Theoretical models and statistical modeling of linguistic epicentres

Tobias Bernaisch[1] | Stefan Th. Gries[2] | Benedikt Heller[3]

[1] Department of English, Justus Liebig Universität, Giessen, Germany
[2] Department of Linguistics, UC Santa Barbara, USA & Department of English, Justus Liebig Universität, Giessen, Germany
[3] Department of English, Justus Liebig Universität, Giessen, Germany

Correspondence
Tobias Bernaisch, Department of English, Justus Liebig Universität, Giessen, Germany.
Email: Tobias.J.Bernaisch@anglistik.uni-giessen.de

Abstract

Motivated by a fundamental discussion of the relation between theoretical and statistical modeling, the present paper takes stock of the research history of linguistic epicentres in the world Englishes paradigm and seeks to provide suggestions for future epicentral studies. Based on a review of earlier research into potential epicentral constellations, we provide an overview of the genesis of the concept of linguistic epicentres, describe different methodological approaches that have been chosen and/or recommended for their study and provide statistical comments. With a view to future epicentral research, we aim at a) refining the theoretical construct of linguistic epicentres, b) suggesting empirical methods that allow identifying epicentral constellations more reliably than in the past and c) making statistical recommendations particularly relevant to epicentral research.

1 | INTRODUCTION

In this article, we seek to identify ways to move forward the theory of linguistic epicentres and the methodology of studying them in the world Englishes paradigm. In section 2, we zoom in on central theoretical and methodological pillars in epicentral research. More specifically, in section 2.1, we discuss theoretical considerations involving the role of theory and empiricism and the role of different kinds of frequencies in epicentral model/theory development. In section 2.2, we address the role of sociolinguistic/cultural effects in epicentral research, and in section 2.3, we offer statistical considerations in the light of representative studies of linguistic epicentres. Section 3 offers a short conclusion.

2 | TAKING STOCK

Now that the field of world Englishes in general as well as epicentre research have received decades of academic attention, we feel it is time to take stock of the epicentre model, of associated assumptions as well as of the methodological tools used to study potential epicentral constellations. To us, this is timely because there seems to be room for further improvement, selected aspects of which we will discuss in the next three sections.

2.1 | Theory, empiricism, and different kinds of frequencies

2.1.1 | Theory and empiricism

Our first concern is abstract/fundamental with important implications for what it means to work on varieties/epicentres with theoretical and statistical models. We are reacting to Hundt (2020, p. 1)'s perception of a 'gap between theoretical and statistical modelling' and reference to Gries, Bernaisch and Heller (2018), where we identify a need on the part of the sociolinguists to provide more testable theoretical models: '"it would certainly be useful if such models were formulated with a degree of precision that makes it (more) straightforward to arrive at falsifiable operationalizations to test their claims, not to mention predictions"' (Hundt, 2020, p. 2).[1] However, Hundt seems to disagree and asks 'whether this is actually the purpose of theoretical modelling. Schneider (2004, p. 233), in an earlier attempt at corpus-based verification of his dynamic model, makes clear that "it is the [empirical, MH] researcher's duty to bring forward hypotheses, to ask the right questions". Rather than requiring theoretical models to provide empiricists with operationalisations for testing, one could also ask what the statistical models, in turn, have to offer to the theorists in order to allow them to advance their models' (Hundt, 2020, p. 2).

The implied scenario seems maximally convenient for the theoretical modeler, who does not have to worry about whether a suggested theoretical model is operationalizable and/or specific enough to even be testable or falsifiable because developing such hypotheses is the task of the empirical modeler. More polemically, this is developing a picture of science in which the theoretical modeler can dream up whatever lofty theory they want because they are not required to make their theories operationalizable. Very conveniently, if the empirical modeler then tries to derive testable hypotheses from a theory that has not been formulated in a testable way and

- *does not* find results supporting the theory, the theoretical modeler has the easy recourse of dismissing these empirical falsifications by arguing that the hypotheses tested were invalid operationalizations of the theory anyway;
- *does* find results supporting the theory, the theoretical modeler can promote these as empirical validation of the theory.

It seems unclear what definitions of model and theory Hundt (2020) implies here, but it cannot be the notions of model and theory that form the basis of social/behavioral empirical sciences. In these, both a) the notion of theory implied by the above views and b) an understanding of being a theoretical modeler as an invitation to be unconstrained regarding empirical validation in devising a theory are untenable. For example, Manning (2003, p. 296) points out that any explanatory hypothesis that is 'disconnected from verifiable linguistic data' ought to give rise to some concern or VanPatten, Williams, Keating and Wulff (2020, p. 2) state that the first duty of a theory is to account for or explain observed phenomena. But a theory ought to do more than that. A theory also ought to make predictions about what would occur under specific conditions. Thus, a theory's plausibility/utility is a function of how well it can account for empirical data or make predictions on to-be-collected empirical data – and since a

statistical 'model is a formal representation of a theory' (Adèr, 2008, p. 280, quoting Bollen, 1989, p. 72), it is with statistical modeling (or tools) that we test theories. Thus, the answer to Hundt's (2020, p. 2) – though rhetorical – question of 'what the statistical models, in turn, have to offer to the theorists in order to allow them to advance their models' is simple: quality control. Hundt (2020, p. 3) then exemplifies her argument regarding the separation of theoretical and empirical models with Schneider's (2007) dynamic model: 'One of the reasons why Gries et al. (2018) find it so difficult to derive predictions that can be operationalised for a quantitative approach from Schneider's (2007) dynamic model is that it aims to capture the complex historical, social and cognitive processes involved in the evolution of Postcolonial Englishes (PCEs)'. Similarly, she (2020, p. 9) states '[i]t is obvious that a socio-historically complex model that aims to incorporate aspects of psycholinguistic processes cannot be statistically modelled in its entirety'. Although Hundt (2020) apparently profiles a seemingly unavoidable incompatibility of theoretical and empirical models rooted in the undisputed complexity of Schneider's (2007) dynamic model, Schneider himself contests this incompatibility when he argues that

- 'a monodirectional causal relationship' (2007, pp. 30-31) lets historical, sociocultural and sociolinguistic characteristics of a PCE culminate in observable linguistic/structural effects;
- the 'most promising road to a possible detection of early traces of distinctive features is a principled comparison of performance data collected along similar lines, that is, systematically elicited corpora' (2004, p. 227);
- structural nativization of phrasal verbs in world Englishes can be explored with the following research questions: 'Incidence and frequency of use: Are PVs in general, or certain PVs in particular, preferred in certain WEs? The question may be asked with respect to quality (the range of distinct forms, that is types, found in a given variety) and quantity (the token frequencies of occurrence). [and other empirical research hypotheses follow]' (2004, p. 233).

Thus, Schneider does not seem to be arguing for the kind of separation of theoretical and empirical modeling that Hundt attributes to him – on the contrary. Still, the issue at hand goes way beyond the concrete question of whether Gries, Bernaisch and Heller (2018) derived and tested predictions that Schneider's (2007) model did or did not imply. Our point – just like Hundt's (2020) – relates to a fundamental understanding of what it means to work scientifically. We submit that if a model/theory has not been derived from empirical evidence or does not generate predictions that are empirically falsifiable at any level – micro, macro or somewhere in-between – then it is either not a model/theory or it is a model/theory of dubitable utility for the scientific process of accumulating knowledge. It is in this spirit that we also approach the model of linguistic epicentres.

2.1.2 | Relative frequencies of features and diachrony

The premise seems to have by default been that an epicentre displays higher frequencies of a phenomenon disseminated via epicentral mechanisms than the varieties under epicentral influence (Bernaisch & Lange, 2012; Parviainen & Fuchs, 2018). More technically, the expectation seems to be that

- a certain linguistic feature *F* (a morpheme, word, construction, …) is attested in an assumed epicentral variety *V* (for example Indian English (IndE)) at point of time t_0;
- if *V* is indeed an epicentral variety, then the presence of *F* in it will exert some pressure/temptation on other (maybe geographically neighboring) varieties $W_{1\text{-}n}$ to adopt *F* at a point in time t_x (later than t_0);
- thus, one way to document epicentre status of a candidate variety *V* is
  - to determine the frequency of *F* in *V*;
  - to demonstrate that *F* was not attested in $W_{1(\text{-}n)}$ at t_0 but
    - is attested at a later point in time t_x, for example t_3;
    - increases in frequency in $W_{1(\text{-}n)}$, maybe to the level of frequency of *F* in *V*.

This is represented in Figure 1: Time is on the *x*-axis (in arbitrary units), the frequency of *F* is on the *y*-axis, the blue and red lines represent the frequencies of *F* in two varieties, with *F* getting adopted into *W* at around t_3.

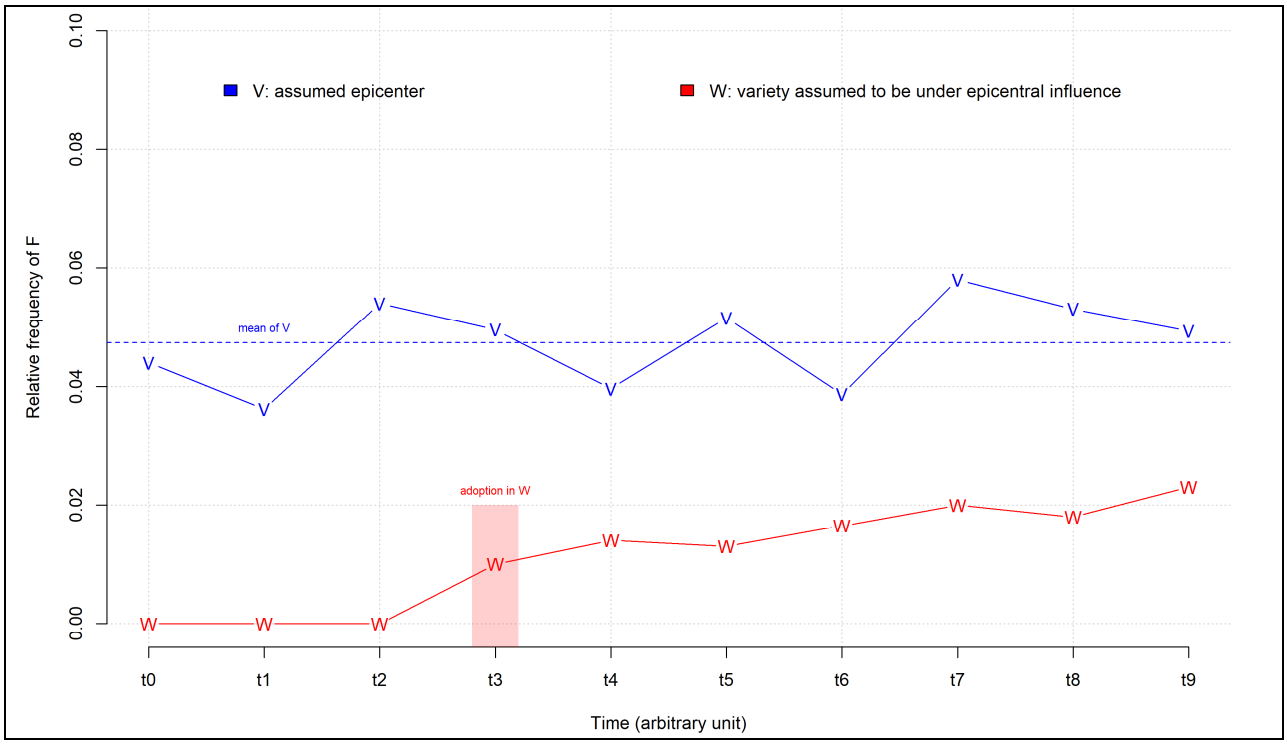

FIGURE 1 Traditionally assumed frequency distribution in epicentral constellations across time.

The assumption seems to be that when a researcher looks at a certain point of time t_x (with *x* being later than the point of time of adoption), then the frequency of *F* in variety *W* will be below the frequency of *F* in the epicentral variety *V*. This assumption becomes evident with Parviainen and Fuchs (2018, p. 11): '[T]he lower frequency of this feature [certain clause-final particles] in HKE [= Hong Kong English] and PhiE [= Philippine English], compared to IndE, is commensurate with the hypothesis of spread from IndE to HKE and possibly PhiE'. Still, other diachronic developments are conceivable in epicentral constellations. There does not seem to be a basis for the assumption that the spread of *F* from *V* to *W* is reflected in a lower frequency

of *F* in *W* at t_x. Given the current absence of truly diachronic data, researchers can usually check at one or two points in time only, which means the relation of the frequency of *F* in *V* to the frequency of *F* in *W* could be completely dependent on when one determines them, as in Figure 2:

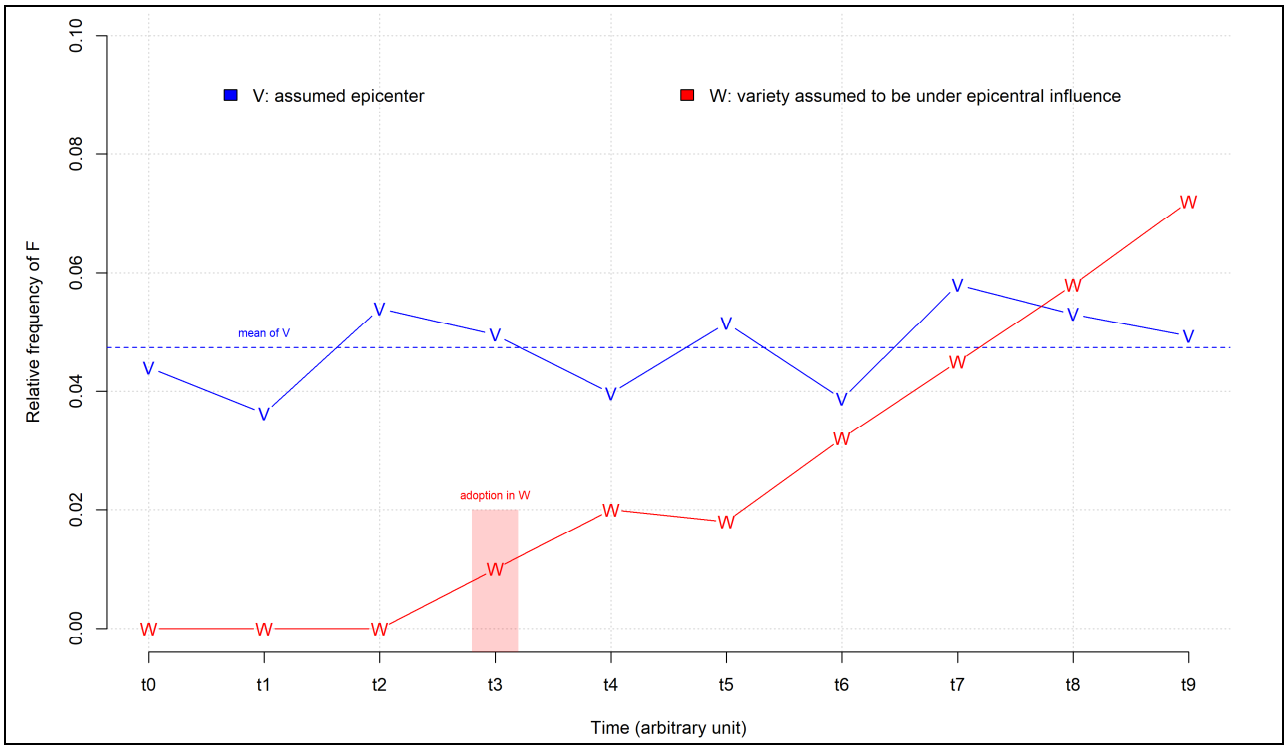

FIGURE 2 Alternative frequency distribution in epicentral constellations across time.

Researcher A, who compared the frequencies of *F* in *V* and *W* at t_0 and t_5 would find indeed that

- at t_0, *F* is attested in *V*, but not in *W*;
- at t_5, *F* is attested in *V* and in *W*, but more frequently in *V*.

But researcher B, who compared the frequencies of *F* in *V* and *W* at t_0 and t_9 would find that

- at t_0, *F* is attested in *V*, but not in *W*;
- at t_9, *F* is attested in *V* and in *W*, but more frequently in *W*.

It is difficult to find arguments why *F* in *W* cannot overtake the frequency of *F* in *V*. As soon as *F* makes its first appearance in *W* (at around t_3), (socio-)linguistic forces might make *F* become more frequent in *W* than in *V*. A case in point are pseudotitles in Bahamian English, which Bahamian English adopted from American English (AmE), but have become more frequent in contemporary Bahamian English than in AmE newspapers (Hackert, 2015). Similarly, light-verb constructions occur more often in IndE than in British English (BrE), although BrE served as the input variety for IndE (Hoffmann, Hundt and Mukherjee, 2011).

While it is indisputable that epicentral influence has a diachronic dimension to it, diachronic corpora for world Englishes are generally not yet available. Unless one has fine-grained diachronic data that allow properly tracking the development of *F* in *V* and *W*, the ratios of the frequencies of *F* in the two varieties might have to be treated more cautiously than assumed. Note, however, that this call for caution also extends to statistically more sophisticated epicentral studies – including our own research – that employ apparent-time comparisons of synchronic corpora (see Gries, Bernaisch and Heller (2018) for how even complex statistical methods are unreliable when the assumptions of the apparent-time method are not compatible with the data analyzed).

Hopefully, the completion of ongoing corpus compilation projects (Bernaisch & Heller, 2020 for South Asian Englishes; Biewer, Bernaisch, Heller, & Berger, 2014 for HKE; Collins, Borlongan, & Yao, 2014 for PhiE; Hoffmann, Sand, & Tan, 2012 for Singapore English (SingE); Kruger, van Rooy & Smith, 2019, for various Englishes) and already available short-term BrE and AmE data will ultimately allow us to get more reliable empirical insights into potential epicentral constellations – while controlling for other diachronic processes such as Americanization, colloquialization or democratization (Leech, Hundt, Mair, & Smith, 2009; Baker, 2017) as well as level-1 contextual predictors as discussed in section 2.3.

2.1.3 | From counting structural features to interpreting underlying norms

The term *epicentre* has its origins in seismic geology and 'means the "outbreaking point of earthquake shocks", that is, the point on surface of earth above the hypocenter, which is the subterranean source of the seismic disturbance' (Peters, 2009, p. 108). Although what is directly observable to the human eye is the epicentre with its horizontal regional spread of seismic waves across the earth's surface, the triggering and underlying powers behind the epicentre have a vertical dimension to them – unobservable subterranean seismic disruptions cause forces to rise up until they crack open the surface of the earth, from which seismic swells subsequently spread.

The linguistic equivalent of such a geological epicentre is contextualized in a dynamic and pluricentric conceptualization of world Englishes. Leitner (1992, p. 225) is the first scholar to employ the term *epicentre* in a linguistic sense to refer to first- and second-language 'norm-setting centres'. Although he did not provide a proper definition, it is evident that Leitner (1992, p. 202) attributes the following characteristics to linguistic epicentres: (i) a relatively high degree of standardization of a variety, also including the acceptance of localized features on the part of the respective variety speaker groups, and (ii) the potential of a variety to influence others. Compatibly, but with reference to Schneider's (2007) dynamic model, Hoffmann, Hundt and Mukherjee (2011, p. 259) – as well as Hundt (see 2013, p. 185) – suggest that 'the concept of epicentre includes two components, an internal and an external one. […] On the one hand, an epicentre is marked internally by endo-normative stabilisation, that is by the wide-spread use, general acceptance and codification of the local norms of English. […] On the other hand, an epicentre should also have the potential to serve as a model of English for neighbouring countries, that is exert an influence on other speech communities in the region'.

In this light, there are certain parallels between epicentres in seismic geology and in linguistics. Just like a hypocenter is only perceivable/observable indirectly by human beings via epicentres that crack open the surface of the earth, underlying linguistic norms constituted by

certain linguistic and extra-linguistic factor constellations are also only perceivable/observable indirectly by human beings via (statistically modeling) the linguistic surface structures speakers use in their daily interactions. The relation between underlying norms and their corresponding surface-structure choices and the resulting implications for epicentre research can be illustrated via the dative alternation as in Table 1.

TABLE 1 Underlying predictors and surface structures with the dative alternation.

Surface structure	He gave [her]$_{\text{RECIPIENT}}$ [a book]$_{\text{PATIENT}}$.	He gave [his daughter]$_{\text{RECIPIENT}}$ [the freedom to come home late]$_{\text{PATIENT}}$.
Underlying predictor constellation	RECIPIENT = *pronominal*	RECIPIENT ≤ 5 words + *non-pronominal* & PATIENT > 3 words + *abstract*

It is to be noted here that, although the underlying factor constellations are different, they trigger the same surface-structure choice – in this case the double-object construction. As it seems generally accepted in epicentral research (Leitner, 1992; Peters, 2009; Hoffmann, Hundt, & Mukherjee, 2011; Hundt, 2013) that linguistic epicentres act as models for other varieties in the region, the question is whether we should focus on surface-structure choices (as the majority of epicentral studies have so far) or their underlying factor constellations when attempting to trace this model character of epicentres empirically.

Not only in linguistics is a model 'a simplified or idealized description or conception of a particular system, situation, or process, often in mathematical terms, that is put forward as a basis for theoretical or empirical understanding, or for calculations, predictions' (OED online, under *model*). Transferring this model notion to world Englishes results in a conceptualization of regional varieties of English as abstract representations of the complex interplay between various underlying factor constellations, where factors, which can be speaker-related (age, ethnicity, gender, social class et cetera), contextual (formality, historical period, interlocutors et cetera) or structural/structure-specific, guide the surface-structure choices speakers make in a given communicative context and region. These variety-specific and highly abstract representations encapsulating factor constellations for surface-structure choices – or, in other words, models – are activated, actualized, and updated whenever variety speakers communicate and produce linguistic structures. It is this set of spoken and written structures that constitutes a linguistic epicentre as a physically perceivable entity, and this set of structures can spread to other regions – the horizontal plane in parallel to a geological epicentre – in that material from the epicentral variety permeates to other varieties in the region via speech/writing and can be consumed by speakers/writers of the other varieties in the region. If the consumption of these permeated structures by speakers/writers living in a region adjacent to that of the epicentre leads to a re-configuration of their variety-specific model, this re-configuration instantiates epicentral influence since the epicentral variety model would have then updated that of the variety under epicentral influence, which becomes evident in structural choices different from those that the variety model under epicentral influence would have produced without epicentral influence. One of the repercussions of this conceptualization of epicentres and epicentral influence is that – at least from our point of view – epicentral research should focus on how well a given variety-specific model assumed to be an epicentre can also account for the structural choices made in varieties assumed to be under

epicentral influence. Surface-structure choices as evident from linguistic corpora should thus be considered pathways to profile the underlying abstract factor constellations constituting variety-specific models, aligning empirical research with the regional model character figuring prominently in theoretical definitions of epicentres. In contrast, epicentral research restricting itself to establishing frequencies of certain surface-structure choices without considering underlying factor constellations has little to offer regarding the model character of linguistic epicentres. This is the case because – when we accept that a variety-specific model is constituted by the sum and interplay of underlying factor constellations for surface-structure choices – an exclusive focus on surface-structure choices might actually mask differences with underlying factor constellations as shown in the dative-alternation example above, where the same surface-structure choice, the double-object construction, is made despite vast differences in the underlying factor constellation for this syntactic choice.

Against this background, epicentres pose methodological challenges and require specific sets of data. Relying on Meyerhoff and Niedzielski's (2003, p. 544) recommendations, Hundt (2013, p. 191) argues that we need to a) establish the structural and functional equivalence of the variant found in two varieties, b) distinguish external influence of the (potential) epicentre by another variety from independent regional (parallel) developments, and c) assess the evaluation of the variable in the variety that is adopting it, that is whether speakers use the feature consciously or unconsciously. In addition, any verification of the epicentre status of a variety has to take the social, economic and cultural context into account and evaluate whether it fosters or hinders epicentric influence of one variety on another.

While a call for common standards in epicentre research is generally welcome, it is worth discussing whether these standards are uniformly applicable and relevant to all objects of investigation related to epicentral studies. Early book-length variety descriptions (for example Bolton, 2002 for HKE; Lim, 2004 for SingE; Gunesekera, 2005 for Sri Lankan English) often sought – even in the absence of large corpora – to document what could in some cases be referred to as linguistic butterflies – exotic surface-structure deviations from standard realizations of, for example, relative clauses as in (1), negation as in (2), or subjects as in (3).

(1) This is the student did it. (Gisborne, 2002, p. 144 on HKE)
(2) We not visit the place. (Fong, 2004, p. 92 on SingE)
(3) Raining no, how to come? (Gunesekera, 2005, p. 129 on Sri Lankan English)

Varietal/epicentral research of these butterflies does benefit from Hundt's (2013, p. 191) and Meyerhoff and Niedzielski's (2003, pp. 544-547) suggestions triggering questions such as:

- Are the structural realizations and the functional scope of the phenomenon studied the same in the varieties covered?
- Was the phenomenon transferred from one variety to the other or did it develop independently in both varieties?
- Are speakers of the varieties aware they are using said structures?

Yet, these butterflies – despite their undisputed status of variety markers – tend to be infrequent (in general and in comparison to their structural alternatives) and statistically

evasive and perform so little grammatical/functional work that one cannot help but wonder how much they can tell us about epicentral configurations. While these butterflies are of course attested in authentic data, they seem somehow reminiscent of the kind of examples on which generative grammarians used to build whole theories even though these examples were never attested and their acceptability was contested (Labov, 1975). Maybe in part due to this recognition, more recent investigations into world Englishes often focus on high(er)-frequency phenomena at the lexis-grammar interface (for example Heller, 2018; Roethlisberger, 2018; or Grafmiller & Szmrecsanyi, 2019 in ESL studies and Gries & Bernaisch, 2016; or Heller, Bernaisch & Gries, 2017 in epicentral research). Phenomena such as alternating structures of datives, genitives or particle placement have a much higher frequency and a much higher grammatical/functional workload, which is why we are referring to these as 'linguistic ants'. Crucially, these differ, in addition to their overall higher frequency, from linguistic butterflies in the following ways: First, from a Construction Grammar perspective, they exhibit a higher degree of schematicity in that they often have multiple (related) functions and allow many different elements to fill their slots. Second, from a varietal perspective, these phenomena/constructions have been core parts of the historical input varieties of world Englishes for a long time, meaning the corresponding surface structures are generally shared across varieties already. In this light, variety/world Englishes research on linguistic ants as just defined cannot use the same guidelines as those for linguistic butterflies because they would lead to partly misleading research questions (here based on the genitive alternation) like these:

- Are *s*- and *of*-genitives structurally and functionally equivalent across the varieties covered?
- Did the *s*- and the *of*-genitive emerge in parallel or did one variety transfer the variants to another?
- Do speakers use the *s*- and the *of*-genitive consciously?

Meyerhoff and Niedzielski (2003) appear to have developed their guiding questions with a view to linguistic butterflies but they apply much less to linguistic ants. Epicentral investigations of linguistic ants do not engage in the empirically extremely delicate endeavor of reliably tracing the spread of surface structures from one variety to another in a region, but seek to understand which variety in an epicentral constellation provides the abstract model best predicting the surface-structure choices in a given region. Conducting epicentral research on linguistic ants in this fashion makes the kind of sociolinguistic and sociocultural considerations, which inter alia Biewer (2015) and Hundt (2020) call for, secondary as discussed in section 2.2.

2.2 | On sociolinguistic/cultural aspects

'The best that we can thus hope for is converging evidence from quantitative modelling of usage data, on the one hand, and supplementary data from surveys, elicitation experiments, ethnographic interviews (not necessarily within the same study), that will hopefully allow us to eventually piece together how speakers negotiate the use of WE in context and what this means for the emergence of varieties of English' (Hundt, 2020, p. 15). We certainly agree with the relevance ascribed to quantitative modeling, and, of course, we are not principally opposed

to the consideration of the 'supplementary data' referred to – we are nevertheless skeptical about how much this kind of data can enrich in particular epicentral research. Even when attitudinal questionnaires and ethnographic interviews succeed in reducing distortions (due to, for example, social desirability and intuition-based self-reporting), the information obtained will document sociolinguistic parameters of the present when epicentral influence might have already manifested itself in structural convergence, and not when one variety starts exerting epicentral influences on others. Put differently, the difficulty/riskiness of inferring diachronic processes from synchronic corpus data also applies to such sociolinguistic information. Current or recent sociolinguistic surveys and experiments are simply too late to be directly relatable to the question 'whether speakers consciously aspire to a particular variety of English and thus adopt certain features from it' (Hundt 2013, p. 184). Although these sociolinguistic surveys of the present might potentially be useful in understanding epicentral mechanisms resulting in regional structural convergence in the future, they can inform currently observable structural convergence in epicentral configurations indirectly at best. Plus, what degree of granularity of information can one reasonably expect here?

As Peters (2009) and Peters, Smith and Bernaisch (2019) show, entries in historical dictionaries feature information on concrete lexical items that can be used to profile Australian English as a linguistic epicentre for New Zealand English (NZE), but (i) such historical dictionaries are unavailable for most world Englishes and (ii) offer little to non-lexical epicentral studies. Even if one had access to the rare historical document commenting directly on non-lexical structural features relevant to epicentres, it is not straightforward how this historical source is supposed to inform epicentral studies. This so far unaddressed methodological issue is – in our view – even more evident with the question how other types of sociohistorical information can be connected reasonably to observed structural convergence in epicentral constellations. How are migration patterns, media consumption or textbook dissemination across national boundaries in a given region quantitatively correlated with a local structural convergence of, for example, presentational *itself* in South Asian Englishes (Bernaisch & Lange, 2012)? While the call for sociolinguistic and sociohistorical information in epicentral studies seems plausible at first, building an empirically reliable bridge between these coarse-grained sociolinguistic/-historical parameters and cross-varietally converging usage patterns of structural features appears daring. Would a creative use of an English expression by an Indian actor famous for starring in movies featuring a reconciliation of India and Pakistan be predicted to not spread to Pakistani English (because of the current tensions between the countries) or be predicted to spread to Pakistani English (because viewers sympathize with the theme of the movies)? There does not seem to be a theoretically motivated way of predicting that because the theoretical scope of linguistic epicentres currently focusses on the level of national boundaries and (still) largely disregards the speaker level within these national boundaries. Nevertheless, multifactorial statistical modeling can operate at various levels and could – as regards the concrete Indian-actor example – in principle incorporate individuals' attitudes towards the two countries, the actor, the linguistic feature and other aspects to check whether these attitudinal factors significantly influence the use of the structural feature in the varieties concerned. In other words, statistical modeling could here – notably in the absence of meaningful predictions derived from a theoretical epicentre model – inform theoretical modeling with regard to the

impact of certain attitudinal aspects which seem unpredictable from a purely theoretical perspective.

That notwithstanding, we think that, currently, these relatively coarse-grained sociolinguistic/-historical aspects still have their say in epicentral theory in that they can contribute meaningfully what in cognitive linguistics has been referred to as *motivation*, which is a post-hoc way of explaining that does not reach the level of being truly predictive. To use an example from Panther (2012): The fact that the concept of *giving* has in many languages developed grammatical functions involving meanings of benefactive or causative (for example a preposition or a case like dative) is plausible since the concept of *giving* prototypically involves an agent causing the transfer of a patient to a recipient/benefactive. This does not mean that we can predict that every language that has a verb meaning give will have such a preposition or such a case, but it increases the probability of that and lends (sometimes a lot of) post-hoc compatibility/credibility to an analysis that invokes GIVE's semantics as an explanation.

2.3 | Statistical issues

With regard to epicentral studies with non-quantitative approaches or descriptive/monofactorial statistics (Leitner & Sieloff, 1998; Hoffmann, Hundt, & Mukherjee, 2011; Bernaisch & Lange, 2012; Parviainen, 2020), we think – often in line with what the authors themselves express in the respective studies – that it is accurate to consider them by definition unable to do justice to the multifactoriality of the studied phenomena. Strictly speaking, such studies explore the effect of factor *X* on some linguistic surface-structure choice *Y*, but do so without controlling for any of the other (linguistic) factors that are also involved in *Y* (and, in observational data, are often collinear with *X*). Even more problematic, the one factor *X* is often an extralinguistic factor – sometimes really only the variety in which instances were observed – which means such studies are generally lacking because of their acontextuality or, put differently, the fact that they are ultimately grounded in nothing but observed frequencies per corpus, condition, speaker. In other words (per Gries, 2018 for learner corpus data) and more technically in the language of mixed-effects models: such studies often concern themselves not with level-1 predictors (i.e. predictors at the level of the individual speaker choice in favor of or against *Y*) but only with level-2 or higher predictors (including extralinguistic one such as the corpus). For example, Bernaisch and Lange (2012) include no linguistic level-1 predictors in their analysis of presentational *itself*, although the choice by a speaker to use this focus marker instead of an alternative focus strategy is probably co-determined by a number of contextual features. Therefore, many such studies will be anticonservative and overestimate the importance of between-variety differences. In addition, virtually none of the studies that use frequencies or frequency differences of features *Y* to make a theoretical point consider the dispersion of *Y* in the corpus although it has been shown that underdispersion can invalidate any corpus statistic (Gries, 2008, 2021b).

In terms of statistically multifactorial studies in epicentre research, which are theoretically in a better position to deal with such data, we also seek to stress three major and widespread issues: First, information loss: This is evident in earlier epicentral studies such as Biewer (2015) or Parviainen and Fuchs (2018). Based on Varbrul analyses, Biewer (2015, p. 306) examines – as she herself points out somewhat inconclusively as regards structural convergence in the region

– the potential epicentral influence of NZE on South-Pacific Englishes in Fiji, Samoa, and the Cook Islands by analyzing existential-*there* constructions. Parviainen and Fuchs (2018) employ the private-conversation parts of the respective ICE components to establish the degree to which the use of the clause-final particles *also* and *only* as in *I do not have to work also* or *I don't get time only* in HKE and PhiE might be due to the influence of an IndE epicentre. Rooted in a linear regression model per particle, they (2018, pp. 11-12) conclude that (i) 'the lower frequency of [clause-final *also*] in HKE and PhiE, compared to IndE, is commensurate with the hypothesis of spread from IndE to HKE and possibly PhiE' and that (ii) 'the diffusion of clause-final *only* is more advanced in IndE than in either HKE or PhiE, and hence may have spread from IndE to HKE, while the results for PhiE remained inconclusive'. Both studies potentially lose valuable information in their analyses because (of how) they bin numeric predictors – in both cases SPEAKER AGE – into categorical ones for their models, although this might occasionally be a pragmatic necessity due to the degree of detail that comes with the metadata of certain corpora. Second, the existence/relevance of interactions: Both publications do not consider interaction effects properly by splitting their data sets up into smaller parts which are then analyzed separately, a well-known but cardinal sin in multifactorial modeling (Nieuwenhuis, Forstmann, & Wagenmakers, 2011; Makin & Orban de Xivry, 2019; Gries, 2021a, sections 5.2.4, 5.2.8). Third, both publications do not take into consideration the repeated-measurements nature of the data: Biewer's (2015) Varbrul analysis ignores the multiple data points provided by each speaker and Parviainen and Fuchs (2018) conduct their regressions on averages that conflate data from anywhere between 1 and 170 speakers; both publications therefore violate the assumptions of their methods and accord too much importance to too few potentially outlier speakers.[2]

Thus, we feel the following considerations should be integrated more systematically not only in epicentral research:

- More comprehensive random-effects structures: random effects for speakers were just mentioned, but other hierarchical structure may be present in corpus data. A widely-known example is using random effects for items meaning for lexically-specific effects in, say, alternation studies. A less well-recognized need for random effects involves the sampling structure of corpora. For example, Gries and Deshors's (2015) analysis of the dative alternation uses a random-effects structure that reflects how (i) files are hierarchically nested into varieties, which are nested into variety types (EFL vs. ESL); similarly, Gries and Bernaisch (2016)'s analysis of datives in South Asian Englishes uses a random-effects structure that reflects how newspapers are hierarchically nested into varieties,
- As mentioned above, numeric predictors should usually not be factorized into categorical predictors but treated as numeric (while allowing for the possibility of curved effects) – if they are factorized, they need to be treated as ordinal predictors or with successive-difference contrasts rather than as unordered categorical predictors;
- Interaction effects need to be taken more seriously both in regression modeling (see above) and in tree-based analyses (Zuur, Ieno, & Elphick, 2010).

To suggest a way of statistically tackling epicentral studies of linguistic ants, Gries, Bernaisch & Heller's (2018) study of the genitive alternation in BrE and SingE using the MuPDAR(F) protocol

ticks many boxes. Though developed by Gries and colleagues in the context of Japanese (Gries & Adelman, 2014) and learner varieties of English (Gries & Deshors, 2014; Wulff & Gries, 2015), MuPDAR(F) is applicable to the study of world Englishes and provides novel analytical perspectives for linguistic epicentres. Using this method means:

1. One applies a first regression/classifier R1 to reference speakers (RS, for example Inner Circle variety speakers);
2. If R1 works well enough, then it is used to impute for each situation each target speaker (TS, for example an Outer Circle variety speaker) was in what a RS would have said in the exact same linguistic context;
3. One determines how the actual TS choices relate to the imputed ones by
   a. just checking whether a TS made the same choice a RS would have made or not or
   b. more precisely quantifying the discrepancy between the two (for example with logloss);
4. One explores the discrepancies between RS and TS choices with a second model/classifier R2.

This MuPDAR(F) protocol was extended to epicentral research questions in Gries & Bernaisch (2016), where – across several iterations of the protocol – each South Asian English was made the reference and it was checked which of the South Asian Englishes could best predict the surface-structure choices in the remaining South Asian Englishes. With the dative and the genitive alternation (Heller, Bernaisch, & Gries, 2017), IndE served as the best model for the constructional choices concerned in South Asian Englishes.

Still, Gries, Bernaisch and Heller (2018) use this protocol (both R1 and R2 involved random effects and interactions) with diachronic BrE and SingE data and show that studies working with synchronic data only are likely flawed because of how they rely on (i) the variety-research equivalent of the apparent-time method (comparing a present variety to a historical source variety) and (ii) the assumption that the historical source variety has not undergone relevant changes itself. True, the current default of the apparent-time method can return partially correct results (true positives), but also returns false positives (effects that show up in apparent-time studies, but not in real diachronic work) and false negatives (effects that do not show up in apparent-time studies, but in real diachronic work).

Thus and as we freely admit, all the statistical sophistication in the world will not help if a statistical analysis is forced on data not meeting the assumptions made by a method – however, this is not a permit to abandon statistics, it is stating the desideratum for more collection of diachronic data to get (even) closer to an empirical identification of linguistic epicentres. The ultimate prize would be if we could use statistical analyses on fine-grained diachronic data to come closer to establishing causal effects, not just correlational/predictive ones, which, however, would require data from numerous diachronic stages of varietal development. Although this data is currently not even on the horizon for most world Englishes, structural equation modeling – an extension of regression modeling and confirmatory factor analyses that allows/forces researchers to model more explicitly the relationship between observed variables and latent constructs – enables a more reliable exploration of causal hypotheses. Granger causality, a hypothesis-testing method that checks whether some time-

series data can predict other time-series data well, deserves more detailed attention here since the notion of causation is central to epicentral modeling from theoretical and empirical angles.

If we observe an association between linguistic behavior in two varieties, this might or might not be the effect of one influencing the other since correlation does not automatically signal causation. For instance, the number of letters in the winning words of national spelling bees is apparently highly correlated with the number of people killed by venomous spiders.[3] Still, it would be far-fetched to assume a causal relationship between the two. The same might apply to epicentral research: Just because linguistic phenomena co-vary, there is not necessarily a causal link.

So how can we get closer to inferring causality, and – by extension – inferring the existence of linguistic epicentres? We should (i) assure temporal precedence of the assumed cause, and (ii) discard alternative explanations. Assuring temporal precedence in epicentral research means making sure that usage patterns in variety A – the assumed epicentre – precede the usage patterns in variety B – a variety assumed to be under epicentral influence. In practice, this also involves taking into account another property of change over time (linguistic or otherwise): autoregression. Time series (be it stock prices, sales, or linguistic usage patterns over time) rarely change radically over night; they mostly evolve gradually. In statistical terms this means that individual data points (for example a probabilistic model of a given surface-structure choice at time *t*) are not independent of their predecessors (meaning their values at *t-1*, *t-2* et cetera). Because of this autoregression, some of the variance is already explained by a time series's history. So in order to check whether one time series (for example that of a probabilistic model of a given surface-structure choice in the assumed epicentre) might be the cause of another time series (for example that of a probabilistic model of a given surface-structure choice in a variety assumed to be under epicentral influence), we have to take the history of both time series into account. Granger (1980) proposed a causality measure that does just that. It is derived from two models: (i) An autoregressive model of a time series that predicts values based on past values and (ii) another model that adds the past values of another time series as a predictor. This other time series is the potential cause of the first one. Next, the errors of the two models are evaluated and, if the second model's error is much smaller – meaning that the addition of the other time series adds explanatory value –, this suggests that this other time series might cause the first one. The Granger causality index summarizing this relationship is defined as the logarithm of the ratio of variances from model 1 and model 2; statistical significance can be determined with a Fisher exact test.

Once temporal precedence has been established, we have to rule out alternative explanations. For instance, instead of variety A influencing variety B, there might – in addition to other potentially pan-varietal forces listed at the end of section 2.1.2 – be another variety C influencing both A and B. Mair (2013), in his *World System of Englishes*, proposes a hierarchy of varieties based on systemic factors such as demographic weight and transnational reach. In principle, all superordinate varieties might be potential candidates that should be controlled for. For instance, when investigating changes in Sri Lankan English (a Central variety in Mair's model), one should also control for IndE (the relevant Super-central variety) and AmE (the Hyper-central variety).

3 | CONCLUSION

A bird's-eye view of epicentre research to date can probably be grouped into two clusters. One cluster has dedicated itself to studying linguistic butterflies, with which results are generally rooted in comparisons of aggregate frequencies of a linguistic feature according to extra-linguistic level-2 characteristics such as VARIETY, lacking the closer inspection of individual examples concerning their linguistic level-1 characteristics. At least nominally, epicentral studies of this type often try to interpret these cross-varietal comparisons of aggregate feature frequencies in the light of sociolinguistic/attitudinal information. This complementation of corpus-linguistic and sociolinguistic data appears to be facilitated by the fact that the level of analysis is usually not the individual feature choice or attitude of a given speaker, but the aggregate behavior of variety-specific speaker groups, for which such sociolinguistic/attitudinal information may be available from earlier research into the varieties concerned.

In contrast, the cluster of epicentral studies with a focus on linguistic ants is based on individual linguistic choices annotated for many level-1 as well as level-2 features (including any extralinguistic characteristics) and involves predictive modeling techniques on the level of individual choices. Sociolinguistic interpretations are restricted to what is rigorously annotatable and part of the predictive modeling technique.

While this classification is a simplification, we hope to have shown why we believe that current epicentre and world Englishes research can benefit more from the latter cluster. While its methods are not perfect and suffer from the absence of diachronic data (just as much as the former cluster), at this point in time, we feel that it is (i) linguistically more informative (given its higher level of resolution and its inclusion of more and linguistically relevant predictors), (ii) statistically more appropriate (if only for avoiding information loss and ignoring important structures in the data) and (iii) theoretically more easily reconcilable with the model character of linguistic epicentres.

NOTES

[1] Our discussion focuses on Hundt (2020, p. 2) because hers is the most recent 'meta-theoretical and meta-methodological' overview, but we feel similarly about any other publication arguing towards a separation of theoretical/model development on the one hand and empirical/statistical modeling on the other.

[2] See Gries (2021a, Ch. 6, exercises) for a detailed discussion of Parviainen and Fuchs (2018).

[3] https://www.tylervigen.com/spurious-correlations

REFERENCES

Adèr, H. (2008). Modelling. In H. J. Adèr & G. J. Mellenbergh (Eds.), *Advising on research methods: A consultant's companion* (pp. 271-304). Huizen: Johannes van Kessel Publishing.

Baker, P. (2017). *American and British English: Divided by a common language?* Cambridge: Cambridge University Press.

Bernaisch, T., & Heller, B. (2020). *Manual for the 2020-update of the South Asian Varieties of English* (SAVE2020) *Corpus*. Version 1.0. Giessen: Justus Liebig University, Department of English.

Bernaisch, T., & Lange, C. (2012). The typology of focus marking in South Asian Englishes. *Indian Linguistics*, *73*, 1-18.

Biewer, C. (2015). *South Pacific Englishes: A sociolinguistic and morphosyntactic profile of Fiji English, Samoan English and Cook Islands English*. Amsterdam: John Benjamins.

Biewer, C., Bernaisch, T., Heller, B., & Berger, M. (2014). Compiling the Diachronic Corpus of Hong Kong English (DC-HKE): Motivation, progress and challenges. Talk at *ICAME 35*, University of Nottingham, 30 April-4 May 2014.

Bolton, K. (Ed.). (2002). *Hong Kong English: Autonomy and creativity*. Hong Kong: Hong Kong University Press.

Collins, P. C., Borlongan, A. M., & Yao, X. (2014). Modality in Philippine English: A diachronic study. *Journal of English Linguistics, 42*, 68-88.

Fong, V. (2004). The verbal cluster. In L. Lim (Ed.), *Singapore English: A grammatical description* (pp. 75-104). Amsterdam: John Benjamins.

Gisborne, N. (2002). Relative clauses in Hong Kong English. In K. Bolton (Ed.), *Hong Kong English: Autonomy and creativity* (pp. 141-160). Hong Kong: Hong Kong University Press.

Grafmiller, J., & Szmrecsanyi, B. (2019). Mapping out particle placement in Englishes around the world: A study in comparative sociolinguistic analysis. *Language Variation and Change*, *30*, 385-412.

Granger, C. W. J. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, *2*, 329-352.

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*, 403-437.

Gries, S. Th. (2018). On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies*, *1(2)*, 276-308.

Gries, S. Th. (2021a). *Statistics for linguistics with R* (3rd revised and extended ed.). Boston: De Gruyter.

Gries, S. Th. (2021b). Analyzing dispersion. In M. Paquot & S. Th. Gries (Eds.), *A practical handbook of corpus linguistics*. Berlin: Springer.

Gries, S. Th., & Adelman, A. S. (2014). Subject realization in Japanese conversation by native and non-native speakers: Exemplifying a new paradigm for learner corpus research. In J. Romero-Trillo (Ed.), *Yearbook of corpus linguistics and pragmatics 2014: New empirical and theoretical paradigms* (pp. 35-54). Cham: Springer.

Gries, S. Th., & Bernaisch, T. (2016). Exploring epicentres empirically: Focus on South Asian Englishes. *English World-Wide*, *37*, 1-25.

Gries, S. Th., & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*, *9*, 109-136.

Gries, S. Th., & Deshors, S. C. (2015). EFL and/vs. ESL? A multi-level regression modeling perspective on bridging the paradigm gap. *International Journal of Learner Corpus Research*, *1(1)*, 130-159.

Gries, S. Th., Bernaisch, T., & Heller, B. (2018). A corpus-linguistic account of the history of the genitive alternation in Singapore English. In S. C. Deshors (Ed.), *Modeling world Englishes: Assessing the interplay of emancipation and globalization of ESL varieties* (pp. 245-279). Amsterdam: John Benjamins.

Gunesekera, M. (2005). *The postcolonial identity of Sri Lankan English*. Colombo: Katha Publishers.

Hackert, S. (2015). Pseudotitles in Bahamian English: A case of Americanization? *Journal of English Linguistics*, *43*, 143-167.

Heller, B. (2018). *Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English* (PhD thesis). KU Leuven, Leuven.

Heller, B., Bernaisch, T., & Gries, S. Th. (2017). Empirical perspectives on two potential epicentres: The genitive alternation in Asian Englishes. *ICAME Journal*, *41*, 111-144.

Hoffmann, S., Hundt, M., & Mukherjee, J. (2011). Indian English – An emerging epicentre? A pilot study on light verbs in web-derived corpora of South Asian Englishes. *Anglia*, *129*, 258-280.

Hoffmann, S., Sand, A., & Tan, P. (2012). *The Corpus of Historical Singapore English: A first pilot study on data from the 1950s and 1960s*. Paper presented at *ICAME 33*, KU Leuven.

Hundt, M. (2013). The diversification of English: Old, new and emerging epicentres. In D. Schreier & M. Hundt (Eds.), *English as a contact language* (pp. 182-203). Cambridge: Cambridge University Press.

Hundt, M. (2020). On models and modelling. *World Englishes*, 1-20.

Kruger, H., van Rooy, B., & Smith, A. (2019). Register change in the British and Australian Hansard (1901-2015). *Journal of English Linguistics*, *47(3)*, 183-220.

Labov, W. (1975). Empirical foundations of linguistic theory. In R. Austerlitz (Ed.), *The scope of American linguistics* (pp. 77-133). Lisse: Peter de Ridder Press.

Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.

Leitner, G. (1992). English as a pluricentric language. In M. Clyne (Ed.), *Pluricentric languages: Differing norms in different nations* (pp. 179-237). Berlin: De Gruyter.

Leitner, G., & Sieloff, I. (1998). Aboriginal words and concepts in Australian English. *World Englishes*, *17*, 153-168.

Lim, L. (Ed.). (2004). *Singapore English: A grammatical description*. Amsterdam: John Benjamins.

Mair, C. (2013). The World System of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide*, *34*, 253-278.

Makin, T. R., & Orban de Xivry, J. J. (2019). Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife*, *2019*, 8:e48175.

Manning, C. D. (2003). Probabilistic syntax. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 289-342). Cambridge, MA: MIT Press.

Meyerhoff, M., & Niezielski, N. (2003). The globalisation of vernacular variation. *Journal of Sociolinguistics*, *7*, 534-555.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*, 1105-1107.

OED Online. (2020). model, n. and adj. Oxford: Oxford University Press. Retrieved from www.oed.com/view/Entry/120577

Panther, K. (2012). Motivation in language. In S. Kreitler (Ed.), *Cognition and motivation: Forging an interdisciplinary perspective* (pp. 407-432). Cambridge: Cambridge University Press.

Parviainen, H. (2020). *Crossing the borders: The influence of Indian English in the Southeast Asian region* (PhD thesis). Tampere University, Finland.

Parviainen, H., & Fuchs, R. (2018). 'I don't get time only': An apparent-time investigation of clause-final focus particles in Asian Englishes, *Asian Englishes*, *21(3)*, 285-304.

Peters, P. (2009). Australian English as a regional epicentre. In T. Hoffmann & L. Siebers (Eds.), *World Englishes: Problems, properties and prospects: Selected papers from the 13th IAWE Conference* (pp. 107-124). Amsterdam: John Benjamins.

Peters, P., Smith, A., & Bernaisch, T. (2019). Shared lexical innovations in Australian and New Zealand English. *Dictionaries*, *40(2)*, 1-30.

Roethlisberger, M. (2018). *Regional variation in probabilistic grammars: A multifactorial study of the English dative alternation* (PhD thesis). KU Leuven, Leuven.

Schneider, E. W. (2004). How to trace structural nativization: Particle verbs in world Englishes. *World Englishes*, *23*, 227-249.

Schneider, E. W. (2007). *Postcolonial English: Varieties around the world*. Cambridge: Cambridge University Press.

VanPatten, B., Williams, J., Keating, G. D., & Wulff, S. (2020). Introduction: The nature of theories. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (pp. 1-17). New York: Routledge.

Wulff, S., & Gries, S. Th. (2015). Prenominal adjective order preferences in Chinese and German L2 English: A multifactorial corpus study. *Linguistic Approaches to Bilingualism*, *5*, 122-150.

Zuur, A., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*, 3-14.