Compiling a “small” corpus: Issues of data and comparability in a corpus of Grenadian English

Ryan Durgasingh (University of Münster)

Chief among the strengths of the ICE suite of corpora is the project’s ongoing emphasis on comparability, both cross-varietally (catering to analyses across national standards) and cross-textually (important to analyses focused on contextually dependent text types). While the geographic spread of national varieties represented in ICE is large, its current lack of any particular regional focus can prove problematic to researchers wishing to explore features which may be sensitive to potential areal properties – not surprising given the continued focus on delineating the corpora according to “L1/ENL varieties or ‘core Englishes’ […] and […] L2/ESL varieties or ‘new Englishes’ […] corresponding to Kachru’s (1985) model of ‘Inner’ and ‘Outer’ Circles of English (Kirk & Nelson 2018: 1). Because of the dearth of material from various small nation states, and a general tendency to treat large, culturally/economically dominant varieties as indicative of their entire regions, it is often difficult to assess whether or not regional varieties may exist in more than a theoretical sense (c.f. Allsopp’s definition of “Standard Caribbean English” (1996: lvi), for example). Clearly, there is a need for more comparable corpora intra-regionally, yet, their compilation poses specific challenges that need to also be addressed. By using the example of the compilation of a “small” (both in size and in terms of the socio-demography of the variety’s location) Grenadian corpus, this paper aims to explore the various issues arising out of spoken data collection and comparability experienced by the present author in order to carry out morphosyntactic analyses of English in the Caribbean using the aforementioned compiled Grenadian data, ICE-JAM, and the as-yet-unreleased ICE-TNT corpora.

The presentation will explore the balancing act between practical concerns (what is feasible given the limitations of a single researcher’s time/resources, and the availability of data in a small nation state), and larger theoretical concerns (how does one gather enough data, given the practical concerns, in order to be representative of a particular variety?). In the case of the latter, the importance of a driving theoretical framework, namely Tannen’s (1982) oral-literate continuum, and the attendant arrangement of text types along this continuum so as to both maintain comparability across varieties, and to be representative of language in typical/important domains of use within the variety, will be discussed.

Also important to the discussion will be the various corpus design specifics of collecting different text types – from exploiting academic/layperson networks, to using social media to gather data. The paper ends with general recommendations of those text types which may be more, or less, difficult for researchers hoping to compile small corpora for comparison with existing ICE corpora.

References

Allsopp, Richard. 1996. *Dictionary of Caribbean English usage*. Oxford: Oxford University Press.

Kirk, John & Gerald Nelson. 2018. The International Corpus of English project: A progress report. *World Englishes* 37(4). 1-20.

Tannen, Deborah. 1982. *Spoken and written language: Exploring orality and literacy*. New Jersey: ABLEX.