

DH and the Evidence for Linguistic Generalizations

DH has played an important role in the development of evidence-based linguistics. While generative linguists have generally avoided the processing of linguistic evidence, sociolinguists and phoneticians have developed sophisticated statistical methods regarding variable rules and acoustic phonetic data, often using the R system. These analyses would have been impossible without digital tools. However, the emergence of large-scale bodies of evidence has brought such statistical methods into question. It is not the methods themselves but the nature of the evidence that raises issues. It has become apparent from survey evidence and from forced-alignment with automatic vowel measurement that the evidence for linguistic generalizations is not normally distributed. It lacks a central tendency, and further, linguistic evidence is scale free, self similar at every level of scale. This means that statistics that assume a normal distribution will not produce correct results. We require new digital tools in order to make appropriate generalizations from large-scale evidence, including the databases used for training AI programs. This paper will suggest appropriate digital GIS-based methods for use on large-scale phonetic data, based on a large-scale American project, and will illustrate alternative digital methods for use with linguistic corpora.

References excluded to keep the abstract anonymous