Profiling verb complementation constructions across New Englishes

A two-step random forests analysis of *ing* vs. *to* complements

Sandra C. Deshors and Stefan Th. Gries
New Mexico State University / University of California, Santa Barbara

In this paper, we explore verb complementation patterns with *to* and *ing* in native English (British and American English) as compared to three Asian Englishes (Hong Kong, Indian, and Singaporean English). Based on data from the International Corpus of English annotated for variables describing the matrix verb and the complement, we run two random forests analyses to determine where the Asian Englishes have developed complementation preferences different from the two native speaker varieties. We find not only a variety of differences between the Asian and the native Englishes, but also that the Asian Englishes are more similar (i.e. 'better predicted by') the American English data. Further, as the first study of its kind to extend the MuPDAR approach from the now frequent regression analyses to random forests analysis, this study adds a potentially useful analytical tool to the often messy and skewed observational data corpus linguists need to deal with.

Keywords: verb complementation, *ing* vs. *to*, New (Asian) Englishes, MuPDAR, random forests

1. Introduction

The English language is rich in syntactic alternations: the dative alternation, particle placement, the genitive alternation, *that*/0-complementation, to name but a few. Many of these alternations exhibit considerable variation and much of that variation is co-determined by a large number of linguistic and cognitive factors that cut across alternations (e.g. constituent lengths, information statuses, animacy statuses, priming, etc.). What is more, some of these alternations are lexically-specific

International Journal of Corpus Linguistics 21:2 (2016), **192–218**. DOI 10.1075/ijcl.21.2.03des
ISSN 1384–6655 / E-ISSN 1569–9811

in the sense that, all other things being equal, particular words in the grammatical context of a given alternation pattern increase the chance of a particular syntactic choice. For a long time now, those alternations have been a point of focus in native language (L1) research (see Green 1974, Ransom 1979, Collins 1995, Gries 2003, Bresnan et al. 2007 for studies on the dative alternation and Noonan 1985, Duffley 1999, Smith & Escobedo 2002 for studies on the *ing vs. to* alternation). However, as scholars have gained knowledge on the governing principles behind syntactic choices in L1, they have only recently started to explore how speakers other than the "traditional L1 speakers" handle the probabilistic uncertainty of these choices. As a result, there is now a fast growing body of corpus-based studies on alternations in EFL (i.e. foreign varieties of English learned in countries where English is not institutionalized; see Deshors 2014a, Gries & Deshors 2015, Gries & Wulff 2013) as well as in ESL (i.e. indigenized varieties of English learned as a second language in countries such as Hong Kong, India or Singapore and where English is institutionalized and learners have daily contact with the English language; see Bernaisch et al. 2014, Deshors 2014b, Schilk et al. 2013, Nam et al. 2013).

With regard to ESL specifically, much of the research on syntactic alternations has focused on the case of Asian Englishes. For instance, recent studies such as Gries & Bernaisch (2016) have investigated the dative alternation with GIVE (*John gave Mary a book* vs. *John gave a book to Mary*) across six south Asian Englishes with a view to (i) identify factors triggering different constructional choices across ESL varieties and (ii) capture the linguistic epicenter of English in South Asia, using a logistic-regression approach with random effects (described below as the 'MuPDAR' approach). The study clearly illustrates that as researchers have probed more deeply into syntactic alternations, their methodology has become more sophisticated. Although early studies in both learner corpus research (targeting EFL speakers and their alternation behaviors) and variety corpus research (targeting ESL speakers and their alternation behaviors) are corpus-based, their "statistical design" remained relatively simple: many studies did little more than cross-tabulation of frequencies and percentages of constructional choices across different L1 backgrounds and varieties, maybe followed up by chi-squared or loglikelihood ratio tests (see Shastri 1996 for an example of an early corpus-based study of *ing* vs. *to* complement patterns in Indian English). A first major improvement was then to apply regression modeling approaches to these questions of cross-variety variation with a view to model constructional choices as a function of several predictors. With this regression approach, scholars have started to do justice to the fact that those constructional choices are multifactorial in nature (i.e. they investigate syntactic variants on the basis of a variety of linguistic and cognitive factors; see Bernaisch et al. 2014, Bernaisch & Gries 2015 and Deshors 2015 for such examples). A second currently ongoing improvement involves computing mixed-effects

regression models. While mixed-effect models account for the multifactorial nature of constructional choices, unlike fixed-effect models, they also allow analysts to account for the hierarchical structure of their corpus data and to control for effects that are speaker-specific. So in sum, the current state-of-the-art in cross-variety variation research is essentially a regression-based approach that, ideally, is (i) multifactorial$_1$ (in the sense of including multiple independent variables at the same time), (ii) multifactorial$_2$ (in the sense of also including interactions of these independent variables), and (iii) involves random effects for variables whose levels in the data do not exhaust the levels observable for the phenomenon in the general population (cf. Gries 2015a).

These methodological strategies have already started to provide the ESL research community with powerful tools to tease apart different populations of ESL speakers and establish, with a great deal of precision and reliability, how native and ESL speakers differ in their linguistic choices. Recent work in learner corpus research and varieties research has taken this methodological development to the next level with the so-called 'MuPDAR' (Multifactorial Prediction and Deviation Analysis with Regressions). MuPDAR involves a two-step regression approach. In a first step, a regression R_1 is fit on the part of a dataset representing a reference level (e.g. native speakers or a historical source variety), and the R_1 is used to make predictions for the other part of the same dataset, namely a target sample (e.g. EFL or ESL varieties). In other words, with MuPDAR, one can determine for every utterance in the target sample what the predicted "canonical" (or native-like) choice would have been, and a second regression R_2 can be run to determine what leads to "non-canonical" choices by speakers. That is to say, this approach answers one of the central questions for comparing different speakers: in the multivariately-annotated situation the non-native speaker is in now, what would the (native) speaker of a certain reference variety say (and why is it that the non-native speaker did not make the "native-like" choice)? In the best of cases, both of these regressions involve mixed-effects models (e.g. Gries & Adelman 2014, Gries & Deshors 2015, Wulff & Gries 2015).[1]

In this paper, we apply a variant of the MuPDAR approach to an alternation which has not been targeted much outside of native English, the *to* vs *ing* alternation in Example (1), and whose study, if not targeting native speakers, has not been corpus-based, has focused mostly on lexically-specific preferences of the two alternating patterns and has rarely included other predictors (with the exception of Deshors 2015 and Khamis 2015).

1. Although the term 'non-native' is often used in the literature to refer to erroneous use, in this paper, the term is simply used to refer to language use that significantly differs from native use; we use NatE as an abbreviation here for British English (BrE) and American English (AmE).

(1) a. She prefers to eat French locust.
b. She prefers eating French locust.

That is, in this paper we are the first to bring several strands and methodological aspects of research together in the study of infinitive and gerundial verb complementation in ESL, namely:

i. a rigorously corpus-based/empirical analysis (as recommended in Kaleta 2012);
ii. a multifactorial statistical design in which we model speakers' constructional choices as a function of several predictors at the same time;
iii. a comparison of native speakers and ESL speakers from three Asian Englishes at different development stages in Schneider's (2007) evolutionary model, specifically, Hong Kong English (HKE; phase II–III), Indian English (IndE; phase III–IV) and Singaporean English (SingE; advanced phase IV); also, for reasons to be explained in detail below, we extend the notion of native speakers to not just British English speakers but also American English speakers;
iv. a MuPDAR-like approach in which we attempt to determine precisely if and how Asian Englishes speakers' uses of *to* and *ing* complements differ from those of native speakers.

Given the absence of ESL research on the two complementation patterns in question and the fact that comparisons between EFL and ESL are not unproblematic (since the two variants are not necessarily directly (qualitatively) comparable; see Szmrecsanyi & Kortmann 2011, Gries & Deshors 2015), the goal of this paper remains largely descriptive. To date, there are just not many studies from which one can derive working hypotheses on the principles that govern ESL speakers' choices of *to* and *ing* complement patterns and that are precise enough to be testable. Further, as will become clear below, this paper also introduces a methodological alternative to the currently prevalent state-of-the-art of regression approaches to address cases where data turns out to be too noisy to be analyzable using regression modeling. In the next section, we will briefly discuss previous work on verb complementation patterns and of course in particular *to* vs. *ing*; the discussion will be brief and focus on the factors that have been claimed to be associated with the complementation choice under investigation and the relevance of grammatical factors for a better understanding of non-finite verb complementation patterns in ESL.

2. *To* and *ing* complementation patterns and their grammatical environments

In the context of a discussion on the influence of grammatical environments over speakers' syntactic choices, it is important to keep in mind that while certain grammatical features associate more strongly than others with the choice of a particular complementation patterns, the extent of this association can also vary depending on the speakers' English variety (i.e. native vs. ESL). Section 2.1 below focuses on the grammatical features associated with the two syntactic variants in focus and Section 2.2 discusses the relevance of exploring verb complementation patterns through the lens of cross-variety variation.

2.1 The contribution of grammatical factors to *to* and *ing* complementation

Traditionally, the relations between predicates and complements, and the choice between infinitival and gerundial complements specifically, have been investigated from a semantic perspective, based on Noonan's (1985: 88) observation that "[c]omplementation is basically a matter of matching a particular complement type to a particular complement-taking predicate" (see also Wierzbicka 1988, Duffley 1999, Langacker 1991, Smith & Escobedo 2002). This line of approach mainly focuses on the types of verbs that license a particular complement. For instance, in infinitival cases expressing some kind of potentiality, complement clauses tend to involve lexical verbs that are compatible with such notion of potentiality (De Smet 2013). In gerundial cases, three major verb groups were observed in De Smet (2013) to attract gerund complements: verbs associated with negative implications (a*void*, *defend*, *omit*, etc.), emotive verbs (*enjoy*, *hate*, *like*, etc.), and aspectual verbs (*begin*, *continue*, *stop*, etc.). However, although semantically-based research in complementation patterns has proved to be relatively fruitful, semantic approaches are not sufficient to accurately predict speakers' choices of complements (Smith & Escobedo 2002).

Against this background, studies such as Noël (2003), Vosberg (2003), Rohdenburg (1995), Mindt (2000), and Mair (2002) have shown the need to include additional non-semantic factors to the analysis of *to* vs. *ing* alternations such as information structure, the *horro aequi* principle, the cognitive complexity principle, social and regional stratification, and register. One particularly illuminating study is Cuyckens et al. (2014), which focuses on the finite vs. non-finite complement patterns of *remember*, *regret* and *deny* in Late Modern English (i.e. complement-taking predicate + *to*-infinitive, gerundial *-ing* and *that*-clause). Cuyckens et al.'s (2014) study investigates complement patterns based on a wide range of language-internal and language-external factors (e.g. the meaning of the matrix

verb, the meaning of the complement clause, the relation between the time reference of the complement and the meaning of the complement-taking predicate, type of subject in the complement clause, structural complexity of the complement clause, the type of subject in the main clause, the animacy of the subject in the complement clause, the voice in the complement clause, the type of complementation, and others). Using a binary logistic regression approach with fixed effects, the authors pinpoint the factors that favor and to some degree characterize non-finite verb complements in native English. For instance, they find that main clauses with a first-, second- and third- person pronoun as well as a noun favor *to* and *ing* complements. Similarly, speakers tend to prefer non-finite clauses with more complex complement clauses (i.e. clauses including verbs with one argument/modifier, verbs with an argument + modifier, or two arguments, or two modifiers). Finally, another factor that favors non-finite complements is passive voice. While Cuyckens et al. (2014) clearly show the benefits of combining a contextualized approach to verb complementation with a sophisticated statistical methodology, their results urge us to take an even closer look at *to* and *ing* complements and assess the contribution of linguistic factors to each construction variant. In the context of indigenized English varieties, this is a particularly important goal to pursue, given recent developments in ESL research showing how specific grammatical contexts can be the loci of differentiation of individual ESL varieties from native English.

2.2 The contribution of grammatical factors to alternating verb complements in ESL

Recent literature on post-colonial Englishes recognizes the value of investigating the interface between lexis and grammar to better characterize ESL varieties. With specific regard to complementation, the field is currently witnessing a fast growing development of corpus-based research aimed to track systematic deviations in the contexts of use of complement constructions (see Olavarría de Ersson & Shaw 2003; Mukherjee & Hoffmann 2006; Mukherjee & Schilk 2008; Schilk et al. 2012, 2013; and Deshors 2014a). One particularly interesting study is Schilk et al. (2012) in which the authors focus on the complementation patterns of the lexical verbs *convey*, *submit* and *supply* (which are typically used in the transfer-caused motion construction [TCM] such as *give something to someone*) across Indian, Sri Lankan and British Englishes. Drawing from the verbs' contexts of use, the authors find that across ESL varieties there are "differences [that] manifest themselves at the level of individual complementation patterns, and in many cases these differences can be explained if we look at the actual linguistic realizations of the patterns and the lexical items used in them" (Schilk et al. 2012: 162). Ultimately, it emerges that a focus on the lexico-grammatical environments of complement constructions

offers new insights into distinctive and so far largely neglected structures of English varieties (Schilk et al. 2012).

As hinted in our introduction, in order to explore cross-variety variation at the lexis-grammar interface, ESL scholars are increasingly starting to adopt regression statistical approaches that allow them (i) to include in their analyses wide ranges of linguistic factors and (ii) to analyze alternating syntactic pairs by cutting across the semantic and morphosyntactic levels. For instance, Nam et al. (2013) use this approach to investigate three complementation patterns of GIVE (ditransitive, prepositional dative and monotransitive) and they begin to unveil the governing principles behind ESL speakers' choices of one construction over another. However, despite their usefulness, multifactorial approaches in ESL have mainly been used to investigate the dative alternation and their application remains to be extended to other syntactic alternations. One recent exception is Deshors (2015) which explores *to* vs. *ing* across native and Hong Kong English using binary logistic regression modeling on the basis of 3,119 occurrences of the two complement patterns. Overall, the study reveals that Hong Kong English speakers overuse *to* complements with a non-finite verb form in the predicate, complement verbs denoting a cognitive process and objects expressed in the form of either a prepositional phrase, a noun phrase or a pronoun. Although Deshors' (2015) study is a first step towards distinguishing the linguistic features that generally contribute to each complement construction in native and Hong Kong Englishes, the study investigates only one ESL variety and does not follow the newly developed MuPDAR protocol.

3. Methodology

For the purpose of this study, we used data from the International Corpus of English (ICE; see Greenbaum 1996). In Section 3.1, we explain how our data were extracted as well as our annotation procedure and in Section 3.2 we present our statistical approach.

3.1 Corpus data and their annotation

The present work contrasts the uses of gerundial and infinitival complement constructions across different subsections of the ICE corpus, as recommended in Mukherjee & Gries (2009). Therefore, our study includes the following sub-parts from the corpus: native English, British (BrE) and American (AmE), and three ESL varieties, Hong Kong (HKE), Indian (IndE) and Singaporean (SingE) Englishes. To some readers, our choice to include American data as representative of native English may seem somewhat surprising, mainly because, traditionally, BrE has been

the default variety considered as the historical source variety for Asian Englishes. However, even though Asian Englishes have been historically most influenced by BrE, in an age in which American culture leaves a mark on cultures everywhere — via music, television, and cinematic culture — it is fair to assume that Englishes all over the world are now beginning to not only look back to BrE as a historical source variety, but also to AmE as a contemporarily perhaps even more influential source variety (see Mair's (2013) article on the world system of Englishes for a justification of why, theoretically, it makes sense to use American English as a reference variety in addition to BrE). Further, an increasing number of contrastive studies on ESL varieties are beginning to go beyond BrE as the only relevant native variety/standard of comparison: for our purposes, more importantly, Edwards (2014) and Koch (2015) include AmE in their analyses, and Hoffmann (2014) uses four native varieties, namely BrE, NZE, CanE, and IrE. Thus, there is clearly the beginning of a trend to look beyond BrE only and we are following this recent development. Finally, as we will show below in Section 4.4, in the present data, it is in fact the AmE data that the Asian Englishes are more similar to, not BrE. Therefore, it emerges that theoretical considerations as well as empirical data converge in their "recommendation" to at least exploratorily also consider AmE in the data (until proven otherwise).

The data were extracted from the relevant written sub-sections of ICE. No distinction across writers was made on the basis of age or educational background and all material was extracted and statistically analyzed using the software *R* (R Development Core Team 2012). With regard to the compilation of the data set, we followed Martínez-García & Wulff's (2012) methodology: for all sub-corpora, we first retrieved all instances of words ending in *-ing* and all instances of the preposition *to*. In a second step, true hits of either complementation construction were checked manually for syntactic relevance, yielding a sample of approximately 7,400 instances of gerundial and infinitival complementation construction from each of the five corpora. This data sample was then trimmed down so as to only include the verbs that were attested in both *to* and *ing* constructions. Table 1 provides an overview of the distribution of the two constructions across all sub-corpora.

Table 1. Summary of the occurrences of *ing* and *to* constructions in the sub-corpora

Complement pattern	ICE-GB	ICE-US	ICE-HK	ICE-IND	ICE-SING	Total
ing	84	187	126	102	128	627
to	990	781	753	531	753	3,808
Total	1,074	968	879	633	881	4,435

Each match was annotated against twelve grammatical factors (see Table 2 for a list of all those factors and their respective levels). To ensure a thorough treatment of the data, each factor was encoded according to a taxonomy established to

allow for its measurement and its consistent treatment across the five sub-corpora. The annotation process was carried out using a careful bottom-up approach and a taxonomy that is theory-consensual in nature. In other words, our annotation is based on the linguistic analysis of the context of use of each complementation pattern extracted from the corpus and it reflects the unique combination of grammatical components included within their context of use. So, similarly to Divjak & Arppe (2013), the annotation scheme accounts for both exemplars (i.e. specific instances of use) and their abstractions. This is an important aspect as no other study (but Deshors 2015) has so far adopted this type of data annotation to contrast gerundial and infinitival complementation patterns in ESL. Table 3 presents an abbreviated sample of the annotation table.

Table 2. Overview of the variables used in the annotation of the native English and ESL data and their respective levels

Variable	Variable levels
COMPLPATTERN (complementation pattern; dependent variable)	gerund, infinitive
LEMMACOMP (lemma in the complement clause)	be, sponsor, work, perform, …
LEMMAMATRIX (lemma in the matrix clause)	appear, seem, accept, adopt, …
COUNTRY (English variety)	british, american, hong kong, indian, singaporean
FINITEMATRIX (finite or non-finite use of the verb in the matrix clause)	finite, non-finite
VOICE (voice of the matrix verb)	active, passive
COMPVERBSEM (semantics of the complement's lexical verb)	abstract, action, communication/informational, copula, cognitive/emotional, perception
MATRIXVERBSEM (semantics of the matrix's lexical verb)	abstract, action, communication/informational, copula, cognitive/emotional, perception
MATRIXVERBTYPE (type of the matrix verb)	state, accomplishment, achievement, process
COMPTVERBTYPE (type of the complement verb)	state, accomplishment, achievement, process
NEG (negation)	neg, affirm
OBJECTFORM (form of the object)	pp (prepositional phrase), np (noun phrase), do (double object), pr (pronoun), no(no object)

Table 3. Abbreviated sample of the annotation table

COMPLPATTERN	COUNTRY	VOICE	COMPVERBSEM	MATRIXVERBTYPE	COMPVERBTYPE	OBJECTFORM
to	sing	active	abstract	state	state	np
ing	hk	passive	abstract	process	process	np
…	…	…	…	…	…	…

It is important to note here that at the semantic level, our taxonomy provides a more fine-grained approach to lexical verbs in predicate and complement clauses than is currently offered in existing studies. This is because the taxonomy distinguishes between *types* of lexical verbs and their *semantics*. Broadly, the MATRIXVERBTYPE and COMPVERBTYPE variables mark the types of lexical verbs used in the predicate and in the complement clause. Conceptually, the two variables follow Vendler (1957: 143) in its recognition that the notion of time is crucially related to the use of a verb and is "at least important enough to warrant separate treatment". Vendler (1957) identifies four types of verbs namely *state*, *accomplishment*, *achievement* and *process*. This verb classification distinguishes between time periods and time instants on the one hand and uniqueness/definiteness and non-uniqueness/indefiniteness of those time periods and time instants on the other hand. As Vendler (1957: 146) notes, "some verbs can be predicated for single moments in time, while others can be predicated for shorter or longer periods of time". In that respect, accomplishment verbs encode verbal statements that imply a unique and definite time period and achievement verbs encode verbal statements that imply a unique and definite time instant. Similarly, process verbs identify statements that reflect non-unique and indefinite time periods and state verbs identify statements that reflect non-unique and indefinite time instants. Like the MATRIXVERBTYPE and COMPVERTYPE variables, the MATRIXVERBSEM and the COMPVERSEM variables target lexical verbs used in the predicate and in the complement clause, and they identify the type of information that lexical verbs convey in terms of abstraction, action, communication, etc.

3.2 Statistical evaluation

The present data proved extremely difficult to analyze by what, over the last few years, has become the standard statistical analysis, i.e. regression modeling of the traditional or the MuPDAR kind, which initially was our preferred choice. However, the data set proved to be extremely recalcitrant. In a first attempt to explore if and how the complementation patterns differ between the native and the Asian Englishes, we considered a multifactorial$_2$ regression approach with a bidirectional model selection procedure based on *AIC*. That is to say, we fit a very small model in which the COMPLPATTERN (*ing* vs. *to*) is predicted only from COUNTRY as well as varying intercepts for FILE and LEMMAMATRIX (varying intercepts for LEMMACOMP were not attempted to fit because of the large number of verb lemmas with extremely low frequencies; for the same reason, no varying slopes were implemented) and then tried to improve the model by adding or subtracting the predictors that improved the model fit most (with *AIC* as an indicator of quality). However, it quickly became obvious that this route was not feasible

because after having added five predictors this way, data sparsity and collinearity led to models that did not converge and whose confidence intervals covered the whole probability interval from 0 to 1.

In a second attempt and to overcome such problems, we used a multimodel inferencing approach (see Burnham & Anderson (2002) for a statistical introduction, Kuperman & Bresnan (2012) for the first approach in linguistics we are aware of, and Gries (2015b) for a recent application in cognitive/usage-based linguistics). However, multiple attempts to use this strategy were not successful either given the large number of possible combinations of predictors and all their interactions with COUNTRY, the variable representing the variety. In fact, a multimodel inferencing script was shut down after running for about 200 hours on 7 cores of an 8-core Intel i7 3.4 GHz processor with Multithreaded BLAS/LAPACK libraries. Given these results, it became clear that the current default of regression modeling was not going to be successful with the present data; that means a lot of the conceptual machinery of regression modeling we would have been interested in was not available to us anymore — interactions of predictors, random-effects/multilevel structure, and user-defined a priori contrasts — and an alternative approach to analyzing the data had to be developed. We decided on an analysis involving random forests of the kind used by Bernaisch et al. (2014) for the dative alternation in Asian Englishes. Random forests is an approach that is similar to classification trees, but also extends it considerably. Classification (and regression) trees are a partitioning approach that consists of successively splitting the data into two groups based on some independent variables such that the split maximizes the classification accuracy regarding the dependent variable within the groups. This process is recursive, i.e. repeated until no further split would increase the classification accuracy sufficiently. Random forests in turn add two layers of randomness to the analysis, which help (i) recognizing the impact of variables or their combinations that a normal classification tree might not register and (ii) protecting against overfitting. On the one hand, the algorithm constructs many different trees (we set that parameter to 2,000), each of which is fitted to a different bootstrapped sample of the full data. On the other hand, each split in each tree could choose from only a randomly-chosen subset of predictors (we set that parameter to three predictors). The overall result is then based on amalgamating all 2,000 trees that have been generated.

While this approach does not provide us with all that regression analyses would have to offer (to those data sets where they work), it comes with several advantages, too: random forests are known to generate quite good accuracies, they can be computed fast, they do not make the kinds of distributional assumptions that regression models do (and that observational/corpus data often violate), and, as explained above, because of the sampling procedures they overfit and overlook important predictors less easily and are better able to handle collinear predictors

(see Matsuki et al. (2016) for similar conclusions and the additional finding that random forests are well suited for many-predictors-few-datapoints problems); we are using the implementation in the R package randomForest (Liaw & Wiener 2015, version 4.6–12). The perhaps thorniest issue of random forests is how to interpret/visualize their results. Since random forests consist of thousands of very different trees, there is no obvious way to compute *p*-values for predictors (main effects or interactions) and there are no obvious ways to determine the effect of a predictor in isolation. The solution we adopt here is a heuristic, but one that worked very well in Bernaisch et al. (2014): we represent average predictions for the (combinations of) variables of interest and plot them in dotcharts. While the potential downside of this approach is that these predicted probabilities are based on more predictors that the one they are visualized for — in other words, they are not based on only the predictor being visualized with every other predictor held constant as in effects plots for regressions — they nevertheless provide a good approximation for the effects of predictors that is convenient and more interpretable than other statistics; in addition, we provide variable importance measures (as provided by Liaw & Wiener's (2015) package).

To sum up, in this paper we are extending Gries and colleagues' (Gries & Adelman 2014, Gries & Deshors 2014, Gries & Bernaisch 2016, Wulff & Gries 2015) MuPDAR approach from one based on (two) regressions to one based on two random forests (MuPDARF, with RF standing for random forests). Specifically, we:

i. do a random forests analysis on only the native BrE and AmE speakers and test whether its fit is good enough to proceed; this analysis uses COMPLPATTERN as the dependent variable and the following as predictors: FINITEMATRIX, VOICEMATRIX, NEGMATRIX, OBJECTFORM, VERBSEMCOMP, VERBTYPECOMP, VERBSEMMATRIX, VERBTYPEMATRIX, and COUNTRY (for the variety);
ii. if the fit is good enough, we apply the results from the first random forests analysis to the HKE, IndE, and SingE speakers to obtain predictions of what native speakers would have said in the contexts that the indigenized variety speakers were in;
iii. compare the native-speaker predictions against the indigenized-variety choices to see how much the two coincide; for that we compute a numeric variable called DEVIATION, which is:
    a. set to zero when the indigenized-variety speaker made the choice a native speakers is predicted to have made;
    b. between -0.5 and 0 when the indigenized-variety speaker chose *to* although the native speaker is predicted to have chosen *ing*;
    c. between 0 and 0.5 when the indigenized-variety speaker chose *ing* although the native speaker is predicted to have chosen *to*.

The exact value depends on how strongly the native speaker was predicted to choose *to*/*ing*. Thus, higher absolute values of DEVIATION indicate that indigenized-variety speaker made choices that are more at odds with what native speakers were predicted to have said.

iv. do a second random forests analysis that models the non-nativelike choices of indigenized-variety speakers, i.e. DEVIATION in all cases where DEVIATION≠0, as a function of, again, FINITEMATRIX, VOICEMATRIX, NEGMATRIX, OBJECTFORM, VERBSEMCOMP, VERBTYPECOMP, ERBSEMMATRIX, VERBTYPEMATRIX, and COUNTRY.

4. MuPDARF: A step-by-step presentation of our results

As explained in Section 3, our statistical approach involves a series of steps. Therefore, in this section we report on our findings in way that reflects the sequence of our various methodological steps. Accordingly, Section 4.1 presents the results of the first random forests analysis (on native-speaker data), Section 4.2 reports on predicted syntactic patterns in ESL, Section 4.3 presents the results of the second random forests analysis (on ESL-speaker data), and Section 4.4, which takes a closer look at BrE vs. AmE, reports on how, with a MuPDARF approach, we are able to pinpoint the specific native English variety that predicts best verb complementation patterns in ESL.

4.1 Random forests 1 on native-speaker data

The first analysis yielded a classification accuracy of 88.5%, which is not much, but significantly higher than the baselines of always choosing the more frequent complementation pattern (i.e. *to*) or choosing proportionally randomly ($p_{\text{binomial test against baseline1}}<0.01$, $p_{\text{binomial test against baseline2}}<10^{-40}$). More illuminating is the analysis's *C*-value, which just about exceeds the usually-assumed threshold value for "good" results of 0.8 with a value of 0.81. We therefore proceeded with the analysis.

4.2 Applying the first results to the indigenized variety data

We then used the above results to compute a random forests-based prediction for every case in the ESL variety data. As in previous MuPDAR analyses, the results here were more mixed in the sense that the classification accuracy measure went down a bit (to 85.2%), which is little surprising given that one would expect the

ESL speakers to not behave completely predictably from the native-speaker data. Correspondingly, the *C*-value also decreased to 0.76, indicating that (aspects of) the choices that ESL speakers make are not completely compatible with those predicted for the native speakers, and it is precisely that difference that the second regression or, here, the second random forests analysis explores. As mentioned above, we then computed the DEVIATION variable that captures the degree, if any, to which the ESL speakers' choices differed from the native-speaker predictions. Descriptively, it was already interesting to notice that there are significant differences between the three Asian Englishes (according to a Kruskal-Wallis rank sum test: χ^2=7.24, *df*=2, *p*=0.027) such that the IndE speakers differ more from the native speaker predictions than the HKE and SingE speakers.

4.3 Random forests 2 on deviations from native-speaker predictions

The final analysis consisted of trying to model non-zero cases of DEVIATION on the basis of the same predictors as before. The overall summary results were very encouraging in the sense that the statistical analysis could predict the DEVIATION-values very well (adjusted R^2=0.87), which is why we felt justified to explore the results further, first, by assessing the importance of individual variables and, second, by looking at how the values of DEVIATION differ for the crossing of every predictor and COUNTRY.

As for the former, Figure 1 shows a plot that summarizes two indices of variable importance: on the *x*-axis, we show a normalized measure of the size of the prediction error, which, with some simplification, results from how much worse predictions become if the variable in question has its values randomly permuted; on the *y*-axis, we show a measure that represents how clean the splits in all classification trees are based on the residual sum of squares, a measure also commonly used in regression analyses; for both axes, high values reflect high variable importance.

As is obvious, both measures of variable importance largely coincide, with VERBSEMMATRIX, VERBTYPEMATRIX, VERBSEMCOMP, and OBJECTFORM having the strongest effects. Interestingly, COUNTRY on its own does not appear to have a strong effect but, on the other hand, just like it would be in a regression context, it is mostly the interactions of COUNTRY with other predictors that one would be interested in anyway. In what follows, we will discuss those roughly in order of variable mention in Figure 1.

Figure 2 shows what is roughly equivalent to the interaction VERBSEMMATRIX : COUNTRY in a regression context. Both panels show the same result with predicted DEVIATION-values on the *x*-axis, but in the top one, the levels of VERBSEMMATRIX are nested into COUNTRY (to facilitate comparisons between semantic classes per

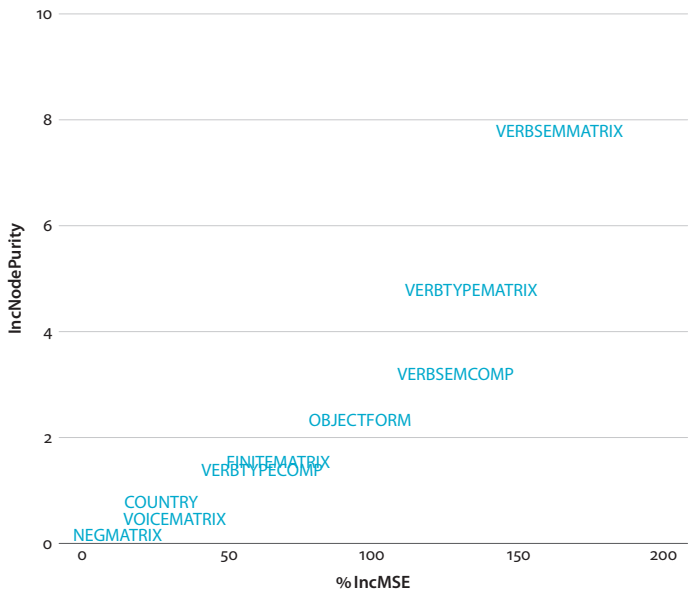

Figure 1. Variable Importance measures from the second random forests analysis

country/variety), whereas in the bottom one, the levels of COUNTRY are nested into VERBSEMMATRIX (to facilitate comparisons of countries per semantic class).

The results show that SingE is most similar to NatE, closely followed by HKE, with IndE deviating from NatE more. Overall, this variable-specific result is compatible with the overall result mentioned above. However, it is also obvious from the result that IndE differs most from NatE in three semantic contexts:

i. with action verbs in the matrix verb slot, IndE speakers are much more likely to use *to* than native speakers;
ii. with abstract verbs in that slot, IndE speakers are more likely to use *ing* than native speakers;
iii. with communication verbs, IndE speakers are a bit more likely to use *to* than native speakers.

Figure 3 is a corresponding representation of the results for VERBTYPEMATRIX : COUNTRY. In this case, although the results for the three ESL varieties are much more similar to each other, the same ordering of similarity to NatE is observed: SingE > HKE > IndE. However, here, the more interesting aspect is that variation across ESL varieties is limited to specific semantic classes: for states and accomplishments, the three varieties do not differ much from NatE, nor do they differ from each other. In the specific case of states, all Asian speakers slightly overuse *ing*. For accomplishments, most Asian speakers slightly overuse *to*. For processes

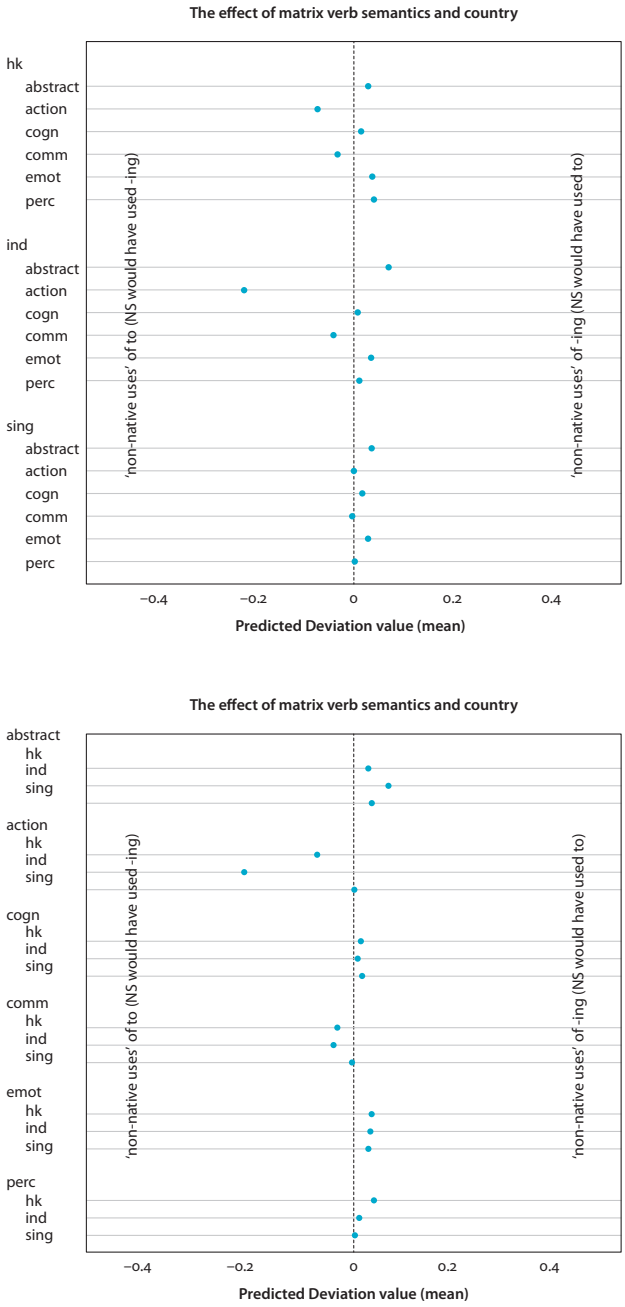

Figure 2. The effect of matrix verb semantics and country across HKE, IndE and SingE

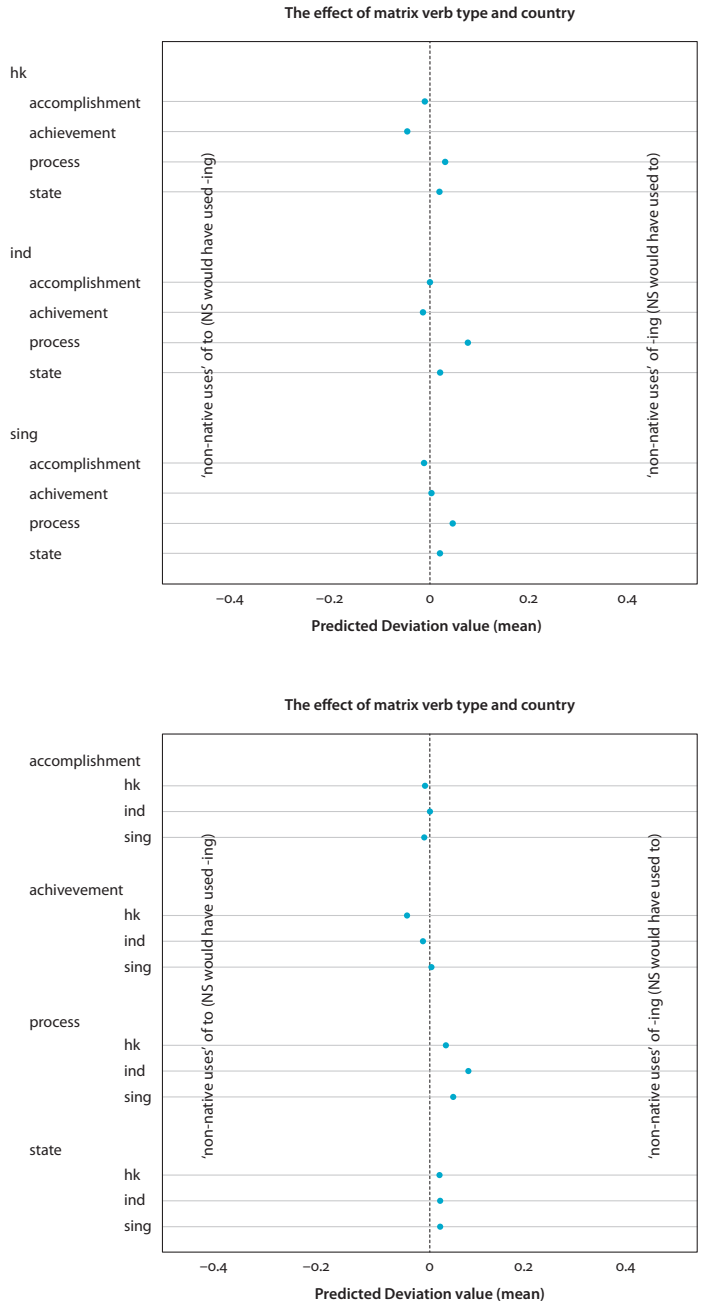

Figure 3. The effect of matrix verb type and country across HKE, IndE and SingE

verbs (e.g. *try*, *resist*, *contemplate*, *seek*), all Asian varieties use *ing* more often than native speakers would whereas, for achievements (e.g. *decide*, *stop*, *die*), HKE and IndE speakers use *to* more than native speakers would.

Figure 4 is concerned with VERBSEMCOMP : COUNTRY. The results are somewhat similar to those of VERBSEMMATRIX : COUNTRY. Again, HKE and SingE are more similar to the native-speaker choices (although this time HKE is slightly more similar to NatE than SingE) compared to IndE, which exhibits some higher deviations. As before, however, this deviation of IndE is not found across the board — it is most pronounced with complement verbs denoting emotions (e.g. *suffer*, *hate*, *enjoy*), actions (e.g. *perform*, *create*, *build*), and communication (e.g. *request*, *inform*, *discuss*). Interestingly, now that we are looking at the complement clause, some tendencies are reversed: while IndE speakers again overuse *ing* with emotion verbs (this time in the complement clause, not as above in the matrix clause), with action and communication verbs, they now, in the complement clause, overuse *to* relative to native speakers.

Figure 5 represents the results for OBJECTFORM : COUNTRY. The upper panel is not particularly informative, but it does draw attention to the fact that, on the whole, the seven object forms pattern somewhat similarly across the three varieties. The lower panel shows more clearly that for each object form, IndE exhibits the largest deviations (by overusing *ing*), and that the level OBJECTFORM: *no* leads to the least nativelike choices. However and as maybe expected for a variable with much less importance than the previously discussed ones, there seems to be much less systematic patterning here.

For reasons of space, we will discuss the remaining results, which are associated with lower variable importance scores anyway, just summarily (and we will not discuss COUNTRY as a main effect, given that it participates in the above interactions):

i. for FINITEMATRIX : COUNTRY, we find that (i) SingE is closer to NatE than HKE, which in turn is closer to NatE than IndE, and (ii) in HKE and IndE, finite matrix verbs lead to slight overuses of *ing* whereas non-finite ones lead to much more pronounced overuses of *to*.
ii. for VERBTYPECOMP : COUNTRY, the results are very similar to those of VERBTYPEMATRIX : COUNTRY: IndE differs from NatE than the other two Asian Englishes, and particular so in their overuse of *ing* with process verbs.
iii. for VOICEMATRIX : COUNTRY, the Asian English speakers overuse *ing* more with active matrix verbs — with passive ones, they are very close to the predicted NatE choices.
iv. for NEGMATRIX : COUNTRY, no particularly strong pattern emerges, the only potential effect is that, for HKE and IndE speakers, they are more nativelike with negated matrix verbs.

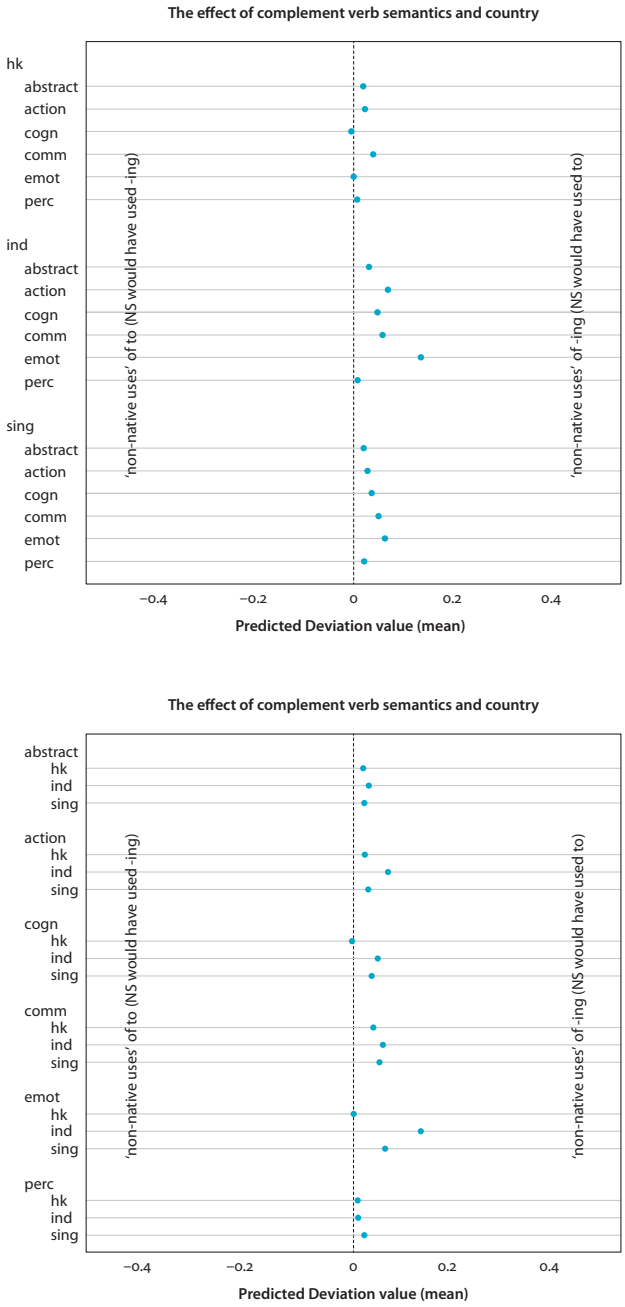

Figure 4. The effect of complement verb semantics and country across HKE, IndE and SingE

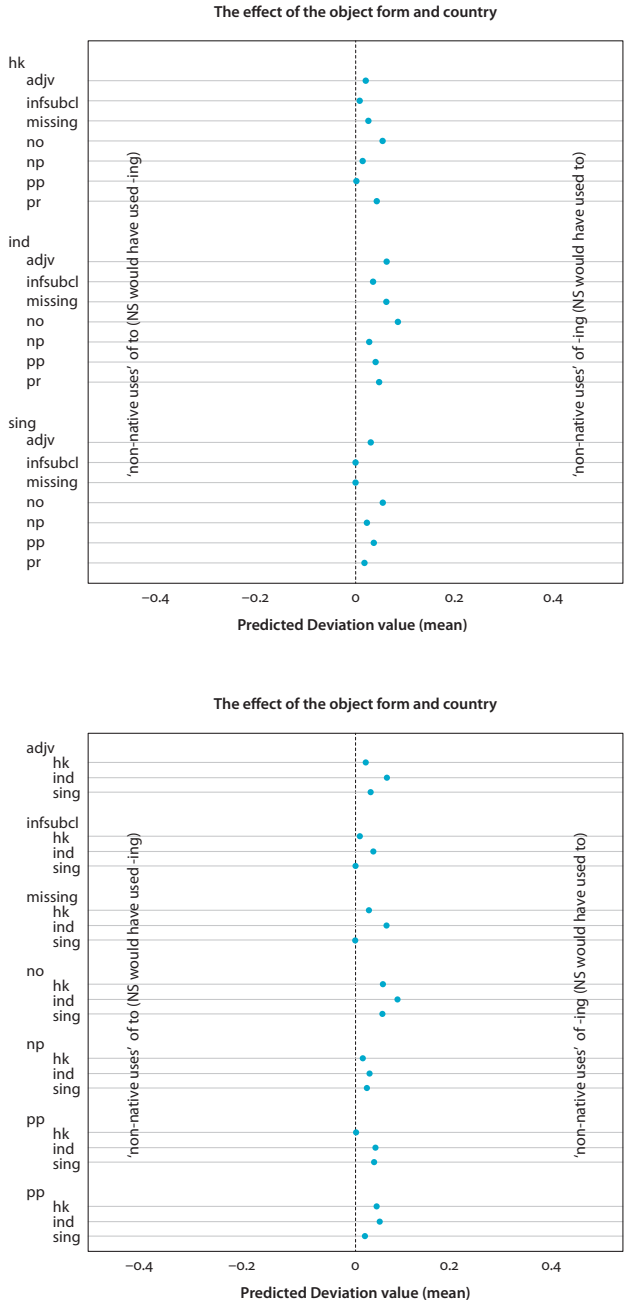

Figure 5. The effect of object form and country across HKE, IndE and SingE

4.4 Excursus: BrE vs. AmE

As we discussed briefly above, including AmE among the native-speaker data is not totally uncontroversial. While we have already mentioned several reasons why we remain convinced that this is a good idea, we are now also in a position to provide empirical data to that effect. In particular, we apply Gries & Bernaisch's (2016) bottom-up strategy to identify epicenters to our present data. We:

i. split up the native-speaker data into the BrE and the AmE data;
ii. ran separate random forests analyses on each of them;
iii. used both the BrE random forests results and the AmE random forests results separately to make predictions for the Asian Englishes data;
iv. then computed how well the BrE predictions and the AmE predictions predicted what the ESL speakers would say and (i) explored that correlation in *C*-values as well as (ii) tested statistically whether the Asian Englishes deviation values from the BrE predictions were significantly different from those of the AmE predictions.

The results were very clear: using BrE to predict the Asian Englishes data led to an extremely poor *C*-value (0.51) whereas using AmE to predict the Asian Englishes data led to a much better one (0.76). In addition, the mean of the absolute deviations from the BrE examples (0.082) are nearly twice as high as those from the AmE examples (0.044), a difference that is statistically significant both in a *U*-test (W=3215000, $p<10^{-10}$) and a Kolmogorov-Smirnov test (D=0.124, $p<10^{-10}$). In other words, it is the American English data that are more similar to the Asian Englishes data. While that does of course not provide evidence with regard to the exact causal patterning, it does provide *prima facie* evidence against discarding any AmE data in an *a priori* fashion.

5. Conclusions

As mentioned above, given the absence of testable hypotheses regarding the *to* vs. *ing* alternation with indigenized variety speakers, this study has a decidedly descriptive slant, which is also compatible with the current random forests analysis (rather than the use of the hypothesis-testing approach of regression modeling). After having provided detailed representations of the results above, we do not reiterate those here, but believe it is justified to summarily state that (i) the set of formal and semantic characteristics we annotated do distinguish reliably between the two complementation constructions and that (ii) the overall method is applicable and yields results with an overall good degree of prediction accuracy.

It is therefore interesting to look at the results from the perspective of how they relate to widely-used theories such as Schneider's (2007) dynamic-evolutionary model. In a nutshell, according to that model, New English (i.e. ESL) varieties follow a uniform pattern of evolution world-wide as "there is a shared underlying process which drives [the formation of postcolonial Englishes], accounts for many similarities between them, and appears to operate whenever a language is transplanted" (Schneider 2007: 29). Crucially, indigenized Englishes evolve following a sequence of five characteristic stages that are associated with linguistic changes and the gradual emergence of locally characteristic linguistic patterns. According to the model, HKE, IndE and SingE all represent different stages in the evolution of postcolonial Englishes: HKE is transitioning between phases II and III (i.e. while the input language still determines language standards and norms in the variety, it is nonetheless becoming an integral part of the local linguistic repertoire), IndE is transitioning between phases III and IV (i.e. as the input language is becoming part of the local linguistic repertoire and the number of competent bilingual L2 speakers increases, the new English variety is beginning to develop accepted local standards), and SingE is at an advanced stage of phase IV (i.e. as the input language may have been retained as a (co-)official language and is used for intra-national contexts, the variety is undergoing in stabilization process).

Bearing this context in mind, much previous alternation research has yielded results such that, the more advanced a stage a variety is at, the more it becomes different from the historical source variety (see Mukherjee & Gries 2009). Interestingly, this is not the case here: on the whole, the stagewise more advanced variety of SingE is more, not less, similar to the native speaker data, conversely, the stagewise less advanced varieties of HKE and IndE are less, not more, similar to the native speaker data (but see also below). While our results do in no way disprove Schneider's (2007) model, they raise the interesting possibility that, as varieties become increasingly emancipated as a whole from some source variety (in previous work, typically BrE), which in many earlier studies has manifested itself in them becoming more different, it seems — and it would be premature to state this any more strongly — this does not always have to be the case. More abstractly, emancipation can, but need not always, result in unidirectional pathways away from the historical source variety. HKE and IndE are different from NatE, but whatever crystallization of patterns emerged for SingE on its way to stage IV did not make SingE even more different from NatE, and it will be interesting to see whether similar pathways can be identified for other phenomena.

That being observed and hypothesized, our result may also have to do with one other relative innovation in this study, the potentially controversial inclusion of AmE into the NatE data. While we believe we have provided three good reasons for why we have done this, this may affect the results such that there would still

just be a unidirectional development away from BrE (even for SingE), but the data do not show that because SingE is also (more?) evolving in the direction of AmE, and indeed our excursus exploration of which of the two native varieties the Asian Englishes are more similar to returned AmE. In other words, this seems to us to be an encouragement for future research to not just look at what ESL varieties are developing away from (BrE), but also where "they are headed", which on the one hand will be shaped by local L1s, but on the other potentially also by contemporary influential cultures (influential on world-wide scale, that is). While, based on our results, it is clearly too soon to claim that ESL varieties have started to undergo a process of Americanization, our results certainly stress the importance of rigorously accounting for the globalization of English and its effect on ESL (Bolton 2008) in our corpus studies. In turn, our results raise the important question of what constitutes an appropriate native-speaker yardstick in the late 20th and early 21st century. Given the currently available data, a more precise analysis contrasting BrE and AmE separately with all the Asian Englishes data awaits much larger data sets, ideally manually-annotated data sets that then also allow for regression modeling again, given the interesting methodological tools that provides for the analyst (in particular, controlling for lexically-specific effects).

Finally, there is a trivial sounding but nonetheless important methodological lesson to be learnt here, namely that sometimes data sets do not permit use of the method(s) that has/have emerged as the state-of-the-art. This has two consequences. First and again seemingly trivially, we need to be familiar (enough) with a range of tools that allow us to squeeze information out of our data sets that are often limited in size and annotation (for obvious reasons having to do with the availability of representative (!) corpora and annotation manpower); in this case, while random forests do not provide all the "machinery" that regression modeling provides, it is more compatible with the descriptive/exploratory approach we adopted, and it also offers us some interesting results and in fact even some advantages (minimizing the risks of overfitting and collinearity, e.g.). Second, we need more authors testing and reporting in their papers whether the methods they used were in fact applicable/appropriate given the data. Observational corpus data are often extremely skewed and often highly collinear, but there are very few papers out there that mention these facts and, more importantly even, how the authors dealt with those threats. For instance, it is temptingly easy to proceed without testing assumptions or to just ignore convergence warnings and high variance inflation factors, but that comes at the cost of the validity of the results, and being able to test for these assumptions and, if so needed, switch to an alternative method, is therefore important for the discipline that is slowly beginning to be more statistical as a whole. We hope that the questions we raise above, and the recommendations

we just made, will lead to an increasingly sophisticated exploration of what it means for (English) varieties to emancipate themselves.

References

Bernaisch, T., Gries, St. Th., & Mukherjee, J. (2014). The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide*, *35*(1), 7–31. doi: 10.1075/eww.35.1.02ber

Bolton, K. (2008). Varieties of World Englishes. In B. B. Kachru, Y. Kachru & C. L. Nelson (Eds.), *The Handbook of World Englishes* (pp. 289–312). Singapore: Wiley-Blackwell.

Bresnan, J., Cueni A, Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bourne, I. Kraemer & J. Zwarts (Eds.), *Cognitive Foundations of Interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Science.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach* (2nd ed.). New York, NY: Springer.

Collins, P. (1995). The indirect object construction in English: An informational approach. *Linguistics*, *33*(1), 35–49. doi: 10.1515/ling.1995.33.1.35

Cuyckens, H., Frauke D., & Szmrecsanyi, B. (2014). Variability in verb complemention in late modern English: Finite vs. non-finite patterns. In M. Hundt (Ed.), *Late Modern English Syntax* (pp. 182–204). Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139507226.014

De Smet, H. (2013). *Spreading Patterns: Diffusional Change in the English System of Complementation*. Oxford: Oxford University Press.

Deshors, S. C. (2014a). Towards an identification of prototypical non-native modal constructions in EFL: A corpus-based approach. *Corpus Linguistics and Linguistic Theory*, *11*(1), 19–50.

Deshors, S. C. (2014b). A case for a unified treatment of EFL and ESL: A multifactorial approach. *English World Wide*, *35*(3), 279–307. doi: 10.1075/eww.35.3.02des

Deshors, S. C. (2015). A constructionist approach to gerundial and infinitival verb-complementation patterns in native and Hong Kong English varieties. *English Text Construction*, *8*(2), 207–235. doi: 10.1075/etc.8.2.04des

Divjak, D. S., & Arppe, A. (2013). Extracting prototypes from exemplars: What can corpus data tell us about concept representation? *Cognitive Linguistics*, *24*(2), 221–274. doi: 10.1515/cog-2013-0008

Duffley, P. (1999). The use of the infinitive and the *-ing* after verbs denoting the beginning, middle and end of an event. *Folio Linguistica*, *23*(3), 295–331.

Edwards, A. (2014). The progressive aspect in the Netherlands and the ESL/EFL continuum. *World Englishes*, *33*(2), 173–94. doi: 10.1111/weng.12080

Green, G. M. (1974). *Semantic and Syntactic Irregularity*. Bloomington, IN: Indiana University Press.

Greenbaum, S. (Ed.) (1996). *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Gries, St. Th. (2003). *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London: Continuum Press.

Gries, St. Th. (2015a). Quantitative linguistics. In J. D. Wright (Ed.), *International Encyclopedia of the Social and Behavioral Sciences* (2nd ed.) (pp. 725–732). Oxford: Elsevier. doi: 10.1016/B978-0-08-097086-8.53037-2

Gries, St. Th. (2015b). The role of quantitative methods in Cognitive Linguistics: Corpus and experimental data on (relative) frequency and contingency of words and constructions. In J. Daems, E. Zenner, K. Heylen, D. Speelman, & H. Cuyckens (Eds.), *Change of Paradigms - New Paradoxes: Recontextualizing Language and Linguistics* (pp. 311–325). Berlin: De Gruyter Mouton.

Gries, St. 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, St. Th., & Bernaisch, T. (2016). Exploring epicenters empirically: Focus on South Asian Englishes. *English World-Wide*, *37*(1), 1–25. doi: 10.1075/eww.37.1.01gri

Gries, St. Th., & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*, *9*(1), 109–136. doi: 10.3366/cor.2014.0053

Gries, St. 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. doi: 10.1075/ijlcr.1.1.05gri

Gries, St. Th., & Wulff, S. (2013). The genitive alternation in Chinese and German ESL learners: Towards a multifactorial notion of context in learner corpus research. *International Journal of Corpus Linguistics*, *18*(3), 327–356. doi: 10.1075/ijcl.18.3.04gri

Hoffmann, T. (2014). The cognitive evolution of Englishes: The role of constructions in the Dynamic Model. In S. Buschfeld, T. Hoffmann, M. Huber & A. Kautzsch (Eds.), *The Evolution of Englishes: The Dynamic Model and Beyond* (pp. 160–180). Amsterdam: John Benjamins. doi: 10.1075/veaw.g49.10hof

Kaleta, A. (2012). The English gerund vs. the to-infinitive: The case of aspectual constructions. Selected papers from *UK-CLA Meetings*. Retrieved from http://www.uk-cla.org.uk/files/proceedings/Kaleta.pdf (last accessed June 2014).

Khamis, A. (2015, July). Cross-varietal variation in English verb complementation: A multivariate corpus analysis. Paper presented at the *International Cognitive Linguistics Conference 2015*, Newcastle upon Tyne.

Koch, C. (2015). Routines in lexis and grammar: A 'gravity' approach within the International Corpus of English. Paper presented at the *ICAME 36 conference*, Universität Trier, 27-29 May 2015.

Kuperman, V., & Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language*, *66*(4), 588–611. doi: 10.1016/j.jml.2012.04.003

Langacker, R. (1991). *Cognitive Grammar 2*. Stanford, CA: Stanford University Press.

Liaw, A., & Wiener, M. (2015). randomForest. Version 4.6-12. A package for R. Retrieved from https://cran.r-project.org/web/packages/randomForest/index.html (last accessed February 2016).

Mair, C. (2002). Three changing patterns of verb complementation in Late Modern English: A real-time study based on matching text corpora. *English Language and Linguistics*, *6*(1), 105–131. doi: 10.1017/S1360674302001065

Mair, C. (2013). The world system of English. *English World-Wide*, *34*(3), 253–278. doi: 10.1075/eww.34.3.01mai

Martínez-García, M. T., & Wulff, S. (2012). Not wrong, yet not quite right: Spanish ESL students' use of gerundial and infinitival complementation. *International Journal of Applied Linguistics*, *22*(2), 225–244. doi: 10.1111/j.1473-4192.2012.00310.x

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading*, *20*(1), 20–33. doi: 10.1080/10888438.2015.1107073

Mindt, D. (2000). *An Empirical Grammar of the English Verb*. Berlin: Cornelsen.

Mukherjee, J., & Gries, St. Th. (2009). Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide*, *30*(1), 27–51. doi: 10.1075/eww.30.1.03muk

Mukherjee, J., & Hoffmann, S. (2006). Describing verb-complementational profiles of New Englishes: A pilot study of Indian English. *English World-Wide*, *27*(2), 147–173. doi: 10.1075/eww.27.2.03muk

Mukherjee, J., & Schilk, M. (2008). Verb-complementational profiles across varieties of English: Comparing verb classes in Indian English and British English. In T. Nevalainen, I Taavitsainen, P. Pahta & M. Korhonen (Eds.), *The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present* (pp. 163–181). Amsterdam: John Benjamins. doi: 10.1075/silv.2.14muk

Nam, C., Mukherjee, S., Schilk, M., & Mukherjee, J. (2013). Statistical analysis of varieties of English. *Journal of the Royal Statistical Society*, *176*(3), 777–793. doi: 10.1111/j.1467-985X.2012.01062.x

Noël, D. (2003). Is there semantics in all syntax? The case of accusative and infinitive constructions vs. that-clauses. In G. Rohdenburg & B. Mondorf (Eds.), *Language Typology and Syntactic Description: Vol.2, Complex Constructions* (pp. 52–150). Cambridge: Cambridge University Press.

Noonan, M. (1985). Complementation. In T. Shopen (Ed.), *Language Typology and Syntactic Description. Vol. 2. Complex Constructions* (pp. 42–110). Cambridge: Cambridge University Press.

Olavarría de Ersson, E., & Shaw, P. (2003)). Verb complementation patterns in Indian standard English. *English World-Wide*, *24*(2), 137–161. doi: 10.1075/eww.24.2.02ers

R Development Core Team. 2012. R: A Language and Environment for Statistical Computing. Foundation for Statistical Computing. Vienna, Austria. <http://R-project.org> (last accessed July 2012)

Ransom, E. (1979). Definiteness and animacy constraints on passives and double object constructions in English. *Glossa*, *13*(2), 215–240.

Rohdenburg, G. (1995). On the replacement of finite complement clauses by infinitives in English. *English Studies*, *76*(4), 367–388. doi: 10.1080/00138389508598980

Schilk, M., Bernaisch, T., & Mukherjee, J. (2012). Mapping unity and diversity in South Asian English lexicogrammar. In M. Hundt & U. Gut (Eds.), *Mapping Unity and Diversity*

World-wide: Corpus-based Studies of New Englishes (pp. 137–166). Amsterdam: John Benjamins. doi: 10.1075/veaw.g43.06sch

Schilk, M., Mukherjee, J., Nam, C., & Mukherjee, S. (2013). Complementation of ditransitive verbs in south Asian Englishes: A multifactorial analysis. *Corpus Linguistics and Linguistic Theory*, *9*(2), 187–225. doi: 10.1515/cllt-2013-0001

Shastri, S. V. (1996). Using computer corpora in the description of language with special reference to complementation in Indian English. In R. J. Baumgardner (Ed.), *South Asian English: Structure, Use, and Users* (pp. 70–81). Urbana & Chicago, IL: University of Illinois Press.

Schneider, E. (2007). *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511618901

Smith, M. B., & Escobedo, J. (2002). The semantics of *to*-infinitival vs. *-ing* verb complement constructions in English. In M. Andronis, C. Ball, H. Elston & S. Neuvel (Eds.), *Proceedings from the Main Session in the Chicago Linguistics Society's Thirty-Seventh Meeting* (pp. 549–564). Chicago, IL: Chicago Linguistic Society.

Szmrecsanyi, B., & Kortmann, B. (2011). Typological profiling: Learner Englishes versus L2 varieties of English. In J. Mukherjee & M. Hundt (Eds.), *Exploring Second-language Varieties of English and Learner Englishes: Bridging the Paradigm Gap* (pp. 167–207). Amsterdam: John Benjamins. doi: 10.1075/scl.44.09kor

Vendler, Z. (1957). Verbs and times. *Linguistics in Philosophy*, *66*(2). 143–160.

Vosberg, U. (2003). The role of extractions and horror aequi in the evolution of *-ing* complements in modern English. In G. Rohdenburg & B. Mondorf (Eds.), *Determinants of Grammatical Variation in English* (pp. 329–345). Berlin: Mouton de Gruyter.

Wierzbicka, A. (1988). *The Semantics of Grammar*. Amsterdam: John Benjamins. doi: 10.1075/slcs.18

Wulff, S., & Gries, St. Th. (2015). Prenominal adjective order preferences in Chinese and German L2 English: A multifactorial corpus study. *Linguistic Approaches to Bilingualism*, *5*(1), 122–150. doi: 10.1075/lab.5.1.05wul

Authors' addresses

Sandra C. Deshors
Department of Languages and Linguistics
New Mexico State University
MSC 3L
Las Cruces, NM 88003
United States of America

deshors@nmsu.edu

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
Santa Barbara, CA 93106–3100
United States of America

stgries@linguistics.ucsb.edu