A new text statistical measure and its application to stylometry

Felix Golcher
Institut für deutsche Sprache und Linguistik
Humboldt-Universität zu Berlin
felix.golcher@hu-berlin.de

Abstract

We define a simple, purely surface frequency based measure $S(T, t)$ which quantifies the similarity of a training text T with a test text t. S can be decomposed into three factors: one depending on the training text, one depending on the test text, and one nearly constant residual factor. The slight variations of this near constant allow us to measure stylistic differences between T and t with high accuracy. The defined quantity S is unique among other stylometric measures in that it uses the full frequency information of all substrings in both texts. Its applicability for stylometric classifications was tested in a variety of experiments.

1 Introduction

Stylometry aims at quantifying linguistic style. Style can be understood as the subtle but regular differences between texts which ideally share language, genre and topic, but differ with respect to authorship, the gender of the author or similar parameters.

The present paper develops a new text statistical measure S and uses it successfully for stylometric classification on a variety of data sets and in the framework of four different languages. S quantifies how much of a training text is repeated in a test text. It can be factorised into three independent parts, two of which describe the dependency on test and training text respectively, and one which is nearly constant. Subtle variations of this near constant can be used for precisely measuring the stylistic similarity between test and training text.

S has intriguing theoretical properties and using it for stylometry constitutes a completely new approach to this field. Stylometry usually relies on a vector description of documents focussing on the distribution of function

words while omitting content words (*cf.* Section 2). Contrarily, the computation of S is based on the frequency of all substrings in both the test text and the training text.

The paper is arranged as follows. In Section 2, modern approaches to stylometry are briefly reviewed. Section 3 defines S and motivates this definition with the help of simple examples. Section 4 discusses three empirical investigations. Firstly, Section 4.1 describes an introductory experiment run on data assembled specifically for an authorship attribution contest (Juola, 2004). Secondly, in Section 4.2, experiments from Baroni and Bernardini (2006) are reproduced. The objective here is to classify texts either as original Italian or as translations to Italian. Thirdly, section 4.3 reviews the famous problem of the twelve disputed federalist papers. I conclude with a discussion in Section 5.

2 Previous Work

Stylometry as a scientific discipline is more than a hundred years old. The usual timeline of citations contains the cornerstones Mendenhall (1887), Yule (1938), and Mosteller and Wallace (1964). Today's literature on the subject is abundant. Several reasons for this plethora might exist: firstly, literature in general is abundant today. Secondly, there is a vivid and growing interest in forensic applications of stylometry. Thirdly, the research question of stylometry is more clear cut than in other fields of text classification: the question "who wrote this text?" has, in general, a much more definite answer than a question like "what is the topic of this text?".

In what follows, I present a short overview and a classification of the wealth of ideas existing in stylometry today. This overview shall in no way be exhaustive, but rather capture the main currents in the stream of publications.

Most people in stylometry work on tokenised text. There are however some exceptions to this rule. The authors of (Benedetto, Caglioti and Loreto, 2002a; Baronchelli, Caglioti and Loreto, 2005; Cilibrasi and Vitanyi, 2005) exploit the characteristics of the standard data zipping algorithm LZ77 (Ziv and Lempel, 1978) to measure text similarity. Since this algorithm has no knowledge of tokens or words, their method uses all substrings of the untokenised text[1]. The data basis of Forsyth (1999) is a (small) subset of

[1]Fierce battles were fought about the appropriateness of this method (Goodman, 2002;

the character n-grams of a text. Khmelev and Tweedie (2002) used markov chains on the character level for various problems of authorship attribution.

Nearly all existent approaches use the frequency information contained in their corpora. The only exception I came across is the above mentioned zip community around Benedetto, Caglioti and Loreto (2002a). To compress a file, it is not necessary, to know how often a string repeats in a text. Only the fact of the repetition as such is used by LZ77.

Only a few of the available stylometric methods use syntactic information. Examples for this minority include Stamatatos, Fakotakis and Kokkinakis (1999), and Baayen et al. (1996). For the layman, this might seem odd, since the normal human test person would probably pin down stylistic differences to higher level features of language use than character n-grams and token frequencies. Obviously, the reason for our restraint in using such information is its ongoing invisibility to the computer.

A seemingly unbroken rule of stylometry is the following: everybody who uses n-gram frequencies in her research presets some fixed n. I found no n higher than 10 (Keselj et al., 2003).

A very common research practice is the use of vectors as document representations[2]. The dimensions of these vectors are frequencies of unigrams or n-grams with a small n of word forms or POS tags. Exceptions to this rule are, on the one hand, the above mentioned zip-branch of approaches, and, on the other hand, those researchers who develop measures of vocabulary richness. Algorithms of this flavour compress the complete frequency statistics of a text in one single number (*cf.* Tweedie and Baayen (1998) for an excellent review).

A near must of stylometric investigations is to exclude content words from the start. The reason for this is obvious: the use of content words depends on content, and the content of a text ("topic") is, by definition, not covered by stylometry. There are very few exceptions to this rule besides the generally exceptional zip community. Whenever content bearing words are not excluded from the start, considerable argumentative effort is undertaken to show that the results are nevertheless not dependent on them (Baroni and Bernardini, 2006).

Whatever representation of the data is chosen, a mechanism to classify

---

Benedetto, Caglioti and Loreto, 2002b; Khmelev and Teahan, 2003; Benedetto, Caglioti and Loreto, 2003).

[2]I cannot deny myself the nitpicking remark that maybe no one ever checked the compliance of these representations with the formal definition of a vector.

documents is needed in order to get results. Some examples for this include neural networks (Tweedie, Singh and Holmes, 1996), principal component analysis (PCA) (Burrows, 1987; Baayen et al., 1996; Holmes, Robertson and Paez, 2001), and support vector machines (SVMs) (Diederich et al., 2003; Fung, 2003; Baroni and Bernardini, 2006).

I conclude this section with a more detailed description of research presented by Baroni and Bernardini (2006). It is a very appropriate example for leading research on the field of stylometry and I will repeat its main experiments in Section 4.2.

Baroni and Bernardini (2006) analyse the subtle regular deviations between translations to Italian and texts originally written in Italian. They set up a data set of 813 articles from the high standard Italian geopolitical journal *limes*. 569 of these are original Italian, the remaining 244 are translations to Italian from various languages[3].

The documents are condensed into vectors whose dimensions are frequencies of unigrams, bigrams and trigrams of word forms, lemmata or POS tags. The authors test their classification method on a variety of such vector representations and of combinations thereof. Classification is carried out by means of support vector machines (SVMs, for a description *cf.* Scholkopf and Smola (2002)).

For the task of classifying a test document as translation or original, the authors report a performance considerably higher than the figure they cite for human individuals.

With their work, the authors establish empirical backup for the hypothesis that something like *Translationese* exists, that is, that there are systematic deviations between texts originally written in a language and texts translated into that language (*cf.* Gellerstam (1986) and follow up literature for a discussion).

More details of corpus setup and the experiments conducted by Baroni and Bernardini (2006) will be given in Section 4.2 where main experiments will be repeated.

3 Definition of S

The stylometric method proposed in this paper violates all common sense rules of stylometry as described in the previous section, except the one, not

[3]English, Arabic, French, Spanish, Russian.

to take syntactic information into account.

It makes use of the full frequency information for all substrings in the training and in the test text, i.e. in terms of character n-grams, n is not fixed. Up to my knowledge, these data have never been used for stylometric investigations, presumably because of their extreme size.

Content words are not omitted. No vector representation is involved. No linguistic knowledge whatsoever enters the computation. The method of classification is kept very simple and is mostly limited to the simple comparison of numbers.

In short, the proposed measure S quantifies the extent to which character strings from the training text are repeated in the test text. The formal definition of S might sound a bit abstract. In order not to overshadow the underlying simplicity of the concept with formulas, we present a short example first.

3.1 An example

Let $T =$ `abrakadabra` be the training text. $t =$ `abar` is the test text. We now define – step by step – a numeric index $S(T, t)$ which quantifies the similarity of t and T.

The complete frequency list of all substrings s of T is:

substring s of T	frequency $F_T(s)$ of s in T
`a`	5
`abra, abr, ab, bra, br, b, ra, r`	2
all other	1

Now, for every substring s' of the test text t, we look up its frequency $F_T(s')$ in the training text T:

substring s' from t	frequency $F_T(s')$ of s' in T
`abar, aba, bar, ba, ar`	0
`ab`	2
`a`	5
`b`	2
`a`	5
`r`	2

Here $F_T(s')$ is the frequency of s' in the training corpus T. Note that `a` is counted double, as it occurs twice. The sum of all training frequencies would seem to be a natural choice for measuring the similarity of the texts T and t (after appropriate normalisation). Experimental tests disproved this: for realistic texts, the extremely high frequencies of the very short strings overshadow the lower but meaningful frequencies of the longer strings. In order to make frequencies of different order of magnitude comparable, the summation is not done over the frequencies themselves, but over their logarithms. Furthermore, since the logarithm of 1 is 0, the logarithm is applied to $F_T(s') + 1$ instead of $F_T(s')$, otherwise strings of the test corpus which appear only once in the training corpus would be ignored. Thus we get:

$$S(\texttt{abrakadabra}, \texttt{abar}) = \frac{3\log(2+1) + 2\log(5+1)}{4} = \frac{6.88}{4} = 1.72 \qquad (1)$$

The division by test text length is done in order to get rid of the expectable proportionality of the sum of logarithms on text length.

The table below gives $S(\texttt{abrakadabra}, t)$ for some other test texts t. It gives an impression in which way $S(T, t)$ relates to the intuitive concept of text similarity.

t	$S(T,t)$	
`abarx`	1.38	x does not occur in $T =$ `abrakadabra`, but the length of t is now 5. Thus we get $S(\texttt{abrakadabra}, \texttt{abarx}) = 6.88/5$
`abarxabarx`	1.38	A simple reduplication of t does not change S.
`abra`	3.09	A prefix of T
`xyz`	0	Nothing repeats.
`elanp`	0.36	$S = log(5+1)/5$

3.2 Formal definition

Let T and t be two texts, t having length L. $s'_{m,n}$ with $1 \le m \le n \le L$ is the substring of t running from text position m to n. The *similarity index* $S(T, t)$ is defined as:

$$S(T,t) = \frac{\sum_{m,n} \log(F_T(s'_{m,n}) + 1)}{L} \ , \qquad (2)$$

where $F_T(s'_{m,n})$ is the number of occurrences of $s'_{m,n}$ in T. It follows from this definition that multiple occurrences of the same character string in t are counted separately, as is the case with the substrings $s'_{1,1} = s'_{3,3} = a$ of t in the above example.

In realistic applications, the number of substrings of a text quickly gets very large. We keep this abundance manageable by using the index structure of suffix trees (Gusfield, 1997). To build this elaborate data structure, the algorithm of Ukkonen (Ukkonen, 1995) was implemented in `c++`.

The apparent asymmetry between T and t in Definition 2 stems from this method of handling the immense data pool: first a suffix tree of the training text is built, functioning as a data base of substring frequencies. Afterwards, we run through the test text looking up the frequencies of encountered substrings one by one, without access to earlier substrings. Hence we cannot take logarithms of test file frequencies but only of training file frequencies.

4 Experiments

In the following sections, I will report three experiments exploring the properties of S. The data set investigated in Section 4.1 was set up in the framework on an authorship attribution contest (Juola, 2004). The task of the second experiment (4.2) is the discrimination between original Italian texts and translations to Italian as described in Baroni and Bernardini (2006). Section 4.3 investigates the authorship of the 12 disputed essays of the famous federalist papers.

4.1 Authorship attribution with laboratory data

The data used in this first investigation were set up as the training set of an authorship competition held in 2004 (Juola, 2004). They are very suitable for a showcase demonstration of the method since the solution is always known and the data are highly controlled. The data set is split into 13 different problems – labeled from A to M – simulating different stylometric standard setups.

The first investigation was done with the data of problem M. It consists of 48 training files, containing 6 files from each of a set of 8 authors. The language is Dutch. I concatenated 5 files of author i to be the training corpus T_i and left the 6th as the test file t_i. Figure 1 shows the similarity values $S(T_i, t_j)$ for all combinations of T_i and t_j.

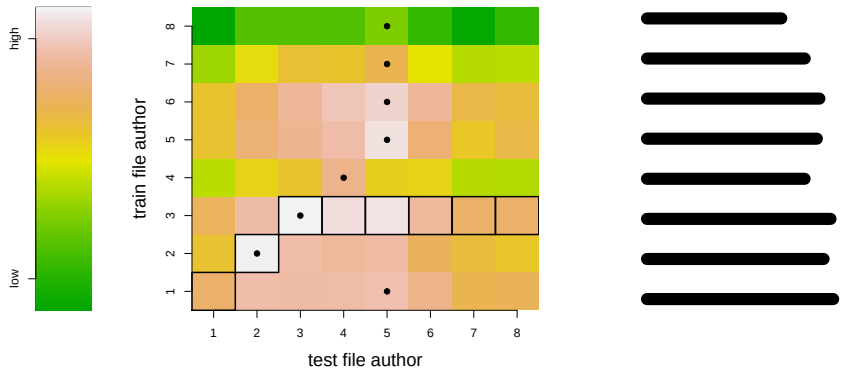

Figure 1: Pseudocolour representation of the matrix $S(T_i, t_j)$ comparing the training file of author i with the test file of author j. The colour code is given as a bar on the left. The matrix values range from 17.2 to 19.5. The boxes indicate maximum values in the vertical direction, while the black dots show horizontal maxima. The training file lengths are visualised by the thick black lines on the right. These range from 20289 characters for file T_8 to 28133 characters for file T_1.

If we assign to each test file t_j the author i of the training file which has the highest value of $S(T_i, t_j)$, we get the assignment visualised by the square boxes. Nearly all test files would be classified as being written by author 3. T_3 is seemingly most similar to 8 out of 10 test files.

This is partly due to differences in training text length. Equation 2 corrects for test file length, but not for training file length. It was designed that way because the sum of logarithms in the numerator of this definition can be expected to be proportional to the text file length while its dependency on training file length is unknown at this point[4]. In any case, $S(T, t)$ should grow monotonously with the length of T.

However, this still does not fully account for the domination of training file T_3, since it is not the longest file, this being T_1. We get a complementary picture if we look at the data the other way round and mark, for each training file T_i, which test file t_j had the highest $S(T_i, t_j)$ relative to it. These maxima are indicated by the black dots in the figure. Again we get the result that one test file (t_5) is more similar to most training files than the other test files.

[4]Recent experiments suggest its proportionality to $(L_T)^b$ with the training file length L_T and $b \approx 1/5$.

The exact reason for the fact that some of the test files are closer to all training files than the other test files, and vice versa, is not yet known. It seems as if some of the texts are somewhat more typical Dutch than the others. Presumably they contain a slightly higher proportion of the already common elements of Dutch.

As a consequence of these observations, authorship attribution is not possible by means of the raw values of $S(T,t)$. It will be necessary to isolate the parts of S which depend only on T and t separately. These parts must be split off in order to filter out the fraction of S which hopefully is suitable for measuring the similarity between both texts.

In order to remove the factor from $S(T_i,t_j)$ which only depends on T_i, we normalise S by averaging over all test files $t_{j'}$ relative to one training file. I define the *test set normalised similarity index* S_{test} as:

$$S_{test}(T_i,t_j) = \frac{S(T_i,t_j)}{\frac{1}{N_t}\sum_{j'=1}^{N_t} S(T_i,t_{j'})} \tag{3}$$

In terms of the matrix visualised in Figure 1, the operation leading from $S(T_i,t_j)$ to $S_{test}(T_i,t_j)$ is equivalent to dividing each row of the matrix by its mean. This is shown in the left panel of Figure 2.

If we now reassign to each test file t_j the author i whose training file T_i has the highest $S_{test}(T_i,t_j)$, we get a much better result, shown again by the square boxes. 7 out of 8 authors are assigned correctly. The black dots which indicate, for the training file T_i, the test file t_j with the highest value of $S_{test}(T_i,t_j)$, did not move, since this maximum is not affected by the averaging procedure.

This changes if we repeat the normalisation procedure, now averaging over columns instead of rows. I define the *fully normalised similarity index* $S_{full}(T_i,t_j)$ as:

$$S_{full}(T_i,t_j) = \frac{S_{test}(T_i,t_j)}{\frac{1}{N_T}\sum_{i'=1}^{N_T} S_{test}(T_{i'},t_j)} \tag{4}$$

N_T is accordingly the number of training files, here coincidentally the same as N_t, the number of test files. This operation corresponds to the division of the columns of the matrix of the left panel in Figure 2. The result is given in its right panel.

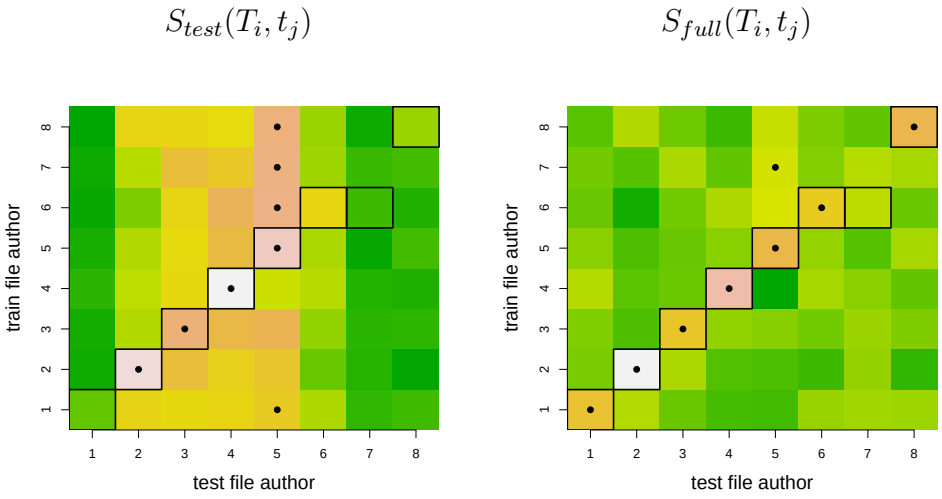

Figure 2: The effect of normalisation. The left part visualises the matrix of Figure 1 with the rows being divided by their means (i.e. the *test set normalised similarity index* $S_{test}(T_i, t_j)$). The right part shows the left matrix with each column divided by its mean (i.e. the *fully normalised similarity index* $S_{full}(T_i, t_j)$).

Now the diagonal – where the author of the training file and of the test file are the same – comes out visibly higher than the rest (except for author 7, where the method fails). Figure 3 explicitly shows the distribution of the 64 values for $S_{full}(T_i, t_j)$.

This empirically motivated decomposition can be formalised as:

$$S(T, t) = A_T B_t (C + \epsilon \delta_{a_T a_t} + f) \tag{5}$$

A_T is only dependent on the training file T and can be identified with the denominator in Equation 3. B_t is a factor only dependent on the test file t and can be identified with the denominator in Equation 4. The third factor $(C + \epsilon \delta_{a_T a_t} + f)$ is equivalent to S_{full}. a_T and a_t are the authors of T and t, and $\delta_{a_T a_t}$ is defined as:

$$\delta_{a_T a_t} = \begin{cases} 1, & \text{if } a_T = a_t \\ 0, & \text{if } a_T \neq a_t \end{cases} \tag{6}$$

Its prefactor ϵ has, in this case, a value close to 0.013. f is a Gaussian noise component. Here, this Gaussian has a standard deviation of about 0.0035. The fact that ϵ is considerably larger than the standard deviation σ_f of f

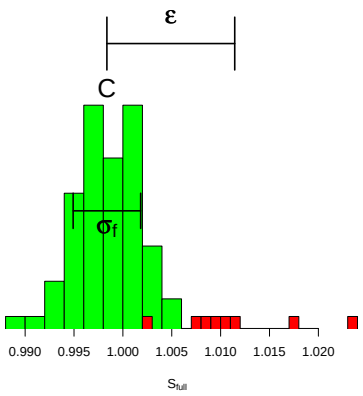

Figure 3: Distribution of the values of the fully normalised matrix as depicted in the right part of Figure 2. Matrix fields, where the author of the training file and the author of the test file are not the same, are depicted in green, otherwise they are red. Note the extremely small fluctuations. Nevertheless, the file pairs where the authors match are clearly separated. The symbols C, ϵ, σ_f are explained in the text.

makes the separation of authors in the data set of problem A possible. C, σ_f, and ϵ are added to Figure 3.

4.1.1 Interference with topic related effects

The method, as described above, leaves the text untouched and does not exclude content words. This inevitably leads to interference with topic based variation if this parameter is not controlled in the corpora used. Up to now, this feature did not surface, since the training documents used so far contain five texts each with different topics, the same topics in each training file T_i. The test files t_j, on the other, hand also share the same topic.

I turn now to problem A of the data of Juola (2004) to demonstrate the topic specific behaviour of S_{full}. For this subset of data, thirteen authors contributed three essays (in American English) about the same three topics. Since author 2 did not come up with a paper concerning topic 1, there are $12 \cdot 3 + 2 = 38$ files, which I label t_{Ai} with $1 \leq i \leq 38$.

Figure 4 compares all files t_{Ai} in the data set of problem A with each other. As can easily be seen, the similarity arising from topic identity is

larger than the similarity resulting from author identity.

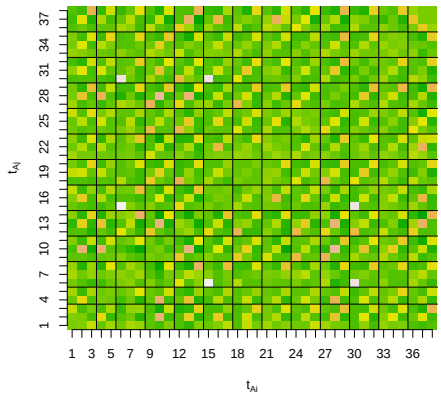

Figure 4: Analysing Problem A of Juola (2004). All files t_{Ai} are compared with each other. The fields show the *fully normalised similarity index* $S_{full}(t_{Ai}, t_{Aj})$. The meaningless diagonal fields $S_{full}(t_{Ai}, t_{Ai})$ are coloured neutrally. The framed fields sort together the comparison of all files of one author to all files of another author. The diagonal fields of these boxes represent comparisons of two files with the same subject.

Figure 5 makes this relation quantitatively visible. As can be seen, the shift in S_{full} resulting from topic identity is about twice as large as the shift stemming from authorship identity. However, we will see in Sections 4.2 and 4.3 that the performance of the method is not disturbed by these effects if the corpus at hand is known to be rather narrow or balanced with respect to topic.

4.1.2 Cross validation

I performed cross validation experiments on the problems A, C, D, F, G, I, K, and M of Juola (2004). The problems B, E, and J were excluded, since they have identical training sets with other (included) problems. I used only the training sets, because I could not get hold of the correct classification of the test files in time. Problems H and L are missing, since they are too small.

In each cross validation run, one file from each category[5] was removed

[5]In all but one cases the categories were authorship by different authors. In one case

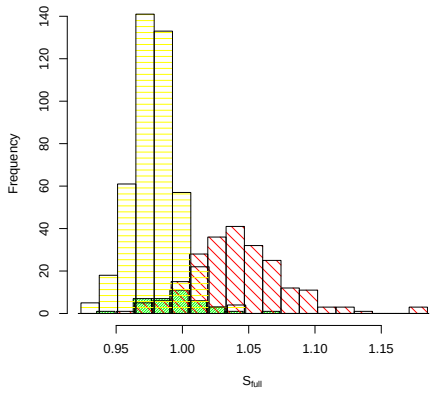

Figure 5: Distribution of the values in the matrix shown in Figure 4. The yellow histogram is comprised of $S_{full}(t_{Ai}, t_{Aj})$ values where the compared files t_{Ai} and t_{Aj} share neither subject nor author. The red histogram shows file pair comparisons where the subjects are identical, but not the author. The green histogram, on the other hand, shows file pairs with the same author, but not the same subject.

from the files of the problem at hand. The removed files were used as test file. The test file t_j was (re)assigned to the category i which had the highest value of $S_{full}(T_i, t_j)$. Results are shown in Table 1.

The catastrophic performance on problem G is due to topic effects similar to the one analysed in Section 4.1.1. In this problem, all files are by the same author, Edgar Rice Burrows. He wrote series, *Tarzan*, to given an example. As a consequence, the thematic similarities of his earlier and later works completely screen possibly existing stylistic differences.

The Results on the other problems look rather encouraging. Unfortunately they are not fully comparable with the ones given for the original competition (Juola, 2006), since the test sets were not used.

4.2 Translationese

As mentioned in Section 2, I repeated the main experiments run by Baroni and Bernardini (2006). The aim of this investigation is to show that – despite its conceptual simplicity – stylometry based on the quantity S is fully competitive with very elaborate modern methods.

(Problem G) the categories were the age of the author.

P	N_c	N_r	g	description
A	13	2	0.65 ± 0.05	American English essays
C	5	2	1	British English Novels
D	3	3	1	English Plays
F	3	20	0.95 ± 0.16	English letters (*c.* 1470)
G	2	3	0	American English novels
I	2	2	1	French novels
K	3	2	0.67	Serbian-Slavonic texts
M	8	6	0.83 ± 0.13	Dutch essays

Table 1: The fraction g of correctly identified documents. P is the label of the problem, N_c is the number of categories and N_r is the number of cross validation runs. Details about the corpus can be found in Juola (2006).

In order to make my results comparable with Baroni and Bernardini (2006), I copied their experimental setup as closely as possible. The data were given to me by the authors of the mentioned paper.

As described in Section 2, the corpus consists of 813 Italian articles, 569 original Italian texts and 244 translations to Italian. I split the corpus into 16 sections, each made of 15 random original documents and 15 random translated documents. This left a remainder of 329 original texts and 4 translated texts. The 30-document sections were used in a series of 16-fold cross-validation experiments; the articles in the remainder were used as part of the training data in each fold, but never as test data. Thus, within each fold, the training set contained 229 translated texts and 554 original texts; the test sets contained 15 translations and 15 originals[6].

On the available computer, with its 1.25 gigabyte of working memory, it is not possible to run the suffix tree programme mentioned in Section 3.2 on more than 8 megabyte of text. This limitation forced a breakup of the corpus of originals into two chunks. Thus we have three corpora: one corpus of translations (T_T) of 5.7 megabyte and two corpora of originals, one (T_{O1}) of approximately 7.6 megabyte and one (T_{O2}) of approximately 5.5 megabyte.

A given test file t is classified as translated if the *fully normalised similarity index* S_{full} was higher relative to the translated corpus T_T than to the average result for the original corpora, i.e. $S_{full}(T_T, t) > (S_{full}(T_{O1}, t) +$

[6]The last paragraph is a close paraphrase of Baroni and Bernardini (2006) with only minor typos on the numbers corrected.

$S_{full}(T_{O2}, t))/2$. This results in a precision of 0.81 ± 0.09, a recall of 0.83 ± 0.11 and an F-score of 0.82 ± 0.09.

These figures can directly be compared with the results of Baroni and Bernardini (2006). The authors cite a precision of 0.893, a recall of 0.833, and, accordingly, an F-score of 0.862. No error margins are given. Thus, the results of Baroni and Bernardini (2006) are included by the error margins of the approach presented in the present paper. Baroni and Bernardini (2006) experimented with a set of twelve different document representations, consisting of unigrams, bigrams and trigrams of word forms, lemmata, and POS tags. Several of these representations were combined to reach the result cited above. The approach proposed in the present paper does not have this flexibility and does not use any linguistic knowledge, i.e. works on untokenised and unlemmatised text. The fact that it can compete so well with Baroni and Bernardini (2006), is quite encouraging.

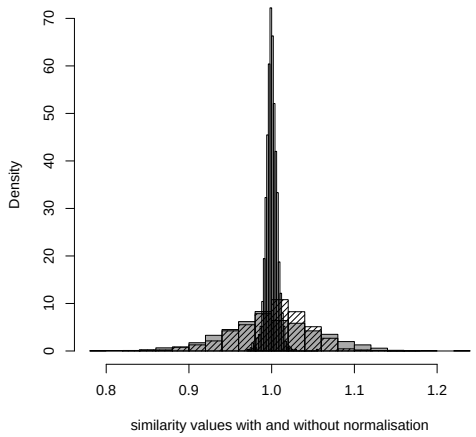

Figure 6: The grey histogram shows the "raw" values of $S(T, t)$ for the Translationese data set from Baroni and Bernardini (2006). Standard deviation is 0.056. The shaded histogram displays the *test set normalised similarity index* $S_{test}(T, t)$. Its standard deviation is only slightly lower at 0.045. The black histogram represents the *fully normalised similarity index* $S_{full}(T, t)$. Its standard deviation is lowered by an order of magnitude to 0.0068. The x-values of the distribution of $S(T, t)$ are divided by its mean in order to make it comparable to the two other distributions.

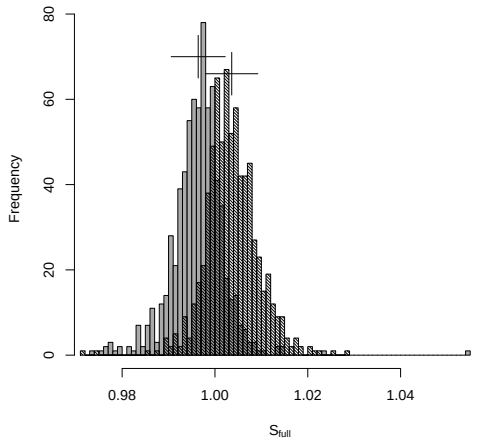

Figure 7: This figure shows the data of the black histogram in Figure 6 enlarged. The shaded histogram counts those instances of $S_{full}(T, t)$ for which T and t both are translations or originals. The other cases are shown in grey. The crosses give the mean and the standard deviation. The overlap of the two distributions is responsible for the roughly twenty percent of misclassifications.

Figures 6 and 7 demonstrate that two opposing properties dominate the behaviour of S_{full}: it is constant at a very high level, while its dependency on style (ϵ in Equation 5) is sufficiently large to differentiate between translations and originals. The observed constancy of S_{full} might be a very interesting property of human language texts, especially if recent work on a similar subject is taken into account (Golcher, to appear). This paper states the suspected constancy of the level of repetitions in human language texts across texts from different languages.

It is very important to make sure that the high quality of the results is not due to topic related effects (*cf.* Section 4.1.1): it is not improbable that, in geopolitic articles translated to Italian, names and places from outside of Italy dominate, while, in originally Italian articles, Italian proper nouns are more frequent. Baroni and Bernadini provide different representations of their data which make it possible to check for such effects.

I used the following versions of the corpus, which, gradual and in different ways, take all content out of the text:

full: The full text.

tok: Proper nouns and numerals are replaced by placeholders to circumvent the problem mentioned above. The placeholders are numbered (i.e. NPR1, NPR2, ...). The same placeholder is used for the same expression within one document, but the counters are reset for each new document.

mix: Baroni and Bernadini (Baroni and Bernardini, 2006) call this the *mixed* representation: "in the mixed representation, function words are left in their inflected wordform, whereas content words are replaced by the corresponding tags"

mix,rand: Like the **mix** representation, but the sequence of tokens (i.e. function words and content word tags) is randomly scrambled.

func: Only function words. The content word tags of the **mix** representation are removed.

tag: Only content word tags. The function words of the **mix** representation are removed.

tag,rand: Like the **tag** representation, but the sequence of tokens (i.e. content word tags) is randomly scrambled.

Figure 8 visualises the resulting F-score for these representations. The seven representations can be grouped into three categories: **full**, **tok**, **mix** and **func** are essentially at the same level of performance, i.e. at 80 percent or slightly above. **tag** is significantly lower, but seemingly above chance level[7]. The randomised representations **mix,rand** and **tag,rand** are at chance level.

These results allow for some conclusions:

- As long as the function words are kept in the text and as long they are kept in their original sequence, performance stays at its maximum.
- If the token ordering is destroyed, performance is also destroyed.

[7]Detailed statistical analysis would be necessary to confirm this thoroughly.

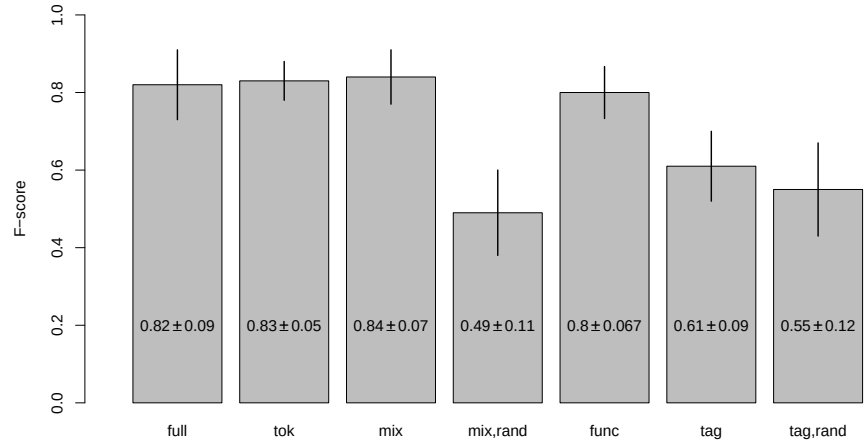

Figure 8: The F-score for the different representations of the *limes* data as described in the text. The vertical lines show the standard deviation over 16 runs of cross validation.

- Using only content word tags still results in some residual discriminatory power. This is maybe a surprising result.

Thus the experiment assessed that no content dependent effects are responsible for the good performance of the presented approach.

In addition to the *limes* data, Marco Baroni provided me with a similar data set, the *giallisti*[8] corpus. It contains 28 texts, of which 10 are translations and 18 original Italian. The texts are written by 26 different authors, two authors contributed two texts each. I used the *limes* corpus as training data and the *giallisti* texts as test files to be classified. Again, the original part of the *limes* corpus had to be split into two files T_{O1} and T_{O2}, due to space limitations. Note that the three training corpora now consist of all 813 *limes* files. As before, a file t was classified as translation if $S_{full}(T_T, t)$ was larger as the mean $(S_{full}(T_{O1}, t) + S_{full}(T_{O2}, t))/2$. Averaging was done over the 28 *giallisti* test files.

As a result, 8 out of 10 translations were classified and 12 out of 18 originals. The *test of equal proportions* was used to assess statistical significance

[8]Ongoing work by Baroni, Bernardini, Castagnoli, Piccioni and Zanchetta; *Giallisti* – authors of crime stories.

(Walpole and Myers, 1972, pp. 261). It results in a significance level of $\alpha \leq 0.01$.

4.3 The federalist papers

In the two investigations described so far in Sections 4.1 and 4.2, the test set always contained the same number of files from each category: in the case of the "laboratory data" in Section 4.1, we had eight test files, one for each author. In the Translationese experiment in Section 4.2, a test set of thirty files was used in each run, fifteen originals and fifteen translations. This gave us the possibility of defining S_{full} in a meaningful way such that the style dependent ϵ of Equation 5 could be isolated.

The averaging procedure must break down, if the "true" categories of the test corpus are not balanced. A look at the left part of Figure 1 makes this obvious: let us suppose that the first impression were correct and all test files, or most of them, had been written by author 3. Then, the normalisation of the matrix rows would level this natural dominance of the third row down and the assignment to authors would be random.

The following investigation explores a strategy to avoid this trap. It addresses the old problem of the federalist papers. These are a series of 85 articles written in 1787/88 in New York. Three authors wrote the articles: Alexander Hamilton, James Madison, and John Jay. For most of the texts, authorship is an established fact: Jay wrote 5 of them, Madison wrote 14 of them alone and 3 in cooperation with Hamilton. Hamilton wrote 51 essays alone. For the remaining 12 articles, authorship is disputed. But, over the years, many researches worked on the problem (an incomplete list is Mosteller and Wallace (1964), Holmes and Forsyth (1995), Fung (2003), Tweedie, Singh and Holmes (1996), and Khmelev and Tweedie (2002)). If nowadays a stylometric method cannot establish the authorship of Madison for the disputed essays, it can be considered seriously flawed.

The training corpora T_H and T_M consist of the texts known to be written by either Hamilton or Madison. The test set consists of the 12 disputed papers. A priori, it is not known if they all were written by Hamilton, all by Madison, or if they are of mixed authorship. Thus the averaging method which produced good results in sections 4.1 and 4.2 is not directly applicable.

Instead, I set up a corpus of 100 pseudo test documents τ_r ($1 \leq r \leq 100$). These where chosen at random from the BNC (Burnard, 2000) and cut down to 35,000 characters each. Now the raw values of $S(T_H, t_i)$ and $S(T_M, t_i)$ for

the 12 disputed texts t_i are gauged by the mean similarity $S(T, \tau_r)$ of the pseudo test files relative to T_H and T_M. I give the set of the 100 pseudo test texts τ_r the collective name τ. Then the τ *gauged similarity index* S_τ is defined as:

$$S_\tau(T, t_i) = \frac{S(T, t_i)}{\sum_{r=1}^{100} S(T, \tau_r)} \tag{7}$$

Behind this gauging operation stands the assumption that the mean *distance* – how ever it might be defined – between modern English texts of random subject and our two training corpora should be independent of the style of Madison or Hamilton. If a test file t_i is compared with the training corpora, we expect $S_\tau(T_H, t_i)$ and $S_\tau(T_M, t_i)$ to both be considerably above 1 since they will both be much closer to this elder form of American English than the BNC files. But, likewise, we expect $S_\tau(T_H, t_i)$ to have a slightly higher value if Hamilton is the author and vice versa.

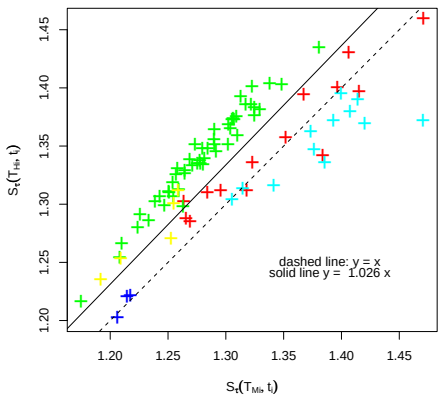

Figure 9: The federalist papers. Each data point represents the comparison of one of the disputed papers θ_i with the two corpora T_{Hi} and T_{Mi}. The figure shows the *BNC gauged* values $S_\tau(T_{Hi}, \theta_i)$ and $S_\tau(T_{Mi}, \theta_i)$. Texts written by Hamilton are coloured green, texts by Madison are red. Jay is yellow, and the three cooperations between Hamilton and Madison are coloured blue. The disputed papers are represented by the cyan data points. The Identity $S_\tau(T_{Hi}, \theta_i) = S_\tau(T_{Mi}, \theta_i)$ is shown by the dashed line, while the solid line separates the texts of Madison and Hamilton (with one remaining exception).

To test this hypothesis, the procedure was as follows. For 70 of the

federalist papers the authorship by either Hamilton or Madison in known. These files are named θ_j ($j \leq 1 \leq 70$), to differentiate them from the 12 disputed essays t_i. Now, each of them is, in turn, removed from the training set and than reattributed to an author. In each of the 70 runs, the reduced training set was rearranged into two files T_{Hj} and T_{Mj}. If Hamilton wrote θ_j, the Madison training set stays unaffected, that is $T_M = T_{Mj}$, while θ_j is missing from T_{Hj}. Otherwise $T_H = T_{Hj}$ and θ_j is missing from T_{Mj}. The τ *gauged similarity indices* $S_\tau(T_{Hj}, \theta_j)$ and $S_\tau(T_{Mj}, \theta_j)$ are compared in order to assign an author to θ_j.

The result can be seen in Figure 9. If the above assumption were correct, all texts by Hamilton, that is the green data points, should lie below the dashed line. They do. On the other hand, the texts by Madison should exclusively lie below the dashed line, and they do not. It seems to be the case that the Hamilton training corpus is more typical for American English of its time than it is typical for modern British English. This plausible assumption could easily introduce the skew in the distribution of S_τ which we can observe in Figure 9. However, the line $S_\tau(T_{Mi}, \theta_i) = 1.026 \cdot S_\tau(T_{Hi}, \theta_i)$ separates the data correctly with only one exception. All disputed files end up on the Madison side of the line. This matches the results reported in Fung (2003), Bosch and Smith (1998) and Mosteller and Wallace (1984).

5 Discussion and outlook

In this paper, a novel text statistical measure $S(T, t)$ is defined which functions as a measure of the similarity between a training text T and a test text t. This quantity can be split into a product of three factors: one (A) dependent on T, another (B) dependent on t, and one being nearly constant. The factors depending on training and test text can be interpreted as measuring the "typicality" of a text, since a high value of A or B is equivalent to a high similarity index $S(T, t)$. That S indeed measures similarity is guaranteed by its definition which is directly and solely based on the surface frequencies of the two texts. Differently from other related measures, S is computed on grounds of the complete frequency list of all substrings of both corpora. Most often, stylometry is done by means of statistics which use only a very small subset of the data actually available, usually the frequencies of function words or POS tags.

The near constant third factor in S can be used to measure stylistic differences. This was tested on two problems of authorship attribution

(4.1 and 4.3) and on the task of separating original Italian documents from translations (4.2). It was shown that the approach of using S for stylometric classifications is fully competitive with state of the art methods of stylometry.

Although this paper shows the unique characteristics of S and establishes a method for using them for stylometric classifications and measurements, a set of intriguing questions still await their answer.

First and foremost, it is of importance for the feasibility of stable applications that the impact of the text topic is exactly known and that it is brought under control.

Furthermore, it could be assessed with more certainty, which elements are responsible for the discriminatory power of S.

As $S(T,t)$ is defined to date, it is asymmetric in the training and test texts T and t. This is due to technical restrictions (see the remarks following Equation 2). In principle, it is easy to come up with symmetric definitions which might have an even greater discriminatory power. It might be a worthwhile task to devise an algorithm capable of circumventing the mentioned restrictions. Furthermore, it could be helpful to find a parametrisation of S which – for example – explicitly incorporates its dependency on training corpus length. Such a parametrisation would possibly make the normalisation of S, that is the computation of S_{test}, S_{full}, and S_{τ}, redundant. Similarly, it would be desirable to have a definition for $S(T,t)$ – and an algorithm to compute it –, such that $S(T,t) = 1$ if $T = t$.

Apart from practical applications, the theoretical implications of the properties of S deserve a closer look. Its very regular behaviour fits well with recent research results concerning the level of repetitions in human text and its suspected constancy (Golcher, to appear). The fact that S uses the complete surface string statistics might be considered a flaw by many. It opens the door for the interference of dimensions of text similarity which are often not wanted to be intermixed with stylometric problems, for example authorship attribution. On the other hand, it could allow for subtle and novel investigations into the statistical properties of human text.

6 Acknowledgements

I thank Anke Lüdeling for supervising my thesis and for her relentless support, Marco Baroni for providing me the invaluable Translationese corpus, Patrick Juola for creating his extremely useful test suite for authorship attribution (Juola, 2004), Anna McNay for the hard work of proof reading, the

R-team for its great statistical software (R Development Core Team, 2006).

References

Baayen, H. et al. (1996): Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, 11, Nr. 3, 121–131.

Baronchelli, Andrea, Emanuele Caglioti and Vittorio Loreto (2005): Artificial sequences and complexity measures. *J. Stat. Mech.*, P04002.

Baroni, Marco and Silvia Bernardini (2006): A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21, Nr. 3, 259–274.

Benedetto, D., E. Caglioti and V. Loreto (2002a): Language Trees and Zipping. *Physical Review Letters*, 88, Nr. 4, 048702.

Benedetto, D., E. Caglioti and V. Loreto (2002b): On J. Goodman's comment to "Language Trees and Zipping". No address in Available on-line from http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0203275.

Benedetto, Dario, Emanuele Caglioti and Vittorio Loreto (2003): Benedetto, Caglioti, and Loreto Reply:. *Phys. Rev. Lett.* 90, Nr. 8, 089804.

Bosch, R. A. and J. A. Smith (1998): Separating hyperplanes and the authorship of the disputed federalist papers. *American Mathematical Monthly*, 105, Nr. 7, 601–608.

Burnard, Lou (2000): The British National Corpus Users Reference Guide. Available on-line from http://www.natcorp.ox.ac.uk/docs/userManual – accessed: 06/30/2007.

Burrows, J. (1987): Word-patterns and Storyshapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, 2, Nr. 2, 61–70.

Cilibrasi, Rudi and Paul M. B. Vitanyi (2005): Clustering by Compression. *IEEE Transactions on Information Theory*, 51, Nr. 4, 1523–1545 Available

on-line from http://homepages.cwi.nl/~paulv/papers/cluster.pdf, corrected version.

Diederich, J. et al. (2003): Authorship attribution with support vector machines. *Applied Intelligence*, 19, Nr. 1–2, 109–123.

Forsyth, R. S. (1999): Stylochronometry with Substrings. *Literary and Linguistic Computing*, 14, Nr. 4, 467–477.

Fung, Glenn (2003): The disputed federalist papers: SVM feature selection via concave minimization. In TAPIA '03: Proceedings of the 2003 conference on Diversity in computing. New York, NY, USA: ACM Press, 42–46.

Gellerstam, M. (1986): Translationese in Swedish Novels Translated from English. In L. Wollin and H. Lindquist, editors: Translation Studies in Scandinavia. Lund: CWK Gleerup, 88–95.

Golcher, Felix (to appear): A stable statistical constant specific for human language texts. In Recent Advances in Natural Language Processing 2007 (RANLP-07)..

Goodman, Joshua (2002): Extended Comment on Language Trees and Zipping. No address in, cond-mat/0202383 Available on-line from http://front.math.ucdavis.edu/0202.0383.

Gusfield, Dan (1997): Algorithms on strings, trees, and sequences: computer science and computational biology. Cambridge University Press.

Holmes, D. and R. Forsyth (1995): The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, 10, Nr. 2, 111–127.

Holmes, D. I., M. Robertson and R. Paez (2001): Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, 35, Nr. 3, 315–331.

Juola, Patrick (2004): Ad-hoc Authorship Attribution Competition. Available on-line from http://www.mathcs.duq.edu/~juola/authorship_contest.html – accessed: 12/06/06.

Juola, Patrick (2006): Questioned Electronic Documents : Empirical Studies in Authorship Attribution. In Olivier and Shenoi, editors: Research Advances in Digital Forensics II. Heidelberg: Springer.

Keselj, Vlado et al. (2003): N-gram-based Author Profiles for Authorship Attribution. In Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03. Halifax, Nova Scotia, Canada, 255–264.

Khmelev, D. V. and F. J. Tweedie (2002): Using markov chains for identification of writers. *Literary and Linguistic Computing*, 16, Nr. 4, 299–307.

Khmelev, Dmitry V. and William J. Teahan (2003): Comment on "Language Trees and Zipping". *Phys. Rev. Lett.* 90, Nr. 8, 089803.

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

Mosteller, Frederick and David L. Wallace (1964): Inference and disputed authorship, the Federalist. Reading, MA: Adison Wesley.

Mosteller, Frederick and David L. Wallace (1984): Applied Bayesian and Classical Inference: The Case of the Federalist Papers. New York, Heidelberg: Springer, Springer series in statistics.

R Development Core Team (2006): R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2006 Available on-line from http://www.R-project.org, ISBN 3-900051-07-0.

Scholkopf, B. and A. J. Smola (2002): Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. The MIT Press.

Stamatatos, E., N. Fakotakis and G. Kokkinakis (1999): Automatic authorship attribution. In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 158–164.

Tweedie, Fiona J. and R. Harald Baayen (1998): How Variable May a Constant Be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, 32, 323–352.

Tweedie, F.J., S. Singh and D.I. Holmes (1996): Neural network applications in stylometry: the *Federalist papers*. *Computers and the Humanities*, 30, 1–10.

Ukkonen, Esko (1995): On-line construction of suffix-trees. *Algorithmica*, 14, Nr. 3, 249–260.

Walpole, Ronald E. and Raymond H. Myers (1972): Probability And Statistics For Engineers And Scientists. New York: Macmillan Publishing Co..

Yule, G.U. (1938): On Sentence Length as a Statistical Characteristic of Style in Prose with Application to Two Cases of Disputed Authorship. *Biometrika*, 30, 363–390.

Ziv, J. and A. Lempel (1978): Compression of individual sequences by variable rate coding. *IEEE Trans. Inform. Theory*, IT-24, 530–536.