

Linguists are divided on the question of language universals and language diversity. Some have argued that the grammars and sound systems of the world's languages exhibit striking similarities, down to minute details, even when the languages are not historically related, suggesting an explanation for this universality in terms of our shared cognitive mechanisms. Others have instead underscored language diversity, arguing that languages differ in puzzling ways, suggesting that there may not be any limits to possible human languages and that the observed shared properties are just accidents of history. One difficulty in adjudicating this debate is that language universals can be statistical and studying them requires quantitative data as well as a deep understanding of individual languages. We propose a new approach to linguistic universals that tackles the problem simultaneously from two angles. Building on recent work in theoretical linguistics we first derive candidate universals from categorical and stochastic linguistic models through sophisticated mathematical analysis of the structure of these models. We then validate the resulting predictions on large linguistic databases. This two-pronged approach brings together mathematical, theoretical, and empirical linguistics and promises to shed new light on the contentious question of the place of language in human cognition.