DATA QUALITY AND THE FALLACY OF COMPOSITION

Martijn Wieling and John Nerbonne, University of Groningen
{m.b.wieling,j.nerbonne}@rug.nl

Abstract

The primary data on pronunciation variation, e.g., dialect atlas data, is often recorded incommensurably, i.e., in different ways in different atlases, and even in different ways within the same atlas when teams of fieldworkers and transcribers are involved. In particular, these data collections differ in the detail in which pronunciations are recorded, using between 40 and 100 different basic symbols. This study shows that transcription system detail (understood in this sense) increases the linguistic distance measured and therefore must be regarded as a source of bias in assessing pronunciation differences and comparing them across languages. A method is therefore introduced to reduce transcription system complexity, even while retaining faithful assessments of aggregate pronunciation differences. The technique introduced is relevant when comparing within sets that have been transcribed very differently and also when comparing different dialectological datasets, e.g., with respect to the dependence of linguistic difference on geography.

Key words

Pointwise Mutual Information; Linguistic versus geographic distance; Transcription system reduction; Sublinear curve; Normalizing transcriber differences