Validation of Results in Linguistic Science and Technology: Terminology, problems, and solutions

Mark Liberman

Invited talk abstract:

Everyone agrees that there's a problem: very often, results and conclusions in experimental science and some areas of engineering turn out to be unreliable or false. And everyone agrees that the solution is to put more effort into verifying such results and conclusions, by having other people re-do aspects of the research and analysis.

There can be many reasons for unreliability: outright fraud, programming errors, "p-hacking", the "file drawer effect", or unrecognized co-variates in complex situations. And there are many types of solutions, from checking or re-writing the scripts used to analyze the original data, to trying new analysis techniques on the original data, to redoing human or machine coding of raw specimens or recordings, to collecting new datasets using the original techniques, to collecting new relevant data in new ways or in new contexts.

Unfortunately, the terminology in this area is a mess. The two ends of this spectrum of "doing over" are commonly described using the terms replicate/replication/replicability vs. reproduce/reproduction/reproducibility — but different groups use these terms in diametrically opposite ways. This makes discussion of the issues confused and confusing, in a situation where we need to be precise about diagnosing possible problems and prescribing best practices for different types of research and in different subdisciplines.

With respect to the "reproducibility crisis", under whatever name, corpus-based speech and language analysis is decades ahead of psychology, biology, and medicine. Everyone agrees that researchers should make various types of validation easier by publishing their data and methods, and by using well-defined evaluation techniques that are resistant to over-fitting — and we've (mostly) been doing this for 30 years! But there's still room for improvement. In this talk, I'll try to clarify the terminology, assess the remaining problems in our field, and suggest directions for improvement.