 383 authors, only those who enabled us to create two groups of messages that do not chronologically overlap and that each consist of 2500 tokens were further selected to meet the experimental specifications detailed later in this subsection. This resulted in 115 authors and their messages being selected for the FTC experiments that were carried out.

The 115 authors were separated into three mutually exclusive sub-databases: the test database (39 authors); the background database (38 authors); and the development database (38 authors). The test database was used to simulate the various offender-suspect comparisons by means of which the performance of the FTC system was assessed. The background database was used as a reference to determine typicality for calculating LRs. The development database was used to calculate weights for calibrating the derived LRs of the SA and DA comparisons generated from the test database. As the test database contained material by 39 authors, 39 same author (SA) and 1482 independent different author (DA) comparisons (= 741 author pairs (= ${}_{39}C_2$) × 2 comparisons for each author pair) were possible. Given their identical origins, the LRs estimated for the 39 SA comparisons were anticipated to be greater than LR = 1, to the extent that the system works. Likewise, given their different origins, the LRs estimated for the 1482 DA comparisons were expected to be smaller than LR = 1. Four different sample sizes or, in other words, token lengths (500, 1000, 1500 and 2500 tokens), were used to model the attributes of each message group. This allowed us to investigate how different sample sizes (e.g. "sample size 500", for a sample size of 500 tokens, etc.) influence the performance of a system.

The small size of the selected database (115 authors) could be seen as a weakness. However, practically speaking, to the best of the author's knowledge, there are no publicly available databases that are appropriate for the FTC experiments designed for the current study. There are some databases of electronically-generated texts (e.g. SMS, chatlog) in the public domain, such as the "Ubuntu chat corpus" (http://daviduthus.org/UCC/), the "NPS chat corpus" (http://faculty.nps.edu/cmartell/NPSChat.htm) and the "NUS SMS Corpus" (http://www.comp.nus.edu.sg/entrepreneurship/innovation/osr/corpus/). The "Ubuntu chat corpus" is a large archive that consists of chatlogs from Ubuntu's Internet Relay Chat technical support channels. However, the messages of the archive are not verified for authorship, and it is well known that the users of public channels often change their usernames. It goes without saying that this is a critical issue for authorship attribution studies (the existence of multiple usernames for a single author is reported for the "Ubuntu chat corpus" by Uthus and Aha (2013:99)). The other two corpora are considerably smaller than the selected database in terms of the number of appropriate authors for the FTC experiments designed for the current study. This left us with no viable alternative, apart from the chatlog database selected for this study. The fact that the messages stored in the database were actually used as text evidence in criminal cases is a bonus; however, the selection of a small database has a downside relating to the instability of the resultant model (Vergeer et al. 2016). This point will be addressed by referring to the outcome of the FTC experiments, and a possible solution will be implemented in §4.2.

3.2 Tokenisation

For the MVKD and the token *N*-grams procedures, the different ways of breaking up message lines into word tokens impinge on the results of the experiments. In this study, the chatlog messages were tokenised using the tokenisers stored in the *Natural Language Tool Kit* (*NLTK*) (version 2.0) (http://www.nltk.org/). Three *NLTK* tokenisers were tested to see which tokenisation function works

best for authorship attribution. The functions that were tested are: word_tokenize() (hereafter "word tokeniser"), wordpunct_tokenize() (hereafter "punctuation tokeniser") and WhitespaceTokenizer() (hereafter "whitespace tokeniser"). The word tokeniser performs tokenisation by finding the word tokens and punctuation in the input string. The punctuation tokeniser splits the input string on the basis of whitespace and punctuation. The whitespace tokeniser carries out its task based on whitespace. As can be seen in the examples given in

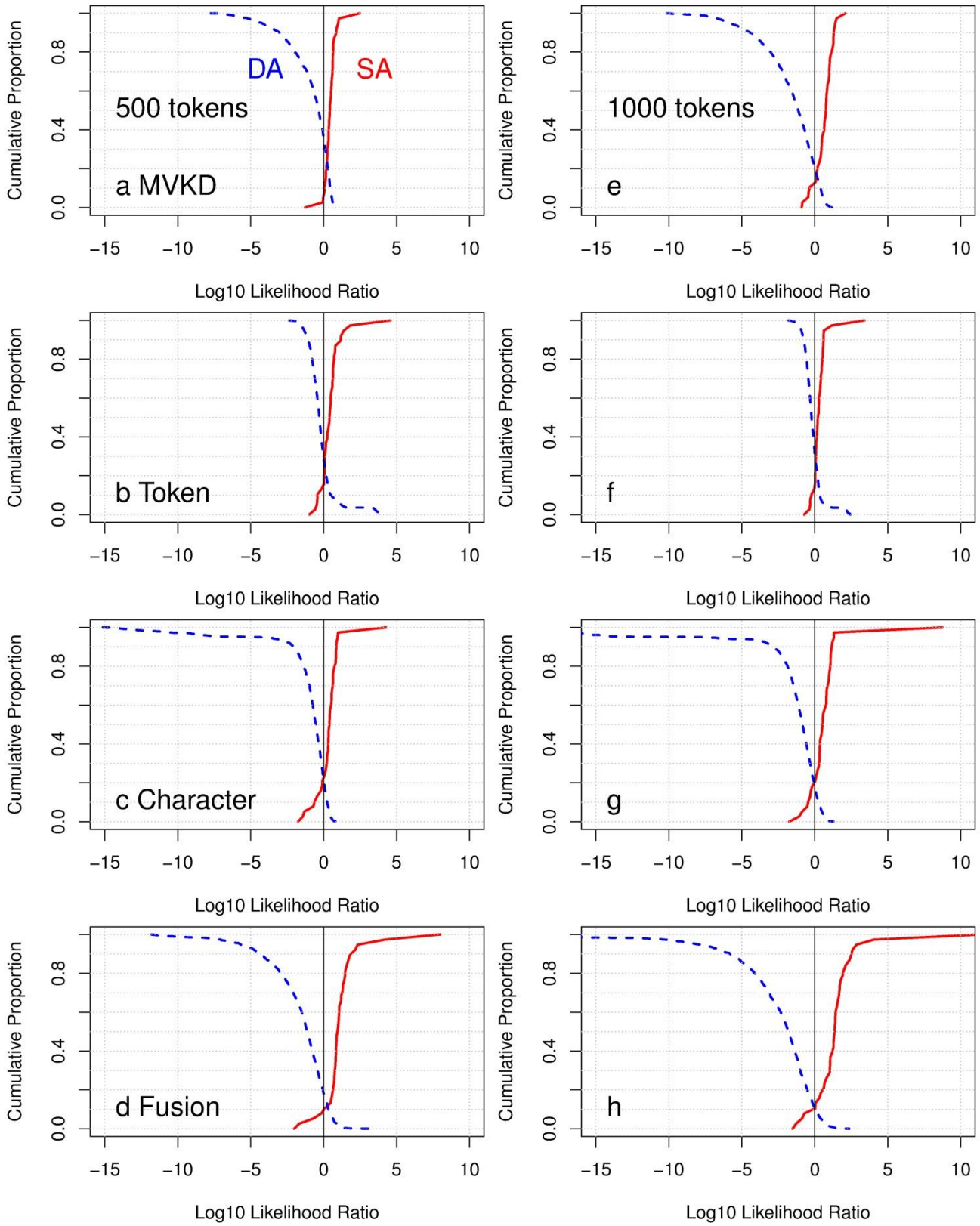

Figure 1: Tippett plots showing LRs for the best-performing configurations of the MVKD procedure (panels a and e), the token *N*-grams procedure (panels b and f), the character *N*-grams procedure (panels c and g) and the fused system (panels d and h). Red (solid) = SA comparisons; blue (dashed) = DA comparisons. Left panels = sample size 500; right panels = sample size 1000. Note that some curves extend beyond the range between $\log_{10}$LRs = -15 and $\log_{10}$LRs = 10.

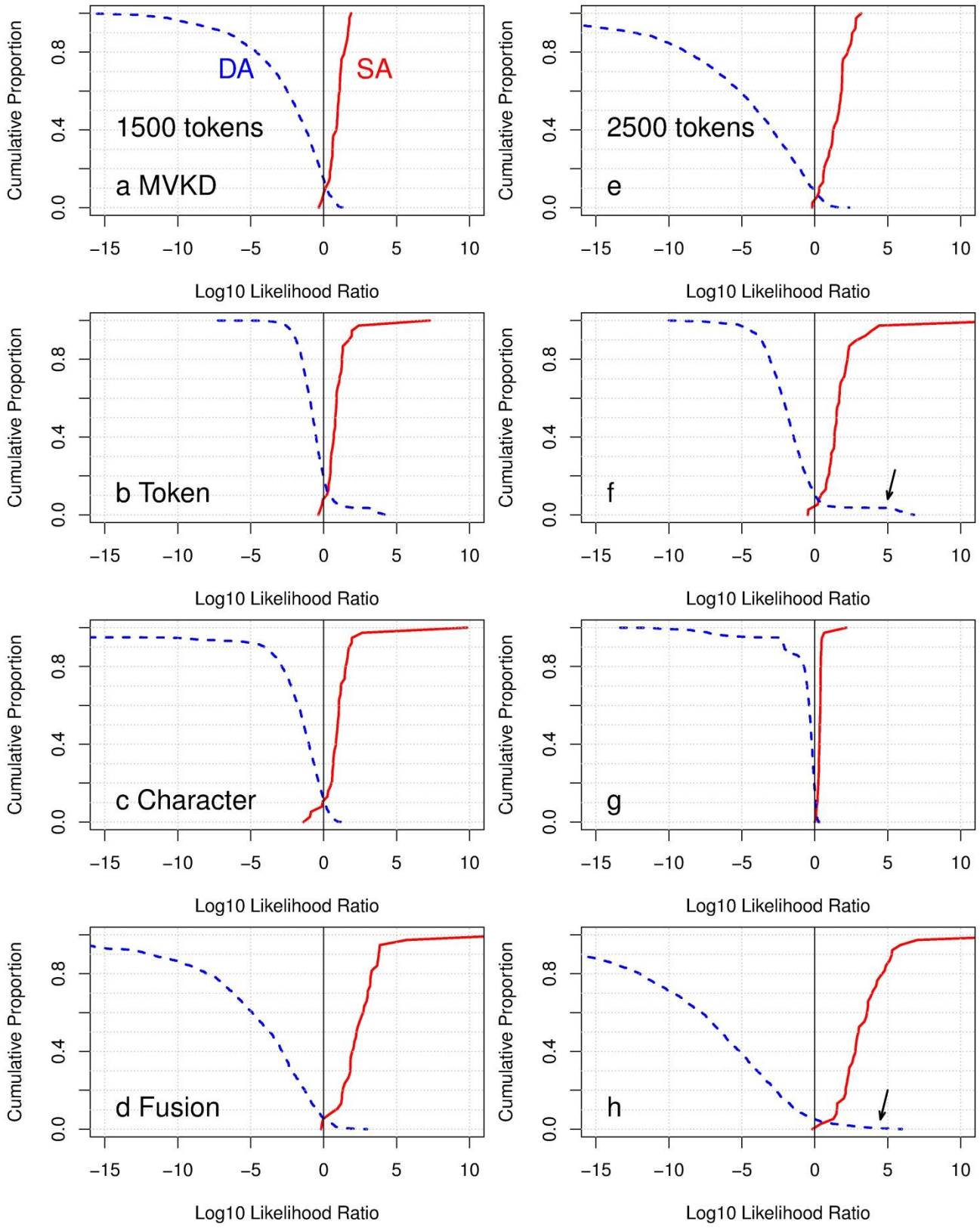

Figure 2: Tippett plots showing LRs for the best-performing configurations of the MVKD procedure (panels a and e), the token *N*-grams procedure (panels b and f), the character *N*-grams procedure (panels c and g) and the fused system (panels d and h). Red (solid) = SA comparisons; blue (dashed) = DA comparisons. Left panels = sample size 1500; right panels = sample size 2500. The arrows in panels f and h indicate large contrary-to-fact LRs of some DA comparisons.

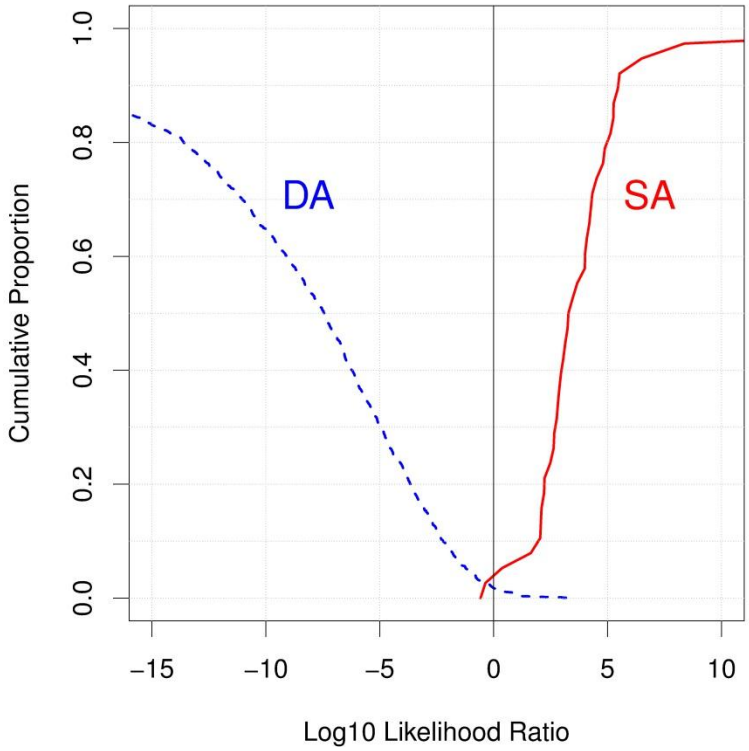

Figure 3: Tippett plot of the system that fused the best results of the three different procedures: the MVKD procedure on 2500 tokens, the token *N*-grams procedure on 2500 tokens and the character *N*-grams procedure on 1500 tokens.

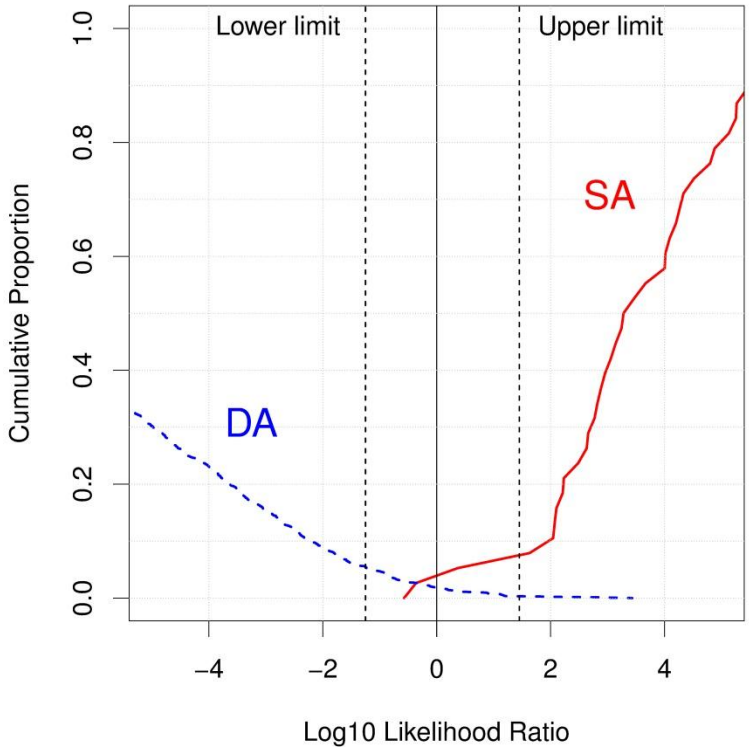

Figure 4: Replotted Figure 3 with the $\log_{10}LR_{min}$ and $\log_{10}LR_{max}$. The vertical dashed lines indicate the $\log_{10}LR_{min}$ and $\log_{10}LR_{max}$ values for the LRs. Note that a narrower range is used for the x-axis in Figure 4 compared to Figure 3.

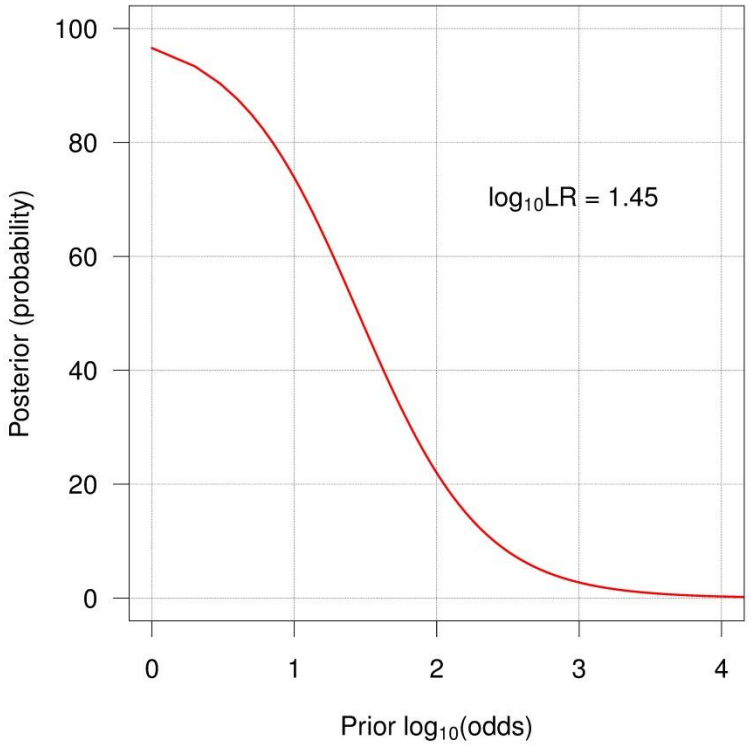

Figure 5: A graph showing how prior odds plays an important role when deriving the posterior probability from an LR. A $\log_{10}$LR of 1.45 is used as an example.

Table 1, the three tokenisers in question produce fairly different outputs, for example, for the strings "couldn't" and "$3.50" in the sentence "I couldn't even buy a muffin of $3.50 when I was a student".

3.3 Features, Modelling Techniques and Likelihood Ratio Calculations

For the purpose of LR estimations, the three different procedures (MVKD, token *N*-grams and character *N*-grams) were employed independently to enable comparison of their performance. Subsequently, the outcomes of the three procedures were fused to see how the fused system performs in comparison to each of the three procedures viewed individually. In this section, the three procedures are explained in detail.

3.3.1 Authorship Attribution Features and Multivariate Kernel Density Likelihood Ratio Formula

In the MVKD procedure, each group of messages needs to be modelled as a vector based on a certain set of authorship attribution features. Following the results of previous authorship analysis studies (De Vel et al. 2001, Iqbal et al. 2010, Zheng et al. 2006), the authorship attribution features used in the MVKD procedure were those listed in Table 2. Although there are some different ways of classifying authorship attribution features, the features listed in Table 2 can be largely categorised as token-based (F1~F5) and character-based (F6~F12). The first three token-based features (F1~F3) are for vocabulary richness, which attempts to quantify the diversity of vocabulary of a text. Various measures for vocabulary richness have been proposed (e.g. Honoré 1979, Yule 1944), and have virtually been used as one of the standard features in authorship analysis (Zheng et al. 2006:380). The other token-based features (F4 and F5) and the character-based features (F6~F12) are straightforward and self-explanatory.

The formulae of the vocabulary richness features (VRFs), i.e. *Yule's I*, Type-token ratio (*TTR*) and *Honoré's R* (F1~F3), are reproduced below, in (2), (3) and (4), respectively (Baayen 2001, Oakes 1998).

$$\textit{Yule's } I = (\mathrm{M1} \times \mathrm{M1})/(\mathrm{M2} - \mathrm{M1}) \tag{2}$$

In (2), M1 is the total number of word tokens observed in a text, and M2 is the sum of the products of each observed frequency to the power of two and the number of words observed with that frequency (Oakes 1998:204). To illustrate this by means of a short text, let us consider the sentence "The cat in the hat is not my cat", which consists of nine word tokens. Five words ("in", "hat", "is", "not" and "my") appear once and two words ("the" and "cat") appear twice. Therefore, M1 is 9 and M2 is 13 (= $(5\times1^2)+(2\times2^2)$). Thus, *Yule's I* for this example text is 20.25 (= 9×9/(13-9)).

$$\textit{TTR} = \mathrm{V}/\mathrm{N} \tag{3}$$

In (3), V is the number of word types, and N is the total number of word tokens appearing in a text. For the same example text given above, V is 7 and N is 9, resulting in the *TTR* being ca. 0.78 (= 7/9).

$$\textit{Honoré's } R = 100 \times \log \mathrm{N}/(1 - (\mathrm{V}_1/\mathrm{V})) \tag{4}$$

N is the number of word tokens in a target text, V_1 is the number of hapaxes (words occurring only once) and V is the number of word types. For the same example as above, N is 9, V_1 is 5 and V is 7. Thus, *Honoré's R* is ca. 769.03 (= 100×log(9)/(1-(5/7))).

Three different VRFs (F1~F3) were used in the experiments because the use of multiple VRFs reportedly contributes to a better performance (Holmes 1992).

The ratio of punctuation characters (F11) can be classified as a syntactic feature, as such characters provide syntactic information of a particular kind on sentences or phrases (cf. Zheng et al. 2006). However, in this study, it was classified as a character-based feature, since punctuation marks are character-based (Stamatatos 2009).

Different combinations of the features listed in Table 2 were tested. This is because i) it is generally understood that good feature combinations play an important role in the performance of an authorship analysis system (Zheng et al. 2006:380), and ii) while the favourable cumulative effect of less informative features was also pointed out (Aizawa 2001), the inclusion (as noise) of less informative features may deteriorate system performance (De Vel et al. 2001). Since testing all possible feature combinations with various dimensions of a feature vector is time-consuming, only some of the possible combinations were selected and implemented for testing. First of all, all 66 possible combinations of two features [f_1,f_2] were trialled; then, on the basis of the outcomes of the 66 experiments, the five best performing bi-features were selected for the next process. Using these five best performing bi-features as bases, the performance of the tri-features [f_1,f_2,f_3] was tested by adding one of the remaining features to them, one by one. Again, the five best performing tri-features were selected for the next set of experiments. This process was repeated for feature vectors of a higher dimension.

If none of these features were correlated with each other, the estimated LR from each of the features given in Table 2 could be combined via simple multiplication, viz. naïve Bayes. However, the assumption that there is no correlation is obviously untenable (Ishihara 2014d); the three VRFs, for example, are inherently correlated, and they would increase as the token and character numbers increase. The issue of estimating LRs from correlated variables was addressed by Aitken and Lucy (2004), resulting in a newly defined multivariate kernel density (MVKD) formula for the calculation of LRs. Following the initial application of the formula to data from glass fragments, the validity of the procedure was tested on voice (Rose et al. 2004), handwriting (Marquis et al. 2011) and text (Ishihara 2012b). In a nutshell, the MVKD formula accepts multiple continuous features, such as those given in Table 2, as inputs and takes their correlations into account to estimate a single overall LR for a comparison. The MVKD formula is described mathematically in (5). The following conventions were used: m is the number of groups (e.g. authors) in the background data; p is the number of assumed correlated variables measured on each object (e.g. message); n_i is the number of objects in each group in the background data; x_{ij} are the measurements constituting the background data; $\bar{x}_i$ are the within-object means of the background data; y_{lj} are the measurements constituting offender ($l = 1$) and suspect ($l = 2$) data; $\bar{y}_l$ are the offender and suspect means; U are the within-group variance/covariance matrices; C are the between-group variance/covariance matrices; D_l are the offender and suspect variance/covariance matrices; and h is the optimal kernel smoothing parameter. For a full mathematical exposition of the formula, see Aitken and Lucy (2004)).

$$LR = \frac{\begin{array}{c}(2\pi)^{-p}|D_1|^{-1/2}|D_2|^{-1/2}|C|^{-1/2}(mh^p)^{-1}\left|{D_1}^{-1}+{D_2}^{-1}+(h^2C)^{-1}\right|^{-\frac{1}{2}} \\ \times exp\left\{-\frac{1}{2}(\bar{y}_1-\bar{y}_2)^T(D_1+D_2)^{-1}(\bar{y}_1-\bar{y}_2)\right\} \\ \times \sum_{i=1}^{m} exp\{-\frac{1}{2}(y^*-\bar{x}_i)^T\left\{\left({D_1}^{-1}+{D_2}^{-1}\right)^{-1}+(h^2C)\right\}^{-1}(y^*-\bar{x}_i)\}\end{array}}{\begin{array}{c}(2\pi)^{-p}|C|^{-1}(mh^p)^{-2}\times\prod_{l=1}^{2}[|D_l|^{-1/2}\left|{D_l}^{-1}+(h^2C)^{-1}\right|^{-\frac{1}{2}} \\ \times \sum_{i=1}^{m} exp\{-\frac{1}{2}(\bar{y}_l-\bar{x}_i)^T(D_l+h^2C)^{-1}(\bar{y}_l-\bar{x}_i)\}]\end{array}} \quad (5a)$$

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}, \quad (5b)$$

$$x_{ij} = (x_{ij1}, \ldots, x_{ijp})^T, i \in \{1, \ldots, m\}, j \in \{1, \ldots, n_i\}, \quad (5c)$$

$$\bar{y}_l = \frac{1}{n_l}\sum_{j=1}^{n_l} y_{lj}, \tag{5d}$$

$$y_{lj} = (y_{lj1}, \dots, y_{ljp})^T, l \in \{1,2\}, j \in \{1, \dots, n_l\}, \tag{5e}$$

$$D_l = {n_l}^{-1}U, l \in \{1,2\}, \tag{5f}$$

$$h = (4/(2p+1))^{1/(p+4)} m^{-1/(p+4)}, \tag{5g}$$

$$y^* = \left({D_1}^{-1} + {D_2}^{-1}\right)^{-1}\left({D_1}^{-1}\bar{y}_1 + {D_2}^{-1}\bar{y}_2\right). \tag{5h}$$

The numerator of the MVKD formula in (5a) calculates a probabilistic likelihood influenced by the similarity between the offender and the suspect samples (e.g. the similarity between a group of messages produced by the offender and a group of messages produced by a suspect) when it is assumed that both of them share the same origin (e.g. both message groups were produced by the same author, or the prosecution hypothesis (H_p) is true). It requires the mean vectors of the offender and suspect samples ($\bar{y}_1, \bar{y}_2$), and the within-group (= within-author) variance, which is given in the form of a variance/covariance matrix. The same mean vectors of the offender and suspect samples ($\bar{y}_1, \bar{y}_2$) and the between-group (= between-author) variance (C) are used in the denominator of the formula in (5a), to estimate the likelihood of getting the same evidence when it is assumed that they are of different origins (e.g. the defence hypothesis (H_d) is true). These within-group and between-group variances (U and C) are estimated using the background database, which, in the present study, consists of 38 authors (m = 38).

The difference between the two feature vectors is evaluated using a *Mahalanobis* distance – the general form is the product $(\bar{X} - \bar{Y})^T(\Sigma)^{-1}(\bar{X} - \bar{Y})$ in the formula (e.g. the difference between offender and suspect means $(\bar{y}_1, \bar{y}_2) = (\bar{y}_1 - \bar{y}_2)^T(D_1 + D_2)^{-1}(\bar{y}_1 - \bar{y}_2)$). In the MVKD formula, the covariance matrices (D_l, $l \in \{1,2\}$) of the offender (l = 1) and the suspect (l = 2) samples are assumed to be constant, and they are estimated from the pooled within-speaker covariance matrix (U) of the background database scaled by the number of samples (n) (see (5f)). That is, only the suspect and offender means are used in the calculation of the LRs in the MVKD formula. The MVKD formula assumes normality for within-group variance, while it uses a kernel density model for between-group variance. The remaining complexities of the formula result mainly from modelling a kernel density for the between-group variance.

As explained above, the within-group and between-group covariance matrices (U and C, respectively) are estimated using the background database of 38 authors, which is small, and it may cause the matrices to become ill-conditioned. To see the adequacy of the matrices, the reciprocal condition numbers (RCNs) of the U and C were computed using the background database (38 authors) based on a sample size of 2500 tokens and the 12 authorship attribution features. For comparative purposes, the RCNs were also calculated with all available data (115 authors), under the same conditions. The more ill-conditioned a matrix is, the closer to 0 the RCN should become. The RCNs are given in Table 3.

The RCN is a continuous value, and the determination of the threshold as to whether the matrix is ill-conditioned or well-conditioned is arbitrary. However, it can be seen from Table 3 that the RCNs are very similar between the 38 authors and the 115 authors for both the U and C. That is, the matrices based on the 38 authors share the same level of adequacy with the matrices based on the 115 authors. As will be shown in §4, the MVKD procedure does not require as many as 12 features to yield the best result; it works best with 5~7 features. In those cases, the RCNs are far bigger than those given in Table 3.

Theoretically speaking, the LRs estimated by the MVKD formula should be well-calibrated. However, this is not always the case, including in the present study. Thus, the poorly-calibrated LRs estimated by the MVKD formula, which are customarily referred to as *scores*, need to be calibrated (refer to §3.4 for a detailed explanation of calibration).

3.3.2 Token-based and Character-based N-grams

In languages, word tokens follow one another, making up a clause, a sentence or an even higher unit. The occurrences of tokens or the sequences of tokens are not even. That is, a language can be described by probabilistically predicting the next token(s) in a sequence. This is a basic concept of the (token) *N*-grams, which is a statistical language model based on the probability distribution over sequences of items. Although word tokens were used as an example above, the items appearing in a sequence can be phonemes, syllables, characters and so on. Theoretically, the concept of *N*-grams can even be applied to DNA and music, for instance (Keselj et al. 2003:257). In the current study, an author's attributes are modelled using the (token-based) token *N*-grams procedure and the (character-based) character *N*-grams procedure.

The formula given in (6) (Doddington 2001) is the first step in estimating the LRs from *N*-grams. Note that the outcome of the formula is a *score*, not an LR, and the score later needs to be converted to an LR by means of calibration (refer to §3.4 for a detailed explanation of calibration).

$$score_{i.j} = \frac{log_{10}\frac{\Lambda^{i}_{author}(j)}{\Lambda_{background}(j)}}{N_j} \tag{6}$$

In short, the $score_{i,j}$ of the message groups (*i* and *j*) is expressed as the $\log_{10}$ ratio of the similarity between message group *i* and message group *j* to the typicality of message group *j* against the relevant background data, further normalised by the number of tokens or characters appearing in message group *j* (N_j) (Ishihara 2011). To determine the similarity between message group *i* and message group *j*, the former first needs to be modelled by *N*-grams (represented by Λ^{i}_{Author}), after which it is probabilistically assessed to what extent message group *j* is similar to the model (represented by $\Lambda^{i}_{Author}(j)$). The better message group *j* fits the model, the higher probability the $\Lambda^{i}_{Author}(j)$ returns. The typicality of message group *j* is measured against the background model ($\Lambda_{background}$), which was built based on the background database. Here, too, the more typical message group *j* is, the higher probability the $\Lambda_{background}(j)$ returns. For the actual implementation in this study, the ngram-count and ngram functions of the *Speech Technology and Research Laboratory Language Modelling Toolkit (SRLM)* (http://www.speech.sri.com/projects/srilm/) were used. Different minimal counts of 1, 2, 3, 4, 5, 6 7, 8, 9 or 10 were set for the *N*-grams. Thus, all *N*-grams with frequencies smaller than the set count were discounted to 0. *Good-Turing* discounting – the default smoothing method in *SRLM* – was applied (Jurafsky and Martin 2000:214-216).

An important aspect of the *N*-grams procedure is the determination of the *N*-gram length (*N*). The length (*N*) needs to be decided language-dependently according to the amount of data available. Stamatatos (2007) observed that, for an average text length of 1000 tokens, the so-called Common *N*-grams (*CNG*) method based on the character *N*-grams procedure usually returns the best results with an *N*-gram length between $N = 3$ and $N = 5$. Kajarekar et al. (2009), too, reported that, with differently selected sets of character *N*-grams, performance starts converging with $N = 3$ for texts of various lengths (ca. 1200~20000 tokens) when *N* is increased from 2 to 15. Based on the results of these previous studies, $N = 3$ was selected for the character *N*-grams procedure in this study. On the other hand, authorship attribution studies based on token *N*-grams are significantly less common than studies based on character *N*-grams. However, Ishihara (2014b) reported promising results stemming from FTC experiments of SMS messages consisting of 200~3000 tokens, in which token-based

Trigrams (N = 3) were used for modelling. Thus, the length of the token *N*-grams procedures was also set as N = 3 in the current study.

As explained in §3.3.1, a text can be modelled as a feature vector. It is probably the most traditional and prevalent model in authorship analysis, and has been often referred to as a *bag-of-words* model (Stamatatos 2009) or a *static* model (Layton et al. 2012). Although the modelling technique is popular and successful in authorship analysis, it has also been criticised for its disregard of contextual information (e.g. word order and collocations), which is disadvantageous since writing is a time-varying human activity and contextual information is likely to carry author-specific information. On the other hand, the token *N*-grams, which probabilistically model a sequence of tokens, can naturally take advantage of contextual information (Stamatatos 2009). Additionally, the character *N*-grams are considered to capture complicated stylistic information of an individualising nature at various linguistic levels (e.g. lexical, morphological, syntactic and structural), which may be overlooked by the token *N*-grams and the bag-of-words model. Thus, the three different modelling procedures that were tested in this study are designed to be able to extract different types of authorial attribution, which should theoretically contribute to an improvement in the performance of an FTC system when the experimental results are fused together.

3.4 Calibration and Fusion

As explained earlier, the outcome of the three different procedures implemented in this study are *scores*, not LRs. Those scores need to be converted to LRs, a process for which, in this study, logistic-regression calibration (Brümmer and du Preez 2006, Gonzalez-Rodriguez et al. 2007) was employed. Logistic-regression *conversion* offers calibration of a single set of scores (e.g. a set of scores derived from the MVKD procedure) or simultaneous fusion and calibration of multiple sets of scores (e.g. three parallel sets of scores derived from the three target procedures). Logistic-regression *calibration* is operated by applying linear shifting and scaling to the scores, in the log odds space, relative to a decision boundary; its aim is to minimise the magnitude and incidence of scores that are known to misleadingly support the incorrect hypothesis, and also to maximise the values of scores correctly supporting the hypotheses. A logistic-regression line, the weights of which are estimated on the basis of the scores derived from a training database, is used to monotonically shift and scale the scores of the testing database to the log LRs (logLRs). By way of exemplification, assuming a logistic-regression line of the type $y = ax + b$ (where x is the score and y is the logLR, and the weights, a and b, are estimated on the basis of the scores derived from the development database), the formula $y = ax + b$ is used to shift by the amount of b, and scale by the amount of a, the scores of the test database to the logLRs.

The score-to-logLR conversion or calibration explained above is for a single set of scores (i.e. univariate score-to-logLR conversion). Fusion enables us to combine and calibrate multiple parallel sets of scores from different sets of features or even different forensic detection systems (e.g. the scores of the three target procedures applied in this study), with the output being logLRs. The idea behind fusion is essentially the same as what was described above for a single set of scores; however, for multiple sets of scores, logistic-regression weight needs to be calculated for each set of scores. The way to implement fusion is mapped out in (7).

$$log(fused\ LR) = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n + b \quad (7)$$

In (7), $x_1, x_2, x_3 \ldots x_n$ are the scores of the first through nth set, and $a_1, a_2, a_3 \ldots a_n$ are the corresponding logistic-regression weights for scaling. The logistic-regression weight for shifting is b. These logistic-regression weights are obtained from a training database. In the context of the present study, x_1, x_2 and x_3 are the sets of scores obtained from the MVKD, token *N*-grams and character *N*-grams procedures, respectively. The scaling weights a_1, a_2 and a_3 and the shifting weight b are obtained from the scores for the SA and DA comparisons generated from the development database.

The *FoCal Toolkit* (https://sites.google.com/site/nikobrummer/focal) was used to perform calibration and fusion (Brümmer and du Preez 2006) in this study. For a more detailed exposition of the topic, see Morrison (2013), a tutorial paper on logistic-regression and fusion.

3.5 Evaluation of Performance: Log-likelihood-ratio Cost

Error or accuracy rate is naturally and intuitively considered as a way of assessing the performance of a detection system. In fact, the obtained LR can be used as a discriminatory function (LR = 1 as the threshold), and the system can be assessed by saying that, for example, out of 100 SA comparisons, 90 were correctly judged as being from the same author, producing an SA accuracy rate of 90%. Similarly, if the same system correctly judges 90 out of 100 DA comparisons as being from different authors, its DA accuracy rate is again 90%. Although accuracy or error rate (e.g. equal error rate) provides a very useful piece of information, the use of LR for binary classification, as illustrated above, does not elucidate the envisioned disposition of LR as quantified strength of evidence.

In this study, the log-likelihood-ratio cost (C_{llr}) – a gradient metric, not a categorical metric based on error vs. non-error – is used as the evaluation metric for the performance of the LR-based FTC system. C_{llr} is calculated as in (8) (Brümmer and du Preez 2006).

$$C_{llr} = \frac{1}{2}\left(\left[\frac{1}{N_{SA}}\sum_{i}^{N_{SA}} log_2\left(1+\frac{1}{LR_{SA_i}}\right)\right] + \left[\frac{1}{N_{DA}}\sum_{j}^{N_{DA}} log_2\left(1+LR_{DA_j}\right)\right]\right) \quad (8)$$

In (8), the term between square brackets on the left assesses the quality of all SA LRs; N_{SA} and LR_{SA_i} are the number of SA comparisons and the derived SA LRs, respectively. The term between square brackets on the right assesses the quality of all DA comparisons; N_{DA} and LR_{DA_j} are the number of DA comparisons and the derived DA LRs, respectively. In this metric, all LRs are given penalties. However, the LRs that support the counter-factual hypotheses (SA and DA LRs that are smaller and greater than 1, respectively) are more severely penalised according to the degree of deviation from unity (LR = 1); an LR supporting a counter-factual hypothesis with greater strength will be given a higher cost than the ones that are closer to unity, because the LR is more misleading. The C_{llr} measures the overall performance of a system based on a cost function in which there are two main components of loss: namely discrimination loss and calibration loss (Brümmer and du Preez 2006, Drygajlo et al. 2015, van Leeuwen and Brümmer 2007). The former is the minimum C_{llr} value, which is obtained after the application of the so-called pooled-adjacent-violators (PAV) transformation – an optimal non-parametric calibration procedure. The latter is obtained by subtracting the former from the C_{llr}. That is, C_{llr} can be decomposed into a discrimination loss (C_{llr}^{min}) and a calibration loss (C_{llr}^{cal}).

The C_{llr} was originally developed for use in the area of automatic speaker recognition (van Leeuwen and Brümmer 2007). It was subsequently used in forensic voice comparison (Gonzalez-Rodriguez et al. 2007, Kinoshita and Ishihara 2014, Morrison 2011), virtually as the standard metric. In practice, the C_{llr} can be tapped as a performance assessment metric for any LR-based detection system. Once again, for calculating C_{llr} (including C_{llr}^{min} and C_{llr}^{cal}), the *FoCal Toolkit* was used (Brümmer and du Preez 2006).

In this study, the performances of the different procedures are compared by means of their C_{llr} values. However, their equal error rate (*EER)* values are given as well, as references for discriminability, since some readers may be interested in them. *EER* is the error rate at the threshold where its false acceptance rate and false rejection rate become equal. In addition, the magnitude of the derived LRs is visually presented via Tippett plots.

4 Results and Discussion

4.1 System Performance and Strength of Evidence

The best-performing configurations for each of the three procedures and the result achieved by the fused system (based on a fusion of the best-performing results of the three individual procedures) are given in Table 4, which provides separate C_{llr} and *EER* readings for each of the four sample sizes (500, 1000, 1500 and 2500 tokens). C_{llr}^{min} and C_{llr}^{cal} values are also included in Table 4; they are useful indicators of whether the overall loss (C_{llr}) is due to a lack of discrimination or a problem of calibration or both.

As far as the C_{llr} values are concerned, out of the three procedures, the MVKD procedure constantly performed best, regardless of the sample size (C_{llr} = 0.68; 0.53; 0.35 and 0.21, respectively), while the procedure based on token *N*-grams performed worst for sample sizes of 500, 1000 and 1500 tokens (C_{llr} = 0.97; 0.90 and 0.65, respectively). For sample size 2500, the procedures based on token and character *N*-grams performed similarly (C_{llr} = 0.57). The underperformance of the token *N*-grams procedure was previously reported (Forsyth and Holmes 1996, Grieve 2007); on the whole, the results of the current study therefore support previous findings.

A slightly different outcome was obtained with respect to discriminating power. In terms of *EER*, it can be seen from Table 4 that the discriminating performance of the MVKD and the character *N*-grams procedures is comparable for sample sizes 500 (*EER* = 0.23), 1500 (*EER* = 0.15) and 2500 (*EER* = 0.05). Compared to previous authorship analysis studies, where the character *N*-grams procedure generally performed better than the other procedures (e.g. bag-of-words and token N-grams) (Coyotl-Morales et al. 2006, Sanderson and Guenter 2006), the MVKD procedure and the character *N*-grams performed equally well in the current study – in fact, for sample size 1000, the *EER* (= 0.17) of the MVKD procedure is better than the *EER* (= 0.20) of the character *N*-grams procedure. As explained in §3.3.1, this may be due to the fact that the MVKD formula uses an adaptation technique by which the covariance matrices of the offender and the suspect samples are estimated on the basis of the background database. An adaptation technique is known to work well to model, for example, speakers in automatic speaker recognition (Reynolds et al. 2000). Most importantly, however, the differences in assessment between C_{llr} and *EER* indicate that systems with superior discriminability do not necessarily provide good quality LRs. This point will be further discussed later in this section.

As for the tokenisation types, the whitespace tokeniser constantly achieved the best results with the MVKD procedure. However, as in the case of the token *N*-grams procedure, no consistency was observed in that sample size appeared to impact on which tokenisation type yielded the best results. This observation indicates that, for casework, it is worthwhile to try different tokenisation types for the token *N*-grams procedure to see how the performances would be affected by different sample sizes.

Although there are some slight differences (depending on sample size) in well-performing features in the MVKD procedure, five features appear to be robust regardless of the sample size; they include the VRFs (*Yule's K* (F1), *TTR* (F2) and *Honoré's R* (F3)), Punctuation character ratio (F11) and Special character ratio (F12). The results furthermore demonstrate the benefits of multiple VRFs (Holmes 1992) and the usefulness of punctuation marks (Chaski 2001) in a feature vector. It is also good to know that inclusion of all 12 features is not necessary to achieve the best result. However, it is important to point out that the features listed in Table 2 are only a small set of a large number of potential authorship attribution features (Abbasi and Chen 2008, Holmes 1994, Stamatatos 2009). Authors use all sorts of devices at every linguistic level (lexical, morphological, syntactic, semantic and so on) to produce their texts, and their authorial uniqueness is therefore encoded across a variety of levels. As the features used in this study are linguistically low level features, the FTC system is anticipated to further improve with the inclusion of linguistically higher level features and other linguistically low level features that were not tested in this study. Approximately 1000 features have

so far been proposed in the literature (Rudman 1997:360). Needless to say, this warrants further study.

As for the minimal count of *N*-grams, generally speaking, it can be judged from Table 4 that it is not necessary to set it high to achieve the best result. The low minimal count of 1~3 produced the best outcome, except in the case of the character *N*-grams procedure on a sample size of 1000 tokens, which produced its best outcome with a minimal count of 5.

It is evident from Table 4 that the fusion of the scores derived from the three different procedures brought about an improvement in performance for all four sample sizes, in that the C_{llr} values (0.54; 0.42; 0.15 and 0.20, respectively) of the fused systems are all smaller than those of the three single procedures for the corresponding sample sizes. In particular, when the C_{llr} value of the fused system is compared to that of the MVKD procedure, which is the best procedure for all four different sample sizes, for the corresponding sample size, it can be seen that the improvement is larger for sample sizes 500, 1000 and 1500 than for sample size 2500. The improvements in C_{llr} values are bigger than 0.1 in C_{llr} for 500 (0.14 = (0.68-0.54)), 1000 (0.11 (= 0.53-0.42)) and 1500 (0.20 (= 0.35-0.15)) tokens, while the improvement is minimal for 2500 tokens (0.01 (= 0.21-0.20)). The improved performance resulting from the fusion indicates that the three different procedures provide different pieces of individualising information, and also that the pieces of information are complementary.

It can be seen from Table 4 that the C_{llr}^{cal} value is notably smaller than the corresponding C_{llr}^{min} value in many of the experimental results, which indicates that the derived LRs are well calibrated.

The sample size unsurprisingly influences the performance of a system; as one would have expected, more is better. Even so, the improvement is not linear; there is a large improvement between sample sizes 1000 and 1500 in that, for example, between these sample sizes, the C_{llr} value of the fused system decreased from 0.42 to 0.15. The same observation can be made in terms of *EER*; the *EER* value of the fused system dropped by half as the sample size was increased from 1000 tokens (*EER* = 0.10) to 1500 tokens (*EER* = 0.05). The discriminability of the system continues to improve with more data, even after sample size 1500 – for example, with a sample size of 2500, the *EER* of the fused system goes down from 0.05 to 0.02. On the other hand, the performances of the character *N*-grams procedure and the fused system deteriorated with a sample size of 2500 tokens (C_{llr} = 0.57 and 0.20, respectively) compared to a sample size of 1500 tokens (C_{llr} = 0.41 and 0.15, respectively). Although this point will be looked into below when the Tippett plots are presented, the results given in Table 4 entail that the fused system of the MVKD, character and token *N*-grams procedures appears to be able to obtain optimal results when 1500-token samples are used.

The derived LRs for the best-performing configurations of the three procedures as well as for the fused system are given as Tippett plots in Figure 1 and Figure 2. Figure 1 is for sample sizes 500 and 1000, and Figure 2 is for sample sizes 1500 and 2500. In the Tippett plots, the $\log_{10}$ LRs ($\log_{10}$LRs) are accumulatively plotted, separately for the SA (red solid curve) and DA (blue dashed curve) comparisons. That is, Tippett plots present the magnitude of the derived LRs (= the strength of evidence), regardless of whether they support the correct hypothesis; the further away from unity ($\log_{10}$LR = 0), the stronger the support for either hypothesis. Taking Figure 1e as an example, it can be seen that the greatest SA $\log_{10}$LR is 2.47, which correctly supports the prosecution hypothesis (i.e. consistent-with-fact LR), whereas the smallest SA $\log_{10}$LR is -1.27, which incorrectly supports the defence hypothesis (i.e. contrary-to-fact LR). The crossing-point of the SA and DA $\log_{10}$LR curves is *EER* along the y-axis.

Figure 1 and Figure 2 show that the magnitude of the derived consistent-with-fact LRs becomes more pronounced as the sample size increases: the more tokens in a sample, the further the consistent-with-fact parts of the SA and DA curves move away from unity ($\log_{10}$LR = 0). They also show that the magnitude of the consistent-with-fact LRs derived from the fused system is greater than that of any of the single procedures. Furthermore, it is favourable to see that the magnitude of the contrary-to-fact LRs becomes weaker with more tokens for the SA comparisons. However, in

many cases, the magnitude of the contrary-to-fact DA LRs tends to be strengthened with more tokens; otherwise it remains more or less unchanged. For example, the token *N*-grams procedure returned some strong contrary-to-fact DA LRs with sample size 2500, which can be observed as a long-stretched counter-factual DA LR curve (indicated by the arrow in Figure 2f). When it comes to the character *N*-grams procedure, the change in magnitude of the LRs as a function of the sample size follows the general description provided above, up until sample size 1500.

Although it still returns the best *EER* value (= 0.05) of all single procedures, as mentioned earlier, the C_{llr} value is worse for the character *N*-grams procedure with sample size 2500 (= 0.57) than it is with sample size 1500 (= 0.41). It can be seen from Figure 2g that the magnitude of the derived LRs (both factual and counter-factual) is considerably weaker with sample size 2500 in comparison to other sample sizes (Figure 1cg and Figure 2c), which has a perceptibly negative effect on the C_{llr} value.

A close observation of the SA and DA scores of the development database and those of the test database for sample size 2500 revealed, for the character *N*-grams procedure, that there are fairly strong counter-factual scores for both the SA and DA comparisons carried out on the development database; they are nearly as great as the strongest factual SA and DA scores of the same database, while it was found that the counter-factual scores of the test database are very weak (close to $\log_{10}$LR = 0) both for the SA and DA comparisons. Unlike the counter-factual scores, which were of a different magnitude in the test and development databases for sample size 2500, the other scores appeared to be very similar in magnitude. It is quite possible that those strongly misleading scores of the development database, which is a training data set for estimating calibration weights, may have brought about extensive scaling for the logistic-regression calibration, resulting in more conservative LRs for the character *N*-grams procedure. This is a problem of inaccurate calibration, which in fact can also be seen from the poor C_{llr}^{cal} value (= 0.45) of the character *N*-grams procedure with 2500 tokens; it is far greater than the corresponding C_{llr}^{min} value (= 0.12). In many of the other results given in Table 4, the C_{llr}^{cal} is far better (smaller) than the C_{llr}^{min}. The above observation entails that some regularisation strategies to the logistic-regression objective may avoid the divergence in the training of the calibration, preventing an inappropriately large shifting transformation in calibration. However, this is outside the scope of the current study; it is a potential area for future research.

As pointed out above, the fused system performed better on sample size 1500 (C_{llr} = 0.15) than it did on sample size 2500 (C_{llr} = 0.20). The relevant Tippett plots given in Figure 2d and Figure 2h provide insight into this. They show that, as envisaged, the magnitude of the derived consistent-with-fact LRs is in fact stronger for sample size 2500 than for sample size 1500. For example, 39.1% of the DA LRs are smaller than $\log_{10}$LR = -5 for sample size 1500, whereas no less than 61.4% of the DA LRs are smaller than $\log_{10}$LR = -5 for sample size 2500. Likewise, only 5.1% of the SA LRs are greater than $\log_{10}$LR = 5 for sample size 1500, whereas 15.3% of the SA LRs are greater than $\log_{10}$LR = 5 for sample size 2500. This greater magnitude of consistent-with-fact LRs for sample size 2500 is clearly an end result of the amount of input material, which should contribute to a better C_{llr} value. However, it is notable from Figure 2h that, on a sample size of 2500 tokens, the fused system unduly generated a few strong contrary-to-fact DA LRs (indicated by the arrow in Figure 2h). These strong contrary-to-fact DA LRs generated by the fused system may have been a consequence of the strong counter-factual DA LRs produced by the token *N*-grams procedure carried out on a sample size of 2500 tokens (indicated by the arrow in Figure 2f). Nevertheless, as explained in §3.5, these strong contrary-to-fact DA LRs are heavily penalised, resulting in an uninvited higher C_{llr} value for sample size 2500 than for sample size 1500.

The best result of each procedure was also fused regardless of sample size. The C_{llr} (including C_{llr}^{min} and C_{llr}^{cal}) and *EER* values are given in Table 5. The Tippett plot appears in Figure 3. The C_{llr} and *EER* values of the fused system are as low as 0.09 and 0.02. As can be seen in Figure 3, the LRs are well calibrated, which is reflected in the very small C_{llr}^{min} value (= 0.04). Although they are outside the range of the x-axis, the strongest consistent-with-fact SA and DA LRs are $\log_{10}$LRs of

22.66 and -53.69, respectively. 20.5% of the SA LRs are greater than $\log_{10}LR = 5$ and 69.8% of the DA LRs are smaller than $\log_{10}LR = -5$.

.

In the 1990s, it was considered very difficult to reliably attribute a text of less than 500 tokens to an author (Forsyth and Holmes 1996, Ledger and Merriam 1994). On account of recent developments in text mining and machine learning, it is reported that some state-of-the-art systems can now reliably perform even with a limited amount of tokens. Layton et al. (2010), for instance, report that their system can perform with an accuracy significantly better than chance even with as few as the 140 characters of a Twitter message. However, in the context of the LR framework, these highly advanced systems still need to be assessed with regard to the quality of the LRs. This is because a system with high discriminating power is not necessarily able to derive good quality LRs. As pointed out earlier, the *EER* of the character *N*-grams procedure is as low as 0.05 on a sample size of 2500 tokens, which is the lowest *EER* amongst the non-fused systems. However, as shown in Figure 2g, the derived LRs are fairly weak: 97.4% of the SA LRs and 84.4% of the DA LRs are within the range between $\log_{10}LR > -1$ and $\log_{10}LR < 1$. That is, according to the verbal interpretation of the magnitude of $\log_{10}LR$ given in Champod and Evett (2000), 97.4% of the SA LRs and 84.4% of the DA LRs provide only *limited* support for either hypothesis, which practically means that they are not of much use as evidence.

4.2 A Limiting Strategy for Substantially Strong Likelihood Ratios

In §4.1, the strength of the derived LRs was described by referring to the Tippett plots (Figure 1, Figure 2 and Figure 3). However, it is important to explicitly point out here that some of the estimated LR values in the current study are unrealistically large. For example, the strongest consistent-with-fact SA and DA LRs of the best-performing fusion system, the results of which are given in Table 5 and Figure 3, are $\log_{10}$LRs of 22.66 and -53.69, respectively; these values are greater than even DNA cases. These immensely strong LRs (and possibly other very strong LRs) are highly likely to be due to the instability of the model trained by the small database of the current study.

Two probability models are involved in an LR system: the probability model relevant to the prosecution hypothesis (H_p) and the one relevant to the defence hypothesis (H_d). If the LR system has high discriminating power, the probability distribution under H_p and the one under H_d do not understandably coincide much; this is an inherited feature of a well-discriminating LR system. That is, when the evidence occurs in the modal area of the probability density under one hypothesis (that concurrently means that the same evidence occurs in the tail area of the probability density under the competing hypothesis), a strong LR value (either supporting H_p or H_d) naturally follows. However, an issue here is the fact that there is often very little data in the corresponding tails of the distributions. That is, the density in the tail areas of the distributions is only weakly supported by data, and the model is unavoidably based on extrapolation (Vergeer et al. 2016:482-483). When a database is small, the model is even more compromised in the tail regions, causing severe extrapolation errors, which further leads to unrealistically strong LR values.

Vergeer et al. (2016) proposed a solution to the above-explained extrapolation problem; namely the “empirical lower and upper bound LR” (ELUB) method. In brief, the ELUB method is a limiting strategy for the LRs: it sets a minimum (LR_{min}) value and a maximum (LR_{max}) value for the LRs of a system according to the size of the database and the discrimination and calibration properties of the system. Any system-LRs that fall outside the range should be limited to the LR_{min} or LR_{max} value. These limits are determined based on the normalised Bayes error-rate (NBE) (Brümmer 2010) in

combination with the introduction of misleading LRs with increasing strength. A mathematical exposition and a theoretical justification of the ELUB method are given in Vergeer et al. (2016).

Using the ELUB method, the $\log_{10}LR_{min}$ and $\log_{10}LR_{max}$ values were estimated for the best-performing fusion system. They are given in Table 6. Figure 3 is drawn here again as Figure 4 with the $\log_{10}LR_{min}$ and $\log_{10}LR_{max}$ values. Note that a narrower range is used for the x-axis in Figure 4 in comparison to Figure 3.

It can be seen from Table 6 and Figure 4 that the ELUB LR values (-1.25 and 1.45 for the $\log_{10}LR_{min}$ and $\log_{10}LR_{max}$ values, respectively) are significantly more conservative than the unlimited LRs that were originally estimated for the best-performing fusion system. However, it is highly likely that the range of the ELUB LR will be widened with more data, particularly given the promising discriminating and calibration performance of the system (C_{llr}^{min} = 0.05 and C_{llr}^{cal} = 0.04).

4.3 Likelihood Ratio, Prior Odds and Posterior Odds

As explained in §2, to quantify the strength of a hypothesis, the LR needs to be combined with the prior odds of the case concerned. That is, the magnitude of the derived LR has different implications depending on the prior odds of the case. Although forensic experts should not/cannot refer to the posterior odds, it is useful for them to keep in mind how the posterior odds is subject to the prior odds of the case, given an LR. This point will be explained below, using as an example the empirical upper limit LR ($\log_{10}LR_{max}$ = 1.45) of the best-performing fusion system (cf. Rose 2013:100). For this example case, if the prior odds is one to one (1:1), which is the most advantageous prior odds for the prosecution hypothesis, the posterior odds would be 28.184 to 1 (≈ $10^{1.45}$:1*1:1) in favour of the prosecution hypothesis; the posterior probability ≈ 96.6% (≈ 28.184/(28.184+1)*100). However, if the prior odds is one to one thousand (1:1000) in favour of the prosecution hypothesis, the posterior odds would be 0.028 to 1 (≈ $10^{3.18}$:1*1:1000). The posterior probability is ca. 2.7% (≈ 0.028/(0.028+1)*100) for the same strength of evidence. Figure 5 illustrates how the posterior probability is subject to prior odds, using a $\log_{10}$LR of 1.45 as an example.

According to Figure 5, for example, the prior odds needs to be 1:7.045957 (prior $\log_{10}$ odds = 0.84794) or lower in order to obtain a posterior probability of 80% or higher from a $\log_{10}$LR of 1.45.

5 Conclusion

It was empirically demonstrated in this study that logistic-regression fusion on the resultant LRs, which were separately estimated from three different procedures (MVKD, token *N*-grams and character *N*-grams), will result in an improvement in the quality of LRs and discriminability of the system. The employment of logistic-regression fusion appears to be more beneficial when the sample size is small (e.g. 500~1500 tokens). This is advantageous for casework in which the scarcity of data is a common problem; for example short online messages (e.g. SMS, Twitter) used in cybercrimes. The Tippett plots showed that the magnitude of the fused LRs is generally far greater than that of the LRs that were separately estimated by the three different procedures.

On the other hand, unrealistically strong LRs, most likely triggered by a small database and extrapolation errors, pointed to the instability of the model. To deal with this issue, a limiting strategy for the LRs of a system, namely the "empirical lower and upper bound LR" (ELUB) method was trialled and applied to the derived LRs of the best-performing fusion system. Although the range of the upper and lower limits is substantially more conservative than the magnitude of the original LRs, it can well be expected that the range will become wider with more data, particularly considering the very good discriminating and calibration performance of the system.

As emphasised in §1, the LR framework has not made any inroads yet into forensic authorship analysis, even though more and more different fields consider or have started considering it as the

legally and logically correct conceptual framework for assessing and presenting forensic scientific evidence (Marquis et al. 2011, Morrison 2009, Neumann et al. 2007, Zadora 2009). In forensic voice comparison (FVC), for example, which is probably the closest field to forensic authorship analysis, The European Network of Forensic Science Institutes released new guidelines in 2015 that recommend an LR approach for FVC (Drygajlo et al. 2015). Besides a handful of other FTC studies (e.g. Ishihara 2014a, Ishihara 2014c), the current paper further demonstrates that the LR framework does work for authorship text evidence. However, for the LR framework to be reliably implemented for the analysis and presentation of forensic text evidence in legal proceedings, extensive fundamental research is a prerequisite, as is the compilation of necessary databases of decent size that can be used both for research and casework.

Acknowledgements

This paper is a revised and extended version of papers presented at the 3rd Cybercrime and Trustworthy Computing Conference on 24-25 November, 2014 in Auckland, New Zealand and the 5th International Conference on Advances in Computing, Communications and Informatics on 24-27 September, 2014 in Delhi, India. The author wishes to thank Peter Vergeer for providing an R implementation of the ELUB method. Finally, the author thanks the reviewers for their constructive comments that have significantly enhanced the quality of the final work.

References

Abbasi, A. C. and Chen, H. C. (2005) Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems 20*(5): 67-75. https://doi.ieeecomputersociety.org/10.1109/MIS.2005.81

Abbasi, A. C. and Chen, H. C. (2008) Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems 26*(2): 1-29. https://doi.acm.org/10.1145/1344411.1344413

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. http://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; London: Ellis Horwood.

Aitken, C. G. G. and Taroni, F. (2004) *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester: John Wiley & Sons.

Aizawa, A. N. (2001) Linguistic techniques to improve the performance of automatic text categorization. *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*: 307-314.

Baayen, R. H. (2001) *Word Frequency Distributions*. Dordrecht; London: Kluwer Academic Publisher.

Balding, D. J. and Steele, C. D. (2015) *Weight-of-evidence for Forensic DNA Profiles*. Chichester: John Wiley & Sons.

Bozza, S., Taroni, F., Marquis, R. and Schmittbuhl, M. (2008) Probabilistic evaluation of handwriting evidence: Likelihood ratio for authorship. *Journal of the Royal Statistical Society, Series C (Applied Statistics) 57*(3): 329-341. https://dx.doi.org/10.1111/j.1467-9876.2007.00616.x

Brümmer, N. (2010). *Measuring, Refining and Calibrating Speaker and Language Information Extracted from Speech.* Ph.D. Dissertation, University of Stellenbosch, Stellenbosch, South Africa. Retrieved from http://hdl.handle.net/10019.1/5139

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

Champod, C. and Evett, I. W. (2000) Commentary on A. P. A. Broeders (1999) 'Some observations on the use of probability scales in forensic identification', Forensic Linguistics 6(2): 228-41. *International Journal of Speech Language and the Law 7*(2): 238-243. https://dx.doi.org/10.1558/ijsll.v7i2.239

Champod, C. and Meuwly, D. (2000) The inference of identity in forensic speaker recognition. *Speech Communication 31*(2-3): 193-203. https://dx.doi.org/10.1016/S0167-6393(99)00078-3

Chaski, C. E. (2001) Empirical evaluations of language-based author identification techniques. *Forensic Linguistics 8*(1): 1-65. https://dx.doi.org/10.1558/sll.2001.8.1.1

Cohen, F. (2009) Bulk email forensics. In G. Peterson (ed.), *Advances in Digital Forensics V* 306: 51-67. New York: Springer.

Corney, M. W., Anderson, A. M., Mohay, G. M. and De Vel, O. (2001) *Identifying the Authors of Suspect Email*. Unpublished paper, available from http://core.ac.uk/download/pdf/10878359.pdf

Coyotl-Morales, R. M., Villaseñor-Pineda, L., Montes-y-Gómez, M. and Rosso, P. (2006) Authorship attribution using word sequences. In J. F. Martínez-Trinidad, J. A. Carrasco Ochoa and J. Kittler (eds.), *Progress in Pattern Recognition, Image Analysis and Applications*: 844-853. Berlin, Heidelberg: Springer.

Curran, J. M. (2003) The statistical interpretation of forensic glass evidence. *International Statistical Review 71*(3): 497-520. https://dx.doi.org/10.1111/j.1751-5823.2003.tb00208.x

De Vel, O., Anderson, A., Corney, M. and Mohay, G. (2001) Mining e-mail content for author identification forensics. *ACM Sigmod Record 30*(4): 55-64. https://dx.doi.org/10.1145/604264.604272

Doddington, G. R. (2001) Speaker recognition based on idiolectal differences between speakers. In P. Dalsgaard, B. Lindberg, H. Benner and Z. H. Tan (eds.), *Proceedings of Eurospeech 2001*: 2521-2524.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. and Niemi, T. (2015) *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*: European Network of Forensic Science Institutes.

Evett, I. W. (1998) Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice 38*(3): 198-202. https://dx.doi.org/10.1016/S1355-0306(98)72105-7

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

Evett, I. W., Scranage, J. and Pinchin, R. (1993) An illustration of the advantages of efficient statistical-methods for RFLP analysis in forensic-science. *American Journal of Human Genetics 52*(3): 498-505.

Foreman, L., Champod, C., Evett, I., Lambert, J. and Pope, S. (2003) Interpreting DNA evidence: A review. *International Statistical Review 71*(3): 473-495. http://dx.doi.org/10.1111/j.1751-5823.2003.tb00207.x

Forsyth, R. S. and Holmes, D. I. (1996) Feature-finding for text classification. *Literary and Linguistic Computing 11*(4): 163-174. https://dx.doi.org/10.1093/llc/11.4.163

Fuhrman, C. P. (2008) Forensic value of backscatter from email spam. In T. Tryfonas and P. Thomas (eds.), *Proceedings of the 3rd International Annual Workshop on Digital Forensics and Incident Analysis*: 46-52.

Gao, Y. B. and Zhao, G. (2005) Knowledge-based information extraction: A case study of recognizing emails of Nigerian frauds. In A. Montoyo, R. Munoz and E. Metais (eds.), *Proceedings of the 10th Natural Language Processing and Information Systems*: 161-172.

Gonzalez-Rodriguez, J., Rose, P., Ramos-Castro, D., Toledano, D. T. and Ortega-Garcia, J. (2007) Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio Speech and Language Processing 15*(7): 2104-2115. https://dx.doi.org/10.1109/tasl.2007.902747

Grieve, J. (2007) Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing 22*(3): 251-270. https://dx.doi.org/10.1093/llc/fqm020

Holmes, D. I. (1992) A stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society, Series A (Statistics in Society) 155*(1): 91-120. https://dx.doi.org/10.2307/2982671

Holmes, D. I. (1994) Authorship attribution. *Computers and the Humanities 28*(2): 87-106. https://dx.doi.org/10.1007%2FBF01830689

Honoré, A. (1979) Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin 7*(2): 172-177.

Iqbal, F., Binsalleeh, H., Fung, B. and Debbabi, M. (2010) Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation 7*(1): 56-64. https://dx.doi.org/10.1016/j.diin.2010.03.003

Ishihara, S. (2011) A forensic authorship classification in SMS messages: A likelihood ratio based approach using N-gram. In D. Molla and D. Martinez (eds.), *Proceedings of the Australasian Language Technology Workshop 2011*: 47-56.

Ishihara, S. (2012a) A forensic text comparison in SMS messages: A likelihood ratio approach with lexical features. In N. Clarke, T. Tryfonas and R. Dodge (eds.), *Proceedings of the 7th International Workshop on Digital Forensics and Incident Analysis*: 55-65.

Ishihara, S. (2012b) Probabilistic evaluation of SMS messages as forensic evidence: Likelihood ratio based approach with lexical features. *International Journal of Digital Crime and Forensics 4*(3): 47-57. https://dx.doi.org/10.4018/jdcf.2012070104

Ishihara, S. (2014a) A fused forensic text comparison system using lexical features, word and character N-grams. In D. E. Comer, S. M. Thampi, P. Mueller, D. Krishnaswamy, B. Mallick, A. Sikora and S. Mukherjea (eds.), *Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics*: 2762-2768.

Ishihara, S. (2014b) A likelihood ratio-based evaluation of strength of authorship attribution evidence in SMS messages using N-grams. *The International Journal of Speech, Language and the Law 21*(1): 23-50. https://dx.doi.org/10.1558/ijsll.v21i1.23

Ishihara, S. (2014c) A likelihood ratio-based forensic text comparison in predatory chatlog messages. In L. G. a. J. Vaughan (ed.), *Proceedings of the 44th Conference of the Australian Linguistic Society*: 39-57.

Ishihara, S. (2014d) Predatory Chatlog messages as forensic evidence in court: A comparison of two different procedures for estimating the weight of evidence. In M. Harvey and A. Antonia (eds.), *Proceedings of the 45th Australian Linguistic Society Conference*: 131-152.

Jurafsky, D. and Martin, J. H. (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, N.J.: Prentice Hall; London: Prentice-Hall International.

Kajarekar, S. S., Scheffer, N., Graciarena, M., Shriberg, E., Stolcke, A., Ferrer, L. and Bocklet, T. (2009) The SRI NIST 2008 speaker recognition evaluation system. *Proceedings of*

International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009): 4205-4208.

Keselj, V., Peng, F., Cercone, N. and Thomas, C. (2003) N-gram-based author profiles for authorship attribution. In V. Kešelj and T. Endo (eds.), *Proceedings of the 3rd Conference Pacific Association for Computational Linguistics*: 255-264.

Khan, S. R., Nirkhi, S. M. and Dharaskar, R. V. (2012) Author identification for E-mail forensic. *Proceedings of the National Conference on Recent Trends in Computing (NCRTC)*: 29-32.

Kinoshita, Y. and Ishihara, S. (2014) Background population: How does it affect LR-based forensic voice comparison? *International Journal of Speech, Language and the Law 21*(2): 191-224. https://dx.doi.org/10.1558/ijsll.v21i2.191

Kuchler, B., Champod, C. and Evett, I. W. (2001) Earmarks as evidence: A critical review. *Journal of Forensic Science 46*(6): 1275-1284. http://dx.doi.org/10.1520/JFS15146J

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C. and Can, F. (2006) Chat mining for gender prediction. In T. Yakhno and E. J. Neuhold (eds.), *Advances in Information Systems*: 274-283. New York: Springer-Verlag Berlin/Heidelberg.

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C. and Can, F. (2008) Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management 44*(4): 1448-1466. https://dx.doi.org/10.1016/j.ipm.2007.12.009

Layton, R., Watters, P. and Dazeley, R. (2010) Authorship attribution for Twitter in 140 characters or less. In L. O'Connor (ed.), *Proceedings of the 2nd Cybercrime and Trustworthy Computing Workshop (CTC)*: 1-8.

Layton, R., Watters, P. and Dazeley, R. (2012) Recentred local profiles for authorship attribution. *Natural Language Engineering 18*(3): 293-312. https://dx.doi.org/10.1017/S1351324911000180

Ledger, G. and Merriam, T. (1994) Shakespeare, Fletcher, and the two noble kinsmen. *Literary and Linguistic Computing 9*(3): 235-248. https://dx.doi.org/10.1093/llc/9.3.235

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*(Supplement 1): S238-242. https://dx.doi.org/10.1111/j.1556-4029.2010.01602.x

Mendenhall, T. C. (1887) The characteristic curves of composition. *Science 9*(214S): 237-249. https://dx.doi.org/10.1126/science.ns-9.214S.237

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. (2011) Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice 51*(3): 91-98. https://dx.doi.org/10.1016/j.scijus.2011.03.002

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

Mosteller, F. and Wallace, D. L. (1964) *Inference and Disputed Authorship, The Federalist*. Reading, Mass.,: Addison-Wesley.

National Research Council (U.S.). (2009) Strengthening forensic science in the United States: A path forward. Retrieved on 19 November 2011 from http://www.nap.edu/catalog.php?record_id=12589

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 Sciences 52*(1): 54-64. https://dx.doi.org/10.1111/j.1556-4029.2006.00327.x

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

Pillay, S. R. and Solorio, T. (2011) Authorship attribution of web forum posts. *eCrime Researchers Summit (eCrime)*: 1-7. https://dx.doi.org/10.1109/ecrime.2010.5706693

Reynolds, D. A., Quatieri, T. F. and Dunn, R. B. (2000) Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing 10*(1-3): 19-41. https://dx.doi.org/10.1006/dspr.1999.0361

Robertson, B. and Vignaux, G. A. (1995) *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Chichester: John Wiley & Sons.

Rose, P. (2013) More is better: Likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *International Journal of Speech, Language and the Law 20*(1): 77-116. https://dx.doi.org/10.1558/ijsll.v20i1.77

Rose, P., Lucy, D. and Osanai, T. (2004) Linguistic-acoustic forensic speaker identification with likelihood ratios from a multivariate hierarchical random effects model: A "non-idiot's Bayes" approach. In S. Cassidy, F. Cox, M. Mannell and S. Palethorpe (eds.), *Proceedings of the 10th Australian International Conference on Speech Science and Technology*: 492-497.

Rudman, J. (1997) The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities 31*(4): 351-365. https://dx.doi.org/10.1023/A:1001018624850

Sanderson, C. and Guenter, S. (2006) Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In D. Jurafsky and E. Gaussier (eds.), *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*: 482-491.

Son, P. T., Du, L., Jin, H., de Vel, O., Liu, N. and Caelli, T. (2008) A simple WordNet-ontology based email retrieval system for digital forensics. In C. C. Yang, H. Chen, M. Chan, K. Chang, S. D. Lang, P. S. Chen, P. Hsieh, D. Zeng, F. Y. Wang, K. Carley, W. Mao and J. Zhan (eds.), *Intelligence and Security Informatics*: 217-228.

Stamatatos, E. (2007) Author identification using imbalanced and limited training texts. In A. M. Tjoa and R. R. Wagner (eds.), *Proceedings of the 18th International Conference on Database and Expert Systems Applications*: 237-241.

Stamatatos, E. (2009) A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology 60*(3): 538-556. https://dx.doi.org/10.1002/asi.v60:3

Stolfo, S. J., Hershkop, S., Wang, K., Nimeskern, O. and Hu, C. W. (2003) Behavior profiling of email. In H. Chen, D. D. Zeng, J. Schroeder, R. Miranda, C. Demchak and T. Madhusudan (eds.), *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics*: 74-90.

Uthus, D. C. and Aha, D. W. (2013) The Ubuntu chat corpus for multiparticipant chat analysis. In E. Hovy, V. Markman, C. Martell and D. Uthus (eds.), *Proceedings of the 2013 Association for the Advancement of Artificial Intelligence Conference Spring Symposium*: 99-102.

van Leeuwen, D. and Brümmer, N. (2007) An introduction to application-independent evaluation of speaker recognition systems. In C. Müller (ed.), *Speaker Classification I: Fundamentals, Features, and Methods*: 330-353. Berlin; New York: Springer.

Vergeer, P., van Es, A., de Jongh, A., Alberink, I. and Stoel, R. (2016) Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? *Science & Justice 56*(6): 482-491. https://doi.org/10.1016/j.scijus.2016.06.003

Wei, C., Sprague, A., Warner, G. and Skjellum, A. (2008) Mining spam email to identify common origins for forensic application. In R. L. Wainwright and H. M. Haddad (eds.), *Proceedings of the 2008 ACM Symposium on Applied Computing*: 1433-1437.

Yule, G. U. (1939) On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika 30*(3/4): 363-390.

Yule, G. U. (1944) *The Statistical Study of Literary Vocabulary*. New York: Cambridge University Press.

Zadora, G. (2009) Evaluation of evidence value of glass fragments by likelihood ratio and Bayesian network approaches. *Analytica Chimica Acta 642*(1): 279-290. https://dx.doi.org/10.1016/j.aca.2008.10.005

Zheng, R., Li, J. X., Chen, H. C. and Huang, Z. (2006) A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology 57*(3): 378-393. http://dx.doi.org/10.1002/asi.20316

Zipf, G. K. (1932) *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press.

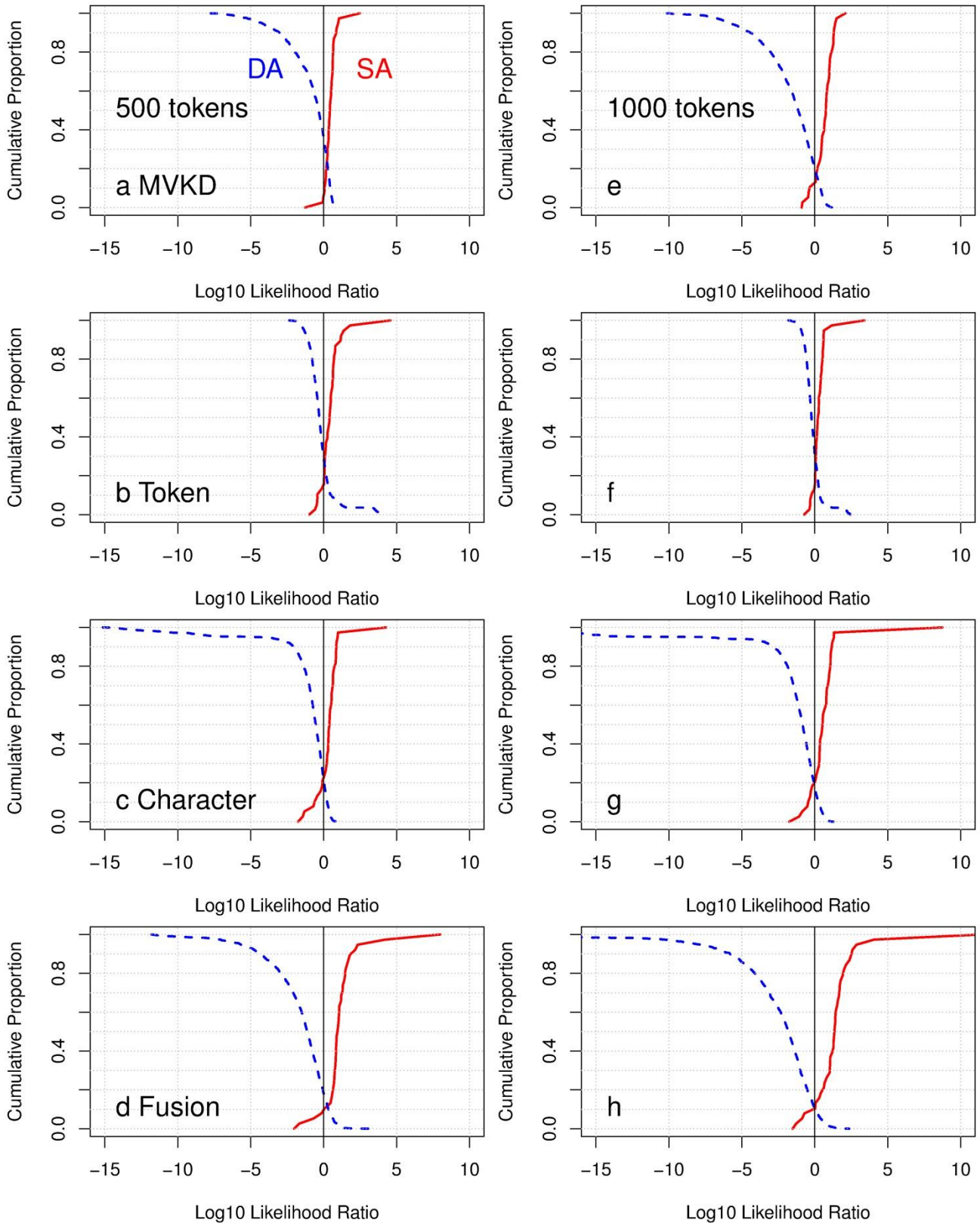

Figure 1: Tippett plots showing LRs for the best-performing configurations of the MVKD procedure (panels a and e), the token *N*-grams procedure (panels b and f), the character *N*-grams procedure (panels c and g) and the fused system (panels d and h). Red (solid) = SA comparisons; blue (dashed) = DA comparisons. Left panels = sample size 500; right panels = sample size 1000. Note that some curves extend beyond the range between $\log_{10}$LRs = -15 and $\log_{10}$LRs = 10.

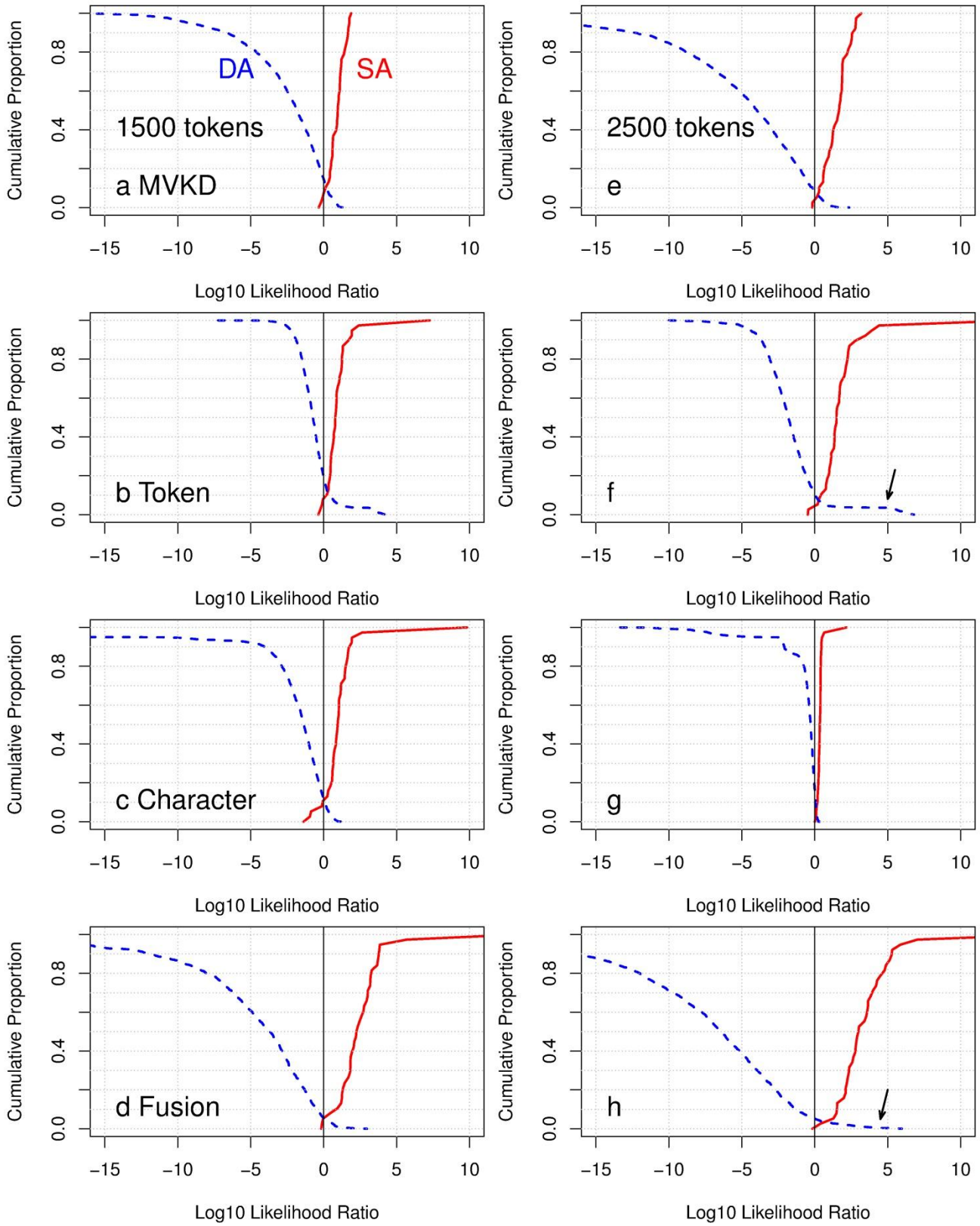

Figure 2: Tippett plots showing LRs for the best-performing configurations of the MVKD procedure (panels a and e), the token *N*-grams procedure (panels b and f), the character *N*-grams procedure (panels c and g) and the fused system (panels d and h). Red (solid) = SA comparisons; blue (dashed) = DA comparisons. Left panels = sample size 1500; right panels = sample size 2500. The arrows in panels f and h indicate large contrary-to-fact LRs of some DA comparisons.

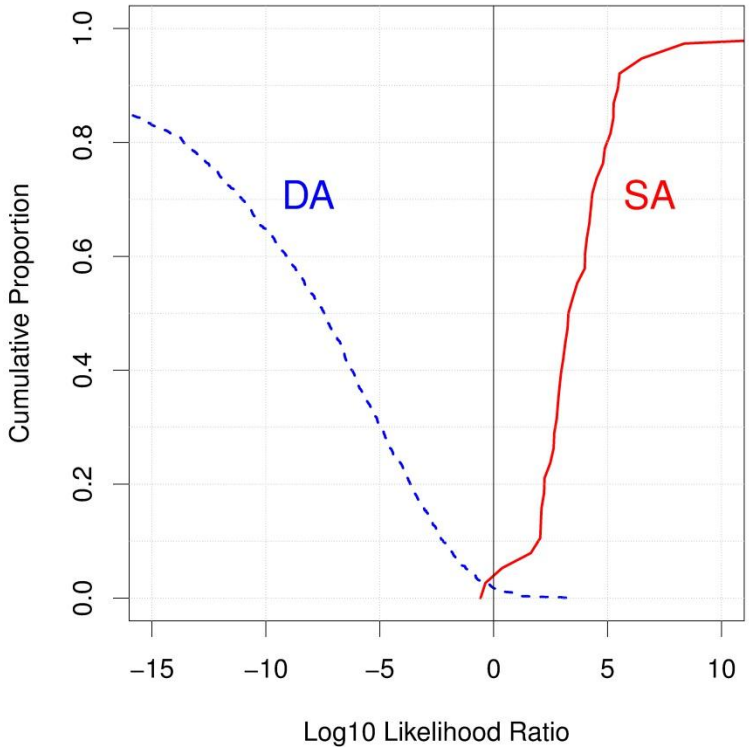

Figure 3: Tippett plot of the system that fused the best results of the three different procedures: the MVKD procedure on 2500 tokens, the token *N*-grams procedure on 2500 tokens and the character *N*-grams procedure on 1500 tokens.

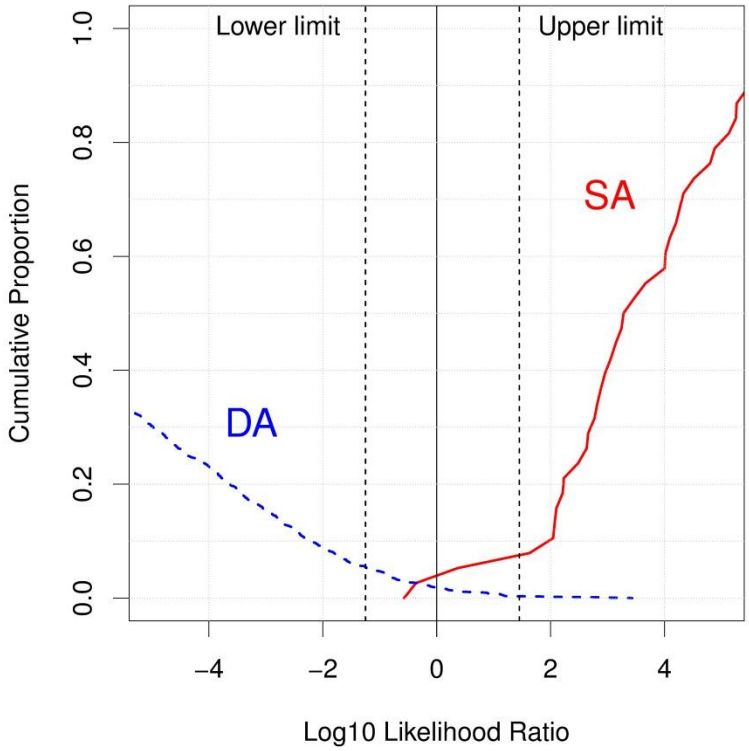

Figure 4: Replotted Figure 3 with the $\log_{10}LR_{min}$ and $\log_{10}LR_{max}$. The vertical dashed lines indicate the $\log_{10}LR_{min}$ and $\log_{10}LR_{max}$ values for the LRs. Note that a narrower range is used for the x-axis in Figure 4 compared to Figure 3.

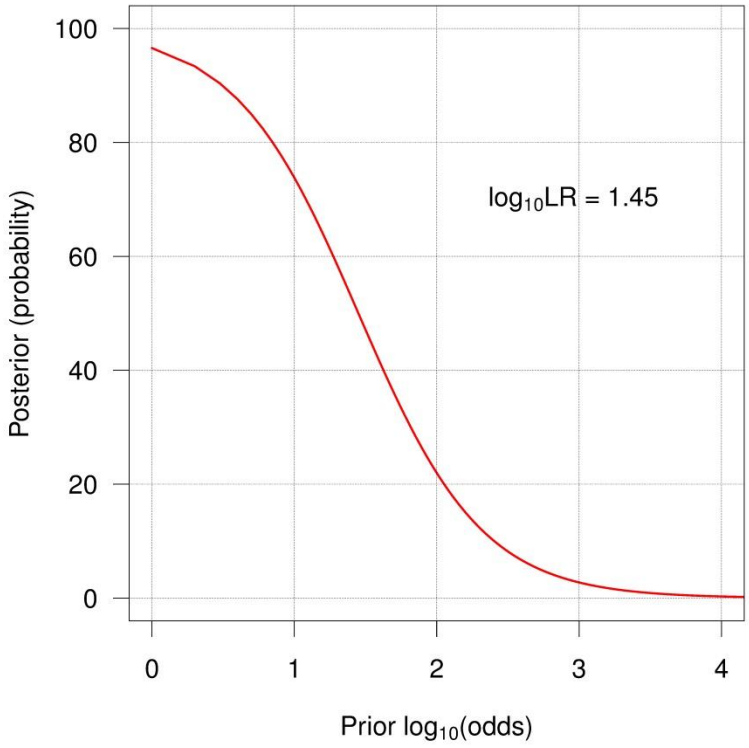

Figure 5: A graph showing how prior odds plays an important role when deriving the posterior probability from an LR. A $\log_{10}$LR of 1.45 is used as an example.

Table 1: The three different tokenisers and their outputs for the sentence “I couldn’t even buy a muffin of $3.50 when I was a student

Tokenisers	Outputs
Word	[“I”, “could”, “n't”, “even”, “buy”, “a”, “muffin”, “of”, “$”, “3.50”, “when”, “I”, “was”, “a”, “student”, “.”]
Punctuation	[“I”, “couldn”, “'”, “t”, “even”, “buy”, “a”, “muffin”, “of”, “$”, “3”, “.”, “50”, “when”, “I”, “was”, “a”, “student”, “.”]
Whitespace	[“I”, “couldn't”, “even”, “buy”, “a”, “muffin”, “of”, “$3.50”, “when”, “I”, “was”, “a”, “student.”]

Table 2: List of authorship attribution features (F1~F12) for the MVKD procedure

Feature type		Authorship attribution features
Token-based features	F1.	*Yule's I* (Inverted *Yule's K*)
	F2.	Type-token ratio (*TTR*)
	F3.	*Honoré's R*
	F4.	Average token number per message
	F5.	SD of token number appearing in message
Character-based features	F6.	Average character number per message
	F7.	SD of character number appearing in message
	F8.	Upper case ratio
	F9.	Digits ratio
	F10.	Average character number in a token
	F11.	Punctuation character ratio (, . ? ! ; : ' ")
	F12.	Special character ratio (< > % | [ ] { } \ / @ # ~ + - * $ ^ & =)

Table 3: RCNs of the within-author and between-author covariance matrices (*U* and *C*, respectively) estimated with the samples from 38 authors and 115 authors. All of the 12 authorship attribution features are included

	38 authors	115 authors
U	7.65e-10	1.87e-9
C	1.69e-10	1.53e-10

Table 4: The best-performing results of the three procedures in terms of C_{llr} and the result of the fused system for each sample size (500, 1000, 1500 and 2500 tokens). *EER* is also given as an added reference of discriminability of the systems. Within each sample size, the C_{llr} (including C_{llr}^{min} and C_{llr}^{cal}) and *EER* values for the fused system are printed in bold.

Sample size	Procedures	Tokenisation types	Features	Minimal count	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}	*EER*
500	MVKD	Whitespace	1,2,9,11,12	-	0.68	0.60	0.08	0.23
	Token	Punctuation	-	2	0.97	0.70	0.27	0.25
	Character	-	-	2	0.77	0.64	0.13	0.23
	Fused	-	-	-	**0.54**	**0.40**	**0.14**	**0.10**
1000	MVKD	Whitespace	1,2,4,6,7,11,12	-	0.53	0.46	0.07	0.17
	Token	Whitespace	-	1	0.90	0.72	0.18	0.26
	Character	-	-	5	0.65	0.56	0.09	0.20
	Fused	-	-	-	**0.42**	**0.33**	**0.09**	**0.10**
1500	MVKD	Whitespace	1,2,4,6,11,12	-	0.35	0.29	0.06	0.10
	Token	Word	-	1	0.65	0.40	0.25	0.10
	Character	-	-	3	0.41	0.33	0.08	0.10
	Fused	-	-	-	**0.15**	**0.11**	**0.04**	**0.05**
2500	MVKD	Whitespace	2,3,4,6,11,12	-	0.21	0.17	0.04	0.05
	Token	Punctuation	-	1	0.57	0.25	0.32	0.07
	Character	-	-	2	0.57	0.14	0.43	0.05
	Fused	-	-	-	**0.20**	**0.12**	**0.08**	**0.02**

Table 5: The C_{llr} (including C_{llr}^{min} and C_{llr}^{cal}) and *EER* values (in bold face) of the system that fused the best-performing results of the three procedures

Sample size	Procedures	Tokenisation types	Features	Minimal count	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}	*EER*
2500	MVKD	Whitespace	2,3,4,6,11,12	-	0.21	0.17	0.04	0.05
2500	Token	Punctuation	-	1	0.57	0.25	0.32	0.07
1500	Character	-	-	3	0.41	0.33	0.08	0.10
	Fused	**-**	**-**	**-**	**0.09**	**0.05**	**0.04**	**0.02**

Table 6: The $\log_{10}LR_{min}$ and $\log_{10}LR_{max}$ values for the best-performing fusion system.

Lower limit ($\log_{10}LR_{min}$)	Upper limit ($\log_{10}LR_{max}$)
-1.25	1.45