CLAUDIA LANGE

'Mine is bigger than yours' or rather 'Small is beautiful'? (Re)contextualizing Postcolonial Englishes

1. Introduction

There can be no doubt that the field of Postcolonial Englishes (PCEs) has been advanced on an unprecedented scale by, firstly, the International Corpus of English (ICE) project and subsequently many other corpus projects aimed at representing the global unity and diversity of the English language. Indeed, the majority of articles published in the main journals serving the field are now corpus-based in one way or another, subjecting their data to more or less sophisticated statistical techniques to arrive at valid correlations between dependent and independent variables. However, Kilgarriff has a point when he concedes that "the fact that a relation between two phenomena is demonstrably non-random, does not support the inference that it is not arbitrary" (2005, 272). Kilgarriff's proposal to overcome this problem is "to use more data" (273), now readily available when we turn to the web as the source for corpus data (cf. Hundt et al. 2007; Davies and Fuchs 2015).

Mair (2013), on the other hand, has demonstrated that size is not everything in corpus linguistics and that for specific research questions, an approach which combines quantitative and qualitative perspectives may be more appropriate, specifically with regard to spoken language as the locus of language change from below. This paper, then, will attempt a bird's eye view of some current research on Postcolonial Englishes with respect to the two conflicting paradigms – 'Mine is bigger than yours' vs. 'Small is beautiful.' Of special interest will be the question of how we contextualize our findings, i.e. in how far we can be certain that a correlation is not only non-random and not arbitrary, but actually represents a salient linguistic pattern within a specific speech community.

The argument presented in this paper will proceed as follows. Chapter 1 will spell out the rationale for relying so heavily on quantitative methods and quantified correlations in research on PCEs. Chapter 2 will juxtapose the prototypical 'big' corpus with the prototypical 'small' corpus and make a case for the inclusion of the kind of data that is most difficult to process in corpus projects and therefore unlikely to be 'big,' namely of spoken language. Chapter 3 will point out some temptations and conceptual traps lurking behind the ready availability of all kinds of corpora for our research. Finally, the conclusion will draw together the arguments put forward in the course of the paper in a plea for (re-)contextualizing our research on PCEs.

2. Quantitative, Not Qualitative

Edgar Schneider's Dynamic Model of the evolution of Postcolonial Englishes has provided the scholarly community working in the field with an invaluable research tool, as witnessed by the countless publications whose research agendas rely on his model. The Dynamic Model offered a unified framework to classify, describe and

analyse PCEs which used to be categorized according to binary labels such as native vs. non-native, 'normal' languages vs. Pidgins and Creoles, first vs. second language etc.

A substantial body of empirical work has contributed to the issue of the linguistic correlates of a PCE's position on Schneider's evolutionary trajectory, mostly (but not exclusively) based on the parallel corpora provided by the ICE-project. Finding or confirming evidence for assigning a PCE to stage 3, structural nativization ("the central stage during which cultural and linguistic transformations take place" (Schneider 2014, 11)), typically involves an emphasis on divergence from the historical input variety: if a specific PCE can be shown to have incorporated a set of innovative features in different realms of its structure, then it can securely be placed in stage 3. 'Innovative' in this context does not necessarily mean 'unique' or 'novel;' Schneider explicitly notes that differences between the 'Old' and the 'New' Englishes are more likely to be quantitative rather than qualitative:

Innovations and distinctive structural properties of PCEs are frequently positioned at the interface between lexis and grammar, i.e. certain words but not others of the same word class prefer certain grammatical rules or patterns. The patterns as such are not new, nor are the words, but what is novel is the habitual association between them in specific varieties. (2007, 83)

[…] grammatical innovations, also in the process of structural nativization, typically start out where the regular meets the chaotic, i.e. at the intersection of grammar and lex- is. Therefore, many of the characteristic innovations of PCEs can be located at this boundary; they concern the co-occurrence potential of certain words with other words or specific structures. (86)

Researchers working on a particular feature in a particular variety then typically set out to prove two points: firstly, that the feature under scrutiny was not a learners' or performance error but a stable, recurring feature, and secondly, that it diverged from BrE as the historical input variety in form, meaning, and/or distribution. The arguments for both points typically – and crucially – rely on frequency: for a feature to be considered a natural part of speakers' repertoires, it has to be quite frequent across different speaker groups stratified by age, gender etc. as well as across different registers. Secondly, both higher or lower frequencies of a PCE feature as compared to BrE indicate structural nativization (cf. Schneider 2004 as an example).

Many studies explicitly incorporate Schneider's evolutionary perspective in comparative studies by pitting PCEs against each other that are supposed to be on different stages of the cycle. Again, the aim is to identify the linguistic forms that correlate systematically, in frequency and distribution, with a PCE's position, e.g. in stage 3 or 4. However, Schneider's overview of relevant research (2014, 14-15) displays a certain ambiguity when it comes to the exact nature of this correlation. Mukherjee and Gries, for example, found "a steady cline of dissimilarity to the historical input varie- ty" (2009, 47) in their study of verb complementation patterns in 3 PCEs, and are quoted with their explicit reference to divergence as the indicator for the evolution of a PCE: "Thus, the picture at large is that the more advanced a New English variety is in its evolution, the more dissimilar it is to British English at the level of collo- structions" (2009, 48). The same authors came to a rather different conclusion for their subsequent corpus-driven study of n-grams in the same 3 PCEs: even though "[b]y their very nature, n-grams are also located at the interface between lexis and syntax"
(Gries and Mukherjee 2010, 523), the patterns arising from the analysis failed to replicate the results from the earlier study. The authors thus note:

In a wider setting, our observations and conceptual considerations raise general questions about the appropriate level of descriptive granularity at which evolutionary stages of the development of New Englishes manifest themselves. It seems to us that the grouping of New Englishes into evolutionary stages is a categorization at a very high level of abstraction; against this background it is obvious that a neat alignment of evolutionary stages on the one hand and linguistic features, their frequencies and their distribution on the other can only be found at the level of rather abstract linguistic configurations based on a wide range of linguistic forms. (Gries and Mukherjee 2010, 542)

The authors highlight an important point: they emphasize that large-scale generalizations about a variety's position on Schneider's cline should ideally be based on "a wide range of linguistic forms" rather than on only one feature. This point naturally leads to the question which features should be investigated – a basic question about the salience of individual features and their potential for becoming linguistic variables which will be discussed in more detail in section 3.

Further, Schneider's model crucially combines a linguistic and a social perspective, as captured in Table 1. Corpuslinguistic studies of PCEs mostly tackle frequency (and sometimes historical depth if corpora of earlier stages of a specific variety are available), for the reasons outlined above; investigating the other factors would require a quite different approach as well as a lot of fieldwork. To my knowledge, there has been no general discussion in the research community how to conceptualize this multidimensional model: it is unclear how the factors listed in Table 1 should be ranked or weighed against each other, whether they can be quantified at all or be studied in isolation from each other.

<table>
<thead>
<tr>
<th>Linguistic factors</th>
<th>Social factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency</td>
<td>demography</td>
</tr>
<tr>
<td>markedness</td>
<td>historical depth</td>
</tr>
<tr>
<td>salience</td>
<td>status of speakers</td>
</tr>
<tr>
<td>transparency and regularity</td>
<td>identity-marking of linguistic forms</td>
</tr>
<tr>
<td>similarity or difference between L1 and L2</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Relevant factors for diffusion and selection of an innovative form in the process of structural nativization in PCEs (summarized from Schneider (2007, 110-111))

In a sense, the focus on frequency is an outcome not only of the research paradigm informed by Schneider's model, but also of the advances in corpus linguistics in general and of those in the field of PCEs in particular. The next section considers in how far more, or 'big' data, contributes to our understanding of PCEs.

3. Big Data and Bad Data

The internet and technologies related to it have had a profound impact on the way we retrieve, use and store information. The web provides access to ever more (re)sources from all over the world: millions of digitized books are made available by companies such as Google, state and university libraries around the world collaborate in digitizing their holdings (cf. https://archive.org), manuscripts can be scrutinized online as

---

2 Schneider points out that "no claim is made that this list is exhaustive or mutually exclusive (2007, 110)."
fully searchable high-resolution images, and many publishers of scholarly journals retro-digitize their back issues well back into the 19th century (e.g. Anglia). Anybody wanting to create a corpus of written registers now has a staggering amount of sources at their disposal just by opening a browser and entering a few search terms. Moreover, web-related technologies have also given rise to new registers, frequently subsumed under the label 'computer-mediated communication' (CMC) and studied in their own right (cf. e.g. Herring et al. 2013). Linguists have embraced the research possibilities facilitated by the web, creating new tools and methodologies along the way (cf. Hundt et al. 2007). The – literally – biggest achievement to date in terms of size is the Corpus of Global Web-Based English (GloWbE), comprising 1.9 billion words culled from websites from 20 different English-speaking countries and accessible via a search interface (http://corpus.byu.edu/glowbe/). The advantages and disadvantages of this particular resource have already been discussed in the peer commentaries related to Davies and Fuchs (2015); in the context of this special issue, it is worth pointing out that a billion-word-corpus is far from solving the problem sketched by Kilgarriff and mentioned in the introduction: on the contrary and paradoxically, the sheer abundance of data may have the same effect as its scarcity for older stages of the language – that is, 'big data' turns into 'bad data':

The fundamental methodological fact that historical linguists have to face is that they have no control over their data. […] The great art of the historical linguist is to make the best of this bad data – 'bad' in the sense that it may be fragmentary, corrupted, or many times removed from the actual productions of native speakers. (Labov 1972, 100)

A sample search in GloWbE for the well-known Singaporean discourse marker *lah* may serve to illustrate the bad data problem which necessarily arises with a corpus consisting of very large decontextualized data sets that were extracted from the web: the search for *lah* turned up 806 hits for Singapore and 1520 for Malaysia, but also and surprisingly 31 tokens for India, where the form so far has not been attested. On closer inspection, these hits represent misspellings, e.g.

(1) Over 11 *lah* students have written this prestigious entrance exam.1

Quite a number of further hits constitute a misrepresentation of *Allah*; since only the main source website is still available, but not the individual text which contributed this example, we have no way of finding out how *Allah* became *Al??lah*:

(2) This is most generous gift of *Al??lah*.

Given this creative way of spelling *Allah*, the even higher number of *lah*-tokens listed for Pakistan may not come as a surprise. A closer examination of tokens from blogs provides another, not altogether positive surprise: 6 examples are contributed by a pornographic blog detailing sex with "Jam?*lah* Aurangzeb's daughter." The few remaining *lah*-tokens that appear to be genuine cannot be traced back with any confidence to actual Indian speakers/writers: "in GloWbE we only know that a website is

---

1 The context makes it clear that 'lah' should have been the numeral 'lakh,' i.e. 100,000.
from a particular country, but there might be speakers from other countries who have posted to that website" (Davies and Fuchs 2015, 26).

Creating such a huge corpus, then, involves relinquishing control over the data. Two important consequences of this lack of control have already been pointed out by Mair (2015). Firstly, automatic tagging is highly likely to produce quite misleading results where informal and/or nonstandard language is concerned, as the search results for lah-tokens have shown; and "[s]uch problems with the POS tagging are of an order of magnitude far greater than the usual errors found when CLAWS is used to tag more mainstream data" (30). Secondly, since "genre and style are the biggest source of unwanted noise in corpus-based synchronic comparisons of regional varieties of English" (30), GloWbE's rough distinction between 40% of 'general' web-derived content and 60% blogs may turn out to be too vague to actually be considered a genre distinction in any way that can be made operational. The data are further 'corrupted' in the sense that manual checking for spelling mistakes or typographical misrepresentations as witnessed in examples (1) and (2) above is not feasible for such a large project.

Labov's last point about data that are "many times removed from the actual productions of native speakers" is trivially true for any corpus of written language, and GloWbE is no exception, as Davies and Fuchs readily concede (2015, 26), even though the inclusion of 60% of material from blogs ("fairly informal language" (2015, 26)) is supposed to mirror the 60% of spoken language represented in the ICE corpora. The sample search for lah may be used to illustrate how far removed spoken language remains from 'informal language,' however defined: lah occurs 1,605 times in the direct conversation files of ICE-Singapore (216,913 words), that is with a frequency of 7,399 per million words (cf. Lange 2012a, 242). The Singaporean sub-corpus in GloWbE comprises 42,974,704 words, which yields a frequency of 0.2 tokens per million words – a ubiquitous phenomenon of face-to-face communication more or less vanishes in the vast Singaporean subcorpus including 60% of purportedly 'informal' language. 4

That obscure object of desire for socio- and other linguists, namely the vernacular, is typically only available in corpora of the contrasting 'Small is beautiful'-paradigm, probably best exemplified by the ICE-project. 5 With just one million words each, the individual ICE-corpora are indeed quite small, but their design has set a benchmark with respect to the notions of 'balancedness,' representativeness, and comparability in corpus design as discussed by Leech (2007) – even the 'big' GloWbE pays tribute to the rough genre distinction of 60% spoken and 40% written material. The successive generations of ICE-family members have also demonstrated an increasing awareness for contextualization of their data: while some ICE-projects of the first generation unfortunately do not include any metadata on speakers' backgrounds (e.g. ICE-Singapore), all the more recent ICE-projects provide such information. The researcher may thus go beyond the quantification of differences between national varieties and investigate further correlations between dependent and independent variables, eventually providing a more comprehensive account of language use in a particular speech community – but I will return to this point in section 3 below.

---

4 The figure of 806 lah-tokens for Singapore were taken at face value, i.e. not examined in context as to whether they represented genuine instances.

5 To my knowledge, the largest corpus of spoken language is the spoken part of the British National Corpus (BNC), comprising around 10 Mio. words and 10% of the overall corpus.
Even the small but balanced corpora of the ICE-family struggle with the vernacular: direct face-to-face conversation as the most informal and most immediate register of spoken language is notoriously difficult to compile and further extremely time-consuming to transcribe – the fact that the release of ICE-Sri Lanka, ICE-USA and ICE-Nigeria does not include any spoken data yet testifies to the difficulties inherent in creating corpora of spoken language. However, a focus on spoken language and in particular informal conversation is supported by several related theoretical considerations. These methodological assumptions are derived from linguistic subdisciplines which overlap in the study of PCEs. Firstly, the perspective from language variation and change informed by the sociolinguistic paradigm crucially relies on the vernacular as the locus of innovation, and on patterns of interaction between members of speech communities for the spread of these innovations. In societies with standard languages, there is a great gap between the relative stasis of the standard language and the flexibility and dynamism of the spoken language. Mair in his sample study of recent change in specificational cleft constructions (2013) has made a point for including spoken language in order to capture the directionality of change:

the inclusion of transcribed dialogic speech proved essential in two ways: it drew attention to variants largely confined to the spoken medium, and it helped to interpret the development in the written data as a genuine grammatical change rather than a change in formality expectations in particular written genres (2013, 183)

Secondly, the perspective from language contact that I would like to advocate here equally gives centre stage to face-to-face conversation as the locus where multilingual speakers will exploit the full range of their repertoires for successful communication. Language contact happens first and foremost in multilingual encounters; when speakers cease to police the clear demarcation of languages at their disposal and freely make use of the entire range of their communicative repertoire we may be afforded a glimpse of the actuation of contact-induced language change (cf. Matras and Sakel 2007). In general, however, we will only be able to trace the propagation of such changes, either because the linguistic feature in question is already well established or because the corpus design as such discourages too much of multilingual exchange. The ICE-project is, after all, a corpus of national varieties of English; "speak in English" is a comment or command one contributing Indian speaker received for codeswitching even though she and the other contributing speaker share their multilingual background:

(3) B: You've become more wise after going to Sangli Walchand College of Engineering
<,> uhn
A: (Marathi:) Tumcha awaz Tumcha awaz mala lagech odhako
B: Uhn no no <,> speak in English <,> speak in English
A: Okay (ICE-IND:S1A-096#232-337)

Even with access to genuinely multilingual face-to-face conversations, the question of how to pin down contact as the main motivation for change remains a thorny one; Thomason (2010) provides an overview of the issues involved. Let us circum-

---

6 'Immediate' in the sense of Koch and Oesterreicher's notion of 'language of immediacy,' cf. Koch and Oesterreicher (2012).
7 ICE-India S1a-96; speaker A: female student, speaker B: retired professor; both have Kannada as their mother tongue and Hindi and Marathi as other languages.
vent this problem for the time being and just assume for the sake of argument that we have found an innovative feature in a variety under scrutiny. If we are interested in the trajectory of change with regard to this specific feature, we again have to consider both spoken and written language. Two examples may serve to illustrate this point. The Singaporean discourse marker *lah*, introduced above, has already entered the *OED* and has become so iconic as a marker of colloquial Singapore English (Singlish) that it has found its way onto T-shirts, fridge magnets and other Singaporean souvenirs. Its high frequency in the spoken language, however, is not matched by its frequency of occurrence in writing: *lah* occurs a mere 13.9 times per million words in the written part of ICE-Singapore as against 7,399 times per million words in the direct conversations sampled by the corpus. This discrepancy is to be expected since discourse markers are by definition a feature of spoken language and more specifically conversational interaction. Singlish may well become a marker of a local Singaporean identity, providing the clue for many researchers to claim that it has moved on to stage 4 of Schneider's Dynamic Model, the stage of endonormative stabilization. Most of these claims are based on the forms and functions of Singlish, many of which are unlikely to cross the threshold to the written language and thus also unlikely to be codified, another crucial factor for assigning a PCE to stage 4.

My next example probes deeper into the tension between the spoken and the written language when it comes to the propagation of innovative features in PCEs. In Indian English (IndE), both *only* and *itself* are used as presentational focus markers, as in:

(4) Somehow like *no* it's very hot
   Like in the afternoon *only* it is burning
   The skin is burning like <,> whereas in Goa we have the cool climate (ICE-IND: S1A-001#18:1:A)

(5) I think you should start going to the gym from now *itself* (ICE-IND: S1A-061#86:1)

Both forms are used in more or less the same linguistic contexts (cf. Lange 2012b, 182-195) and thus qualify as linguistic variables in the classical sociolinguistic sense: "The definition of a linguistic variable is the first and also the last step in the analysis of variation. It begins with the simple act of noticing a variation – that there are two alternative ways of saying the same thing" (Labov 2004, 6).

While presentational *only* is quite common in the spoken language, it is also stigmatized within the Indian English speech community and thus absent from writing. *Itself* is much rarer overall, but has been established in the written language and might become part of an emerging Indian English endonormative standard (Lange 2012b, 190, Bernaisch and Lange 2012). Without access to carefully balanced corpora including spoken language, neither Labov's first nor the last step in an account of variation would have been feasible.

We might draw an interim conclusion at this point by considering what the two preceding examples have in common: the analysis of their distribution across the subgenres of prototypically 'small,' carefully balanced corpora exposed quite different degrees of entrenchment in everyday spoken interaction vs. more formal/written genres. This difference – a purely linguistic factor – then served as the point of departure for forming hypotheses about the acceptance and possible codification of the innovative forms – purely social factors in Schneider's classification as depicted in table 1 above. Further corroboration of these hypotheses would have to come from attitude
studies within the respective speech communities rather than from corpus studies, and certainly not from more or 'bigger' data.

4. Conceptual Traps

So far, I have discussed the issue of decontextualization exclusively with respect to corpus compilation: I have argued for the 'Small is beautiful'-approach which operates in the creation of carefully balanced corpora including spoken language and documenting speakers' metadata. The 'Mine is bigger than yours'-approach certainly has its merits, and it is left to the individual scholar to work with the kind of corpus that best matches his/her research interest. However, decontextualization may also occur at other stages of the research process, e.g. when formulating the actual research question, or at the subsequent level of interpreting the results of corpus-based analyses. This will be discussed in the following.

First of all, corpora predispose us to pursue certain questions/issues more readily than others, in two important respects. On the one hand, we tend to study and compare those varieties which are already represented in corpora, and on the other hand, we tend to study what can be found in corpora, that is, the structure and markup of corpora influences our research questions. To illustrate: we know a lot about the lexicogrammar of IndE thanks to the availability of ICE-India and the earlier Kolhapur Corpus, but we know next to nothing about English in Pakistan and even less about English in Bangladesh, the two neighbouring South Asian countries which share their colonial past with India and therefore also the historical British English input variety. Sri Lankan English has secured a firm position on the linguistic research agenda by the release of ICE-Sri Lanka (written) and by its inclusion in the SAVE (South Asian Varieties of English)-Corpus. Both Pakistan and Bangladesh are also represented in the SAVE-Corpus, but due to our general lack of knowledge about these speech communities, the danger of decontextualizing findings looms particularly high – a point that will be taken up further below.

We also tend to study what can be found in corpora, or rather what can be retrieved easily. 'Easy' should be understood here as a relative notion: corpus-based research projects may range from straightforward lexical searches which require very little manual checking of the data (e.g. the recent study by Parviainen (2016) on invariant isn't it in Asian Englishes) to search strings which incorporate the search for absences, for example Huber's research on relative clauses in Ghanaian English (2014); in this case, the "two or more ways of saying the same thing" include the overt relativizers who and that as well as the zero relativizer. In these and many other cases the variable and its linguistic instantiations are known in advance, can be retrieved from the corpus and then subjected to further analysis. Such studies have contributed enormously to our understanding of the range of variation within a single variety or across varieties, and will continue to do so with the steady growth of the ICE-family. Nevertheless, we should be aware of what we might overlook in our data if we rely too much on automatic retrieval and predefined categories. A case for a more comprehensive approach has been made by Cheshire (2005), who investigated the marking of discourse-new entities in conversations with reference to gender and social class. Her database consisted of interviews from 96 adolescent speakers from three different English towns, the independent variables collected with the interviews

---

8 The SAVE-Corpus was compiled at Justus-Liebig-University Gießen.
were gender, social class and location. Cheshire first focused on existential constructions, which are typically used to introduce discourse-new entities; when her data did not display any correlation between the linguistic and the independent variables, she went back to the interviews for a more holistic picture, including all other possibilities of indicating that the conversation turns towards something not previously mentioned. Apart from syntactic constructions such as the existential, she found a wide range of forms, such as tags and discourse markers which create speaker involvement (Cheshire 2005, 486), a specific articulation (488), and others. That is, the forms enlisted by speakers to express a specific function in discourse range over more than one level of linguistic structure, and this range only becomes apparent through a careful qualitative analysis of the data:

As we have seen, the adolescents in the study reported here harnessed a wide range of forms to mark discourse-new referents, taking them from all components of their language. The forms were of many different kinds, including constructions normally considered to be generated by the grammar, forms stemming from the demands of online performance, and pragmatic particles with functions in the interpersonal dimension of discourse. In other words, the speakers happily performed the same discourse function using phenomena that are normally analysed separately from each other. Perhaps we may learn more about the way speakers make use of syntactic variation in the construction of discourse if as linguists we follow their lead and in our analyses consider what might be gained by abandoning a rigid separation between the different components of language. Such an approach would not fit well, if at all, within a conventional variationist framework, and the analysis would no longer focus purely on syntactic variation. [...] although it may be revealing to pluck syntactic forms from their discourse context in order to enter them into a multivariate analysis, in this case the only aspects of the discourse context that will be taken into account will be those that we decide ahead of time to be potential constraints on the variation. Other aspects of the discourse contexts will remain unexplored. (Cheshire 2005, 502-3)

Such a broader approach to the linguistic variable as envisaged by Cheshire requires a thorough qualitative and context-sensitive investigation of the data to discover the relevant forms that map onto a specific function. This perspective would also prevent us from decontextualization at a later stage of our analysis, when we might be tempted to overgeneralise and to abstract away from the speech community. Brezina and Meyerhoff (2014) is an impressive demonstration of the pitfalls of the standard method of deriving sociolinguistic conclusions from corpus data:

Such analyses often reproduce stereotypes about language and society rather than contribute to our understanding of genuine sociolinguistic variation since, as we have shown, similar results can be produced with socially meaningless groupings. (2014, 24)

One striking aspect of their study is the distribution of the "dummy variables" the and some which were chosen "to test the distribution of forms that are not hypothesized to have any social indexicality" (2014, 9). These dummies also turned out to pattern in statistically significant ways with the social variables gender and/or age. This result casts serious doubt upon the enterprise of corpus-based variation studies, highlighting "the limits of useful generalization" (2). In order to arrive at valid observations, the authors recommend a mix of quantitative and qualitative methods as well as a generally more context-sensitive approach, both to the speech community under scrutiny and to variation between individual speakers: "A corpus (no matter how large) should never be regarded as a magic box which is able to provide fast and
ready-made answers to any sociolinguistic questions we chose to ask" (23). Ignoring this caveat is likely to produce misleading results: "The differences reported in the studies based on this methodology are thus more likely to be a function of the method used rather than an indication of sociolinguistically meaningful variation" (18).

The question remains how "sociolinguistically meaningful variation" can be captured in corpus-based research on PCEs. Croft's (2006) functional-typological model of language change distinguishes three levels of variation in language which roughly correspond to Labov's distinction between indicators, markers and stereotypes. First-order variation is "the variation that occurs in individual occasions of language use […] Second-order variation is built on first-order variation: certain first-order variant productions take on sociolinguistic value in the speech community" (Croft 2006, 71-72). Finally, third-order variation is defined as "variation across languages and across varieties of language. Third-order variation is second-order variation that has become fixed conventions in different speech communities as these have diverged and/or entered into contact over the history of humankind" (2006, 71-72). Croft explicitly emphasizes the overarching importance of social factors for the propagation of change:

Social factors – the structure of the speech community – drive the change forward, and do vary over time (leading to propagation or even the halting of a change in progress). [...] The socially motivated propagation process begins only after a novel variant acquires a social value (that is, it is identified with a social group). This is the transition from first order variation to second-order variation. (2006, 81)

What is depicted in corpus-based studies of PCEs then is third-order variation in the sense of "variation across languages and across varieties of language" – the question that needs to be raised at this point is whether it is also variation that emerged from second-order variation, that is from variation that is indexical of local social meanings and thus amenable to propagation as a socially motivated process of language change. This question leads to the more fundamental question whether there can be third order variation without second order variation, or to put it differently: do the variety-specific differences in the distribution of, e.g., the dative alternation (cf. the recent study by Szmrecsanyi et al. 2016) really tell us something about salient choices within and across speech communities, or do they just reflect first-order variation on a grander scale? Unless we (re-)contextualize our data and take the actual speech community into account, we will have no way of knowing for sure.

5. Conclusion

It is not without irony that the continuous advances in corpus linguistics – ever 'bigger' data and more refined statistical tools – have also contributed to the emergence of a new kind of armchair linguist, namely the armchair corpus linguist. He or she can now write a corpus-based monograph about a variety of English without ever having talked to a member of the targeted speech community, let alone visited the country where the variety is spoken. In this paper, I have sided with Cheshire (2005; 2007), Meyer (2012), Mair (2013) and many others not explicitly mentioned here to argue for a combination of quantitative and qualitative approaches to the study of variation. I am fully aware that (a) a lot of my own research has stumbled head-on and with gusto into the conceptual traps outlined in section 3, and (b) that I have raised more
questions than I could answer. Nevertheless, I am convinced that taking the time to contemplate these questions is time well spent.

Works Cited


