
Hermann Moisl
University of Newcastle upon Tyne

Corpus-based generation of linguistic hypotheses using quantitative methods

In linguistics as in other sciences, the de facto standard methodology is based on Karl Popper's concept of the falsifiable hypothesis, whereby a hypothesis is proposed in answer to a research question about some domain of interest and then tested by observation of the domain. Because of their centrality, it is natural to ask how hypotheses are generated.

The consensus in philosophy of science is that hypothesis generation is non-algorithmic, that is, not reducible to a formula, but is rather driven by human intellectual creativity in response to a research question using a combination of deductive inference from existing axioms and theorems, and inductive inference of generalizations from observation of the domain. Hypothesis generation by deductive inference has long been dominant in generative linguistics and, to a lesser extent, in other subdisciplines like historical linguistics and sociolinguistics. The advent in recent decades of large amounts of digital electronic text amenable to computational analysis has, however, made hypothesis generation by inductive inference viable. This paper shows how mathematical and statistical techniques such as cluster analysis and singular value composition can be used for that purpose.