Darwin's Language Evolution: Detecting Lexical Change in Darwin's Correspondence

Nicole Tamer[*1,2], Barbara McGillivray[3], and Elizabeth Smith[4]

*Corresponding Author: nicole.tamer@uzh.ch
[1]Department of Comparative Language Science, University of Zurich, Switzerland
[2]Center for the Interdisciplinary Study of Language Evolution, University of Zurich, Switzerland
[3]Department of Digital Humanities, King's College London, United Kingdom
[4]Cambridge University Library, University of Cambridge, United Kingdom

Abstract

Charles Darwin's evolutionary theory was highly influential. To better understand his theories and Darwin as an individual, we investigated the language of his private correspondence. The present article detects and analyses the usage of scientific words in Charles Darwin's correspondence using a combination of corpus linguistics and computational approaches. With a correspondence corpus of nearly 15,000 letters, we developed a computational model based on distributional semantics to track lexical semantic change over time. The findings provide new insight into Darwin's language use.

1 Introduction

Darwin is best known for the development of his evolutionary theory. His highly influential work formed the basis of modern biology and medical research (Mayr, 2000; Bowler, 1996; Engels and Glick 2008). Beyond the scientific world, the influence of Darwinism was felt in almost all areas of thought, not least the politics and literature of the late nineteenth and twentieth centuries (Beer 2009; Levine 2011). Based on the collection of Darwin's letters from the Darwin Correspondence Project, this article focuses on the evolution of Charles Darwin's vocabulary, particularly the time periods 1821-1870 and 1871-1882. This work is the first in-depth quantitative textual analysis undertaken with the resources of this corpus and focused on Darwin's lexical usage.

2 Darwin's Letters

The dataset used in this article was provided by the Darwin Correspondence Project team based at Cambridge University Library. This team of experts has been editing his letters since 1974 and accumulated an in-depth knowledge of Darwin's life and letters. The first letters in the collection date to 1821, when he was still a child, and continued to his death in 1882. Most of his vocabulary was used in either a precise scientific context which we focused on, or a more relaxed rhetorical sense. Letters were an important method for Darwin to collect data, get new insights from other scientists and to substantiate his theories. They have been described as a crucial part of Darwin's scientific archive.

3 Computational Methods

To investigate semantic shifts in the usage of words, we experimented with various parameters for building Word2vec embeddings using the skipgram with negative sampling algorithm (Mikolov et al., 2013). We trained both skip-gram and continuous-bag-of-words embeddings with 300 dimensions and considered window sizes of 5 or 10 and a minimum frequency count of 0 or 1. To choose the best combination of parameters, we compiled a list of pairs of words that we expected to display distributionally similar characteristics (verification approach). We then chose the model that led to the highest average cosine similarity between the synonym pairs, which was a skipgram model with a window size of five and a minimum frequency count of 1. The final step in this process consisted in identifying those words by the algorithm that could be considered as candidates for semantic change in the corpus. We trained Word2vec models with the parameters chosen above in each of the two subcorpora: before 1870 (t1) and after 1870 (t2). The wide range of Darwin's interests and his tendency to nurture them for much of his adult life make it

impossible to cleanly divide his life into specific time periods, but we chose two time intervals, the time period before 1870 and the other one ranging from 1870 until Darwin's last letter in 1882. We

then aligned the two spaces using the Orthogonal Procrustes method (Schönemann, 1966), already used in the semantic change detection literature by Hamilton et al. (2016) and many others since. By using this method, the cosine similarity score of a lemma is higher if the embeddings of the lemma in the two time periods are most similar

4 Corpus Analysis Results

Several factors influenced the way Darwin used language in his letters. Darwin had a wide range of interests about which he corresponded extensively, and the shift of his interests and the subjects on which he was writing and publishing is reflected in the letters. This explains the increase of usage of the words of interest before two of his most influential publications, On the Origin of Species in 1859 and The Descent of Man in 1871. The topics in these publications triggered multiple words' semantic change. The most common semantic change was broadening. The narrow meaning of many scientific words got adapted to a wide range of metaphorical meanings or became more flexible by relating to more extensive scientific fields over time. This discovery is supported by how he approached his research and readily integrates new knowledge and discoveries and according to our findings, new meanings to his scientific and casual terms. A future area of interest could be to compare the findings of language change in other sets of correspondence to a representative selection of publications.

5 Computational Results

The computational analysis demonstrated that an algorithmic model based on word embeddings can be applied to the Darwin Correspondence Corpus to detect words which could have changed over time. Based on qualitative and quantitative analysis as well as in-depth expert knowledge of the corpus, we have demonstrated that change could be detected with the verification as well as the discovery methods. However, when applied to a restricted corpus such as Darwin's correspondence, and over a relatively limited time period, the type of changes detected are more likely to involve the context of usage of the terms rather than their semantic profile. This study is the first to adopt recent computational linguistics research into semantic change to investigate questions in the history of science. It provides novel insights into Darwin's language usage with a specific focus on lexical semantic change and usage change. The computational model confirmed these lexical semantic changes which were demonstrated in topdown and bottom-up analyses. This study's results can be replicated in future corpus work and more generally in other areas of digital humanities research to elucidate lexical usage of individuals.

6 References

Charles Darwin. 1968. *On the origin of species by means of natural selection.* 1859. *London*: Murray Google Scholar.

Charles Darwin. 1871. *The descent of Men.* New York.

Ernst Mayr. 2000. Darwin's influence on modern thought. *Scientific American*, 283(1), 78-83.

Eve-Marie Engels and Thomas F Glick. (2008). *The reception of Charles Darwin in Europe*. 2 vols. London: Continuum.

George Levine. (2011). *Darwin the writer*. Oxford University Press.

Gillian Beer. 1983. *Darwin's Plots: Evolutionary Narrative in Darwin, George Eliot, and Nineteenth-Century Fiction.* Cambridge University Press.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. "Cultural shift or linguistic drift? comparing two computational measures of semantic change." *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing.* Vol. 2016. NIH Public Access.

Peter J. Bowler. 1996. *Charles Darwin: the man and his influence.* Cambridge University Press.

Peter. H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1), 1-10.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems.*