ICE corpora, register, and omitted variable bias: A multi-dimensional perspective

Axel Bohmann (University of Freiburg)

The development and continuing growth of the International Corpus of English (ICE) project (Greenbaum & Nelson 1996) has marked a substantial improvement of the empirical base for the description of many New Englishes. Over the past twenty years, this has led to a proliferation of solid, corpus-based studies (e.g. contributions in Seoane & Suárez Gómez 2016; Hundt & Gut 2012). While the importance of the ICE corpora can hardly be overstated, their pioneering role and the amount of research conducted on their basis have turned them into highly normative resources for establishing descriptive facts about individual varieties as well as for drawing cross-varietal comparisons.

The present study takes a critical perspective on this normative status. The focus is specifically on one aspect that has been part of the ICE sampling framework from the outset but has rarely been addressed in explicit statistical modeling: the role of register (Biber & Conrad 2009). The study compares the statistical power of two different extra-linguistic factors in predicting multivariate patterns of linguistic variation in ten ICE corpora: a)the variety a given ICE corpus text represents, and b)the ICE text category label of a text. Results indicate that text category labels are more effective at predicting multidimensional patterns of variation than national variety labels, usually by at least one order of magnitude.

This finding leads to a second problem, namely: omitted variable bias in ICE-based research that does not include register directly in its statistical modeling. In a case study on verb complementation, it is shown that apparent cross-varietal differences diminish in statistical significance once register is systematically accounted for. In other words, cross-varietal differences are heavily mediated by situational and genre context and tend to be exaggerated if register is not directly considered as a predictor of variation.

The results of the study should not be interpreted as a wholesale rejection of ICE-based research so much as a contribution towards methodological refinement. They suggest that, when comparing New Englishes on the basis of ICE data, register relations are an important level of explanation that deserves more explicit attention. Moreover, the multidimensional patterns of variation established on the basis of the ICE corpora can be used to gauge the register differentiation of bigger, messier corpora (e.g. GloWbE; Davies & Fuchs 2015), thus yielding a promising marriage of established, carefully-curated and more recent mega-corpora.

References

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Davies, Mark & Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide* 36(1). 1-28.

Greenbaum, Sidney & Gerald Nelson. 1996. The International Corpus of English (ICE) project. *World Englishes* 15(1). 3-15.

Hundt, Marianne & Ulrike Gut (eds.). 2012*. Mapping unity and diversity world-wide: Corpus-based studies of New Englishes*. Amsterdam: Benjamins.

Seoane, Elena & Cristina Suárez Gómez (eds.). 2016. *World Englishes: New theoretical and methodological considerations.* Amsterdam: Benjamins.