Scaling Rhetorical Move-Step Analysis: An HPC Pipeline for the Royal Society Corpus

Lead Supervisor: Dr Charles Lam (Language Centre, University of Leeds) <c.lam@leeds.ac.uk>

1. Context and Problem Statement

While the rhetorical structures of modern scientific writing are well-documented, there remains a significant knowledge gap regarding the diachronic evolution of these structures over the last three centuries. Most genre studies in applied linguistics are synchronic, focusing on contemporary articles to assist modern scholars. Consequently, we lack a large-scale, empirical understanding of how the "scientific voice" transitioned from the descriptive, personal narratives of the 17th century to the highly structured, objective frameworks used today. The Royal Society Corpus 6.0 Open (RSC6o) provides a unique opportunity to bridge this gap, containing approximately 17,520 texts from the Philosophical Transactions and Proceedings of the Royal Society. However, analyzing the shifts in "moves" and "steps" (e.g., how authors establish research territory or report findings) across 300 years of data is impossible through manual annotation. This project seeks to fill this historical void by applying machine learning to the RSC, mapping the rhetorical structure in science writing from its inception to the modern era.

2. Overarching Goals and Themes

- Automate the identification of rhetorical intent in historical scientific writing.
- Provides the necessary data architecture to feed a pre-existing visualization dashboard, allowing for comparative studies between historical and modern scientific genres.
- Bridge the gap between raw corpus data and high-level visual analysis.

3. Computational Requirements and HPC Integration

Leveraging parallel architecture with GPUs, the pipeline can decompose millions of individual sentences into concurrent batches for simultaneous classification, transforming the task sequential on a local computer into a vectorized operation. Such a transition requires:

- Classifying millions of sentences across the 17,000-article corpus.
- HPC scaling to manage the memory-intensive inference and data structuring.
- Data pipeline optimization to transform complex XML schemas into structured outputs compatible with downstream data visualization web applications.

4. Role and responsibilities of the Intern

- Developing the ingestion pipeline for RSC6o
- Optimizing machine learning classification models for deployment on N8 CIR HPC clusters
- Structuring the processed output to feed into an existing dashboard for final visualization

5. Expected outcome

- The project will enhance RSC6o with classification of rhetorical structure, showing how scientific persuasion and argumentation have evolved. This will identify "pivot points" in history where descriptive narratives were replaced by the modern objective structures we see today.
- A HPC-pipeline will be delivered as a documented workflow for the N8 community, which can be adapted for other historical corpora that require sentence-level rhetorical classification.
- By feeding the processed RSC data into a visualization dashboard, other researchers can perform real-time queries, contrasting rhetorical strategies in historical and contemporary science writing.

References

Omotola, O., Nnamoko, N., **Lam, C.**, Korkontzelos, I., Altham, C., & Barrowclough, J. (2025). Automatic detection of the CaRS framework in scholarly writing using natural language processing. Electronics: Special Issue on Emerging Theory and Applications in Natural Language Processing, 14(14), 2799. https://doi.org/10.3390/electronics14142799

Swales, J. M. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge University Press.

Universität des Saarlandes. (2024). The Royal Society Corpus (RSC) - Version 6.0. Retrieved from https://fedora.clarin-d.uni-saarland.de/rsc_v6/