Shakespeare Review 60.4 (2024): 695-715
DOI: 10.17009/shakes.2024.60.4.003

Some Problems in Using Numbers to Represent the Writing Styles of Shakespeare and His Contemporaries*

Gabriel Egan
(De Montfort University)

As was shown by George Lakoff and Mark Johnson, we habitually use the metaphor of distance when we want to express similarity. We say that one film adaptation is much closer to its source novel than another. We say that one writer's style is far from the norm. Of course, we know that the notion of distance has no direct application to language. The two domains of language and space are incommensurable. Yet the metaphor of distance is one we easily fall into when discussing texts. And it ceases to be a metaphor once we start to count things in texts, which is my concern in this essay.

We are familiar with the notion of distance between numbers. How far is it from 4 to 7? It is 3. This use of distance is not metaphorical if we think of the numbers lying along a line, as they do on a ruler. Along a ruler, the difference

* I would like to record the assistance of my PhD student Nathan Dooner in helping me to understand some of the concepts and problems introduced in this essay. Any errors about them are of course mine not his.

of 4 and 7 is literally a distance of 3. This is the absolute magnitude that follows from a subtraction, which is absolute in the sense that we discard the sign of the answer. 7 minus 3 is 4 and 3 minus 7 is -4, so if we discard the sign, it matters not which of the two terms we put first. In mathematics, a vertical bar before and after a term means that we take its absolute value, so |3-7| is 4.

One of the simplest things to count in language is the number of words. We can ask “What is the distance between the number of words in Shakespeare’s dramatic canon and the number of words in Christopher Marlowe’s dramatic canon?” First, we must agree that here we mean by “words” the tokens, so that “never, never, never” counts as three-word tokens not one word type. Next, we must agree on exactly which plays Shakespeare and Marlowe wrote. To assist in the work of the New Oxford Shakespeare editors in 2011, Hugh Craig made this calculation based on an agreed set of attributions to Shakespeare—leaving out the disputed *Arden of Faversham*, *Double Falsehood*, and the Additions to *The Spanish Tragedy*—and came up with the number 740,209 (Taylor 247). Assisting the same project, Paul Brown calculated word counts of plays by other dramatists of Shakespeare’s time and if we agree that the Marlowe canon is *Doctor Faustus*, *Edward II*, *The Jew of Malta*, *The Massacre at Paris*, and Parts One and Two of *Tamburlaine* then the total from Brown’s counts is 101,146 (Brown).

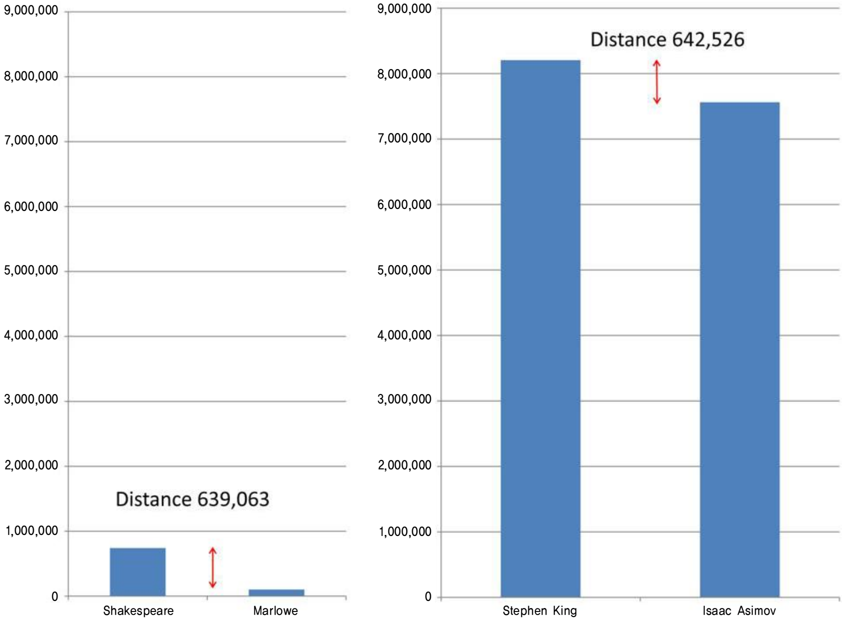


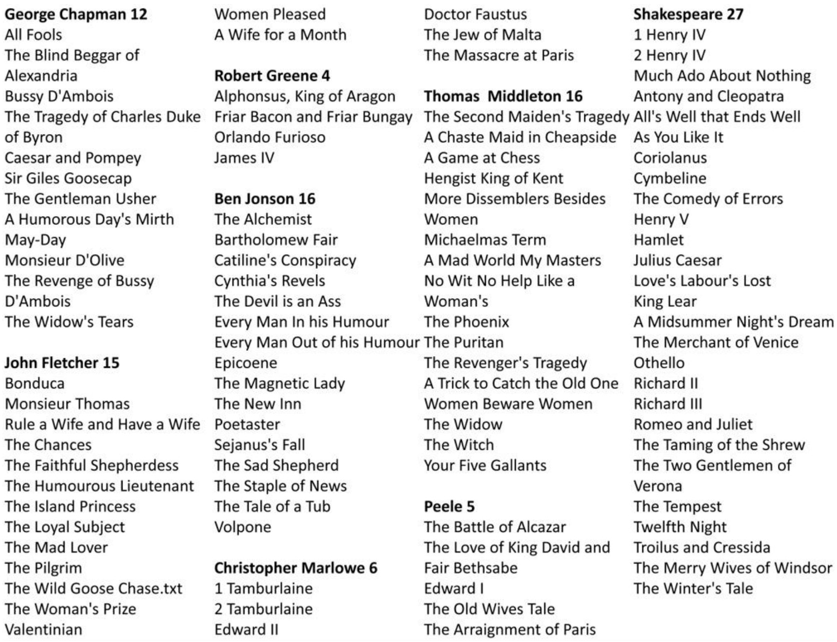
Fig. 1. "The size of the Shakespeare, Marlowe, King, and Asimov canons on the same scale"

By these scholars' calculations, Shakespeare's canon is 639,063 words bigger than Marlowe's canon. We can do the same calculation for the complete works of the modern writers Stephen King (just over eight million words) and Isaac Asimov (around seven-and-a-half million words). The distance between these numbers is 642,526, almost exactly the same as the distance between the size of the Shakespeare canon and that of the Marlowe canon. In Figure One, we put the two histograms on the same scale. Can we say that, regarding canon size, Shakespeare is to Marlowe as King is to Asimov, because the distance between the sizes of the two canons is, in each pairing, about the same? No, because the King and Asimov canons are much bigger than the Shakespeare and Marlowe canons, making the same distance of around 640 thousand words have different significances in the two cases. For every 10 words that Marlowe left us, Shakespeare left us about 70, whereas for every 10 words that Asimov left us, King has given us about 11. Sometimes the correct measure of difference is not distance but proportion, and the correct operation is division not subtraction. Instead of subtracting one number from another we divide one number by another, and we get their relative proportions.

Proportions are meaningful when the numbers arise from counting actual objects in the world. But when our numbers arise from an arbitrary scale that humans have invented, such as temperature, the reverse is true. There is no meaningful way to divide one temperature by another because in any temperature scale we simply invent the magnitude of the unit and set the zero point arbitrarily. This, then, is Common Methodological Error #1. A surprising number of studies that quantify aspects of language get this wrong and use a measure of distance where proportion is the right measure or use proportion where distance is the right measure.

* * *

The 740 thousand words in the Shakespeare canon are not 740 thousand different words, of course, since many of these are repetitions of the same word. Shakespeare used 740 thousand word tokens, but many fewer distinct word types. To figure out how many words Shakespeare knew—his vocabulary—we can begin by considering how many different word types he used in his plays. To count the types in Shakespeare, I will use the corpora of sole-authored well-attributed plays by Shakespeare and seven of his fellow dramatists for whom more than a few plays survive, as shown in Figure Two. For each of these corpuses I used the transcriptions of the plays from the ProQuest *One Literature* database. After each name I have in the figure recorded how many plays are in that dramatist's corpus.



George Chapman 12
All Fools
The Blind Beggar of Alexandria
Bussy D'Ambois
The Tragedy of Charles Duke of Byron
Caesar and Pompey
Sir Giles Goosecap
The Gentleman Usher
A Humorous Day's Mirth
May-Day
Monsieur D'Olive
The Revenge of Bussy D'Ambois
The Widow's Tears

John Fletcher 15
Bonduca
Monsieur Thomas
Rule a Wife and Have a Wife
The Chances
The Faithful Shepherdess
The Humourous Lieutenant
The Island Princess
The Loyal Subject
The Mad Lover
The Pilgrim
The Wild Goose Chase.txt
The Woman's Prize
Valentinian
Women Pleased
A Wife for a Month

Robert Greene 4
Alphonsus, King of Aragon
Friar Bacon and Friar Bungay
Orlando Furioso
James IV

Ben Jonson 16
The Alchemist
Bartholomew Fair
Catiline's Conspiracy
Cynthia's Revels
The Devil is an Ass
Every Man In his Humour
Every Man Out of his Humour
Epicoene
The Magnetic Lady
The New Inn
Poetaster
Sejanus's Fall
The Sad Shepherd
The Staple of News
The Tale of a Tub
Volpone

Christopher Marlowe 6
1 Tamburlaine
2 Tamburlaine
Edward II
Doctor Faustus
The Jew of Malta
The Massacre at Paris

Thomas Middleton 16
The Second Maiden's Tragedy
A Chaste Maid in Cheapside
A Game at Chess
Hengist King of Kent
More Dissemblers Besides Women
Michaelmas Term
A Mad World My Masters
No Wit No Help Like a Woman's
The Phoenix
The Puritan
The Revenger's Tragedy
A Trick to Catch the Old One
Women Beware Women
The Widow
The Witch
Your Five Gallants

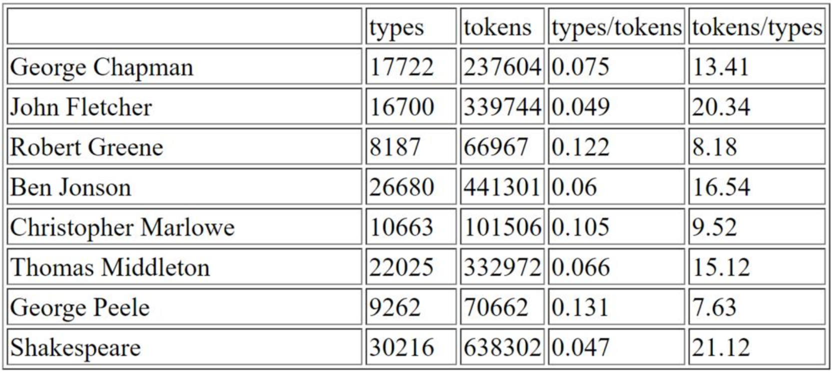
Peele 5
The Battle of Alcazar
The Love of King David and Fair Bethsabe
Edward I
The Old Wives Tale
The Arraignment of Paris

Shakespeare 27
1 Henry IV
2 Henry IV
Much Ado About Nothing
Antony and Cleopatra
All's Well that Ends Well
As You Like It
Coriolanus
Cymbeline
The Comedy of Errors
Henry V
Hamlet
Julius Caesar
Love's Labour's Lost
King Lear
A Midsummer Night's Dream
The Merchant of Venice
Othello
Richard II
Richard III
Romeo and Juliet
The Taming of the Shrew
The Two Gentlemen of Verona
The Tempest
Twelfth Night
Troilus and Cressida
The Merry Wives of Windsor
The Winter's Tale

Fig. 2. "The sole-authored well-attributed plays of Shakespeare and seven of his contemporary dramatists"

A complicating factor is that in a small sample of writing we will, simply because it is small, find fewer word types than in a longer piece of writing by the same author. Word types that we rarely use simply do not get the opportunity, as it were, to appear in a short sample of language. To adjust for the different sizes of the corpora we might decide to divide the number of word types in a writer's corpus (the count of how many different words) by the number of word tokens in that corpus (the count of how large the corpus is). In this division, a small corpus that uses many different words will get a result, a quotient, larger than a big corpus that uses few different words. This types-to-tokens ratio is a measure of the richness of variety in a writer's language. For our eight dramatists the ratios are shown in Figure Three.

$$\text{variety of language} = \frac{\text{number of word types}}{\text{number of word tokens}}$$



	types	tokens	types/tokens	tokens/types
George Chapman	17722	237604	0.075	13.41
John Fletcher	16700	339744	0.049	20.34
Robert Greene	8187	66967	0.122	8.18
Ben Jonson	26680	441301	0.06	16.54
Christopher Marlowe	10663	101506	0.105	9.52
Thomas Middleton	22025	332972	0.066	15.12
George Peele	9262	70662	0.131	7.63
Shakespeare	30216	638302	0.047	21.12

Fig. 3. "The type-to-token ratios for Shakespeare and seven of his contemporary dramatists"

The smaller the types/tokens value, the less the variety in the writing. In the last column of Figure Three I have flipped these values to give the reciprocal, the ratio of tokens to types, and on this measure the higher the value the less the variety in the writing. It is noticeable that the highest three values in this column, the dramatists with the least varied writing, are Fletcher, Jonson, and Shakespeare: the dramatists for whom we have the most surviving plays. And it is noticeable that the lowest three values in this column, the dramatists with the most varied writing, are Greene, Marlowe, and Peele: the dramatists for whom we have the fewest surviving plays.

In fact, this calculation of language variety or richness is misleading. While the result is legitimately and objectively a measure of the linguistic variety or richness of these texts, it is not a good measure of anything we want to consider as a writer's style. The reason is that dividing the number of different word types by the size of the canon measured in tokens overcompensates for the effect of some writers having large canon sizes, making their style seem less varied than that of writers with small canons. Anyone's speech or writing starts to appear less varied the more we listen to them or read their writings. Simply scaling one's counts by the size of a dramatist's canon would be effective if the relationship between the two values—number of types and number of tokens—were linear. But it is not. This is Common Methodological Error #2: assuming that a relationship is linear when it is not.

Rather than a straight line, the type/token relationship is a characteristic curve. To illustrate it, Gilbert Youmans (588) noted how many different types had been encountered (and recorded on the *y* axis) as he read through, from first word to last, the 5000 tokens of a particular text (recorded on the *x* axis). His illustration is reproduced here as Figure Four. He chose as his text the simplified story of Shakespeare and Middleton's *Macbeth* as told in Charles Lamb's *Tales from Shakespeare* and rendered into the Basic English system invented by Charles Kay Ogden, in which only 850 different word types are allowed. With so few types at the writer's disposal, it soon became necessary, after writing just a few sentences, to heavily reuse types that had already been used. Thus, each new sentence is increasingly made up of repetitions of previously used word types and the curve soon starts to plateau. That is, as the token count rises steadily, the type count—which is increased only by the use of new words not previously seen in the text—goes up by ever smaller amounts.

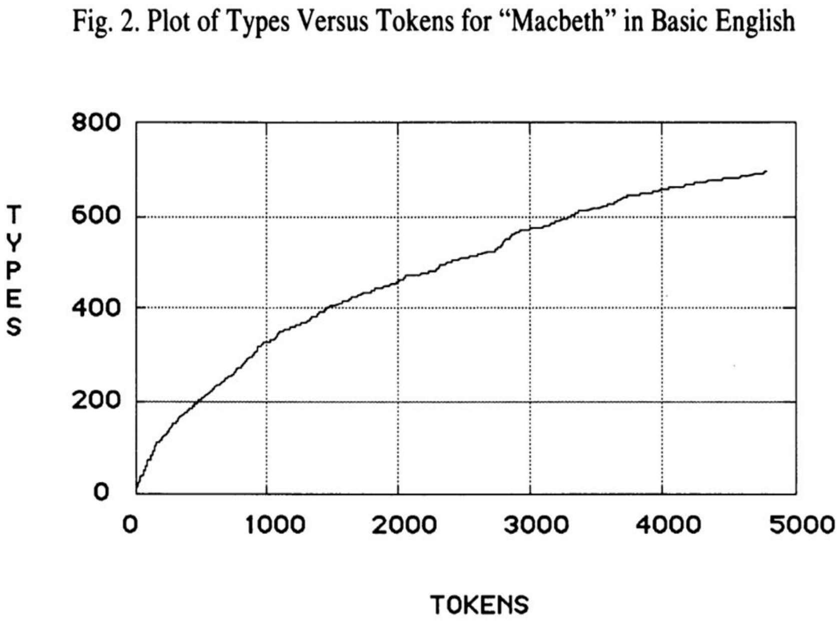


Fig. 4 From Gilbert Youmans, 'Measuring lexical style and competence: The type-token vocabulary curve.' (588)

The same principle of plateauing applies in real-world language that is not artificially constrained to using Ogden's Basic English. The limit in the real world is not Ogden's 850 words but the complete set of words known by the writer, her vocabulary. This plateauing effect is the reason that large canons such as Shakespeare's, Jonson's, and Fletcher's tend to produce overall a lower type/token ratio than small canons. In the large canons, the writers have more fully exhausted their entire vocabulary and are forced to repeat themselves. These large canons have more types than the smaller ones, but not proportionally more. This tapering off of new type deployment allows us to estimate the size of a writer's vocabulary from the shape of this type/token curve. The trick is to extrapolate the curve until it becomes perfectly flat and then read off from the *y* axis the number of types in the vocabulary.

The mathematical calculations for doing this are complex but the principle is straightforward. We do it by observing the rate of change of the slope of the curve as we move along the *x* axis, from steep at the beginning to less steep as we read more of the canon. Each tangent to the curve shows the slope of the curve at a particular point in the *x* axis. The rate at which these tangents slow down their clockwise rotation is constant, so we can predict future tangents at higher *x* values by applying decreasing clockwise rotation, and plot the *y* values that each new projected tangent gives us. When the tangent is horizontal, when the author has used every word she knows, and any further writing can be made only of repetitions of words used before, the *y* value is the writer's vocabulary size.

In a landmark study of 2011, Hugh Craig produced this kind of curve for Shakespeare and for 12 of his contemporary playwrights. What we have depicted as moving our attention along the *x* axis, taking in more and more writing by the author, was in Craig's study implemented as considering what is added to the type count by each successive new Shakespeare play as Craig added it to the experiment. When comparing what each new Shakespeare play added to the Shakespeare type-count with what each new play by one of the other dramatists added to that dramatist's type-count, Craig found that Shakespeare was in the middle of the pack and entirely average. If we want to know what makes Shakespeare's writing extraordinary, we will not find it in the area of vocabulary richness because in that he is not unusual. Shakespeare seems to use a greater variety of words than his rival dramatists, but that is an illusion caused by his leaving us more writing than they did.

Craig pursued his analysis to consider how often in standard-size chunks of his writing Shakespeare used commonplace words versus rare words. Again, Shakespeare was absolutely like his peers in this regard, not exceptional. Craig measured how often Shakespeare used the 100 most common words, compared to his rival dramatists. Again, Shakespeare came out as utterly ordinary. Indeed, Craig concluded, "If anything his linguistic profile is exceptional in being unusually close to the norm of his time" (Craig 68). After all his careful adjustments and his awareness of the mathematical caveats, Craig was entitled to his metaphor of Shakespeare being "close" to the norm.

* * *

To produce the figures, I have presented here for various dramatists' types/tokens ratios, I used a simple computer program in the language called Python. It takes a typical Humanities scholar about half a day of training to get good enough to write a program like the simple one used here, which is available from the author. This program not only produces the type and token counts, but also prints a frequency table showing for any text how often each of the types it contains appears in that text. This table gives some surprising results. We normally expect the word *the* to be the most frequent in any large body of writing, but in the plays of Chapman and those of Fletcher, the word *and* instead takes that first place in the rank order.

When we work in just one dimension, as along the number line on a ruler, the notion of the distance between two numbers resolves, without complication, to the absolute value of the result of subtracting one of the numbers from the other, in either order. But the notion of distance starts to get more complicated when we work in two dimensions instead of one, as when we simultaneously count two features of their writing instead of one. In Figure Five, the counts for *the* and *and* are used to plot points on an *x*/*y* scatterplot.

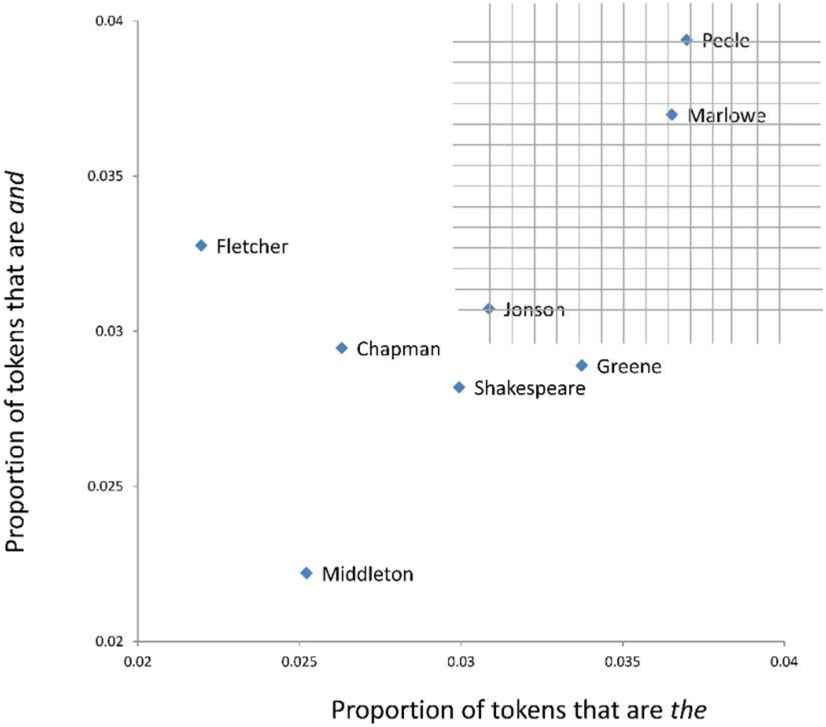


Fig. 5. "Frequency of use of the words *the* and *and* by Shakespeare and seven of his contemporary dramatists"

In Figure Five, each dot represents a corpus of plays by a different dramatist and each dot has a position in the picture that represents two numbers. How far the dot is along the *x* axis shows what proportion of that dramatist's tokens are the word *the*. How far the dot is along the *y* axis shows what proportion of the dramatist's tokens are the word *and*. So, we can read off from Fletcher's dot that 0.022 (that is, 2.2%) of his tokens are the word *the* and 0.033 (that is, 3.3%) of his tokens are the word *and*. Any dot in the top left corner of the scatterplot has many more *and*s than *the*s and any dot in the bottom right corner has more *the*s than *and*s. The dots for Peele and

Marlowe in the top-right corner show that these two dramatists use *the* and *and* at rates that are higher than the rates shown by the other dramatists.

We can see that Peele and Marlowe in the top-right corner are far from the other dramatists. But how far are they, precisely, from, say, Jonson? The answer might seem simple to calculate: we draw a line directly from the Peele dot to the Jonson dot and measure its length, and then do the same for the distance from the Marlowe dot to the Jonson dot. This as-the-crow-flies measurement is called the Euclidean distance. But another way to measure the same thing is to imagine how a taxi driver might make the journey from Jonson to Marlowe or Peele if she had to drive along roads laid out in a grid of city blocks. So long as the driver does not overshoot the destination in either the north-south or east-west direction, all the different routes have the same total length: 17 city blocks for Marlowe and 22 city blocks for Peele. Named after a city famed for its grid layout, this measure is known as Manhattan distance.

The Euclidean and Manhattan measurements give different distances for the same journeys. In this example, they at least agree that the Marlowe data point is nearer to the Jonson data point than the Peele data point is. But it is possible for the Euclidean and Manhattan measurements to give different answers about which of two points is the one nearer to a third point. Consider the three points A, B, and C shown in Figure Six. Which of B and C is nearer to A? By Manhattan Distance, the drive from A to B is 10 city blocks south followed by 28 city blocks west for a total distance of 38 blocks, while the drive from A to C is 20 blocks east and 20 blocks north for a total of 40 blocks, so B is nearer to A than C is. But when we calculate the Euclidean Distance as the crow flies, using Pythagoras's theorem for right-angled triangles, we find that C is nearer to A than B is. Both answers are correct and they disagree because they are based on differing notions of distance. People generally find the Euclidean distance to be more intuitively right, but no one would dispute a taxi fare on the grounds that Manhattan distance is actually wrong.

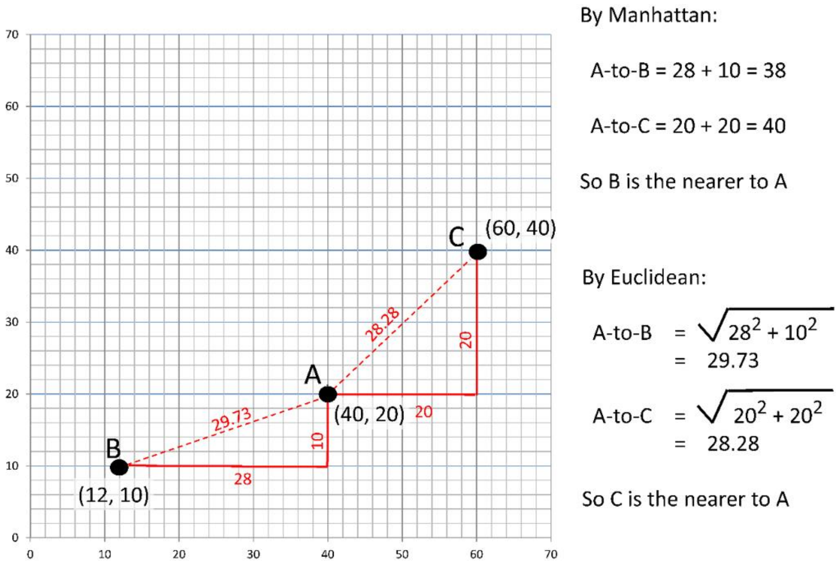


Fig. 6. When Euclidean and Manhattan Distance differ.

Yet a third way to measure the distance between these authors' habits regarding the use of *the* and *and*, shown in Figures Five and Seven, is to say that we do not care about how often the words are used overall but rather we care about the relative preferences for one of these words over the other. Fletcher clearly prefers *and* over *the*, using more *and*s. Middleton clearly prefers *the* over *and*, using more *the*s. And Shakespeare falls somewhere between Fletcher and Middleton, but nearer to Middleton. To represent these preferences, we can use not the Cartesian coordinates of the data points but the angles between lines drawn from the data points to the origin of the scatterplot. These angles, used in a measure called Cosine Distance, are shown in Figure Seven. (Strictly speaking, for measuring these angles the origin of this scatterplot should be x=0 and y=0 not x=0.02 and y=0.02, as shown here; but for illustrative purposes it was desirable to reuse a single diagram across multiple distance measures.)

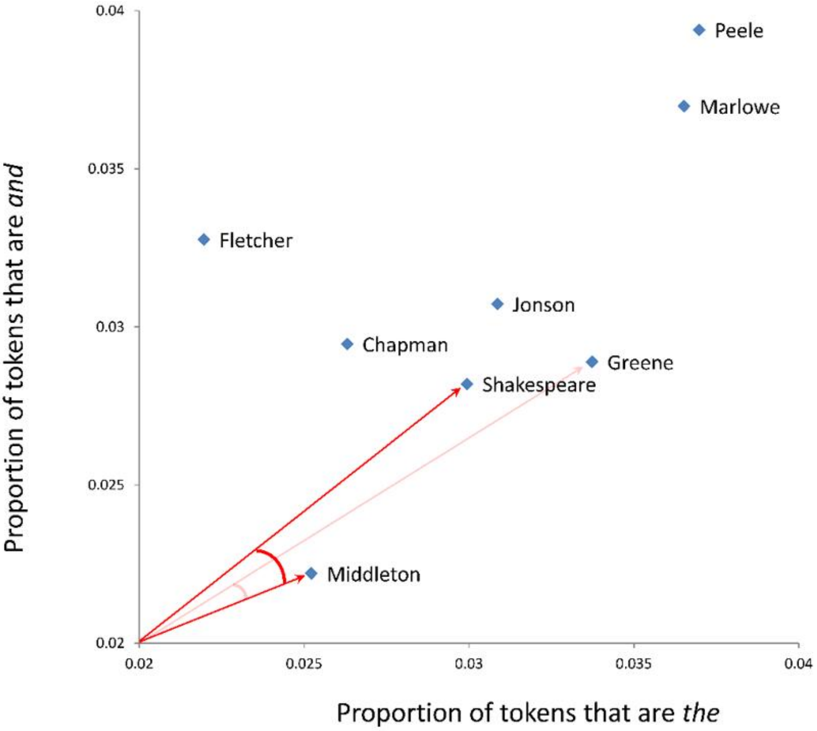


Fig. 7. Cosine Distance disagreeing with Euclidean Distance and Manhattan Distance.

Cosine Distance can easily disagree with Euclidean and Manhattan Distance about how different two writers' styles are. Consider in Figure Seven the Cosine Distance from Greene to Middleton, which is smaller than the Cosine Distance from Shakespeare to Middleton, although by Euclidean and Manhattan Distance the Shakespeare data point is closer to the Middleton data point than the Greene data point is. What this tells us is that although Greene uses many more *the*s and *and*s than Middleton does, using them about as liberally as Shakespeare does, Greene's strong preference for *the* over *and* is much like Middleton's strong preference for *the* over *and* and is different from Shakespeare's habit which only slightly prefers *the* over *and*. This shows us

Common Methodological Error #3: assuming that all distance measures will agree on how far apart are the data points that we derive by counting features of various writers' writings. In fact, different distance measures can give us different answers about whose writing styles are most alike.

The importance of our choice of distance measure increases as we use more dimensions. If we count not only the occurrences of *the* and *and* but also the occurrences of the verb *to be* in the canons of multiple authors, we end up with three data values for each canon. We can visualize in a three-dimensional scatterplot the results of such counting, but when we show such a three-dimensional plot on a two-dimensional plane (such as a page of a book or an image on a flat computer screen) it becomes impossible to read off the values for each data point or even to see which data point is nearest to which others. Although at three dimensions such visualizations cease to help human eyes discern the patterns we are trying to discover (unless we use specialized three-dimensional visualization equipment such as the still-rare goggles), the underlying mathematics of Euclidean, Manhattan, and Cosine Distance still work and their values can be calculated.

We need not stop at counting three features of a text and treating our counts as the coordinates of points in three-dimensional space. We might count occurrences of the 100 most-common words in a set of texts and treat the resulting numbers as coordinates in 100-dimensional space. Even at this level of abstraction from everyday reality, the various means of measuring distance still work in principle—at least the various mathematical formulas still give us apparently meaningful results—but unexpected complications make the results potentially deceptive.

In experiments with multi-dimensional data, we often want to ask just which of several clouds of data points, each derived from one author's works, is the cloud from which a new data point is least distant. The nearest cloud will, we expect, be a set of data points for the writing that is closest in style to that of the writing that generated the new data point. Unfortunately, as we increase the number of dimensions, something odd happens to the data points: they spread out so that the kind of clustering by author that we see in lower dimensions ceases to appear.

To understand why this happens, let us return to the simple number line we began with. In the top-left corner of Figure Eight, we see that in a one-dimensional universe—corresponding to measuring just one feature in our texts, such as the frequency of one word—there are 10 possible places that a data point can fall, and hence we need only 10 data points to fill that universe. The top-right corner of Figure Eight conveys that if we add one more dimension to make a two-dimensional universe—corresponding to measuring the frequencies of, say, *the* and *and*—then we need 100 data points to fill the 100 places in that universe. (For the sake of illustration, we are thinking here of all our counts falling along a fixed range of integers, but the problem we are approaching also applies if our numbers include decimal fractions.)

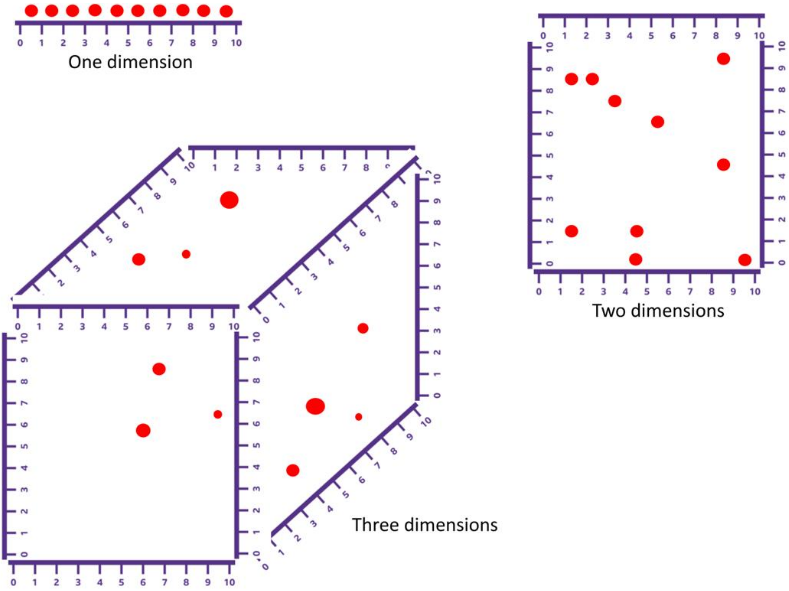


Fig. 8. The curse of dimensionality.

If in the two-dimensional universe represented in the top-right corner of Figure Eight we have only 10 data points from our experiments－our counts of the features in texts－then those data points will be more widely spaced out than they were in one dimension. If we add a third dimension (see the bottom-left corner of Figure Eight) then we need 1000 data points to fill the space. And if we have only 10 data points from our experiments, then they will be even more widely distant from one another. Every new dimension multiplies by 10 the number of data points we need to fill the space and if we have only a few data points they become ever more widely separated. At 100 dimensions, which is the space we are using if we count the 100 most-commonly-used words, as we often do in authorship studies, we need an extraordinary number of data points to fill the space. We need this many: 1,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000. For a sense of perspective, this is greater than the number of atoms in the known universe. Even if our experiments give us tens of billions of data points－which rarely happens－these data points will be hugely distant from one another in 100-dimensional space. The notion of nearness－the notion of how far Shakespeare's writing is from that of his contemporaries that we started with － stops making sense when our data points are so far apart.

This problem is known as the Curse of Dimensionality and it is a significant obstacle across data science. Just measuring more features in language does not necessarily give us more knowledge, because our distance measures are less discriminating as we move into the higher dimensions. This is Common Methodological Error #4: thinking that the more features of writers' writings that we count, the more data we derive, the more knowledge we have about that writing. In important ways, counting more gives us less knowledge. The problem of measuring distance does not affect all distance measures equally. Euclidean, Manhattan, and Cosine Distance, and various more esoteric measures, are more or less discriminating of authorship depending on precisely what we are measuring and how many dimensions our data have. Unfortunately, most published papers on computational analysis of writing style pay little or no attention to the choice of distance measure. Investigators typically accept the default distance measures provided in such software packages as the popular *R Stylo* and neglect to consider how the choice of distance measure affects their results.

* * *

What can we conclude from the problems described above? The metaphor of distance has no direct application to the comparison of texts. Once we start counting features of texts, the notion of distance has some validity, but it is not simply a matter of subtracting one number from another. Sometimes division rather than subtraction gives the more meaningful sense of distance. As readers of publications that apply numerical methods to the analysis of writing, we must always ask ourselves if we agree that a meaningful notion of distance is being used by the investigator. At its simplest, for one-dimensional comparisons, this comes down to asking whether absolute differences (from the mathematical operation of subtraction) or relative proportions (from the mathematical operation of division) are the more meaningful. When the operation of division is applied to data in order to put them on the same scale —as when adjusting for different canon sizes—we must always ask whether the relationships we are concerned with are truly linear; if not, a simple division will distort them.

When we create multi-dimensional data from language, as when we simultaneously count the frequencies of occurrence of two or more words, a new problem emerges from the multiple ways of counting distance in multi-dimensional space. No one method is inherently more correct than another, so as readers we should check whether the investigators have shown an awareness that selecting a distance-measuring method is an active choice that can change experimental outcomes. Ideally, investigators should conduct multiple experimental validation runs using different distance-measuring methods in order to firmly establish that for the particular element of style they wish to pursue, or discrimination they want to make, they have found the distance-measuring method that gives their experiments the greatest power. (Discriminatory power is a quantifiable value from statistics that should be stated for each experiment in any quantitative investigation of writing style.)

Finally, our ability to generate more and more numbers from texts should not lead us to conclude that we are gaining more and more information about them, on account of the Curse of Dimensionality. It is a human foible to be impressed by large numbers, so that studies counting many features across many texts seem to our intuition more likely to produce reliable results than studies counting fewer features in fewer texts. Having a lot of data is, in general, better than having only a little. The problem emerges when we treat the features that we are counting as if they are dimensions in multi-dimensional space and then try to measure the distances between data points. Studies using these methods should show that they have collected sufficient data points to substantially fill their multi-dimensional spaces. Readers should look out for discussions that show investigators being aware of the problem of sparseness that can occur when studies collecting large amounts of data treat it in this way.

▸Key Words: quantitative language analysis, computational stylistics, linguistic distance, style measurement methodology, dimensionality

Works Cited

Brown, Paul. *Play Word Counts*. A contribution to the website *Shakespeare's Early Editions: Computational Methods for Textual Studies*, hosted by De Montfort University, and funded by the Arts and Humanities Research Council from grant AH/N007654/1, 2018.

Craig, Hugh. "Shakespeare's Vocabulary: Myth and Reality." *Shakespeare Quarterly*, vol. 62, no. 1, 2011, pp. 53-74.

Lakoff, George, and Mark Turner. *Metaphors We Live By*. U of Chicago P, 1980.

Taylor, Gary. "Did Shakespeare Write *the Spanish Tragedy* Additions?" *The New Oxford Shakespeare Authorship Companion*, edited by Taylor, and Gabriel Egan, Oxford UP, 2017, pp. 246-60.

Youmans, Gilbert. "Measuring Lexical Style and Competence: The Type-token Vocabulary Curve." *Style*, 1990, pp. 584-99.

Some Problems in Using Numbers to Represent the Writing Styles of Shakespeare and His Contemporaries

Abstract

Gabriel Egan

The quantitative study of writing styles—sometimes called stylometry or computational stylistics—has in the past two decades been enhanced by the widespread availability of large digital textual corpora and easy-to-use software tools that lower the technical obstacles for participation in this field. For the study of early modern drama, the availability of the raw text datasets called ProQuest *One Literature* (formerly *Literature Online* (*LION*)) and *Early English Books Online* (*EEBO*) makes it easy to compare Shakespeare's writing with that of his contemporaries. The result has been a boom in quantitative studies of early modern drama. Certain aspects of language, such as authorial preferences for particular words and phrases, are especially easy to quantify. But there are problems attendant on the quantitative analysis of language that are easily overlooked because language is a more complex subject than it first appears. This essay surveys four kinds of problems that can distort our perspective when we start using numbers to represent writing styles.

Gabriel Egan (sole author)
Professor of Shakespeare Studies, De Montfort University
The Gateway, Leicester, LE1 9BH UK
gegan@dmu.ac.uk

Submitted: 30 June 2024
Review Completed: 31 August 2024
Accepted: 31 October 2024