Hartmut Ilsemann

R Stylo and the authorship determination of *Henry V*

Abstract

Over 25 years, Thomas Merriam has argued that *Henry V* was co-authored by Shakespeare and Christopher Marlowe, and in his most recent publication 'Is it time to reconsider *Henry V'* (2023) he established differences in word length which give clear evidence. This paper makes use of the R Stylo suite of stylometric tools and employs the Rolling Delta, Rolling Classify and the General Imposters methods, all of which obtain the same result that Shakespeare used a Marlowe pretext in his composition of *Henry V*.

Introduction

In his most recent publication 'Is it time to reconsider *Henry V'* Thomas Merriam (2023) maintained that differences in word length can reflect differences in authorship. Over 25 years, he argued that *Henry V* was co-authored by Shakespeare and Christopher Marlowe. It is tempting to use 'this excellent exploratory software suite' (Merriam) R Stylo (Eder, M., Kestemont, M. and Rybicki, J., 2016) to undergird his findings, and it is Rolling Delta in particular which gives clear evidence of co-authorship.

Rolling Delta

The first advantage of Rolling Delta is that more than 100 reference texts were used in a unique selection process to find those texts with the lowest delta values. The target text was of course *Henry V*, but it was rearranged in such a way that all the prose parts came first and were followed by the verse parts.[1] Whereas the normal *Henry V* text yielded a clear Shakespeare result, the combined prose plus verse text came to the following conclusion, in which the lowest deltas, which denote the smallest stylistic difference, were printed in bold white against a black background. The second lowest delta are in white with a dark grey background, and the third lowest deltas have a light grey background. All other plays and authors that did not have a single one of the three lowest delta values were disregarded.

Table 1 *Henry V* arranged in prose and verse

	A	B	C	D	E	F	G	H	I
1	Worte	**Rolling Delta attribution in prose and**							
2	0		**verse parts of *Henry V***						
3	250		Variable: mf3c						
4	500		Window size: 5000 words						
5	750		Step size: 250 words						
6	1000		Culling value: 70 %						
7	1250		analysed with 123 reference texts						
8	1500								
9	1750								
10	2000								
11	2250								
12	2500	32.2	31.3	27.1	35.2	**25.3**	25.8	26.8	29.0

13	2750	31.2	30.8	26.6	34.5	**25.0**	25.6	26.9	28.7
14	3000	31.2	30.8	27.0	34.6	**24.9**	25.8	27.0	29.1
15	3250	31.5	31.4	27.2	34.9	**25.2**	26.4	27.4	29.4
16	3500	32.0	31.8	27.4	35.2	**25.5**	26.7	27.8	29.6
17	3750	31.3	30.9	26.9	34.7	**25.3**	26.6	27.9	29.2
18	4000	31.3	31.0	26.9	34.9	**25.2**	26.7	27.8	29.1
19	4250	31.7	31.5	27.7	35.3	**25.0**	26.6	27.6	28.9
20	4500	31.2	31.1	27.1	35.1	**24.5**	26.1	27.1	28.5
21	4750	31.2	31.2	26.8	34.9	**24.7**	26.2	27.1	28.6
22	5000	31.4	31.3	26.7	35.0	**25.5**	27.0	27.9	28.9
23	5250	31.0	30.6	26.2	34.5	**24.8**	26.4	27.6	28.6
24	5500	31.0	30.5	26.0	34.4	**24.4**	25.9	27.2	27.9
25	5750	31.0	30.6	26.0	34.6	**24.5**	26.2	27.6	28.0
26	6000	31.3	30.9	26.9	35.0	**24.7**	26.3	27.7	28.1
27	6250	31.4	31.1	27.2	35.1	**24.1**	25.8	26.9	27.8
28	6500	32.2	31.6	27.6	35.6	**24.2**	25.9	26.9	27.8
29	6750	32.8	32.3	28.1	36.4	**24.1**	25.4	26.5	27.5
30	7000	34.1	33.4	29.4	37.6	**24.7**	26.0	27.1	28.3
31	7250	34.4	33.6	29.6	37.6	**25.0**	26.4	27.6	28.7
32	7500	34.1	33.3	29.7	37.2	**24.9**	26.1	27.1	28.3
33	7750	34.2	33.4	29.4	37.4	**25.0**	25.8	26.7	27.9
34	8000	34.4	33.5	29.4	37.8	**25.0**	25.5	26.2	27.7
35	8250	34.8	33.8	30.3	38.1	26.0	**25.7**	26.6	28.2
36	8500	34.6	33.6	30.5	38.2	26.1	**25.5**	26.9	28.3
37	8750	35.7	34.8	32.0	39.2	27.6	**26.8**	28.2	29.4
38	9000	36.2	35.0	32.7	39.4	28.0	**27.2**	28.6	30.0
39	9250	35.8	34.5	32.4	38.7	27.7	**26.8**	28.3	29.6
40	9500	35.0	33.5	31.6	37.7	26.9	**26.3**	27.7	29.0
41	9750	34.8	33.2	31.3	37.4	26.7	**26.4**	27.7	28.9
42	10000	34.3	32.7	30.9	36.9	26.4	**25.9**	27.7	28.7
43	10250	34.0	32.5	30.5	36.8	25.6	**25.1**	26.6	27.6
44	10500	33.4	31.7	30.1	36.3	25.0	**24.7**	26.1	27.4
45	10750	32.4	30.7	29.4	35.2	24.8	**24.6**	26.1	26.9
46	11000	31.5	29.8	28.5	34.4	24.4	**24.4**	26.3	26.6
47	11250	30.6	29.2	28.3	33.6	24.9	**24.7**	27.0	26.7
48	11500	29.0	27.5	27.3	32.4	24.3	**24.0**	26.9	25.9
49	11750	28.0	26.6	26.7	31.3	**23.9**	24.3	27.1	25.7
50	12000	27.1	25.7	26.3	30.3	**24.1**	24.9	27.6	26.1
51	12250	26.6	25.2	26.1	29.5	**24.0**	24.8	27.9	26.2
52	12500	25.9	24.3	25.7	28.8	**24.1**	25.0	28.3	26.7
53	12750	25.7	**24.1**	26.0	28.2	24.5	25.7	28.9	27.1
54	13000	25.4	**23.4**	26.3	27.5	25.0	26.4	29.7	27.6
55	13250	24.8	**23.1**	26.5	27.0	25.6	27.2	30.4	28.4
56	13500	25.5	**23.8**	26.9	27.6	26.0	27.9	30.7	28.8
57	13750	25.3	**23.9**	27.3	27.2	25.8	28.1	30.6	28.5
58	14000	24.7	**23.3**	26.7	26.6	25.8	28.3	30.8	28.6
59	14250	24.7	**23.4**	26.8	27.0	26.3	28.8	31.2	29.0
60	14500	25.2	**23.7**	27.0	27.1	26.0	28.4	31.1	28.6
61	14750	25.0	**23.8**	27.0	26.6	26.2	28.3	31.3	28.8

62	15000	25.4	**24.1**	27.3	27.0	26.4	28.5	31.2	28.9
63	15250	24.9	**23.8**	27.2	26.6	26.5	28.7	31.5	29.0
64	15500	24.6	**23.5**	27.5	26.4	26.9	29.0	31.9	29.5
65	15750	24.5	**23.7**	27.4	26.4	26.0	28.3	31.1	28.8
66	16000	24.4	**23.6**	27.4	26.4	26.3	28.5	31.3	28.9
67	16250	24.3	**23.4**	27.3	26.7	26.1	28.3	31.2	28.7
68	16500	24.0	**23.3**	27.1	26.7	25.6	28.0	30.6	28.3
69	16750	24.4	**23.8**	27.5	26.8	25.6	28.2	30.8	28.6
70	17000	24.2	**23.7**	27.2	26.6	25.6	28.2	30.7	28.7
71	17250	23.9	**23.6**	27.0	26.6	25.8	28.8	31.1	29.1
72	17500	24.4	**23.9**	27.4	26.7	26.0	28.9	31.2	29.2
73	17750	24.2	**23.7**	27.0	27.0	25.6	28.2	30.7	28.8
74	18000	23.9	**23.5**	27.0	26.8	25.8	28.7	31.2	29.4
75	18250	24.3	**23.8**	26.9	27.2	25.9	28.9	31.2	29.4
76	18500	23.8	**23.2**	26.9	26.7	25.9	29.0	31.4	29.3
77	18750	23.3	**22.7**	26.3	26.4	25.9	29.0	31.2	29.3
78	19000	23.8	**23.0**	26.8	26.6	26.4	29.5	31.8	29.9
79	19250	23.8	**23.2**	26.7	26.9	26.6	29.4	32.0	30.1
80	19500	23.7	**23.4**	27.2	27.0	27.7	30.4	33.0	31.2
81	19750	24.0	**23.3**	27.0	27.4	27.4	30.0	32.6	30.8
82	20000	23.6	**23.0**	26.7	27.0	27.6	29.8	32.7	30.8
83	20250	23.9	**22.8**	26.4	27.4	27.2	29.1	32.1	30.2
84	20500	24.0	**22.9**	26.4	27.8	27.5	29.2	32.1	30.2
85	20750	24.5	**23.3**	26.2	28.5	27.7	29.4	32.2	30.5
86	21000	24.5	**23.3**	26.1	28.7	27.2	29.2	31.7	30.3
87	21250	24.6	**23.2**	25.9	28.7	27.0	28.7	31.2	30.1
88	21500	25.2	**23.7**	26.0	28.9	27.2	29.1	31.5	30.4
89	21750	25.6	**24.0**	26.1	29.3	27.3	29.2	31.7	30.3
90	22000	25.5	**23.7**	26.2	29.2	27.2	29.2	31.7	30.2
91	22250	25.4	**23.3**	25.7	29.0	26.6	28.6	31.1	29.7
92	22500	25.6	**23.6**	26.1	29.1	26.8	29.2	31.4	29.6
93	22750	24.8	**23.1**	26.1	28.5	26.9	29.3	31.6	29.6
94	23000	24.4	**23.1**	25.2	28.6	26.3	28.4	30.8	28.7
95	23250	24.1	**22.8**	25.3	28.8	26.1	27.8	30.7	28.4
96	23500	23.8	**22.6**	25.5	28.3	26.3	27.8	31.0	28.5
97	23750	B	C	D	E	F	G	H	I
98	24000		**44**			**27**	**14**		
99	24250	42	1	3		16	23		
100	24500	1	2	25	2	24	5	23	3
101	24750								%
102	25000	B =	Marlowe. Tamburlaine 1						
103	25250	C =	Marlowe. Tamburlaine 2					44	51.8
104	25500	D =	Nashe. Summers Last …						
105	25750	E =	Peele. The Battle of …						
106	26000	F =	Shakespeare. Hamlet					27	31.8
107	26250	G =	Shakespeare. Merchant of …					14	16.5
108	26500	H =	Shakespeare. Othello						48.2
109	26750	I =	Shakespeare. Winters Tale						

According to the prose/verse combination, the 5000-word windows that went through the text with a 250-word overlap returned, down to line 52, the lowest deltas, which belonged to the Shakespeare texts, followed by Marlowe's *Tamburlaine 1* and *2*. The ratio was 48.2 Shakespeare and 51.8 Marlowe. Whereas Merriam had used the 276 most frequent word bigrams in his former investigations (Merriam, 2017), the present study employed 70% (the culling value) character trigrams of all the reference texts and the target text. This is a clear result that, with its wider approach, supports Merriam's claim.

Rolling Classify

Rolling Classify, too, with its nearest shrunken centroid (NSC), support vector machine (SVM) and delta classifier, also examined the combined prose/verse text and, similarly to Rolling Delta, gave an attribution for each 250-word section. Due to the mathematical kernel of the classifiers, the results vary, but the overall tendency is clear enough. Instead of endless tables with attributions for each 250-word segment, please find below three representative charts that are provided by the program.

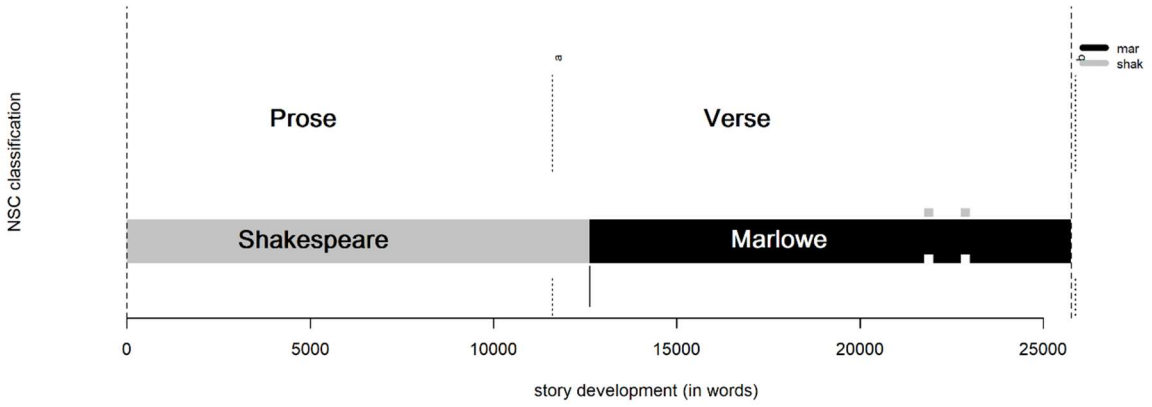

Fig. 1 NSC classification result with 5000-word windows

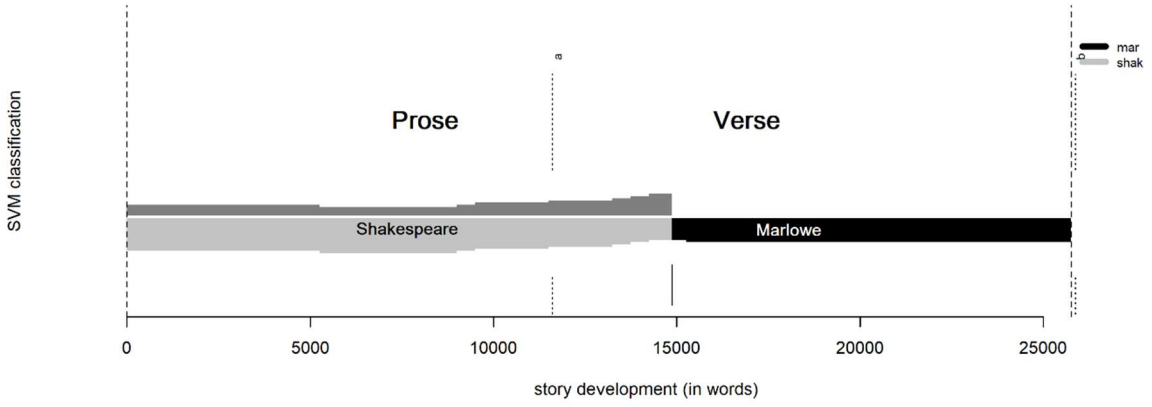

Fig. 2 SVM classification results with 8000-word windows

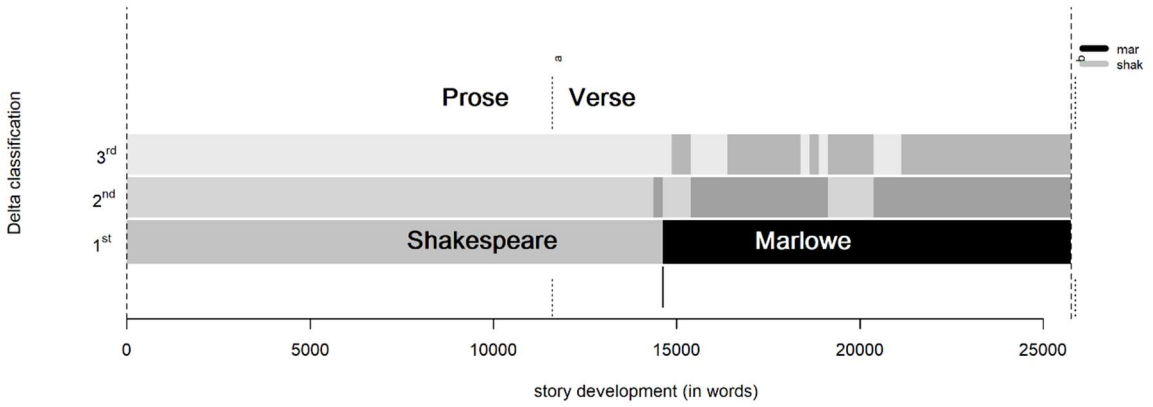

Fig. 3 Delta classification results with 8000-word windows

Prima facie there seems to be a reverberation of assessments that came first when analysing with classifiers. It is the NSC classifier that is closest to Rolling Delta results when the adherence to the prose/verse division of the text was identical. However, considering the Marlowe evaluation of the General Imposters Method below, it is entirely possible that the reduced Marlowe part of the charts in comparison to the prose/verse division is indicative of the lower degree of influence that Marlowe exerted.

The General Imposters method

In 2018 Maciej Eder added the General Imposters Method to the program features of R Stylo. It makes sense to apply this relatively new approach to the verse and prose parts of *Henry V*. In his 2018 blog on the webpage of the Computational Stylistics Group (https://computationalstylistics.github.io/blog/) Maciej Eder gives a detailed account of the new method, referring to its introduction by Koppel and Winter (2014) and Kestemont et al.'s (2016a) application to the study of Julius Caesar's disputed writings. He also quotes the authors' description of the capacity of the new feature:

> the 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 (2018) then describes the prerequisites necessary to use the 'imposters ()' function, namely that all the texts '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.' Initially Eder had claimed that on theoretical grounds, any score above 0.5 would suggest that the authorship verification for a given candidate was successful. However, the latest development is an optimized procedure that checks the grey area of doubtful attributions. Jan Rybicki developed a so-far-unpublished script which gives the boundaries of the grey area. Values above the upper boundary (column C of Table 2) indicate

authorship, values below the lower boundary (column B of Table 2) exclude authorship. The investigations were carried out with the delta classifier to which Eder had added two more distance measures, cosine delta (wu), developed by the Würzburg Computational Stylistics Group, and Ružička metrics (ru). The latter requires a very long computation time, but is regarded as highly reliable. Kestemont et al. (2016b) who had reported on the role of nearest neighbours in determining the authorship of anonymous texts, and of the metrics used 'to calculate the distances between vector representations of texts in a higher-dimensional space' (246) reached the following conclusions in the evaluation of the Ružička distance: 'Comparative evaluations across a variety of benchmark corpora show that this metric yields better, as well as more consistent results than previously used metrics' (246). The tests comprised words (mf1w), word bigrams (mf2w), character bigrams (mf2c), and character trigrams (mf3c), which means that, in combination with delta, wu and ru, each of the analysed text segments of *Henry V* undergoes twelve evaluations.[2]

Table 2 GI and prose and verse parts of *Henry V*

	A	B	C	D	E	F	G
1							
2	**delta**	low	high	**mar**	**row**	**shak**	
3	h5prose	0	0.99	0.14	0	**1**	mf1w
4	h5prose	0	0.99	0.08	0	**1**	mf2w
5	h5prose	0	0.99	0.03	0.31	0.96	mf2c
6	h5prose	0	0.99	0	0.41	**1**	mf3c
7							
8	**wu**	low	high	**mar**	**row**	**shak**	
9	h5prose	0	0.99	0.14	0	0.89	mf1w
10	h5prose	0	0.99	0.18	0	0.92	mf2w
11	h5prose	0	0.97	0.14	0.02	0.62	mf2c
12	h5prose	0	0.99	0.15	0	0.71	mf3c
13							
14	**ru**	low	high	**mar**	**row**	**shak**	
15	h5prose	0	0.99	0	0.28	**1**	mf1w
16	h5prose	0	0.99	0	0.09	**1**	mf2w
17	h5prose	0	0.99	0	0.51	0.98	mf2c
18	h5prose	0	0.99	0	0.31	**1**	mf3c
19							
20	**delta**	low	high	**mar**	**row**	**shak**	
21	h5verse	0	0.99	0.26	0	**1**	mf1w
22	h5verse	0	0.99	0.92	0	0.73	mf2w
23	h5verse	0	0.99	0.62	0.07	0.87	mf2c
24	h5verse	0	0.99	0.65	0.03	0.95	mf3c
25							
26	**wu**	low	high	**mar**	**row**	**shak**	
27	h5verse	0	0.99	0.22	0	0.34	mf1w
28	h5verse	0	0.99	0.69	0	0.23	mf2w
29	h5verse	0	0.99	0.18	0	0.22	mf2c
30	h5verse	0	0.99	0.15	0	0.41	mf3c
31							

	ru	low	high	mar	row	shak	
33	h5verse	0	0.99	0.72	0	0.65	mf1w
34	h5verse	0	0.99	0.53	0.03	**1**	mf2w
35	h5verse	0	0.99	0.98	0.03	0.19	mf2c
36	h5verse	0	0.99	0.85	0	0.61	mf3c

The prose parts of *Henry V* are listed in lines 2 to 18 and the verse parts in lines 31 to 36. As so often happens, cosine delta (wu) was very reluctant to acknowledge clear attributions. Otherwise Shakespeare dominates clearly in the prose parts of *Henry V*. The evaluation of the verse parts is more complicated. Word frequencies of delta (F21) and word bigrams of the Růžička metric (F34) still opt for Shakespeare, but, in the grey area of doubtful attributions, we also find Marlowe, represented by *Tamburlaine 1* and *2*. This means that there is some stylistic influence of Marlowe, and, even though the final version of *Henry V* is by Shakespeare, it is more than likely that he used a pretext by Marlowe. It is certainly an asset of GI that, in contrast to classifiers, which return only one author, the figures indicate collaborations as well.

Conclusion

In the light of the R Stylo results one can state that Merriam's suggestion of the need to reconsider *Henry V* was more than justified. Simultaneously, the results have made it clear that the bigger task of reconsidering apocryphal plays by Shakespeare with R Stylo is also more than overdue.

References

Eder, M., (2018). "Authorship verification with the package 'stylo'", Blog of the Computational Stylistics Group, https://computationalstylistics.github.io/blog/ (accessed 25.03.2023).

Eder, M., Kestemont, M. and Rybicki, J. (2016). Stylometry with R: A package for computational text analysis, R Journal, 16(1): 107–121, https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

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.

Merriam, T. (2017). Verse and prose in *Henry V.* Notes and Queries, 262: 267–9.

Merriam, T. (2023). Is it time to reconsider *Henry V*?, Digital Scholarship in the Humanities, advance publication fqad015, https://doi.org/10.1093/llc/fqad015

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.

Notes

[1] I am very grateful to Thomas Merriam for providing me with the prose and verse texts of *Henry V*.

[2] The corpus files from which the target texts and the authors columns were derived were h5prose.txt, h5verse.txt, mar_tamburlain1.txt, mar_tamburlain2.txt, row_whenysee.txt, shak_hamlet.txt, shak_romjul.txt. Rowley's text was needed to fulfil the minimum of three authors.