Sverker Johansson

Dalarna University

Patterns of preposition use across World Englishes

English is probably the most widespread language in the world, with by far the largest number of L2 speakers. Apart from the handful of countries where English is the majority L1, English is also in common use as a high-status *lingua franca* in a number of multi-lingual countries around the world, from Philippines to South Africa. The pattern of language use varies between countries, but typically English is the language of commerce and higher education, and sometimes of government.

The English varieties spoken in these countries differ from American (AE) or British English (BE) in numerous ways, and they are sufficiently established and accepted that they are increasingly recognized as valid varieties on a par with AE or BE (e.g. Kachru et al., 2009). Lexical differences have received the most attention, with both loanwords from local languages and idiosyncratic use of regular English words being common – “barangay” and “comfort room” are two Philippine English (PE) examples – but grammar, phonology, and usage also differs (e.g. Kachru & Nelson 2006).

Instead of studying individual words, I have chosen as my research question the patterns of word choice in different Englishes, with focus on prepositions. Prepositions are fairly resistant to borrowing, and form a semi-closed class of manageable size so that a near-exhaustive analysis is feasible. At the same time an English speaker has a fair amount of leeway in the choice of prepositiowns, leading to variation even within the same variety of English. And prepositions are notoriously difficult to learn for adults, leading to even larger variation between L2 speakers. My hypothesis is that the local language substrate will color the pattern of preposition choice in different Englishes, which will show up as differences in the frequencies of various prepositions in corpora from different countries.

Method

For this study, I have used the GloWbE corpus (Davies & Fuchs 2015), which contains 1.9 billion words harvested from websites in 20 different countries with established varieties of English. The individual country corpuses vary in size from 35 million to 400 million words. I used the POS tagging of the corpus to seek all occurrences of prepositions in the corpus, with results separated by country. This provided me with a large table containing the 1000 most common prepositions in the corpus, with both raw numbers and word frequencies for each country. In the actual analysis, I used the 300 most common prepositions, which means down to a frequency of about 0.01 per million words, or 20-odd occurrences in the whole corpus.

The null hypothesis used is that the true frequency of each preposition is the same in every country, with all differences in the corpora attributable to sampling effects. To test this hypothesis, I calculated the expected raw number for each preposition in each country sample, given the frequency of that preposition in the entire corpus. The expected number in a sample from a population of size N where the frequency is *p* will be a Poisson distributed stochastic variable with mean n_{exp} = *pN* and standard deviation $\sqrt{n}$. So in order to test the null hypothesis, I calculated the number of standard deviations between the actual and expected number $z = \frac{n_{actual} - n_{exp}}{\sqrt{n_{exp}}}$ for each preposition in each language, as well as for the total number of prepositions.

The resulting table with 301 standard deviation values (300 individual prepositions, plus one for the total number), for each country sample was used as input to a neighbor-joining algorithm (Saitou & Nei 1987) in order to find which countries most resemble each other in preposition-use pattern. The 301 values for a country was treated as a vector $\bar{v}$ in a 301-dimensional space, and countries were clustered according to their Euclidean distance in this space, in the following procedure:

1. For each pair of countries *i,j*, calculate the distance *d_{ij}* as $\sqrt{\sum_{k=0}^{300} (v_{ik} - v_{jk})^2}$
2. Find the smallest of all distance values.
3. Merge the country vectors $\bar{v}_i, \bar{v}_j$ of the smallest-distance pair, by replacing the two vectors with a single vector $\bar{v}_{merged} = 0.5(\bar{v}_i + \bar{v}_j)$.
4. Repeat 1-3 until everything is merged into a single vector.

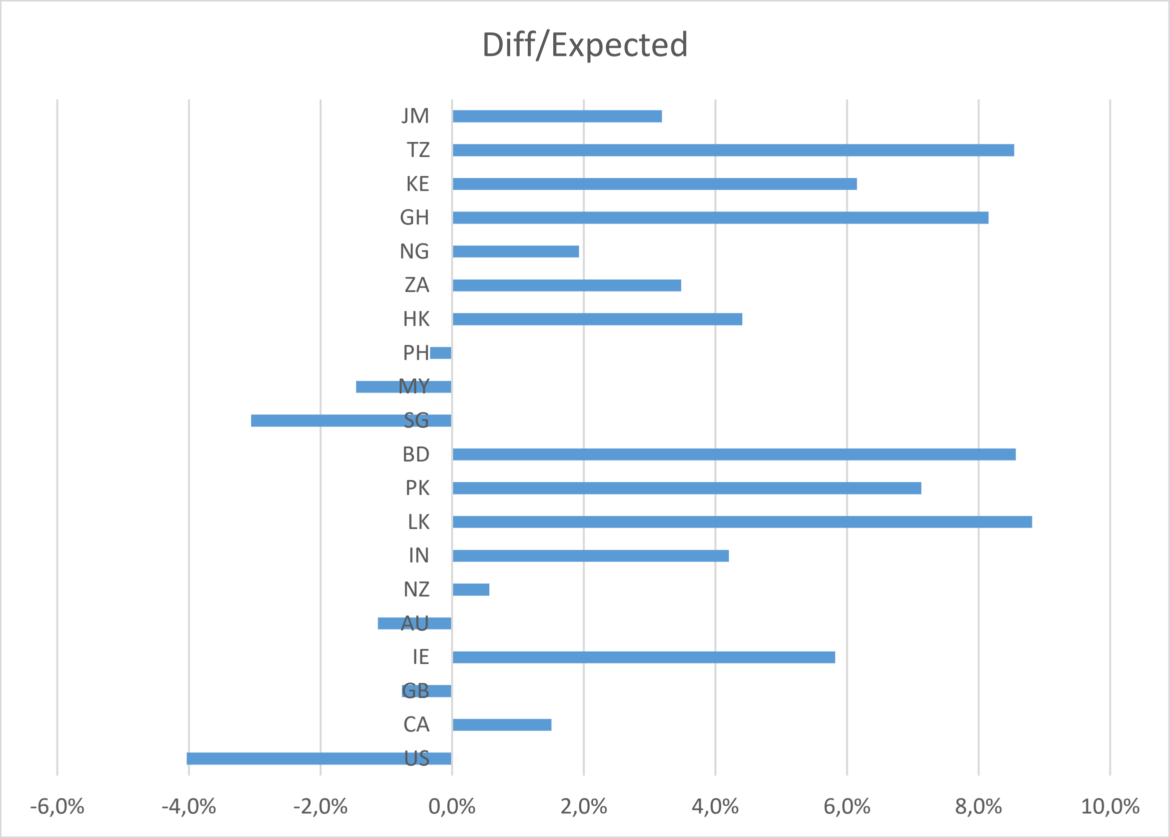
The resulting hierarchical clustering was fed into a tree-drawing algorithm. The full table is too large to include here, but can be found in the attached Excel document.

I assume here that the various country corpora are equally balanced, so that any differences between them can be attributed to the country, rather than to any other differences between the corpora. But since the whole point with the GloWbE corpus is to enable between-country comparison, this is probably a safe enough assumption – the corpus creators have presumably done their best to balance the corpus in this way.

It is also assumed that the country tagging of each sample is accurate. A quick check searching the corpus for “barangay” and “comfort room” reveals that the fraction of erroneous country attributions is not negligible. But any such errors in the corpus will serve to wash out statistical country differences; making the assumption is thus conservative with respect to the present analysis.

Results and discussion

The total frequency of prepositions in the corpus is about 10% of all words, with substantial variation between countries. The chart below shows the difference between the actual number of prepositions in a country corpus and the number expected from the total corpus frequency, as a fraction of the expected number.

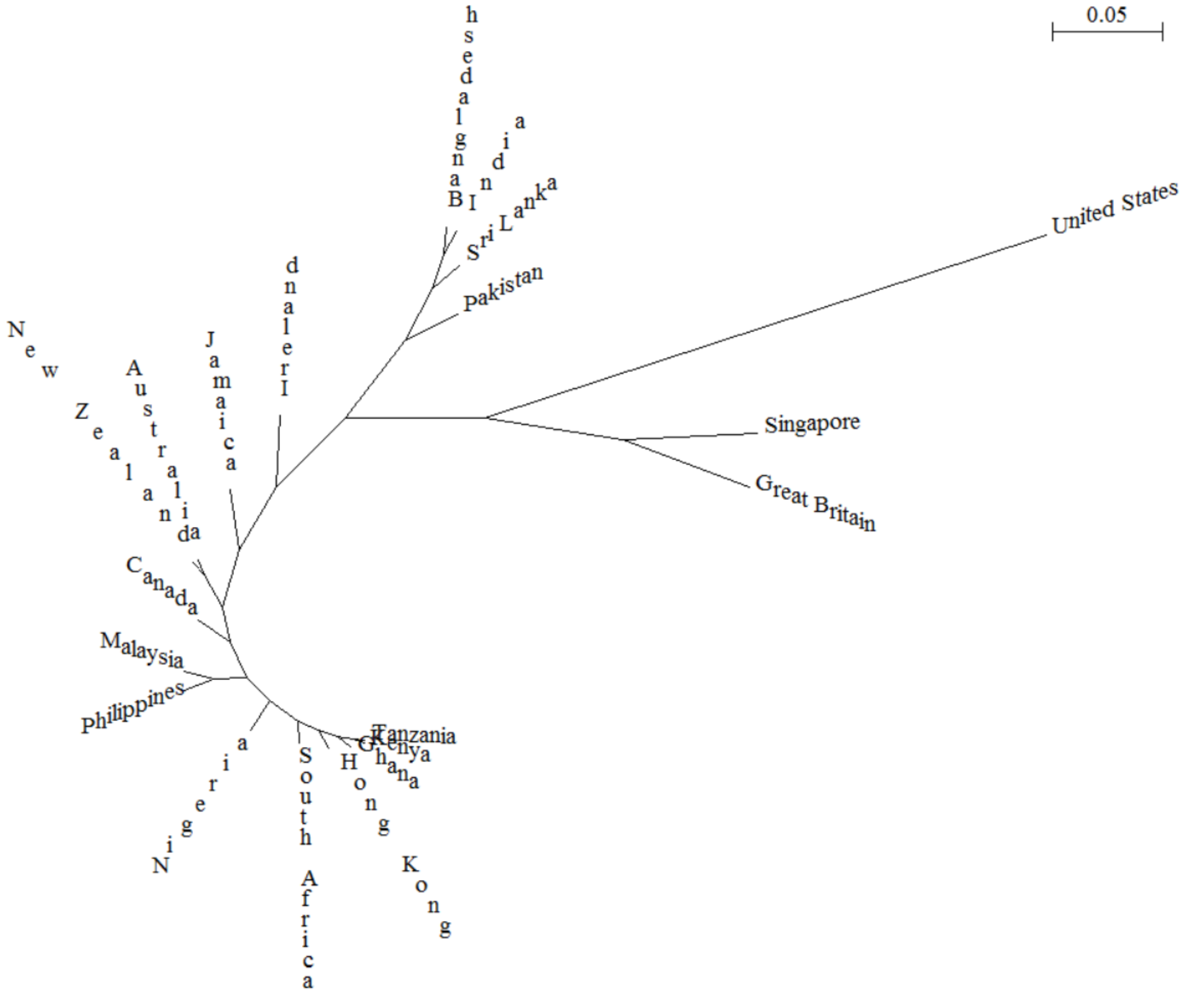


As can be seen, Americans use less prepositions than anybody else, whereas both African and South Asian Englishes are preposition-heavy. The statistical significance of these differences is solid, in the tens or hundreds of standard deviations; detailed statistical calculations are superfluous.

When the frequency patterns for individual prepositions is used to cluster countries with similar patterns, as described in the Methods section, the result is the tree which is shown below. The length of each line is proportional to the difference in preposition use between the countries joined by the line. As expected from the total frequency above, the United States is again an outlier. In the rest of the diagram, one can observe a clear South Asian cluster at the top, and an African cluster at the bottom. Somewhat surprisingly, Hong Kong is in the middle of the African cluster. Australia and New Zealand are close together, as are Malaysia and Philippines.

Some of these results make sense in the light of the linguistic demographics of the countries involved. The South Asian countries all have local languages from the Indo-Aryan and Dravidian families, and they all have a history as British colonies for a similar period of time. The Malaysia-Philippines cluster is united by Austronesian local languages, but their history is totally different. The African cluster is more complicated; Kenya and Tanzania have related local languages, and both have Swahili as one official language; having them cluster closely together makes sense. Other languages from the

Niger-Congo family are widely spoken in all these African countries. What Hong Kong is doing in the same cluster is less clear; I can see no similarity except a common history as British colonies. My prediction would have been to find Hong Kong and Singapore together, as they have both historical and demographic similarities, with Chinese/English bilingualism common in both. Possibly the much larger population of English L1-speakers in Singapore serves to unite Singaporean and British English?



Concerning individual prepositions, the table below shows for each country which preposition is the most over-used and most under-used (most significant deviations from the global frequency), plus a list of any prepositions that are predominantly used in that country and not elsewhere. For all the “local specials”, more than half of all occurrences in the world corpus are from the country listed. Not all words here look like prepositions, but all are POS-tagged as prepositions in the corpus; how reliable is the tagging?

Country	Most overused	Most underused	Local specials
United States	ABOUT	AS	favor, in-n-out, off-and-on
Canada	PURSUANT	AGAINST	goals-against, o.p.p.
Great Britain	RATHER	OF	outwith, in/out
Ireland	IN	LIKE	hyper-v, politics-from
Australia	VIA	IN	give-part-with, in-divide-u-all-s
New Zealand	AROUND	ACCORDING	
India	TILL	ABOUT	on/upon, mid-off
Sri Lanka	OF	ABOUT	
Pakistan	OF	AT	a.s., here-at
Bangladesh	OF	ABOUT	40-per, insured-against
Singapore	TAE	OF	tae, pre-a, step-into, ball-less
Malaysia	UPON	OF	pre-u
Philippines	ASIDE	ON	being-with, bel-at, in-order-to
Hong Kong	OF	ABOUT	15-per, 6-per, milter-limit-from
South Africa	TERMS	LIKE	
Nigeria	UNTO	ABOUT	as/as, sleeping-with
Ghana	OF	ABOUT	in-out-out-out
Kenya	IN	ABOUT	within/as, in/as, hold-onto, give-into
Tanzania	IN	ABOUT	one-per
Jamaica	WID	ABOUT	wid, 10-per

Some of the local specials are just orthographic variants (in-n-out vs. in/out), but other are more difficult to interpret. “Terms” in South Africa is often part of a multi-word prepositional expression “in terms of”, each component of which is apparently POS-tagged as a preposition. In any case, the whole expression remains overused in South African English. The Singaporean “tae” appears to be a tagging error – the examples I have looked at all are all components of Chinese-sounding names. Jamaican “wid” is presumably a back-borrowing from the English-based creole that is the main local language. The various “N-per” prepositions, with different N in different countries (the examples above are not exhaustive) are harder to interpret.

The next step would be to look deeper into the actual occurrences of these unusual prepositions, to check whether they are actually used as prepositions, and if so, what they mean. But that is beyond the scope of this preliminary study.

Conclusions

The main hypothesis of this study is supported, in that there are significant differences between countries in preposition usage patterns, and that countries with related local languages also have similar usage patterns. This constitutes a *prima facie* case for substrate influence, but many checks remain to be done before solid conclusions can be drawn.

References

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

Kachru, Y. & Nelson, C. (2006) *World Englishes in Asian Contexts.* Hong Kong University Press.

Kachru, B., Kachru, Y., Nelson, C., eds. (2009) *The Handbook of World Englishes.* Wiley.

Saitou N & Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.