Martin Mueller on Brian Vickers and the Kyd canon

DATA

DIGITALLY ASSISTED TEXT ANALYSIS

Home › Blogs › Martin Mueller's blog

(1) N-grams and the Kyd canon: a crude test

Martin Mueller — Tue, 08/18/2009 – 15:09

Last year Sir Brian Vickers argued in *TLS* (April 18, 2008) that we should think of Thomas Kyd not only as the author of *The Spanish Tragedy* but as an author with a substantial oeuvre that in addition to his translation of Garnier's *Cornelia* and the Turkish tragedy *Soliman and Perseda* includes *Arden of Faversham*, *The Chronicle of King Leir*, *Fair Em the Miller's daughter*, and substantial parts of *1Henry VI.*

Vickers based his argument on trigrams that occurred in known works by Kyd and only in the plays that he would like to include in the Kyd corpus. He is not the first to argue for Kyd as the author of those plays, and in fact his goal is to use the modern tool of plagiarism detection software, in particular Pl@giarism, to produce more conclusive evidence for attributions that various early 20th century scholars made independently and on different grounds.

I found the argument attractive and largely persuasive. At the same time I was troubled by the loose use of probability arguments. Vickers relies heavily on subjective judgments of the kind that if 'thrust his sickle in' occurs only in *Soliman* and *Arden*, this is unlikely to be a coincidence. He knows that it takes more than one swallow to make a summer, and he has found a lot of swallows, but he has no guide beyond common sense to determine how many swallows are enough.

My mother liked to say that three hairs on your head are relatively few but three hairs in the soup are relatively many. Informal statistical judgments may rest on strong practical foundations, and the judgment of an experienced reader is not to be discarded lightly. But can we create a quantitative context that gives even more authority to such judgments?

With that question in mind, I ran an experiment on 318 early modern plays in the MONK corpus. These plays come from the TCP-EEBO texts. They include most of the surviving plays of the playwrights who wrote or began writing before 1642, when the playhouses closed for almost two decades.

From those texts, linguistically annotated with Phil Burns's MorphAdorner, I extracted lemma n-grams from bigrams to heptagrams that were repeated at least once. There are about 1.1 million of them, and they add up to 8.7 million occurrences. With a little patience, you can manage such data quite comfortably in Microsoft Access. My counts exclude lemma strings that only occur as part of longer strings as in thrust his sickle in

his sickle in

sickle in

In this experiment I pay no attention to the length or content of n-grams, but simply count their distribution across plays. What is the distribution of repeated n-grams that are restricted to one play? What is the distribution of n-grams that are shared between two plays but occur nowhere else? Since plays may differ in their length by a factor of two, it is helpful to express frequencies as occurrences per 10,000 words. For n-grams that are shared between plays, the word count is the composite count of words in the two plays

Here are the descriptive statistics for one-play n-grams:

mean	stdv	min	q1	med	q3	max
61.45	37.15	7.82	36.27	53.37	77.57	269.2

Here, by contrast are the descriptive statistics for n-grams that occur in two plays (there are 49,593 combinations of two plays that share at least one n-gram):

mean	stdv	min	q1	med	q3	max
2.14	1.37	0.17	1.33	1.96	2.68	63.3

On average, one-play n-grams will exceed by a factor of 30 n-grams shared exclusively with some other play. The difference between the interquartile ranges makes that point even more emphatically. Unsurprising as these figures are, they provide a firm background for making sense of outliers. They also suggest that a relatively small number of shared n-grams can have considerable evidentiary value.

Before looking at the outliers in detail, let us look at the distribution of n-grams that are shared by two plays of the same author. There are 2303 two-play combinations. 'Same' has to be taken with a grain of salt, because there are quite a few plays that cannot be assigned unambiguously to a single author.

mean	stdv	min	q1	med	q3	max
4.34	3.33	0.21	2.63	3.73	5.15	61.9

The interquartile range is quite telling here, and clearly shows that two plays by the same author may be expected to share about twice as many unique n-grams as two plays by different authors.

If we now look at the outliers among two-play shared n-grams, the top of the list is uninteresting in that it consists mainly of two-part plays, such Killigrew's *Cicilia and Clorinda*, or the two parts of Heywood's *Iron Age*. *Richard III* and the three parts of *Henry VI* are also at the top. What else do you expect? There are 110 two-play combinations with relative frequencies between 10 and 63, well above the 99th percentile for two-play combinations but with three exceptions well below the median value for one-play n-grams.

How does Vickers's proposed Kyd canon fit into the framework of expectations raised by the distribution of two-play n-grams? There are 28 two-play combinations of Vickers's seven plays, plus *The first part of Hieronimo* (1602), which I include because of its obvious status as a 'prequel', whoever wrote it.

Of these 28 combinations, six place in the top quartile for shared two-play n-grams by the same author (figures in parentheses are frequencies per 10K and percentile rank for two-play shared n-grams):

1. Spanish/Hieronimo (15.4, 99.9%)
2. Spanish/Soliman (13.3, 99.9%)
3. Soliman/Arden (9.2, 99.7%)
4. Leir/Arden (6.2, 99%)
5. Leir/Em (5.2, 98%).

Another three place above the median:

1. Leir/Soliman (4.5, 96.5%)
2. Spanish/1Henry VI (4.3, 96%)
3. Cornelia/Soliman (3.85, 93.5%).

These tests are extremely crude and pay no attention to the salience of particular phrases, especially longer ones. But read properly, they have some subtlety of their own. On the surface, it seems that *The Spanish Tragedy* is most like *The First Part of Hieronimo*. But when you observe from the ranking of the 28 pairs that *Hieronimo* is least like *Cornelia* and *Soliman and Perseda* you conclude that the resemblances between *Hieronimo* and *The Spanish Tragedy* probably result from continuities in subject matter rather than characteristic patterns of authorial usage.

On balance, my figures lend support to Vickers's argument, although they are not conclusive. There are many plays by different authors that share more n-grams than the plays in the putative Kyd canon. On the other hand, if you play a ranking game with each play in the Kyd canon and list the plays with which it shares the most n-grams, some of the other "Kyd" plays will appear in the top five. Something is going on here. More work seems worth doing, and it appears to me that in the increasingly digital philology of the future readerly observation will go hand in hand with quantitative routines, ranging from the simple counting procedures used here to more complex algorithms.

DATA

DIGITALLY ASSISTED TEXT ANALYSIS

Home › Blogs › Martin Mueller's blog

(2) Vickers is right about Kyd

Martin Mueller — Sun, 08/23/2009 - 09:36

In a previous blog I discussed Sir Brian Vickers's argument that the old Leir play, *Arden of Faversham*, *Fair Em*, and large chunks of *1 Henry VI* should be attributed to Thomas Kyd. Based on some tests with common trigrams I think they provide compelling corroborative evidence that Vickers is right about the *Leir* play, *Fair Em*, and *Arden*. Authorship arguments provide huge yawns in English departments. But, as Vickers argued in his *TLS* piece, there is good reason to look with interest at an expanded Kyd canon. Shakespeare and Marlowe suddenly acquire a slightly older and gifted contemporary whose oeuvre has some size and considerable thematic and generic range.

What does it matter whether the author of *Fair Em* is Kyd or "anon."? It is the same play after all. Or is it? At the minimum, a play by a known author receives a lot more attention. I recently used Papers, a software program for managing and searching pdf files, to look for mentions of *Lust's Dominion* in JStor. There were some fifty. Then I looked for mentions of *Titus Andronicus*. Papers crunched away, and after 30 seconds its brilliantly simple interface showed well of 2,000 hits. Is *Titus Andronicus* forty times as interesting as *Lust's Dominion*?

To ask the question is to recognize the attention a text gets from being part of a canon. If we can establish a Kyd canon, the prospects are good that some bright graduate students will do interesting work based on the assumption that half a dozen plays from the late 1580's and early nineties are the work of a single author.

My corroborative evidence comes from applying discriminant analysis to lemma trigrams that occur at least 500 times in 318 early modern plays. There are 56 of them, and they range from “I will_not” (2332) to “what_do_you” (508). Riveting fragments of speech, but remarkable informative when you look at their distribution.

How does discriminant analysis work?

Skip this section if you already know or don't want to know about the conceptual underpinnings of discriminant analysis. The math is well beyond me.

Discriminant analysis is a form of multivariate analysis, which is obviously different from univariate analysis. ANOVA or Analysis of Variance is a well-known form of the latter, and Fisher's lilies are a good example of it. Ronald Fisher, an influential statistician of the early twentieth century, measured the length of the petals in different patches of lilies, and developed a formula for determining whether the variance within each group was smaller than the variance between groups.

In multivariate analysis you try to establish variance between groups on the basis of the combined effect of multiple variables. In discriminant analysis, these multiple comparisons involved take a particular form. You have a set of items --whether lilies or plays-- and you 'classify' or assign them to different groups, whether a particular lily patch or author. Because of this initial assignment of each item to a group by a human, this type of technique is known as 'supervised classification'. You then provide the program with data about the frequency of the different variables in each of your items.

The statistical algorithm now crunches through your data and assigns each item to a group with a 'confidence value'. Think of it as a conversation in which the program says: "You said that *Poetaster* was by Jonson, and I think there's a 99% chance you're right." Or "You said that *The Massacre at Paris* is by Marlowe, but I think there's only a 5% chance that this is right. There's a 42% chance that it's by John Bale and a 35% chance that it's by Thomas Kyd."

It should be obvious from this sketchy introduction that for a literary critic discriminant analysis is interesting as much for the disagreements as the agreements it produces. For instance, if you use discriminant analysis to distinguish between Shakespeare's tragedies and comedies you are very likely to find that the algorithm 'misclassifies' *Othello* as a comedy. This begins to make a lot of sense if you remember an old essay by Douglas Stewart about *Othello* as a Roman comedy turned nightmare.

Discriminant analysis, lemma trigrams, and the Kyd corpus

Vickers made his argument for an expanded Kyd corpus on the basis of shared rare repetitions. In the old *Leir* play, *Fair Em*, and *Arden*, there are a lot of phrases that occur in the *Spanish Tragedy*, *Soliman and Perseda*, and the translation of *Cornelia*, but nowhere else. My test ignores this evidence and looks instead at the most common trigrams, which show up in at least half (164) or more than 90% (297) of all plays. Vickers's and my conclusions therefore rest on an entirely different evidentiary basis. To the extent that we agree, the case is greatly strengthened: the evidence of rare and of common phenomena support each other.

Discriminant Analysis misclassifies 50 or 16% of 318 plays. It gets 84% right. Of 37 plays by Shakespeare, it gets 34 right, 24 of them with a confidence value of over 90%. The misclassifications are instructive. Discriminant analysis is uncertain about *1 Henry VI* and assigns it to Heywood (44.5%) or Dekker (24%) and gives it only a 4% chance of being by Shakespeare, as opposed to the second and third parts of *Henry VI*, which get a 91% confidence score as Shakespearean.

Discriminant analysis thinks that there is a 92% chance that John Lyly wrote *Love's Labor's Lost* as opposed to an 8% chance for Shakespeare. It is also puzzled by *Pericles*, giving it only a 5% chance of being Shakespearean, and assigning it tentatively to Jonson (44%) or Kyd (21%).

You don't need to be much of a Shakespearean to recognize that these are not random errors. They pick out two plays whose authorship has often been contested, and the assignment of *Love's Labour's Lost* to Lyly's Euphuistic world is a stroke of statistical brilliance in its own right.

Here are the scores for the correct assignments for other playwrights:

- Beaumont: 7/7
- Chapman: 10/14
- Fletcher: 9/13
- Ford: 8/8
- Goffe: 4/4
- Heywood: 16/22
- Jonson: 14/20
- Killigrew: 10/11
- Lyly:7/7
- Marlowe: 5/6
- Marston: 5/7
- Massinger: 11/14
- Middleton: 13/16
- Shirley: 31/32
- Marston: 6/8
- Webster: 4/5

For the purposes of my analysis each of the plays has been assigned to a primary author, which may be questionable with heavily collaborative authors like Dekker or Middleton. But the results are fairly consistent. The more clearly an author is established as a single author, the better Discriminant Analysis works. Jonson is a special case, because his corpus includes masques or other texts that are plays only by a stretch.

Now to Kyd. I assigned to Kyd all plays that Vickers assigns to him, except for *1Henry VI*, because Vickers assigns parts of it to Shakespeare. And I added the 'prequel', *The first part of Jeronimo* because it shares a lot of rare repetitions with the *Spanish Tragedy*, although it shares relatively few rare repetitions with the other plays in the Kyd canon. Here are the results:

- *The first part of Jeronimo* (30%)
- *Cornelia* (79.7%)
- *Soliman and Perseda* (85.3%)
- *The Spanish Tragedy* (96.1%)
- *Arden of Faversham* (97.4%)
- *The true chronicle history of King Leir* (99.3%)
- *Fair Em* (99.5%)

These are striking results. First, Discriminant Analysis rejects the prequel as Kyd's. It assigns it to the grab bag of anonymous plays with a 57.4% chance. So it is not fooled by the presence of many shared repetitions between it and *The Spanish Tragedy*. Secondly, Discriminant Analysis very strongly confirms that the other plays come from the same stable. Indeed, if *The Spanish Tragedy* is the clearest case of a play by Kyd, the three English plays are, so to speak, a little more Catholic than the Pope.

If you combine my evidence from common trigrams with Vickers's evidence from rare shared repetitions, you would have to be very skeptical about the power of quantitative analysis not to acknowledge the fact that the claim for an expanded Kyd canon rests on quite solid evidence.

What about *1 Henry VI*, one of the three Shakespeare plays about which Discriminant Analysis has doubts? A good test here might be to separate the scenes that Vickers argues belong to Kyd, treat Kyd's and Shakespeare's scenes as separate plays, and see how they line up when compared with the Kyd canon and Shakespeare's plays before 1593. I will do that test at some point.

It would also be useful to run a test with part-of-speech trigrams -- sequences like 'determiner-adjective-noun' or 'past modal - have - past participle'. These have considerable discriminating power as Harald Baayen showed in *Analyzing Linguistic Data*. And I know from my experiments that they discriminate sharply between Shakespeare's comedies and tragedies.