Joybrato Mukherjee
Justus Liebig University Giessen

With the Global Web-based English Corpus (GloWbE), Mark Davies and Robert Fuchs have launched an unprecedented resource for corpus-based analyses and the comparison of national varieties of English worldwide. Because of its immense size, enormous range and free availability, GloWbE will no doubt expand the horizons for research into World Englishes. The pilot studies that Mark Davies and Robert Fuchs offer in their paper exemplify the extent to which GloWbE can be utilised in future research — these are enticing and fascinating prospects.

In Giessen, we are particularly interested in English in South Asia, the world's largest postcolonial anglophone *Sprachraum*; we have already profited from the South Asian components of GloWbE (currently including India, Pakistan, Sri Lanka and Bangladesh) by re-running previous analyses of smaller corpora, in particular the Indian and Sri Lankan components of the International Corpus of English (ICE-IND, ICE-SL$_{\text{written}}$) and the six national components of the South Asian Varieties of English Corpus (SAVE; cf. Bernaisch et al. 2011), and by identifying many more instances of low-frequency phenomena in lexis and lexicogrammar.[1] For example, there is a tendency in research on South Asian Englishes to miscategorise unusual formations, e.g. new prepositional verbs such as *cope up with* (for British English 'cope with'), as dominant forms in new varieties of English. The SAVE Corpus, including 18 million words of acrolectal newspaper English from South Asia, shows, however, that South Asian speakers too prefer *cope with* to *cope up with* (i.e. in more than 90 per cent of all instances in each of the four components from India, Pakistan, Sri Lanka and Bangladesh). *Cope up with* is a relevant South Asian innovation, but a minority variant in less than ten per cent of the cases — a tendency clearly corroborated by GloWbE. However, while in any single SAVE component we are dealing with far fewer than 100 instances of both variants, the national components of GloWbE include several thousand relevant instances in total — a richness of data that enhances the confidence with which we can identify quantitative differences and preferences across World Englishes. This certainly is one of the major strengths of GloWbE.

Mark Davies and Robert Fuchs view GloWbE as an important addition to the "toolbox" of resources for studying World Englishes — and rightly so. Yet for various reasons, GloWbE cannot simply replace smaller and tidier corpora such

[1] I am grateful to Tobias Bernaisch for various discussions about GloWbE in general and the examples included in this response in particular.

English World-Wide 36:1 (2015), 34–37. DOI 10.1075/eww.36.1.02muk
ISSN 0172–8865 / E-ISSN 1569–9730 © John Benjamins Publishing Company

as ICE and SAVE. By focusing on South Asian Englishes, I would like to address two interrelated reasons in particular as to why we will continue to need corpus resources above and beyond GloWbE, namely the unknown variability of data and the heterogeneity of speakers included in GloWbE.

By its very nature, GloWbE is an extreme example of what Brezina and Meyerhoff (2014) recently referred to as the — largely inevitable — "aggregate data methodology" in corpus design. Corpus linguistics has always been primarily interested in the typical patterns of language use; thus, corpora have always been designed as representative samples of language use with the underlying assumption that corpus findings are characteristic of a prototypical average speaker of the language variety at hand. While small and controlled corpora like ICE (a general corpus) or SAVE (a specialised newspaper corpus) have been construed very carefully in order to ensure the intended representativeness in corpus design, GloWbE is in many regards unspecified or, for that matter, aggregative: apart from the specification that approximately 40 per cent of the corpus is made up of informal blogs, we do not know which types of speakers and which language variants are represented by the national web domains included in GloWbE. This raises a number of questions:

1. How can we discriminate between speakers of the acrolectal variant(s) of postcolonial Englishes on the one hand and speakers of mesolectal variants and basilectal/pidgin variants on the other? Even if we examine the individual contexts of use, the status and competence of speakers may often remain unclear.
2. Which genres are represented in what way in the national web domains of GloWbE? We know from the ICE project that the forms and functions of genres may differ enormously between varieties of English. The same also holds true for English across the national domains on the World Wide Web.
3. To what extent do the national web domains appropriately represent the national varieties of English?
4. Given the unknown heterogeneity of speaker types, language competencies, genres etc., the overall question arises as to what extent the national web domains are comparable in the first place.

Refer, for example, to the distribution of the British and American spelling variants of *colour*/*color* and *theatre*/*theater* in the South Asian domains of GloWbE. While in small and controlled corpora like SAVE, the American variants *color* and *theater* are — unsurprisingly — minority forms in India, Pakistan, Sri Lanka and Bangladesh that occur in 1–5 per cent of all relevant cases only, in GloWbE they account for 23–36 per cent (*theater*) and for 47–69 per cent (*color*) of all cases, respectively.[2] Why is that so? Do mesolectal and basilectal users of English in South Asia that may have been included in GloWbE prefer the American variants? Are the South Asian GloWbE domains characterised by American data/speakers? Is this trend relevant for particular web genres only? Or is this a general symptom of the Americanisation of Web English? One should also keep in mind that non-South Asian domains may include South Asian speakers and their linguistic output. If one looks for instances of presentational *itself*, a form shown to be characteristic of South Asian varieties in general (cf. Bernaisch and Lange 2012), one can infer from the contextual information that several of the seven instances of *today* plus *itself* in the US- and GB-domains of GloWbE must have been produced by South Asians. Obviously, in a globalised world with the Internet as a global network for communication, the fuzzy boundaries between varieties of English (and their speakers) are not identical with the rigid lines between national web domains (and their texts).

Against this background, I strongly recommend following Mark Davies and Robert Fuchs in viewing GloWbE as a very useful addition to the toolbox for corpus linguists interested in World Englishes, but also suggest using this new tool with a measure of caution. More specifically, as a big and aggregative corpus, GloWbE should best be used in combination with small and controlled corpora of varieties of English. For those interested in South Asian Englishes in particular, the advent of GloWbE will make it possible to expand the corpus-cum-web and web-to-corpus approaches that have already become popular in research into English in South Asia (cf., for example, Sedlatschek 2009; Hoffmann, Hundt and Mukherjee 2011; Hundt, Hoffmann and Mukherjee 2012).

References

Bernaisch, Tobias, Christopher Koch, Joybrato Mukherjee, and Marco Schilk. 2011. *Manual for the South Asian Varieties of English (SAVE) Corpus*. Giessen: Justus Liebig University, Department of English.

Bernaisch, Tobias, and Claudia Lange. 2012. "The Typology of Focus Marking in South Asian Englishes". *Indian Linguistics* 73, 1–18.

Brezina, Vaclav, and Miriam Meyerhoff. 2014. "Significant or Random? A Critical Review of Sociolinguistic Generalizations Based on Large Corpora". *International Journal of Corpus Linguistics* 19, 1–28. DOI: 10.1075/ijcl.19.1.01bre

[2] With regard to *color*, we have noted that a substantial part of the metalanguage code, in which the word *color* is frequently used, is included in GloWbE. This may have a distorting effect on query results.

Hoffmann, Sebastian, Marianne Hundt, and Joybrato Mukherjee. 2011. "Indian English – an Emerging Epicentre? A Pilot Study on Light Verbs in Web-Derived Corpora of South Asian Englishes". *Anglia* 129, 258–280. DOI: 10.1515/angl.2011.083

Hundt, Marianne, Sebastian Hoffmann, and Joybrato Mukherjee. 2012. "The Hypothetical Subjunctive in South Asian Englishes: Local Developments in the Use of a Global Construction". *English World-Wide* 33, 147–164. DOI: 10.1075/eww.33.2.02hun

Sedlatschek, Andreas. 2009. *Contemporary Indian English: Variation and Change*. Amsterdam: Benjamins. DOI: 10.1075/veaw.g38

Author's address

Joybrato Mukherjee
Justus Liebig University Giessen
Otto-Behaghel-Str. 10B
35394 Giessen, Germany

Joybrato.Mukherjee@anglistik.uni-giessen.de