Hartmut Ilsemann

Some additions to 'The Marlowe Corpus Revisited' and *Phantom Marlowe*[1]

Introduction

One of the more recent additions to the stylo program package (Eder, Rybicki, Kestemont, 2016) is the General Imposters method (GI), which is also known as the second verification system. Previous investigations, indicated in the headline, had made use of Rolling Delta and Rolling Classify results and together with matching N-grams, adopted from Pervez Rizvi's data base (http://shakespearestext.com/can/index.htm), were able to show that the Marlowe corpus is stylistically inhomogeneous, whereas the style of the two *Tamburlaines* is also contained in a number of other plays of the time. This was summarised in the following figure.

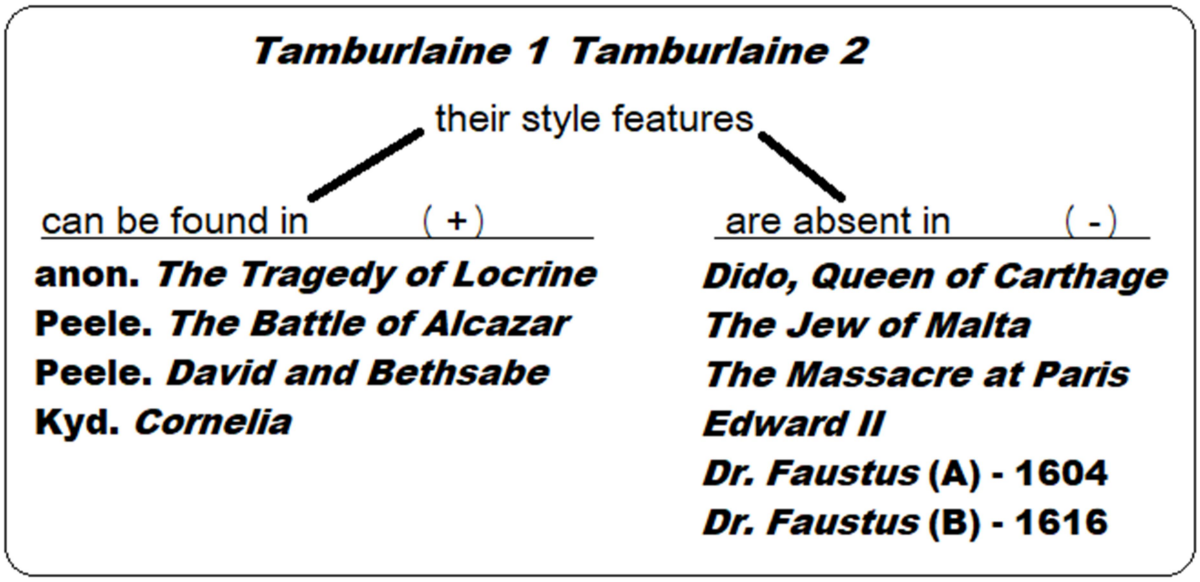


Fig. 1 Stylistic features of the *Tamburlaines*

With the advent of GI an additional check of similarities in the writing style became available. In his post "Authorship verification with the package 'stylo'"[2] Maciej Eder gives a detailed account of the new method, referring to its introduction by Koppel and Winter (2014) and Kestemont's application to the study of Julius Caesar's disputed writings (Kestemont et al., 2016). He also quotes the authors' description of the capacity of the new feature:

---

[1] "The Marlowe Corpus Revisited" was preceded by "Christopher Marlowe: Hype and Hoax", *Digital Scholarship in the Humanities*, Volume 33, Issue 4, 1 December 2018, doi/10.1093/llc/fqy001 and made use of an extended methodological framework to rebut Ros Barber's interim riposte. A more comprehensive investigation was laid down in *Phantom Marlowe* (see References), a monograph, which is currently only available in German.

[2] Maciej Eder "Authorship verification with the package 'stylo'," *Computational Stylistics Group*, 30 May, 2018, https://computationalstylistics.github.io/blog/imposters/, accessed 14.07.2021.

[t]he general intuition behind the GI, is not to assess whether two documents are simply similar in writing style, given a static feature vocabulary, but rather, it aims to assess whether two documents are significantly more similar to one another than other documents, across a variety of stochastically impaired feature spaces (Eder, 2012; Stamatatos, 2006), and compared to random selections of so-called distractor authors (Juola, 2015), also called 'imposters'." (Kestemont et al., 2016a: 88).

Eder then describes the prerequisites necessary to use the function *imposters()*.

It assumes that all the texts to be analysed are already pre-processed and represented in a form of a matrix with frequencies of features (usually words). The function contrasts, in several iterations, a text in question against (1) some texts written by possible candidates to authorship, or the authors that are suspected of being the actual author, and (2) a selection of "imposters", or the authors that could not have written the text to be assessed. Consequently, a given candidate's class is assigned a score between 0 and 1." (Eder, 2018)

Assessments

The reasonable assumption of the procedure is that any result above 0.5 can be seen as a successful verification of authorship. In the case of the Marlowe corpus and associated corpus files the following steps had to be carried out. In a specified folder the eighteen plays to be analysed could be found.

apo_locrine.txt (1); kyd_mscornelia.txt (2); m0_fausta.txt (3); m1_bfaust.txt (4); m2_dido.txt (5); m3_edw2.txt (6); m4_jewmalta.txt (7); m5_massacre.txt (8); mar_tamburlain1.txt (9); mar_tamburlaine2.txt (10); nashe_summer.txt (11); peele_alcazar.txt (12); peele_davbeth.txt (13); shak_12thnight.txt (14); shak_hamlet.txt (15); shak_lear.txt (16); shak_romjul.txt (17); shak_winters.txt (18).

Rather than using the prefix *mar* for all Marlowe plays, the texts were given individual prefixes apart from the *Tamburlaines*. This was necessary to make sure that their real affiliations could be found by the method.

After loading the stylo library and setting the folder the following commands were executed:

```
tokenized.texts = load.corpus.and.parse(files = "all")
features = make.frequency.list(tokenized.texts, head = 2000)
data = make.table.of.frequencies(tokenized.texts, features, relative = TRUE)
imposters(reference.set = data[-c(1),], test = data[1,])
```

The last line would then determine the authorship relations of the first file in the folder (*The Tragedy of Locrine*) to the remaining author texts. To find the affiliations of the second file, Kyd's *Cornelia,* the command changes to:

```
imposters(reference.set = data[-c(2),], test = data[2,])
```

This procedure is then applied to subsequent plays in the list and results in a contingency table (Table 1), where values above 0.5 are accounted for in white letters and a black background. The list of plays appears in a vertical arrangement and author affiliations are reproduced horizontally. In subsequent assessments further distance measures are added so that a prototype command line includes:

```
imposters(reference.set = data[-c(2),], test = data[2,], distance = "wurzburg")
```

and

```
imposters(reference.set = data[-c(2),], test = data[2,], distance = "minmax")
```

Table 1 Verification figures with classic Delta

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Classic Delta**	**anon**	**kyd**	**m0**	**m1**	**m2**	**m3**	**m4**	**m5**	**mar**	**nashe**	**peele**	**shak**
2	anon_locrine		0.21	0.00	0.08	0.10	0.25	0.00	0.40	**0.84**	0.00	0.10	0.05
3	kyd_mscornelia	0.61		0.00	0.00	0.19	0.45	0.00	0.12	**0.62**	0.01	0.09	0.02
4	m0_fausta	0.00	0.00		**1.00**	0.01	0.14	0.47	0.18	0.00	0.01	0.00	0.33
5	m1_bfaust	0.00	0.00	**1.00**		0.02	0.29	0.31	0.01	0.00	0.07	0.00	0.31
6	m2_dido	0.00	0.00	0.00	0.25		**0.96**	0.16	0.02	0.23	0.08	0.08	**0.50**
7	m3_edw2	0.00	0.00	0.00	0.03	0.03		**0.52**	**0.96**	0.01	0.00	0.00	0.40
8	m4_jewmalta	0.00	0.00	0.00	0.04	0.00	**0.72**		0.05	0.00	0.01	0.00	**0.77**
9	m5_massacre	0.02	0.00	0.02	0.05	0.01	**1.00**	0.38		0.07	0.00	0.04	0.36
10	mar_tamburlain1	0.11	0.01	0.00	0.00	0.01	0.28	0.00	0.00	**1.00**	0.00	0.35	0.02
11	mar_tamburlaine2	0.12	0.00	0.00	0.02	0.06	0.19	0.01	0.10	**1.00**	0.00	0.45	0.00
12	nashe_summer	0.00	0.00	0.00	0.02	0.01	0.00	0.06	0.00	0.00	**1.00**	0.00	**0.51**
13	peele_alcazar	0.08	0.02	0.00	0.00	0.00	0.14	0.00	**0.66**	**0.82**	0.00	0.01	0.04
14	peele_davbeth	0.10	0.00	0.00	0.02	0.17	0.36	0.00	0.04	**0.99**	0.00	0.00	0.01
15	shak_12thnight	0.00	0.00	0.00	0.01	0.00	0.00	0.03	0.01	0.00	0.01	0.00	**1.00**
16	shak_hamlet	0.00	0.00	0.00	0.01	0.00	0.02	0.02	0.00	0.00	0.01	0.00	**1.00**
17	shak_lear	0.00	0.00	0.00	0.00	0.00	0.06	0.01	0.00	0.00	0.01	0.00	**1.00**
18	shak_romjul	0.00	0.00	0.00	0.03	0.00	0.03	0.01	0.00	0.00	0.02	0.00	**1.00**
19	shak_winters	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	**1.00**

In all three tables there are several empty cells with a light grey background (B2-I9). They are arranged diagonally in the contingency table and are treated as single plays awaiting attributions. If we look into the *mar* column (J, fourth column from the right) containing the two *Tamburlaines* (J10,11) the stylistic associates are *The Tragedy of Locrine* (J2), Kyd's *Cornelia* (J3), Peele's *The Battle of Alcazar* (J13) and Peele's *David and Bethsabe* (J14) as already indicated in Figure 1. The 1604 A Text of *Dr Faustus* (E4) is awarded identity with the 1616 B Text (D5). The remaining nominal Marlowe corpus shows an intricate system of interrelations in the framed rectangle D4 – I9. It may be a good start to refer to the prevailing identification of *The Jew of Malta* (0.77) with Shakespeare (M8). But *The Jew* is also stylistically related to *Edward II* (0,72) (G8) which has relations to *The Massacre at Paris* (1,00) (G9) and to *Dido, Queen of Carthage* (0,96) (G6), thus confirming a large degree of stylistic similarity. But this does not come from the two *Tamburlaines*), as the dark grey cells underneath and to the right of the rectangle show. Likewise *Dido* (J+K6), supposedly written by Marlowe and Nashe, does not confirm either of them. Instead Delta records strong

Shakespeare references (M6), and Thomas Nashe's play *Summers Last Will and Testament*, being the only play by Nashe in the list, was in prior tests even given the value 1,00, which results from an impaired classification process that occurs when other fitting candidates are missing in the corpus.[3]

At this point one should also mention that the results were achieved with the Delta method, and Eder promised in his post that SVM, NSC, kNN and NaiveBayes would be provided in the next version. In the present version, it is, however, possible to combine Delta with a number of distance measures.

> […] in their paper introducing the imposters method (Kestemont et al., 2016b), the authors argue that the Ruzicka metrics (aka Minmax) outperforms other measures. Similarly, the Wurzburg guys (Evert et al., 2017) show that the Cosine Delta metrics does really well when compared to other distances. It's true that my implementation of the `imposters()` invokes Classic Delta by default, but other measures can be used as well. (Eder, 2018)

The following table is based on adding `distance = "wurzburg"` in the imposters command.

Table 2 Verification figures with Cosine Delta metrics

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Würzburg Distance**	**anon**	**kyd**	**m0**	**m1**	**m2**	**m3**	**m4**	**m5**	**mar**	**nashe**	**peele**	**shak**
2	anon_locrine		**0.76**	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.47	0.00
3	kyd_mscornelia	**0.71**		0.00	0.00	0.02	0.02	0.00	0.00	**0.54**	0.00	0.36	0.00
4	m0_fausta	0.00	0.00		**1.00**	0.03	0.01	0.40	0.13	0.00	0.00	0.02	0.10
5	m1_bfaust	0.03	0.00	**1.00**		0.05	0.08	0.36	0.09	0.00	0.00	0.03	0.02
6	m2_dido	0.16	0.17	0.00	0.00		0.14	0.04	0.00	0.29	0.00	0.44	0.00
7	m3_edw2	0.02	0.02	0.00	0.02	0.12		**0.52**	**1.00**	0.03	0.00	0.13	0.00
8	m4_jewmalta	0.00	0.00	0.13	0.06	0.01	0.39		0.13	0.00	0.00	0.00	**0.50**
9	m5_massacre	0.06	0.02	0.01	0.01	0.00	**0.95**	0.18		0.04	0.00	0.45	0.01
10	mar_tamburlain1	0.13	0.14	0.00	0.00	0.01	0.01	0.00	0.00	**1.00**	0.00	0.45	0.00
11	mar_tamburlaine2	0.14	0.10	0.00	0.00	0.05	0.00	0.00	0.00	**1.00**	0.00	0.47	0.00
12	nashe_summer	0.00	0.01	0.01	0.02	0.01	0.00	0.02	0.00	0.00	**1.00**	0.00	0.30
13	peele_alcazar	0.19	0.07	0.00	0.00	0.00	0.00	0.00	0.07	**0.97**	0.00	0.16	0.00
14	peele_davbeth	0.11	0.06	0.00	0.00	0.02	0.00	0.00	0.01	**0.99**	0.00	0.18	0.00
15	shak_12thnight	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	**1.00**
16	shak_hamlet	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.01	0.00	0.00	0.00	**1.00**
17	shak_lear	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	**1.00**
18	shak_romjul	0.00	0.00	0.01	0.00	0.00	0.01	0.03	0.00	0.00	0.00	0.00	**1.00**
19	shak_winters	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.01	0.00	**1.00**

[3] To overcome a corpus problem like this, Nashe's play text was doubled to yield a Nashe reference, and the doublet was later removed. The fact that a strong Shakespeare reference remained, probably has to do with the stylistic variety of Shakespeare's plays in the corpus build.

In the *mar* column the two Peele plays *The Battle of Alcazar* (J13) and *David and Bethsabe* (J14) are clearly linked to the two *Tamburlaines*, and the same applies to Kyd's closet play *Cornelia* (J3). This is also interrelated with *Locrine* in C2 and B3 which finds its expression in values like 0.76 and 0.71. Shakespeare is once again given influence in *The Jew of Malta* (M8). The remaining nominal Marlowe corpus in the rectangle consists of a number of (often weaker) interrelations, already indicated in the classic Delta verification. Thus the two *Faustus* texts refer to each other (E4, D5), and *Edward II* has close links to *The Jew of Malta* (H7) and *The Massacre at Paris* (H7).

The following table is based on adding `distance = "minmax"` in the imposters command.

Table 3 Verification figures with Ruzicka metrics

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Ruzicka metrics**	**anon**	**kyd**	**m0**	**m1**	**m2**	**m3**	**m4**	**m5**	**mar**	**nashe**	**peele**	**shak**
2	anon_locrine		0.04	0.00	0.02	0.14	0.09	0.00	0.03	**0.98**	0.00	0.07	0.09
3	kyd_mscornelia	**0.56**		0.00	0.00	0.09	0.03	0.00	0.02	**0.88**	0.06	0.01	0.02
4	m0_fausta	0.00	0.00		**1.00**	0.00	0.03	**0.53**	0.00	0.00	0.02	0.00	0.29
5	m1_bfaust	0.00	0.00	**1.00**		0.00	0.18	0.40	0.00	0.00	0.01	0.00	0.33
6	m2_dido	0.03	0.00	0.01	0.16		**0.50**	0.05	0.00	0.03	0.00	0.00	**0.87**
7	m3_edw2	0.00	0.00	0.00	0.21	0.02		0.33	0.45	0.00	0.00	0.00	**0.72**
8	m4_jewmalta	0.00	0.00	0.25	0.26	0.00	0.34		0.00	0.00	0.00	0.00	**0.88**
9	m5_massacre	0.07	0.00	0.02	0.23	0.00	**1.00**	0.14		0.10	0.00	0.00	**0.53**
10	mar_tamburlain1	0.29	0.03	0.00	0.00	0.03	0.10	0.00	0.01	**1.00**	0.00	0.41	0.01
11	mar_tamburlaine2	0.45	0.00	0.00	0.00	0.02	0.06	0.00	0.01	**1.00**	0.00	0.35	0.00
12	nashe_summer	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	**0.63**
13	peele_alcazar	0.19	0.04	0.00	0.00	0.00	0.03	0.00	0.00	**1.00**	0.00	0.22	0.01
14	peele_davbeth	0.22	0.01	0.00	0.00	0.01	0.03	0.00	0.01	**0.98**	0.00	0.18	0.00
15	shak_12thnight	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	**1.00**
16	shak_hamlet	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.00	0.00	0.00	0.00	**1.00**
17	shak_lear	0.00	0.00	0.00	0.00	0.00	0.06	0.03	0.00	0.00	0.00	0.00	**1.00**
18	shak_romjul	0.00	0.00	0.01	0.03	0.04	0.00	0.05	0.00	0.00	0.00	0.00	**1.00**
19	shak_winters	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.00	0.00	0.00	0.00	**1.00**

The *mar* column (J) once again confirms the stylistic equivalent of the two *Tamburlaines* (J10, J11) with *Locrine* (J2) and Kyd's *Cornelia* (J3) as well as Peele's *Battle of Alcazar* (1.00) (J13) and *David and Bethsabe* (0.98) (J14). If we look at the remaining Marlowe corpus from m0 to m5 within the framed rectangle the stylistic identity of the two *Faustus* texts (E7, F8) becomes obvious next to other interrelations, but the decisive information is that the plays in the dark grey cells underneath and to the right of the rectangle have no stylistic identification with the two *Tamburlaines.* The coordinates M6 down to M9 demonstrate a noticeable Shakespeare participation in *Dido, Queen of Carthage, Edward II, The Jew of Malta* and *The Massacre at Paris*.

Summary and Evaluation

The findings of classic Delta, Cosine Delta and Ruzicka metrics are all based on word frequencies. When Jack Grieve evaluated Burrows's Delta in 2007, he came to the conclusion that character trigrams as variables provided clearer results which was most likely due to a higher number of available variables. Here the number of features is 2000, but nevertheless more tests were undertaken with frequencies of character trigrams and word bigrams, resulting in six more tables in the kind of Table 1 to Table 3. Rather than replicating these space-consuming tables evaluation tables were compiled presenting the results as a survey.

Table 4 Evaluation Table of nominal and real Marlowe references

	A									
	Marlowe	mf1w			mf3c			mf2w		
	classic delta	**δ**	**wu**	**Ru**	**δ**	**wu**	**Ru**	**δ**	**wu**	**Ru**
2	anon_locrine	**0.84**	0.38	**0.98**	**1.00**	**0.82**	**1.00**	**0.93**	0.31	**1.00**
3	kyd_mscornelia	**0.62**	**0.54**	**0.88**	**0.88**	0.42	0.48	**0.65**	0.36	**0.83**
10	mar_tamburlain1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**
11	mar_tamburlaine2	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
13	peele_alcazar	**0.82**	**0.97**	**1.00**	**1.00**	**1.00**	**1.00**	**0.71**	0.46	**1.00**
14	peele_davbeth	**0.99**	**0.99**	**0.98**	**1.00**	**0.69**	**0.94**	**0.82**	**0.62**	**0.97**
4	m0_fausta	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	m1_bfaust	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
6	m2_dido	0.23	0.29	0.03	0.00	0.08	0.00	**0.61**	0.32	0.02
7	m3_edw2	0.01	0.03	0.00	0.00	0.00	0.00	0.07	0.03	0.00
8	m4_jewmalta	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	m5_massacre	0.07	0.04	0.10	0.01	0.06	0.00	**0.78**	0.30	0.15

The lines in Table 4 carry the same numbers as in tables 1 to 3 starting with *Locrine* and *Cornelia*, the *Tamburlaines* and Peele's *The Battle of Alcazar* and *David and Bethsabe*. Word frequencies (left columns) are followed by frequencies of character trigrams (middle columns) and word bigrams (right columns). Each of the bundles contains the values of classic delta results (δ), and the results of Würzburg distances (wu) and Ružička metrics (Ru). The overriding majority of measurements in lines 2, 3, 13, and 14 give a clear indication of stylistic identity with Marlowe's main work *Tamburlaine 1* and *2* (lines 10, 11). But equally clear are the figures for the nominal Marlowe corpus (lines 4 to 9) where only two out of 54 measurements have figures above 0.50.

Another interesting result can be seen as we come to the Shakespeare stylistics summarised in Table 5.

Table 5 Evaluation Table of Shakespeare references

	A									
	Shakespeare	words			mf3c			mf2w		
	classic delta	δ	wu	Ru	δ	wu	Ru	δ	wu	Ru
4	m0_fausta	0.33	0.10	0.29	**0.51**	0.15	0.30	0.38	0.15	0.26
5	m1_bfaust	0.31	0.02	0.33	**0.52**	0.07	**0.52**	0.49	0.10	**0.54**
6	m2_dido	**0.50**	0.00	**0.87**	**0.99**	0.10	**1.00**	**0.79**	0.02	**0.95**
7	m3_edw2	0.40	0.00	**0.72**	**1.00**	0.14	**1.00**	**0.65**	0.00	**0.83**
8	m4_jewmalta	**0.77**	**0.50**	**0.88**	**1.00**	**0.95**	**0.99**	**0.96**	**0.50**	**0.91**
9	m5_massacre	0.36	0.01	**0.53**	0.46	0.00	**0.62**	0.05	0.00	0.35
15	shak_12thnight	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
16	shak_hamlet	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
17	shak_lear	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
18	shak_romjul	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**0.99**
19	shak_winters	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Whereas lines 4 to 9 recall the evaluations of the nominal Marlowe corpus, lines 15 to 19 return the results of some core Shakespeare plays. Out of 45 measurements only two do not quite reach the top value of 1.00. The interesting part is certainly Shakespeare's apparent involvement in the nominal Marlowe plays which were already characterised in their Rolling Delta and Rolling Classify results by the absence of stylistic features of the *Tamburlaines.* These plays must be seen as collaborations, and one of the contributors in varying degrees must have been William Shakespeare. His participation stands out in line 8 (*The Jew of Malta*) where horizontally all variables and their evaluations in the subdivisions give a unanimous verdict.

Cross-validations

Last but not least it makes sense to use cross validations, which in all classification processes have the ability to assess the quality of the trained model.[4] It is based on a number of swaps between the training and the testing set, and in a later version has the advantage that the same list of plays used with GI can be taken over. In an abridged form the following commands, which denote the parameters, were applied:

```
library(stylo)
setwd()
```

After calling up the stylo library the main folder is determined.

[4] Comp. Macij Eder. „Cross-validation using the function classify()," *Computation Stylistics Group*, 13 December, 2017, https://computationalstylistics.github.io/blog/cross-validation/, accessed 14.07.2021.

```
texts = load.corpus.and.parse(files = "all", features = "c", ngram.size = 3, corpus.dir = "imp")
```

This command loads and parses the files in the subfolder (here: imp) and character trigrams were chosen for their reliability as opposed to simple word tests.

```
freq.list = make.frequency.list(texts, head = 1000)
word.frequencies = make.table.of.frequencies(corpus = texts, features = freq.list)
results = crossv(training.set = word.frequencies, cv.mode = "leaveoneout", classification.method = "svm")
results$y
```

The result command summarises some valuable information about the classifications carried out previously. The classifier used was *svm*, having a much higher decision level than any of the other classifiers. As could be expected the first eight plays that, due to their author prefix, were characterised as single plays, have been equipped with the non-numeric value (NaN).

Table 6 Cross-validation results of referenced plays

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
0	**mf3c results**				apo	kyd	m0	m1	m2	m3	m4	m5	mar	nashe	peele	shak
1	apo_locrine.txt	NaN	1	apo	0	0	0	0	0	0	0	0	1	0	0	0
2	kyd_mscornelia.txt	NaN	2	kyd	0	0	0	0	0	0	0	0	1	0	0	0
3	m0_fausta.txt	NaN	3	m0	0	0	0	1	0	0	0	0	0	0	0	0
4	m1_bfaust.txt	NaN	4	m1	0	0	1	0	0	0	0	0	0	0	0	0
5	m2_dido.txt	NaN	5	m2	0	0	0	0	0	0	0	0	0	0	0	1
6	m3_edw2.txt	NaN	6	m3	0	0	0	0	0	0	0	0	0	0	0	1
7	m4_jewmalta.txt	NaN	7	m4	0	0	0	0	0	0	0	0	0	0	0	1
8	m5_massacre.txt	NaN	8	m5	0	0	0	0	0	0	0	0	0	0	0	1
9	mar_tamburlain1.txt	1	9	mar	0	0	0	0	0	0	0	0	2	0	0	0
10	mar_tamburlaine2.txt	1	10	nashe	0	0	0	0	0	0	0	0	0	2	0	0
11	nashe_summer.txt	1	11	peele	0	0	0	0	0	0	0	0	2	0	0	0
12	nashe_test.txt	1	12	shak	0	0	0	0	0	0	0	0	0	0	0	5
13	peele_alcazar.txt	0														
14	peele_davbeth.txt	0														
15	shak_12thnight.txt	1														
16	shak_hamlet.txt	1														
17	shak_lear.txt	1														
18	shak_romjul.txt	1														
19	shak_winters.txt	1														

Accordingly the first eight single plays of Table 6 (column A) have no classification, but in the stylistic comparison (G4-H3) the A and B text of *Dr Faustus* are identical. However, the Marlowe column M includes *Locrine* and *Cornelia*. Next we find the two *Tamburlaines*, (M9), and in line 11 the two Peele plays *The Battle of Alcazar* and *David and Bethsabe* are also given to Marlowe (M11). The two Nashe files are confirmed as well in C10/N10, but they are identical anyway. The Peele column (O) has nothing but zeros, but in the Shakespeare column P we find five Shakespeare plays (P12), registered in A15 to A19. Furthermore *Dido, Queen of Carthage* (P5), *Edward II* (P6), *The Jew of Malta* (P7) and *The Massacre at Paris* (P8) contain stylistic elements which can also be found in Shakespeare's

plays (A15 to A19). In sum, the methodological extensions outlined here, complement the findings previously found in 'The Marlowe Corpus Revisited' and elaborated in detail in *Phantom Marlowe*, so that the existing attributions of literary history need to be reviewed urgently by independent scholars, well versed in non-traditional stylometry.

Lengths of speeches in the number of characters

Average speech lengths of plays are normally not suitable for answering questions of authorship conclusively. However, in the case of real and nominal Marlowe plays, some interesting figures turn up which, together with frequency distribution curves confirm previous findings where the style of the two *Tamburlaines* could also be found in *The Tragedy of Locrine*, in Kyd's closet play *Cornelia* and Peele's *The Battle of Alcazar* and *David and Bethsabe*. The nominal Marlowe plays contain divergent figures which correspond not at all to *Tamburlaine 1 and Tamburlaine 2*.

Plays with Marlowe's style **non Marlovian plays**

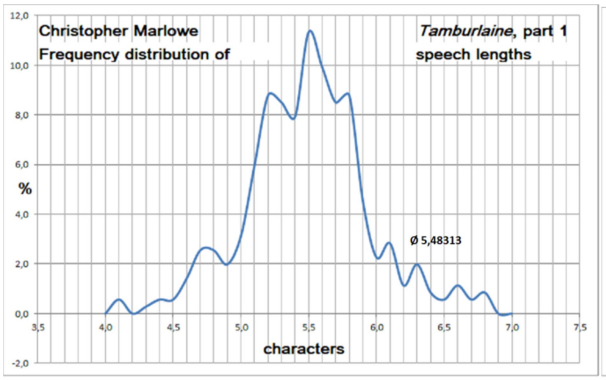


Fig. 3

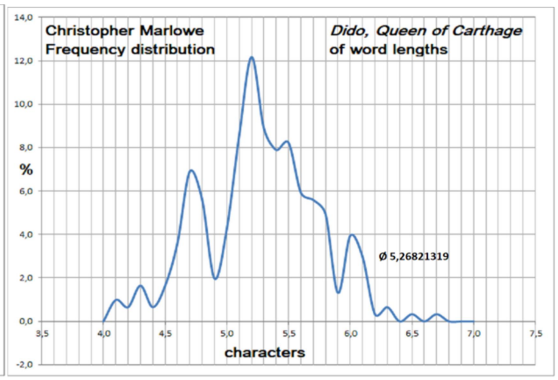


Fig. 4

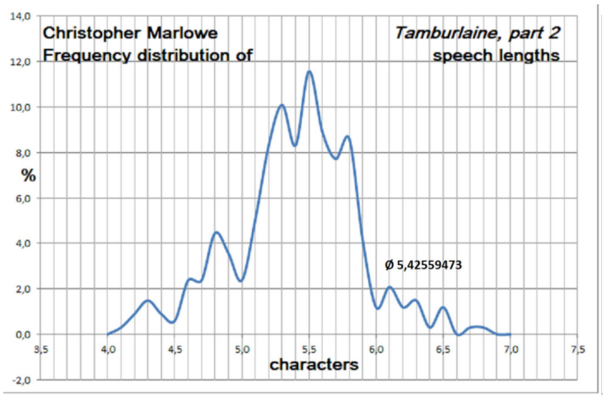


Fig. 5

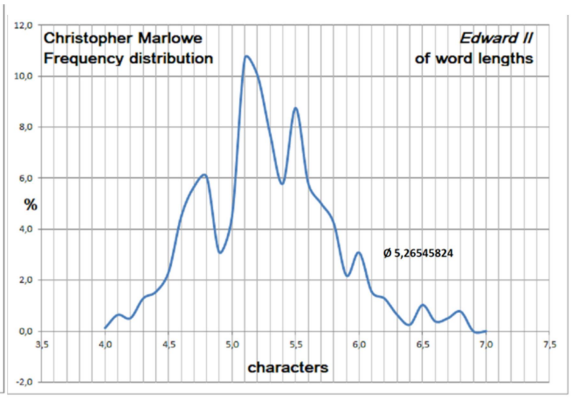
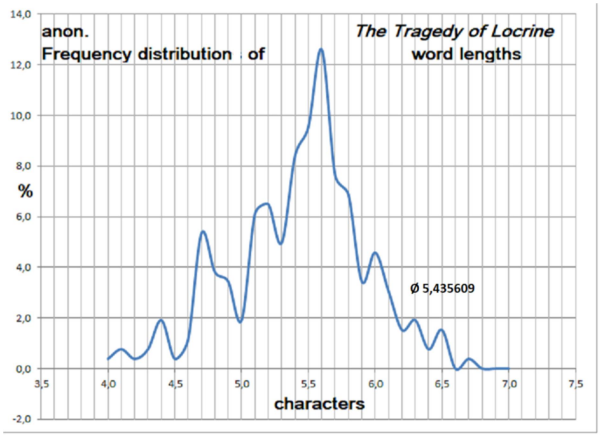
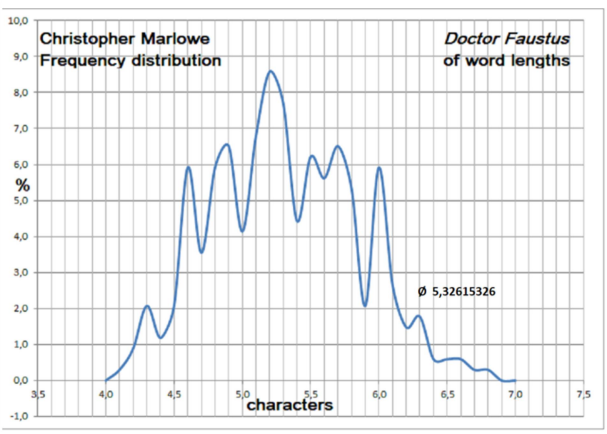
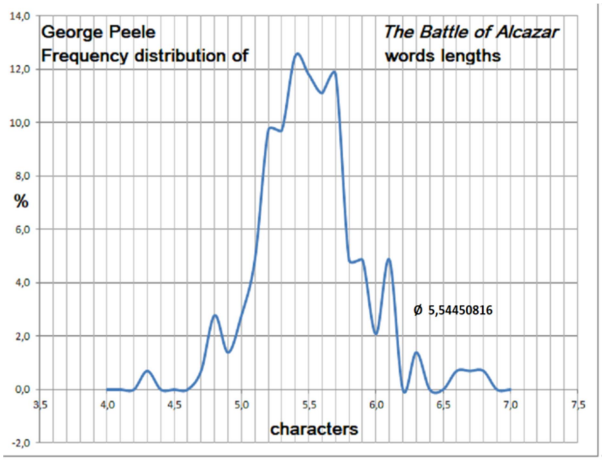
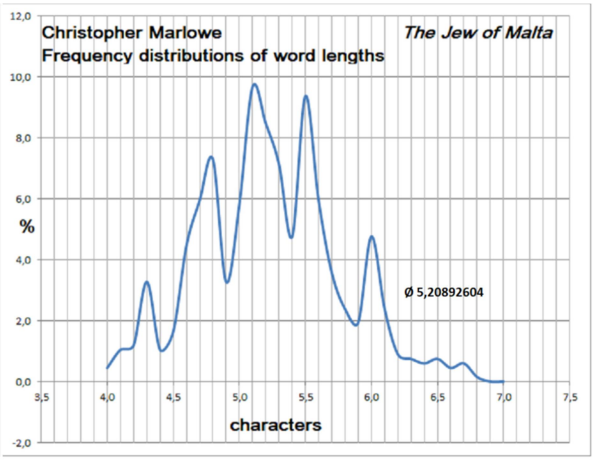


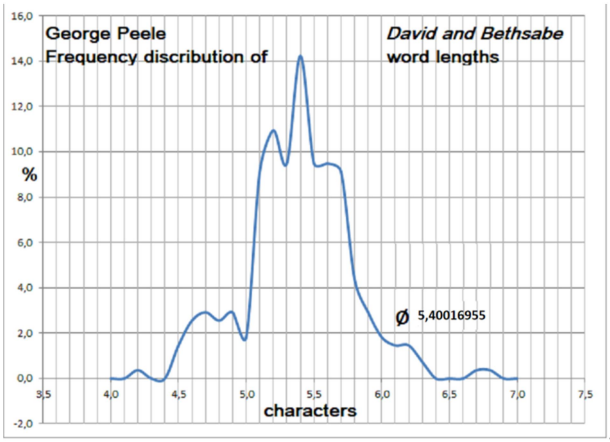
Fig. 6

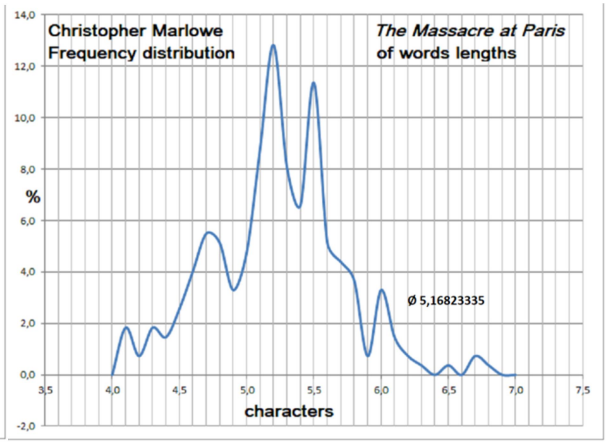

anon.
Frequency distribution of
The Tragedy of Locrine
word lengths
%
characters
Ø 5,435609

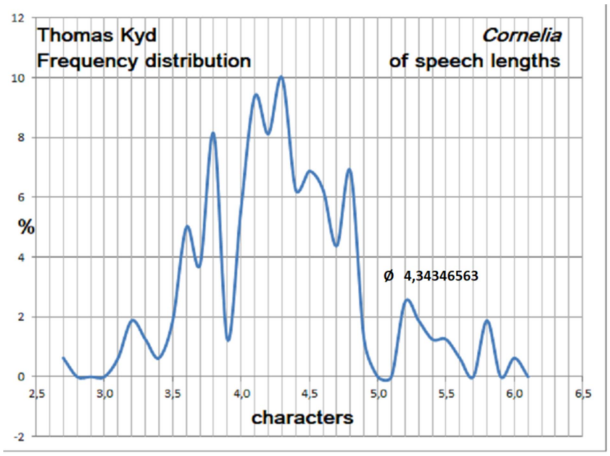

Christopher Marlowe
Frequency distribution
Doctor Faustus
of word lengths
%
characters
Ø 5,32615326


George Peele
Frequency distribution of
The Battle of Alcazar
words lengths
%
characters
Ø 5,54450816


Christopher Marlowe
Frequency distributions of word lengths
The Jew of Malta
%
characters
Ø 5,20892604


George Peele
Frequency discribution of
David and Bethsabe
word lengths
%
characters
Ø 5,40016955


Christopher Marlowe
Frequency distribution
The Massacre at Paris
of words lengths
%
characters
Ø 5,16823335


Thomas Kyd
Frequency distribution
Cornelia
of speech lengths
%
characters
Ø 4,34346563

Figs. 7 to 13

Apart from *Cornelia*[5] the distribution curves on the left are more similar to each other than any of the plays on the right, and none has an average speech length below 5.4 characters whereas the plays on the right are all below 5.4 and often begin with 5.2 or 5.3. This seems to be a trifling matter, but it is not when one considers the amount of text involved. The two columns (Marlowe left, nominal Marlowe right) look indeed as if the individual style of a playwright with a university education faces collaborative writings of two or three contributors, and Germanic vocabulary is opposed to Romance languages. The following table summarises the results already displayed in the charts above.

Marlowe Plays	avg. Length	Non-Marlovian Plays	avg. Length
Tamburlaine 1	5.48313	Dido, Queen of Carthage	5.26821
Tamburlaine 2	5.42559	Edward II	5.26545
Locrine	5.43561	Doctor Faustus	5.32615
The Battle of Alcazar	5.44508	The Jew of Malta	5.20892
David and Bethsabe	5.40017	The Massacre at Paris	5.16823
Cornelia	4.34346		

References

Eder, M. and Rybicki, J. (2013). Do birds of a feather really flock together, or how to choose training samples for authorship attribution. Literary and Linguistic Computing, **28**(2): 229-36.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. R Journal, 8(1): 107–21 https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Eder, M. (2016). Rolling stylometry. Digital Scholarship in the Humanities, **31**(3): 457-469.

Eder, Macij (2017). Cross-validation using the function classify(), Computation Stylistics Group, 13 December, 2017, https://computationalstylistics.github.io/blog/cross-validation/, accessed 14.07.2021.

Eder, Maciej (2018). Authorship verification with the package 'stylo', Computational Stylistics Group, 30 May, 2018, https://computationalstylistics.github.io/blog/imposters/, accessed 14.07.2021.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C. and Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. Digital Scholarship in the Humanities, 32(suppl. 2): 4–16, doi:10.1093/llc/fqx023. http://dx.doi.org/10.1093/llc/fqx023.

[5] The deviating values of *Cornelia* may have to do with its function as a closet play, not intended for performances, but another factor is certainly the stylistic presence of Kyd whose share is at least 10 per cent.

Grieve, Jack (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. Literary and Linguistic Computing, 22(3): 251-270.

Ilsemann, H. (2020). The Marlowe Corpus Revisited, Digital Scholarship in the Humanities, Advance Publication 26.04.2020,1-28. https://doi.org/10.1093/llc/fqaa010.

Ilsemann, H. (2020). Phantom Marlowe: Paradigmenwechsel in Autorschaftsbestimmungen des englischen Renaissancedramas, Düren: Shaker, ISBN 978-3-8440-7412-3.

Ilsemann, H. (…). Phantom Marlowe: Paradigm Shifts in Authorship Attributions of English Renaissance Plays, (as yet unpublished in English).

Juola, P. (2015). The Rowling case: A proposed standard protocol for authorship attribution. Digital Scholarship in the Humanities, 30(suppl. 1): 100–13 doi:10.1093/llc/fqv040.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F. and Daelemans, W. (2016a). Authenticating the writings of Julius Caesar. Expert Systems with Applications, 63: 86–96.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F. and Daelemans, W. (2016b). Authorship verification with the Ruzicka metric. In, Digital Humanities 2016: Conference Abstracts. Kraków: Jagiellonian University & Pedagogical University, pp. 246–49 http://dh2016.adho.org/abstracts/402.

Koppel, M. and Winter, Y. (2014). Determining if two documents are written by the same author. Journal of the Association for Information Science and Technology, 65(1): 178–87 doi:10.1002/asi.22954. http://dx.doi.org/10.1002/asi.22954.

Peñas, A. and Rodrigo, A. (2011). A simple measure to assess non-response. In, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1. Portland, Oregon, pp. 1415–24.

Rizvi, Pervez. (no date). Collocations and N-Grams in Shakespeare's Text: A collection of resources for students of the original texts of Shakespeare's plays, http://shakespearestext.com/can/index.htm, accessed 24.05.2021.

Stamatatos, E. (2006). Authorship attribution based on feature set subspacing ensembles. International Journal on Artificial Intelligence Tools, 15(05): 823–38 doi:10.1142/S0218213006002965.