Abstract

The Lancaster-Northern Arizona Corpus of Spoken American English (LANA-CASE) is a collaborative project between Lancaster University and Northern Arizona University to create a publicly available, large-scale corpus of American English conversation. In this article, we describe the design of LANA-CASE in terms of the challenges that have arisen and how these have been addressed – including decisions related sampling the data, recruiting participants, and selecting instruments for data collection. In addressing these challenges, we were able to draw on and further develop strategies established in the creation of the British English counterpart to LANA-CASE, the Spoken BNC2014 (Love et al. 2017), as well as implement recent theoretical and technical innovations. We hope that this discussion can inform future projects focused on the design and construction of spoken corpora.

Keywords: spoken corpora; conversation; corpus compilation; LANA-CASE