Axel Bohmann

Contrastive Usage Profiling: A Word Vector Perspective on World Englishes

Abstract: This paper introduces Contrastive Usage Profiling (CUP), a method for quantifying relationships among varieties of English based on lexical co-occurrence patterns in large corpora. The approach is situated in relation to similar research and illustrated with a case study from the context of World Englishes. Based on the national sub-corpora of the Corpus of Global Web-based English (GloWbE, Davies/Fuchs 2015), varietal profiles are constructed for twenty varieties of English. Patterns for individual words as well as aggregate patterns for varietal differentiation based on many words are shown to yield theoretically plausible results and to remain robust across different parameter settings of the method. Model interpretability is identified as an important area for future research.

1 Introduction

In this paper, I outline steps towards contrastive usage profiling (CUP), a method for quantifying relationships among varieties of English based on lexical co-occurrence patterns in large corpora. Measuring similarities and differences among varieties in robust, statistically elaborate terms has become a recent focus in both dialectological (Grieve 2016; Szmrecsanyi 2013) and World Englishes research (Bohmann 2019; Szmrecsanyi/Grafmiller/Rosseel 2019). The present paper concentrates on the latter using online discourse from the 20 countries represented in the Corpus of Global Web-based English (GloWbE, Davies/Fuchs 2015) as a case study. However, the method can be extended to any comparison of varieties, whether these be dialects, text types, diachronic snapshots, etc., provided the respective corpora are sufficiently large to allow for the construction of robust word-vector representations.

The procedure I introduce here relies on word embeddings (also known as word vector models or distributional models) to represent word usage. Such models describe individual words by means of their co-occurrence profiles with other words. A separate word embedding model is constructed for each national sub-corpus of GloWbE. Differences between these models, i.e., between the varieties they represent, are then measured by aggregating over differences in the profiles of the most frequent individual words in the corpus on the whole. The resulting inter-varietal distance pattern remains relatively stable even after consideration

https://doi.org/10.1515/9783111017433-002

of only the 100 most frequent words. The picture of differentiation in English worldwide is theoretically plausible and shows traces of both developmental status (according to Schneider 2007) and areal groupings of varieties.

A problem that remains is to identify what factors drive the output of the CUP procedure. Capitalizing on a large number of words, and representing each by its co-occurrence patterns with other words, means the results cannot easily be traced to a single, unified explanation. In the Discussion, I outline some steps to enhance the interpretability of the results, but recognize that these are largely objectives of future research. In its present form, CUP, as detailed below, is a flexible and robust method for comparing large corpora of text and can be adapted by other researchers with relative ease.

2 Quantitative Relations among Varieties: Methodological Approaches

Measuring relationships among varieties has been central to several subfields of linguistics for much of the discipline's history. In traditional dialectology, linguistic atlases are constructed based on the degree of similarity different regions show in relation to individual – lexical, morpho-syntactic, or phonological – features. Where several isoglosses, i.e., geographical boundaries of feature distribution, overlap, borders between dialect areas are drawn. Similarly, in historical linguistics the comparative method (Hoenigswald 1960) uses feature correspondences among varieties at one synchronic stage to reconstruct common ancestor languages. Typological research proceeds along similar lines, but focuses on synchronic comparison and the discovery of pervasive relationships.

Both traditional dialectology and historical linguistics rely on categorical observations about the presence or absence of features. Recently, however, these have been complemented by approaches that employ more sophisticated methods of quantification in probabilistic rather than categorical terms. The quantitative dialectological methods developed in Salzburg (e.g., Goebl 2006) and Groningen (e.g., Heeringa/Nerbonne 2013; Nerbonne 2006) are early examples of this development, which has been more fully realized in recent studies such as Szmrecsanyi (2013) and Grieve (2016). In this perspective, covariance among large sets of linguistic features in different places is used to establish dialect areas statistically.

The comparative method from historical linguistics, likewise, has found application in quantitative terms, as prominently elaborated by Poplack/Tagliamonte (2001). In their comparative sociolinguistic approach, rather than relating

varieties as to the presence or absence of a feature, the full set of constraints conditioning individual linguistic variables is considered. The three lines of evidence in this perspective are the statistical significance of constraints, their relative strength, and the rank ordering of their importance in different varieties or communities. An extension to this approach, which directly quantifies the lines of evidence, is the "variation-based distance and similarity modeling" (VADIS) paradigm developed by (Szmrecsanyi/Grafmiller/Rosseel 2019).

In research on World Englishes, the use case under discussion here, such aggregate quantitative methods are still in the minority. Detailed investigation of isolated features based on comparable corpora remains the dominant paradigm. Such studies have much to offer in relation to the specific variables they consider; however, extrapolating from their findings to general relationships among varieties can be problematic. The assumption that the behaviour of one isolated feature is indicative of difference or similarity among varieties on the whole is often unjustified (Bohmann 2021; Hundt 2009).

A recent contribution towards grounding the description of inter-varietal relations in World Englishes in more robust aggregate terms is Bohmann's (2019) multidimensional analysis. Following the procedures pioneered by Biber (1988), and using ten corpora from the International Corpus of English (ICE) project (Greenbaum/Nelson 1996) representing educated standard English from various countries, this study extracts frequency information about 236 linguistic features for each corpus text. On the basis of this data, dimensions of variation are established that give structure to the range of varieties and registers represented in ICE. Without going into detail about the interpretation of any of these dimensions, the dominant finding is that register – whether a corpus text is a piece of fiction writing, a broadcast interview, etc. – significantly outperforms the country a text is from in structuring variation along these dimensions.

Lexical variation has received comparably sparse attention in the context of World Englishes. Most frequent are attempts to quantify the normative orientation of varieties towards British and/or American English. Gonçalves et al. (2018), for instance, demonstrate overwhelming "Americanization" on a global scale based on a large corpus of Twitter messages. They calculate the proportion of British and American words from a closed set of clearly marked alternants, e.g., *eggplant* and *aubergine*, on "a grid of cells of 0.25 ° × 0.25 ° spanning the globe" (Gonçalves et al. 2018: 4). This approach is well-suited for the specific question it is aimed at addressing, but reducing varieties of English to their relative dependence on British or American norms arguably does not do justice to the full range of differentiation to be found in World Englishes.

Another choice in the literature has been to focus on "cultural keywords," i.e., "words that are revealing of a culture's beliefs or values" (Rocci/Wariss

Monteiro 2009: 66). Mukherjee/Bernaisch (2015) adopt this perspective in an analysis of three South Asian varieties. They establish a set of words that are generally more frequent in South Asian Englishes compared to a reference corpus of British English and narrow this list down to relevant keywords through "a socio-culturally motivated selection" (Mukherjee/Bernaisch 2015: 420). For each keyword established in this way, they contrast collocates in the three varieties under discussion.

To a certain extent, the cultural keyword perspective can be seen as complementary to the one taken in Gonçalves et al. (2018). Whereas the latter subsumes the status of New Englishes under their relative adherence to British/American norms, Mukherjee and Bernaisch's (2015) perspective is firmly focused on nativization, linguistic acculturation, and locally specific usage. Their method, however, pre-selects the most distinctive items and focuses heavily on denotationally rich content words. Yet, the innovation that results from structural nativization is not limited to this level. Nativization can often be seen in collocational preferences, e.g., between verbs and prepositions and other constructional peculiarities (Schneider 2003). The more general collocational analysis presented below is based on common words in general without further pre-selection of relevant items. This choice was motivated by the fact that local innovations may be found not only in the frequency of "big" content words, but in the subtleties of how relatively common function words enter into collocation patterns.

In general, CUP is not proposed as a competitor to the approaches discussed above, each of which achieves a level of sophistication that cannot be matched by simply considering word co-occurrence profiles. Instead, the method should be seen as a complementary view achieved by zooming out from individual items of interest to a bird's eye view of varietal differentiation. The utility of CUP will depend largely on the extent to which it can plausibly be tied back to more particular, fully contextualized analyses.

3 Methodological Procedure

In the present analysis, a similar focus on pervasive patterns beyond individual variables as in Bohmann (2019) is employed. However, whereas that study draws on a catalogue of features that are attested to play a role in register and/or variety differentiation, the selection of relevant features presented here is both more comprehensive and more agnostic in regards to prior expectations. Specifically, the usage profiles of the most frequent 28,341 words in GloWbE are considered (all words that occur in all sub-corpora and with a total frequency of 1000 or more).

The profile for a given word in a given variety is encoded based on its co-occurrence behaviour with other words in that variety (see below for details). This has often been framed in terms of distributional semantics (see Erk 2012 for an overview), but in fact encompasses other aspects of word use as well, "including syntactic, semantic, and pragmatic aspects" (Hovy/Purschke 2018: 4383).

At the heart of contrastive usage profiling are word vector models, also known as word embeddings or distributional word models. These represent individual words as vectors in an N-dimensional space. The dimensions are derived from properties of large amounts of naturally occurring text, usually in the form of co-occurrence profiles. An established approach is to count for each possible pair of unique word types in a corpus how often its two members occur within close proximity to each other, e.g., within a 5-word window. The information derived from this procedure is then used to construct a vector space in which words that show similar co-occurrence behaviours are located close to each other. The mathematical details are beyond the scope of the present paper (see Erk 2012 for more details).

In the resulting vector space, proximity between words is taken to express commonalities. These commonalities can be along semantic dimensions, such as when the equation *king* – *man* + *woman* leads to a point in the vector space whose closest word is *queen*. Likewise, grammatical properties are encoded, allowing for similar calculations as the above in the form of *sitting* – *sit* + *walk* finding the word vector for *walking*. These relationships are usually developed from the patterning of surface forms without recourse to semantic or syntactic knowledge. As such, the method is not predisposed to express a particular kind of linguistic knowledge, whether grammatical, semantic, stylistic, etc.

CUP uses word vector models constructed with the word2vec algorithm (Mikolov et al. 2013) as implemented in the Gensim Python library (Řehůřek/Sojka 2010). Word2vec has seen wide application in computational linguistics due to its computational efficiency and competitive performance. Unlike approaches based on simple co-occurrence frequencies as described above, word2vec works on a predictive basis. This approach has been shown to outperform more traditional, count-based methods (Baroni/Dinu/Kruszewski 2014). Word2vec's objective is to find word vector representations that, given a training sentence, maximize the probability of encountering the sentence's words close to each other. There are, in fact, two separate training algorithms to achieve word vector representations in word2vec: continuous bag-of-words (CBOW) and skip-gram. CBOW is trained by optimizing predictions of words given a set of surrounding context words, whereas the latter attempts to predict context words from an individual target word (see Mikolov et al. 2013 for more detail). CBOW is faster and tends to achieve robust results even with smaller data sets, whereas skip-gram is able to construct

more nuanced word vectors based on very large data. CUP draws on the CBOW algorithm by default; however, choosing skip-gram instead is an option that should be considered depending on the nature of the data.

In the analysis of linguistic variation, word embeddings have not been widely utilized to date. Two notable exceptions are the studies by Hovy/Purschke (2018) and Rosenfeld (2019), both of which use an extension of word2vec, the doc2vec algorithm (Le/Mikolov 2014). Hovy/Purschke reconstruct dialect continua in the German-speaking area (Austria, Germany, and Switzerland). Training their model from a corpus of social media data and employing post-hoc geographic smoothing, the authors are able to reproduce results from established dialect atlases with high accuracy. An advantage of their method is that the results can be scaled to the desired level of granularity, e.g., in terms of how many distinct dialect areas to construct. Rosenfeld (2019), in addition to performing diachronic analyses of word usage, employs similar methods with a different geographic smoothing procedure to establish Texas English dialect regions based on Twitter messages. His research includes discussion of demographic difference as a mediator of linguistic differences.

CUP differs from these two examples, both of which draw on the doc2vec algorithm, in important ways. The latter represents document labels – such as city or district identifiers in the examples cited above – as vectors in the same space in which words are embedded. Consequently, individual words are more or less closely associated with individual cities or geographic regions. The method therefore answers questions about how the frequencies of individual words are associated with varieties. CUP instead quantifies the similarities and differences between varieties in relation to the usage profiles of individual words. It does not ask whether a given word is more or less frequent in a given variety, but whether it tends to enter into the same collocational patterns in one variety compared to another. In order to achieve such a comparison, a separate word2vec model is constructed for each variety.

In the case study below, the varieties considered are the 20 national components of GloWbE (Davies/Fuchs 2015). Comprising a total of about 1.9 billion word tokens sampled from blogs and general web sites in the different countries, there is significant variance in the corpus size for individual countries, ranging from over 380 million words (for the USA and Great Britain) down to 35 million for Tanzania. The median size of national sub-corpora is 44,169,602 words. The choice of GloWbE is opportunistic, as large amounts of data are required to construct word embeddings. This does not mean, however, that it should be seen uncritically. Loureiro-Porto (2017) identifies some important issues in GloWbE's composition, the most relevant for the present context being a tendency to under-represent

genuinely local usage and to over-represent Americanisms. A degree of levelling is therefore expected in the corpus that will make it more difficult to find local differences.

The goal of CUP is to achieve comparability of word usage in the 20 varieties at a general level. While this makes it desirable to include as many individual words as possible, several factors impose restrictions in this regard. Most importantly, words that occur only in a sub-set of the corpora pose problems, since their vector representations cannot be learned for all varieties. This motivates the exclusion of all such items, which are generally low-frequency and often locally specific. In order to keep computational complexity manageable, the additional restriction is imposed that a word has to occur with a total frequency of at least 1,000 (amounting to a normalized frequency of about 0.5 pmw). This threshold is fundamentally arbitrary and subject to further modification, depending on how much or little data CUP requires to arrive at stable inter-varietal distance profiles. After these exclusions, a total of 28,341 unique surface forms are retained for further analysis.

Next is the problem that word embeddings are abstract spaces that are not directly comparable. The vector for a given word in the vector model for Jamaican English cannot immediately be related to that for the same word in New Zealand English, etc., because neither the origins of the coordinate systems for each variety nor the individual dimensions of each vector space are in themselves meaningful. What is comparable across models, however, is the distance between individual words. For instance, if the word *biscuit* is found to be closer in vector space to *tea* in British English than in American English, but closer to *gravy* in the latter, this fact expresses a meaningful aspect of lexical variation. Drawing on this property, CUP represents each word under analysis, for each variety, as the vector of its distances to all other words (according to the selection criteria outlined above) in that variety. For each pair of varieties, then, the cosine distance of the two word-distance vectors for a given word can be calculated. Doing this for each pairing of varieties, a distance matrix can be constructed representing the (dis)similarity of varieties to each other. Tab. 1, for the word *language*, is an example of such a matrix, abbreviated to the alphabetically first nine varieties in GloWbE.

The steps detailed above create separate distance profiles for individual words. These can be visually inspected for qualitative interpretation and utilized for proof-of-concept. However, the profile for any one word retains only isolated information. To arrive at an aggregate view, the general tendency behind many words needs to be quantified. This is achieved by simply summing distance matrices. One question in this regard is how to treat words with different overall frequencies. It is apparent that the profiles of highly frequent words should contribute

Tab. 1: Sample CUP distance matrix for the word *language*.

	AU	BD	CA	GB	GH	HK	IE	IN	JA
AU	0	0.21	0.11	0.09	0.24	0.20	0.12	0.14	0.19
BD	0.21	0	0.21	0.21	0.27	0.24	0.21	0.19	0.24
CA	0.11	0.21	0	0.08	0.24	0.18	0.11	0.14	0.17
GB	0.09	0.21	0.08	0	0.23	0.19	0.09	0.12	0.17
GH	0.24	0.27	0.24	0.23	0	0.28	0.25	0.23	0.24
HK	0.20	0.24	0.18	0.19	0.28	0	0.18	0.19	0.23
IE	0.12	0.21	0.11	0.09	0.25	0.18	0	0.14	0.17
IN	0.14	0.19	0.14	0.12	0.23	0.19	0.14	0	0.19
JA	0.19	0.24	0.17	0.17	0.24	0.23	0.17	0.19	0

more strongly to the aggregate measure of inter-varietal distances than infrequent ones. However, word occurrences generally follow a power-law distribution in which the most common items are so much more frequent than all others that scaling distance matrices by raw frequency amounts to disregarding the majority of words entirely. Instead, as is common practice (e.g., van Heuven et al. 2014), the contribution of individual words is scaled by the natural logarithm of their frequency of occurrence.

The outcome of this analysis is a matrix containing pairwise distances generalized over all of the 28,341 words. These can then be used in hierarchical clustering to represent the relationships among individual varieties. Compared to other clustering solutions, hierarchical clustering has the benefit of not requiring a set number of clusters. Instead, the entirety of the data is represented in a tree diagram (dendrogram) where each branching node corresponds to a subdivision creating an additional cluster. Inspection of such trees can reveal the most basic splits in the data as well as the immediate relationships of individual items to each other. Specifically, CUP, as presented below, uses hierarchical agglomerative clustering with Ward's (1963) minimum variance as a linkage method.

4 Individual Word Usage Profiles

Before discussing the end result of the CUP procedure, i.e., the aggregate picture of cross-variety distance, it is useful to consider the profiles of individual words. Doing so illustrates the results below in more concrete terms and helps to test the plausibility of the method in relation to specific terms. As such, Fig. 1 shows the profiles for six selected words. For illustration purposes, the optimal

number of clusters is calculated by means of the dynamicTreeCut (Langfelder/Zhang/Horvath 2016) package in R (R Core Team 2020), and individual varieties' cluster membership represented by different font colours. Since the colour-coding is illustrative rather than essential for interpretation, and since the procedure for finding the optimal number of clusters would require a lengthier explanation, the reader is referred to Langfelder/Zhang/Horvath (2009).

The top two panels in Fig. 1 show items chosen for their cultural distinctiveness. To the left, *english* shows a first split that may be interpreted in relation to the linguistic situation in each country. The left branch, in red, comprises countries in which English is clearly the dominant language. This is obvious in the case of New Zealand, Great Britain, Australia, Canada, and the United States. The remaining two countries, Ireland and Jamaica, require qualification. The official language of Ireland is Irish, with English constitutionally "recognized as a second official language" (Constitution of Ireland, Article VIII, § 2).

However, despite language policy efforts, Irish continues to have a small native speaker base while English dominates in everyday communication. In Jamaica, Jamaican Creole is more widely spoken than English. However, the distinction between the two languages, descriptively accurate as it may be, is not normally made in everyday discourse. The countries in the right branch all feature more intense levels of societal multilingualism, and in most, the majority of inhabitants are not native speakers of English. The second split in the tree, further differentiating these countries into a Southeast-Asian and an African-South-Asian group, is less relevant here.

In the top right of Fig. 1, *holy* was chosen for its obvious religious meaning. The first split, separating Pakistan from all other countries, requires explanation in terms of a peculiarity of GloWbE. The word *holy* is significantly overrepresented in the Pakistan sub-corpus compared to all other parts of GloWbE, with a per-million-word frequency of 545, i.e., 7.5 times the global average and almost four times as high as the next most frequent country (Philippines).

More interesting is the second split, creating an almost perfect distinction between countries in which Christianity is and those where it is not the dominant religion. Nigeria is an in-between case, with Islam being slightly more widespread than Christianity. However, the material in GloWbE-Nigeria appears to contain more Christian than Muslim references: the search term "god" is about 20 times more frequent than "allah" (60,344 and 3,152 respectively), and "bible" (7,097) occurs about seven times as often as the sum of "quran" (914) and "koran" (205). The only consistently puzzling country remaining is South Africa. The difficulty in relation to this country is not limited to the word *holy*. In all plots below, South Africa and Sri Lanka are the two countries that form the tightest minimal cluster (in other words, the last split to occur is always the one

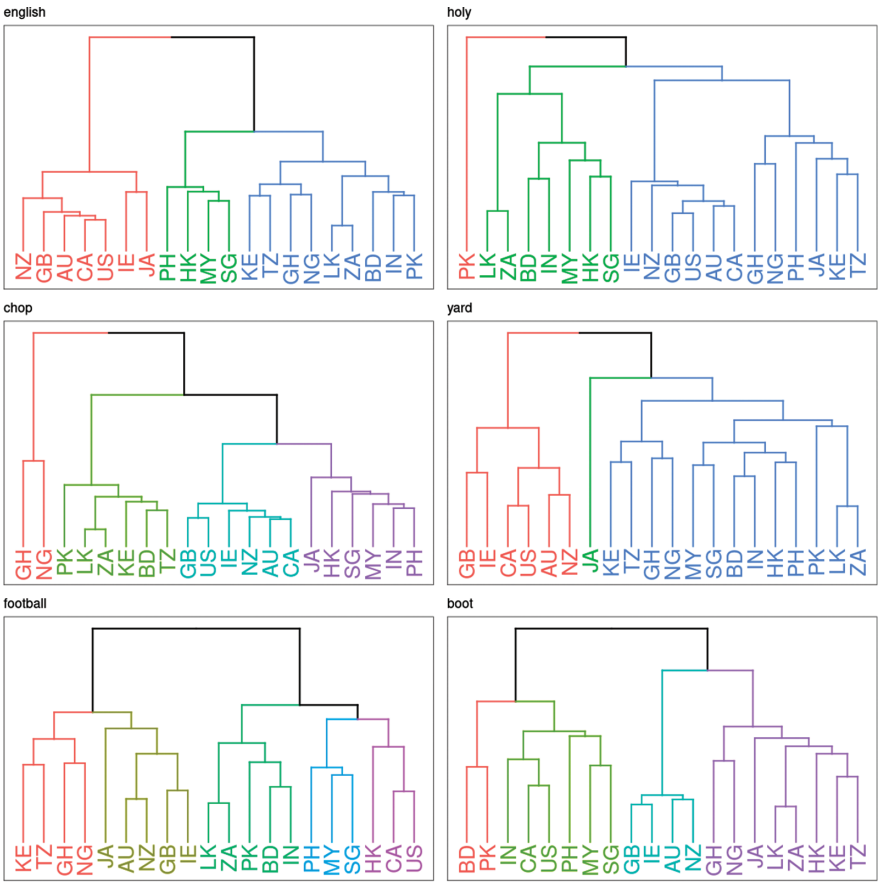

Fig. 1: CUP profiles for six selected content words in 20 varieties of English.

separating these two), often leading to implausible group membership for the former. The reasons are unclear at present and require further analysis.[1]

Moving on to the middle row of Fig. 1, the dendrograms for *chop* and *yard* were chosen because both items represent innovative uses in particular varieties.

1 At the time the article is going into publication, I have been able to identify the reason for this unexpected behavior. The offline version of GloWbE, which has to be purchased from english-corpora.org, contains a file each for South Africa and Sri Lanka, comprising over 3 million words, with completely identical content. This is, then, an obvious problem of corpus compilation, not of the CUP method. Scholars interested in working with the offline version of GloWbE should be aware of this fact and consider whether the issue persists in the version of the corpus they have available.

In West African Englishes, *chop* refers to eating, whereas in Jamaica, the word *yard* has generalized into the meaning of 'home'. Both these local idiosyncrasies are clearly reflected in the CUP dendrograms. An anonymous reviewer points out that *chop* also has an idiosyncratic meaning in Hong Kong, where it means "to stamp a document," and that this should also be reflected in the dendrogram. This point is well taken. Qualitative consideration of the 253 instances of *chop* in GloWbE-HK shows that indeed 53 of them are used with this meaning. The fact that Hong Kong is not clearly shown as separate from other varieties, however, may be explained by the fact that the "stamp" use of *chop* exists in other subcorpora as well. For instance, example 1) is from GloWbE-SG and example 2) from GloWbE-MY.

1) The stamp chop of your company must be affixed (GloWbE-SG)
2) The use of company stamp, chop and personal seal shall be discontinued (GloWbE-MY)

Nonetheless, this example points to several limitations of the CUP method. First, it does not include an option for disambiguating between homographs or polysemous items. It may be that the locally specific usage in GloWbE-HK, despite making up about 20% of all cases of *chop*, gets suppressed by the predominant general usage. Second, once a split (or lack thereof) in the dendrogram is noted, it is possible to look for explanations by considering examples from the corpus data. Yet the precise mathematical link between the structure of the dendrogram and specific kinds of usage cannot easily be established. Developing procedures to make this link more tangible is a major desideratum for future work.

The bottom row of Fig. 1 shows two words that clearly show different usage in British and American English. Whereas *football* refers to two different sports, *boot* in British English refers to the part of a car that would be called *trunk* in American English. The left panel indicates a first split that creates two groups, the left of which contains countries in which football in the British English sense is widely played. The countries in the right branch all feature popular sports other than football. In relation to *boot*, the difference between an American and a British sphere of influence is even clearer in the first split. The British group, to the right, is further divided in a second split, into the core settler varieties and the formerly colonized countries. The American group contains a couple of questionable candidates, notably the South Asian varieties. As with the close link between South Africa and Sri Lanka, more work would need to be done to shed light on this pattern.

These six examples, selected on the basis of theoretical expectations, show that CUP is able to produce plausible results at the level of individual words. Important issues remain in regards to homography/polysemy, and the precise

structure of a given dendrogram cannot easily be attributed to specific explanations. These limitations notwithstanding, the method's strength, discussed in the next section, is its ability to average over many individual words' usage profiles.

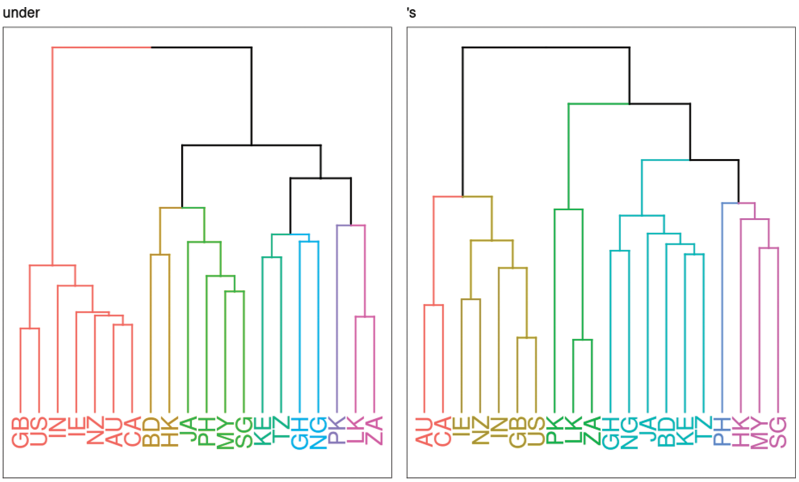

Fig. 2: CUP profiles for the function words "under" and "'s" in 20 varieties of English.

Before moving on to the aggregate picture, Fig. 2 illustrates a different aspect of CUP, once again on the basis of individual items. The two surface forms chosen here, the word *under* and the sequence *'s*, do not come with clear expectations as to the cross-varietal differentiation they create. As a basic function word, *under* is clearly part of the core vocabulary of English everywhere. The *'s* sequence is interesting because it may represent genitive case marking or enclitic versions of *is* and *has*. As such, some register sensitivity may be expected, but not strong cross-varietal differentiation. Yet, Fig. 2 shows that both these items, in fact, create more fine-grained groupings of countries than the content words discussed above. In Fig. 1, the number of clusters identified as optimal ranged between three and five. With seven and six clusters respectively, *under* and *'s* produce more nuanced profiles. This fact underlines two aspects of CUP: first, that the underlying word vector model not only captures semantics, but more general aspects of word usage; and second, that differences between varieties of English should not only be ascribed to lexical words, as cultural keyword analysis tends to do. Instead, the collocational preferences of high-frequency function words like prepositions are a rich area of structural nativization and should be considered alongside denotationally "heavy" items.

5 Aggregating Usage Profiles of Many Words

With these exploratory remarks established, it is now time to consider the big-picture view of cross-varietal differentiation suggested by a CUP for the varieties of English covered in GloWbE. Fig. 3 shows the clustering solutions produced on the basis of combined distance matrices for the most common 100, 1,000, and 10,000 words in the corpus, as well as the final diagram based on all words that meet the inclusion criteria specified in section 3. The number of groups in each tree was kept constant at 4 in order to facilitate the discussion of similarities and differences.

The general impression is one of relative stability. In all four diagrams, there is an important first split, followed at quite a distance in height by two further, almost co-occurring splits. The four groups of countries created in this way appear consistent on the whole, with a few varieties showing inconsistent group membership across the four diagrams.

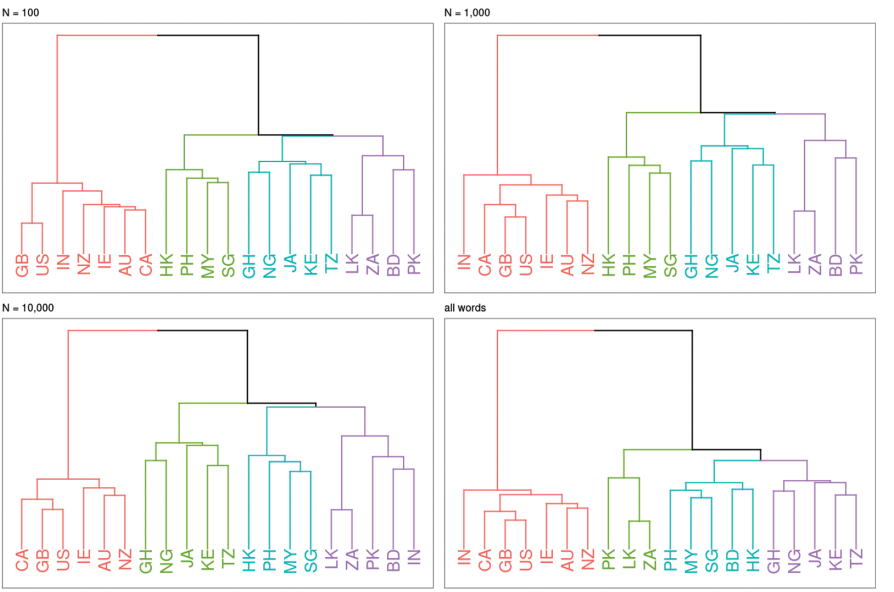

Fig. 3: Aggregate CUP profiles after the most frequent 100, 1,000 and 10,000 words as well as all words in GloWbE.

The first split constitutes a relatively clear division between British English and former settlement colonies to the left and formerly colonized nations to the right. In Schneider's (2007) evolutionary model, the countries to the left are those that

have progressed furthest along the trajectory of postcolonial linguistic independence, having entered the fifth and final stage of the evolutionary process, 'differentiation.' Only one variety troubles this view: India, featuring in the left branch in three out of the four dendrograms. Situated somewhere between the third and fourth stage in Schneider's model and sharing a history of forceful colonization with most countries in the right branch, the inclusion of India among the phase five group is not immediately plausible. One explanation might be that Indian English continues to follow a British normative model closely, but in this case, one would expect India to be closer to Great Britain throughout.

The formerly colonized countries in the right branch are further sub-divided into areal clusters. The order in which these appear in each tree is an effect of how the second and third split separate the data. Given that these two splits occur at almost the same height, differences in the order of the three areal groups across dendrograms are of little consequence. The most robust group shows up consistently in all four clustering solutions and comprises Nigeria, Ghana, Kenya, Tanzania, and Jamaica, showing a clear African profile with Jamaica as the odd variety out. However, with reference to the African ancestry shared by the majority of Jamaicans, including substrate influence from African languages, the patterning of Jamaica among African varieties is not entirely implausible.

Similarly robust is the (South-) East Asian cluster, containing Singapore, the Philippines, Malaysia, and Hong Kong. These countries pattern together in all four dendrograms, being joined by Bangladesh only in the bottom right panel based on the largest number of words.

The least consistent group are the South Asian countries India, Pakistan, Bangladesh, and Sri Lanka. While all four dendrograms show a group that might be interpreted as representing this area, none of these groups is internally pure or consistent. It has already been noted above that South Africa shows up as closely related to Sri Lanka throughout the CUP analysis. This leads to the inclusion of South Africa among the tentatively labelled South Asian clusters in all cases. Similarly, India only makes a brief appearance in the areal cluster at N=10,000, whereas it patterns with the phase five countries in all other panels. Bangladesh and Pakistan appear consistent in their participation in the South Asian cluster with the exception of the dendrogram based on all words, which sees Bangladesh switch groups and join the (South-) East Asian cluster. As a country on the borderline between these two regions, this behaviour is not altogether surprising.

Approaching the dendrograms from the opposite perspective, i.e., looking at the most immediate connections between countries, similarly plausible pairs emerge, with the exception of South Africa and Sri Lanka. Australia and New Zealand, Kenya and Tanzania, as well as Nigeria and Ghana are among the lowest-level clusters, indicating sensitivity to smaller-scale areal patterns than the

ones discussed above. The fact that Great Britain and the United States also form a tight micro-cluster speaks to their shared history as well as their position as globally dominant varieties in the world system of Englishes (Mair 2013). This view also underlines the limited theoretical purchase of attempts to treat other varieties of English as normatively dependent on either British or American English. For the most part, CUP shows other Englishes to be different from both British and American English.

Finally, a brief remark is in order in relation to the parameter settings chosen for the CUP reported here. There are considerable levels of choice in regards to at least the following variables: the frequency cut-off to include words in the analysis, the metric to represent the distance between varieties in their word usage, the question of what kind of item to focus on (surface forms vs. preprocessed data containing lemma and part-of-speech information), and the relative weighting of words by their frequency of occurrence. Space limitations prevent a detailed discussion of each of these choices; yet, it is obvious that a CUP method is preferable that does not produce vastly divergent results depending on how each parameter is set.

In order to explore this aspect of CUP, solutions were run with variations to the parameters mentioned above: once with no frequency cut-off, i.e., including all words that occurred at least once in each national sub-corpus of GloWbE, once with a Euclidean distance measure instead of cosine distances, once with (part-of-speech-tagged) lemmas instead of surface forms, and once with individual word profiles scaled by their raw rather than log frequency. Aggregate distance matrices for each of these were calculated for the first 10, 100, 1,000, and 10,000 most frequent words. A Mantel test for the correlation between the solutions presented in Fig. 3 above and each of the new variations was performed at each of these four steps, with Spearman's rho as the chosen correlation coefficient. Fig. 4 visualizes the results.

With the exception of lemma-based distance profiles for relatively few words, all correlations are strongly positive, with a rho above 0.9. With larger sample sizes, there is a tendency for the correlations to increase in strength, except when word profiles are scaled to their raw instead of log frequency. This is plausible since the effect of the scaling will increase with a wider range of raw frequencies, which in turn increases as more low-frequency words are considered. However, after an initial decrease, at about N=1,000, the correlation stabilizes to a rho of ~0.95. Without going into any further details, Fig. 4 indicates a surprising robustness of the CUP method against manipulation of individual parameters. The results are encouraging, for instance, in relation to developing CUP analyses on the basis of other data, which may not come in lemmatized and part-of-speech tagged form. They also indicate that consideration of a

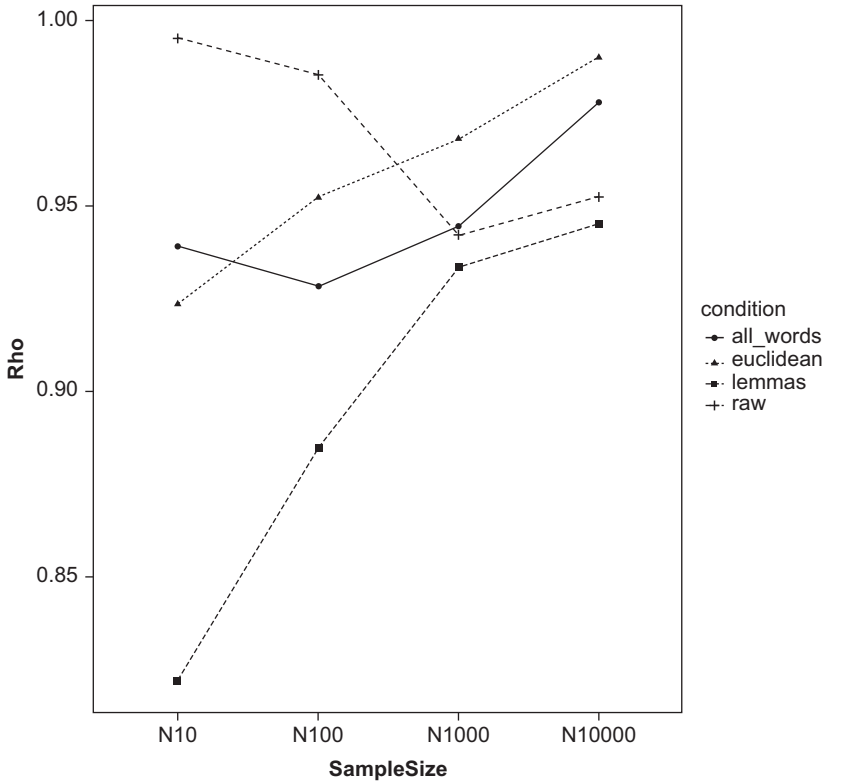

Fig. 4: Spearman's correlations for CUP results with various modifications at different levels of aggregation.

relatively small sample of all words may be enough to reach robust CUP results, thus promising computational efficiency where needed.

6 Discussion

The results presented above are encouraging. Without any information beyond co-occurrences of surface forms, the CUP procedure was able to uncover relationships among varieties, both in regards to individual words and in an aggregate view, in good accordance with theoretical expectations. Against the context of World Englishes research in particular, the results indicate a system of inter-varietal differentiation that is structured along two axes. First, former settlement countries in what is traditionally referred to as the "Inner Circle" (Kachru 1985)

behave significantly differently from formerly colonized ones. Secondly, the latter are not so much differentiated by their linguistic emancipation as per Schneider (2007) but rather pattern into areal groups. The role of British and American English as competing spheres of influence – a popular notion at least since Strevens (1980) – was not confirmed in the CUP analysis. Most other varieties are different both from British and American English rather than being more drawn to one or the other.

A key question that remains is how to explain the aggregate dendrograms in Fig. 3. What motivates the relationships between varieties as shown in these diagrams? Does CUP capitalize on cultural discourse patterns, on structural innovations in different countries, on some hidden aspects not considered so far? The fact that word vector models represent usage in a very general sense, comprising various levels of description like semantics, grammar, and style, is a strength in terms of the comprehensive view provided by CUP. When interpreting the results, however, it turns into a double-edged sword. Figures 1 and 2 certainly seem to indicate that both (culturally specific) semantics and grammatical idiosyncrasies are captured by the method, but the extent to which each plays a role deserves further attention.

To that end, it will be necessary to develop methods for post-hoc analyses of a given CUP solution. These should ideally be able to show which groups of words are most relevant for a particular split. For these relevant words, further, more qualitatively informed collocational analyses could then be constructed, thus re-anchoring the method in contextualized corpus data. For instance, comparing the closest neighbours of a given word in each variety's vector space could give insight into what it is that causes cross-varietal differentiations. I am currently in the process of developing principled steps in this direction.

Beyond the specific context of the present study, CUP as a method may be useful for any research interested in contrastive relationships among varieties broadly conceived. These may be defined historically, stylistically, regionally or otherwise. All that CUP presupposes is a sufficiently large collection of electronic text to represent each variety, and that they share large parts of their respective vocabulary. The question of how large a corpus needs to be for CUP to produce meaningful results is not easy to answer with mathematical precision. Future experience and dedicated simulation studies should be able to shed light on the relationship between corpus size and the robustness of CUP results. The smaller sub-corpora considered are 35 million words large, which can be taken as a preliminary conservative estimate of "large enough." Whether smaller corpora, e.g., the International Corpus of English, which contains 1 million words per variety, may also produce robust CUP results requires further empirical confirmation.

7 Conclusion

Above, I have outlined the methodological steps for an innovative perspective on cross-varietal distance, dubbed contrastive usage profiling (CUP). The method draws on algorithms that have been implemented in popular programming languages like Python and are consequently fairly easily available to the research community at large. The added analytical steps can be computationally expensive, but not prohibitively so. The method can still be implemented on a mid-end personal computer with a couple of hours of runtime.

The results of the case study on differences between national varieties of English have revealed an important differentiation between countries in phase five according to Schneider (2007) and formerly colonized countries that are still in the process of postcolonial linguistic emancipation. The latter further cluster into areal groups. This finding emerges from consideration of relatively few surface forms and remains largely consistent as more forms are considered. It is also robust against manipulation of individual parameters such as the choice of part-of-speech tagged lemmas instead of surface forms or the metric used to calculate the distance between two varieties for a given word. At present, CUP is an experimental method awaiting further methodological refinement and empirical validation. Still, the results so far are promising.

References

Baroni, Marco/Dinu, Georgiana/Kruszewski, Germán (2014): "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors." In *52nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1. Long Papers. Baltimore*. Association for Computational Linguistics, 238–247.

Biber, Douglas (1988): *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Bohmann, Axel (2019): *Variation in English Worldwide. Registers and Global Varieties*. Cambridge: Cambridge University Press.

Bohmann, Axel (2021): "Register in World Englishes research." In: Britta Schneider/Theresa Heyd (Eds.): *Bloomsbury World Englishes*. London: Bloomsbury, 80–96.

Davies, Mark/Fuchs, Robert (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, 1–28.

Erk, Katrin (2012): "Vector space models of word meaning and phrase meaning: A survey." In: *Language and Linguistics Compass* 6, 635–653.

Goebl, Hans (2006): "Recent advances in Salzburg dialectometry." In: *Literary and Linguistic Computing* 21, 411–435.

Gonçalves, Bruno/Loureiro-Porto, Lucía/Ramasco, José J./Sánchez, David (2018): "Mapping the Americanization of English in space and time." *PLoS ONE* 13(5): e0197741.

Greenbaum, Sidney/Nelson, Gerard (1996): "The International Corpus of English (ICE) project." In: *World Englishes* 15, 3–15.

Grieve, Jack (2016): *Regional Variation in Written American English*. Cambridge: Cambridge University Press.

Heeringa, Wilbert/Nerbonne, John (2013): "Dialectometry." In: Frans Hinskens/Johan Taeldeman (Eds.). *Language and Space. An International Handbook of Linguistic Variation (Handbooks of Linguistics and Communication Science)*. Vol. 3. Berlin: Mouton de Gruyter, 624–645.

Heuven, Walter J. B. van/Mandera, Pawel/Keuleers, Emmanuel/Brysbaert, Marc (2014): "SUBTLEX-UK: A new and improved word frequency database for British English." In: *Quarterly Journal of Experimental Psychology* 67, 1176–1190.

Hoenigswald, Henry M. (1960): *Language Change and Linguistic Reconstruction*. Chicago: The University of Chicago Press.

Hovy, Dirk/Purschke, Christoph (2018): "Capturing regional variation with distributed place representations and geographic retrofitting." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 4383–4394.

Hundt, Marianne (2009): "Colonial lag, colonial innovation, or simply language change?" In: Günter Rohdenburg/Julia Schlüter (Eds): *One Language, Two Grammars? Morphosyntactic Differences Between British and American English (Studies in English Language)*. Cambridge: Cambridge University Press, 13–37.

Ireland (1945): *Bunreacht na hÉireann = Constitution of Ireland*. Dublin.

Kachru, Braj (1985): "Standards, codification and sociolinguistic realism: The English language in the Outer Circle." In: Randolph Quirk/H. G. Widdowson (Eds.): *English in the World. Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press, 11–30.

Langfelder, Peter/Zhang, Bin/Horvath, Steve (2016): *DynamicTreeCut: Methods for detection of clusters in hierarchical clustering dendrograms*. Online at: https://CRAN.R-project.org/package=dynamicTreeCut <03.11.2021.>

Langfelder, Peter/Zhang, Bin/Horvath, Steve (2009): *Dynamic Tree Cut: In-depth description, tests and applications*. Online at: https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/BranchCutting/Supplement.pdf <03.11.2021.>

Le, Quoc V/Mikolov, Tomas (2014): "Distributed representations of sentences and documents." arXiv:1405.4053 [cs], 1–9.

Loureiro-Porto, Lucía (2017): "ICE vs GloWbE: Big data and corpus compilation." In: *World Englishes* 36, 448–470.

Mair, Christian (2013): "The world system of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars." In: *English World-Wide* 34, 253–278.

Mikolov, Tomas/Chen, Kai/Corrado, Greg/Dean, Jeffrey (2013): "Efficient estimation of word representations in vector space." *ICLR Workshop*. arXiv:1301.3781 [cs.CL], 1–12.

Mukherjee, Joybrato/Bernaisch, Tobias (2015): "Cultural keywords in context: A pilot study of linguistic acculturation in South Asian Englishes." In: Peter Collins (Ed.): *Grammatical Change in English World-Wide*. Amsterdam: John Benjamins, 411–436.

Nerbonne, John (2006): "Identifying linguistic structure in aggregate comparison." In: *Literary and Linguistic Computing* 21, 463–475.

Poplack, Shana/Tagliamonte, Sali A. (2001): *African American English in the Diaspora*. Malden: Blackwell.

R Core Team (2020): R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Online at: https://www.R-project-org/ <03.11.2021>.

Řehůřek, Radim/Sojka, Petr (2010): "Software framework for topic modelling with large corpora." In: *LREC 2010 Workshop on New Challenges for NLP*. 46–50.

Rocci, Andrea/Monteiro, Márcio Wariss (2009): "Cultural keywords in arguments: The case for interactivity." In: *Cogency* 1, 65–100.

Rosenfeld, Alex B. (2019): *Computational models of changes in language use*. Austin, TX: The University of Texas at Austin dissertation.

Schneider, Edgar W. (2003): "The dynamics of New Englishes: From identity construction to dialect birth." In: *Language* 79, 233–281.

Schneider, Edgar W. (2007): *Postcolonial English. Varieties Around the World*. Cambridge: Cambridge University Press.

Strevens, Peter (1980): *Teaching English as an International Language*. Oxford: Pergamon Press.

Szmrecsanyi, Benedikt (2013): *Grammatical Variation in British English Dialects. A Study in Corpus-Based Dialectometry*. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt/Grafmiller, Jason/Rosseel, Laura (2019): "Variation-based distance and similarity modeling: A case study in World Englishes." In: *Frontiers in Artificial Intelligence* 2, 1–14.

Ward, Joe H. (1963): "Hierarchical grouping to optimize an objective function." In: *Journal of the American Statistical Association* 58, 236–244.