(92)

Ooi, Vincent B. Y. (National University of Singapore, Singapore): Examining the GloWbe corpus as a lexicographic resource for Singapore, Malaysian and Hong Kong English

PANEL: CORPUS-BASED LEXICOLOGY AND LEXICOGRAPHY

With the advent of big data for linguistic research, the GloWbe corpus promises to 'expand horizons in the study of World Englishes' (Davies and Fuchs, 2015 – forthcoming). In turn, would such a corpus offer a reliable evidence base for the incorporation of World Englishes – highlighting the pluricentric nature of English as the leading global language--in the dictionary? This paper offers a modest, exploratory answer to such a question by considering certain phrases important in the contexts of Singapore, Malaysia and Hong Kong - three ESL countries which have a number of striking similarities (and differences).

First, GloWbE is said to be based on "1.9 billion words in 1.8 million web pages from 20 different English-speaking countries. Approximately 60 percent of the corpus comes from informal blogs, and the rest from a wide range of other genres and text types." (Davies and Fuchs 2015). The general idea for GloWbe, as it is for other web corpora, is to regard the 'web as corpus' (Kilgarriff and Grefenstette 2003). But, Sinclair (2004), while welcoming the WWW as 'a remarkable new resource for any worker in language', also notably warns that 'the WWW is not a corpus', if the latter is defined to be maximally representative of the linguistic phenomenon in question. Sinclair's reasons include the 'mysterious' dimensions of the Web and the varying algorithms afforded by the various search engines that do not lead to the right balance and sampling (even if size is exponentially increased).

Examining the GloWbe corpus, the Standard Singapore English phrase "killer litter" is significant in Singapore which is sensitive to the danger/injury posed by heavy objects thrown from high-rise buildings. In the GloWbe corpus, the phrase is remarkably absent in 18 countries and occurs a total of 7 times in Singapore and 1 time in the Sri Lankan context. The one time that it does occur in the Sri Lankan context, as in any given context, is not significant – given the odd migration and diffusion of English use across contexts.

In the Malaysian context, a prototypical 'Manglish' (or colloquial Malaysian English) phrase is "lepak" (meaning 'to skive' or 'to chill out'), borrowed from Malay: A search for the phrase in GloWbe yields a view of frequencies from different countries: U.S.A. (1 occurrence), Canada (2), UK (2), Australia (6), Singapore (5), Malaysia (23). In this case, 'the ability to see the frequency of any word, phrase, or grammatical construction in each of the 20 different countries' (Davies, 2013) may lead to the misleading conclusion that this Manglish phrase is most used in Malaysia, and then productively more used in Australia than in Singapore. In the Malaysian concordance for "lepak", the 'chill out' sense is retained but a closer examination of the Australian concordance shows that it refers to someone's name ('Dennis Lepak') which has no bearing to the Manglish phrase.

Turning to Hong Kong English, "shroff" is a term that many Hongkongers regard as 'standard English' because of its prevalence in parking lots/carparks ("shroff" means 'an office or kiosk, e.g. in car parks' (Bolton 2003: 295). A search for the term gives the impression that it is used across a number of countries: U.S. (4), Canada (3), UK (12), Australia (4), New Zealand (20), India (126), Sri Lanka (15), Pakistan (31), Hong Kong (16). But, unlike the case of "lepak" in which a quick distinction between upper and lower case would do, the lexicographer will have to trawl through the 'parking lot' sense from the proper name sense in the concordance listing for Hong Kong.

In this paper, a number of other linguistic examples will be used to show the internal diglossic nature of these varieties of English (e.g. between 'standard Singapore English' and 'Singlish') that lexicographers will have to sift through the data resource afforded by

the GloWbe corpus. Notwithstanding this, it would be very much apparent that the GloWbe corpus is a much welcome resource for the lexicographer to incorporate World Englishes into the dictionary.

Select References:

Bolton K. 2003. Chinese Englishes: A Sociolinguistic history. UK: Cambridge University Press.

Davies M. 2013. “New corpus: GloWbE -- 1.9 billion words, 20 countries”, in Corpora-List.

Davies M, and R Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English corpus (GloWbe), In English World-Wide 36:1. (forthcoming), pp1-29.

Kilgarriff, A and G Grefenstette. 2003. Web as corpus.

Sinclair, J. 2004. Corpus and text – basic principles. In Developing Linguistic Corpora: A Guide to Good Practice.