Authorship attribution of texts: a review

M.B. Malyutov
Department of Mathematics, Northeastern University, Boston, MA 02115, U.S.A.

Abstract. We study the authorship attribution of documents given some prior stylistic characteristics of the author's writing extracted from a corpus of known works, e.g., authentication of disputed documents or literary works. Although the pioneering paper based on word length histograms appeared at the very end of the nineteenth century, the resolution power of this and other stylometry approaches is yet to be studied both theoretically and on case studies such that additional information can assist finding the correct attribution.

We survey several theoretical approaches including ones approximating the apparently nearly optimal one based on Kolmogorov conditional complexity and some case studies: attributing Shakespeare canon and newly discovered works as well as allegedly M. Twain's newly-discovered works, Federalist papers binary (Madison vs. Hamilton) discrimination using Naive Bayes and other classifiers, and steganography presence testing. The latter topic is complemented by a sketch of an anagrams ambiguity study based on the Shannon cryptography theory.

1. Micro-style analysis

1.1. Introduction

The importance of dactyloscopy (fingerprint) and DNA profiling in forensic and security applications is universally recognized after successful testing of their resolution power and standardization of analyzing tools. Much less popular so far is a similar approach to the attribution of disputed texts based on statistical study of patterns appearing in texts written by professional writers. The best tests and their power are yet to be estimated both theoretically and by intensive statistical examination of stylometric differences between existing canons. If this 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 persons, stylometry will change its status from a *hobby* to a *forensic tool* of comparable importance to those mentioned above. One obstacle for implementing this program is the *evolution and enrichment of styles* during professional careers of writers. Thus *plots of style characters vs. time of production* seem more

relevant tools than constant characters. Rates of change for characters may vary. Also, authors can work in several forms, for instance, prose and verse which may have different statistical properties. Therefore, an appropriate *preprocessing* must be applied to the texts analyzed to avoid heterogeneity of forms in, for example, parts of a dramatic corps. Finally, a reliable stylometry analysis should take into account all available information about a disputed work, say time of its preparation, and thus teams of "classifiers" should consist of specialists in different fields, certainly including literary experts.

Especially appealing are those case studies where the stylometric evidence helps to identify an otherwise unexpected candidate for authorship or deny a popular candidate, if this attribution is confirmed later by 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 recently discovered play "Is he dead" was also attributed to Mark Twain. It would be also interesting to study this play by tools of stylometry surveyed further.

Much more dramatic is the famous Shakespeare controversy with the attribution result so far unavailable. Various stylometry and other tests point to the same person, although much more careful testing is needed. It would be extremely encouraging if credible evidence would prove one day the correctness of the stylometry results in this case study.

1.2. Survey of micro-stylometry tools

The pioneering stylometric study (Mendenhall, 1887, 1901) was based on histograms of word-length distribution of various authors. These papers showed significant difference of these histograms for different languages and also for different authors (Dickens vs. Thackeray) using the same language. The second paper describes the histograms for Shakespeare contemporaries commissioned and supported by A. Hemminway. This study demonstrated a significant difference of Shakespearean histogram from those of all but one contemporaries studied (including the Bacon's), and at the same time it called attention to the practical striking identity of Shakespearean and C. Marlowe's histograms (Marlowe allegedly perished two weeks before the first Shakespearean work was published). The identity was shown by a method close to the contemporary bootstrap. However, Williams (1975) raised some doubts about the validity of the Bacon-Shakespeare divergence of styles, pointing to the lack of homogeneity of the texts that were analyzed (Bacon used different literary forms, which in my opinion only strengthens discrepancy of their styles).

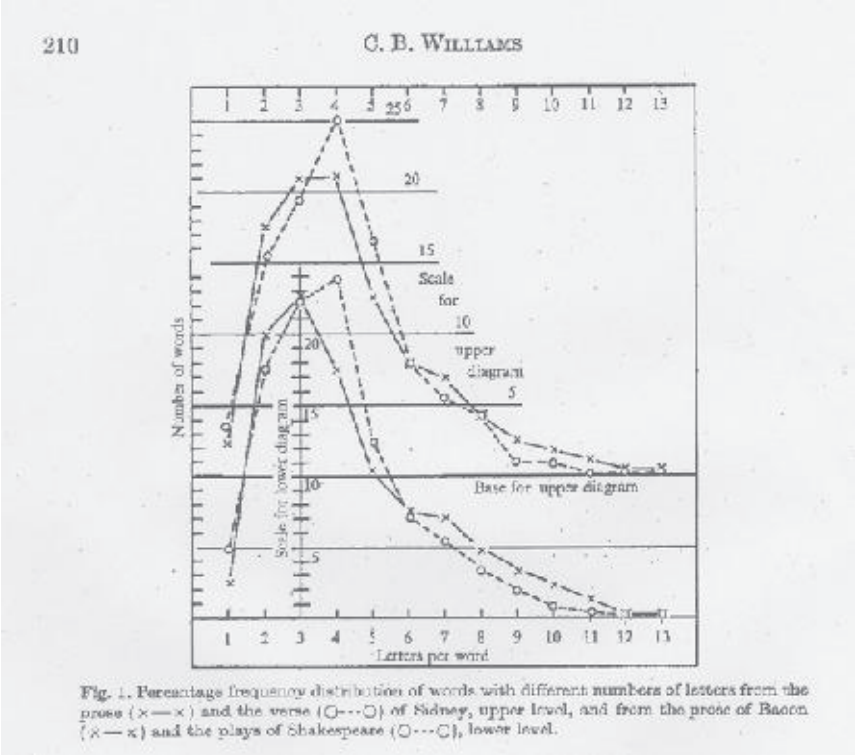


Fig. 1. Percentage frequency distribution of words with different numbers of letters from the prose (×—×) and the verse (○---○) of Sidney, upper level, and from the prose of Bacon (×—×) and the plays of Shakespeare (○---○), lower level.

This objection deserves careful statistical analysis; its cost is now minor (hours vs. months before) because of the availability of software and texts in electronic form. Stability of word-length distribution for a given author also deserves further statistical study.

Ever since T. Mendenhall's pioneering work, word-length histograms have become a powerful tool that has been used to attribute authorship in several case studies including an inconclusive one over a disputed poem (Moore vs. Livingston) controversy, and a successful rejection of Quintus Curtius Snodgrass articles' attribution to M. Twain, as described in Brinegar, 1963.

The frequencies and histograms mentioned above characterize the stationary distribution of words or letters when an author has a large body (canon) of known work. Another popular attribution tool of this kind is a *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* supported by federal funding.

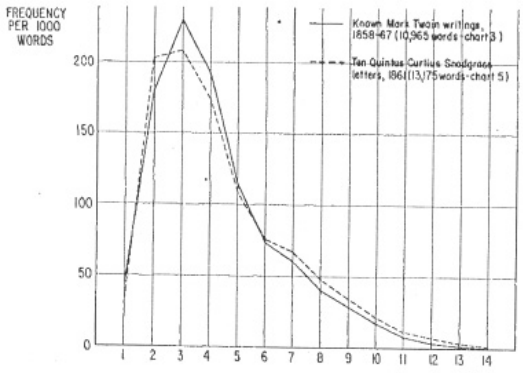


CHART 6. Word frequencies for known Mark Twain Writings and Quintus Curtius Snodgrass Letters.

Histograms of word length in Mark Twain and Quintus Curtius Snodgrass

After fitting appropriate parametric 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 **false**. This premise should be kept in mind when estimating significance of similar studies (see, for example, the attribution study of certain Shakespeare works as a byproduct of cardiac diagnosis software, well-advertised by the Boston Globe on August 5, 2003, or certain Moliere-Corneille controversy studies). The NB-attribution can often be confirmed by other stylometric tests, although the NB-likelihood ratios cannot be taken seriously. The NB-classifier is routinely used also for screening out bulk or junk e-mail messages, see Katirai, 1999, De Vel et all, 2001.

In contrast, Thisted and Efron, 1987, use the *new words usage distribution* in a newly discovered non-attributed anapest poem "Shall I die, shall I fly?", found in the Yale University library, 1985.

I will touch on only one detail in their application of a popular estimation method for the number of unseen biological species (first invented by Turing and Good for breaking the Enigma code), namely neglecting the enrichment of an author's language with time. Thus the distribution of new words in a disputed work *preceding* the canon of an author and that for a text *following* the canon, can be significantly different, for example if Marlowe or Shakespeare wrote the poem. Therefore, this particular application of the Turing-Good method seems

inappropriate. Also, the comparative power of their inference in appropriate cases seems unknown.

More promising and popular now tools use *modeling of long canons as Markov chains of some order composed of English letters and auxiliary symbols*. Given a non-attributed text T and a collection of firmly attributed (to author k) canons T(k) of approximately the same length for training the Markov model of, say, order 1, with transition probabilities $P(k,i,j)$ between symbols i and j, k=1,..., M, the log likelihood of T being written by the k-th author is

$$\sum \log(p(k,i,j))N(i,j) + \log \pi_k(x(1)),$$

where the sum is over all i and j, $N(i,j)$ is the frequency of i followed by j, π_k denotes the stationary probability of the k-th Markov chain, and $x(1)$ is the first symbol in T. Second order Markov chain modeling admits similar expressions for the likelihood. The author with *maximal likelihood* is chosen, which is practically equivalent to *minimizing the cross entropy of empirical and fitted Markov distributions* and to *minimizing the prediction error probability of a next symbol given the preceding text* (Rosenfeld, 1994, 1996, Zhao, 1999), see also Khmelev, 2000, who considers his work as an extension of the classical approach of A. Markov, 1913, 1916. Markov, 1913, introduces the Markov modeling of language. Markov, 1916, rejects an earlier less satisfactory approach to the authorship attribution of Morozov, 1915. *The power of this inference can be approximated theoretically for large sizes of canons T(k) and T under rather natural conditions of asymptotic behavior of their sizes* (Kharin and Kostevich, personal communication). Some regularization of small transition frequencies is worthwhile.

In a canon apparently written jointly by several authors (say, *Edward III*), a Hidden Markov modeling is more appropriate.

Even better attribution performance in certain tests is shown in Kukushkina et al, 2001, by the now very popular *conditional complexity of compression (CCC) minimizing* classifier discussed also by Cilibrasi and Vitanyi, 2003, available from the web-site of the first author. Asymptotic results on CCC are surveyed in Kaltchenko, 2004. The CCC approximates a more abstract *Kolmogorov conditional complexity* concept, which may appear theoretically the best authorship classifying tool based on microstyle. The CCC measures how good compressor adapts to patterns in the training text for better compressing the disputed text.

Let us define *concatenated texts* $C(k) = T(k)T$ as texts starting with $T(k)$ and proceeding to T without stop, and corresponding compressed texts $T'(k)$ and $C'(k)$. Define the conditional compressing complexity (CCC) to be the difference between the lengths of compressed texts

$|C'(k)| - |T'(k)|$ and choose the author with minimal CCC. Certainly, this definition depends on the compressor used. In the tests described in Kukushkina et al, 2001, the best attributing performance was shown to be that of the compressor rar under Windows.

A defficiency of their work is lack of variability study of CCC for different parts of the disputed text. We show in section 2.3.2 an example of this methodology.

A comparable performance is shown by some ad hoc classification methods such as Support Vector Machines, (see Bosh and Smith, 1998, Burges, 1998). These methods are based on sets of characters chosen ad hoc and not unified between different applications which does not permit a valid comparison.

I skip also any discussion of methods based on grammar parsing since these methods are yet not fully automated. Also, their application for classifying very old texts, such as those written by Shakespearean contemporaries, seems doubtful.

2. Shakespeare controversy

2.1. Introduction

Controversy concerning authorship of the works traditionally attributed to W. Shakespeare dates back several centuries. A **bibliography** of material relevant to the controversy that was compiled by 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. Resolving the controversy would certainly aid our understanding of what the author intended to convey in his works and thus would 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. The goal of this part of our rather personal overview is to stimulate further research by scholars with diverse areas of expertise in order to resolve the Shakespeare authorship mystery. My own contribution is minor and concerns the preliminary MCCC-attribution of the first poem in the canon, existence of certain "watermarking" in the sonnets and plausibility of longer steganography (i.e. hidden cryptography) there. I review in more detail the arguments in favor of only one alternative candidate, whom I personally regard as the most likely one referring to other sources listed in http://www.shakespeareauthorship.org/Forum

If additional incentive to undertake this study is needed, note that the Calvin Hoffman prize, presently worth about one million British pounds, will be awarded to the person who resolves this controversy.

The orthodox side, consisting of those who believe the traditional figure to be the true author of these works or simply of those who find it appropriate to maintain this version, mostly keeps silent about arguments put forth against the authorship of W. Shaxper (W.S.) from Stratford on Avon (*this one of several spellings of the name is used to distinguish the traditional figure from the as yet undecided author of the Shakespeare canon*). When not silent, the orthodox accuse the heretics of being lunatics or snobbish. A collection of their arguments can be found in Matus, 1994.

2.2. Documentary and Literary Arguments

Anti-Stratfordian snobbish lunatics (including to some extent M. Twain, S. Freud, Ch. Chaplin, Ch. Dickens, B. Disraeli, J. Galsworthy, V. Nabokov, W. Whitman, R. Emerson, J. Joyce, and H. James: *"divine William is the biggest and most successful fraud ever practiced"*) point out numerous documentary and literary reasons for rejecting or doubting the authorship of W.S.
One early survey of these grave doubts in *several hundred pages* was written by a US presidential hopeful *Donnelly*, 1888. Similar doubts were expressed in many subsequent books including recent ones, see *Mitchell*, 1996, and *Price*, 2001. Scarce documents related to W.S. revealed there allow the following scenario of his career.

His education in Stratford or any literary work there is not documented. A rather ambiguous record about his marriage is kept in the local church. Abandoning Stratford just after the birth of his twins and being warned of severe persecution over next stealing rabbits in the woods of his landlord, he apparently wandered for several years in constant fear of a severe punishment imposed on tramps in the Elizabethan time. Eventually, he was employed in valet horse parking at one of London theaters, later on he was apparently promoted to its security (since he is mentioned in several complaints over his part in assaults against alternative theaters: these were also centers of criminal activities such as gambling, prostitution, etc., there were frequent fights between them which forced the London mayor to transfer them out of City). Being a talented organizer, W.S. has later become an ambitious administrator, producer and shareholder of the theater occasionally performing secondary scenic roles,, and likely also an informer of the ESS (how to explain otherwise that he avoided arraignment after the Essex revolt started by a performance of an allegedly W.S.' play Richard II while all its organizers were executed?). W.S. has probably bought a respect of censors for popular plays to pass smoothly. He used to make around a *thousand pounds a year* for his apparently mostly

undercover activity (compare this to *only a twice larger sum which was paid by Elizabeth to her prime minister W. Cecil!*). He argued fiercely with dramatists for changes in their plays to make them more popular, and he was not sensitive to authors' rights in publications which brought him pennies as compared to his other activities. Thus he apparently cared little if any plays were published under his name. His Last Will clearly shows that he did not keep any printed matters, without mentioning manuscripts. His death was not even noticed by contemporary poets.

There is evidence that W.S. lent money to dramatists for writing plays performed and published under his name and ruthlessly prosecuted those failing to give the money back in time. This is revealed by Mitchell and Price in their discussions of *Groatsworth of Wit* published in 1592 after the death of well-known dramatist R. Green, where apparently W.S. is called *Terence and Batillus* with the obvious meaning of appropriating somebody else's plays. In a recently found manuscript (see
`http://ist-socrates.berkeley.edu/~ahnelson/Roscius.html`)
written during W.S.'s retirement in Stratford prior to 1623 (First Folio) he was called *our humble Roscius* by a local educated Stratfordian author, meaning a famous Roman who profited from special laws allowing him to hawk or sell seats in the theater, and who was not known as an actor/playwright, merely as a businessman who profited on special favor.

As a Russian scholar, I knew several Russian *Terences* in Math Sciences who used their Communist party privileges to produce remarkable lists of publications "borrowed" from others, say persons *condemned as dissidents or enemies of the State*, who were meant to be forgotten in the Soviet Union, and for whom any reference to their work was strictly forbidden. It is not sufficiently remembered that Elizabethan England was an equally closed society with its ruthless censorship and persecution. Well-known, Oscar-winning scenarios written during the McCarthy era in the US by blacklisted authors under false names were milder similar stories.

This concise overview cannot touch on the *hundreds of grave very different questions* raised in the books mentioned above[1]. W.S.'s authorship (WSA) of a substantial part of Shakespeare is hardly compatible with *any* of them and *my subjective log likelihood of WSA* to answer *all of them does not exceed negative 40* (compare with *naive Bayes classifier* discussed before). In my experience as a statistical consultant in forensic cases (especially a disputed paternity) involving

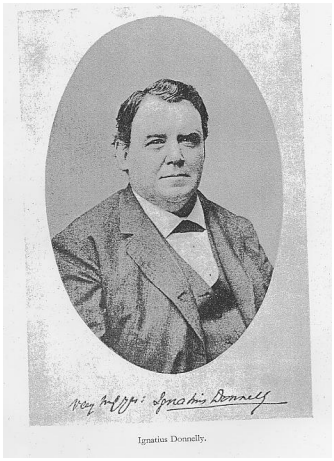
[1] see also a vast recent collection in `http://www2.localaccess.com/marlowe/`

DNA profiling, a much milder mismatch would be sufficient for a court to reject paternity. Some scholars would prefer an explanation of the existing documents not to be based on miracles as holds for WSA. Forensic (in addition to literary) experts must play a decisive role in resolving the controversy as shown further.

The major issues for anti-Stratfordians to resolve are: whose works were published under the Shakespeare name, and why this disguise of authorship happened in the first place and then remained hidden for such a long time.

Francis Bacon became the first candidate for an alternate author, probably because his knowledge of vast areas of culture matched well with that shown in the Shakespeare works.

Francis Bacon



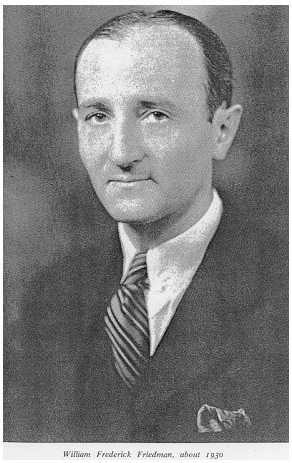
Ignatius Donnelly

The pioneering stylometric study (Mendenhall, 1901) of Shakespeare contemporaries using histograms of their word-length distribution demonstrated the unlikelihood of Bacon's authorship of Shakespeare.

Century-long fruitless mining for cryptography in Shakespeare, allegedly installed there by F. Bacon, and multi-million expenditures for digging the ground in search of the documents proving that F. Bacon wrote Shakespearean works, are brilliantly analyzed in *Friedmans*, 1957. The father of American military cryptography William Friedman and his wife started their careers in cryptography assisting the deceptive (by their opinion) Bacon cryptography discovery in Shakespeare by E. Gallup (which was *officially endorsed by general Cartier, the head of the French military cryptography* those days!). This amusing book, full of historic examples, exercises and humor, should be read by everyone studying cryptography!

Up to now, one of most attractive alternative candidates has been *Edward de Vere, 17th earl of Oxford*. De Vere's life seems by many to be reflected in the sonnets and *Hamlet*. Both de Vere and F. Bacon headed branches of the *English Secret Service* (ESS). De Vere was paid an enormous sum annually by Queen Elizabeth allegedly for heading the *Theater Wing* of the ESS, which was designed in order to prepare plays and actors to serve the propaganda and intelligence collecting aims of the Queen's regime[2]. De Vere's active public support of the corrupt establishment of the official Anglican church in the dramatic Marprelate religious discussions confirms him as one of the principal Elizabethan propaganda chiefs.

Mary Sidney Herbert, countess of Pembroke



William Friedman

Other major candidates for Shakespeare authorship include R. Manners, 5th earl of Rutland, W. Stanley, 6th earl of Derby and several other members of an aristocratic Inner Circle surrounding the Queen and including F. Bacon, Edward de Vere and *Mary Sidney Herbert* (who ran a *literary academy* at her estate in Wiltshire for the University Wits) together with her sons. *Judging by the works that were firmly attributed with reasonable certainty to each of them*, none seems to have been a genius in poetry.

Some from this circle might have been able to produce plots and first versions of plays, but these attempts would need a master in order to be transformed into masterpieces. Some of these people may in fact have done the *editing* work on some of the Shakespeare works (Mary Sidney and her sons). One should also consider that the voluntary hiding of

[2] see www.shakespeareauthorship.org/collaboration.htm referring to Holinshed's chronicles commissioned by W. Cecil, the head of Elizabethan Privy Council.

authorship on any of their parts seems unlikely. Due to the wide extent of the Inner Circle, authorship information would inevitably have become known to everyone. And yet, to the true author of the plays and poems there should have been *dramatic reasons to not claim the works universally recognized as "immortal"*. Note also that the author of the works mastered more than 30,000 English words (as estimated by Efron and Thiested (1975)) compared to about 3000 words used by an average poet. He had also mastered Greek, Latin and several contemporary European languages. In addition, he must have had a profound knowledge of classical literature, philosophy, mythology, geography, diplomacy, court life and legal systems, science, sport, marine terminology and so forth.

The role of paper *Mendenhall, 1901*, may be informally compared with that of a hunting dog. Due to the discovery contained in it, a *famous poet, translator and playwright Christopher Marlowe emerged as one of main candidates*. In an **unprecedented** petition by *Elizabethan Privy Council* Marlowe's important service on behalf of the ESS was acknowledged, and granting him Masters degree by Cambridge University was requested in spite of his frequent long absences (see *Nicholl*, 1992.). *His blank iambic pentameter, developed further in Shakespearean works, remained the principal style of English verse for several centuries*. In 29, ambitious Marlowe was among the most popular London dramatists during his allegedly last 5 years.

Arraigned into custody after T. Kyd's confessions under torture, and let out on bail by his ESS guarantors, Marlowe was allegedly killed by an ESS agent (in the presence of another one responsible for *smuggling agents to the continent*) at their *conventional departure house in Deptford*, owned by a close associate of Elizabeth, almost immediately *after a crucial evidence of Marlowe's heresy* was received by the court, implying an *imminent death sentence*. There is evidence of Marlowe's involvement in the Marprelate affair which made him a *personal enemy of the extremely powerful ruthless archbishop Whitgift of Canterbury*, who did everything possible to expose Marlowe for ages as a heretic and eliminate him (see the well-known *anathema* written by cleric T. Bird, O. Cromwell's teacher, in Nicholl, 1992).

A Marlowe's friend T. Penry, publisher of the Marprelate pamphlets, was hanged previous evening *two miles from Deptford* and *his body has never after been accounted for, in spite of many petitions by Penry's relatives*.

Then, two weeks after Marlowe's supposed demise, the manuscript of the poem *Venus and Adonis*, which had been anonymously submitted to a publisher some months before, was amended with a dedication to the earl of Southampton that listed for the *very first time* the name of

W. Shakespeare as author (any link between the earl and W.S. seems unlikely, Marlowe was likely the earl's tutor in Cambridge).

There are numerous documentary and literary reasons to believe that Marlowe's death was faked by his ESS chiefs (who expected further outstanding service from him) in exchange for his obligation to live and work forever after under alternate names. These arguments are shown on the intelligent and informative web-site
`http://www2.prestel.co.uk/rey/` of a popular Shakespearean actor and former top manager of British Airlines, P. Farey. One of them is *obvious* : spending the whole last day in Deptford (apparently, awaiting the companion), Marlowe defied a strict regulation of daily reporting to the court, hence *he knew beforehand that he would never come back* under his name.

Farey also reviews extracts from the sonnets and other works of Shakespeare hinting at their authorship by Marlowe *after the Deptford affair*. He gives the results of various stylometric tests, showing that the *micro-styles of Marlowe and Shakespeare are either identical, or else the latter's style is a natural development and enrichment of the former*. The micro-style *fingerprint* would give strong evidence for Marlowe's authorship of Shakespearean work, if further comprehensive study confirms that their style patterns are within the natural evolutionary bounds while other contemporary writers deviate significantly in style [3].

Some scholars believe that the ingenious propaganda chiefs of the ESS partly inspired and paid for the production of C. Marlowe, and perhaps of some other politically unreliable dramatists, and directed this production using the Shakespeare pipeline to avoid problems with scrupulous censorship proceedings and also convert dissidents into a kind of unnamed slaves.

During O. Cromwell puritan revolt in forties-fifties of 17th century all theaters were closed, many intelligence documents were either lost or burnt, and the revival of interest to the Shakespearean creative work came only in 18th century making the authorship attribution problematic.

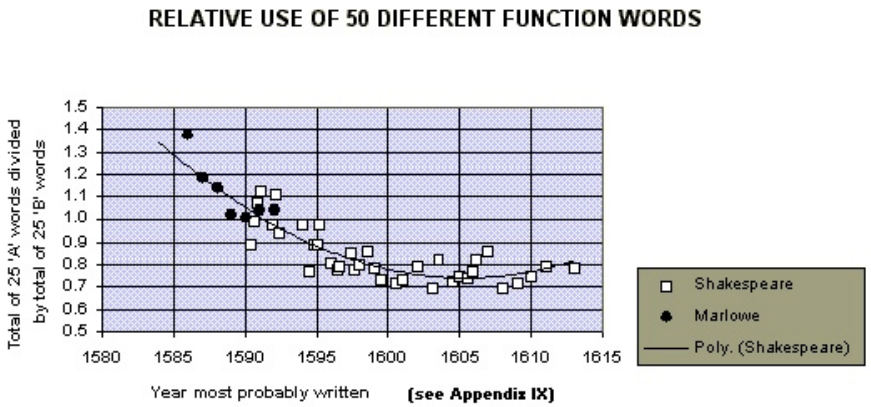
[3] Imagine James Bond let out on bail and reported killed by his colleague soon after a DNA test proved his unauthorized crime. Will you believe in his death if his DNA was repeatedly found later on his victims?

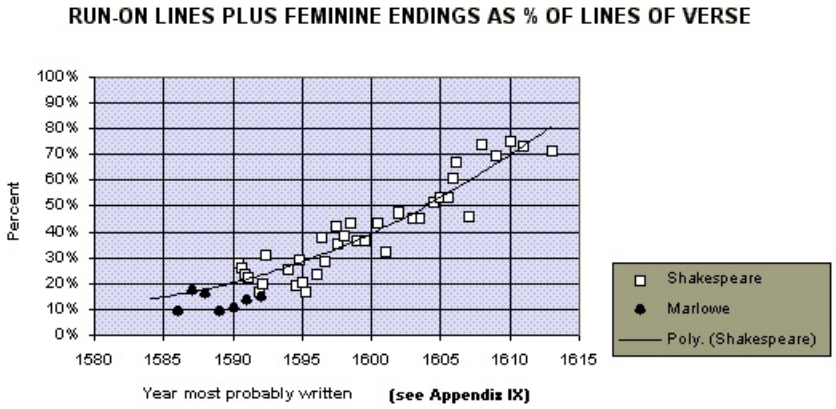
2.3. Micro-Style Analysis

2.3.1. *Introduction*

The stylometric tables in the section Stylometrics and Parallelisms, Chapter *Deception in Deptford*, found on the Farey's website include convincing tables of word-length usage frequencies, including those made by T. Mendenhall, as well as of *function words, feminine endings, run-on lines*, etc., in both Marlowe and Shakespeare as functions of presumable dates of writing corresponding texts.

Again, more careful statistical study of these and more powerful micro-style tests described in section 1 is desirable. Farey's plots clearly show the evolution of styles, which has not been taken into account (or even denied) in many previous studies. For example, some Russian linguists have claimed that the proportion of function words is constant inside the canon during the whole writer's life. This claim was used by them to reject Sholokhov's authorship of the first parts of his Nobel prize-winning novel. This controversy is described in Solzhenitsin, 1974, its study using stylometric methodology outlined in section 2.3.2, is forthcoming.





2.3.2. *Preliminary MCCC-attribution of 'Venus and Adonis'*
My undergraduate student Sufeng Li made the following study under my supervision. She downloaded the following versions of the poems from the internet:

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 (Shall):
http://www.shaksper.net/archives/1997/0390.html

These versions with corrected spelling errors in original versions which were produced by different publishers, were recommended to me by British linguist Peter Bull. The poems were preprocessed according to the rules formulated in Kukushkina et al, 2001 : all words with capital letters and all punctuation signs were removed, new line characters were replaced with space, unless following or preceding a space. After that the poems studied were partitioned into several approximately equal parts of size S. These parts are large enough for averaging the compression complexities (CC) and, at the same time, they are tiny as compared with the training text 'Amores' Thus the self-adapting of a compressor on the disputed texts seems negligible as compared with the adapting to the patterns in the training text.

The quantity of main interest is the Relative Conditional Compression Complexity ($RCCC = CCC/(S)$) tabulated in the last column of the tables for the BWT compressor from
http://www.dogma.net/markn/articles/bwt/bwt.htm.

Compression of VA trained on Amores

N	Precomp	Postcomp	CCC	orig. size	RCCC
1	91,008	33,029	999	2,722	.3670
2	91,010	33,061	1,031	2,724	.3785
3	91,007	33,020	990	2,721	.3638
4	91,004	33,046	1,016	2,718	.3738
5	91,005	33,081	1,051	2,719	.3865
6	91,014	33,035	1,005	2,728	.3684
7	91,005	33,058	1,028	2,719	.3781
8	90,997	33,061	1,031	2,711	.3803
9	91,004	33,046	1,016	2,718	.3738
10	90,995	33,058	1,028	2,709	.3795
11	90,996	33,042	1,012	2,710	.3734

Compression of Hero1 trained on Amores

N	Precomp	Postcomp	CCC	orig. size	RCCC
1	90,986	33,145	1,115	2,700	.4130
2	90,991	33,141	1,111	2,705	.4107
3	90,995	33,157	1,127	2,709	.4160
4	90,990	33,161	1,131	2,704	.4183
5	90,993	33,125	1,095	2,707	.4045
6	90,993	33,172	1,142	2,707	.4219
7	90,990	33,103	1,073	2,704	.3968
8	90,994	33,139	1,109	2,708	.4095
9	90,988	33,174	1,144	2,702	.4234
10	90,996	33,156	1,126	2,710	.4155

The average $R\bar{C}CCs$ are: $R\bar{C}CC(Venus) = 0.375$ with $StD = 0.0066$, $R\bar{C}CC(Hero1) = 0.413$ with $StD = 0.008$. Hence the Kit's translation of Amores helps compressing 'Venus' (**allegedly written by another person**) significantly better than *his own 'Hero' written allegedly at the same time with similar mythological contents and form*! **This seems to be a grave paradox unless the official attribution**

of 'Venus' is flawed! Significance of the difference between $\bar{R}CCCs$ can be shown by both Wilcoxon and two-sample T-tests.

The same partitions of the poems were used to estimate their unconditional $\bar{R}CC$ giving $R\bar{C}C(Venus) = 0.528$ with $StD = 0.0068$ and $R\bar{C}C(Hero1) = 0.528$ with $StD = 0.0065$, which makes the greater $R\bar{C}CC(Hero1)$ even more striking given the practical equality of their unconditional compression complexities. More greedy preprocessing leads to the same conclusion.

A notably smaller $RCC(Amores) = 0.35$ can be explained by its evaluation for the whole huge text without partitioning, so the compressor managed to self-adapt to its patterns.

A smoother version 'Hero2' (easier for reading by a contemporary reader) has smaller $R\bar{C}CC$. Training on both 'Amores' and 'Venus' reduces $R\bar{C}CC(Hero1)$; $R\bar{C}CC(Hero2) < R\bar{C}CC(Shall) < R\bar{C}CC(Hero1)$, when trained on 'Amores'. These results make the Kit's authorship of both 'Venus and Adonis' and 'Shall I die, shall I fly?' very likely.

CCC-test showed similar results under different compressors. The information loss minimization study by choosing appropriate number of partitions and more greedy preprocessing was made and will be reported elsewhere in detail. Unfortunately, CCC-test is extremely sensitive to spelling errors. Thus the *final verdict on attributing 'Venus and Adonis' can be made only after the consensus is reached on the choice of versions analyzed.*

An exciting textual analysis of spectacularly popular at its time erotic poem 'Venus and Adonis' pointing out to its Kit's authorship is made by bishop J. Baker in his essay posted on his site `http://www2.localaccess.com/marlowe/`. Kit cites Venus and Adonis several times in the introduction to his 'Hero and Leander'.

2.4. Macro-Style Analysis

An interesting controversial comparative study of Shakespeare's and Marlowe's *macro-styles*[4] exists on the web-site of late Alfred Barkov

`http://www.geocities.com/shakesp_marlowe/`

Barkov's analysis of the *inner controversies* in Marlowe's and Shakespeare works including *Hamlet*, well-known for a long time, enables him to claim that the texts were intentionally used to encode the story in such a way that the authors' actual messages remain misunderstood by laymen while being understandable to advanced attentive readers. Barkov calls this style *menippea*, considering it similar to the *satira menippea*, a style found in many classical works and discussed by promi-

[4] Namely, semantics and a sophisticated architecture of their works and well-known ambiguity of many statements inside them.

nent Russian philosopher M.Bakhtin (1984). Menippeas often appear in closed societies, since authors tend to use Aesopian language to express their views. This language was very characteristic for Marlowe: he used his poetic genius to provoke Elizabethan enemies by his ambiguous statements to expose their views for subsequent reporting to the ESS (see *Nicholl*, 1992). Similar language was used by Ben Jonson, Kit's ESS colleague and editor of the first folio, in his statements about the canon and its author.

Barkov's analysis of the inner controversies in Hamlet is parallel to the independent analysis of other authors. For instance, the well-known contemporary novelist publishing under the nickname *B. Akunin*, presented recently his version of Hamlet in Russian (available in the Internet via the search inside the web-library www.lib.ru) with a point of view rather similar to that of Barkov, including the sinister decisive role played by *Horatio*.

2.5. Cryptpgraphy Mining

In November 2002, a Florida linguist, R. Ballantine, sent me her decipherment of allegedly *Marlowe's anagrams*, (i.e. sensible permutations of letters) in consecutive bi-lines (that is, pairs of lines) of most of Shakespeare and also of some other works, revealing the author's amazing life story as a *master English spy both in Britain and overseas* up to 1621. Her stunning overview with commentaries based also on her previous 20 years of documentary studies is almost 200 pages long. Her novels covering Marlowe's life until the Deptford affair are more than thousand pages long. I was challenged to make a judgment about the validity of her findings, which stimulated my interest in the topic.

Roberta Ballantine

Irrespective of the authenticity of the historic information conveyed in her overview, the story is so compelling that it might become a *hit of the century* if supplied with dialogues and elaboration and published as a fiction novel by a master story teller (see several chapters of her unpublished novels and anagram examples on the web-site:
`http://www.geocities.com/chr_marlowe/`

Barkov claims that Ballantine's *deciphered anagram texts follow the menippea macro-style of Marlowe's works*. If established as true, this story will constitute a bridge between golden periods of poetry and theater in the South-Western Europe and Britain because in it C. Marlowe is revealed as a close friend of such leading late Renaissance figures as M. Cervantes and C. Monteverdi, as well as the main rival in love and theater of Lope de Vega.

It is almost unbelievable that the author of Shakespearean works could pursue additional goals while writing such magnificent poetry. However, caution is needed: Thompson and Padover, 1963, p. 253, claim that Greek authors of tragedies used to anagram their names and time of writing in the first lines of their tragedies (a kind of *water marking*), which Marlowe could well learn from his teachers in the King's school, Canterbury, and University of Cambridge; a similar tradition was shared by Armenian ancient writers as a protection against plagiarism of copyists, as described in Abramyan, 1974. Also, *first announcing discoveries by anagrams was very popular in those times (Galileo, Huygens, Kepler, Newton among other prominent authors)*; anagrams were certainly used by professional spies.

Attempting to establish cryptographic content in Shakespeare after the discouraging book *Friedmans*, 1957, is very ambitious. Moreover, serious doubts remain concerning the appropriateness of anagrams as a hidden communication (or *steganography*) tool, as will be discussed further.

It is natural to consider two stages in the analysis of the validity of deciphered anagrams. The first question to address is the *existence of anagrams* in the texts. This we have attempted to test statistically starting from our observation that all the anagrams deciphered in Shakespeare contain various forms of Marlowe's signature at the beginning.

R. Ballantine has considered bi-lines as suitable periods for anagramming case-insensitive letters. After deciphering an initial bi-line, she proceeds to the very next one, and so on, until the final signature. In a given play, the first bi-line that begins an anagramming is usually at the beginning of a dialogue, or after a special, but otherwise meaningless sign, a number of which appear in early editions of Shakespearean works.

Following Thompson and Padover, 1963, we mine for Marlowe's signature in the first bi-lines of sonnets, which makes for an easier test, since a disastrous multiplicity-of-decisions problem is avoided in this way. Besides, 154 sonnets, with only a small part of them deciphered so far, constitute a homogeneous sample of 14 lines (7 bi-lines) each (with a single exception). Hence we chose to focus on the sonnets for statistical testing of the presence of anagrams leaving aside almost all other Shakespearean works, which allegedly also contain anagrams.

An important requirement is a careful choice of an accurate published version which has varied over time. I was fortunate to find help from an expert in the field, Dr. D. Khmelev, University of Toronto, who was previously involved in a joint Shakespeare-Marlowe stylometry study with certain British linguists.

For a given bi-line b, let us introduce the event $M = \{b$ contains the set of case-insensitive letters M,A,R,L,O,W,E $\}$ (*event M is equivalent for this name to be a part of an anagram*) Using a specially written code, Khmelev showed (by my request):

THEOREM 1. *The numbers of first, second, etc. bi-lines in the sonnets for which event M occurs are respectively 111, 112, 88, 98, 97, 101, 102 out of 154 sonnets.*

Our first corollary follows:

THEOREM 2. *Let us test the null hypothesis of homogeneity: event M has the same probability for all consecutive bi-lines in sonnets versus the alternative that the first bi-line contains this set of letters more often than subsequent ones. It is also assumed that these events for all bi-lines are independent. Then the P-value of the null hypothesis (i.e. the probability of the frequency deviation to be as large or more under the null hypothesis) is less than four per cent.*

Proof. We apply a standard two-sample test for equality of probabilities based on the normalized difference between frequencies $f_i, i = 1, 2$, of containing the case-insensitive set of letters 'm', 'a', 'r', 'l', 'o', 'w', 'e' inside the first and all other bi-lines respectively which has approximately standard normal distribution for such a big sample; f_1 is near 72.1 per cent, f_2 is almost 65 per cent. Thus the approximate normalized difference of frequencies $(f_1 - f_2)/\sqrt{\bar{f}(1-\bar{f})(1+1/6)/154}$ is around 1.78, where $\bar{f} := (f_1 + 6f_2)/7$, and *the normal approximation to the binomial probability of this or larger deviation (P-value) is near 3.75 per cent which is a rather unlikely event.*[5]

[5] a more detailed study of numbers in theorem 1 ignoring the multiplicity of hypotheses shows that the case insensitive set 'marlowe' is located anomalously often

Apparently, this anomaly in homogeneity of bi-lines signals that the first bi-lines were specially designed to include this set of letters as part of an anagram signature. Note that signatures may vary over sonnets. Thus our estimate is an upper bound for the P-value of bi-lines homogeneity versus several variants of Marlowe's signature in the first bi-line.

Thus, the *existence of anagrams hidden by Marlowe in Shakespeare* looks rather likely.

Of course, *other explanations of this statistical anomaly* might also be possible. To deal with this possibility, I applied to a recognized expert in statistics on Shakespeare and on English verses in general who is with the University of Washington. Unfortunately, she turned out to be a Stratfordian, and so she chose not to reply at all.

A much more difficult task is to study the *authenticity (or uniqueness) of the anagrams deciphered by R. Ballantine*. This is due to a *notorious ambiguity of anagrams* which seems to be overlooked by those who have used anagrams to claim priority, see above. An amazing example of this ambiguity is shown on pp. 110-111, Friedmans, 1957, namely: **3100** different meaningful lines-anagrams in Latin exist for the salutation "Ave Maria, gratia plena, Dominus tecum". These are referred to a book published in 1711.

A *theory of anagram ambiguity* can be developed along the lines of the famous approach to cryptography given in C. Shannon's *Communication Theory of Secrecy Systems* written in 1946 and declassified in 1949. An *English text is modeled in it as a stationary ergodic sequence of letters* with its entropy per letter characterizing the uncertainty of predicting the next letter given a long preceding text. The binary entropy of English turns out to be around 1.1 (depending on the author and style), estimated as a result of long experimentation.

Shannon showed that this value of the entropy implies the existence of around $2^{1.1N}$ meaningful English texts of large length N. Due to the ergodicity of long texts, the frequencies of all letters in all typical long messages are about the same, and so all typical texts could be viewed as almost anagrams of each other. Thus, the *number of anagrams to a given text seems to grow with the same exponential rate as the number of English texts*. We can prove this plausible **conjecture** in a more artificial approximation of English text as an i.i.d. multinomial sequence of symbols. Let us first further simplify the setting for transparency:

---

in *the first two bi-lines* of the sonnets (**the homogeneity P-value is around 0.2 per cent**). Another popular (according to Ballantine) signature 'Kit M.' turns out to be found unusually often (the homogeneity P-value is around 5 per cent) in the *last two bi-lines* concluding the sonnets.

THEOREM 3. *Consider an i.i.d. three-nomial N-sequence of three letters A, B and C with rational probabilities $p(A) = L(A)/N, p(B) = L(B)/N$ such that $Np(A) := L(A)$ and $Np(B) := L(B)$ are integers. Our claim is: Number $N(A, B)$ of N-sequences with $L(A)$ letters A and $L(B)$ letters B satisfies:*

$$\log N(A, B)/N = H(A, B) = -[p(A) \log p(A) + \ldots + p(C) \log p(C)](1 + o(1)).$$

Proof follows immediately from the method of types (see e.g. Cover and Thomas, 1991). The fraction above is asymptotically the number of typical N-sequences as we stated above.

A generalization to a general multinomial case without the condition of all probabilities being multiples of $1/N$ is straightforward. A generalization to a model of stationary ergodic source can be formulated and proved using the techniques also developed in Cover and Thomas, 1991, say their *sandwich* argument, in proving the equipartition theorem.

Thus the number of meaningful English anagrams for n bi-lines is the n-th power of that for a single bi-line, if deciphering is independent for subsequent bi-lines, and also *exponential* in the length of text. This is a discouraging result for considering anagrams as a communication tool beyond other disadvantages of anagrams, namely excessive complexity of encoding and decoding. Moreover, the aim of putative anagrams that would become known to an addressee only after the long process of publication is unclear, unless an ESS editor would pass it directly to an addressee. Again a parallelism: many of *M. Bulgakov's menippeas* with hidden anti-Soviet content were prepared for publication by an active informer of the Stalin secret police!

There still remains hope that R. Ballantine's claim about the uniqueness of the anagrams she deciphered may prove correct due to the following reasons:

Every one of her deciphered anagrams starts with one of the variants of Marlowe's signature, which restricts the remaining space on the first bi-line, and makes the combination of remaining letters *atypical*, thereby narrowing the set of meaningful anagrams. Furthermore, the names and topics conveyed by Marlowe in the hidden text, may be familiar to his intended receiver (say, the earl of Southampton or M. Sidney Herbert with her sons), who might decipher the anagrams using a type of Bayesian inference, looking for familiar names and getting rid of possible anagrams that did not make sense for him/her. Existence of other keys unknown to us is also possible. It should also be noted that the hidden sentence on the first bi-line is usually continued on the next bi-line (run-on line) giving the decipherer additional information as to how to start deciphering the next bi-line, and so forth. Surely, these arguments are rather shaky. Only a **costly experimentation**

in deciphering anagrams by specially prepared experts can lead to sound results about the authenticity of anagrams deciphered from these texts. Various types of software are available to ease the deciphering of anagrams, although it is questionable if any of them is suitable for these archaic texts.

In summary, the anagram problem in Shakespeare remains unresolved, although I regard it as worthy of further study.

C. Shannon himself developed an important theory for breaking codes. His *Unicity theory* specifies the minimal length of encoded messages that admit a unique decoding of a hidden message by a code-breaker due to the redundancy of English. Unfortunately, *his main assumption of the key and message independence, crucial for his results about unicity in cryptography, is obviously not valid for anagrams*, which use special keys for each bi-line depending on the combination of letters in the bi-line.

Our statistical result on the special structure of the first bi-lines shows that the encoding (if it took place at all!) had to be iterative: if the poetic bi-line was not suitable for placing Marlowe's anagram-signature there, the line and hidden message were to be revised in order to make the enciphering possible. This is exactly a situation where *knowledge of an incredible number of English words, demonstrated by Shakespeare, could have been put to perfect use permitting flexibility in the choice of a relevant revised text*!

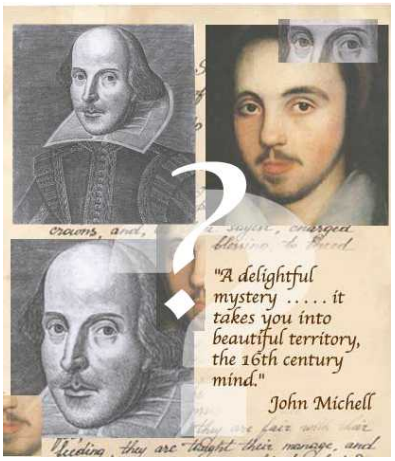
2.6. Possibility of Genetic Evidence

It turns out that the *critical argument against Marlowe's authorship of Shakespeare* is the inquest by Queen Elizabeth's personal coroner found in 1935 (made in violation of several instructions) stating that Marlowe was killed on May 30, 1593. The question of the validity of this inquest is discussed by *Farey* and *Nicholl*, 1992 in detail. If the inquest was faked and C. Marlowe's survival for several more years is proved, then his authorship of Shakespearean works becomes very likely: Marlowe could have written these masterpieces with abundant features to be ascribed to him, and he had more than enough reasons to hide under a fictitious name.

One long-shot way to prove Marlowe's survival is as follows. A mysterious posthumous mask is kept in Darmstadt, Germany, ascribed to Shakespeare by two reasons: The Encyclopaedia Britannica states that it matches perfectly the known portraits of the bard (which are likely actually versions of Marlowe's portraits as shown brilliantly, say, on the title page of the web-site of a recent award-winning documentary film *Much ado about something*. A second reason is the following: this

mask was sold to its penultimate owner-collectioner together with a posthumous portrait of apparently the same dead man in laurels lying in his bed.

The mask contains 16 hairs that presumably belonged to the portrayed person. A specialist from the University of Oxford has claimed in a personal letter to me his ability to extract mitochondrial DNA from these hairs and match it with that from the bones of mothers or siblings of the candidates. As is well-known, mtDNA is inherited strictly from maternal side since sperm does not contain mitochondria. This study is in the planning stage, and serious legal, bureaucratic, financial and experimental obstacles must first be overcome before the study can proceed.



A fragment of the title page of the web-site www.muchadoaboutsomething.com

The posthumous mask ascribed to Shakespeare

3. Conclusion

The problem of Shakespeare authorship is old, and the documents are scarce. Therefore, only a statistical approach, e.g., comparing the likelihoods of hypotheses based on the *fusion of all kinds of evidence*, seems feasible in trying to resolve it.

An explosion in computing power, emergence and development of new methods of investigation and their fusion let me believe that in this framework the Shakespeare controversy will eventually be resolved with sufficient conviction in spite of the four-century long history of puzzles and conspiracies.

The methods that are now developing are promising and could also very well apply in other similar problems of authorship attribution, some of which might even have significant security applications.

4. Acknowledgements

This study was proposed to the very obliged author by R. Ahlswede at the beginning of the author's two-month stay with the program *Information transfer* at ZIF, University of Bielefeld. The author thanks many colleagues (especially R. Ballantine!) for their helpful remarks, E. Haroutunian for citing Abramyan, P. Farey for the permission to show his plots. I am extremely grateful to E. Gover, D. Massey and I. Malioutov for considerable improvement of my English and style and to D. Malioutov for TEX-nical help.

References

Abramyan, A. *The Armenian Cryptography* (in Armenian), Yerevan University Press, 1974.

Bakhtin, M. *Problemy poetiki Dostoevskogo*. English translation, University of Minnesota Press, 1984.

Bosch, R. and J. Smith, J. Separating hyperplanes and the authorship of the disputed Federalist Papers, *American Mathematical Monthly*, **105(7)**:601-608, 1998.

Brinegar, C. Mark Twain and the Quintus Curtis Snodgrass Letters: A statistical test of authorship, *Journal of American Statistical Association*, **58(301)**, 85-96, 1963.

Cilibrasi, R. and Vitanyi, P., *Clustering by compression*, CWI manuscript, 2003, submitted.

Burges, C. A Tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, **2**, no. 2, pp. 955-974, 1998.

Cover, T. and Thomas, J., *Elements of Information theory*, Wiley, N.Y., 1991.

I. Donnelly, *The great cryptogram*, **1**, 1888, reprinted by Bell and Howell, Cleveland, 1969.

Efron, B. and Thisted, R. Estimating the number of unseen species; How many words did Shakespeare know? *Biometrika*, **63**, 435-437, 1975.

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

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

Kaltchenko, A. Algorithms for estimating Information Distance with. Appl. to Bioinformatics and Linguistics.
http://arxiv.org./abs/csCC/0404039/, 2004

Katirai, H. Filtering junk e-mail, 1999, see his web-site
`http://members.rogers.com/hoomank/`

Khmelev, D. and Tweedy, F.J. Using Markov Chains for Identification of Writers, *Literary and Linguistic Computing*, **16**, No. 4, p. 299-307, 2001.

Kukushkina, O., Polikarpov, A. and Khmelev, D. Text Authorship attribution using letter and grammatical information, *Problems of Information Transmission*, **37(2)**, 172-184, 2001.

Markov, A. On application of statistical method, *Comptes Rendus of Imper. Academy of Sciences*, Ser. VI, **X**, 1913, p. 153; 1916, p. 239.

Matus, I. *Shakespeare, in fact*, Continuum, N.Y., 1994.

Mendenhall, T. A. The characteristic curves of composition. *Science*, **11**, 237-249, 1887.

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

Mitchell, J. *Who wrote Shakespeare*, Thames and Hudson Ltd., London, 1996.

Mosteller, F. and Wallace, D. *Inference and Disputed Authorship*, Addison -Wesley, Reading, 1964.

Ch. Nicholl. *The Reckoning*, second edition, Chicago University Press, 1992.

Price, D. *Shakespeare's Unorthodox Biography*, Greenwood Press, London, 2001.

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

Solzhenitsyn, A.I. *Stremya Tikhogo Dona, puzzles of the novel*. YMCA Press, 1974 (In Russian).

Thompson, J.W. and Padover, S.K. *Secret diplomacy; espionage and cryptography, 1500-1815*, F. Ungar Pub. Co., N.Y, 1963.

De Vel, O., Anderson, A., Corney, M. and Mohay, G. Multi-Topic E-mail Authorship Attribution Forensics. *Proc. Workshop on Data Mining for Security Applications, 8th ACM Conference on Computer Security (CCS'2001)*, 2001.

Williams, C. Word-length distribution in the works of Shakespeare and Bacon, *Biometrika*, **62**, 207-212, 1975. Diederich,

Zhao, J. *The Impact of Cross-Entropy on Language Modeling*. PhD thesis, Mississippi State University, 1999.
`http://www.isip.msstate.edu/publications/courses/ece_7000_speech/lectures/1999/lecture_06/paper/paper_v1.pdf`