**Where did the English novel come from? Experimenting with computational approaches to investigate language in seventeenth-century prose fiction**

Dr Mel Evans, School of English, m.evans5@leeds.ac.uk

This project seeks to identify, describe and interpret the key linguistic properties that characterise the early stages of development of literary prose fiction in English: texts written and read in the mid-to-late seventeenth century. The history of the novel in English is an established area of research, but it has traditionally been the preserve of literary criticism and qualitative narratological investigation. More recently, the advantages of 'distant reading', representing the quantitative and computational capabilities of digital humanities, have started to emerge and show potential (e.g. Moretti 2013, Underwood 2019; Menon 2024). However, seventeenth-century literary prose fiction has received less attention, partly because the texts themselves are experimental and heterogeneous, and also because they lack the cultural prestige and name-recognition of their successors. Yet these early works are vital for our understanding of how individual language users, working in a creative context, can converge to produce a shared understanding of a new cultural form. Tracking language and its evolution in early prose is therefore essential for understanding how a new form of narrative communication emerges and stabilises.

The proposed project will act as a proof-of-concept, developing and appraising a workflow and methodology to investigate the language features that make up seventeenth-century prose fiction. This project builds on the academic supervisor's own in-progress research on this area, which provides a baseline of both principles and texts to support the project outlined here. The intern will work with their supervisors to develop and investigate a corpus of English prose fiction published before 1710. During the 8-week internship, it is envisaged that the intern will:

1) collect text-based data from seventeenth-century literary prose publications – including digitizing these texts using OCR tools, where appropriate;
2) apply and evaluate data cleaning techniques;
3) trial approaches for data tagging and data extraction e.g. identifying sections of dialogue, or non-narrative blocks of text;
4) explore the textual data using corpus-based analysis tools and visualization techniques.

Throughout, the intern will have the opportunity to develop their skills and understanding across key stages of the investigation.

The central ethos of the project is one of experimentation and innovation in the use of computational approaches. The computational dimension to this project is, therefore, crucial. In particular, the emerging context of GenAI and LLMs to support language-based research entails that older methodologies for corpus-based language analysis can and should be revisited and augmented, e.g. vibe coding for data cleaning and analysis, NLP visualisations for metadata and analytic outputs. The intern will be supported to question and experiment throughout the project, with the aim of identifying opportunities for efficiency, accuracy, and new insights using computational tools across the workflow.

The student intern will be invited to be a named contributor to any research publications developed by the academic supervisor during the period of the project.