abstract 127

How to design a dataset which doesn't undermine automated analysis

The twenty-first century brings the science of big data. For typologists, two new theoretical tasks have arisen. We must ascertain which new methods are appropriate for the modeling and analysis of language. And we must ascertain which linguistic data, and which methods of representing it, are mathematically appropriate for those methods. We argue that much of data with which typologists are both familiar and comfortable is liable to violate essential preconditions that computational methods regularly place on their input. We review one computational study that employs a dataset which at first glance looks well constructed and examine where, how, and why the dataset undermines the assumptions of the model; we then examine solutions.

The rapid rise of computational statistics presents challenges to a small discipline like linguistics, where even the most outstanding typologists rarely run a lab employing its own information scientist. Moreover, premature studies which do violence to linguistic data but are published in highly visible outlets have muddied the water unnecessarily. The science of the statistics is sound, but the application of it to linguistic data has become contentious. On the positive side, such studies have generated a healthy debate about which statistical models might be appropriate for the analysis of language.

Less attention has focused on the fact that automated methods, even when appropriate, are brittle. They produce meaningful results only if the input data meets stringent mathematical preconditions. Again, distractions abound as attempts are made to analogize from molecular biology to language, when what is at stake for the use of statistical methods is more general and fundamental. Namely, computational statistical methods place strong constraints on the dependencies which may exist between data points; very often, **independence** is required. In linguistics however, a drive for elegance has led us to cultivate analyses in which individual parts are highly interdependent. Without careful scrutiny, these dependencies will carry over into typological datasets, rendering their analysis by most computational methods invalid (or at best, degraded) from the outset.

We illustrate how these issues play out in a study of 121 languages × 160 typological survey-questions (Reesink et al. 2009). We then propose the following methodological guidelines:

1. **Use micro answers rather than macro**. Answers to many 'macro' questions in linguistics, e.g. "are there prenasalised stops?" are arrived at by weighing up answers to multiple, antecedent 'micro' questions, e.g. "does [NC] appear word initially?"; "does /^{N}C/ contrast with /N+C/?". Confusingly, two languages may answer macro questions identically while having none of their micro answers in common. Micro answers are more valuable data.
2. **Identify dependencies**. Macro answers may share underlying micro answers. E.g. "are there prestopped nasals?" and "are there closed syllables?" are both sensitive to micro questions about intervocalic clusters. This gives rise to dependencies.
3. **Minimize and track dependencies**. Dependencies should be minimized, and where they remain, must be kept track of. Doing so enables one to select multiple *subsets* of the data which are independent.

Application of these principles will improve the suitability of linguistic datasets for the advanced statistical methods now dramatically impacting the quantitative sciences.