The role of gender in postcolonial syntactic choice-making: evidence from the genitive alternation in British and Sri Lankan English

Stefan Th. Gries
UC Santa Barbara &
Justus Liebig University Giessen

Benedikt Heller
Justus Liebig
University Giessen

Nina S. Funke
Justus Liebig
University Giessen

Abstract
This paper studies the genitive alternation in British English and Sri Lankan English on the basis of more than 4000 annotated cases of *of*- and *s*-genitives from the British and Sri Lankan components of the International Corpus of English. Specifically, we explore the effects of a variety of language-internal and language-external effects, focusing in particular on how these factors affect genitive choices both on their own, but also in interaction with each other and, a first in this kind of variety research, with the gender of the speakers. Our results corroborate previous findings regarding the language-internal factors, but we also obtain a variety of statistical effects representing interactions of those with variety and gender: For instance, animacy effects are stronger in Sri Lankan English, but animacy and length/weight effects are moderated by speaker gender; we discuss these and other findings with regard to processing, language contact, and gender (in)equality. Methodologically, we are developing two innovations for variationist research, namely a principled way to identify and then also visualize the effect of interactions in random forests.

1 Introduction

1.1 General introduction

There is a rich history of research on the English genitive alternation (Rosenbach 2014), the choice between the *s*-genitive and the *of*-genitive. The two options differ mainly in two respects: the ordering and linking of two constituents. In the *s*-genitive as in (1)a, the possessor (*Sri Lanka*) precedes the possessum (*self-interest*), whereas it follows the possessum (*image*) in the *of*-genitive as in (1)b. To indicate the genitive relation between the constituents, the clitic *'s* (or simply *'* with plurals) is added to the possessor phrase in *s*-genitives, whereas in *of*-genitives, constituents are connected by the preposition *of*. Additionally, the definite article is not explicitly stated in *s*-genitives (since *'s* has a deterministic function), so if we wanted to transform (1)a into an *of*-genitive, we would have to add it (i.e., *the self-interest of Sri Lanka*).

(1) a. So let us now serve **[Sri Lanka]$_{\text{possessor}}$'s [self-interest]$_{\text{possessum}}$** by making use of the potential that this whole region has for massive economic development in the years to come (ICE-SL:S1B-051, 1)

b. You had been manufacturing films to tarnish the **[image]$_{\text{possessum}}$ of [Sri Lanka]$_{\text{possessor}}$** or photographs with the aim of promoting communism, terrorism or creating communal tension […] (ICE-SL:S1B-054, 1)

The present investigation uses a variationist approach (see, e.g., Szmrecsanyi 2017) to grammatical choice making and, in doing so, focuses on genitives that constitute "alternate ways of saying 'the same' thing" (Labov 1972:188; for a discussion of sameness in the genitive

alternation, see Rosenbach 2002). This entails excluding certain forms, which cannot be expressed in the respective other form (see Rosenbach 2002); these forms include appositive genitives (e.g., *the state of California*), descriptive genitives (e.g., *person of color*), double genitives (e.g., *the friend of John's*), noun-noun genitives (e.g., *satellite photographs*; but cf. Szmrecsanyi et al. 2016), partitive genitives (e.g., *one of my friends*), and idiomatic genitives (e.g., *Valentine's Day*). In focusing on the so-called choice context (i.e., on cases that can be expressed in both genitive variants), we seek to produce results that complement recent multifactorial studies of genitive choice (e.g., Grafmiller 2014, Heller, Szmrecsanyi, & Grafmiller 2017, *inter alia*).

Linguists have long been aware that the choice between the *s*-genitive and the *of*-genitive is subject to multiple constraints, the most important of which is possessor animacy. In most studies, this refers to the binary distinction between animate (2)a and inanimate (2)b possessors, but more fine-grained additions to this distinction have been found to be important in explaining genitive choice (e.g., Wolk et al. 2013): collective (2)c, locative (2)d and temporal (2)e possessors. In essence, the higher a possesssor is on the animacy scale, the more likely it is to be used in an *s*-genitive.

(2) a. What form of government will then replace **[Saddam Hussein]$_{\text{possessor}}$'s [dictatorship]$_{\text{possessum}}$**? (ICE-GB:W2E-001, 3)
b. **The [boiling point]$_{\text{possessum}}$ of [coconut oil]$_{\text{possessor}}$** is more conducive for frying (ICE-SL:S1A-006, 1)
c. Had **the [timely arrival]$_{\text{possessum}}$ of [US and British forces]$_{\text{possessor}}$** not prevented this manoeuvre, some 5 million barrels a day of oil production might have been put at risk (ICE-GB:W2E-001, 3)
d. Mahatma Gandhi called India **[Sri Lanka]$_{\text{possessor}}$'s [nearest neighbour]$_{\text{possessum}}$** (ICE-SL:S1B-051, 1)
e. [NAME], now **[today]$_{\text{possessor}}$'s [question]$_{\text{possessum}}$** again; let me repeat that. (ICE-SL:S1A-094, 1)

Almost equally important is the length of the constituents (for a comparison of length and animacy see Rosenbach 2005). If the possessor is particularly long (in relation to the possessum, that is), it is more likely to be used in an *of*-genitive, in which it follows the possessum (3)a. On the other hand, if the possessum is particularly long, it is more likely to be used in an *s*-genitive, in which it, again, is placed in final position (3)b. In other words, genitive choice corresponds to the principle of end-weight (Behaghel 1909), according to which longer constituents are often placed last. Both the effect of possessor animacy and the effect of constituent lengths reflect a more general principle of linguistic choice-making: Easy First (MacDonald 2013), which states that constituents that are more easily retrievable from memory (e.g., animate (see Bock 1982:15 on egocentric bias) and/or short ones) tend to be placed first.

(3) a. Now Mr deputy speaker with **the [conclusion]$_{\text{possessum}}$ of [the conflict situation]$_{\text{possessor}}$** (ICE-SL:S1B-056, 1)
b. And also another interesting thing is **[Ernest Hemingway]$_{\text{possessor}}$'s [last posthumously published novel]$_{\text{possessum}}$** (ICE-SL:S2A-025, 1)

Beyond possessor animacy and constituent length, the present study considers several

additional constraints: three language-internal (i.e., sibilancy, definiteness, and semantic relation) and three language-external ones (i.e., modality, variety, and gender). Sibilancy refers to the presence of a final sibilant at the end of the possessor phrase. If a sibilant ([s], [z], etc.) is present there, language users tend to avoid the *s*-genitive because the combination of the sibilant and the clitic *'s* creates a repetitive sound sequence (4). Definiteness also refers to the possessor and simply distinguishes definite and indefinite ones; if the possessor is definite (5)a, *s*-genitive usage is more probable than with indefinite possessors (5)b. Semantic relation is here operationalized as a binary distinction between prototypical (including part-whole, kinship, and legal) relations (e.g., (6)a and non-prototypical ones (6)b). With prototypical semantic relations, the *s*-genitive is usually more frequent than with non-prototypical ones. The effects of these three language-internal constraints have been found and replicated in many studies (an overview of which can be found in the appendix to Rosenbach 2014).

(4) In **[NAME]$_{possessor}$'s [terms]$_{possessum}$** [...] (ICE-SL:S2B-049, 1)

(5) a. we met **[auntie NAME]$_{possessor}$'s [daughter]$_{possessum}$** […] (ICE-SL:S1A-011, 1)

b. six hundred years ago, **the [delicious flavour]$_{possessum}$ of [mushrooms]$_{possessor}$** intrigued the Pharaohs of Egypt (ICE-SL:S1B-021, 1)

(6) a. **[the slave owners]$_{possessor}$' [children]$_{possessum}$** […] (ICE-SL:S2A-024, 1)

b. you know **the [underlying principles]$_{possessum}$ of [all the grammar issues]$_{possessor}$** (ICE-SL:S1A-015, 1)

The language-external predictors that are included are modality, variety, and gender. Modality (*spoken* vs. *written*) is included as a control because it has been found to have an effect, especially in interaction with language-internal predictors (e.g., Grafmiller 2014). Szmrecsanyi & Hinrichs (2008) report more *s*-genitive usage in spoken texts, which they attribute to a difference in average formality (with spoken language usually being less formal). Low formality, in turn, has long been recognized as a factor that increases the use of the *s*-genitive (Altenberg 1982). Genitive choice across different varieties has only recently been studied in detail (Heller, Bernaisch, & Gries 2017, Szmrecsanyi et al. 2017, Heller 2018), but results show significant differences, especially for variety as moderator of possessor animacy. In essence, it was found that possessor animacy triggers the *s*-genitive more strongly in Inner Circle varieties (e.g., Britain, Canada, etc.; see Kachru 1985) than in Outer Circle varieties (e.g., Heller, Szmrecsanyi, & Grafmiller 2017). Gender has—to our knowledge—so far not been studied as predictor of genitive choice. A study of the (arguably comparable) dative alternation that included gender found non-significant results (Kendall et al. 2011).

The present study seeks to add to the research of syntactic alternations across varieties by investigating genitive choice in Sri Lankan English (SriLE). SriLE is a post-colonial variety with a clear variety-specific structural profile that spans several linguistic levels (e.g., Meyler 2007; Künstler et al. 2009; Bernaisch 2012, 2015). Within Schneider's Dynamic Model, SriLE has thus passed the nativization phase and is arguably on its way toward endonormative stabilization (Mukherjee 2008:361). In order to characterize genitive choice in SriLE, we compare it to its historical input variety, British English (BrE).

Thus far, there has been little research on gender differences in SriLE. In his dictionary of SriLE, Meyler (2007:53) presents words that women use more often, such as child! as "a colloquial term of address". Bernaisch (2012) investigates Sri Lankans' attitudes towards American, British, Indian, and Sri Lankan English without finding a statistically significant

difference between female and male participants. Bernaisch and Revis (to appear) report a significant effect of gender on the choice of filled and unfilled pauses in a conditional inference tree. However, in their general linear mixed-effects model, gender did not reach significance. Gunesekera (2005), in her analysis of the postcolonial identity of SriLE, found the phenomenon of topicalization to be more common in female than in male speech, while sports metaphors and swear words were – at least in public – limited to male speech (Gunesekera 2005: 137).

With respect to genitive choice, SriLE has also not yet received much scholarly attention (pace Heller, Bernaisch, & Gries 2017). Given that previous studies of genitive choice across varieties revealed that language-external factors might influence genitive choice only as moderator of language-internal constraints (e.g., variety, which was found to moderate the strength of possessor animacy), and given the rich history and importance of the study of gender-based differences in language use (see the introduction to this volume), we reckon that the systematic study of gender in genitive choice constitutes a research gap. The present study, therefore, seeks to complement the current body of research on the genitive alternation by (1) including gender as language-external variable, and (2) by investigating gender across varieties.

1.2 Overview of the present paper

In the following section on methods (Section 2), we will first outline the data extraction process and show some descriptive statistics of the above-mentioned predictors (Section 2.1). In Section 2.2, we will present the details of the statistical analysis, whose results are then described and visualized in Section 3. In the final section, Section 4, we will provide a discussion and concluding remarks.

2 Methods

2.1 Data

Data were extracted from a 10% register-stratified sample of the British component of the Corpus of English (ICE; Greenbaum 1996) and a 25% sample of the Sri Lankan components (because at the time of our retrieval, only 25% of the spoken component were available). The extraction of interchangeable genitives was accomplished in several (partly automatized) steps: First, all text units containing genitive markers (i.e., *of*, *'s*, and *-s'*) were automatically extracted. Then, an automatic classification determined interchangeability. Annotation of predictors was done automatically where possible, based on computational work from Heller (2018). In every step, manual corrections were made where necessary. Speaker gender information for each case was taken from the respective metadata in the case of ICE-SL, and from metadata made available by Martin Schweinberger[1] in the case of ICE-GB.

Altogether, we ended up with 4045 cases that are distributed across VARIETY and GENDER as shown in Table 1, with an overall preference for *of*- over *s*-genitives to a degree that raised initial concerns about the class imbalance problem (the problem that if the (two) levels of the dependent variables are very skewed in favor of one option already, it can become problematic to get good results out of a regression).

1 Available at <http://www.martinschweinberger.de/blog/resources/>.

Table 1: Composition of the data with respect to VARIETY and GENDER

VARIETY	GENDER	GENITIVE: *of*	GENITIVE: *s*	Sum
Great Britain	*male*	305	104	409
	female	48	19	67
	unknown	434	153	587
Sri Lanka	*male*	1185	298	1483
	female	580	310	890
	unknown	500	109	609
Sum (baseline %)		3052 (75.5%)	993 (24.5%)	4045

In addition to the variables shown in Table 1, the data were then annotated with regard to several other predictors discussed in the introduction; the following is an overview of these predictors and their levels; the patterns of how genitives are distributed across the levels of previously studied predictors are in line with previous research.

- MODALITY: *spoken* vs. *written*: our sample contains 1878 genitives from spoken texts, 27.10% of which are *s*-genitives. Of the 2167 genitives from written texts, only 22.34% are *s*-genitives. Data sparsity permit neither a more fine-grained division of the two modalities into different registers nor an analysis of how MODALITY interacts with other predictors; however, we are not aware of studies of alternation phenomena in which MODALITY interacted with other, linguistic predictors in a way that led to a reversal of hypothesized effects anyway.
- ANIMACY (of the possessor): *animate* vs. *collective* vs. *locative* vs. *temporal* vs. *inanimate.* In our sample, the distribution of possessor animacy is as follows: *animate* – 1028, *collective* – 695, locative – 280, temporal – 184, inanimate – 1858. The proportion of *s*-genitives is 59.44%, 32.81%, 19.64%, 33.15%, and 2.05%, respectively.
- SIBILANCY (of the final phoneme of the possessor): *absent* vs. *present.* In 3053 cases, there is no final sibilant, but in 992 cases, a final sibilant is present; *s*-genitives are used in 28.15% and 13.53% of the cases, respectively.
- DEFINITENESS (of the possessor): *definite* vs. *indefinite*. 2621 possessors are definite, while 1424 are not. With definite possessors, the *s*-genitive rate is 31.63%, with indefinite ones, it is 11.52%.
- LENGTHDIFF: the difference of $\log_2$ possessor length minus $\log_2$ possessum length, thus an approximately normally distributed numeric predictor ranging from -4.52 to 4.71 (length is measured in words). In 224 genitives, the possessor and the possessum are equally long; the *s*-genitive is used in 31.70% of these cases. When the possessor is longer (i.e, LENGTHDIFF > 0), the *s*-genitive is used in only 14.98 % of the cases. If it is shorter (i.e, LENGTHDIFF < 0), the *s*-genitive is used in 38.32% of the cases.
- SEMRELATION: *prototypical* vs. *non-prototypical*. In our data, we find 242 prototypical and 3803 non-prototypical relations. When prototypical, the *s*-genitive is used in 52.90% of the cases, and when not, in only 22.75%.

2.2 Statistical evaluation

While this kind of alternation question is one that would prototypically be explored with a generalized linear (mixed-effects) model, we did not proceed along that route. This is due to both the skewed distribution (towards the *of*-genitive) already briefly mentioned above in Table 1 and the additional fact that the potential random-effects structure looked as if it would become highly

problematic: There was a fairly high number of speakers who contributed only few data points to the sample (30% of the data points were by speakers contributing only 10 or fewer data points), lowering the chance of proper convergence of our regression models and/or the random effects being particularly relevant. We therefore decided to use an approach based on random forests, an extension of classification and regression trees, here specifically the kind referred to as conditional inference trees (Hothorn et al. 2006) and implemented in R as party::cforest. Random forests add additional layers of randomness to such a tree-based analysis: First, many different conditional inference trees are constructed on different bootstrapped samples of the data. Second, each split in a conditional inference tree is only permitted to choose from a randomly-chosen subset of the available predictors rather than all of them. The predictions of the random forest consist of amalgamating the multitude of trees that were generated and their 'votes' for the out-of-bag cases. Typically, the user has to specify only two hyperparameters (i.e., parameters that are defined before a statistical analysis begins and affect how it is conducted): the number of (randomly-chosen) predictors that may be considered at each split of each tree (we left that at the default value of 5) and the number of trees grown (we set that to 2000).

In order to interpret the results of the random forest analysis, several strategies are available. One that has been in use especially since Tagliamonte & Baayen (2012) is to (i) perform a random forest analysis on the data, (ii) report variable importance scores from the random forest to assess each predictor's importance to the alternation, and (iii) use a single classification/conditional inference tree on the complete data to visualize the predictors' effects. In this study, we are not following this approach. This is for two main reasons that previous research has ignored. First, the practice of interpreting a random forest – i.e. a set of often 500 or even many more trees on randomly resampled data with different sampled predictors at every split – on the basis of a *single* tree on *all the data* with no resampling is highly problematic and can lead to misinterpretation of the patterns in the data. Second, the way in which random forests are often interpreted – variable importance scores and partial dependency scores – can fail dramatically at representing the nature of the effects in the data faithfully in terms of both over- or underestimated variable importance scores and how predictors interact with each other. Space does not permit a more detailed discussion here, suffice it to say that trees and random forests, which are supposed to be very good at detecting and visualizing interactions are not necessarily as good as they are widely believed to be (see Gries, to appear, for more discussion and exemplification and Deshors & Gries, to appear, for another English-varieties application).

In order to address all these issues we follow Gries's (to appear) recommendations: The first step of our statistical analysis consisted of manually creating a number of new predictors that represent what in a regression model would be interaction predictors, i.e. new variables that embody all combinations of the predictors they consist of:

- all two-way interactions of all predictors with GENDER and VARIETY: GENDER:VARIETY, GENDER:MODALITY, VARIETY:MODALITY, GENDER:ANIMACY, VARIETY:ANIMACY, GENDER:SIBILANCY, VARIETY:SIBILANCY, GENDER:DEFINITENESS, VARIETY:DEFINITENESS, GENDER:LENGTHDIFF, VARIETY:LENGTHDIFF, GENDER:SEMRELATION, and VARIETY:SEMRELATION;
- all three-way interactions involving GENDER and VARIETY: VARIETY:GENDER:MODALITY, VARIETY:GENDER:ANIMACY, VARIETY:GENDER:SIBILANCY, VARIETY:GENDER:DEFINITENESS, VARIETY:GENDER:LENGTHDIFF, and VARIETY:GENDER:SEMRELATION.

These were then added as predictors to a forest of all 2000 conditional inference trees.

We then evaluated the forest in two ways: First, we computed the forest's overall prediction accuracy, its precision and recall, and its *C*-score to determine how well the forest identified structure in our data; second, we computed regular variable importance scores but also an alternative one proposed in Janitza, Strobl, & Boulesteix (2013), which is not based on error rates from categorical predictions, which loses important probabilistic information, but in fact on the area under the curve (*AUC*), which does not just rely on categorical predictions, but also uses the probabilistic strength of the predictions.

As for evaluating the directions of effects, multiple options are theoretically available and it does not seem as if there is much of a discussion let alone a consensus yet as to what works best. One could explore effects on the basis of

- the observed percentages of *of*- and *s*-genitives for each level of each predictor (main effects or interaction predictors alike);
- the averages of the predicted percentages of *s*-genitives for every *attested* combination of each level of each predictor;
- the weighted (by frequency of occurrence) averages of the predicted percentages of *s*-genitives for every *theoretically possible* combination of each level of each predictor.

It does not seem that much of the corpus-linguistic literature on random forests topicalizes this issue much but, after some consideration, we ultimately decided to go with the last option: While the first approach would be appealing for its simplicity, it has the huge disadvantage that it shows the differences between levels of a predictor but too simplistically, because this would involve levels of a predictor *without* controlling for all other effects or holding all others constant; thus, one would never know to what degree the effect observed for one predictor is also (in part) due to others, which also often leads to exaggerated and anticonservative results. (This, in fact, is the reason why multifactorial regression models should not be summarized with observed means.)

The second approach would be better in that it would be based on predicted, not observed, probabilities and is the logic behind so-called partial dependence statistics/plots (Friedman 2001). However, it seems as if these averages are still suboptimal in how they would not weight predicted probabilities by the frequencies of predictor levels in the data (see Molnar 2018: Section 5.1), thereby – in unbalanced observational data like the present – this might result in upgrading the impact of infrequent combinations and downgrading the impact of frequent ones.

The third approach, while computationally more complex than both previous ones, seems theoretically most sound and is in fact the logic that underlies Fox's (2003) effect plots for regression models where "values of other predictors [i.e. all those not currently being computed/ visualized] are fixed at typical values: for example, a covariate could be fixed at its mean or median [we do not have any here since, for ease of representation we will factorize LENGTHDIFF], a factor at its proportional distribution in the data" (Fox 2003:1); not only is this much more effective than simple observed results, this approach also leads to easier-to-interpret results than regression tables and visualizes intercepts, main effects, and all interactions nicely, which is why we will adopt those plots here. Applied to the genitive alternation, this means that we will – for each combination of predictors of interest (such as VARIETY:GENDER) – inspect the

effects of these combinations on genitives with otherwise typical values (in the sense of typical distribution) of remaining covariates, such as ANIMACY, SIBILANCY, etc.

3 Results

The random forest of conditional inference trees resulting from the above analysis performed well on the data: The OOB prediction accuracy obtained is 84.5%, which is significantly better than a baseline percentage of the more frequent *of*-genitive (75.5%) and the baseline percentage one would arrive at from random proportional guessing (63%); both $p<10^{-44}$. Precision and recall for *s*-genitives are not particularly high (71% and 62.1% respectively), but this is in part due to the class imbalance: precision and recall for the *of*-genitive are much better (88.2% and 91.7% respectively); with a value of 0.909, the *C*-score for the random forest exceeds the standard threshold value of 0.8.

In terms of variable importance, the top 10 *AUC*-based variable importance values are shown in Table 2.

Table 2: *AUC*-based variable importance scores from the conditional random forest with explicitly coded interaction variables

Predictor	Var. imp.	Predictor	Var. imp.
ANIMACY	0.0912863	VARIETY:ANIMACY	0.0736966
GENDER:ANIMACY	0.0612236	VARIETY:GENDER:ANIMACY	0.0518154
LENGTHDIFF	0.0375256	GENDER:LENGTHDIFF	0.0134294
VARIETY:GENDER:LENGTHDIFF	0.0118362	VARIETY:LENGTHDIFF	0.0113675
VARIETY:GENDER:DEFINITENESS	0.0061491	VARIETY:GENDER:SIBILANCY	0.0053940

Before we look at some of these predictors' effects, it is instructive to compare this set of variable importance values to those that result from a conditional random forest fitted without interaction predictors, as shown in Table 3.

Table 3: *AUC*-based variable importance scores from the conditional inference tree forest without explicitly coded interaction variables

Predictor	Var. imp.	Predictor	Var. imp.
ANIMACY	0.2598940	LENGTHDIFF	0.0797186
SIBILANCY	0.0094320	GENDER	0.0087491
DEFINITENESS	0.0073567	VARIETY	0.0062786
MODALITY	0.0043213	SEMRELATION	0.0038330

The way in which this is instructive is that the forest without interaction variables does not really encourage the analyst to explore variable combinations/interactions that the forest with interaction variables clearly ranks really highly. For instance and to use the language of regression analysis, while both rankings put ANIMACY first, suggesting to researchers to explore this as a main effect with, for instance, a visual representation of the type in Figure 1, the ranking of the forest with interaction variables immediately serves to caution against this, given that ANIMACY appears in interactions with other predictors.

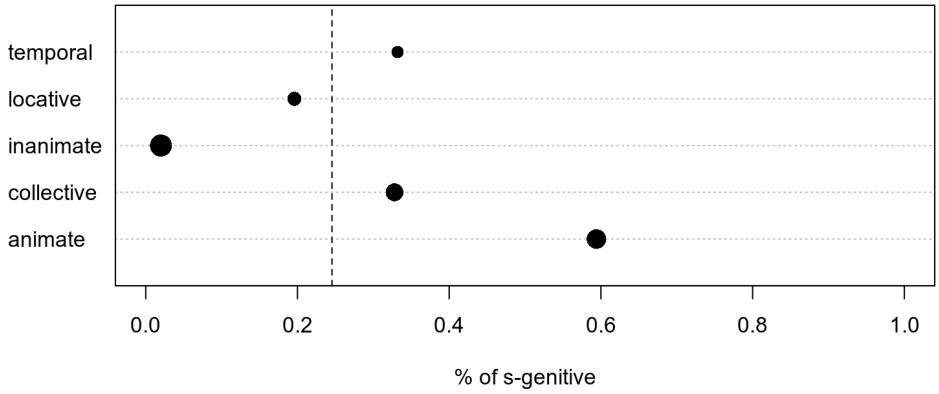

Figure 1: Percentages of GENITIVE: *s* for the levels of ANIMACY (vertical dashed line: overall frequency of *s*-genitives, point sizes are proportional to level frequencies)

This is relevant here because of how the strongest 'main effect' – ANIMACY – but also the main variables of interest in this analysis – VARIETY and GENDER – are all involved in interactions with a high degree of importance. In fact, the interaction predictors either score more highly than the 'main effects' of which they are made up (see e.g. VARIETY and GENDER's main effects are not among the top 10 predictors but they feature in interaction predictors that are) or the interaction predictors immediately follow a main effect predictor (see e.g. LENGTHDIFF). Since it is problematic to analyze a random forest with a single tree fitted on all the data (which, if such a tree was not unproblematic, could reveal interactions in the sequence of splits), exploring interaction variables is, therefore, a possible alternative (we will discuss other alternatives below).

In what follows, we will discuss the following effects: VARIETY:GENDER:ANIMACY (Section 3.1), VARIETY:GENDER:LENGTHDIFF (Section 3.2), VARIETY:GENDER:DEFINITENESS (Section 3.3), and VARIETY:GENDER:SIBILANCY (Section 3.4); the reason we are focusing on these is that these are the predictors with variable importance scores among the top 10 and the ones with the highest order of interactions for every predictor; for instance, ANIMACY has the highest value, but it participates in an interaction in the second most important predictor VARIETY:ANIMACY, but then these two predictors as well as the third most important one, GENDER:ANIMACY, all are involved in the three-way interaction VARIETY:GENDER:ANIMACY, which is therefore the first effect to be discussed.

3.1 The effect of VARIETY:GENDER:ANIMACY

The first interaction is shown in Figure 2. The *y*-axis shows the predicted percentage of *s*-genitives for the combinations of five levels of ANIMACY and two levels of GENDER (abbreviated versions of *male* and *female, unknown* is not shown) shown across the *x*-axis. The two varieties are shown as filled circles: BrE in red and SriLE in blue.

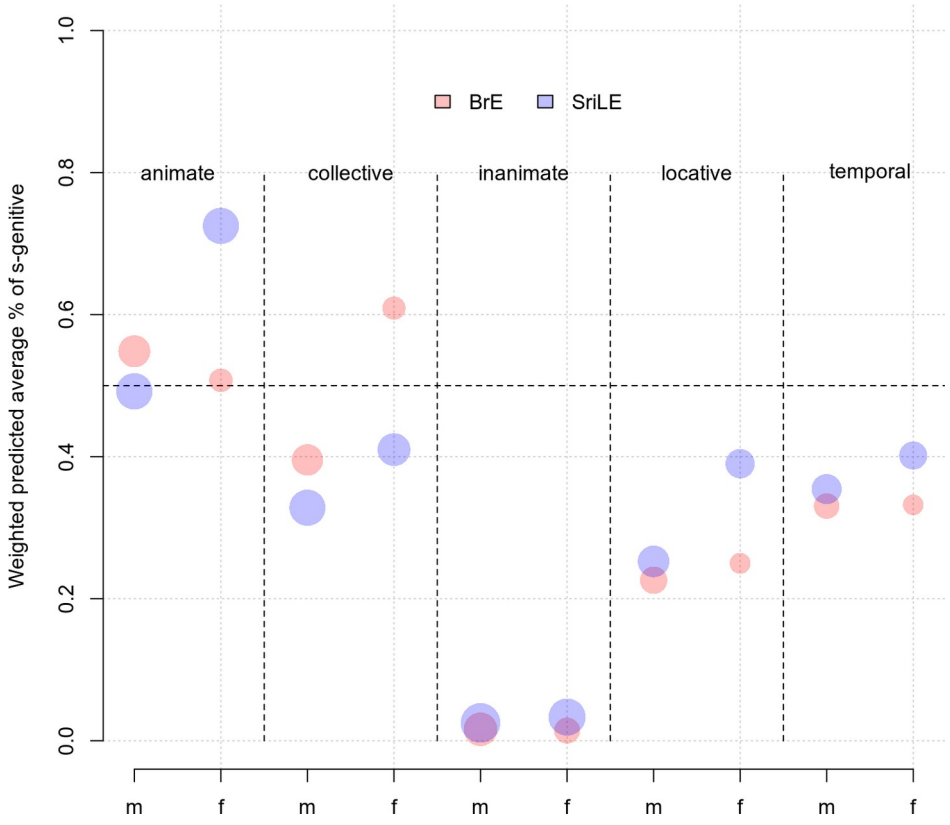

Figure 2: The effect of VARIETY:GENDER:ANIMACY on GENITIVE shows that gender differences are most pronounced for genitive choice with animate possessors

One immediately obvious result is that, with inanimate possessors, there is essentially no difference between varieties and/or genders: *s*-genitives are just very strongly dispreferred (ever so slightly more in BrE). Another fairly strong result is that the interaction is most pronounced for animate possessors: With animate possessors, male speakers of both varieties use the *s*-genitive half the time, but the female SrLE speakers use *s*-genitives much more than the female BrE speakers, who in turn use it less than male speakers; in other words, the female SrLE speakers are exhibiting the 'canonical'/expected patterning more than the female BrE speakers. Finally, we find that collectives behave differently from the other animacy levels: (i) *s*-genitives are rarer in SrLE than in BrE and (ii) compared to the other combinations, there is a very high percentage of *s*-genitive use among female BrE speakers. While this pattern should not be overinterpreted, given the small number of data points for exactly that combination, female BrE speakers seem more advanced in adopting the expansion of the *s*-genitive to possessors that are lower on the animacy scale (see Wolk et al 2013 for a diachronic account).

Another interesting finding is that, in most combinations of predictors, SriLE speakers use *s*-genitives more than BrE speakers: the blue dots are nearly always higher up than the red dots. Also, usually, the differences between the varieties are greater with the female speakers (with a slight exception of inanimate possessors, but with these there is nearly a floor effect anyway). With locative and temporal possessors, the results are not that marked: SriLE speakers produce similarly more *s*-genitives than BrE speakers and this is much more noticeable with the female speakers.

3.2 *The effect of* VARIETY:GENDER:LENGTHDIFF

The second relevant interaction is represented in Figure 3.

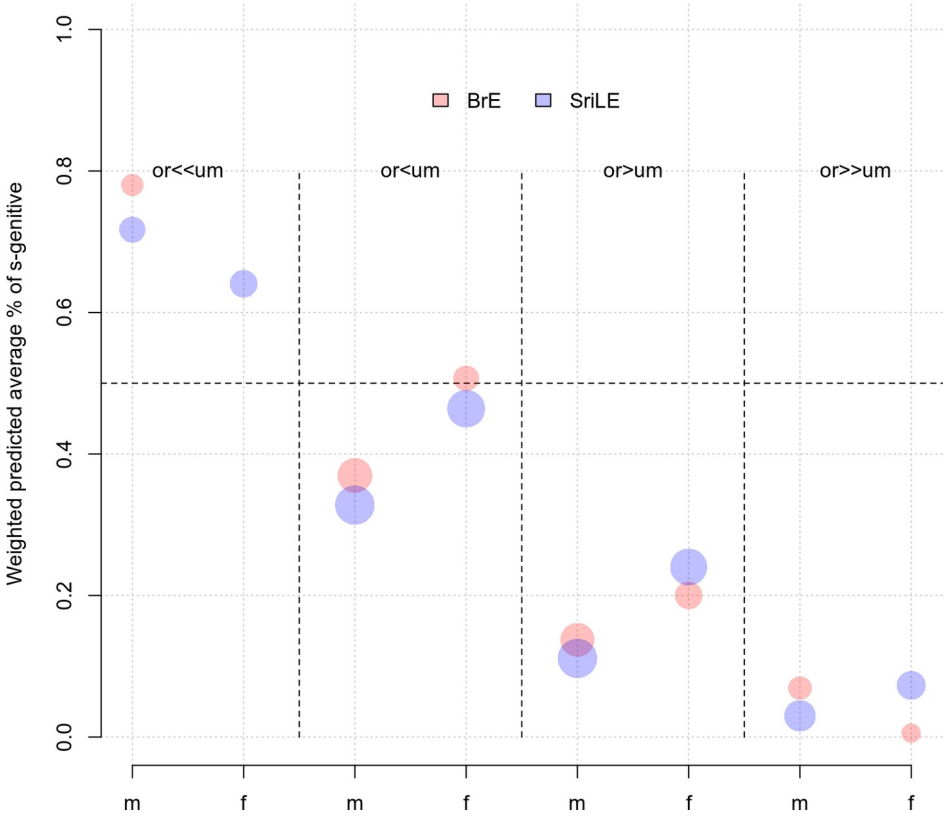

Figure 3: The effect of VARIETY:GENDER:LENGTHDIFF on GENITIVE (LENGTHDIFF is factorized into four levels for expository purposes) is subject to stronger gender differences in SriLE (blue) than in BrE (red)

The most obvious result here is the very strong expected main effect of LENGTHDIFF: As the possessor becomes longer relative to the possessum, the *s*-genitive becomes less and less preferred across all combinations of VARIETY and GENDER. The two outer quarters with the rarer situations of big discrepancies between possessor and possessum lengths show less in terms of differences between genders and varieties, but the middle two quarters of the plot, where most of the cases are located, are more interesting: They show that the differences are small between varieties but more noteworthy between men and women because women simply use many more *s*-genitives than men in both varieties.

3.3 *The effect of* VARIETY:GENDER:DEFINITENESS

The third relevant interaction is represented in Figure 4. We can see a main effect such that definite possessors (left panel) lead to higher numbers of *s*-genitives than indefinite possessors (right panel), as might be expected. Also, there is another main effect such that, with one exception, women use more *s*-genitives than men. However, these main effects are qualified by cross-over effects. First, the one exception of the main effect just mentioned: female BrE speakers use s-genitives less than male BrE speakers, but only with indefinite possessors. Second, the differences between men and women are more pronounced in SrLE than in BrE, and

that is especially true for definite possessors, where female SrLE speakers exhibit a much higher proportion of *s*-genitives than any other combination, which indicates that SriL women seem to react more to the grammatical cue of DEFINITENESS than BrE women: For them, the difference from definite to indefinite is bigger – male speakers show less of an impact there.

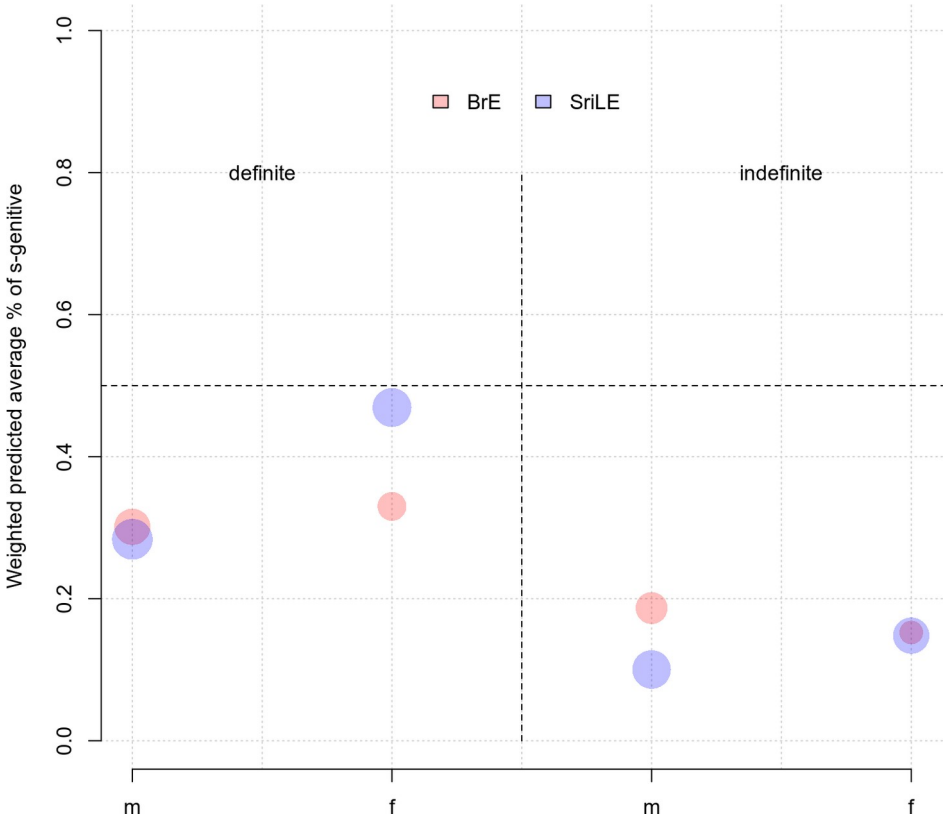

Figure 4: The effect of VARIETY:GENDER:DEFINITENESS on GENITIVE shows that SriLE speakers and females are more sensitive to the definiteness constraint

3.4 The effect of VARIETY:GENDER:SIBILANCY

The final interaction to be discussed is shown in Figure 5. There is the overall expected main effect of SIBILANCY, according to which SIBILANCY: *present* (right panel) should reduce the presence of the *s*-genitive, and if there is a sibilant, then both genders use *s*-genitives correspondingly rarely. However, female speakers react more to sibilancy than men (and especially so in SriLE) and SriLE speakers react more to sibilancy than BrE speakers (and especially so the female speakers).

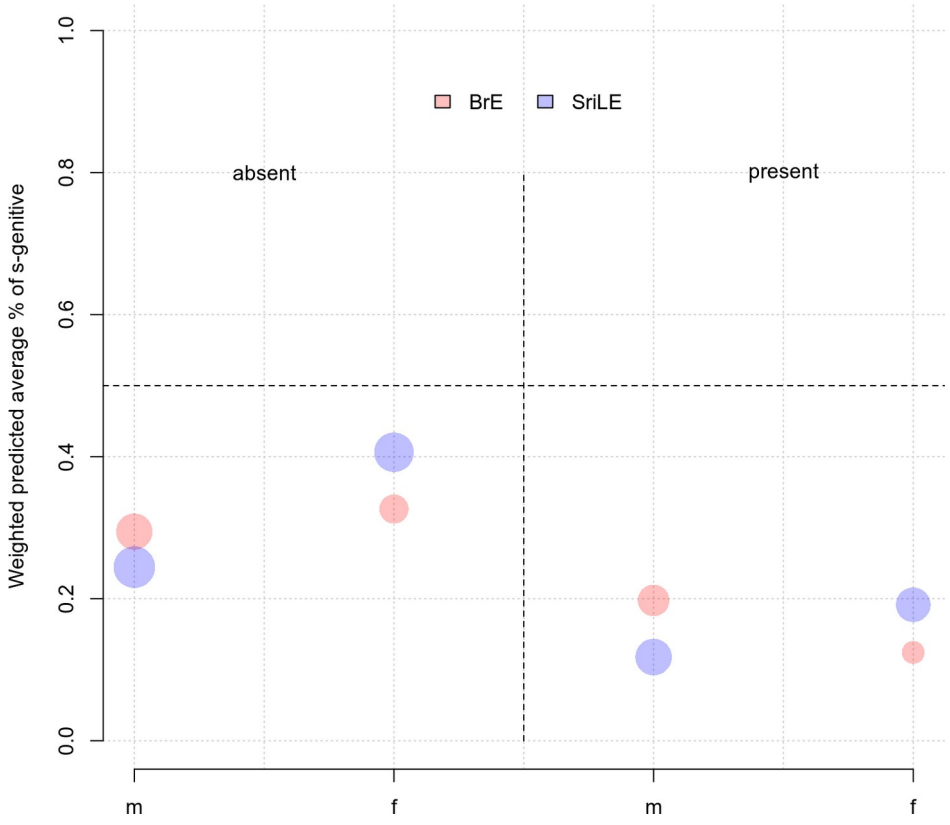

Figure 5: The effect of VARIETY:GENDER:SIBILANCY on GENITIVE shows particular sensitivity of SriLE speakers and females

4 Discussion and concluding remarks

4.1 Implications for the genitive alternation

On the whole and on a general linguistic level, our results are largely compatible with previous studies: we find a strong effect of (possessor) ANIMACY that is in line with previous findings; the same is true of the strong effect of LENGTHDIFF. While weaker, the effects of DEFINITENESS and SIBILANCY do not contradict prior research either. Reassuringly, we find that a factor such as LENGTHDIFF seems to be interacting less with GENDER, which makes sense since one would expect male and female speakers to have a very similar cognitive architecture and, thus, react similarly to the higher degree of processing pressure that arises from high length differences. However, there are also more noteworthy differences between the genders.

The results displayed in Figures 2 to 5 above show that the language-external factors GENDER and VARIETY moderate the effects of language-internal constraints on genitive choice. As stated in the introduction, this conforms to our expectations derived from recent research, which uncovered how VARIETY modulates the effect of ANIMACY (e.g., Heller, Szmrecsanyi, & Grafmiller 2017). However, our results also go against and beyond the expected in that (i) the effect strength of the animacy constraint across the varieties differs from expectations, and (ii) the effect of GENDER on genitive choice had not yet been investigated in a multifactorial design.

Regarding (i), we find that the effect of ANIMACY – the most important language-internal

predictor of genitive choice – more precisely, the difference between animate and inanimate possessors, is stronger in SriLE than in BrE. This is unexpected because recent research (Heller, Szmrecsanyi, & Grafmiller 2017, Heller 2018) found the effect of ANIMACY to be weaker in Outer Circle varieties (i.e., Hong Kong, Indian, Jamaican, Philippine, and Singapore English) than in Inner Circle varieties (i.e., British, Canadian, Irish, and New Zealand English). Since SriLE arguably qualifies as an Outer Circle variety (e.g., Schneider 2011), it was expected that the effect of ANIMACY be weaker in SriLE. Counter-intuitively, however, Figure 2 shows bigger differences between the animate condition and the inanimate condition in SriLE.

Regarding (ii), we present first findings on how GENDER enters the equation: GENDER moderates the effects of length difference and variety and also further qualifies the interactions between VARIETY and ANIMACY (see the previous paragraph) as well as between VARIETY and DEFINITENESS. First, there appears to be a slight gender difference in the effect of length in that males show up as more sensitive to the condition in which the possessum is much longer than the possessor (see the leftmost panel in Figure 3). In this condition, males use the *s*-genitive more frequently. In all other conditions, females seem to be more drawn to the *s*-genitive (see other panels in Figure 3). Closer inspection of these differences reveals that on aggregate, males respond to length difference in a more categorical fashion: while females respond to length difference fairly linearly (i.e., the stronger the cue, the stronger their reaction), males mostly default to using the *of*-genitive, but as soon as possessum length outgrows possessor length by a certain threshold, they prefer the *s*-genitive and, within this extreme range, do so even more than females. GENDER also interacts with VARIETY since most of the time we see that red m/f dots are closer to each other than the blue m/f dots (Figures 2 to 5). Depending on which predictor we interpret as focal, we can describe the observed preference for the *s*-genitive in two ways: either (1) in SriLE, we find a stronger effect of this preference in females, which presumes a stronger gender difference in SriLE; or (2) in females, we find a stronger effect of this preference in SriLE, which presumes that cross-varietal differences emerge more strongly in female language use. In other words, SriL females use the *s*-genitive more than Sri Lankan males, and gender differences are more pronounced in SriLE. Finally, GENDER mediates the VARIETY-ANIMACY and VARIETY-DEFINITENESS interactions: females appear to be more sensitive to both. Once GENDER is taken into account, we see that the stronger effect of ANIMACY in SriLE goes back to female language users only; Sri Lankan males, on the other hand, behave fairly similar to BrE speakers. A similar effect emerges in the possessor definiteness constraint. SriLE-speaking females respond to definiteness more strongly than SriLE-speaking males or speakers of BrE. Thus, Sri Lankan females show higher sensitivity to both animacy and definiteness constraints.

How can we make sense of these patterns? Although there are obvious limitations to our study (e.g., partly exploratory design and limited sample size), it seems reasonable to propose a contact-linguistic explanation of our findings. Since our study covers uncharted territory by focusing on the role of GENDER in the probabilistic grammar of English genitive choice, we cannot inform any World Englishes models (and vice versa) because these models do not make predictions along the lines of GENDER (e.g., models by Kachru, McArthur, Schneider, or more recently by Mair or Buschfeld & Kautzsch). Therefore, and based on previous studies such as Brunner 2014, we turn to a more specific contact-linguistic explanation of the gender difference in the strength of the possessor animacy constraint.

The stronger inclinations of (female) SriLE speakers to use the *s*-genitive might be caused by a transfer of structures in Sinhala, Sri Lanka's most prevalent native language. This transfer might work in two ways – directly and/or in a more abstract way. In Sinhala, the

possessor always precedes the possessum (Chandralal 2010:10), which corresponds to the *s*-genitive. This constituent ordering might carry over directly from Sinhala to English, equivalent to the transfer found by Brunner (2014), who observed that noun phrase modification patterns in Singapore English and Kenyan English correlate with preferences in the countries' respective native languages. There could also be a more abstract transfer of cue strength. Rosenbach (2017) showed that this is indeed the case with genitive choice in the L2 English of Afrikaans speakers. However, the relation between Sinhala and SriLE is different because there is no genitive alternation in Sinhala – the transfer of the animacy constraint could thus only be plausible on a more abstract level. We propose that it might be the high salience of the constraint in Sinhala that carries over to English. Sinhala is special in that it has a different inflectional morphology for animate/inanimate and definite/indefinite nouns (Chandralal 2010:45). The distinctions are thus ubiquitous. Because of the high salience of the animacy and definiteness constraints, speakers could more easily pick up on the constraints in English and also use these constraints in a more categorical fashion, resulting in higher usage frequencies of the *s*-genitive with animate/definite possessors. High salience might also cause an overcorrective use of the "animacy rule" (i.e., use the s-genitive with animate possessors); this tendency to overcorrect might be further enhanced by the perceived high status of English in postcolonial societies (Schneider 2007).

However, it remains unclear why these cross-varietal differences in the effects of animacy and definiteness mostly rely on preferences found in Sri Lankan females. A tentative search for explanations might include societal factors, such as gender equality. According to the United Nations Development Programme (2014), the societies of Sri Lanka and Great Britain are vastly different in terms of labor market participation. In Sri Lanka, only 35% of women participate in the labor market (males: 76.4%), whereas in the UK, 55.7% of women participate (males: 68.8%) (United Nations Development Programme 2014:172-173). Higher participation in the workforce might require more use of English and wider social networks, which might explain the more BrE-like patterns of Sri Lankan males. Sri Lankan women's relative absence in the labor market might further explain why we observe less convergence between the genders than we do in Britain.[2] Lower participation in the workforce, however, is likely to be associated with less use of English and narrower social circles, which might pave the way for more influence from Sinhala in the English of Sri Lankan females.

4.2 Methodological implications

This study has methodological implications as well. First, it is one of the first studies following Gries's (to appear) recommendations regarding classification trees and random forests (see also Deshors & Gries, to appear). To avoid using summarizing a random forest on the basis of a single tree and an interpretation biased towards main effects, we used Gries's new random forest protocol, which involves including interaction predictors and *AUC*-based variable importance scores to determine whether interactions are relevant, too, and which predictors (main effects or interactions) to discuss. We believe that the comparison between Figure 1 and Figure 2 clearly shows that main effects, even if they are the most highly-ranked predictors of the analysis, underestimate, or at least do not at all highlight, the complexity of a data set (and, again, a single tree cannot be used reliably to determine the 'interaction structure' in a whole random forest).

Second, we are the first to implement an alternative to partial dependency scores, which are not yet readily implemented for the experimental forests of conditional inference trees coming with the party package for R, namely an exploration of predicted probabilities that mirror

2 We would like to thank Reviewer 2 for this suggestion.

those of the widely-used effects plots à la Fox (2003) for regression modeling. We think that this is a useful way to proceed given how much using observed percentages can lead to anticonservative overestimates of effects especially in infrequent (combinations of) levels of predictors and in cases with moderate to high collinearity in the data, in which case the effect that a plot returns for one predictor will also contain the effect of many other correlated predictor values.

To see that effect in the present data set, consider Figure 6, which is a somewhat complex extension of Figure 3. Specifically, the filled circles are the points from Figure 3 (with one default size), but, crucially, the red and blue triangles represent the simple observed percentages of *s*-genitives in BrE and SriLE; in other words, the further red and blue triangles are away from red and blue circles respectively, the more the simple observed percentages differ from the effects-type predictions computed here that control for other predictors.

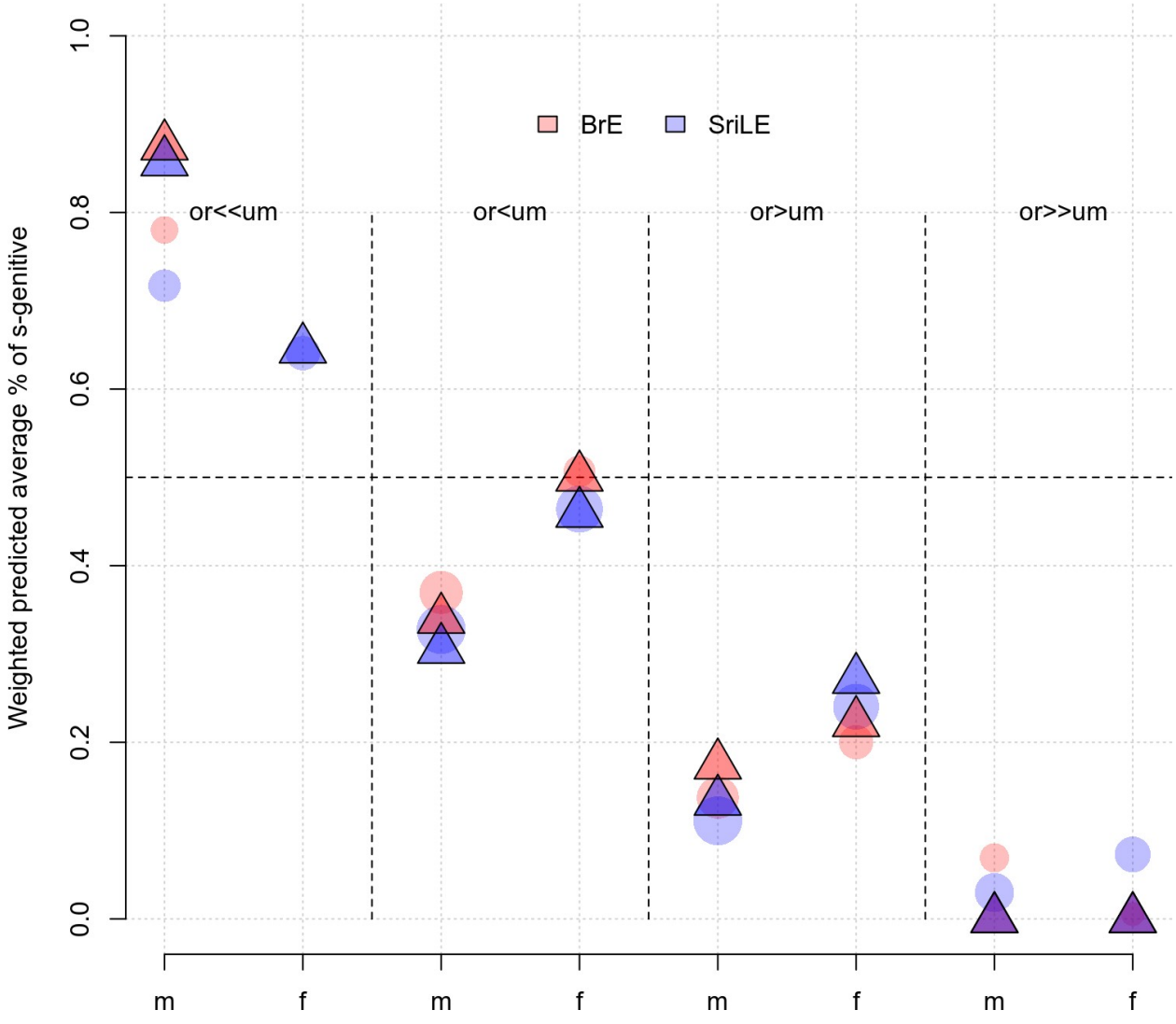

Figure 6: The effect of VARIETY:GENDER:LENGTHDIFF on GENITIVE (circles) vs. observed percentages of *s*-genitives for VARIETY:GENDER:LENGTHDIFF (triangles)

The interpretation is relatively straightforward: In the middle two quarters of the plot, which represent length differences that are fairly frequent in the data, the two estimation methods lead to similar results: the percentages are a bit off and sometimes the ratio of BrE to SriLE is incorrect in the observed percentages display, but the differences are not too dramatic. However, in the outer two quarters of the plot, which represent much fewer cases – because huge discrepancies between possessor and possessum length are rarer – the triangles/observed percentages exaggerate the effects much more.

For instance, in the rightmost quarter, the observed percentages are all 0 whereas the effects predictions are (sometimes quite a bit) higher. It is important to realize that while there are no *s*-genitives when the possessor is much longer than the possessum, which the triangles represent, the triangles/observed percentages are still not a good guide towards understanding the effect of LENGTHDIFF. This is because, for instance, these cases where the possessor is much longer than the possessum also have

- a much higher number of indefinite possessors, which also favor *of*-genitives;
- a much higher number of possessors with final sibilants, which also favor *of*-genitives.

Thus, the triangle positions at y=0 reflect *multiple* variables' effects, not just, like the figure caption would have one believe, (VARIETY:GENDER:)LENGTHDIFF.

The same is true of the or<<um cases on the left, which involve many more definite and non-sibilant-ending possessors than overall so here, too, the triangles/observed percentages overestimate what the graph has a reader attribute to (VARIETY:GENDER:)LENGTHDIFF. Also, the effects-like computation for random forest predictions we pioneer here is, we believe, a nice way of extending a tried-and-true logic from regression modeling to random forests to lead to a better understanding of their effects (one that is comparable to the use of global surrogate models for random forests, see Gries, under revision).

4.3 Where to go from here

While our study is an exercise in World Englishes scholarship, its results do not straightforwardly inform popular models in the field because World Englishes models do not make predictions on our subject matter. This study is concerned with an abstract syntactic alternation and the slight probabilistic differences in its conditioning factors and in particular their interactions with GENDER. Current models of World Englishes do not make predictions about these things; in fact, before our study, there has been hardly any indication that male and female speakers make different genitive choices. In this sense, our findings comply equally with all World Englishes models.

That being said, this study has shown that GENDER is an important determinant of probabilistic grammatical choice-making and we do offer a potential explanation for our strongest finding – the gender difference in the strength of the possessor animacy constraint – by referring to a possible L1 transfer. While L1 transfer, again, is compliant with, and thus does not distinguish between, basically all popular models of World Englishes (e.g., by Kachru, McArthur, Schneider, or more recently by Mair or Buschfeld & Kautzsch), seeing these gender differences is instructive both in and of itself but also in terms of how GENDER qualifies other interactions and in terms of how it may force the analyst to face the sociocultural realities 'on the ground' and how they impact linguistic choices and language change. Further studies of genitive choice should thus do their best to take this influence into account.

In order to understand the role of gender on grammatical alternations more fully, researchers should also look at its effects in a larger scope. This might involve (1) looking at genitive choice in more than two varieties or (2) looking at the effect of gender in other positional alternations such as the dative alternation or the particle placement alternation. To facilitate this, we recommend the use of additional ICE metadata provided by Martin Schweinberger (see above) or Beke Hansen (Hansen 2017).

Further, we see potential of further research in the field of contact-induced probabilistic differences in grammatical alternations across varieties. Although previous studies have suggested a transfer of constituent order (Brunner 2014) and probabilistic weights (Rosenbach 2017), contrasting findings that show opposite patterns exist as well (Heller 2018). Further, the present study suggests that high salience of certain distinctions might already spark probabilistic differences in English, a hypothesis that remains to be tested. This might be achieved by inspecting genitive choice across a range of varieties with differing degrees of salience and different probabilistic weights of predictors like animacy and definiteness in the respective countries' L1s.

Finally, we do feel that the statistical methodology we applied here merits more attention and future use. While in particular mixed-effects modeling has been taking much of linguistics by storm, its applicability to observational data is still often quite difficult so it is understandable that alternatives such as tree-based methods and/or random forests are becoming prominent alternatives. However, as alluded to above, if only briefly, there are scenarios in which the deceptive simplicity of these methods is counter-productive and hides some of the interesting variability in the data – exploring interactions and visualizing effects are extremely rare in random forest studies and we hope to have shown how and why this matters and what the discipline has to gain from such steps; the main effects of many phenomena are already well understood so the ability to add interactions with, for instance, speaker-level effects or other language-external factors is one of the things we need to move things to the next level.

References

Behaghel, O. (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, 25, 110-42.

Bernaisch, T. (2012). Attitudes towards Englishes in Sri Lanka. *World Englishes*, 31(3), 279-91.

Bernaisch, T. (2015). *The Lexis and Lexicogrammar of Sri Lankan English*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Brunner, T. (2014). Structural nativization, typology and complexity: noun phrase structures in British, Kenyan and Singaporean English. *English Language and Linguistics*, 18(1), 23-48.

Chandralal, D. (2010). *Sinhala*. Amsterdam & Philadelphia: John Benjamins.

Deshors, S.C. & St.Th. Gries.(to appear). Mandative subjunctive vs. *should* in world Englishes: A new take on an old alternation. *Corpora*.

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, 8(15), 1-27.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-232

Grafmiller, J. (2014). Variation in English genitives across modality and genres. *English*

Language and Linguistics, 18(3), 471-96.
Gries, St. Th. (to appear). On classification trees and random forests in corpus linguistics: some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory*.
Gunesekera, M. (2005). *The Postcolonial Identity of Sri Lankan English*. Colombo: Vijitha Yapa Publications.
Hansen, B. (2017). The ICE metadata and the study of Hong Kong English. *World Englishes*, 36(3), 471-86.
Heller, B. (2018). Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. Unpublished Ph.D. dissertation, KU Leuven.
Heller, B, T. Bernaisch, & St.Th. Gries. (2017). Empirical perspectives on two potential epicenters: The genitive alternation in Asian Englishes. *ICAME Journal*, 41, 111-44.
Heller, B, B. Szmrecsanyi, & J. Grafmiller. (2017.) Stability and fluidity in syntactic variation world-wide. *Journal of English Linguistics*, 45(1), 3-27.
Janitza, S., C. Strobl, & A.-L. Boulesteix. (2013). An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics*, 14, 119.
Kachru, B.B. 1985. Standards, codification and sociolinguistic realism: the English language in the outer circle. In R. Quirk & H.G. Widdowson, eds., *English in the World: Teaching and Learning the Language and Literatures.* Cambridge: Cambridge University Press, pp. 11-30.
Künstler, V., D. Mendis, & J. Mukherjee. (2009). English in Sri Lanka: Language functions and speaker attitudes. *International Journal of English Studies*, 20(2), 57-74.
Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Philadelphia Press.
Meyler, M. (2007). *A Dictionary of Sri Lankan English.* Colombo: Mirisgala.
Molnar, C. (2018. *Interpretable machine learning: a guide for making black box models explainable*. URL: <https://christophm.github.io/interpretable-ml-book/index.html>, version of 12 Dec 2018.
Mukherjee, J. (2008). Sri Lankan English: Evolutionary status and epicentral influence from Indian English. In K. Stierstorfer, eds., *Proceedings / Anglistentag 2007 Münster.* Trier: Wissenschaftlicher Verlag Trier, pp. 359-68.
Revis, M. & T. Bernaisch. (to appear). Pragmatic nativisation in Asian Englishes? Evidence from filled and unfilled pauses. *World Englishes*.
Rosenbach, A. (2002). *Genitive variation in English: conceptual factors in synchronic and diachronic studies*. Berlin & New York: Mouton de Gruyter.
Rosenbach, A. (2005). Animacy versus weight as determinants of grammatical variation in English. *Language*, 81(3), 613-44.
Rosenbach, A. (2014). English genitive variation – the state of the art. *English Language and Linguistics*, 18(2), 215-62.
Rosenbach, A. (2017). Constraints in contact: animacy in English and Afrikaans genitive variation – a cross-linguistic perspective. *Glossa*, 2(1), 72.
Schneider, E. (2007). *Postcolonial English: varieties around the world*. Cambridge: Cambridge University Press.
Schneider, E. (2011). *English around the world: an introduction*. Cambridge: Cambridge University Press.
Szmrecsanyi, B. (2017). Variationist sociolinguistics and corpus-based variationist linguistics: overlap and crosspollination potential. *Canadian Journal of Linguistics/Revue*

canadienne de linguistique, 62(4), 1-17.
Szmrecsanyi, B. & L. 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, & M. Korhonen, eds., *The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present.* Amsterdam & Philadelphia: John Benjamins, pp. 291-309.
Szmrecsanyi, B., D. Biber, J. Egbert, & K. Franco. (2016). Toward more accountability: Modeling ternary genitive variation in Late Modern English. *Language Variation and Change*, 28(1), 1-29.
Szmrecsanyi, B., J. Grafmiller, J. Bresnan, A. Rosenbach, S.A. Tagliamonte, & S. Todd. (2017). Spoken syntax in a com perspective: the dative and genitive alternation in varieties of English. *Glossa*, 2(1), 86.
Tagliamonte, S.A. & R.H. Baayen. (2012). Models, forests and trees of York English: *Was*/*were* variation as a case study for statistical practice. *Language Variation and Change*, 24(2), 135-78.
United Nations Development Programme. (2014. Human development report 2014): Sustaining human progress: Reducing vulnerabilities and building resilience. Retrieved from http://hdr.undp.org/sites/default/files/hdr14-report-en-1.pdf.
Wolk, C., J. Bresnan, A. Rosenbach, & B. Szmrecsanyi. (2013). Dative and genitive variability in Late Modern English: exploring cross-constructional variation and change. *Diachronica*, 30(3), 382-419.