REVIEW OF VERONIKA STRAKOVÁ'S THESIS

CLAUDIO J. RODRÍGUEZ H.

Visualizing assumptions based on efficient, data-based methods is an essential task in how we ought to approach communicative intent. Veronika Straková's thesis is a genuine attempt at producing a clear and well-documented methodology and application to understand character relations in fictional works as an analytical tool for authorship attribution. In this sense, the thesis displays a remarkable thoroughness regarding its process, painstakingly documenting the technical work done, including precision evaluation and minor fine-tuning.

The author's clear writing and well-document efforts are greatly appreciated and should be commended. The thesis is well structured, concrete in its theory, exhaustive in its application and sober in its presentation. Overall, it is an excellent demonstration of knowledge and application over a narrow field–the exact sort of work that is best suited for academic research.

As the author recognizes, there are no particular features explored in the thesis that would aid in robust authorship attribution, but this is not a particularly discouraging conclusion for two reasons: Negative results are indeed results, and the process of testing hypotheses is essential for the development of new methodologies. The author did find some interesting traits to explore that could, in fact, aid with the process of finding properties that are correlated with authorship traits.

Interestingly, because the author believes that the information compiled in the first part will be somehow correlated to a secondary property of a work–namely, that certain properties (which we may call semantic, to some degree) within the work are correlated with a property that is external to the work (its author)–she states that a larger corpus will produce more precise results (72), but this is still constrained by the assertion that length is correlated to network features. There is something amiss here, as a larger corpus would be limited by authorial decisions in the first place, meaning that the methodology, even when amplified, would still require a strong selection of an ad-hoc corpus,

which in turn would make the work done less effective when applied unless used in specific cases, such as trying to pinpoint authorship in the corpus of a single writer with a large number of works, under the suspicion that some of those works may have different properties and thus different authorship. The author would do well in being more specific about the exact application of the suggested methods to see whether the long-term aims of the thesis are feasible and if the theory employed is in fact useful. The reason to mention this is that the assumption that certain textual properties are correlated with external (or even epiphenomenal) properties of the text (assuming, of course, that authorship is not an internal feature, which could of course be a perfectly acceptable view if the author of the thesis wishes to hold it) is still contentious, particularly when dealing with works of fiction. It would, in this vein, be interesting to analyze the difference in distributions in works of writers who subvert genre conventions or those who are prolific in different kinds of narrative styles. The intention of this comment is, however, positive: Straková employs techniques to understand some aspects of the text, and these aspects may not necessarily be correlated to authorship, but they may shed light on other features of literature, allowing us to see concrete data about intuitions that we may have about the content of a piece of literature. This is, by itself, a very positive development and an unexpected way forward in her work–providing literary analysis with quantitative tools that can enrich our knowledge of the contents of a piece of literature.

Some lingering questions that may interest the author can be formulated thus:

- Is there any relevant amount of text necessary to make the analysis work? For instance, a comic book may be analyzable within a corpus, but its multimodality would probably render the analysis much more complicated to perform, despite the fact that much of the textual mode is conversational.
- Could there be more noise in the data derived from using novels from different era, dialect and sociolect, and do we even need to account for the latter?
- Is there any more robust way for entity identification in cases of ambiguity, as for instance with shared last names, than simplification to the heaviest weight?

I want to state once more that this is a very good thesis, carefully written and complete to a fault, well documented, clearly written and,

overall, a great demonstration of careful consideration of both theoretical and technical aspects. The author is commended by her work and documentation, and my recommendation is for the thesis to receive the highest grade.

Claudio J. Rodríguez H.