

## Mapping Genre Space via Random Conjectures

Patrick Juola // Duquesne University, Pittsburgh PA // [juola@mathcs.duq.edu](mailto:juola@mathcs.duq.edu)

One of the key problems facing digital humanities today is the increasing number and size of digital repositories and the relative lack of tools for studying them. A collection of a million books (Crane, 2006) is no more useful than a collection of ten thousand if you can't read more than a hundred of them in a realistic timeframe. Scholars like Moretti (2005) have proposed a new analysis method, termed "distant reading," to enable computer-aided large-scale analysis of such collections. In previous work (2009), we have proposed using a conjecture generator (Conjecturator, see also <http://www.twitter.com/conjecturator>) as another computer-aided analysis method.

Like its predecessor and inspiration *Graffiti* (Fajtlowicz, 1988), the conjecturator generates template-based "conjectures" that might or might not be true about the repository and the texts in it. A sample conjecture might be something like:

- The word group (concept) of "pain" appears more in marriage plot novels than in Gothic novels.

This particular conjecture is probably not true, but if it were true, it would represent a previously undiscovered and interesting difference between the two genres (as represented in the repository).

In this paper, we demonstrate one way to extend this conjecture-based analysis to a large-scale "distant reading" and visualization of genre differences. Repeated generation of conjectures will create a large catalogue of potential differences between any particular category pair, some true/supported, and some false. The number of "true" differences, or alternatively, the percentage of true differences, can be viewed as a distance between the categories, a distance measuring the degree of difference between the concepts commonly written about in those genres.

To aid in the study of such differences, we compile the differences into a matrix and apply multidimensional scaling (MDS) (Cox and Cox, 2001). This statistical technique takes a high-dimensional data set defined by interpoint distances and embeds/rescales it to fit a smaller number of dimensions (in this case, two) while minimizing distortion. The resulting two-dimensional coordinates can be plotted to give a visual "map" of the space of genres. We demonstrate this technique using the same set of Victorian novels analyzed in (Juola, 2009).

The resulting images clearly indicate that this method is a new and viable way of performing large-scale distant reading. It is easy and relatively efficient to apply and almost entirely document-agnostic; it can be applied as easily to journal articles (and map the space of scholarship) or to newspaper corpora (perhaps mapping the space of editorial policies and politics) as to novel genres.

- Cox, T.F., Cox, M.A.A., (2001), *Multidimensional Scaling*, Chapman and Hall.
- Crane, Gregory. (2006), "What Do You Do With a Million Books?" *D-Lib Magazine* 12(3)
- Fajtlowicz, Siemion, (1988), On conjectures of *Graffiti*. *Discrete Mathematics*, 72
- Juola, Patrick, Bernola, Ashley, (2009), "Conjecture Generation in the Digital Humanities," *Proc. DH-2009*
- Moretti, Franco, (2005), *Graphs, Maps, Trees: Abstract Models for a Literary History*, Verso