SFB 649 Discussion Paper 2007-057

Conditional Complexity of Compression for Authorship Attribution

Mikhail B. Malyutov*
Chammi I. Wickramasinghe*
Sufeng Li**

* Northeastern University Boston, USA
** Stanford University, USA

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

CONDITIONAL COMPLEXITY OF COMPRESSION FOR AUTHORSHIP ATTRIBUTION

M.Malyutov[1], I. Wickramasinghe[1] and S. Li[2]

[1] Mathematics Department, 567 Lake Hall, Northeastern University, Boston, MA 02115

[2] EE Dept, Stanford Univesrsity, 161 Packard Bldg, 350 Serra Mall, Stanford, CA 94305-9505

ABSTRACT

We introduce new stylometry tools based on the *sliced conditional compression complexity of literary texts* which are inspired by the nearly optimal application of the incomputable Kolmogorov conditional complexity (and presumably approximates it). Whereas other stylometry tools can occasionally be very close for different authors, our statistic is apparently strictly minimal for the true author, if the query and training texts are sufficiently large, compressor is sufficiently good and sampling bias is avoided (as in the poll samplings). We tune it and test its performance on attributing the Federalist papers (Madison vs. Hamilton). Our results confirm the previous attribution of Federalist papers by Mosteller and Wallace (1964) to Madison using the Naive Bayes classifier and the same attribution based on alternative classifiers such as SVM, and the second order Markov model of language. Then we apply our method for studying the attribution of the early poems from the Shakespeare Canon and the continuation of Marlowe's poem 'Hero and Leander' ascribed to G. Chapman.

JEL codes: C12, C15, C63

Keywords: compression complexity, authorship attribution.

Acknowledgement. This research was supported by the Deutshe Forschungsgemeinshaft through the SFB 649 Economic Risk.

1 INTRODUCTION

At the time that early stylometry methods were developed, the field of dactyloscopy emerged with the goal of establishing identification through the study of fingerprints. After initial successful tests and a process of standardization, the dactiloscopy is now universally recognized in forensic and security applications.

Statistical methods for attributing authorship, on the other hand, have not been widely accepted, even though literary critics are often brought to bear on the problem of distinguishing between literary styles of different writers.

Effective statistical attributors still need to be studied both in terms of their theoretical properties and also in empirical terms through exhaustive examination of their performance. If such work will prove that conscious and unconscious style features of different *professionals* can be discriminated as well or nearly as well as fingerprints of different individuals, stylometry will change its status from a mere *hobby* to a professional *forensic tool*.

One obstacle for implementing these methods is the *evolution and enrichment of styles* during professional careers of writers. So, unless we perform an analysis of the stylistic features of authors across time, we can only compare texts written sufficiently close to one another.

The rates of change for different stylistic features may vary. Also, authors can work in different *literary forms* (for instance, prose and verse) which may have different statistical properties. Therefore, appropriate *preprocessing* must be applied to the texts to avoid heterogeneity of forms. Misspellings need to be removed to preserve consistency. Also, *annotated texts* (e.g. verses with stressed vowels indicators) can be more useful resources for computer analysis than bare texts.

Finally, reliable stylometry analysis should take into account all available information about a disputed work (e.g. *time of its preparation* apparently overlooked in Thisted and Efron (1987)). So, it is desirable for teams of specialists in several fields (including literary experts) to be involved in these studies.

It is instructive to tune and test new exploratory stylometry tools on classical case studies such as disputed Federalist papers, where the attribution is either given or abundant previous studies agree on the same author. Our *first detailed tuning of CCC-based technology* (to the best of our knowledge) explores the conditions of its resolution power to be sufficient for discrimination between the authors, if this was done by other acknowledged tools.

Since distribution of CCC over slices has not yet been firmly established, we illuminate our results mainly by transparent histograms, plots and by few tables describing results in their uncompressed form. We also show examples of the P-values (tables 1,4) based on normality, the latter is partly justified by the q-q plots of CCC (figures 6 and 9 and many more in Wickramasinghe (2005)). Asymptotics obtained by modeling language as an ergodic stationary stochastic process have dubious accuracy although they support our approach for large samples. We sketch some mathematical and extensive simulation results justifying the CCC-based methodology on mixtures of IID-sequences in our Appendix B.

Especially intriguing are those case studies where the stylometric evidence helps to identify an otherwise unexpected candidate for authorship or to deny a popular candidate, if this attribution is confirmed by alternative credible evidence. One example of such success is the denial of Quintus Curtius Snodgrass articles' attribution to Mark Twain, later confirmed by credible documents, see section 1. A recent attribution of "Funeral elegy" to Shakespeare by Dan Foster provoked heated debates. It would be instructive to study this poem with tools of stylometry surveyed further. The previous example is a very small part of the famous controversy about the authorship of the Shakespeare Canon (SC) with the attribution result so far disputable. Various stylometry arguments and other historical and literary evidence point to the same person, although more careful analysis is needed. It would be extremely encouraging if credible evidence would confirm one day the accuracy of the stylometry results in this case study.

Starting with the attribution of early poems seems worthwhile due to the availability of substantial historical evidence, proximity of their publication dates to those of the competing candidate, and comparative homogeneity of poems as compared to plays which makes their pre-processing easier.

The first author is responsible for the design of our study and Appendices B,C. Other authors implemented appropriate computer programs and performed data analysis on a corpus of poetry and prose.

Three Appendices at the end discuss respectively poor performance of several alternative tools (including the popular Li et al method (2004) proposed for discriminating between libraries of texts) for authorship attribution, and two sketches of extensions of our study. An extended LZ-index of query text slices construction algorithm outlined in Appendix C gives us the statistics of slices' typical patterns in feasible time with moderate memory size enabling its comparison with that in the training texts and thus facilitating discussions with linguists and making our approach more robust w.r.t. spelling errors and less context-free.

2 Brief survey of micro-stylometry tools

We focus our attention on *context-free statistics of texts* calling this "microstyle". Moreover, we restrict our methods further by *taking aside microstyle methods based on grammatics* since grammatical parsing is not yet fully automated, and grammatical rules in early texts vary considerably. Context-free attributors are equally applicable to any language, even to the encoded messages which are not yet decoded. However, these methods are not always robust w.r.t. spelling errors and their resolution may be worse than that of attributors using the semantics of the text.

The pioneering stylometric study (Mendenhall, 1887, 1901) was based on histograms of word-length distribution of various authors computed for 5 different text strings of length 1000 words from each author. These papers showed significant difference of these histograms for different languages and also for some different authors (Dickens vs. Mill) using the same language. At the same time, histograms of Dickens were close to those of Thackeray in terms of their statistical variability estimated from repeated samples. The second paper describes the histograms for Shakespeare contemporaries commissioned and funded by A. Hemminway (a related study by Wilburg Zeigler, 1895, is also cited). This study demonstrated a significant difference of SC-histogram from those of all (including F. Bacon) contemporaries studied but one, calling attention to the striking practical identity of C. Marlowe's and SC histograms (Marlowe allegedly perished in May 1593 being 29 years old, under extremely suspicious circumstances (see e.g. Nicholl (1992)), two weeks before the dedication was amended into the already published anonymously submitted poem claiming it to be the very first work from SC). This identity was shown by evaluating partial histograms for certain portions of the corpora studied and comparing their inter- and intra-deviations. In an unpublished honors project (available by request) S.Li used a certain modification of Mendenhall's method for attributing popular poem ' 'Twas the night before Christmas' to H. Livingstone rather than to its official author C. Moore supporting the claim in Foster (2000).

Another distinction between the authors is in the numbers of English words they used: 8000 in Bacon's works vs. 31500 in SC including 3200 invented ones in SC which is more than the Bacon's, Jonson's and Chapman's joint contribution. Thisted and Efron's (1987) attribution of a newly discovered poem ' Shall I die...' to SC presumes identity of rates of acquiring new words and forgetting others. Thus their approach appears questionable.

Next to mention is the *Naive Bayes (NB) classifier of Mosteller and Wallace* (1964) developed during their long and very costly work over binary authorship attribution (Madison vs. Hamilton) of certain *Federalist papers* generously supported by the federal funding. After fitting appropriate para-

metric family of distributions (Poisson or negative binomial), they follow the Bayes rule for odds (*posterior odds is the product of prior odds times the likelihood ratio*), when multiplying the odds: Madison vs. Hamilton, by the sequence of likelihood ratios corresponding to the frequencies of a certain collection of relatively frequent function words, obtaining astronomical odds in favor of Madison.

This classifier presumes independence of function words usage, which is obviously **unjustified** and 'NB-likelihoods' should not be taken seriously. Among many NB-applications is Labbe's (2004) attribution of Molliere plays to Corneille, attribution of parts of 'Edward III' to SC and Fletcher, and sorting out spam e-mails (Corney (2003)).

Skipping discussion of other popular attributors emerging after NB-approach based on the SVM (see Bosh (1998)) and modeling language as a Markov Chain of order n (n-MC, see Rosenfeld (1996)), we pass directly to the CCC-attributors which demonstrated even better performance on certain applications (see e.g. Kukushkina et al (2001)) before their tuning and improvement in our study. See also Cilibrasi and Vitanyi (2005), giving a survey of numerous previous approaches such as Li et al (2004), Benedetto et al (2002), etc., where CCC-like methods were applied to classifying, clustering or categorizing languages or libraries. *The main distinction of our method* (outlined further) of all the previous approaches is compressing of *many slices of the query text* enabling the applied statistical analysis of their derived conditional complexities in terms of their location centers and spread. In this way we can judge about statistical significance of CCC-differences similarly to Mendenhall (2001).

3 Approximations of Kolmogorov complexity

3.1 Kolmogorov Complexity

To measure the algorithmic complexity of a binary string, Kolmogorov, Chaitin and Solomonoff independently proposed (around 1965) a quantity which is now called 'Kolmogorov complexity' (KC) since Kolmogorov and his pupils additionally developed a comprehensive theory which showed in particular that the algorithmic complexity contains Shannon's statistical complexity as a particular case for random strings.

We present here a rather informal sketch of KC since the methods of our paper are conceptually close and presumably can approximate Kolmogorov's approach, when better and better compressors are used.

Functions below are always understood as those computable by the so-

called Turing machine, i.e. partially recursive. Given a binary string x of length $|x|$, it can be computed ($F(y) = x$) by various binary strings y and functions $F(\cdot)$. KC is defined as

$$K(x) = \min\{|y| : F(y) = x\}.$$

The minimum above is extended over the finite set of all functions $F(\cdot)$ and finite strings $y : |y| \leq |x|$ such that the above condition holds. We have $K(x) \leq |x|$, since the trivial pair $F(x) = x$ is always an option. Thus the optimal y may be interpreted as the best **compression** (code) of x admitting the restoration (decoding) of x by means of the optimal decoder $F(\cdot)$.

Further we always use comparisons between complexities of different strings and the *same compressor is applied to both strings.*

An important additional notion is the Conditional KC (CKC) given a finite string z, defined as

$$K_F(x|z) = \min\{|y| : F(z, y) = x\} \tag{1}$$

The minimization over functions theoretically can be omitted for very long strings x, since Kolmogorov et al proved the existence of a universal function U such that the KC with respect to U, is the lower bound for the complexity:

$$K_U(x) \leq K_F(x) + c, \tag{2}$$

where c is an incomputable constant which can be arbitrarily large, making this inequality impractical for any given string.

Bennett et al (1998) started a series of papers on information distances between libraries based on CKC($x|z$) interpreting it as the length of the shortest program generating x given library z. It is natural to think that composing new text in style of z is easiest for the author of z.

CKC of particular strings can only be approximated by running better and better compressors.

In a series of papers reviewed in Cilibrasi and Vitanyi (2005), the classification and clustering of text libraries of comparable size was proposed using 'similarity metrics' mimicking information distances of Bennett et al and based on classes of commercial universal compressors satisfying certain properties. Symmetry of distance was an issue in these papers in contrast to ours. Our aim of attributing a paper of moderate size given the text libraries (corpora) of candidates for authorship does not require such a symmetry.

3.2 Analogs of Kolmogorov Complexity as attributors

P is the class of stationary ergodic sources approximated by n-MC's. Compressor family $\phi = \{\phi_n : \mathbf{B}^n \to \mathbf{B}^\infty, n = 1, 2, \ldots\}$ is universal, if for any

$P \in \mathbf{P}$ and $\epsilon > 0, \mathbf{B} = \{0,1\}$), it holds:

$$\lim_{n\to\infty} P(x \in \mathbf{B}^n : |\phi(x)| + \log P(x) \leq n\epsilon) = 1, \tag{3}$$

where $|\phi(x)|$ is the length of $\phi(x)$. Influenced by Kolmogorov (1965), Fitinhof (1966) constructed the first universal compressor followed by more practical developments including that of Ziv and Lempel. We survey further their compressor LZ78 which modifications including a sliding window became indispensable computer tools in around ten years becoming the core of a family of commercial compressors under various names including 'zip'. The economy of compression is achieved by sequentially constructing the binary tree of LZ-patterns from the text in computer RAM. If this LZ-pattern tree is stored in the external memory forming the so called extended LZ index (an algorithm of ELZ index construction and its applications are described in our Appendix C) then LZ-algorithms become a tool enabling the evaluation of patterns' statistics for stylometry study i.e. producing *authors' stylometric signatures*. In the attribution applications the original LZ-algorithms seem more appropriate than commercial ones since, for these applications, economy of processing time and memory is an excessive price for the distortion of patterns by sliding windows. The LZ78 original algorithm is as follows: the string x is sequentially parsed into phrases of minimal length that have not appeared before. Then each phrase is coded by a pair of numbers. The second of them is the last bit, while the first is the consecutive number of the substring of this phrase preceding the last bit. All phrases are separated by commas. The universality of LZ78 was proved by Wyner and Ziv implying

$$\lim_{n\to\infty} P(|\phi(x)|/|x| \to h) = 1 \quad as \quad |x| \to \infty \tag{4}$$

for $P \in \mathbf{P}$, where h is the binary entropy rate (per symbol) proved to be the asymptotic lower bound for compressing an ergodic source by Shannon (1949), establishing stationary ergodic strings as popular models of natural language. The accuracy of Shannon's asymptotic approximations is not necessarily satisfactory for moderately sized texts. Ziv (1988) applies universal compressors for statistical discrimination.

3.2.1 Sliced Relative Conditional Complexity of Compression

We define $|A|$ and $|A_c|$ as the lengths of respectively text A and its compression A_c. Their ratio is

$$CCr = |A_c|/|A| \tag{5}$$

The *concatenated* text $S = AB$ is the text starting with A and proceeding to text B without stop.

The *Conditional Complexity of Compression* and more transparent *Relative Conditional Complexity of Compression*, $0 < CCCr < 1$, of text B given text A are respectively

$$CCC(B|A) = |S_c| - |A_c|, CCCr(B|A) = CCC(B|A)/|B| \tag{6}$$

The CCC mimics a more abstract CKC in our settings and measures how adapting to patterns in the training text helps compressing the disputed text.

In our case studies we average sliced $CCCr$ of text $Q_i, i = 1, \ldots, m = [|Q|/L]$, given the firmly attributed text A, dividing the *query text* Q into slices of equal length L. Universal compressors used are the same for all sizes of texts.

$$\overline{CCCr(Q|A)} := \sum_{i=1}^{m} \frac{CCCr(Q_i|A)}{m}, \overline{CCr(Q)} := \sum_{i=1}^{m} \frac{CCr(Q_i)}{m}. \tag{7}$$

We call the last two empirical quantities '*Mean* $CCCr(Q)$ *and Mean* $CCr(Q)$*'* respectively.

Consider $H(Q, A) = CCC(Q|A) - |Q_c|$. If $\overline{CCr(Q)} \geq \overline{CCr(Q')}$ and $CCC(Q|A) < CCC(Q'|A)$ significantly, then $H(Q, A) < H(Q', A)$ asymptotically for large samples. Quantity $H(Q, A)$ *mimics the homogeneity of two strings test statistic introduced in Ryabko and Astola (2006).* Namely, we *replace their empirical Shannon entropy* h^* of the concatenated sample S (based on n-MC approximation) with $|S_c|$. Their test statistic is invariant w.r.t. interchanging A, Q and *asymptotically strictly positive* for *different* laws of A, Q, if $a < |A|/|Q| < 1/a, a > 0$). The last but not the first property holds also for $H(Q, A)$ in some range of $|A|/|Q|$ due to the lower bound for piecewise-stationary sources compression rate (Merhav (1993)). To adequately model literary texts, the order n of MC must be at least a couple of dozens, if a binary compressor is used. This makes evaluation of h^* *several orders of magnitude more intensive computationally* than that of $|S_c|$ and *requires a regularization of null or small conditional frequencies of patterns.* In our applications $|A|/|Q|$ is large to statistically assess reliability of non-asymptotic attribution. Both our case studies and statistical simulation in Appendix B show that the CCC-attribution has a good discrimination power in this range although further tuning and checks seem necessary.

3.2.2 Pre-Processing

The following steps were used to preprocess all the documents:

- **Removing nouns**. All the words beginning with capital letters except the first word of each line were copied into another file. This file was reviewed carefully and words determined to not be proper nouns were

deleted since proper nouns are not related to the style of the author. This file was used as a dictionary for this analysis. Examples of the words in this dictionary are Ovid, Venus, Neptune, etc in the literary work. Words were added to this dictionary with every new document analyzed. All the words in the dictionary were removed from the documents before compression.

- **Removing punctuation**. We remove all the punctuation from the texts in the cases studied in this paper except hyphenated words (e.g. Hard-hearted, Vine-trees, etc.) and apostrophes used to indicate the omission of letters (e.g. I'll, won't, etc.), although keeping, say dots would give us an info about the mean length of sentences (one of main tools used in Kjetsaa et al (1986)).

- **New line characters were replaced with spaces**.

3.2.3 Methodology

Firmly attributed corpora are referred as *training texts* for training the compressor and the text under investigation will be referred as the *query text*. *Query texts* may be disputed ones or those used for estimating the performance of attributors.

We usually fix the equally sized slices Q_i, Q_i of *query and training texts*, $Q, T(k)$ and calculate the averages over slices $\overline{CCr(T(k))}, \overline{CCCr(Q|T(k))}$ and their empirical standard deviations for each training text $T(k)$. Comparing $\overline{CCCr(T(k)|T)}$ of few *query texts* is also used sometimes keeping the *training text* T fixed. Although $\overline{CCCr}$ is not symmetric, it may be viewed as a *generalized distance*. Cutting texts into equal slices has proved not influencing the performance of our attributors significantly as compared to shifting the beginning of slices to include whole collections of words.

If Mean $CCr(Q)$ is significantly different from Mean $CCr(T)$, the author of the training text T can hardly be the author of Q. If **Means CCr's of Q and T are not significantly different**, then smaller is the Mean $CCCr(Q|T)$, the stronger appears the evidence for the similarities in style between two texts under certain conditions that we address later, and we expect Mean $CCCr$ to be the smallest trained on the *training text* written by the author of the query text. The necessity of keeping unconditional complexities of query text approximately equal to that of the training text is seen from the extreme case of a long query text consisting of repeated identical symbol. Its $CCCr$ is smallest for whatever training text.

We compared the $CCCr's$ for different query or training texts using two-sample t-test and non-parametric Wilcoxon test.

Suppose Mean CCr and Mean $CCCr$ (the latter is trained on his own corpus and applied to slices of his own texts different from those included into the training one) do not significantly differ from that of the query text. Then attributing (with less certainty) the query text to this author is also plausible.

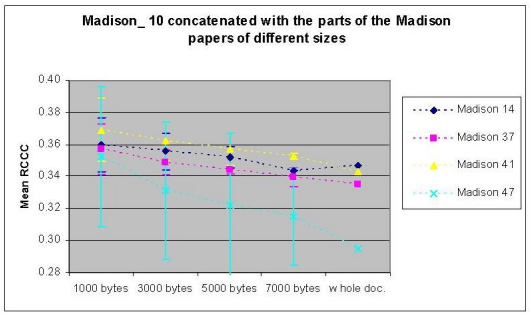
For the evolution of authors' styles to not influence our results, we compare documents presumably written during the same period of time.

Both the pre-processing, compression, statistical and graphical analysis were carried out with the codes developed in C by I. Wickramasinghe and S. Li (see Appendix D in Wickramasinghe (2005)). Earlier essentially the same attribution results of S.Li obtained with compressor BWT were reviewed in Malyutov (2005) . A sample of similar results with compressors pkzip, winzip are shown further. Thus the CCC-method appears robust w.r.t. good universal compressors.

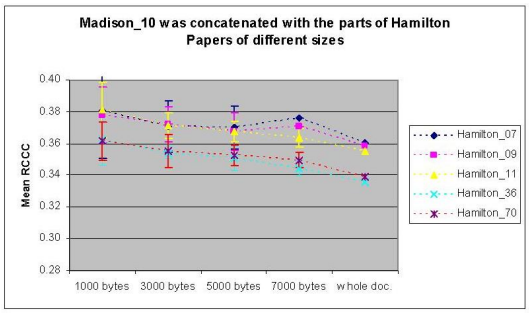
4 Attribution of Federalist Papers

4.1 The Federalist Papers

The Federalist Papers written by Alexander Hamilton, John Jay and James Madison appeared in newspapers in October 1787-August 1788 to persuade the citizens of the State of New York to ratify the U.S. Constitution. Seventy seven essays first appeared in several different newspapers all based in New York and then eight additional articles written by Hamilton on the same subject were published in a booklet form. Since then, the consensus has been that John Jay was the sole author of five (No. 2-5, No. 64) of a total 85 papers, that Hamilton was the sole author of 51 papers, that Madison was the sole author of 14 papers (No. 10,14,37-48) and that Madison and Hamilton collaborated on another three (No. 18-20). The authorship of the remaining 12 papers (No. 49-58, 62,63) has been in dispute; these papers are usually referred to as the *disputed papers*. It has been generally agreed that the *disputed papers* were written by either Madison or Hamilton, but there was no consensus about which were written by Hamilton and which by Madison. It was agreed in all previous stylometry studies that all disputed papers were written by Madison. This part of our study is therefore mostly methodological for tuning up our tool, evaluate its performance and compare it with that of alternative attributors in Appendix A. We studied first how size of the training text influences the certainty of attribution. Detailed tables are in Wickramasinghe (2005). We present a tiny sample of those results because of limitations on the size of our paper. Training on one of

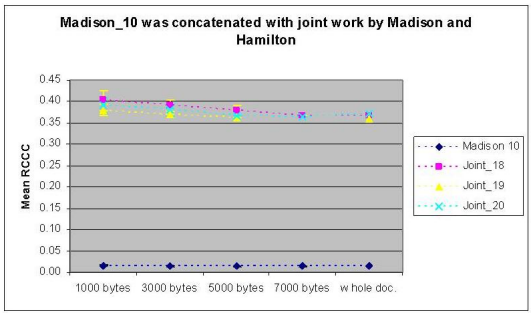


(a)

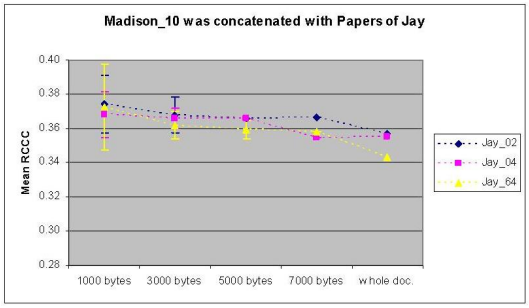


(b)

Figure 1: CCCr's when trained on one paper



(a)



(b)

Figure 2: CCCr's when trained on one paper

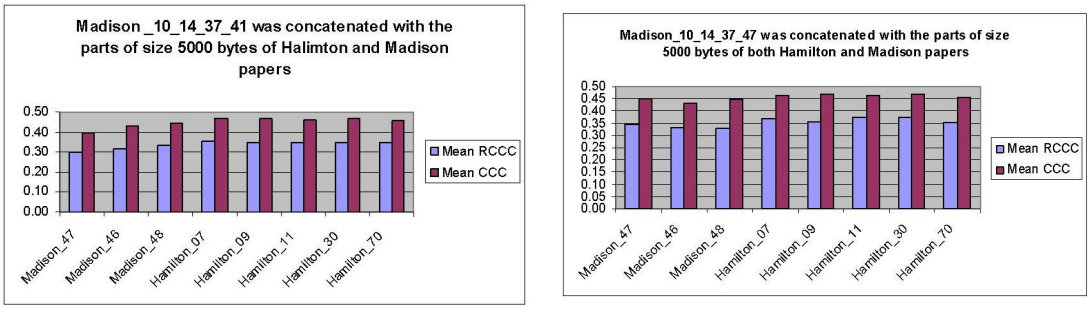
papers was not sufficient for reliable attribution (see figures 1-2).

4.2 'Leave one out' Madison's Essays as *Training Text*

Five essays written by Madison (No. 10, 14, 37, 41, 47) were taken and four of them were combined leaving one out. Five documents of the size of about 62,000-72,000 bytes after preprocessing were obtained. The compressor was *trained* on concatenation of these documents with the slices of other essays. For our *query texts* we chose 12 disputed papers as the test set and 5 of the Hamilton essays (No. 07, 09, 11, 30, 70), 2 more of Madison essays (No. 46, 48) as well as the other Madison paper we left out when we combined them as the learning set. More impressive results than those in the previous section were obtained as shown in figures 3-5.

The Combined Madison files are,

- (a) : Madison10-Madison14-Madison37-Madison41
- (b) : Madison10-Madison14-Madison37-Madison47



(a) (b)

Figure 3: Leave one out CCCr's

- (c) : Madison10-Madison14-Madison41-Madison47
- (d) : Madison10-Madison37-Madison41-Madison47
- (e) : Madison14-Madison37-Madison41-Madison47

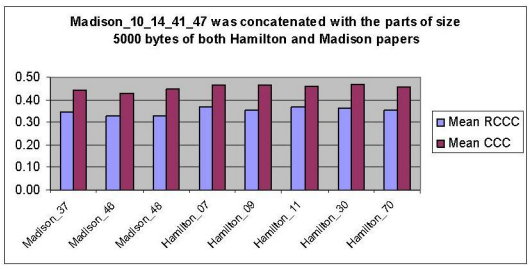
4.2.1 Comparison of $CCCr$ between Madison and Hamilton

We computed the following $CCCr$:

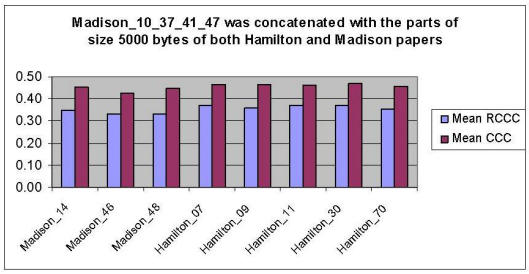
- We *trained* the compressor on M, $M \in ((a), (b), (c), (d), (e))$
- Then we applied the compressor to the concatenated file xy_i where $x \in M$, and y belongs to the set of five Hamilton papers or 2 Madison and the left-out Madison essay and y_i is the i^{th} part of the essay y
- We carried out this study by dividing the y in slices of sizes of 2000, 3000 and 5000 bytes.

We reproduce 5 similar histograms from Wickramasinghe (2005) corresponding to these training collections.

These figures show a substantial difference between the $CCCr$ obtained for the two authors. $CCCr$ for Madison was always lower than that of Hamilton in all 5 cases. The unconditional compression ratio, CCr, in general, was higher for Hamilton who seems less consistent in his style. It may be argued that lower unconditional complexity (CCr) of *query texts* of Madison helps to obtain lower $CCCr$. However, it can be seen that with higher or relatively close CCr for the *query text*, Madison No. 48 compared to the *query texts* of Hamilton No. 07 and 70, the $CCCr$ for Madison was lower than Hamilton. The next plot shows that $CCCr$ empirical distributions are close to Normal.



(a)



(b)

Figure 4: Leave one out CCCr's

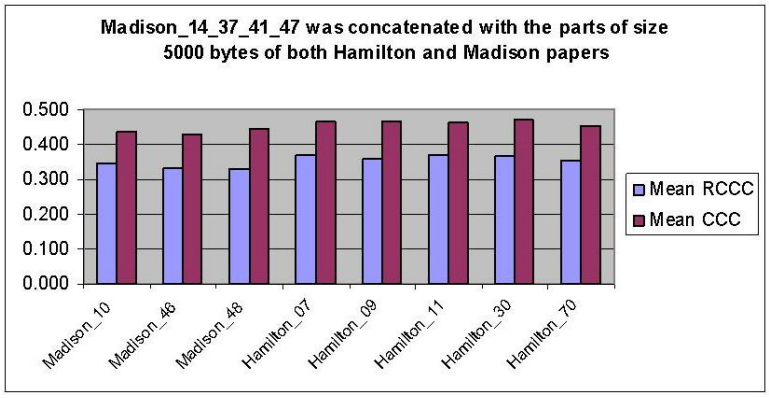


Figure 5: Leave one out CCCr's.

The following table gives the p-values for the two sample t-test for the $CCCr$ under the Alternative that means are different, when the slice sizes of the *query text* are 3000 bytes.

This and other tables in Wickramasinghe (2005) show that $CCCr$ do not differ significantly among the Madison papers. Hamilton differs significantly from Madison in most cases. If we take the unconditional complexity into account and compare only the *query texts* that do not differ significantly in CCr, it can be seen that Madison 48 has lower $CCCr$, than Hamilton 07 and 70 except in the case, where the *training text* is (a).

Comparison of $CCCr$ is more reliable when the *training text* is several times longer than the *query text*.

4.3 Comparisons with disputed papers

- We *trained* the compressor separately on each of the documents belonging to $M = \{(a), (b), (c), (d), (e)\}$. These are the combined Madison essays as described above

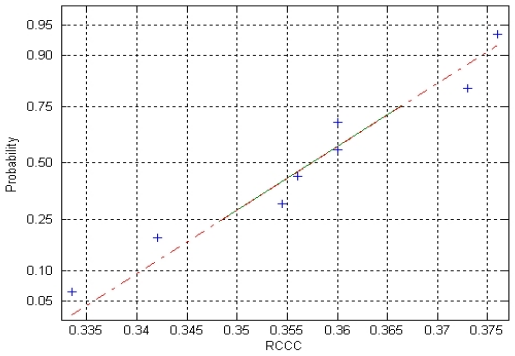


Figure 6: Normal Probability Plot of $CCCr$ for slices of Hamilton No. 70 of size 2000 bytes trained on Madison essays No. 10, No 14, No. 37 and No. 47.

Table 1: P-value of the two sample t-test
for *disputed paper No. 49*

Other documents	(a)	(b)	(c)	(d)	(e)
Hamilton 07	0.027	0.019	0.018	0.024	0.014
Hamilton 09	0.063	0.065	0.084	0.079	0.077
Hamilton 11	0.010	0.009	0.018	0.021	0.015
Hamilton 30	0.011	0.012	0.027	0.033	0.025
Hamilton 70	0.010	0.053	0.073	0.078	0.069
Madison left-(L)	0.209	0.078	0.200	0.157	0.141
Madison 46	0.215	0.486	0.373	0.371	0.433
Madison 48	0.177	0.342	0.378	0.383	0.400

- We applied the compressor on the concatenated file xy_i, where $x \in M$ and $y \in \{$*disputed papers*$\}$ and y_i is the i^{th} part of the essay y
- We carried out this study by dividing the *disputed papers* into file sizes of 2000, 3000, and 5000 bytes.
- $CCCr$ obtained for disputed papers were compared with that for *query texts* of two authors.

The next table below consists of the CCr for slices of *disputed papers* of three different sizes 2000, 3000 and 5000 bytes. The unconditional Mean compression complexity of these disputed texts is substantially lower than that of Hamilton essays and closer to that of Madison papers.

CCr decreases with increasing slice size for any given *query text* since the compressor *self-adapts to its patterns* better with larger slice size, causing

a *bias* in Mean $CCCr$. For very small slice sizes the variability of $CCCr$ becomes excessive.

4.4 Training on 13 Madison essays

We use here the same technique as before to study attribution for larger *training text*. The following documents were obtained by concatenating all the Federalist Papers attributed to Madison leaving only one out. We combined the essays in ascending order of the number of the paper. The federalist papers used are No. 10, No. 14, No. 37 - No 48 which are written by Madison. Sizes of the *training text* varied from 208,000 to 216,500 bytes.

- (a1) : Concatenate all except No. 10
- (a2) : Concatenate all except No. 14
- (a3) : Concatenate all except No. 37
- (a4) : Concatenate all except No. 38

and so on.

The following Madison's documents written between 1787-1793, to avoid evolution of the author's style were used to enlarge *training texts*

- (s) : Concatenated four papers out of five (Number 1-4) called "Helvidius papers", written in reply to series by Hamilton called "Pacificus papers" (24 Aug. - 14 Sep. 1793) on executive powers
- (t) : Concatenated eight papers from 1791-1792 Congress and republican opposition : (Mad 1 : Population and Emigration, *National Gazette*, Nov 21, 1791), (Mad 2 : consolidation, *National Gazette*, Dec. 5, 17910, (Mad 3 : Universal Peace, *National Gazette*, Feb. 2, 1792), (Mad 4 : Government of the United States, *National Gazette*, Feb 6, 1792), (Mad 5 : Spirit of Governments, *National Gazette*, Feb 20, 1792), (Mad 6 : A Candid State of Parties, *National Gazette*, Sep 26, 1792), (Mad 7 : Fashion, *National Gazette*, March 22, 1792), (Mad 8 : Property, *National Gazette*, March 29,1792)

For collections (a1)-(a4) and (s) of size 71,010 bytes as *training text*, the $CCCr$ for Madison and for disputed papers were significantly lower than that of Hamilton (we show only one of many tables in Wickramasinghe (2005)). We could not see a significant difference in $CCCr$ between the two authors when the *training text* was (t) of size only 42,316 bytes.

Table 2: Mean and StD of CCr and the number of slices for *disputed papers*

Disputed papers		size 2000	size 3000	size 5000
No. 49	Mean	0.5176	0.4750	0.4474
	StD	0.0142	0.0156	
	No of parts	4	3	1
No. 50	Mean	0.5248	0.4877	0.4484
	StD	0.0196	0.0207	
	No of parts	3	2	1
No. 51	Mean	0.4941	0.4570	0.4212
	StD	0.0129	0.0210	0.0082
	No of parts	5	3	2
No. 52	Mean	0.5286	0.4728	0.4375
	StD	0.0130	0.0359	0.0092
	No of parts	5	3	2
No. 53	Mean	0.5121	0.4696	0.4243
	StD	0.0153	0.0153	0.0177
	No of parts	6	4	2
No. 54	Mean	0.4911	0.4551	0.4165
	StD	0.0092	0.0077	0.0089
	No of parts	5	3	2
No. 55	Mean	0.5126	0.4741	0.4344
	StD	0.0126	0.0094	0.0088
	No of parts	5	3	2
No. 56	Mean	0.5019	0.4612	0.4142
	StD	0.0030	0.0110	
	No of parts	4	3	1
No. 57	Mean	0.5113	0.4753	0.4392
	StD	0.0126	0.0096	0.0008
	No of parts	6	4	2
No. 58	Mean	0.5181	0.4774	0.4308
	StD	0.0199	0.0 228	0.0130
	No of parts	6	4	2
No. 62	Mean	0.5248	0.4868	0.4453
	StD	0.0108	0.0156	0.0151
	No of parts	6	4	2
No. 63	Mean	0.5336	0.4901	0.4543
	StD	0.0120	0.0127	0.0089
	No of parts	8	5	3

Table 3: Mean and StD of $CCCr$, CCr for *disputed papers* trained on (a)

	size 2000	size 3000	size 5000	size 2000	size 3000	size 5000
	CCr	CCr	CCr	CCCr	CCCr	CCCr
No 49	0.5176	0.4750	0.4474	0.3313	0.3276	0.3260
	0.0142	0.0156		0.0064	0.0063	
	4	3	1	4	3	1
No 50	0.5248	0.4877	0.4484	0.3460	0.3438	0.3366
	0.0196	0.0207		0.0222	0.0148	
	3	2	1	3	2	1
No 51	0.4941	0.4570	0.4212	0.3120	0.3051	0.3033
	0.0129	0.0210	0.0082	0.0096	0.0076	0.0030
	5	3	2	5	3	2
No 52	0.5286	0.4728	0.4375	0.3336	0.3188	0.3172
	0.0130	0.0359	0.0092	0.0073	0.0272	0.0102
	5	3	2	5	3	2
No 53	0.5121	0.4696	0.4243	0.3213	0.3166	0.3083
	0.0153	0.0153	0.0177	0.0175	0.0158	0.0171
	6	4	2	6	4	2
No 54	0.4911	0.4551	0.4165	0.3141	0.3146	0.3074
	0.0092	0.0077	0.0089	0.0144	0.0079	0.0150
	5	3	2	5	3	2
No 55	0.5126	0.4741	0.4344	0.3273	0.3211	0.3193
	0.0126	0.0094	0.0088	0.0152	0.0037	0.0103
	5	3	2	5	3	2
No 56	0.5019	0.4612	0.4142	0.3099	0.3086	0.2954
	0.0030	0.0110		0.0105	0.0143	
	4	3	1	4	3	1
No 57	0.5113	0.4753	0.4392	0.3293	0.3276	0.3262
	0.0126	0.0096	0.0008	0.0114	0.0081	0.0023
	6	4	2	6	4	2
No 58	0.5181	0.4774	0.4308	0.3268	0.3234	0.3126
	0.0199	0.0228	0.0130	0.0225	0.0252	0.0150
	6	4	2	6	4	2
No 62	0.5248	0.4868	0.4453	0.3321	0.3303	0.3214
	0.0108	0.0156	0.0151	0.0176	0.0188	0.0124
	6	4	2	6	4	2
No 63	0.5336	0.4901	0.4543	0.3407	0.3365	0.3361
	0.0120	0.0127	0.0089	0.0179	0.0165	0.0110
	8	5	3	8	5	3

5 Shakespeare controversy

5.1 Introduction

The controversy concerning authorship of the works ascribed to W. Shakespeare dates back several centuries due to the fact that rare documents related to his life are hard for many to reconcile with his authorship (see e.g. http://shakespeareauthorship.org/). Many influential writers, scholars, actors, scientists, statesmen, etc. continue to be non-believers. A **bibliography** of material relevant to the controversy that was compiled by Prof. J. Galland in 1947 is about **1500 pages** long (see *Friedmans*, 1957). A comparable work written today might well be at least several times as large. A substantial part of research moved to the Internet, since publishing works contradicting the official version in academic journals is practically unlikely. The main problem for 'heretics' is that they do not agree on the alternative author.

Resolving the controversy would yield the Hoffmans' prize of around 1000000 English pounds, aid our understanding of what the author intended to convey in his works and contribute to a better insight into the history of culture. Methodology developed during this investigation would also be useful in other applications, including the attribution of newly discovered non-attributed texts or terrorist letters. Our contribution is minor: we discuss rather striking CCC-relation between the first two poems in SC and few other poems written by alternative contemporary authors continuing research of Mendenhall (1901).

5.1.1 CCC-attribution of some Elizabethan poems

We studied the following versions of poems with corrected spelling errors:

- SC: Venus and Adonis (1593), Rape of Lucrece (1594) (we refer to these as Venus and Rape in this study).
- Kit Marlowe's: translation of Ovid's Elegies (Amores).
- Kit Marlowe's: a version of Hero and Leander (Hero 1) both published posthumously in 1598.
- Marlowe's smoother version of Hero and Leander (Hero 2).
- disputed anapest poem 'Shall I die...' earlier attributed in Thisted and Efron (1987).

Kit's translation of Ovid's Elegies (Amores):
http://www2.prestel.co.uk/rey/ovid.htm,

Venus and Adonis (Venus): http://etext.lib.virginia.edu/etcbin/toccer-new2?id= MobVenu.sgm&images=images/modeng&data=/texts/english/modeng/parsed&tag=public&part=all

Hero and Leander (Hero1):
http://darkwing.uoregon.edu/ rbear/marlowe1.html

Hero and Leander (Hero2):
http://www2.prestel.co.uk/rey/hero.htm

Shall I die, shall I fly :
http://www.shaksper.net/archives/1997/0390.html

These versions with corrected spelling errors in original versions (produced by several publishers in two countries), were recommended to us by British linguist Peter Bull.

First, comparatively very long Amores was used for training text which we concatenated with equally-sized slices of the other poems that were used as *query text*. Thus, the size of the training text was not an issue unlike our treatment of *Federalist Papers*. We studied attribution under different sizes of slices, keeping a reasonable number of slices for estimating StD of their CCC thanks to large sizes of the poems analyzed. Later we used also the concatenated text of the two poems Amores and Venus as a *training text*.

5.1.2 $\overline{CCr}$ for the poems

We calculated the $\overline{CCr}$ for each poem divided into slices of various sizes.

In contrast to the Federalist Papers, the unconditional complexities for all four poems are **surprisingly close** for any partitioning, which shows an extraordinary consistency of the authors' style. $\overline{CCr}$ decreases with the increasing slice size as we discussed in the previous section.

5.1.3 Comparison of $CCCr$ for the poems

The plots show that in terms of $CCCr$ Marlowe's translation of Amores (the first English translation published 5 years later than Venus) helps compress Venus significantly better than his own Hero and Leander (written allegedly at around the same time as Venus before his alleged 'untimely demise'). Kit and W. Shakespeare belonged to quite different layers of the society. According to Baker's findings
(`http://www2.localaccess.com/marlowe/`), master of Cambridge University Kit was a high level spy working for two generations of Cecils ruling over Elizabethan England. Kit was employed in their covert operations in several countries and for educating a likely successor to the throne. As a rule, some

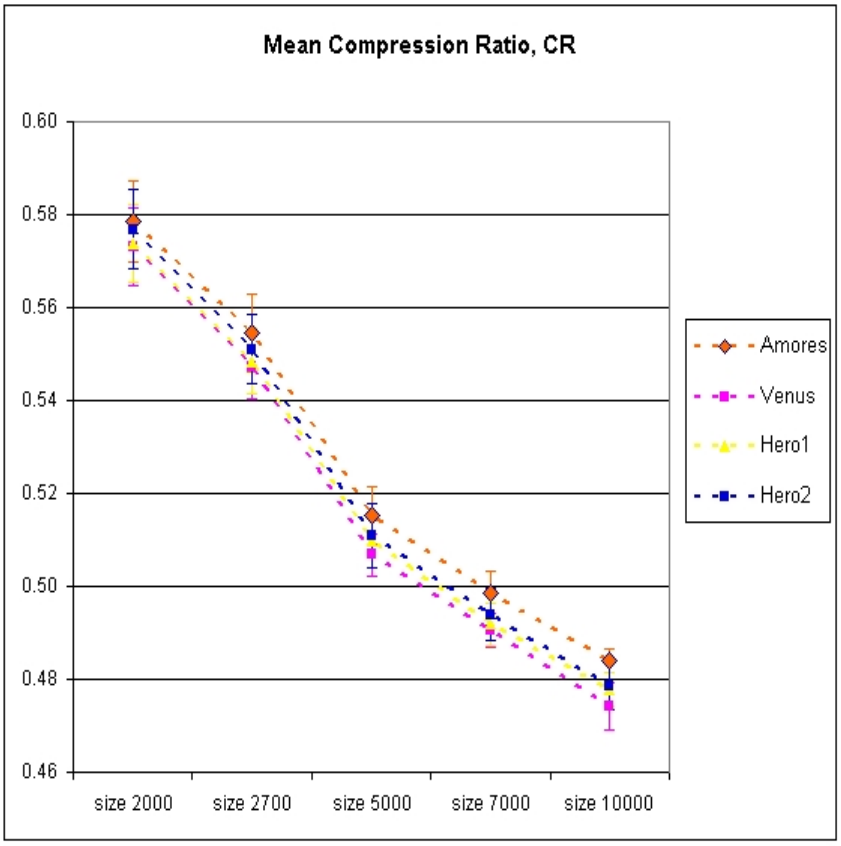


Figure 7: Mean Compression Ratio CCr for Amores, Venus, Hero 1 and Hero 2

of his patrons provided him with lodging in their estates. Any interaction with commoner W. Shakespeare associated with a competing theater is not documented and unlikely.

The normal probability plots shown support asymptotic normality of the CCC for slices.

In the p-values table below we observe that Mean $CCCr$'s are significantly different, the p-value increases with the increasing file size of the disputed text.

Thus patterns in Amores help compressing Venus significantly better than Hero 1 registered by Marlowe in 1593 and published first separately in 1598 and then (the same year) together with its twice larger continuation ascribed to G. Chapman. Amores was printed in the Netherlands in 1598 and all its copies brought to England were immediately burnt by the orders of Marlowe's deadly foe archbishop Whitgift.

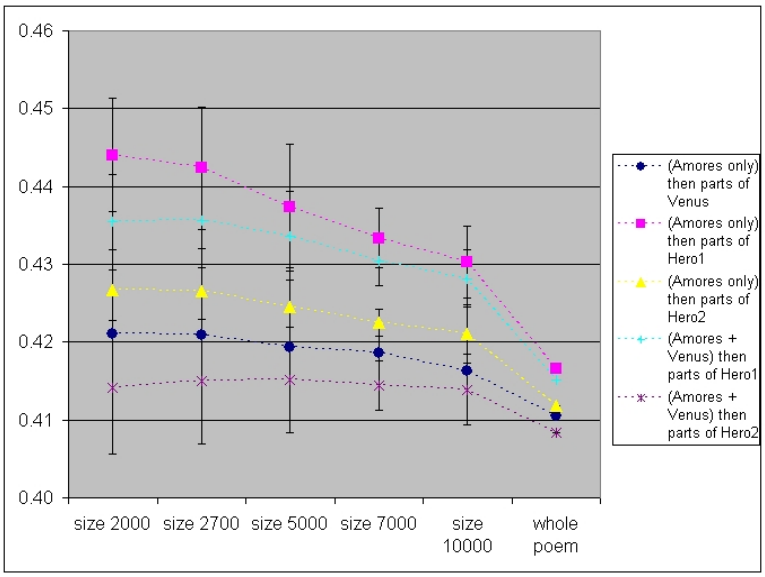


Figure 8: Mean $CCCr$ for the concatenated poems

Table 4: P-value for the two sample t-test, comparing $CCCr$

	size 10000	size 5000	size 2700	size 2000
trained on Amores only				
Venus vs Hero 1	0.00973	0.00113	$1 * 10^{-6}$	$2 * 10^{-10}$
Venus vs Hero 2	0.07148	0.03004	0.00421	0.00334
Hero 1 vs Hero 2	0.0274	0.0057	$6 * 10^{-6}$	$1 * 10^{-8}$
trained on Amores-Venus only				
Hero 1 vs Hero 2	0.00671	0.00021	$2 * 10^{-7}$	$5 * 10^{-9}$

5.2 Amores et al versus Rape of Lucrece

The second work in SC 'Rape of Lucrece' was prepared and published in haste (1594) thanks to an extraordinary success of Venus which was reprinted around ten times during 1593!

Here we compared three versions of Rape of Lucrece with the poems we studied before: Amores, Venus and Hero 1 using two different compressors winzip and pkzip and dividing our *query text* 'Rape of Lucrece' into parts of size 5000 bytes . Essentially the same results were obtained for both compressors.

We see that Venus helps compressing Rape of Lucrece significantly better than others, the concatenated *training text* 'Amores and Venus' helped even

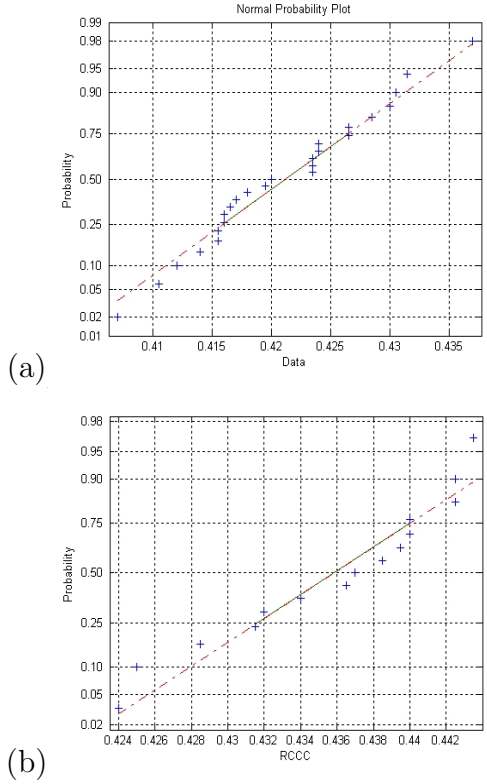


Figure 9: Normal probability plot for $CCCr$ of **(a)**:Venus trained on Amores , **(b)**: Hero 1 trained on the concatenated text of Amores, Venus

more significantly. Our *query text* 'Rape of Lucrece' was fixed using different *training texts* different in size. Whereas Amores is 102,161 bytes, Venus is 51,052 bytes and Hero1 is 33,507 bytes after pre-processing.

One of explanations of the above results would be that styles of poems following each other almost immediately are closer than those of more timely Amores which eventually was the source for both, while the final editing of Hero took place several years later.

$\overline{CCCc}(Hero2) < \overline{CCCr}(Shall) < \overline{CCCr}(Hero1)$, when trained on 'Amores'. These results make the Marlowe's authorship of both 'Venus and Adonis' and 'Shall I die, shall I fly?' likely.

5.3 Hero and Leander versus its continuation

We applied our method to compare the following poems,

- Hero1 (same as in section 3.2) vs HeroChapman 1, a continuation of Hero and Leander written by George Chapman

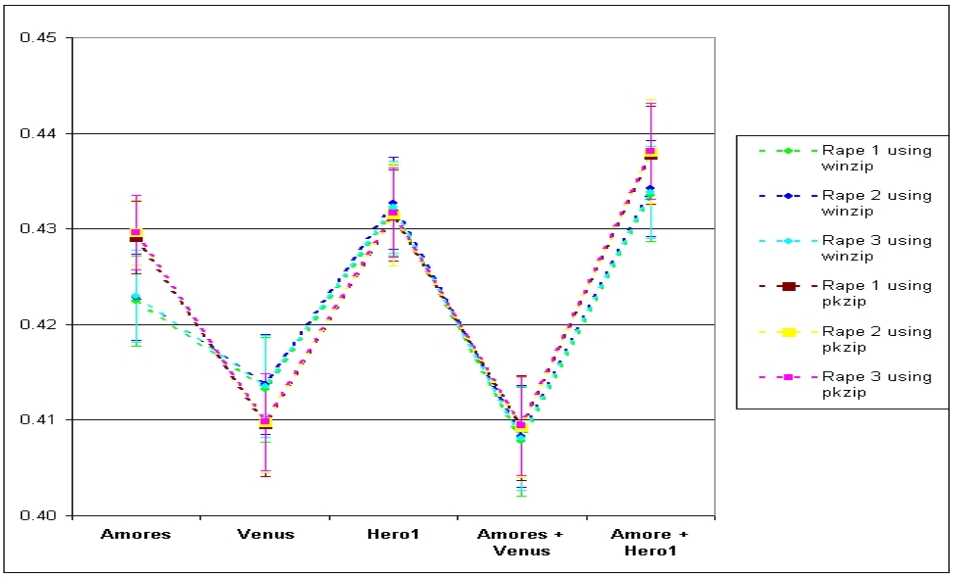


Figure 10: Mean, *StD* of *CCCr* for three versions of Rape of Lucrece with *training texts*: Amores, Venus, Amores and Venus, Amores and Hero1

- Hero 1598(another version of Hero and Leander) vs HeroChapman 1598, another version of continuation of Hero and Leander written by G. Chapman

The following two plots show *CCCr*, when query texts are Hero 1, HeroChapman 1, Hero 1598 and HeroChapman 1598 divided into parts of 2700 bytes.

5.4 Comparison with poems Chapman i, $i = 1, 2, 3$

For style comparison, Peter Bull recommended three poems: Chapman $i = 1, \ldots, 3$, namely 'The Shadow of Night', 'Ovid's Banquet of Sense' and 'The Tears of Peace' written by G. Chapman around the same time. Their Mean *CCr* are lower than those in Figure 7. We use also the poems from 5.3 as *training* for 'query' Chapman $i = 1, \ldots, 3$ which were divided into parts of size 3000 bytes. It can be seen from the table that mean *CCCr* is lower when both the *training* and the *query text* are firmly Chapman's.

One of explanations of the above results would be that Amores and Venus were written by the same author earlier (perhaps, several years before 1593), while the final editing of Hero took place several years later. While a further analysis is needed, our results do not exclude that Chapman helped publishing the Kit's Hero by putting his name on its continuation, or Chapman

Table 5: Mean,StD for $CCCr$ of Chapman $i, i = 1, 2, 3$ for slice size 3000

Training text	parts of Chap 1	parts of Chap 2	parts of Chap 3
Chapman 1			
Mean	0.2374	0.4154	0.4159
StD	0.1846	0.0089	0.0053
Chapman 2			
Mean	0.4158	0.2071	0.4127
StD	0.0077	0.2005	0.0058
HeroChapman 1			
Mean	0.4271	0.4255	0.4287
StD	0.0081	0.0082	0.0060
HeroChapman 1598			
Mean	0.4203	0.4193	0.4207
StD	0.0092	0.0095	0.0069
Hero 1			
Mean	0.4289	0.4275	0.4296
StD	0.0060	0.0070	0.0063
Hero 1598			
Mean	0.4312	0.4268	0.4320
StD	0.0062	0.0074	0.0050

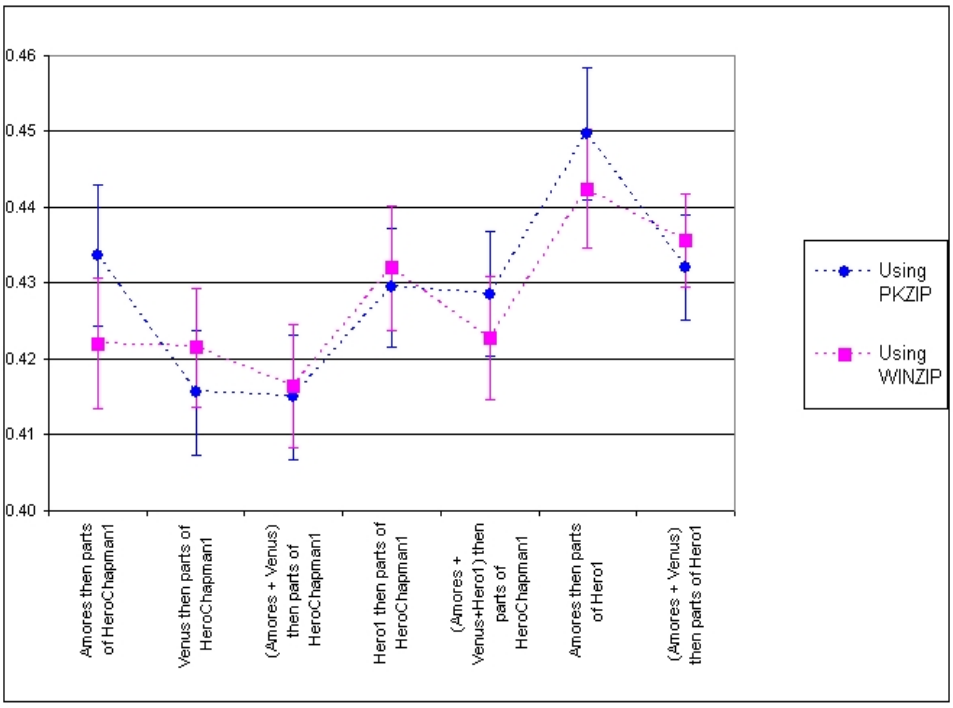


Figure 11: Mean, StD of $CCCr$ for Hero 1 and HeroChapman 1 and some training texts

edited both 'Hero' and its continuation. To distinguish between these alternatives, a further more detailed interdisciplinary comparative style analysis of Kit and G. Chapman is desirable.

An exciting textual and historical analysis of spectacularly popular at its time erotic poem 'Venus and Adonis' pointing out to its Kit's authorship is made by an University of Cambridge PhD program graduate reverend J. Baker in his essay 'Likelihood of Marlowe's authorship of 'Venus and Adonis" posted on his site
`http://www2.localaccess.com/marlowe/`. Kit cites Venus and 'rose-cheek'd' Adonis several times in the introduction to his 'Hero and Leander' (officially thought as written before his demise and thus before 'Venus and Adonis' appeared). The latter contains epigraph from 'Amores' first published 5 years later. Baker (1988) in his early stylometry study, points to the amazing constancy of numbers of English words used in 'Hero 1' and every half of its continuation.

Remark. Malyutov (2005) argues that the P-value of the homogeneity of 154 SC sonnets' lines versus the presence of anagrammed Marlowe's signatures in first two (four) of them is less than 0.0375 (respectively 2/1000).

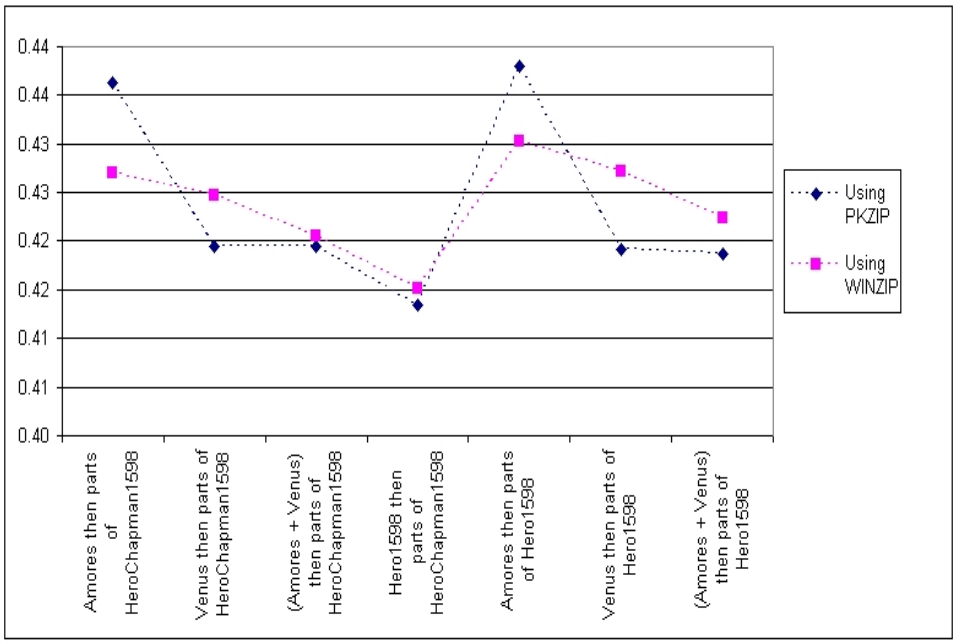


Figure 12: Mean *CCCr* taking Hero 1598 and HeroChapman 1598 as disputed poems

CONCLUSION

Our CCC case studies of literary texts show significantly different mean CCC when training on moderately sized texts of different authors. Our attribution of Federalist papers agrees with previous ones based on other acknowledged tools. Thus CCC-attributor appears a promising tool for both authorship attribution and checking chronology of texts. Its further study, tuning and comparison of its resolution power with other tools on more case studies is desirable.

Our application results can help focusing on the most likely candidates in further interdisciplinary authorship studies.

REFERENCES

Arroyuelo, D. and Navarro, G. (2005): Space-Efficient Construction of LZ-Index. In: *Lecture Notes in Computer Science. Algorithms and Computation: 16th International Symposium, ISAAC 2005, Sanya, Hainan, China, 2005. Editors: X. Deng, D. Zhu*, **3827**, 1143 - 1152.

Benedetto, D. Caglioti, E. and Loreto, V. (2002): Language Trees and Zipping. *Physical Review Letters*, **88**, No. 4, 28 January 2002, p. 048702.

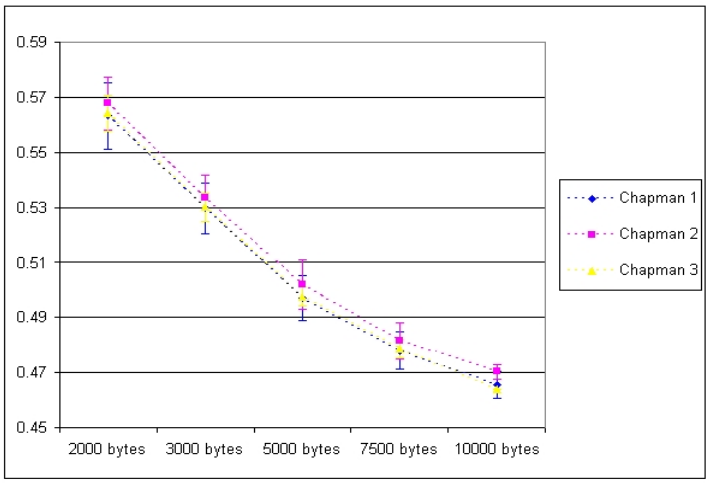


Figure 13: Mean CCr for several Chapman's poems

Bennett, C.H., Gács, P., Li, M., Vitányi, P.M.B., Zurek, W. (1998): Information Distance.*IEEE Trans. Inform. Theory*, **IT-44:4**, 1407–1423.

Bosch, R. and Smith, J. (1998): Separating hyperplanes and authorship of the Federalist papers. *Amer. Math. Monthly*, **105(7)**, 601–608.

Brinegar, C. (1963): Mark Twain and the Quintus Curtis Snodgrass Letters. *Jour. American Statistical Association*, **58(301)**, 85-96.

Cilibrasi, R. and Vitanyi, P. (2005): Clustering by Compression, *IEEE Transaction of Information Theory*, **IT-51:4**, 1523–1545

Corney, M. (2003) *Analyzing E-mail Text Authorship for Forensic Purposes*, Master Thesis, Queensland Uni. Tech. Australia.

Feller, W. 1968: *Introduction to Probability Theory and its Applications*, volume 1, 3rd edition, Wiley, N.Y.

Foster, D. (2000): *Author unknown.* H. Holt, N.Y.

Friedman, W. and Friedman, E. (1957): *The Shakespearean Ciphers exposed*, Cambridge University Press.

Kjetsaa, G., Gustavsson, S. (1986): *Authorship of Quiet Don.* Solum, Norway.

Kolmogorov A.N. (1965): Three approaches to the quantitative definition of information, *Problems of information transmission*, **1**, 3–11

Kukushkina, O., Polikarpov, A. and Khmelev, D. (2001): Text Authorship attribition using letters and grammatical information, *Problems of information transmission*, **37(2)**, 172-184

Labbe, D. (2004): *Corneille dans l'ombre de Molliere?* Les Impressiones Nouvelles, Paris-Bruxelles.

Li, M., Chen, X., Li, X. Ma, B. and Vitaniy, P. (2004): The similarity metric. *IEEE Transaction of Information Theory*, **IT-50:12**, 3250–3264.

Malyutov, M.B. (2005): Review of methods and examples of Authorship Attribution of texts. *Review of Applied and Industrial Mathematics*, TVP Pres, Vol. 12, No.1, 2005, 41-77 (In Russian). Short version: *Springer L. Notes in Comp. Sci. 4123, R. Ahlswede et al eds*, 362-380.

Mendenhall, T.A. (1887): The characteristic curves of composition, *Science*, **11**, 237–249

Mendenhall, T. A. (1901): A mechanical solution to a literary problem. *Popular Science Monthly*, **60**, 97-105.

Merhav, N. (1993): The MDL principle for piecewise stationary sources, *IEEE Trans. Inform. Th.*, **39-6**, 1962-1967.

Mosteller, F. and Wallace, D. (1964): *Inference and Disputed authorship: The Federalist papers*, Addison-Wesley.

Nicholl, Ch. (1992): *The Reckoning*, Chicago University Press.

Rocha, J., Rosella, F. and Segura, J. (2006). The Universal Similarity Metric does not detect domain similarity, arXiv:q-bio.QM/0603007 v1 6 Mar 2006

Rosenfeld, R. (1996): A Maximum Entropy Approach to Adaptive Statistical Language Modeling. *Computer, Speech and Language* **10**, 187–228.

Ryabko, B. and Astola, Y. (2006). Universal Codes as a Basis for Time Series Testing *Statistical Methodology*, **3**, 375-397.

Shannon, C. (1949): Communication Theory of Secrecy Systems. *Bell System Tech. J.*, **28**, 656-715.

Szpankowski, W. (2001): *Average Case Analysis of Algorithms on Sequences*, Wiley, N.Y.

Thisted, R. and Efron, B. (1987): Did Shakespeare write a newly discovered poem? *Biometrika*, **74**, 445–455

Wickramasinghe, C.I. (2005): PhD dissertation, Mathematics Department, Northeastern University, Boston, MA.

Ziv, J. (1988): On classification and universal data compression. *IEEE Trans. on Inform. Th.*, **34:2**, 278-286.

6 Appendices

6.1 A: Testing alternative tools

Some appealing alternatives to our CC-measures of compression complexity were also tested statistically on our case studies, namely:

A ratio-type CCC-measure $R(B|A)$:

$$R(B|A) = CCC(B|A)/|B_c|, \tag{8}$$

Relative Distance of Complexity, RDC of text B given text A

$$RDC(B|A) = |B_c| - CCC(B|A) \tag{9}$$

The *Relative Distance of Complexity as a ratio*, $RRDC$:

$$RRDC(B|A) = \frac{RDC(B|A)}{|B|}, \tag{10}$$

Normalized Compression Distance in Li et al (2004) and Cilibrasi and Vitanyi (2005). They apparently mean discrimination between libraries of texts of apparently comparable length without explicitly formulating prerequisites about the sizes. After a lengthy discussion of the ways to mimic the symmetric variant of the conditional Kolmogorov complexity of x given y

$$\max(K(x|y), K(y|x))/\max(K(x), K(y)),$$

they choose

$$NCD = (\max(|(xy)_c|, |(yx)_c|) - \min(|x_c|, |y_c|))/\max(|(xy)_c|, |(yx)_c|).$$

as an appropriate symmetric distance based on a universal compressor satisfying several asymptotic conditions. In our applications (and notation) $|(B_i)_c| < |A_c|$ and $NCD = \max\{NCD^1, NCD^2\}$.

Table 6: Mean and StD for $NCD's$, when *training texts* are (a1)-(a4)

	(a1) NCD^1	(a1) NCD^2	(a2) NCD^1	(a2) NCD^2	(a3) NCD^1	(a3) NCD^2	(a4) NCD^1	(a4) NCD^2
Ham 07	0.995	0.994	0.994	0.994	0.994	0.994	0.995	0.994
	0.000	0.001	0.000	0.001	0.000	0.001	0.000	0.001
	3	3	3	3	3	3	3	3
Ham 11	0.995	0.994	0.994	0.994	0.995	0.994	0.995	0.994
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	4	4	4	4	4	4	4	4
Ham 70	0.994	0.994	0.994	0.993	0.994	0.993	0.994	0.994
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	5	5	5	5	5	5	5	5
Ham 81	0.994	0.994	0.993	0.994	0.993	0.993	0.994	0.994
	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000
	7	7	7	7	7	7	7	7
Left-M	0.994	0.993	0.994	0.995	0.994	0.993	0.994	0.993
	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000
	4	4	5	5	5	5	4	4
Helvi	0.995	0.994	0.994	0.994	0.994	0.994	0.995	0.994
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	23	23	23	23	23	23	23	23

$$NCD^1(B_i, A) = \frac{|(AB_i)_c| - |(B_i)_c|}{|A_c|}$$

$$NCD^2(B_i, A) = (|(B_iA)_c| - |(B_i)_c|)/|A_c|$$

Our $RDC = |A_c| - (NCD^1 * |A_c|)$

In the last 2 tables we show a lack of discrimination between several sliced Hamilton and Madison query essays based on RDC, Z, NCD with (a1) to (a4) being training texts, (which were perfectly attributed by our sliced CCC-attribution). This performance makes their application for authorship attribution *in our settings* hopeless. More tables with similar outcomes are in Wickramasinghe (2005). See also Rocha et al (2006). By our opinion, the comparatively small size of slices of the query text make the $CCCr$ variation exceedingly high to put the $CCCr$, say, in the denominator of R.

6.2 B: Nonasymptotic study of CCC-attributor

We study nonasymptotic performance of CCC-attributor in two ways: i. justify it for very different distributions of query and training IID sequences and

Table 7: $RRDC, Z$, their StD when training text are (a1)-(a4)

	(a1)	(a2)	(a3)	(a4)	(a1)	(a2)	(a3)	(a4)
	RRDC	RRDC	RRDC	RRDC	Z	Z	Z	Z
Left a	0.159	0.143	0.136	0.145	0.677	0.671	0.708	0.685
	0.005	0.010	0.010	0.009	0.009	0.014	0.019	0.011
(s)	0.125	0.131	0.130	0.127	0.722	0.708	0.710	0.718
	0.010	0.009	0.010	0.010	0.018	0.015	0.020	0.019
(t)	0.116	0.132	0.119	0.117	0.757	0.724	0.750	0.755
	0.012	0.013	0.013	0.013	0.024	0.025	0.027	0.027
Ham 06	0.123	0.143	0.131	0.126	0.761	0.720	0.745	0.753
	0.007	0.003	0.006	0.007	0.011	0.003	0.010	0.011
Ham 07	0.114	0.134	0.124	0.119	0.768	0.728	0.749	0.758
	0.007	0.007	0.006	0.006	0.013	0.018	0.012	0.011
Ham 11	0.117	0.140	0.122	0.119	0.763	0.717	0.752	0.760
	0.005	0.003	0.006	0.007	0.016	0.008	0.013	0.014
Ham 15	0.127	0.146	0.132	0.128	0.749	0.712	0.739	0.746
	0.007	0.009	0.009	0.010	0.014	0.018	0.018	0.019
Ham 26	0.144	0.160	0.147	0.145	0.715	0.683	0.709	0.713
	0.012	0.007	0.009	0.011	0.024	0.010	0.016	0.020
Ham 30	0.130	0.155	0.139	0.134	0.744	0.693	0.725	0.735
	0.007	0.004	0.009	0.009	0.013	0.007	0.020	0.021
Ham 34	0.130	0.148	0.136	0.133	0.735	0.699	0.724	0.729
	0.017	0.011	0.017	0.015	0.031	0.020	0.031	0.029
Ham 69	0.143	0.146	0.148	0.144	0.699	0.694	0.689	0.697
	0.010	0.008	0.012	0.012	0.023	0.014	0.027	0.029
Ham 70	0.127	0.147	0.140	0.136	0.739	0.700	0.714	0.721
	0.009	0.006	0.008	0.008	0.023	0.018	0.021	0.023
Ham 81	0.148	0.152	0.156	0.150	0.691	0.684	0.675	0.687
	0.015	0.010	0.012	0.011	0.030	0.018	0.024	0.022
Ham 84	0.132	0.145	0.139	0.135	0.728	0.701	0.713	0.721
	0.008	0.009	0.009	0.009	0.017	0.022	0.012	0.015

ii. show the results of simulation, when these distributions can be close. Consider a training binary Bernoulli(1/100) sequence $X_1^{10000000}$ with $P(X=1) = p = 1/100$ and the query Bernoulli(0.99) sequence Y_1^{1000} with the opposite distribution $P(Y=0) = 1/100$ and compare the lengths of LZ-compressed sequences $X_1^{1001000}$ and $X_1^{10000000}Y_1^{1000}$. Note that the entropies h of X and Y are the same and thus both belong to $M(h), h = -p\log p - (1-p)\log p$. Let us support discussion of asymptotic performance of CCC in section 3.2.1 by direct arguments.

The classical von Mises's results state that the number of rare patterns in a Bernoulli(p) sequence of length N consisting of r ones has the Poisson(λ) distribution , if $Np^r(1-p) = \lambda$ for large N (see Feller, 1968, problem 11.26. The cardinality of patterns is understood there in a slightly different sense which does not influence our argument significantly).

Thus $X_1^{10000000}$ contains only the Poisson(1) distributed number of 111-patterns (i.e. only one such pattern in the mean) and *much less likely patterns with larger number of ones*. The additional length of compressed $X_1^{1001000}$ w.r.t. the length of compressed $X_1^{10000000}$ is due most likely to few occurrences of *large size patterns consisting mostly of zeroes* in the continuation of the sequence.

The length of LZ-compressed file is approximately $c\log c$ bits, where c is the number of distinct patterns in the initial string. The concatenated sequence $X_1^{10000000}Y_1^{1000}$ contains most likely **more than hundred new patterns** w.r.t. $X_1^{10000000}$ *consisting mostly of ones*, and thus the compressed $X_1^{10000000}Y_1^{1000}$ contains hundreds of additional bits w.r.t. compressed $X_1^{1001000}$ most likely.

Remark. Von Mises (see Feller, 1968, section 13.7) and Szpankowski (2001) prove the asymptotic normality of the patterns' cardinality in Bernoulli sequences which agrees with our empirical normality plots for CCCr .

A MATLAB simulation (with the code written by D. Malioutov (MIT) using the commercial update of LZ78 for UNIX systems) compared the CCC (denoted further as δ_1) of I.I.D. binary query strings of length N_2 generated first for the same randomization parameter p_1 as for the training string of length N_1, and CCC δ_2 for the second query string with the complementary randomization parameter $p_2 = 1 - p_1$ (*having the same unconditional CC*). For every training string, CCC was computed 'No. Repeats' times, all these series were performed 10 times for averaging the CCC's. We tabulate below the empirical means and standard deviations of the two δ's:

6.3 C: Outline of extended LZ index

The statistics of patterns contributing to a good compression seems desirable for optimizing the performance of attributors and for discussing the results

Table 8: CCC of complementary IID strings

Trials	Rep.	N_1	N_2	p_1	δ_1	$S(\delta_1)$	δ_2	$S(\delta_2)$	Signif.
No.10	20	$5(10)^4$	1000	0.1	104.45	6.69	117.65	5.34	High
Mean 10 tr.	20	$5(10)^4$	1000	0.1	104.32	1.27	116.45	1.64	High
No.10	20	$5(10)^4$	1000	0.4	170	2.43	170.55	2.42	No
Mean 10 tr.	20	$5(10)^4$	1000	0.4	168.95	1.19	170.96	1.55	Brink
Mean 10 tr.	20	$5(10)^4$	1000	0.45	169.64	1.54	169.66	1.65	No
No.9	20	$5(10)^5$	10000	0.4	1601.55	7.31	1603.05	6.34	No
No.10	20	$5(10)^5$	10000	0.4	1601.75	8.45	1604.8	8.52	No
Mean 10 tr.	20	$5(10)^5$	10000	0.4	1603	1.74	1604.86	1.48	Brink
No.1	20	$5(10)^5$	10000	0.4	1599.32	7.48	1600.15	7.28	No
No.10	100	$5(10)^5$	10000	0.4	1604	8.04	1605.42	7.24	No
Mean 10 tr.	100	$5(10)^5$	10000	0.4	1602.05	2	1603.46	2.08	No
No.10	100	$5(10)^5$	333	0.4	488	4.58	490.45	4.75	No
Mean 10 tr.	100	$5(10)^5$	333	0.4	487.3	1.68	488.25	2.18	No

with linguists. This goal *is different* from that pursued in construction of so-called LZ-indexes in Arroyuello and Navarro (2005). However, an economical subroutine SUBR finding for a given string S: how many out of a certain string collection S is a prefix of, was prepared by Arroyuello and Navarro (2005). SUBR is used further in our construction of extended LZ index aimed at comparison of binary LZ-trees for long training texts and small slices of a query texts each of size less than 5000 bytes. The following is an outline of the extended LZ index construction feasible for these small slices.

1. The FORWARD PATH ends up with construction of the preliminary LZ-index. Inductively, after LZ-78 algorithm finds a new pattern of minimal length, store it at the right side of the new file-string in the external memory, then a divider (say, comma), then the binary expansion of its starting location in the original file, then a second divider, THEN CONTINUE to the next new pattern of minimal length.

COMMENTARY. For strings of size less than 3000 bytes (which are shown most relevant for stylometry as slice sizes of query texts) a rude estimate gives the length of the preliminary form of the LZ-index of order of tens Mb.

2. The BACKWARD PATH. After completing construction of this preliminary form of the LZ-index, parse it from its end to its beginning: for its every pattern starting from the penultimate one check if it is the prefix of any pattern already found using SUBR. If YES, DO NOT ENUMERATE it.

3. If NOT, then this pattern is called MAXIMAL (or LEAF of the LZ-tree). They are ENUMERATED from the end and the binary expansion of their length is written from the right of the substring related to this leaf after

a divider.

4. Next ERASE all non-enumerated patterns.

5. RENUMBER leaves in the descending order of their length and place each of them (together with their length and binary expansion of the starting location) on separate lines. The string obtained is called extended LZ-index or ELZ-index.

5. Using the starting location of a leaf decode its content in English or as a sequence of English words surrounded by several bits-artifacts from its beginning and end. Some artifacts may have phonetic meaning.

6. At the end, a HISTOGRAM summarizing the HISTOGRAMS' SEQUENCE of length n-patterns' frequencies IS CONSTRUCTED, $n = 1, \ldots,$ for visualizing the ELZ-index as follows.

For every $n > 0$ a binary pattern of length n corresponds to the binary subinterval of length $2^{(-n)} \in [0, 1]$ starting at the point with binary expansion such as that of the pattern, while the multiplicity of this pattern is exactly the number of LZ-leaves such that this pattern is a prefix of those.

Instead of this sequence of 'multi-resolution' histograms, a *smoother plot* obtained by replacing the Haar base in the previous approach with smooth wavelet base such as Daubechi' base might be more transparent for representing the patterns' frequencies as coefficients of the wavelet.

The **last step** would be studying the statistics of the patterns in the long training text *matching those in the ELZ-index of query text* using the same subroutine SUBR and comparing these matches for competing candidates for authorship.

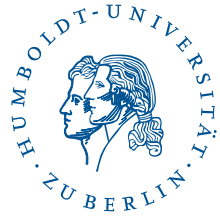
SFB 649 Discussion Paper Series 2007

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Trade Liberalisation, Process and Product Innovation, and Relative Skill Demand" by Sebastian Braun, January 2007.
002 "Robust Risk Management. Accounting for Nonstationarity and Heavy Tails" by Ying Chen and Vladimir Spokoiny, January 2007.
003 "Explaining Asset Prices with External Habits and Wage Rigidities in a DSGE Model." by Harald Uhlig, January 2007.
004 "Volatility and Causality in Asia Pacific Financial Markets" by Enzo Weber, January 2007.
005 "Quantile Sieve Estimates For Time Series" by Jürgen Franke, Jean-Pierre Stockis and Joseph Tadjuidje, February 2007.
006 "Real Origins of the Great Depression: Monopolistic Competition, Union Power, and the American Business Cycle in the 1920s" by Monique Ebell and Albrecht Ritschl, February 2007.
007 "Rules, Discretion or Reputation? Monetary Policies and the Efficiency of Financial Markets in Germany, 14th to 16th Centuries" by Oliver Volckart, February 2007.
008 "Sectoral Transformation, Turbulence, and Labour Market Dynamics in Germany" by Ronald Bachmann and Michael C. Burda, February 2007.
009 "Union Wage Compression in a Right-to-Manage Model" by Thorsten Vogel, February 2007.
010 "On σ−additive robust representation of convex risk measures for unbounded financial positions in the presence of uncertainty about the market model" by Volker Krätschmer, March 2007.
011 "Media Coverage and Macroeconomic Information Processing" by Alexandra Niessen, March 2007.
012 "Are Correlations Constant Over Time? Application of the CC-TRIG$_t$-test to Return Series from Different Asset Classes." by Matthias Fischer, March 2007.
013 "Uncertain Paternity, Mating Market Failure, and the Institution of Marriage" by Dirk Bethmann and Michael Kvasnicka, March 2007.
014 "What Happened to the Transatlantic Capital Market Relations?" by Enzo Weber, March 2007.
015 "Who Leads Financial Markets?" by Enzo Weber, April 2007.
016 "Fiscal Policy Rules in Practice" by Andreas Thams, April 2007.
017 "Empirical Pricing Kernels and Investor Preferences" by Kai Detlefsen, Wolfgang Härdle and Rouslan Moro, April 2007.
018 "Simultaneous Causality in International Trade" by Enzo Weber, April 2007.
019 "Regional and Outward Economic Integration in South-East Asia" by Enzo Weber, April 2007.
020 "Computational Statistics and Data Visualization" by Antony Unwin, Chun-houh Chen and Wolfgang Härdle, April 2007.
021 "Ideology Without Ideologists" by Lydia Mechtenberg, April 2007.
022 "A Generalized ARFIMA Process with Markov-Switching Fractional Differencing Parameter" by Wen-Jen Tsay and Wolfgang Härdle, April 2007.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".



023 "Time Series Modelling with Semiparametric Factor Dynamics" by Szymon Borak, Wolfgang Härdle, Enno Mammen and Byeong U. Park, April 2007.
024 "From Animal Baits to Investors' Preference: Estimating and Demixing of the Weight Function in Semiparametric Models for Biased Samples" by Ya'acov Ritov and Wolfgang Härdle, May 2007.
025 "Statistics of Risk Aversion" by Enzo Giacomini and Wolfgang Härdle, May 2007.
026 "Robust Optimal Control for a Consumption-Investment Problem" by Alexander Schied, May 2007.
027 "Long Memory Persistence in the Factor of Implied Volatility Dynamics" by Wolfgang Härdle and Julius Mungo, May 2007.
028 "Macroeconomic Policy in a Heterogeneous Monetary Union" by Oliver Grimm and Stefan Ried, May 2007.
029 "Comparison of Panel Cointegration Tests" by Deniz Dilan Karaman Örsal, May 2007.
030 "Robust Maximization of Consumption with Logarithmic Utility" by Daniel Hernández-Hernández and Alexander Schied, May 2007.
031 "Using Wiki to Build an E-learning System in Statistics in Arabic Language" by Taleb Ahmad, Wolfgang Härdle and Sigbert Klinke, May 2007.
032 "Visualization of Competitive Market Structure by Means of Choice Data" by Werner Kunz, May 2007.
033 "Does International Outsourcing Depress Union Wages? by Sebastian Braun and Juliane Scheffel, May 2007.
034 "A Note on the Effect of Outsourcing on Union Wages" by Sebastian Braun and Juliane Scheffel, May 2007.
035 "Estimating Probabilities of Default With Support Vector Machines" by Wolfgang Härdle, Rouslan Moro and Dorothea Schäfer, June 2007.
036 "Yxilon – A Client/Server Based Statistical Environment" by Wolfgang Härdle, Sigbert Klinke and Uwe Ziegenhagen, June 2007.
037 "Calibrating CAT Bonds for Mexican Earthquakes" by Wolfgang Härdle and Brenda López Cabrera, June 2007.
038 "Economic Integration and the Foreign Exchange" by Enzo Weber, June 2007.
039 "Tracking Down the Business Cycle: A Dynamic Factor Model For Germany 1820-1913" by Samad Sarferaz and Martin Uebele, June 2007.
040 "Optimal Policy Under Model Uncertainty: A Structural-Bayesian Estimation Approach" by Alexander Kriwoluzky and Christian Stoltenberg, July 2007.
041 "QuantNet – A Database-Driven Online Repository of Scientific Information" by Anton Andriyashin and Wolfgang Härdle, July 2007.
042 "Exchange Rate Uncertainty and Trade Growth - A Comparison of Linear and Nonlinear (Forecasting) Models" by Helmut Herwartz and Henning Weber, July 2007.
043 "How do Rating Agencies Score in Predicting Firm Performance" by Gunter Löffler and Peter N. Posch, August 2007.

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

044 "Ein Vergleich des binären Logit-Modells mit künstlichen neuronalen Netzen zur Insolvenzprognose anhand relativer Bilanzkennzahlen" by Ronald Franken, August 2007.
045 "Promotion Tournaments and Individual Performance Pay" by Anja Schöttner and Veikko Thiele, August 2007.
046 "Estimation with the Nested Logit Model: Specifications and Software Particularities" by Nadja Silberhorn, Yasemin Boztuğ and Lutz Hildebrandt, August 2007.
047 "Risiken infolge von Technologie-Outsourcing?" by Michael Stephan, August 2007.
048 "Sensitivities for Bermudan Options by Regression Methods" by Denis Belomestny, Grigori Milstein and John Schoenmakers, August 2007.
049 "Occupational Choice and the Spirit of Capitalism" by Matthias Doepke and Fabrizio Zilibotti, August 2007.
050 "On the Utility of E-Learning in Statistics" by Wolfgang Härdle, Sigbert Klinke and Uwe Ziegenhagen, August 2007.
051 "Mergers & Acquisitions and Innovation Performance in the Telecommunications Equipment Industry" by Tseveen Gantumur and Andreas Stephan, August 2007.
052 "Capturing Common Components in High-Frequency Financial Time Series: A Multivariate Stochastic Multiplicative Error Model" by Nikolaus Hautsch, September 2007.
053 "World War II, Missing Men, and Out-of-wedlock Childbearing" by Michael Kvasnicka and Dirk Bethmann, September 2007.
054 "The Drivers and Implications of Business Divestiture – An Application and Extension of Prior Findings" by Carolin Decker, September 2007.
055 "Why Managers Hold Shares of Their Firms: An Empirical Analysis" by Ulf von Lilienfeld-Toal and Stefan Ruenzi, September 2007.
056 "Auswirkungen der IFRS-Umstellung auf die Risikoprämie von Unternehmensanleihen - Eine empirische Studie für Deutschland, Österreich und die Schweiz" by Kerstin Kiefer and Philipp Schorn, September 2007.
057 "Conditional Complexity of Compression for Authorship Attribution" by Mikhail B. Malyutov, Chammi I. Wickramasinghe and Sufeng Li, September 2007.

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".