DE GRUYTER OPEN

ICAME Journal, Volume 41, 2017, DOI: 10.1515/icame-2017-0002

Empirical perspectives on two potential epicenters: The genitive alternation in Asian Englishes

Benedikt Heller (KU Leuven)
Tobias Bernaisch (Justus Liebig University Giessen)
Stefan Th. Gries (University of California Santa Barbara)

Abstract

The present study seeks to contribute to two sparsely examined areas of World Englishes research by (i) quantitatively evaluating two potential linguistic epicenters in Asia (Indian and Singapore English) while (ii) investigating the English genitive alternation in a cross-varietal perspective. In a corpus-based bottom-up approach, we evaluate 4,200 interchangeable genitive cases of written English from Great Britain, Hong Kong, India, the Philippines, Singapore and Sri Lanka, as represented in the International Corpus of English. We use a new method called MuPDARF, a multifactorial deviation analysis based on random forest classifications, to evaluate to what extent and with which factors the Asian varieties differ from British English in their genitive choices. Results show conspicuous differences between British English and the Asian varieties and validate the potential epicenter status of Indian English for South Asia, but not unanimously that of Singapore English for Southeast Asia.

1 Introduction

1.1 Linguistic epicenters in Asia and their exploration

In his account of Sri Lankan English (SLE) lexis, Meyler (cf. 2007: xiv) describes vocabulary-related similarities across national variety boundaries in South Asian Englishes. Examples of these pan-South Asian English lexemes, which can also be shown to occur significantly more frequently in South Asian Englishes than in BrE (cf. Bernaisch 2015: 110), include *rupee* (cf. Meyler 2007: 225) denoting the respective currencies in India, Nepal, Pakistan and Sri Lanka or *tank* as a reference to "an artificial lake or reservoir" (Meyler 2007: 254).

Empirical evidence also suggests that the lexis-grammar interface displays characteristics absent from BrE, but shared across several South Asian Englishes. Example (1), taken from the South Asian Varieties of English (SAVE) Corpus (cf. Bernaisch *et al*. 2011), shows the verb SUBMIT in a double-object construction, i.e. a lexicogrammatical configuration that cannot be found in comparable BrE data (cf. Schilk *et al*. 2012: 153f.; Koch and Bernaisch 2013: 78).

(1) [...] the North 24-Parganas SP to submit him a detailed report on the incident. (SAVE-IND-SM_2004-01-21)

Mukherjee and Hoffmann (2006: 157) refer to these newly developed constructions with the label "new ditransitives", which are also productive in South Asian Englishes with other verbs such as EXTEND, RETURN or ADVISE (cf. Koch and Bernaisch 2013: 78). Another pan-South Asian lexicogrammatical innovation is the presentational focus marker *itself* used to pragmatically focus the entity preceding it as in (2). More precisely, the use of *itself* in (2) puts (restrictive) focus, i.e., "that string of expressions which is set off from the rest of the sentence by prosodic prominence and which is specifically affected semantically by the particle" (König 1993: 979), on *today* and thus excludes from consideration other potential – and in the example implicit – alternative points in time when the Supreme Court's ruling is expected. Presentational focus marking with *itself* is likely the product of processes of L1 transfer into the second-language varieties used across the Indian subcontinent (cf. Bernaisch and Lange 2012). Similarly, Lange (2016: 133) finds what she coins the "intrusive *as*" construction, i.e. the use of *as* in complex-transitive constructions, exemplified in (3) in the entire South Asian *Sprachraum*.

(2) He was expecting Supreme Court's ruling on this writ application today itself. (SAVE-NEP-NT_2003-10-24; quoted from Bernaisch and Lange (2012: 8))

(3) [...] we come across a teacher of literature named as Mr. Keating, acted by Robin Williams. (SAVE-SL-DN_2002–05–07; quoted from Lange (2016: 140))

Against this background of pan-South Asian lexical and lexicogrammatical features differentiating South Asian Englishes from their historical input variety BrE, the obvious question is how to account for these parallel cross-national developments in generally distinct South Asian Englishes. Variety-specific independent developments – particularly since South Asia represents a prototypical

instance of a *Sprachbund* where millennium-long contact between local Indo-Aryan and Dravidian languages has triggered structural convergence between them (cf., e.g., Emeneau 1956; Masica 1976) – can certainly serve as explanations of features transferred into English from usage patterns in the respective first languages. In this vein, languages of the Indo-Aryan and the Dravidian language families share, e.g., empathic clitics attachable to sentence elements in order to put pragmatic focus on them (cf. Bernaisch and Lange 2012: 5), which may serve as a template for the usage of presentational *itself* in South Asian Englishes exemplified in (2).

An alternative explanation for the pan-South Asian features delineated above – one that is also compatible with processes of structural transfer from local L1s – is based on the notion of linguistic epicenters. In this account, regional centers functioning as lead varieties in the respective loci structurally influence varieties in their surroundings, triggering postcolonial Englishes squared (cf. Bernaisch and Lange 2012: 13) since for these varieties, another postcolonial English – and no longer the historical input variety BrE – serves as an actuator for (new) developmental cycles.

Hundt (cf. 2013: 185) identifies two central characteristics of linguistic epicenters: they a) are endonormatively stabilized in Schneider's (2003, 2007) terminology (cf. also Peters 2009: 108) and b) fulfill a model function for other varieties. It is to be expected that – possibly prior to a global reach – this modeling effect is most prevalent with other regional varieties in physical proximity of the epicenter since "the waves emanating from an earthquake epicenter have a more or less immediate (and damaging) effect on the adjacent surroundings" (Hundt 2013: 189).

This regional spread of linguistic structures can, for example, be achieved "through face-to-face contact of speakers or because it provides textbook material for teaching English as a second language in a neighbouring country" (Hundt 2013: 189). In the light of this, the setting up of an Indian English (IndE) language teaching institution, the *Sri Lanka-India Centre of English Language Training* (SLICELT) in Peradeniya, Sri Lanka (cf. Lim and Ansaldo 2015: 180), is indicative of the relevance of IndE for Sri Lankan English (SLE) and provides a potential avenue for epicentral influence in South Asia to manifest itself.

Generally, epicenters can be found and are currently developing in specific sociolinguistic constellations all around the globe. In contradistinction to old (e.g., American English) and potentially new epicenters (e.g., Australian or New Zealand English), Hundt (2013: 186) also elaborates on "'emerging' epicenters [...] that have developed their own endo-normativity (e.g., IndE and SinE) but whose status as a potential (local) norm-providing centre has only recently

attracted linguists' attention". While the definition of an emerging epicenter should probably be based on factors other than the immediacy of linguists' interest in them, Hundt (cf. 2013: 186) as well as Leitner (cf. 1992: 225) explicitly assign IndE and Singapore English (SinE) a potential epicentral status in South and Southeast Asian Englishes respectively. The evolutionary cycles the individual varieties in South and Southeast Asia have already completed certainly warrant this perspective.

With a focus on South Asia, IndE can be characterized as an endonormatively stabilized variety, having established and now following localized norms for English language use. These norms find structural reflection in recurrent frequency-related as well as categorical differences to British English (BrE) on the level of phonology (cf., e.g., Fuchs 2016), lexis and morphosyntax (cf., e.g., Sedlatschek 2009) as well as lexicogrammar (cf., e.g., Mukherjee 2007; Schilk 2011; Lange 2012). SLE is another South Asian variety that can unambiguously be profiled as an endonormatively stabilized variety based on its distinct sound system (cf., e.g., Senaratne 2009) as well as its variety-specific lexical and lexicogrammatical structural profile (cf., e.g., Gunesekera 2005; Bernaisch 2015). Consequently, both IndE and SLE are – based on their evolutionary status – candidates for linguistic epicenters of South Asian Englishes, while the remaining varieties in South Asia, i.e. Bangladeshi, Maldivian, Nepali and Pakistani English, cannot be considered equally evolved.

In Southeast Asia, SinE can probably be regarded as the most advanced variety in terms of its evolutionary development.

> By now Singapore has clearly reached phase 4 of the cycle. The country's unique, territory-based, and multicultural identity construction has paved the way for a general acceptance of the local way of speaking English as a symbolic expression of the pride of Singaporeans in their nation. (Schneider 2007: 160)

As the remaining Southeast Asian Englishes do not yield another similarly advanced postcolonial English – e.g., Hong Kong (HKE, cf. Schneider 2007: 138) and Philippine English (PhilE, cf. Schneider 2007: 143) show more traits of phase 3 than of phase 4 varieties – SinE appears the only viable Southeast Asian epicentral candidate.

With the help of corpus data, the present paper simultaneously studies potential epicentral configurations in South and Southeast Asian Englishes with IndE or SLE and SinE as the respective epicentral candidates.

Finally, Hundt stresses the necessity to complement corpus-based evidence with attitudinal considerations to assess "whether speakers consciously aspire to

a particular variety of English and thus adopt certain features from it" (2013: 184). For South Asia, it has been shown that Indian as well as Sri Lankan speakers of English give the least positive evaluation to their respective neighbor variety (cf. Bernaisch and Koch 2016), which may present an attitudinal hurdle to epicentral influences. Still, the relevance of attitudes for epicentral influence to take effect seems closely connected to the degree of speaker awareness of the feature that may be spread in an epicentral fashion. True, there are some easily noticeable features of IndE as listed (and criticized in the tradition of outer-circle linguistic complaints) in Sanyal *et al*. (2006; e.g., the insertion of *good* in *What is your good name?*) with which attitudes may exhibit a strong influence on their adoption. However, with more subtle processes of structural nativization yielding quantitative (instead of overtly marked categorical) differences flying under the radar of speakers' linguistic awareness, attitudinal considerations must be assumed to take a back seat.

A corpus-based bottom-up examination of such a probabilistic object of investigation in South Asian Englishes – in this case the dative alternation, i.e. the alternation between the double-object construction as in *John gave Mary a book* and the prepositional dative as in *John gave a book to Mary* – highlighted that IndE provided the best model for the factor-guided mostly subconscious selection process of either of the two constructions in the remaining five South Asian Englishes (cf. Gries and Bernaisch 2016). Still, given that claims of epicentral influence should be substantiated with multiple examples (cf. Peters 2009: 109), this study investigates the genitive alternation.

1.2 The English genitive alternation

The English genitive alternation is defined as the choice between the *s*-genitive as exemplified in (4) and the *of*-genitive as exemplified in (5). In the *s*-form, the possessor precedes the possessum, whereas the possessor follows the possessum in the *of*-form.

(4) The additional commerce secretary […] has called for increasing [the country]$_{\text{possessor}}$'s [share]$_{\text{possessum}}$ in the world jewellery trade. (ICE-IND:W2C-001)

(5) a person who goes abroad for higher studies should come back […] and serve the [people]$_{\text{possessum}}$ of [his own country]$_{\text{possessor}}$. (ICE-SL:W1A-011)

This study takes a variationist perspective, assuming that the two variants constitute two "ways of saying 'the same' thing" (Labov 1972: 188). Since geni-

tives are used to express a multitude of functions (e.g., Quirk *et al*. 1985; Biber *et al*. 1999), the assumption of sameness requires a solid definition of what constitutes the variable context (see Rosenbach 2002: 22–25 for a discussion of sameness in the English genitive alternation; also compare Szmrecsanyi *et al*. 2016 who model genitive choice as ternary alternation).

In order to obtain all genitives that are variable (or interchangeable), we subtracted from the entirety of genitive constructions all occurrences that are not variable (i.e. categorical; see Rosenbach 2002: 27–28). The proper name *Devil's Backbone* in (6), for example, cannot be expressed in the *of*-form as **the Backbone of the Devil*, and the partitive genitive in (7) cannot be paraphrased as **the suspect's one*. Examples (4) and (5), on the other hand, are interchangeable. The *s*-genitive in (4) could potentially be expressed as *the share of the country*, and the *of*-genitive in (5) could be expressed as *his own country's people*.

(6) A common and attractive foliage plant, the Devil's Backbone is so named because of its zig-zag stem. (ICE-SIN:W2B-021)

(7) One of the suspects was apprehended […]. (ICE-PHI:S2B-001)

(8) the USSR rejected a draft of the union treaty (ICE-SIN:S1B-041)

Apart from proper names (6) and partitive genitives (7), we excluded the following cases from our analysis: double genitives (e.g., *a friend of Mary's*), fixed expressions (e.g., *Valentine's Day*), descriptive genitives (e.g., *people of color*), and appositive genitives (e.g., *the city of New York*). A further requirement was that all possessums be definite in order for the genitive to be interchangeable. Since the *s*-genitive occupies the same grammatical slot as the definite article, it has a deterministic function, which only allows definite possessums in *s*-genitives. *Of*-genitives, therefore, in order to be interchangeable, were required to contain a definite possessum. Consider the example in (8), which contains an indefinite possessum. The alternation *the union treaty's draft* expresses a different meaning since it indicates that there is just one single draft, whereas the *of*-genitive refers to a broader range of either one or many.

The genitive alternation is very well researched; an exhaustive overview can be found in Rosenbach (2014). A lot of this research has been dedicated to investigating historical changes in the distribution of the *s*- and the *of*-genitive (e.g., Thomas 1931; Rosenbach and Vezzosi 2000; Wolk *et al*. 2013) and its diachronic trajectory has been described in meticulous detail. While the *of*-form was hardly ever used in Old English, it took over almost completely in Middle English times. In the following Early Modern English period, the *s*-genitive experienced a renaissance, which was unexpected since it went against the gen-

eral trend of the language's development from a synthetic case system towards a more analytic one. Rosenbach (2002: 189) argues that the re-emergence of the *s*-genitive does not contradict this general trend because the *s*-genitive was not inflectional in nature, but rather a clitic.

Next to the historical perspective, a lot of work has been dedicated to describing how a multitude of predictors influence the choice between the two variants. Animacy of the possessor, syntactic weight of the constituents and topicality (see below) commonly range among the most important factors (see Gries 2002 for a study pitting the three against each other). More recent studies of the genitive alternation also make use of state-of-the-art statistical techniques such as logistic regression (e.g., Hinrichs and Szmrecsanyi 2007) and mixed-effects modeling (e.g., Wolk *et al*. 2013).

However, in spite of this impressive body of research, comparatively little attention has been paid to cross-varietal differences in the genitive alternation. Most of the studies that looked at genitives in different varieties so far focused on the distinction between British and American English (e.g., Jahr Sohrheim 1980; Rosenbach 2002, 2005; Hinrichs and Szmrecsanyi 2007; Szmrecsanyi and Hinrichs 2008: Szmrecsanyi 2010). One exception is Hundt and Szmrecsanyi (2012), who looked at differences between early New Zealand English and early British English; they found only marginal effects of variety, but uncovered interesting interactions, e.g., between the factors variety and possessor animacy. According to their findings, writers of early New Zealand English were significantly less likely to use *s*-genitives unless the possessor was animate (Hundt and Szmrecsanyi 2012: 252). A more recent study, Szmrecsanyi *et al*. (2016), looked at the genitive alternation alongside the dative and the particle placement alternation across four varieties of English (Canadian, British, Indian and Singaporean English) and found, among other things, a higher relative importance of variety compared to other predictors such as genre and thematicity.

1.3 Overview of the present paper

In the course of this paper, we consequently set out to answer two central research questions. First, are there identifiable effects of structural nativization in the genitive alternation in the South and Southeast Asian varieties under scrutiny? Second, if so, do the variety-specific models of the structural norms for the genitive alternation substantiate a potential influence of IndE on varieties in South Asia and of SinE on varieties in Southeast Asia?

2 Methods

In this section, we will discuss the corpus data, their annotation, and the two statistical analyses that we undertook. Section 3 will then discuss the results.

2.1 Data and their annotation

The present investigation is based on a sample taken from six components of the International Corpus of English (ICE) and amounts to a total of 4,200 cases of interchangeable genitives. Every ICE component consists of one million words, and is divided into 600,000 words of spoken English and 400,000 words of written English (Greenbaum 1996). The corpora consist of a variety of balanced subgenres, which allows for valid comparisons among varieties (Nelson 1996). Since for the Sri Lankan component only the written part was available, we restricted our analysis to the written parts of all ICE components. The following written components were analyzed: ICE-Great Britain (ICE-GB) as reference variety, ICE-India (ICE-IND) and ICE-Sri Lanka (ICE-SL) representing South Asia, and ICE-Singapore (ICE-SIN), ICE-Philippines (ICE-PHI) and ICE-Hong Kong (ICE-HK) representing Southeast Asia.

The first step of the data extraction process was the isolation of all interchangeable genitive instances. First, a Perl script was used to extract all occurrences of the genitive markers *'s*, *s'*, and *of* automatically. Secondly, categorical instances were filtered by applying lexical, part-of-speech-related, and grammatical exclusion criteria. Interchangeability decisions were then manually checked to ensure maximum precision and recall. The resulting list of interchangeable genitives was then annotated with multiple language-internal as well as language-external factors. In the course of previous research on the English genitive alternation, scholars have identified a multitude of predictors that significantly influence the choice between the *s*- and the *of*-genitive. In this study, all cases were annotated for animacy and length of the constituents, givenness, thematicity and final sibilancy of the possessor, lexical density of the immediate context, variety and genre. Further, we added a factor that captures the overall frequency of the possessor.

Many studies found animacy of the possessor to be the most important factor determining genitive choice (e.g., Hundt and Szmrecsanyi 2012; Wolk *et al*. 2013). Even though possessor animacy usually explains a substantial part of variance in the genitive alternation and the direction of the effect is undisputed, *s*-genitives do not exclusively occur with animate possessors, e.g., example (4). Diachronic research shows that since the mid-1800s approximately, *s*-genitives have increasingly been used with collective, locative, and temporal possessors

(Wolk *et al*. 2013). In order to assess if this spread affected different varieties equally, we employed an animacy classification following guidelines in Wolk *et al*. (2013) with an additional distinction between humans and animals: human (e.g., *woman*), animal (e.g., *dog*), collective (e.g., *team*), inanimate (e.g., *book*), locative (e.g., *India*), and temporal (e.g., *today*). Animacy of possessor (POR_ANIMACY) and possessum (PUM_ANIMACY) were annotated semi-automatically. First, a Perl script checked the Germanic possessive *-s* database[1] and the WordNet database[2] and transformed the annotations in these databases into the six-fold classification introduced above (levels: *a* for human, *a2* for animal, *c* for collective, *i* for inanimate, *l* for locative, and *t* for temporal). After that, all instances were corrected manually.

Another prominent factor that has a crucial influence on genitive choice is syntactic weight, operationalized here as the length of possessor (POR_ LENGTH_WORDS) and possessum phrases (PUM_LENGTH_ WORDS). It is common ground among researchers that this factor affects genitive choice along the lines of what Behaghel (1909) described as "Das Gesetz der wachsenden Glieder" (the principle of end-weight), the tendency of long constituents to be placed at the end of an utterance. In case of the genitive alternation this means that if the possessor is relatively long, the construction is less likely to be realized as an *s*-genitive since the possessor is the first element in an *s*-genitive (example 9); if the possessum, on the other hand, is relatively long, chances for an *s*-genitive realization are higher (example 10). The influence of end-weight on the genitive alternation has been confirmed in many studies (e.g., Altenberg 1982; Hinrichs and Szmrecsanyi 2007; Ehret *et al*. 2014). POR_LENGTH_ WORDS and PUM_LENGTH_WORDS were both annotated using a function that returns the number of characters in the respective phrase. The function counted all letters and digits, but did not count special characters like spaces and hyphens. ORUM_LENDIFF_LOG captures the difference between possessor length and possessum length, which indicates which one of the two constituents is longer and by how much.

(9) In fact, one of the first military acts after the [occupancy]$_{\text{possessum}}$ of [a town or village]$_{\text{possessor}}$ was the establishment of the public school (ICE-PHI:W2A-001)

(10) [Japan]$_{\text{possessor}}$'s [biggest mobile-phone operator]$_{\text{possessum}}$ NTT DoCoMo offers similar services (ICE-HK:W2B-031)

Also of importance for the genitive alternation is the discourse accessibility of the constituents, which is used as an umbrella term for the factors givenness,

thematicity, and overall frequency here. If a possessor has been mentioned in the previous context of a genitive construction (i.e. it is *given*), it is more likely to be expressed in the *s*-form than if the possessor is discourse-new (e.g., Hinrichs and Szmrecsanyi 2007; Jankowski 2013; Grafmiller 2014). Besides givenness, the degree to which a possessor constitutes a central topic of a text – its thematicity – also makes a difference. In a physics textbook, for example, where *laser* is the central topic, the *s*-genitive *the laser's light power* is more likely than in other texts (Osselton 1988). Since there is no evidence in previous research that givenness and thematicity of the possessum play any role in the alternation, we restricted our attention to the possessor. The corpus file in which a genitive occurs was searched to determine the givenness (POR_GIVENNESS) and thematicity of the possessor concerned (POR_THEMATICITY). Givenness was automatically annotated for by checking the previous context of the genitive instance in the respective corpus file; it was set to *given* instead of *new* if the lemma[3] had been mentioned before. For thematicity not only the previous context but the whole corpus text (excluding the genitive instance in question) was searched and thematicity was determined by counting the frequency of the possessor head lemma. In addition, another measure of accessibility was included in this study: overall frequency. This factor represents the rate with which the constituent heads are used in English overall, a factor which has been shown to influence other syntactic choices (Gahl and Garnsey 2004; Hilpert 2008). To annotate the overall frequency of the constituent heads of possessors (POR_HEAD_FREQ) and possessums (PUM_HEAD_FREQ), we investigated the heads' frequencies in the respective components of the GloWbE corpus (Davies and Fuchs 2015), a corpus of 1.9 billion words of English online. For the genitives in ICE-GB, for example, a script checked the frequencies in the British component of GloWbE, and the same approach was used for the other varieties. ORUM_FREQDIFF_LOG, similar to the difference in length, contains the difference in overall frequency.

The factor final sibilancy is a phonological variable. If the sound at the end of the possessor phrase, which is immediately followed by *'s* in *s*-genitives, is a final sibilant (i.e. [s], [z],[ʃ], [tʃ], [dʒ], or [ʒ]), this produces a sound sequence that is harder to pronounce (Zwicky 1987). Therefore, genitives like *The Mercedes's headlights* are more often paraphrased as *the headlights of the Mercedes* than are cases without final sibilancy of the possessor. For the annotation of POR_FINAL_SIBILANCY, a script checked the phonetic transcription of the possessor-final words in the CMU Pronunciation Dictionary[4]. If the last phoneme was found to be a sibilant, the annotation was set to *true*, otherwise to

false. If a word could not be found in the dictionary, the script relied on the word's orthography to determine the presence of a final sibilant.

Finally, the lexical density of a text has also been shown to favor *s*-genitives (Hinrichs and Szmrecsanyi 2007; Szmrecsanyi and Hinrichs 2008; Szmrecsanyi 2010; Jankowski 2013). Lexically dense environments usually favor *s*-genitives because they are more compact. It is operationalized here as type-token ratio (TTR). Since the measure is highly sensitive to text size, TTR was calculated for the immediate context of 100 words of each genitive instance; usually, 50 words of previous context and 50 words of following context were considered. If the genitive in question was close to the beginning of a corpus text and, therefore, there were less than 50 words of previous context available, the script considered additional words from the following context until a total word count of 100 was reached; if instances were located toward the end of a text, it considered more words from the previous context.

To indicate the ICE component from which each case originates, a column VARIETY was added; its levels are *gb*, *sin*, *phi*, *hk*, *ind*, and *sl*. Further, GENRE_COARSE lists the written subgenre of every genitive (*printed* and *non-printed*). VARIANT contains the genitive choice of each case (*s* or *of*).

2.2 Statistical evaluation, part 1: Quantifying distance in genitive choice between BrE and the South (East) Asian varieties

For many years, much of the research on lexical or morphosyntactic as well as lexicogrammatical alternations in learner corpus research (LCR) and corpus-based indigenized-variety (IV) research has been based on the counting of features in the data of a reference variety – often native speakers in LCR and BrE speakers in IV research – and comparing these frequencies to the corresponding frequencies in target varieties – learner data in LCR and indigenized variety data in IV research. However, by now it is well known that such approaches are often insufficient in that they do not control for a potentially vast number of features that co-determine a particular word/construction's use both by speakers of the target and reference variety. Thus, where sufficient research on the factors influencing the variable under scrutiny is available, the field has moved on in the direction of multifactorial regression modeling so as to be able to take a larger number of determinants of variation into consideration statistically. In the best of such approaches, multiple predictors of a particular phenomenon would be taken into consideration as well as (i) a predictor coding the L1/native language of the speaker and (ii) crucially, minimally all pairwise interactions between all linguistic predictors and the L1/native language predictor since it is only these

interactions that reveal whether certain linguistic/contextual determinants differ across L1s/varieties/etc. in their effect on the speaker choice.

In this paper, we are using and extending a new approach of this regression-based modeling. The new approach we are referring to is called MuPDAR (for Multifactorial Prediction and Deviation Analysis with Regressions). This approach aims at facilitating comparisons between one or more reference varieties and one or more target varieties by answering the following question: "In the situation that the target variety speaker is in now, what would the reference variety speaker have done, and if the two choices differ, how so and why?" This method has been developed in Gries and Adelman (2014), Gries and Deshors (2014) and Wulff and Gries (2015) for subject realization in Japanese, *may* vs. *can* by native speakers and French/Chinese learners of BrE, and prenominal adjective order by native speakers and German/Chinese learners of BrE respectively. MuPDAR involves the following three steps:

- fit a regression R_1 that predicts the choices that speakers of the target/reference level (typically, native speakers of the reference variety) make with regard to the phenomenon in question;
- apply the coefficients resulting from R_1 to the other speakers in the data (typically, learners or speakers of institutionalized second-language varieties) to predict for each of their data points what the native speaker of the reference variety would have done in their place;
- fit a regression R_2 that explores how the other speakers' choices differ from those of the speakers of the target/reference variety.

Following the above work, Gries and Bernaisch (2016) were the first to apply this approach, which was initially only applied in LCR settings, to IV research by looking at how the dative alternation with GIVE is used differently in BrE compared to six indigenized varieties. In the current study, we also apply the general logic of this approach to our genitive data, but, following Deshors and Gries (2016), use two random forests instead. Random forests is an approach that is similar to classification (and regression) 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 or some other quality criterion (deviance, *Gini*, …) regarding the dependent variable within the groups. This process is recursive, i.e. repeated until no further split would improve the quality criterion 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 regis-

ter and (ii) protecting against overfitting. On the one hand, the algorithm constructs many different trees (we set that parameter to 2000), each of which is fitted to a different bootstrapped sample of the full data. On the other hand, each split in each tree chooses from only a randomly-chosen subset of predictors (we set that parameter to what in our case amounts to the default of three predictors). The overall result is then based on amalgamating all 2000 trees that have been generated.

It is useful to briefly comment on what the change from regressions to random forests entails. On the one hand, random forests do not provide all the 'machinery' and results that a regression analysis can provide in the best of cases. For instance, Gries and Bernaisch's (2016) regression analyses are multilevel/mixed-effects models that are instructive in how they can include random effects such as the hierarchical structure of the corpus. In addition, some kinds of effects are probably easier to explore with regressions than with random forests – e.g., slopes of polynomial to a degree of 2 or greater. That being said, the kind of data that corpus linguists often work with do not always allow for the use of these powerful regression models, given that these data often involve heavily skewed (Zipfian) frequency distributions (i.e., sparse data for many combinations of predictors and/or random-effects levels), predictors that are multicollinear, etc., and in fact our attempts to run the relevant regression models on the present data set were unsuccessful (exceeding the default tolerance threshold by a factor of 20 in R_1 and lack of convergence in R_2). Finally, in the absence of careful cross-validation, regression models are often likely to suffer from overfitting, i.e. fitting very well the particular training data set but performing much more poorly on different test data sets.

Random forests, on the other hand, are fast and easy to generate, they usually achieve very good prediction accuracies, they do not make distributional assumptions of the kind that regression models usually make, the sampling components make them much less likely to overfit, cross-validation is inbuilt into the algorithm, and they are good at handling many-predictors-few-datapoints problems – the main downside is how to understand and potentially visualize the effects predictors have on the dependent variable, given how the final results are based on the aggregation of in this case 2000 separate classification trees (using different predictors and different data points). Given the inapplicability of regression modeling to our data, we are here using random forests in place of regressions; in particular, we are using the implementation in the R package randomForest (Liaw and Wiener 2015, version 4.6-12). As for the interpretation of the results, we are following Bernaisch, Gries and Mukherjee (2014) and Deshors and Gries (2016): we compute predicted probabilities for all cases and

then report averages for each combination of the predictor VARIETY and each other predictor. While this is a heuristic in how the resulting plots do not control for the effects of all other predictors at the same time, the above studies have used this successfully and comparisons of such plots with corresponding effects plots of regressions have been very encouraging.

In sum, we

- do a random forests analysis on only the native BrE data and test whether its fit is good enough to proceed; this analysis uses VARIANT as the dependent variable and the following as predictors: PUM_HEAD_FREQ, PUM_LENGTH_WORDS, PUM_ANIMACY, POR_HEAD_FREQ, POR_LENGTH_WORDS, POR_ANIMACY, POR_FINAL_SIBILANCY, POR_GIVENNESS, POR_THEMATICITY, ORUM_FREQDIFF_LOG, ORUM_LENDIFF_LOG, GENRE_COARSE, and TTR;
- if the fit is good enough, we apply the results from the first random forests analysis to the HKE, IndE, PhilE, SinE, and SLE speakers to obtain predictions of what native speakers would have said in the contexts that the IV speakers were in;
- compare the BrE predictions against the IV choices to see how much the two coincide; for that we compute a numeric variable called DEVIATION,[5] which is
  - set to zero when the IV speaker made the choice a BrE speaker is predicted to have made;
  - between -0.5 and 0 when the IV speaker chose *of* although the BrE speaker is predicted to have chosen *s*;
- between 0 and 0.5 when the IV speaker chose *s* although the BrE speaker is predicted to have chosen *of*.

  (The exact value depends on how strongly the native speaker was predicted to choose *s*/*of*. Thus, higher absolute values of DEVIATION indicate that indigenized-variety speakers made choices that are more at odds with what native speakers were predicted to have said.) On the basis of these deviation values we then do a second random forests analysis that models whether IV speakers make BrE-like choices; the dependent non-BrE-like choices of indigenized-variety speakers, i.e. whether DEVIATION is 0 or not (a variable referred to as BRELIKE) as a function of all the above predictors and VARIETY.

2.3 Statistical evaluation, part 2: Determining accuracies of mutual predictions

The second analysis we are pursuing here is concerned with which varieties predict which other varieties best. As described in Section 2.2 above, the general logic follows that of Gries and Bernaisch (2016): for each variety V_x of the six varieties $V_{1\text{-}6}$,

- we fit a random forest with all predictors;
- we then applied that random forest from V_x to the data from $V_{\neg x}$ to generate predictions of genitive choices;
- we then compared whether the speakers of varieties $V_{\neg x}$ made the same choices predicted from V_x and computed the prediction accuracy from V_x to each of $V_{\neg x}$;
- we summed up the results from all six iterations (i.e., from when each variety was the target).

Unlike in Gries and Bernaisch (2016), we then analyzed these data in two ways: first, we computed a cluster analysis to determine degrees of similarity between varieties and to see which, if any subgroups, emerge; second, we tested which South Asian and which Southeast Asian variety predicted the others in those groups best to see what that would reveal about each variety's potential epicenter status.

3 Results

3.1 Random forests 1 on native-speaker data

The first analysis RF_1 yielded a classification accuracy of 87.8 percent, which is significantly higher than the baselines of always choosing the more frequent genitive (i.e., *of*) or choosing proportionally randomly ($p_{\text{binomial test against baseline1}}<10^{-13}$, $p_{\text{binomial test against baseline2}}<10^{-44}$). More illuminating is the analysis's *C*-value, which exceeds the usually-assumed threshold value for 'good' results of 0.8 with a value of 0.934. We therefore proceeded with the analysis.

3.2 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 IV data. The prediction accuracy measure went down a bit (to 82.8%), and, correspondingly, the *C*-value also decreased to 0.887. 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.

3.3 Random forests 2 on deviations from native-speaker predictions

The final analysis consisted of trying to model BRELIKE as a function of the same predictors as before plus VARIETY in RF_2. The overall summary results were very encouraging in the sense that the statistical analysis could predict the presence/absence of BrE-like decisions very well (classification accuracy=0.86, C=0.875), 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 predictors and VARIETY; on the basis of the variable importance measures returned by the RF_2, we will discuss five predictors and summarily comment on the remaining ones.

3.3.1 The interaction VARIETY : POR_ANIMACY

In RF_2, the strongest predictor of BRELIKE is POR_ANIMACY, and Figure 1 represents the way in which DEVIATION (on the x-axis) varies as a function of POR_ANIMACY across varieties. Recall that DEVIATION values close to 0 represent BrE-like choices and note that the left and the right panel show the same results just from different perspectives: the left panel facilitates comparisons of varieties, the other comparisons of animacy levels.

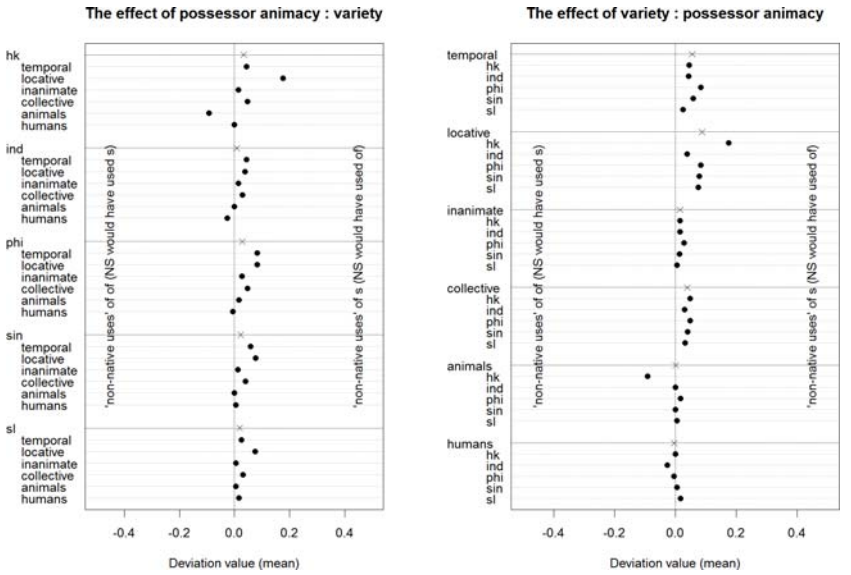

Figure 1: The effect DEVIATION ~ VARIETY : POR_ANIMACY

The most noteworthy results are that, with certain kinds of possessors, such as humans, animals, and inanimates, the IV speakers are very close to the BrE speakers – with the others, i.e. temporal, locative, and collective possessors, the IV speakers use more *s*-genitives. In addition, some varieties are noteworthy: HK displays structural choices different from BrE ones in particular with locative and animal possessors, whereas PHI and SIN do so mostly with temporal and locative possessors; SLE's main variation-inducing factor is also locative possessors, and structural choices in IndE are generally compatible with those in BrE.

41.5 percent of locative possessor heads in HKE (N=53) are *Hong Kong*, which is used in *s*-genitives in over 90 percent of the cases. Other words that are categorically used in *s*-from in HKE are *Beijing* and *China*. Animal possessor heads are extremely rare in our sample. We only find three examples, two of which are lions, and one of which is *pork*.[6] The low frequency of animate possessor heads in HKE might be related to the geographical characteristics of this very densely populated city, where encounters with animals might be lower than in the other countries under investigation.

3.3.2 The interaction VARIETY : ORUM_FREQDIFF_LOG

The second strongest effect is ORUM_FREQDIFF_LOG, which is very similar in strength to just POR_HEAD_FREQ. Since the former includes more information than the latter, we represent the former here. In Figure 2, the frequency differences are on the *x*-axis, the deviation scores are on the *y*-axis; the overall trend for all five varieties is represented by the thicker black smoother, and the other five lines code different varieties as indicated.

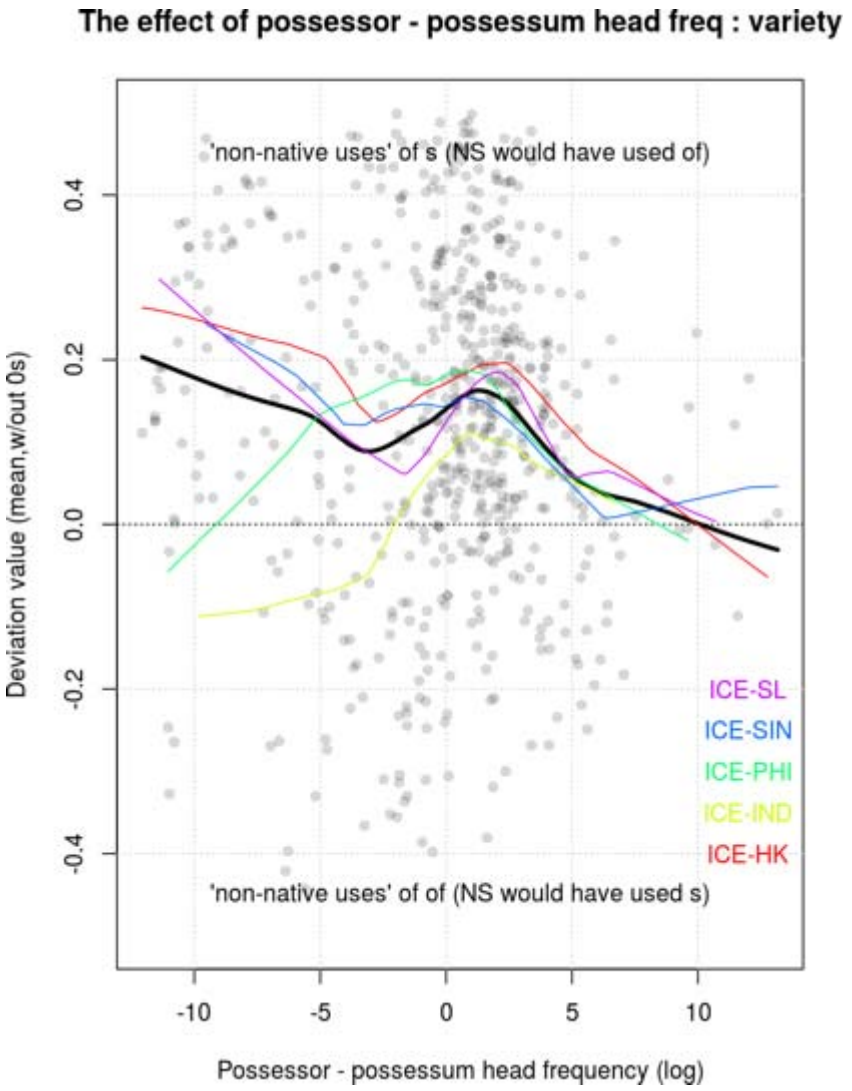

Figure 2: The effect DEVIATION ~ VARIETY : ORUM_FREQDIFF_LOG

This result is interesting in how it relates to previous MuPDAR analyses, in which a frequent finding was that, as a particular cue provides less information, non-nativelike choices become more frequent. In this case, IND and PHI exhibit that behavior: as the frequency difference between possessor and possessum approaches 0, they make the highest number of non-native *s*-genitive choices, and as the frequency differences de-/increase, the choices become, on the whole, more BrE-like. Interestingly, the other three varieties and the overall trend are different: as the possessor becomes more frequent than the possessum, these IV speakers' choices become more BrE-like, most likely because these IV speakers generally tend to employ a higher number of *s*-genitives, but since once the possessor becomes quite more frequent than the possessum, BrE speakers also choose *s*-genitives more, then the IV speakers' choices are compatible with the relevant BrE preferences.

3.3.3 The interaction VARIETY : ORUM_LENDIFF_LOG

The next effect involves the length difference between possessor and possessum; it is represented in Figure 3. The effect is rather straightforward and does not differ much between varieties: the longer the possessor is relative to the possessum, the less BrE-like the choices of the IV speakers. This reflects the fact that native speakers are typically expected to exhibit a short-before-long effect, which means choosing the *of*-genitive when the possessor is longer than the possessed. Apparently, the IV speakers use more *s*-genitives across the board – note how all smoothers are above $y=0$ – but their generally higher use of *s*-genitives becomes particularly noticeable in the right side of the plot where BrE speakers are particularly likely to use *of* instead; the cue length difference is evidently of lower importance for IV than for BrE speakers in genitive choices.

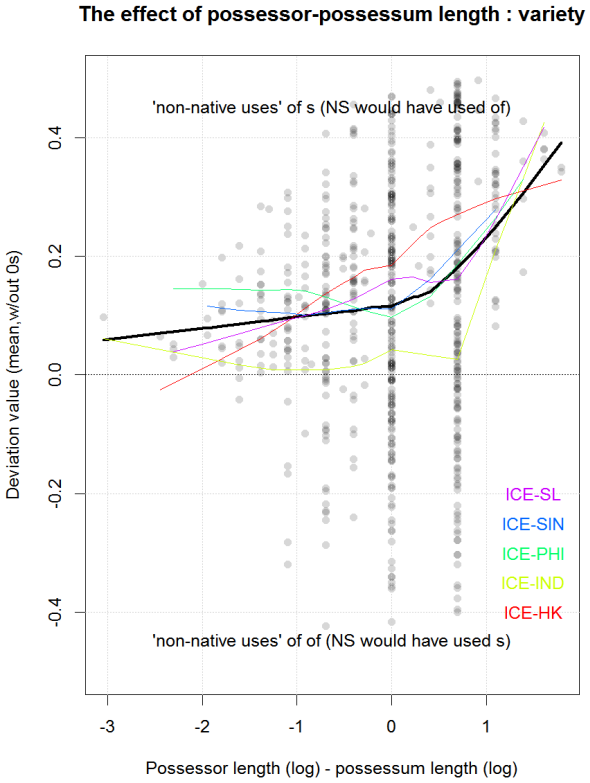

Figure 3: The effect DEVIATION ~ VARIETY : ORUM_LENDIFF_LOG

3.3.4 The interaction VARIETY : POR_THEMATICITY

The final relatively strong effect involves the thematicity of the possessor, which is represented in both panels of Figure 4; the left panel represents thematicity on a log scale of the values in the right panel. On the whole, there is again not a big difference between the five varieties or the overall trend: the more thematic the possessor is, the more BrE-like the IV speakers' choices are because then their generally higher frequency of *s*-genitives meets the increasing tendency of the BrE speakers to also choose *s*-genitives with thematic possessors.

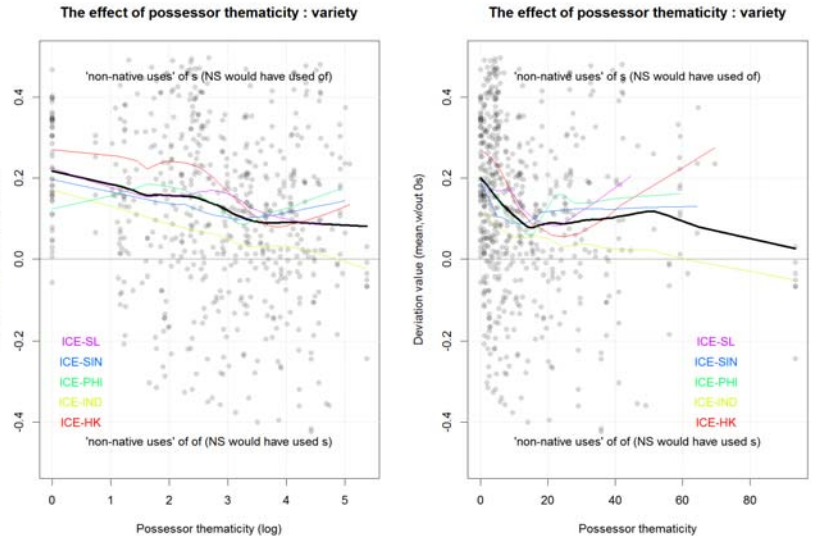

Figure 4: The effect DEVIATION ~ VARIETY : POR_THEMATICITY

What about the other predictors? Some predictors not discussed so far are 'included' in the ones we discussed (e.g., the individual length and frequency values of possessors and possessums), others just scored considerably lower than the ones we did discuss with regard to either one or even both variable importance measures we calculated (mean decrease in accuracy and mean decrease of the *Gini* coefficient). For instance, final sibilancy of the possessor or the coarse genre difference *printed* vs. *non-printed* made very little contributions to whether IV speakers made BrE-like choices; for the former, this is likely due to the fact that final sibilancy has the same *s*-genitive avoidance effect fact across all speakers in our data; for the latter, the resolution was probably just too coarse. That picture changed when we started exploring the more fine-grained genre distinctions. While there seemed to be few systematic and interpretable differences between varieties, it was interesting to note how genres differed from each other: the IV speakers' choices were most often compatible with BrE choices in academic and instructional writing; the largest amount of genitive choices particular to IV speaker contexts were found in reportage as well as pop-

ular and persuasive writing. The givenness of the possessor, the animacy of the possessum, and TTR likewise did not add much to the analysis.

3.4 Clustering and random forests 2 on deviations from native-speaker predictions

The second analysis exploring the similarities of the varieties' predictive power and their epicenter status can be summarized more briefly.

The results of the first part of this analysis are straightforward. If one sums all varieties' predictive accuracies and compares them to each other, then, by a small margin, BrE has the least degree of predictive power whereas SIN has the highest. Once all pairwise predictive accuracies are submitted to a cluster analysis, the picture in Figure 5 emerges:

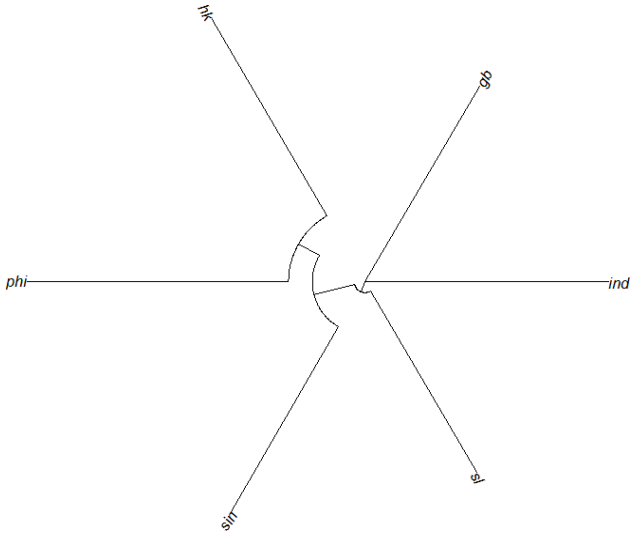

Figure 5: Phylogenetic cluster analysis of similarities between varieties' predictive power

This is an interesting result because (i) we find the Southeast Asian varieties grouped together on the left, (ii) we find the two South Asian varieties together with BrE on the right, and (iii) BrE is closest to IND.

The final part of the analysis consisted of exploring which South Asian and which Southeast Asian variety predicted the others in those groups best to test

our hypotheses that IND and SIN would be the varieties with the highest predictive accuracies respectively. The resulting classification accuracies are summarized in Table 1, where the values represent the prediction accuracy from the variety in the row to the variety in the column.

Table 1: Predictive accuracies (from row variety to column variety)

South Asian

	IND	SL
IND		**0.8526**
SL	0.8476	

Southeast Asian

	HK	PHI	SIN
HK		*0.8279*	0.8325
PHI	0.8502		0.8466
SIN	**0.8453**	*0.8401*	

In the left panel representing the results for the two South Asian varieties, IND predicts SL better than the other way round, as is indicated by the bold percentage of 0.8526 (vs. 0.8476); this is as expected. In the right panel representing the results for the three Southeast Asian varieties, SIN predicts HK slightly better than HK does SIN (see the bold 0.8453 vs. 0.8325), which is also as expected. However, SIN predicts PHI less well than vice versa (see the italicized 0.8401 vs. 0.8466).

In the next section, we will discuss and contextualize our findings and conclude.

4 Discussion and concluding remarks

In this section of the paper, we provide a short summary of the central empirical findings and discuss them in the light of earlier research on the genitive alternation, Asian Englishes and epicenter theory. Against this background, we will also delineate what we perceive to be promising avenues for future research.

4.1 Interim summary

With a focus on South and Southeast Asian Englishes, the present paper studied the genitive alternation in varieties of English spoken in Great Britain, Hong Kong, India, the Philippines, Singapore and Sri Lanka with particular regard to processes of structural nativization and epicentral configurations. Via MuPDARF, we documented that the generally stronger inclination in Asian Englishes to use *s*-genitives in contexts where BrE speakers would use *of*-genitives is driven by four actuators of structural nativization. In decreasing order of

importance, these actuators are possessor animacy, head frequency differences, length differences between possessor and possessum, and possessor thematicity. An iterative approach to the model potential of the respective (so far for the most part assumed) epicentral configurations in South and Southeast Asia highlighted that a) in South Asia, IndE provided better predictions for SLE than the other way round and b) SinE was a better model for HKE than vice versa, but the predictive accuracy of PhilE was higher for SinE than that of SinE for PhilE.

4.2 Implications

Figures (1)–(4) show that the South Asian and Southeast Asian varieties under investigation tend to use the *s*-genitive more frequently than BrE speakers. Only when possessors are either animate or inanimate, highly frequent, highly thematic, or when the possessum is much longer than the possessor, characteristics which are exemplified in (11), do speakers of South Asian and Southeast Asian varieties make genitive choices comparable to BrE speaker choices.

(11) because of [the state]$_{\text{possessor}}$'s [intervention in economy]$_{\text{possessum}}$, the poor will be given a chance to also participate in the economy. (ICE-PHI:W1A-001)

The effects of the individual predictors (animacy, thematicity, etc.) on the English genitive alternation are well-studied and can be explained by referring to models of processing constraints, for example, MacDonald's (2013) Easy First principle (for genitives, see also Rosenbach 2014: 237; for an early and more general psycholinguistic account, see Bock 1982). Easy First states that entities that are more easily retrievable (i.e. animate, thematic, etc.) will be placed first. Easy possessors are, therefore, more frequent in the *s*-genitive, where they precede the possessum. Our cross-varietal investigation of genitive choice did not contradict this pattern (e.g., animate possessors favoring *of*-genitive use instead of *s*-genitive use in any of the varieties). However, compared to BrE speakers, the ESL varieties seem to be less guided by processing constraints. Although ESL speakers use *s*-genitives more often in general, their use drops with highly frequent or highly thematic possessors; when the possessor is relatively long, which according the end-weight principle would make *of*-genitive usage more probable, the ESL speakers use *s*-genitives even more. The data show that the cases where genitive choice contradicts what processing models like MacDonald's (2013) Easy First principle or Behaghel's (1909) principle of end-weight would predict exclusively stem from the ESL varieties (see examples 12–14).

(12) At the same time, the President appealed to the delegates for sobriety in keeping track of the goal of [the Asian-Pacific Parliamentarians]$_{possessor}$' [Union]$_{possessum}$ to uphold the principles of freedom, sovereignty, and territorial integrity (ICE-PHI:W2C-011)

(13) [T]he last time he had met with the Cousin was at the wedding of a cousin – [his father' cousin's sister-in-law]$_{possessor}$'s [grandson]$_{possessum}$ – and the big man had emotionally put his arm over Hairy's shoulders (ICE-IND:W2F-011)

(14) [I]t is perceived as the CFA reclaiming a measure of its autonomy put into jeopardy by [the National People's Congress Standing Committee]$_{possessor}$'s [interpretation]$_{possessum}$ in June 1999 (ICE-HK:W2B-011)

This suggests that the Asian varieties have developed variety-specific preferences of the relative weight of the animacy constraint and the syntactic weight constraint. Whereas weight was found to be the dominant factor with long possessors in British English (Rosenbach 2005), in Asian varieties, animacy seems to be the more important constraint. Further research is therefore encouraged that systematically investigates the variable importance of possessor animacy and syntactic weight in a cross-varietal perspective.

In this study, animacy was operationalized following guidelines in Wolk *et al*. (2013), who found that – although the *s*-genitive was almost exclusively used with animate possessors in Early Modern English times – collective, locative, and temporal possessor heads have over time become less disfavored by the *s*-form. It is revealing that it is exactly with these possessor animacy groups (see Figure 1) that Asian users choose *s*-genitives when British English speakers would have chosen *of*-genitives, while Asian users make choices comparable to ENL speakers with the remaining possessor animacy groups.

This historical trajectory of the expansion of collective, locative, and temporal possessors to the *s*-genitive (Wolk *et al*. 2013) in conjunction with the preference of Asian Englishes for *s*-genitives with exactly these animacy groups leads to the hypothesis that the role the possessor animacy categories play in the individual Asian Englishes may be related to the importance of these categories in the diachronically distinct historical input varieties of the individual Asian Englishes. For example, the historical input variety of IndE, dating back to the beginning of the 17th century when the British used Early Modern English, is markedly different from the historical input variety of Singapore English, whose developmental cycle started in 1819 (cf. Schneider 2007) with a variety of Late Modern English as the basis. These differences in the respective historical input

varieties are *inter alia* reflected in the degree of positive influence that the three possessor animacy categories collective, locative and temporal exhibit on genitive choice. At the beginning of the 17th century, temporal possessors exhibited a stronger influence toward *s*-genitives than collective possessors followed by locative possessors, while the picture changed around 1819 in the sense that, although temporal possessors still showed the strongest tendency towards *s*-genitives among the three discussed, locative possessors had become a stronger cue for *s*-genitives than collective possessors (cf. Wolk *et al*. 2013: 410).

There are some noteworthy parallels between our and Wolk *et al*.'s (2013) study with respect to the importance of the three possessor groups. With IndE and PhilE, temporal possessors show the strongest tendency towards *s*-genitives among the three groups scrutinized (see example 15), which is a parallel to the importance temporal possessors had in relation to collective and locative ones at the beginning of the 17th and 20th centuries, i.e. the times when the respective historical input varieties were used in Britain.

(15) If comparisons are to be made, [today]$_{possessor}$'s [English teacher]$_{possessum}$ is likely to receive a less satisfactory rating [...] (ICE-PHI:W2A-001)

However, based on this rough-and-ready comparison across different studies, the parallels between the importance of the three possessor animacy groups in contemporary Asian Englishes and their respective historical input varieties appear too few and far between to construct a convincing case for the present-day reflection of the historical importance of these possessors in the Asian Englishes concerned – particularly because animate possessors played a much more important role for *s*-genitives in the history of English (cf. Wolk *et al*. 2013: 410) compared to present-day Asian Englishes.

In the light of this admittedly unsophisticated diachronic comparison, the probabilistic genitive profiles of the Asian Englishes scrutinized seem to be different from their respective historical input varieties. At this point it appears more conducive to argue that this trend of the three possessor groups collective, locative and temporal toward *s*-genitives was continued in the ESL varieties when it had already leveled off in the ENL varieties rendering possessor animacy one of the actuators of structural nativization in relation to the genitive alternation in Asian Englishes. Consequently, the process of structural nativization does not only seem to take place on the level of surface structures, but also on the more concealed level of underlying norms triggering the surface structure choices and thus altering the structural profile of the varieties concerned.

This more concealed level of underlying norms clearly relates to the notion of epicenters. Based on their two central criteria, i.e. a) endonormative stabilization and b) model character for varieties in their surroundings (cf. Hundt 2013: 185), we are able to further substantiate the status of IndE as the South Asian linguistic epicenter as already empirically validated with regard to the dative alternation (cf. Gries and Bernaisch 2016). IndE is undoubtedly an endonormatively stabilized variety (cf. e.g., Mukherjee 2007) and its structural model of the genitive alternation provides accurate predictions for other varieties of the Indian subcontinent such as Sri Lankan English. The relevance of epicentral variety users' attitudes to 'rupture zone' varieties, i.e. those varieties under epicentral influence, and – more importantly – vice versa for the cross-regional spread of linguistic items can be expected to vary in accordance with the structural features under scrutiny and the degrees to which users in the speech communities concerned are aware of and associate them with a certain regional origin. With globally shared common-core (cf. Quirk *et al*. 1985: 16) elements of English, such as the constructions constituting the genitive or dative alternation without a strong – if any – association with a certain regional variety, it seems plausible to assume that regional attitudinal profiles play a marginal role at best. Similarly, with region-specific constructions shared across epicentral and rupture-zone varieties – for South Asia features such as presentational *itself* in (2) or intrusive *as* in (3) used in India and Sri Lanka – cross-varietal attitudes can be expected to play a comparably peripheral role as well since the feature spreading in an epicentral fashion would probably be perceived to be local in the rupture zone already and thus not be subject to attitudinal discrimination. Cross-varietal attitudinal considerations should take center stage with epicentral investigations of features which a) are used in the epicentral, but not yet in rupture-zone varieties, b) rupture-zone variety users can clearly identify as an element of the epicentral variety and c) rupture-zone variety users have an active awareness of and can recognize in discourse. The genitive alternation belongs to the first group of common-core phenomena despite significant regional probabilistic adjustments, which, however, are also more than likely to escape users' linguistic awareness. As a consequence, although earlier empirical research profiled SLE speakers' slightly negative attitude towards IndE (cf. Bernaisch 2012: 286), the relevance of rupture-zone variety users' attitudes – in the present study users of SLE – towards the epicentral variety – here IndE – must be assumed to be no more than marginal.

The status of SinE as a linguistic epicenter for Southeast Asia is not similarly undisputable in the light of sociolinguistic considerations and the empirical evidence of this study. True, among the Southeast Asian Englishes studied, SinE

certainly is the only variety that can reasonably be profiled as an endonormatively stabilized variety (cf., e.g., Schneider 2007: 155). However, while SinE provided better predictions for HKE than the other way round, the PhilE model had a higher predictive accuracy for SinE than the SinE model for PhilE. This, of course, raises the question related to the reach of the seismic waves of a linguistic epicenter. In South Asia, IndE is located at the very heart of and surrounded by all other South Asian Englishes making it – also geographically speaking – an ideal candidate for a linguistic epicenter given the physical proximity to the remaining South Asian Englishes. Singapore – in contrast – is physically distant from the other varieties it is supposed to influence, making epicentral spreads of features via face-to-face contact and also textbook material (cf. Hundt 2013: 189) simply more unlikely. Still, the fact that PhilE is the one variety for which SinE cannot be profiled as an epicentral model deserves special attention. PhilE is the only variety among the ESL varieties studied which is "not a product of British but of American colonial expansion" (Schneider 2007: 140). Against this background, one could hypothesize that the structural profile of PhilE is (still) more influenced by its American English historical input variety than by the potential British English-based regional epicenter Singapore English.

4.3 Where to go from here

Although our dataset contained many predictors that influence English genitive choice, there are others that we did not include in our analysis yet, e.g., semantic relation and definiteness (for an exhaustive overview see Rosenbach 2014).

In future research, it would be highly desirable to find answers to the questions that this discussion has prompted. While epicentral influences can certainly provide explanatory avenues for structural similarities across South and Southeast Asian Englishes, the degree of influence the respective historical input varieties (still) exert on the individual Asian Englishes has so far not been approached systematically from a norm-oriented perspective. Studies to come should probably account for the different historical input varieties in two ways – diachronically and regionally. Particularly with BrE-based varieties, it would be highly relevant to compare their present-day structural profile to those of their diachronically distinct historical input varieties. While IndE must be assumed to have emerged from a variety of Early Modern English, SinE stems from a Late Modern English variety. Further, it should also be considered that American instead of BrE served as the input for a number of present-day varieties such as PhilE and should consequently also be included as a potential norm provider for structures of Asian Englishes.

We also hope to have shown that MuPDARF is a feasible method to study the regional model character of linguistic epicenters. It would be desirable to further substantiate the status of IndE as a linguistic for South Asian Englishes (and possibly also that of SinE for Southeast Asian Englishes) on the basis of other structural characteristics, such as particle placement. Further, the present study focused on two epicentral constellations in Asian Englishes, but other potential epicenters in the Pacific region, Africa or America would certainly also benefit from further model-oriented empirical research to add to earlier, partly anecdotal epicentral accounts.

Notes

1. Available at http://www.llc.manchester.ac.uk/research/projects/germanic-possessive-s/**.**
2. Available at https://wordnet.princeton.edu**.**
3. Lemmas were determined using Yasumasa Someya's lemma list available at http://lexically.net/downloads/BNC_wordlists/e_lemma.txt**.**
4. Available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict**.**
5. In keeping with the terminology used in the MuPDAR and MuPDARF methods, the expression *DEVIATION* is employed in this paper to neutrally describe (quantitative) structural preferences which have evolved in the individual Asian Englishes under scrutiny without any normative native-speaker-centered underpinnings.
6. Usually, *pork* is inanimate. In this case (i.e. *slaughter and distribution of pork*), it is by definition animate because it is impossible to *slaughter* inanimate entities.

References

Altenberg, Bengt. 1982. *The genitive v. the of-construction: A study of syntactic variation in 17th century English*. Lund: Gleerup.

Behaghel, Otto. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25: 110–142.

Bernaisch, Tobias. 2012. Attitudes towards Englishes in Sri Lanka. *World Englishes* 31(3): 279–291.

Bernaisch, Tobias. 2015. *The lexis and lexicogrammar of Sri Lankan English*. Amsterdam: John Benjamins.

Bernaisch, Tobias and Christopher Koch. 2016. Attitudes towards Englishes in India. *World Englishes* 35(1): 118–132.

Bernaisch, Tobias and Claudia Lange. 2012. The typology of focus marking in South Asian Englishes. *Indian Linguistics* 73(1–4): 1–18.

Bernaisch, Tobias, Christopher Koch, Joybrato Mukherjee and Marco Schilk. 2011. *Manual for the South Asian Varieties of English (SAVE) Corpus: Compilation, cleanup process, and details on the individual components*. Giessen: Justus Liebig University.

Bernaisch, Tobias, Stefan Th. Gries and Joybrato Mukherjee. 2014. The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide* 35(1): 7–31.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Bock, J. Kathryn. 1982. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review* 89(1): 1–47.

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

Deshors, Sandra C. and Stefan Th. Gries. 2016. Profiling verb complementation constructions across New Englishes: A two-step random forests analysis of *ing* vs. *to* complements. *International Journal of Corpus Linguistics* 21(2): 192–218.

Ehret, Katharina, Christoph Wolk and Benedikt Szmrecsanyi. 2014. Quirky quadratures: On rhythm and weight as constraints on genitive variation in an unconventional data set. *English Language and Linguistics* 18(2): 263–303.

Emeneau, Murray B. 1956. India as a linguistic area. *Language* 32(1): 3–16.

Fuchs, Robert. 2016. *Speech rhythm in varieties of English. Evidence from educated Indian English and British English*. Singapore: Springer.

Gahl, Susanne and Susan Garnsey. 2004. Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language* 80(4): 748–775.

Grafmiller, Jason. 2014. Variation in English genitives across modality and genres. *English Language and Linguistics* 18(3): 471–496.

Greenbaum, Sidney. 1996. Introducing ICE. In S. Greenbaum (ed.). *Comparing English worldwide: The International Corpus of English*, 3–12. Oxford: Clarendon.

Gries, Stefan Th. 2002. Evidence in linguistics: Three approaches to genitives in English. In R.M. Brend, W.J. Sullivan and A.R. Lommel (eds.). *LACUS Forum XXVIII: What Constitutes Evidence in Linguistics?*, 17–31. Fullerton, CA: LACUS.

Gries, Stefan Th. and Allison S. Adelman. 2014. Subject realization in Japanese conversation by native and non-native speakers: Exemplifying a new paradigm for learner corpus research. *Yearbook of Corpus Linguistics and Pragmatics 2014: New empirical and theoretical paradigms*. Cham: Springer. 35–54.

Gries, Stefan Th. and Sandra C. Deshors. 2014. Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora* 9(1): 109–136.

Gries, Stefan Th. and Tobias Bernaisch. 2016. Exploring epicenters empirically: Focus on South Asian Englishes. *English World-Wide* 37(1): 1–25.

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

Hilpert, Martin. 2008. The English comparative – language structure and language use. *English Language and Linguistics* 12(3): 395–417.

Hinrichs, Lars and Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 11: 437–474.

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

Hundt, Marianne and Benedikt Szmrecsanyi. 2012. Animacy in early New Zealand English. *English World-Wide* 33: 241–263.

Jahr Sohrheim, Mette-Catherine. 1980. *The* s*-genitive in Present-day English*. PhD dissertation, University of Oslo.

Jankowski, Briget Lynn. 2013. *A variationist approach to cross-register language variation and change*. PhD dissertation, University of Toronto.

Koch, Christopher and Tobias Bernaisch. 2013. Verb complementation in South Asian English(es): The range and frequency of "new" ditransitives. In G. Andersen and K. Bech (eds.). *English corpus linguistics: Variation in time,*

space and genre – selected papers from ICAME 32, 69–89. Amsterdam: Rodopi.

König, Ekkehard. 1993. Focus particles. In J. Jacobs, A. von Stechow, W. Sternefeld, T. Vennemann and H.E. Wiegand (eds.). *Syntax: Ein Internationales Handbuch zeitgenössicher Forschung/An international handbook of contemporary research*, 978–987. Berlin: de Gruyter.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Philadelphia Press.

Lange, Claudia. 2012. *The syntax of spoken Indian English*. Amsterdam: John Benjamins.

Lange, Claudia. 2016. The 'intrusive as'-construction in South Asian varieties of English. *World Englishes* 35(1): 133–146.

Leitner, Gerhard. 1992. English as a pluricentric language. In M.G. Clyne (ed.). *Pluricentric languages: Differing norms in different nations*, 179–237. Berlin: Mouton de Gruyter.

Liaw, Andy and Matthew Wiener. 2002. Classification and regression by randomForest. *R News* 2(3): 18–22.

Liaw, Andy and Matthew Wiener. 2015. randomForest. Version 4.6-12. A package for R. (See Liaw and Wiener 2002).

Lim, Lisa and Umberto Ansaldo. 2015. *Languages in contact*. Cambridge: Cambridge University Press.

Masica, Colin P. 1976. *Defining a linguistic area: South Asia*. Chicago: University of Chicago Press.

Meyler, Michael. 2007. *A dictionary of Sri Lankan English*. Colombo: Mirisgala.

Mukherjee, Joybrato. 2007. Steady states in the evolution of New Englishes: Present-day Indian English as an equilibrium. *Journal of English Linguistics* 35(2): 157–187.

Mukherjee, Joybrato and Sebastian Hoffmann. 2006. Describing verb-complementational profiles of New Englishes: A pilot study of Indian English. *English World-Wide* 27(2): 147–173.

Nelson, Gerald. 1996. The design of the corpus. In S. Greenbaum (ed.). *Comparing English worldwide: The International Corpus of English*, 3–12. Oxford: Clarendon.

Osselton, Noel E. 1988. Thematic Genitives. In Nixon, Graham and John Honey (eds.). *An historic tongue: Studies in English Linguistics in memory of Barbara Strang*, 138–144. London: Routledge.

Peters, Pam. 2009. Australian English as a regional epicenter. In T. Hoffmann and L. Siebers (eds.). *World Englishes – problems, properties and prospects*, 107–124. Amsterdam: John Benjamins.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London and New York: Longman.

Rosenbach, Anette. 2002. *Genitive variation in English: Conceptual factors in synchronic and diachronic studies*. Berlin, New York: Mouton de Gruyter.

Rosenbach, Anette. 2005. Animacy versus weight as determinants of grammatical variation in English. *Language* 81(3): 613–644.

Rosenbach, Anette. 2014. English genitive variation – the state of the art. *English Language and Linguistics* 18(2): 215–262.

Rosenbach, Anette and Letizia Vezzosi. 2000. Genitive constructions in Early Modern English: New evidence from a corpus analysis. In R. Sornicola, E. Poppe and A. Shisha-Halevy (eds.). *Stability, variation and change of word-order patterns over time*, 285–307. Amsterdam: John Benjamins.

Sanyal, Jyoti. 2006. *Indlish: The book for every English-speaking Indian*. Delhi: Viva Books.

Schilk, Marco. 2011. *Structural nativization in Indian English lexicogrammar*. Amsterdam: John Benjamins.

Schilk, Marco, Tobias Bernaisch and Joybrato Mukherjee. 2012. Mapping unity and diversity in South Asian English lexicogrammar. Verb-complementational preferences across varieties. In M. Hundt and U. Gut (eds.). *Mapping unity and diversity world-wide: Corpus-based studies of New Englishes*, 137–165. Amsterdam: John Benjamins.

Schneider, Edgar W. 2003. The dynamics of New Englishes: From identity construction to dialect birth. *Language* 79(2): 233–281.

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

Sedlatschek, Andreas. 2009. *Contemporary Indian English: Variation and change*. Amsterdam: John Benjamins.

Senaratne, Chamindi Dilkushi. 2009. *Sinhala-English code-mixing in Sri Lanka: A sociolinguistic study*. Utrecht: LOT.

Szmrecsanyi, Benedikt. 2010. The English genitive alternation in a cognitive sociolinguistics perspective. In D. Geeraerts, G. Kristiansen and Y. Peirsman (eds.). *Advances in cognitive sociolinguistics*, 141–166. Berlin/New York: Mouton de Gruyter.

Szmrecsanyi, Benedikt and Lars Hinrichs. 2008. Probabilistic determinants of genitive variation in spoken and written English: A multivariate comparison across time, space, and genres. In T. Nevalainen, I. Taavitsainen, P. Pahta and M. Korhonen (eds.). *The dynamics of linguistic variation: Corpus evidence on English past and present*, 291–309. Amsterdam: Benjamins.

Szmrecsanyi, Benedikt, Douglas Biber, Jess Egbert and Karlien Franco. 2016. Toward more accountability: Modeling ternary genitive variation in Late Modern English. *Language Variation and Change* 28: 1–29.

Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller and Melanie Röthlisberger. 2016. Around the world in three alternations. Modeling syntactic variation in varieties of English. *English World-Wide* 37(2): 109–137.

Thomas, Russel. 1931. *Syntactical processes involved in the development of the adnominal periphrastic genitive in the English language*. PhD dissertation, University of Michigan.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach and Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3): 382–419.

Wulff, Stefanie and Stefan Th. Gries. 2015. Prenominal adjective order preferences in Chinese and German L2 English. A multifactorial corpus study. *Linguistic Approaches to Bilingualism* 5(1): 122–150.