

On the Origin of Theories: The Semantic Analysis of Analogy in a Scientific Corpus

Devin Griffiths
Center for Cultural Analysis, Rutgers University

In my talk, I describe using textual analysis to examine the role of analogy in the 1859 text of Charles Darwin's *On the Origin of Species*. The *Origin* redefined several key biological concepts, including "selection," "variation," "chance," and especially, "evolution" and "species." One way to look at this is to suggest that Darwin's major accomplishment was to change the meaning of these key words, and in order to do this, he had to break apart many of their earlier associations and forge new relationships and connections to new observations. Historians of science have long noted that major changes in scientific understanding are accompanied by shifts in how scientists talk about the world; William Whewell, a contemporary of Darwin's father, pointed out that the dissemination of new theories are marked by "some familiar axiom, or perhaps by some new word or phrase, which becomes part of the current language of the philosophical world" [1]. In order for a phrase like "natural selection" to catch on, it was necessary for Darwin to clarify what this perhaps oxymoronic metaphor meant – by forging connections between phenomena which had previously seemed distinct. I argue that Darwin achieved this shift through the extensive use of analogy, and use Latent Semantic Analysis (LSA) [2] to measure the frequency of single-sentence analogical constructions over the course of his work. Researchers have shown that LSA can be used to evaluate the presence of analogies within a synthetic data set, but to date, no one has explored LSA of analogy upon a natural language corpus [3,4].

Analogy is particularly suited to such analysis because it both has strong semantic features (it draws connections between divergent relationships), and because, in its fully explicit form, it entails complex syntactic forms. In order to evaluate the LSA measurements, I use regression analysis and an independent hand count of analogy frequency within samples from the text (relying upon a strict syntactic definition of which phrases would be counted as analogy). This syntactic definition also allowed me to generate automated syntactic counts of analogy using the Morphadorner part of speech tagger [5]. I show that the semantic LSA-based count of analogy has a strongly significant correlation with the syntactic count, even when controlling for factors like sentence length. Moreover, this survey of analogy allows me to demonstrate that analogies are far more frequent in the initial and concluding chapters of the work, sections that are instrumental in establishing and reiterating the core relationships that define natural selection.

This automated analysis of analogy also allows me to characterize the ratio of semantically analogous to non-analogous sentences in Darwin's work, a ratio which I compare with several contemporary works of popular fiction, non-fiction, and science. This analogy ratio suggests that analogical mode of Darwin's prose style is more closely related to contemporaneous works of popular philosophy, particularly, the scandalous *Notes and Queries*, than more traditional tracts of the biological, medical, and physical sciences (a characteristic that perhaps underlines the deep influence exerted by his careful reading of Milton during those long years at sea).

I conclude with a discussion of analogy and its relationship to narrative coherence. Because analogies, of necessity, involve a transfer of characteristics between two distinct

semantic systems (often described as “domain” to “target” transfer), analogies tend to disrupt the semantic coherence of sequential sentences, a tendency I verify by examining the negative correlation between sentence-to-sentence semantic coherence in the *Origin* and the proximity of analogies. Phrases like “natural selection” reflect the condensation of analogy into metaphor, as the transformative connections which establish the intellectual framework for a new theory modulate into less syntactically rambunctious metaphors – metaphors which can then be incorporated smoothly as agents of extended scientific narratives, and take their place as “part of the current language of the philosophical world.”

[1] *History of the Inductive Sciences*, (New York: D. Appleton and CO., 1875).

[2] LSA developed and described T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge," *Psychology Review* 104 (1997): 211-240.

[3] Michael Ramscar and Daniel Yarlett, "Semantic grounding in models of analogy: an environmental approach," *Cognitive Science* 27 (2003): 41-71.

[4] For the synthetic story set they used, see M. Redington, N. Chater, and S. Finch, "Distributed information: A powerful cue for acquiring syntactic categories," *Cognitive Science* 22 (1998): 425-469.

[5] Developed by developed by the Wordhoard, Monk, and VOSPOS workgroups at the University of Northwestern <<http://morphadorner.northwestern.edu/morphadorner/>>.