Appendix F: Do findings drawn from the field of statistics fall within the province of linguistic theory?

Appendix to "Deriving the Wug-shaped curve:
A criterion for assessing formal theories of linguistic variation

Bruce Hayes
UCLA

March 2021

This brief document addresses the idea that linguistics might "turf out" some of the work of their theorizing. Substantive content and theoretical architecture should remain our task, but the job of finding the right mathematical system into which we embed our linguistic ideas might better done if we are willing to make substantial imports from statistics and computer science, both of which are fields that have much to say on the latter topic.

When I was a linguistics student in the 1970's, I would have found it unimaginable that statistics could become part of linguistic theory. This was partly the consequence of the statistics that was taught to undergraduates at the time, which seemed (to me at least) to be little more than a set of methods experimentalists might use to avoid error — a matter of good scientific practice but hardly of theoretical interest.[1]

However, statistics kept evolving; in the intervening decades it seems to have become a far more lively, exploratory research activity.[2] Nowadays, statistics does better at extracting valid conclusions from noisy data than it used to; the real world, including even baseball, attests to this. And extracting valid conclusions from noisy data is precisely what young human language learners must do. If there exist rational and effective mathematical modes of inductive reasoning, then it is not unreasonable to suppose that natural selection has equipped human beings to use some analogue of these principles.

[1] I might add that my training in linguistics covered virtually no probabilistic phenomena; this rendered the gap between linguistics and statistics complete, assisting my failure to appreciate the work being done by sociolinguists at this time (§**Error! Reference source not found.**).

[2] A popularization I have enjoyed is McGrayne (2011), which conveys some of the liveliness of modern statistical inquiry as well as a sense of drama about how the field came to evolve into its present form.

On the other side of the gap, linguistics has developed in important ways that let it engage more close and more naturally with statistical methods. A clear example is Optimality Theory and its probabilistic descendents, which share the crucial trait that linguistic knowledge is deliberately "atomized" to the point that statistical principles can engage with it effectively. Such theories leave plenty for the linguists to do, for instance in understanding the families of constraints and their origin or in understanding the form of linguistic representations. But we should be happy to import our forms of probabilistic reasoning if that is what works best.

Approaching the same issue from the other side, I feel there is also a need to incorporate linguistic theory into statistical data analysis. Some of the research I have read for purposes of writing this article strikes me as unusually agnostic with regard to theory: it is quite easy in working practice to adopt purely-empirical classifications of the facts, plug them into a good statistical model, and obtain accurate results. I hope to see future work use constraints that are themselves the result of extensive theoretical development and typological testing. Such scaling up of the inquiry will, I think, ultimately make the research results more explanatory and more convincing.

McGrayne, Sharon Bertsch (2011) *The Theory That Would Not Die.* New Haven: Yale University Press.