Hartmut Ilsemann

Methodological Observations concerning Word Rankings and Z-Score Refinements.

Abstract

This paper evaluates word rankings suggested by Ary L. Goldberger, Albert C. Yang, and C. Peng as a means of establishing the authorship of texts in the light of Delta, developed by John Burrows at about the same time. The tests carried out with high ranking function words and results established with the more modern approaches of Rolling Delta, Rolling Classify, and the General Imposters method give clear evidence that word rankings only return crude and unreliable results that cannot keep up with non-traditional modern methods. Even though the stylistic difference between Marlowe and Shakespeare plays could be stated, word rankings failed to recognise Shakespearean stylistics in *The Jew of Malta, Edward II,* and *Doctor Faustus*. It was only through the use of z-scores that a wider vocabulary provided a larger degree of differentiation.

Evaluation

In 2003 Goldberger, Yang, and Peng introduced a new method to solve an old problem, namely the Marlowe-Shakespeare authorship question. In their approach they started from the assumption that "each author has his/her vocabulary 'database', related to education, culture, and life experience" (Goldberger et al., 2003, p.8). These word preferences become apparent in the frequency with which words are used. The highest frequency is regularly reached by function words like the, and, but, to, and from, which have a large number of contextual possibilities. A particular word like "not," for example, may be placed in different positions in the ranking order of various texts. Thus in Marlowe's *Tamburlaine, Part 1* it holds position 24, and in *Tamburlaine, Part 2*, position 23, in *The Tragedy of Locrine*, it has position 28, whereas in *Sir John Oldcastle*, it takes position 16 in the word ranking order. In texts written by the same author there is not much difference in the ranking of words. Thus *Tamburlaine 1* and *2* return a somewhat diagonal line when words are plotted with their coordinates.

Table 11 Word ranking order of *Tamburlaine 1* and *2*

#Word Types: 3140			#Word Types: 3489		
#Word Tokens: 17609			#Word Tokens: 17647		
#Tamburlaine 1			#Tamburlaine 2		
RANK	Frequ.	WORDS	RANK	Frequ.	WORDS
1	765	and	1	803	the
2	701	the	2	776	and
3	519	of	3	579	of
4	424	to	4	404	to
5	341	my	5	341	my
6	254	with	6	254	i
7	232	in	7	234	in
8	229	d	8	228	with
9	227	that	9	221	a
10	216	i	10	209	that
11	196	his	11	149	his
12	187	a	12	148	as
13	160	as	13	141	your

14	151	for	14	138	our
15	130	be	15	135	for
16	126	your	16	126	shall
17	125	you	17	120	all
18	120	their	18	120	be
19	119	shall	19	119	this
20	108	our	20	110	thy
…	…	…	…	…	…

Table 1 returns the first 20 positions of word ranking lists derived from Laurence Anthony's (2022) AntConc program. Five of twenty words occupy the same ranks and the remainder do not differ much from their respective equals. The conjunction and was used most often in Tamburlaine 1, but only holds position 2 in Tamburlaine 2, with the definite article the, it is the other way round. Other words like you and their are outside the range of 20 words and would only show up in a larger selection. The possessive pronoun "our" in Fig. 1 holds position 14 in Tamburlaine 2, but position 19 in Tamburlaine 1, a difference of five places. In Fig. 1, Table 1 is transformed as follows.

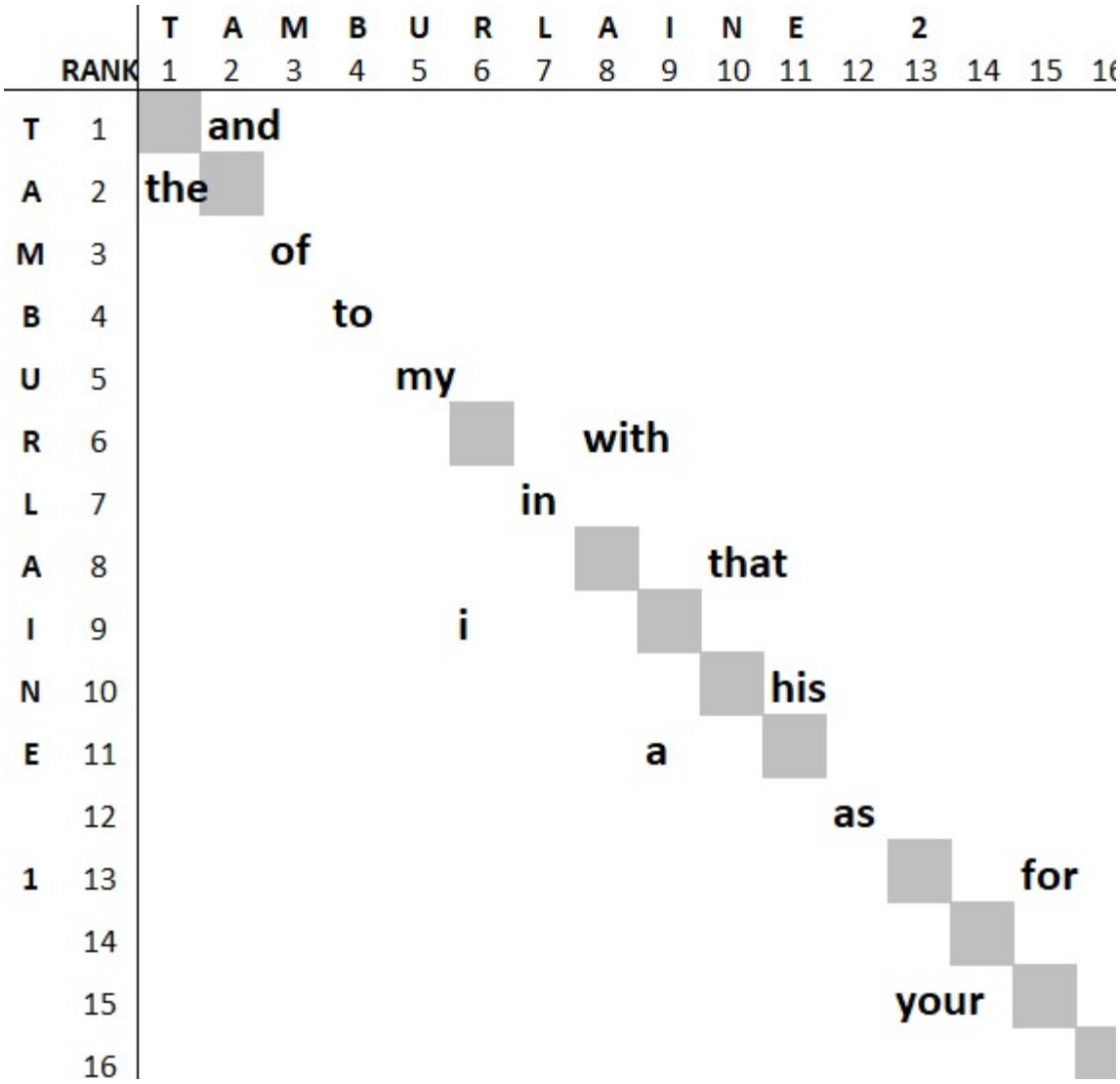

Fig. 1 Diagram with differences in the ranking order of words

It is obvious that starting with the lower distances of the most frequent words the distances between words grow tremendously as the whole range of less frequent words comes into play. It is almost impossible and does not make any sense at all to display the whole vocabulary of plays in this way. However, if the accumulated sum of absolute differences up to, let's say, the first fifty most frequent words is displayed, similarities and sizes of differences become visible. In Fig. 2, the curves containing the differences between *Tamburlaine 1* and *Tamburlaine 2* are presented. The accumulated difference in the positioning of words between *Tamburlaine 1* and *Tamburlaine 2* amounts to 271 at position 50, that between *Tamburlaine 2* and *Tamburlaine 1* to 317. Noticeable differences become visible from position 23 where the two curves start to split.

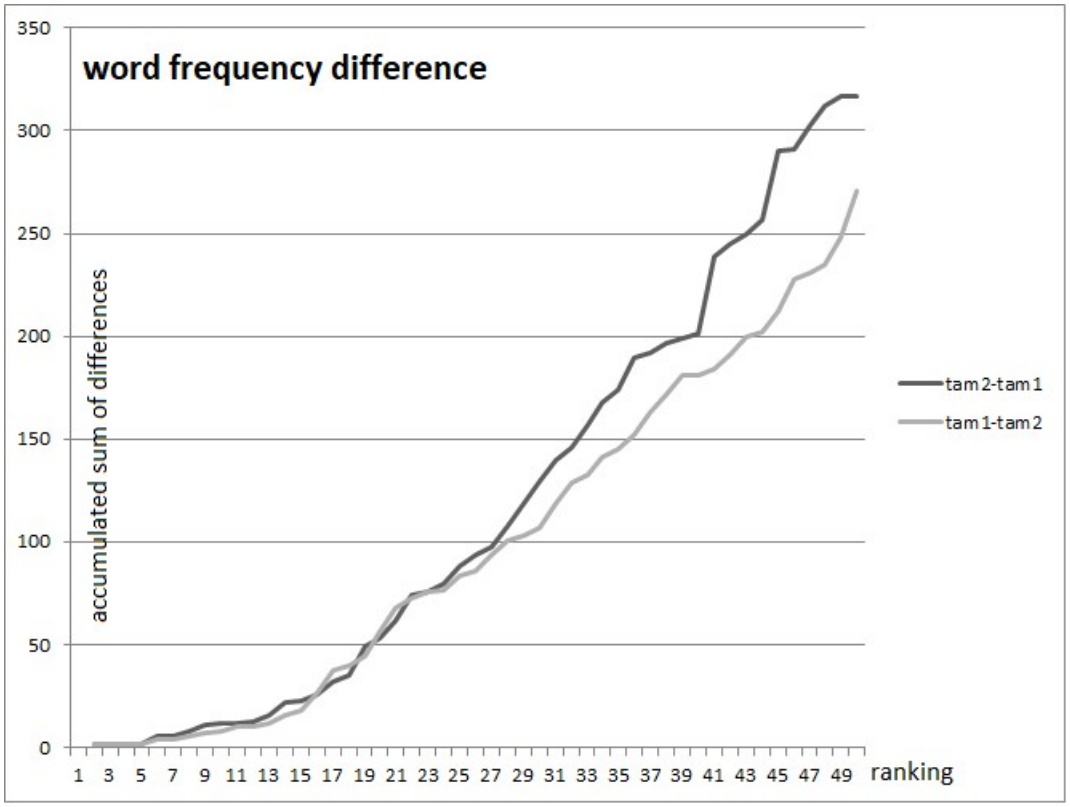

Fig. 2 Word frequency differences between *Tamburlaine 1* and *Tamburlaine 2*

The crucial question is what happens if other curves come in, derived from the ranking differences between a Marlowe and a Shakespeare play? The accumulated sum of differences between *Tamburlaine 1* and *1 Henry IV* at frequency position 50 goes up to 661 and in the case of *1 Henry IV* minus *Tamburlaine 1* even to 1,271. Their curves, integrated into the diagram above, produce the relationships in Fig. 3.

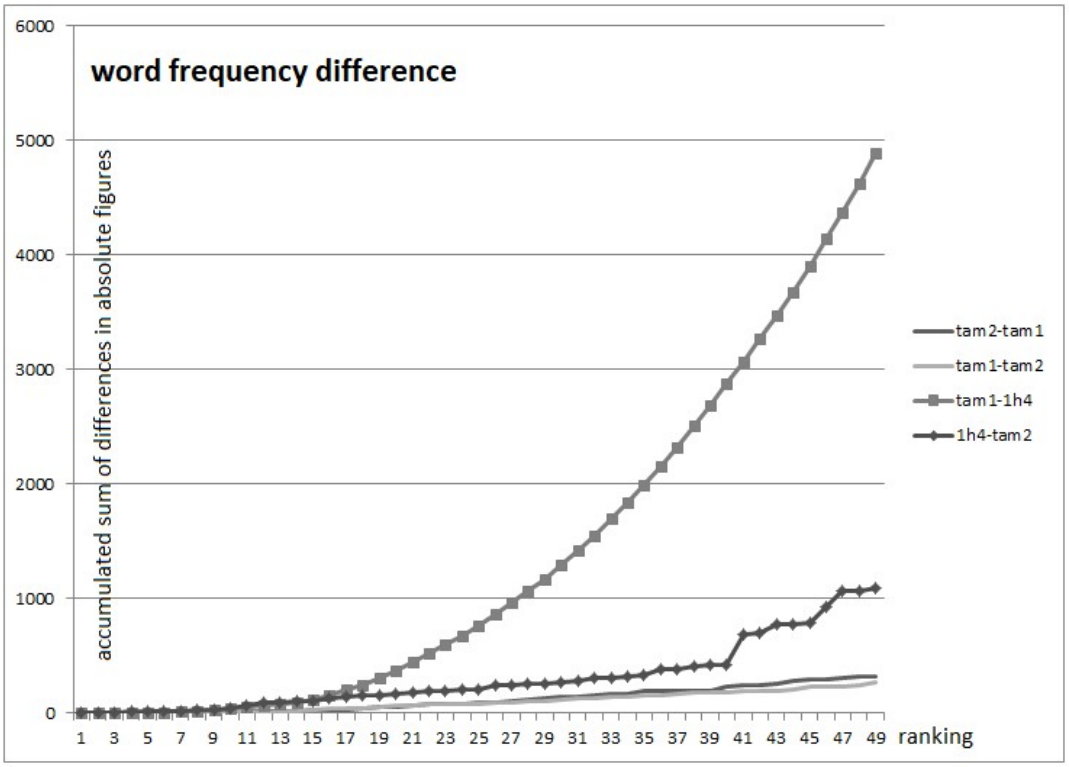

Fig. 3 Word frequency differences between Marlowe's *Tamburlaines* and Shakespeare's *1 Henry IV*

The larger scaling in Fig. 3 shows the differences between Marlowe's *Tamburlaines* and Shakespeare's *1 Henry IV*. Even though Goldberger and his co-authors used the culturally transmitted Marlowe corpus they came to the same conclusion, namely that "the major dramatic works attributed to William Shakespeare are clearly distinct from those of Christopher Marlowe" (Goldberger et al., 2003, pp.22–23).

In the last few years a number of different new approaches like Rolling Delta, Rolling Classify, and the General Imposters method have all confirmed the stylistic inconsistency of the official Marlowe corpus. If we take the accumulated sum of absolute differences in the frequency of words at position 50, a lot of information can be derived from a comparison of plays.

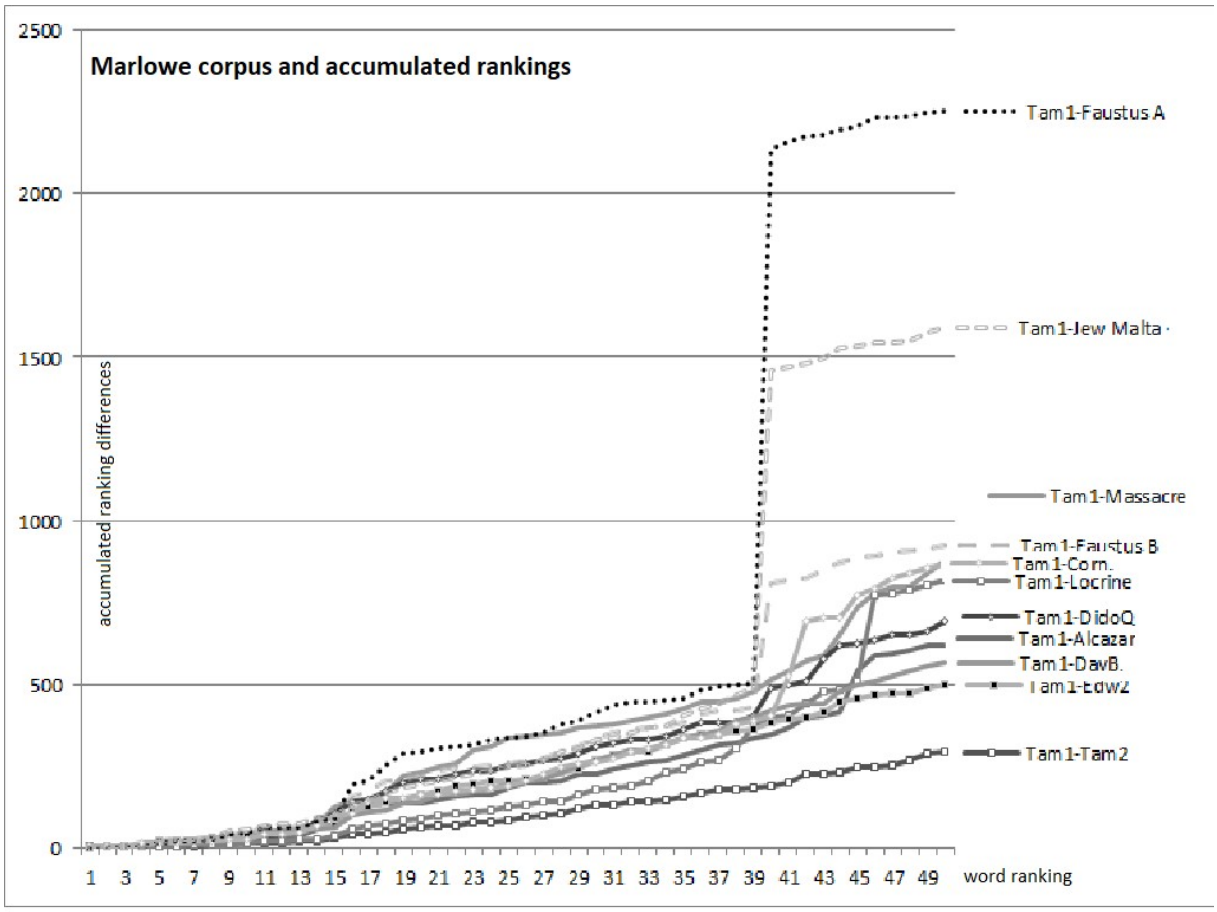

Fig. 41 Word ranking differences in the Marlowe corpus at position 50

The smallest stylistic difference derived from word rankings is that between *Tamburlaine 1* and *Tamburlaine 2*, which is not surprising as Marlowe became famous with them after their performances in 1587 and 1588. That Marlowe was indeed the author of these plays is confirmed by empirical evidence, as it was Marlowe who was questioned after the Dutch Church Libel in May 1593, when the pamphlet was signed 'Tamburlaine'. The next closest figures refer to the word ranking comparison between *Tamburlaine 1* and *Edward II*. No wonder that many scholars, when they compared function word frequencies, came to the apparently safe conclusion that *Edward II* was a play by Marlowe. The next closest plays which do not deviate too much from the *Tamburlaines* are Peele's *David and Bethsabe* and *The Battle of Alcazar*, followed by *Dido, Queen of Carthage* and *The Tragedy of Locrine*, with Kyd's closet play *Cornelia*. It was only with the advent of Delta (Burrows 2002) and later Rolling Delta that Peele's *David and Bethsabe* and *The Battle of Alcazar*, the anonymous play *The Tragedy of Locrine* and Kyd's *Cornelia* were identified as plays by Marlowe. The remaining plays in the corpus, however, were in no way coherent with the style of the *Tamburlaines*, and *Edward II* came out as a play by Shakespeare and Kyd (see Table 2).

Table 2 Marlovian and non-Marlovian style

Tamburlaine 1 *Tamburlaine 2* their style features	
<u>can be found in (+)</u>	<u>are absent in (-)</u>
anon. *The Tragedy of Locrine*	*Dido, Queen of Carthage*
Peele. *The Battle of Alcazar*	*The Jew of Malta*
Peele. *David and Bethsabe*	*The Massacre at Paris*
Kyd. *Cornelia*	*Edward II*
	Dr. Faustus (A) – 1604
	Dr. Faustus (B) – 1616

When Burrows developed Delta just before 2002, he used a total of 150 words, and it was only due to the introduction of z-scores that the basis for word comparisons was enlarged. If we look at word ranking differences at position 100, we note an advancement in the method.

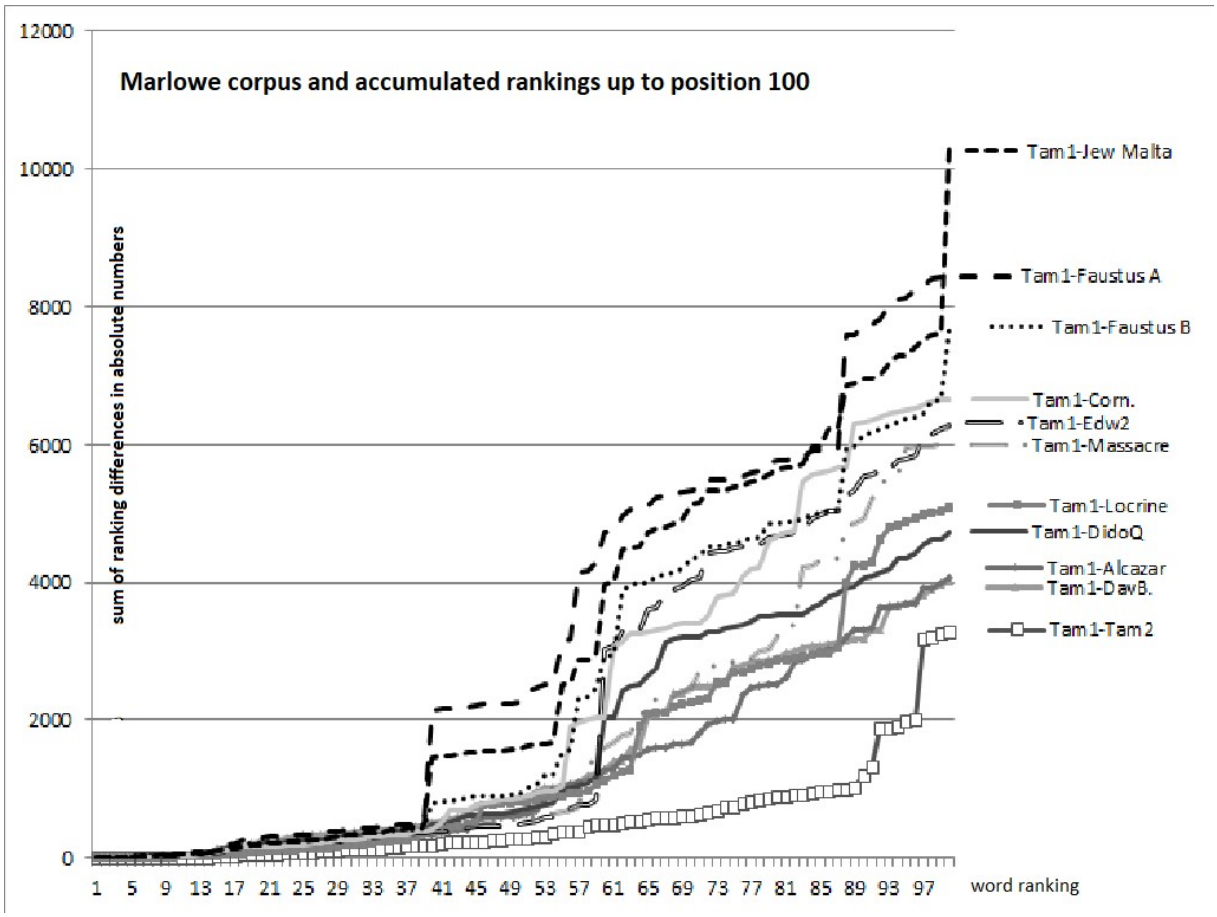

Fig. 5 Word ranking differences in the Marlowe corpus at position 100

The most distant play from the *Tamburlaines* is now *The Jew of Malta*, followed by the two *Faustus* texts A (1604) and B (1616). *Edward II* is no longer within the closer circle of real Marlowe plays, and Kyd's *Cornelia* has also moved to a more distant position. This may have to do with the influence of Kyd, who apparently revised the play Marlowe left behind (see Ilsemann, 2019). *Dido, Queen of Carthage* is still within the compound of Marlowe plays. Rolling Delta demonstrated that the beginning of the play may be by Marlowe, but widening the approach with more vocabulary showed its overall non-Marlovian style. It is within the context of the present analysis of word frequencies and difference of rankings that one begins to

appreciate Delta and its successor Rolling Delta (Eder, M., Rybicki; J. and Kestemont, M., 2016). Delta was based on the observation that the frequency of occurrences in each word list decreases rapidly and the difference from the mean frequency increases with each word. To ensure that the rapidly decreasing frequency of words is equally included in the rating, z-scores were calculated by dividing the difference between the mean and the actual frequency by the standard deviation. The result has a positive or negative value, depending on whether the word is above or below the mean. The absolute difference between the z-scores of the search text and the reference texts then gives the Delta value, which is the expression of the stylistic difference between the texts. Thus, Delta is based on a much wider span of vocabulary than ranking comparisons, and even though Goldberger and his co-authors were able to prove an overall Shakespeare-Marlowe difference, their results emerged from measuring a larger corpus in a crude and unsophisticated way and not from direct, minute play by play comparisons.

References

Anthony, L. (2022). AntConc (Version 4.1.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software (accessed 22.08.2022).

Burrows, J. (2002). 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. Literary and Linguistic Computing, 17(3): 267–287.

Eder, M., Rybicki; J. and Kestemont, M. (2016). "Stylometry with R: A Package for Computational Text Analysis", Contributed Research Articles, The R Journal 8(1): 107–121.

Goldberger, A.L., Yang, A.C.C. and Peng, C.K. (2003). The Marlowe-Shakespeare Authorship Debate: Approaching an Old Problem with New Methods. https://www.medievalists.net/ 2011/11/the-marlowe-shakespeare-authorship-debate-approaching-an-old-problem-with-new-methods/ (accessed 20 August 2022).

Ilsemann, H. (2019). Forensic Stylometry. Digital Scholarship in the Humanities, 34(2):335–349. https://doi.org/10.1093/llc/fqy023.