

## **Data sets from independent sources**

CHRISTOPHER CIERI

*University of Pennsylvania Linguistic Data Consortium*

MALCAH YAEGER-DROR

*University of Arizona*

The fieldwork required for re-recording the same speakers over time is time-consuming and resource intensive, so it can be efficient to make use of existing longitudinal recordings from independent sources instead. 'Found data' has obvious advantages and well-known, but often underestimated, disadvantages in various of the fields that sociolinguists find important. This paper will discuss various 'found' corpora, and both 'advantages' and 'disadvantages' which arise when making use of these corpora. What shortfalls in coding are inevitable, and should be given in the adoption of such a corpus? How and to what extent can one supplement previous short-falls in coding rigor? Projections based on previous experience, and advice for carrying out supplemental coding where necessary (and possible) will be discussed.