RSA 2022, Scholarly Editing in the Digital Age
Mel Evans and Elaine Hobby, Editing Aphra Behn in the Digital Age and Investigating Attribution

In 2016 the General Editors of The Cambridge Edition of the Works of Aphra Behn were awarded major funding by the UK Arts and Humanities Research Council to support the initial stages of our work. Our goal was not a digital edition: we were committed to print publication of the 8-volume set (each volume running to c. 1000 pages). We had though particular digital enhancements and extensions to the edition in mind, and were convinced that electronic means – including the establishment of a digital Workbench where our twenty editors from around the world might collaborate – would form an essential dimension of our working process. I aim here to give a flavour of what happened – including to the Workbench – and then to give the part of this paper that Mel Evans has written, which outlines our digital work on attribution.

slide 2: Today's paper is jointly written with my fellow editor and General Editor Mel Evans. The names of other members of the core team are on the slide: the work of all of them is essential to the edition every day. The collaborative nature of the paper is typical of how our project works. One of its great joys to me has been the constant sharing of ideas. Originally Mel was to have been here in person and giving a whole second paper. I'm very grateful to Cathy for agreeing, late in the day, to my trying to speak for us both. And of course: if you have any properly technical questions about Mel's work on attribution, she is the person to ask, and she's happy to respond to enquiries and comments

slide 3, In the beginning: I did my first editing on an Amstrad with no internet – no WYSIWYG

- no email exchanges
- no online searches
- no digital copies of anything

slide 4, EEBO and a Digital Workbench were the key components of what we planned from the start

- EEBO is a fantastic resource
    - making collation of our kind possible
    - making it easy to search for echoes and sources (and finding e.g. evidence of *The Decameron* and other Behn reading in *The Rover*)
    - also its dangers, e.g. the EEBO copy might well be an uncorrected state, or misidentified, or otherwise problematic
    - and people use EEBO stupidly more generally –
        - not thinking about the religious alignments of works that happen to show up in searches
        - the really useful edition of a work might not exist on EEBO full-text
        - shortcomings in the TCP version leads to omissions
- And a Digital Workbench where editors would work, tagging their text as they went, ensuring consistency, having access to one another's work-in-progress
    - what happened to the workbench

slide 5, high skills of the CUP's digital printers making it possible for us

- to have text, commentary notes and textual notes all on the same page,
- complex layout for songs
    - stage directions
- to key textual notes by line-number at the first-proof stage, thereby reducing clutter on the page

slide 6, and Behn wrote a lot – thus the size of the team

- items in red are in The Cambridge Edition even though they might not be by Behn
- about to decide how to signal that questionable attribution in Volume II
- Volume IV is out – and is passing around the room

slide 7, digital photography as a fantastic resource

- e.g. to record what we find
- and for collators to communicate with editors
- but also revealing things that I failed to see with my human eye
    - such as Anne Brandon's signature on the Princeton copy of *The Emperor of the Moon*, which had almost been erased
    - Anne Brandon – famous divorcee and, the successful 18c writer Richard Savage claimed, his mother – owned this copy
    - digital catalogues and photography made it easy to get a comparison signature from a personal letter of hers held at Wiltshire Record Office
        - this is our work-in-progress; please respect that

So, our goal is to make new research possible, and there are many dimensions to how the digital world has made editing Behn possible

And over to Mel, **slide 9**

Digital Perspectives: Authorship, Attribution and Editing Behn

1. Questions and Problems for the Edition – what are the dubia?

In an edition of the size and nature of a writer such as Behn, there are inevitably questions about the authorship, reliability and reputation of the works associated with that author. Behn's works span a period of two decades and represent a diverse range of genres, and within this polymorphous set are plays, poems, and prose fictions – among others – that have acquired an association with Behn, for better and for worse.

Authorship attribution is not always a hard-and-closed affair, and for the edition we felt it was important to address and acknowledge the challenges presented by Behn's 'traditional' canon of works. In brief, as shown on the slide [slide 10], we have a set of securely attributed literary works – secure in the sense that they were published in her lifetime with her name on the title page. The second column indicates, however, the quantity of material that doesn't meet this criteria. Some of it is well-known, such as the three-part epistolary novel *Love Letters between a Nobleman and his sister* which was recently called into question by Leah Orr. Other works [slide 11], such as the play *The Counterfeit Bridegroom*, which is an adaptation of Middleton's 1611 *No Wit No Help Like a Woman's*, has a less consistent association with Behn. The theatre historian John Genest makes the connection in his 1832 publication *Some Account* (see Challinor, forthcoming), on the basis of it being 'so much improved' over Middleton's original, although it is not quite clear why that should definitely make it Behn's responsibility! Other dubia questions surround the set of short prose works [slide 12] published after Behn's death in 1689, which were purportedly discovered by the bookseller, Samuel Briscoe, and appear for the first time in collections of 1698 and 1700, collecting stories of pirates and shipwrecks, thwarted suitors, incest, and a good dose of melodrama [slide 13]. The verse corpus represents a substantial challenge, with their editor, Gillian Wright, ranking about half of the extant works on a scale of 'moderately dubious' to 'very unlikely'. And, finally, Behn's holograph correspondence, raises its own questions, Nadine Akkerman recently proposing that whilst Behn's letters are secure, it may be that Behn in fact fabricated the contents of those letters she purportedly copied out from the originals of her informant, William Scott.

2. What digital methods and approaches are available to address these questions and problems?

[slide 14] So what's an edition to do? Our approach to exploring the likelihood to Behn's involvement with the array of dubia was to draw on the many and various developments in computational approaches to authorship in order to see what light they could provide. Now we appreciate that computational attribution is often an antagonistic area of scholarship, but our approach has always been to do best by Behn. As it happens, computational stylistics and stylometric methods appear to work best when bespoke, tailored to the eccentricies and limitations of the materials under analysis. We have sought to develop an approach that can capture the key stylistic traits of Behn's literary works, and then compare these with the characteristics of the dubia. If enough of the findings point in the same direction (either for or against), then we feel able to say something confident about Behn's likely involvement. Whether or not the results indicate Behn's authorship will not necessarily affect its inclusion in the edition – works like *Love Letters* are an important association with Behn's legacy as an author and should continue to be viewed as such.

The kinds of approaches we've been working with as those that fall under the remit of computational stylistics and under stylometry. These techniques, developed by scholars such as John Burrows, Hugh Craig, Brian Vickers, David Hoover and others, work on the premise that authorial markers are embedded within (literary) texts at a level of patterning not easily observed by the human eye, but which are readily captured by the computer and its capacity to count. Burrows, Craig, Greatly-Hirsch and others have show that word frequency is a reliable marker of one author's style versus another, and that this can be used to assess the potential authorial provenance of a questionable text. Even most common words – the grammatical small words that comprise the glue of language - are not used to the same extent and in the same ways by different authors, even those working in the same genre at the same point in time.

Other techniques have looked at larger chunks of language – repetitive 2- or 3-word sequences (known as n-grams or lexical bundles) – noting how certain sequences may recur more frequently in one author's corpus than another.
The implementation of these frequency counts can be either descriptive – shaking up the results to see how they pattern and distribute – or as a classifier – train an algorithm based on the frequency profile of a given set of texts and then see what it does with those texts it hasn't seen before.

One thing we've found important to address in developing these methods for Behn's works and Restoration literature more broadly is to acknowledge the specific permutations of literary writing in the period. Thus, there are some characteristics of Restoration genre that shift over the period (the move towards a more sentimental dialogic style), and certain topics and tropes (nuns, bed tricks) that reoccur across different works. The kinds of methods devised for Shakespearean drama, say, or for 19th century literature and Henry James, are not necessarily going to be cut-and-pasted effectively into the chaotic realm of Restoration literature.

3. Shifting our expectations and perspectives: questions of style, as well as authorship

Our initial forays into Behn's authorship was very authorship focussed: looking for clear-cut answers in terms of whether a dubious text was like, or not like, Behn. However, it became rapidly clear that this perspective was unsatisfactory from our perspective as editors, and as literary and language scholars of the early modern period. To that end, our application of the digital tools became more about style: Behn's style – as a whole, and within and across genres – and the style of her contemporaries, her period. Making this the focus of much of our investigations has allowed us to respond flexibly to the curiosities and challenges of the material (some examples of which we'll get to), and to take a more 'data driven' approach that responds to what the evidence *is* rather than, perhaps, what we would like it to be. This doesn't always yield the answers we're after (or any answers, necessarily, at all) but has shaped our approach and raised some very interesting questions about what digital approaches to style can do, and what it can't.

4. What are the challenges and opportunities with these methods and approaches?
    a. Scholarly scepticism/divergent perspectives, motives and values

As you are likely aware, there is a tendency for some scholars in areas of early modern studies to get cross, cynical and at times confrontational about digital approaches to style and authorship. Some of you here today may indeed fall into the 'sceptics' camp, feeling that the computational evidence is unconvincing and unnecessary. It's fair to say that, at times, those of us involved with the digital linguistic analysis of Behn's works have felt similarly cross, cynical and confrontational (sometimes, all at once). However, one of the objectives of the edition was to use the available knowledge and resources to produce a fully rounded perspective on Behn as a writer – and, as our work has progressed, we have indeed found some fruitful features that we may not have identified without recourse to a digital approach.

a. Data availability, data preparation, and bespoke configurations

In order to evaluate the dubia, it is first necessary to have a reliable corpus of Behn's writing: reliable in the sense of secure attribution, but also in terms of its formatting and editorial characteristics. Computers, as we know, do not cope well with spelling or typographic variation. Much of our time on the edition has been spent preparing Behn's texts for analysis: putting into TEI XML, regularising the spelling - an iterative process, it turned out, as questions about whether to preserve 'o' and 'oh' as distinct forms, for instance, took up more time and debate than anyone might anticipate – and generally making clean, machine-readable texts. Because of the nature of the dubia, and our interest in Behn's style more broadly, it was necessary to have comparison corpora: texts by authors other than Behn who can provide a stylistic counterpoint to Behn's writing. These corpora allow us to put Behn in context, but also allow us to examine other, possible authorial candidates. So a lot of time has also been spent in preparing a Restoration drama corpus (68 plays), a Restoration prose fiction corpus (18 texts, and still growing) and a Restoration verse corpus (472 texts, from Ayres to Wycherley. In due course, we hope to make these resources available for other researchers. Whilst early modern studies has led the way, in some respects, in digital text creation, the Restoration has been somewhat neglected compared to the Elizabethan and Jacobean periods, and our work here will hopefully help to address that.

5. What have we learned?

In order to give a sense of the kinds of results the digital investigations have produced, both their strengths and weaknesses, I wish to finish with two short case studies. The first looks at some of the challenges and findings from the analysis of Behn's drama and the dubia. The second, which is still a work in progress, is based on an examination of Behn's prose style.

b. Behn's drama: style developments, likely authorship of *The Revenge* and *Counterfeit Bridegroom*

Our analysis of Behn's works using computational approaches started with the drama: this was because it's a genre looked at extensively within EM attribution studies, providing important theoretical and methodological points for comparison. It's also a genre that spans the greatest temporal range in Behn's career (1670-1690), including 12 comedies, one tragedy and 3 tragi-comedies.

Some initial work, exploring Behn's plays on their own terms, showed that Behn's style develops over the course of her career: there are linguistic markers that distinguish her earlier plays from those later in her career. These developments can be seen to align with broader changes in dramatic fashions and the literary market in the 1670s and 1680s, but they attest to Behn's engagement with the theatrical world and her potential ability to read and anticipate what would be of interest to an audience: a valuable skill

for a professional writer, in any period. The means of assessing these developments focussed firstly on the most frequent words in Behn's plays, documenting the distribution of these frequencies across the individual play texts to see which plays shared similar profiles [slide 15]. Interpreting a descriptive analysis such as this one, using Principal Components Analysis, is not a hard science, and it's about making sense of the patterns that are produced. In this figure, the earliest plays group at the top of the y-axis, mid-career in the middle, and latest plays at the bottom.
Because of this descriptive/interpretative scope for ambiguity, as well as insight, we add other perspectives on Behn's linguistic style – for example, looking at keyword. [slide 16] In corpus linguistics, keywords are words that occur more frequently than they should do in a particular corpus or sub-corpus, based on a comparison (reference) corpus)). Keywords provide both topical and stylistic insights into a corpus, and provide a more 'readable' set of findings about Behn's drama than a most-frequent-wordlist can. When producing keywords for Behn's plays, the results suggest that Behn's dramatic style focussed increasingly on character interaction and interpersonal meaning, and less on character subjectivity. Characters talk more to each other, and less to themselves or the skies. There are also dominant gender themes within the keywords for each sub-period, but the kinds of gender roles and labels shift with Behn's career, from domestic, interior, familial to more hierarchical externally-facing social identities.

With this broad chronological baseline of Behn's dramatic style – of which much more could be said – we've been able to assess the likeness of two dubious works to Behn's profile: *The Counterfeit Bridegroom* – the Middleton adaptation mentioned earlier – and another play, *The Revenge*, which was also first performed in 1680. This play is lodged under Thomas Betterton's name in EEBO/ESTC.

Starting with *The Revenge*, this play is something of a success story for our attribution investigations. We're in the process of writing up these results, but the various measures (using word frequency) suggest that the play's profile is consistently more like Behn than that of her contemporaries [slide 17]. The slide shows the results of a Delta analysis (a method developed by John Burrows and analysed here using David Hoover's Excel Macro files). The lower the score, the more similarity there are in word frequencies between Behn's style and *The Revenge*. There is still more work to do here: we want to know why and how these word frequencies are alike, what characteristics of the dialogue might underlie these similarities and whether that means that there are hallmarks of Behn's dramatic style here – but at least the results are coherent, and promising!

The findings for the Counterfeit Bridegroom are less straightforward. As mentioned, this play is an adaptation of Middleton, and the impact of his authorial style is something that has to be identified and attended to in this kind of analysis. As argued in Evans and Hogarth (2020), this palimpsest approach to style is a key and complicating factor for stylistic analyses of this kind, and one we need to better understand. As shown on the slide [slide 18], you can see how the likeness between acts in Counterfeit Bridegroom and Middleton's style differs: the lower the score, the greater the similarity to the Counterfeit Bridegroom text. This maps clearly onto the scenes with the greatest and least amount of new Restoration-era material. This is not, really, that surprising and in fact offers a proof of concept for the validity of these kinds of tests.

However, it complicates our interest in Behn's authorship: we have to get 'past' Middleton, in order to see what traces might be left of the editor and authorship of the new content. The results, it has to be said, are not particularly conclusive. Some sections of the play show a likeness with Behn more than the other Restoration authors in the dataset, but other sections do not. And remember, that this is only out of the authors we have included. In an open authorship query such as this, we can only make assessments based on the evidence available. The real Restoration author may still be out there. They probably are. If you have any hunches, please let us know!

c. Behn's prose: genre and style; implications for Behn's involvement with the posthumous fictions (1698-1700)

The second case study we'd like to discuss relates to the investigation into Behn's prose style, and the likely authorship of the posthumous fictions. These short fiction works were published a decade after Behn's death and might be viewed as an attempt to capitalise on Behn's marketability at this time. Prose fiction is a fascinating emerging genre within the Restoration literary sphere, and Behn has been identified as a key player in developing the style and characteristics towards its more recent forms. Of course this teleological view isn't particularly helpful: as critics have noted, conventions surrounding prose fiction were broad and flexible, with different writers experimenting with what was possible and what was saleable. Behn's contributions to short prose fiction – separate to the three *Love-letters* volumes in the mid-1680s – are four texts, with *Oroonoko* being the best known. Our investigation therefore, proceeds similarly to the drama: what characterises Behn's prose fiction when compared to her other writing i.e. her drama, and when compared with her contemporaries? Given the nascent work on Restoration prose, we're also seeking to build up a linguistic profile of Restoration prose more broadly – putting Behn in her stylistic context, as it were. Only from this basis can we then assess the posthumous prose fiction.

What have we found? Well, prose and drama are not alike, both in terms of the textual evidence available and in their stylistic characteristics. We have only 4 texts for Behn (c.80,000 words) with no temporal depth, as all were published in 1688-9. So this affects the methods we can use: what works for drama does not, it appears, work so well for prose. Some initial explorations, however, have foregrounded some interesting similarities and differences between Behn and her contemporaries. [slide 19] Using a keyword analysis – words more frequent than they should be, based on a comparison corpus - Behn's prose works are more focussed on temporality, and on sequencing of events, than many of her fellow Restoration writers. The ability, or need, to narrate events underway in different places at the same time, seems to be something more characteristic of her prose – perhaps echoing the multiple narrative strands we find so often in her drama. Her prose works also show variation within them, with Oroonoko profiling differently to the other three texts. This could simply be the unusual nature of its narrative, subject-matter and characters – based in Surinam, rather than Europe, for instance – but it may also reflect, somehow, the innovative qualities identified in this work. [slide 20]

What insights does this analysis provide into the posthumous prose fiction? Provisional results suggest that there are differences between the posthumous prose and Behn's fiction. The dubia groups together – with the exception of *The Dumb Virgin* – suggesting there is internal coherence, possibly authorial coherence. But no real evidence, thus far, that this is a result of Behn's own literary activities.

6. Looking forward: some self-directed advice

[slide 21] What this brief survey of our work on Behn's style and authorship has shown, perhaps, is the potential of computational techniques to develop our appreciation and understanding of literary style, with some value for the editing of those literary works. But at the same time, our initial naïve optimism that we'd be able to secure and resolve the issues surrounding Behn's dubia has rapidly shifted: digital tools are one part of the scholarship puzzle, and like other research approaches (qualitative, bibliographic, and so on), need to be tailored and adapted to the materials at hand. Some of our questions cannot, probably, be answered. But that doesn't mean we can't use these approaches to get new information, and identify new questions, that we might otherwise not have considered.