

Final Report – ZKS-Lendrum Postdoctoral Fellowship, 2022 - Matthew Evan Davis

In my proposal for the ZKS-Lendrum fellowship I had outlined a project that had two main goals: the first was to transcribe the quasi-dramatic works of John Lydgate typically known as his mummings or disguisings, in all witnesses, and make them available online for users of my virtual archive of Lydgate's works.¹ The second, more nebulous goal was to use computational methodologies surrounding natural language processing to discover whether or not there were works alongside those known to have a quasi-dramatic life that might also have been performed in a similar manner. Progress on both of these goals were made over the course of my time at Durham, although not at the level which I would have liked due to my suffering a broken hip shortly after my arrival in the city, as outlined in my mid-fellowship report.

I have transcribed from the available manuscript witnesses the mummings at Eltham, London, Hertford, and Windsor, as well as the mummings for the Goldsmiths of London and the Mercers of London. Additionally, as a control for the second portion of the project I have transcribed most of the witnesses of the Testament of John Lydgate, although as explained below this will likely be unnecessary for that portion of the work. The second stage, proofing, has been done for those items at the British Library, but a follow-up trip the summer of 2023 will be necessary to move those items at Trinity College, Cambridge, from the draft to proofed stage. I am also in the process of ensuring I have proper rights to display the image files I used for my transcription work, as they remain the property of the holding institutions. Once that is done, the items will go up live on the site and be available for the use of scholars and students free and without charge. The reason for this delay is that after my injury I determined my time might best be used to get as much of the original transcription work done as I possibly could and then worry about those elements that did not require me to be in-country in 2023. This work also illustrated several places where what I am calling "stealth edits" occur in the most common scholarly edition of these works, Henry Noble MacCracken's *Minor Poems of John Lydgate*. This has led to some thinking regarding the intervention of editors in the presentation of texts, which is work I would like to continue in collaboration with IMEMS.

For the second part, I consulted with Peter Heslin regarding possible ways of using natural language processing² as a stylometric tool³ with Middle English works. The primary problem with that language, in particular, is that it has a significant amount of orthographic and a lesser amount of dialectal variance. This can be a problem for NLP models as they require set, regularized patterns (or coding that accounts for irregular patterns) to work properly. Thinking further on the problem I determined that the best way to handle the issue would be to do a cluster analysis⁴ of a corrected and normalized version of the Lydgate corpus, which would require a significant amount of preparatory

¹ The site in its current form can be found at <http://www.minorworksoflydgate.net>

² Natural Language Processing, broadly defined, is a set of tools and statistical methodologies designed to allow computers to parse language "naturally," that is to say in the ways that humans do so.

³ Stylometry is the statistical analysis of variations in literary style between authors or, in this case, genres.

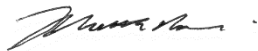
⁴ A cluster analysis attempts to group works together by similarity algorithmically.

work to normalize the orthographic variance in the words.⁵ When that is completed, then the frequency distribution of words in the corpus⁶ can be determined and the aforementioned cluster analysis generated based on the works. To do this work properly, however, will require a larger corpus than my set of transcribed works will allow. For this reason, during my travel time to archives while at Durham I began normalizing the Early English Text Society editions of the *Fall of Princes*, *Temple of Glass*, *Secrees of old Philosophers*, *Reson and Sensualyte*, and *Two Nightengale Poems* alongside MacCracken's *Minor Poems of John Lydgate*. This is work that continues to the present. This has the additional benefit, should the analysis prove valid, of allowing us to correct the categories that MacCracken, places Lydgate's works within in the *Minor Poems*. There are some inconsistencies regarding how these categories are defined and a computational analysis of the works would help support correcting that.

Finally, I was asked to put together a small workshop of like-minded scholars for July, 2023 to further think through both the problems inherent in digital presentation of works online and the possibilities of clustering work at NLP to reassess what have become canonical genres of Lydgate's works. The scholars able to attend should be finalized by the end of this month, but I anticipate it occurring either towards the end of the final week of June or the beginning of the second week of July.

I'd like to reiterate what an absolute privilege and pleasure it was to be at Durham for the year 2022. It sparked a lot of thinking regarding my work in Middle English and larger questions about how to deal with manuscripts as cultural artefacts, and I look forward to continuing to work with IMEMS in whatever capacity I can in the future.

Sincerely,



Matthew Evan Davis

⁵ The orthography of the words will have to be corrected so that every variant of a word is spelled in the same way across all of the witnesses. This can be done to some degree through find-and-replace tools but some words, such as "the," can mean either the article "the" or the pronoun "thee" and must be rendered by context. As there is no current multi-dialect, orthographically consistent model of Middle English to allow a computer to assume that context this work must be done by hand.

⁶ Frequency distribution is the comparison of all of the instances of a particular word to the total number of words present in the corpus, generally reduced to a number between 0 and 1. This number then becomes numeric basis for the algorithms that generate cluster analyses, for example.