One of the principle signs that speech is a complex system (Kretzschmar 2009) is the non-linear arrangement of frequencies of variant lexical responses and pronunciations in linguistic survey data. When the counts of word or variant tokens are charted by frequency, they form an asymptotic hyperbolic curve, or A-curve, at every scale of analysis, from the survey overall down to small geographic subsamples. However, the shape of the curve is sensitive to size of the sample and the categorization of the data. A small sample is unlikely to show an A-curve if the number of tokens is insufficient to differentiate the top-ranked variants from those in the tail of the curve. As for categorization, counting F1/F2 coordinates from acoustic phonetic analysis is unlikely to show a curve because there are so many possible categories (on the order of a million, from an F1 range of c. 500 Hz and F2 range of c. 2000 Hz), so the frequency of any given coordinate pair is most likely to be one or two tokens even for a large sample. On the other hand, allowing too few categories will also not show a curve: frequency data from two categories, say for whether or not speakers pronounce postvocalic -r, gives us a line, not a curve. The A-curve can only be observed when the number of categories into which the data is sorted lies between these two extremes. Common practice in dialectology and sociolinguistics has been to establish a small number of possible categories into which variants could be grouped, such as phonemes for pronunciation, or to notice only the most frequently occurring variants and to ignore the remaining variation. Such methods cannot address the underlying complexity of the data. In this paper, we will discuss tools used in economics (Gini Coefficient, Lorenz Curve), and report the results of an experiment in which lexical and pronunciation data from survey research is analyzed in various geographic subsamples. We will demonstrate thereby that A-curves do exist in the data in all cases, and establish parameters for the interaction of sample size and number of categories in the design of valid and reliable experiments.

Reference