English Language and Linguistics, 24.2: 413–440. © Cambridge University Press 2019
doi:10.1017/S1360674319000133

Probabilistic indigenization effects at the lexis–syntax interface[1]

IVÁN TAMAREDO
University of Santiago de Compostela
MELANIE RÖTHLISBERGER
University of Zurich
JASON GRAFMILLER
University of Birmingham
BENEDIKT HELLER
Continental AG
(Received 20 December 2018; revised 1 April 2019)

Szmrecsanyi *et al.* (2016) define probabilistic indigenization as the process whereby probabilistic constraints shape variation patterns in different ways, which eventually leads to more heterogeneity in the constraints governing syntactic variation across different varieties of English. The present study extends our knowledge of the heterogeneity of probabilistic grammars by sketching a corpus-based variationist method for calculating the similarity between varieties thereby drawing inspiration from the comparative sociolinguistics literature. Based on linguistic material from the *International Corpus of English*, we ascertain the degree of regional variability of five probabilistic constraints on the genitive, dative, particle placement and subject pronoun omission alternations across three varieties of English, namely British, Indian and Singapore English. Our results indicate that, of the four alternations under study, the genitive alternation is the most homogeneous one from a regional perspective, followed – by increasing order of heterogeneity – by subject pronoun omission, dative and particle placement alternations. On the basis of these findings, we evaluate claims in the literature according to which the extent of probabilistic indigenization is proportional to the lexical specificity of the syntactic phenomenon under study, a hypothesis that is borne out by our data.

Keywords: probabilistic grammar, World Englishes, probabilistic indigenization, alternation-internal homogeneity, lexical specificity

[1] Generous financial support from the following institutions is gratefully acknowledged: Regional Government of Galicia (grants no. ED431B 2017/12 and ED431D 2017/09); Spanish Ministry of Innovation, Science and Universities (grants no. FFI2017-86884-P, FFI2014-52188-P and BES-2015-071233); European Regional Development Fund; and the Research Foundation Flanders (grant no. G.0C59.13N). We would further like to express our gratitude to the editors and copy-editors of *English Language and Linguistics*, and three anonymous reviewers for their helpful suggestions. Thanks are also due to Benedikt Szmrecsanyi and Daniela Pettersson-Traba for their valuable comments on earlier versions of this article. The usual disclaimers apply.

1 Introduction

The present study continues a line of research initiated in Szmrecsanyi *et al.* (2016) which combines the main principles of the probabilistic grammar framework (e.g. Bresnan 2007), which argues that grammatical knowledge has a probabilistic component shaped by speakers' linguistic experience, with work on postcolonial varieties of English (e.g. Schneider 2007). Our overarching goal is to determine how similar or dissimilar the probabilistic knowledge of grammar is on the part of speakers with different regional backgrounds and to assess the extent to which the degree of probabilistic indigenization corresponds to an alternation's lexical specificity (as has been claimed in the literature). More specifically, we propose a corpus-based variationist method for quantifying the extent to which syntactic constraints that influence the choice between competing variants behave homogeneously across varieties of English and compare these results to the degree to which the variation between the competing variants depends on the lexical items that instantiate the constituents (i.e. the alternation's lexical specificity). As a case study, we discuss similarity patterns between three varieties of English around the world, namely British English (BrE), Indian English (IndE) and Singapore English (SgE), in four syntactic alternations that offer speakers a binary choice: the genitive alternation (e.g. Rosenbach 2014), as in (1); the dative alternation (e.g. Bresnan & Hay 2008), illustrated in (2); particle placement (e.g. Gries 2003), as in (3); and subject pronoun omission (e.g. Torres Cacoullos & Travis 2014), exemplified in (4).

(1) (a) The *s*-genitive
[Singapore]$_{\text{possessor}}$'s [small size]$_{\text{possessum}}$ meant it could be quick to respond to changes in economic conditions. (ICE-SIN:W2C-011)
(b) The *of*-genitive
The [size]$_{\text{possessum}}$ of [the eyes]$_{\text{possessor}}$ is to help them at night […]. (ICE-GB:W2B-021)

(2) (a) The ditransitive dative variant
That will give [the panel]$_{\text{recipient}}$ [a chance]$_{\text{theme}}$ to expand on what they've been saying. (ICE-GB:S1B-036)
(b) The prepositional dative variant
[…] and that gives [a chance]$_{\text{theme}}$ [to Bhupathy]$_{\text{recipient}}$ to equalise the points at thirty all. (ICE-IND:S2A:019)

(3) (a) Verb-object-particle (or discontinuous) order
[…] you can just [cut]$_{\text{verb}}$ [the tops]$_{\text{direct object}}$ [off]$_{\text{particle}}$ and leave them. (ICE-GB:S1A-007)
(b) Verb-particle-object (or continuous) order
[Cut]$_{\text{verb}}$ [off]$_{\text{particle}}$ [the flowers]$_{\text{direct object}}$ as they fade. (ICE-CAN:W2B-023)

(4) (a) Overt pronominal subject
The vision$_i$ was not very clear. It$_i$ was murky or rather uh foggy or misty. (ICE-IND: S1B-006)
(b) Omitted pronominal subject
Oh, be4 I forget, "Chitra"$_i$ sends you her love. Ø$_i$ Has been asking about you since you left. (ICE-SIN:W1B-003)

Previous research on probabilistic indigenization effects has largely focused on only three of these alternations (dative, genitive and particle) in a similarly small set of varieties (e.g. Szmrecsanyi *et al.* 2016) or has analyzed variable patterns in one alternation but across

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

several varieties (e.g. Heller *et al.* 2017; Röthlisberger *et al.* 2017; Grafmiller & Szmrecsanyi 2018; Hundt *et al.* to appear). While these studies provide various explanations for probabilistic indigenization effects – drawing on general cognitive processes of language acquisition, language contact and dialect drift – the degree to which the lexical items used in each variant might influence variant choice differently in the alternations has received little attention (but see Röthlisberger *et al.* 2017: 698–9). The variationist approach adopted here has previously been proposed in Grafmiller & Szmrecsanyi (2018) and Szmrecsanyi *et al.* (MS) and is extended in the current study to four syntactic alternations, providing thus a more comprehensive view of morphosyntactic probabilistic indigenization effects than has hitherto been attempted. Overall, our findings suggest that the three varieties we examine share a common probabilistic grammar in all four alternations in that the constraints that influence the outcome of syntactic variation behave, for the most part, in a homogeneous manner across varieties. Probabilistic indigenization effects, however, can be observed to different degrees, largely depending on the lexical specificity of the alternation involved.

The rest of this article is structured as follows. Section 2 summarizes the main theoretical issues which the study addresses, with a focus on the connection between the emergence of cross-varietal probabilistic indigenization effects and the lexical specificity of syntactic alternations. In section 3, we describe the datasets and methods used. Section 4 deals with the results of the study, followed by a discussion of their implications in section 5. Finally, section 6 concludes with some final remarks and suggestions for future research.

2 Theoretical background

This article lies at the interface of two well-known research paradigms. On the one hand, it adheres to the probabilistic grammar framework in that it assumes that grammatical knowledge is partially probabilistic and that multiple constraints operate simultaneously, sometimes with opposite effects, on the alternation between competing variants (e.g. Bresnan 2007; Bresnan *et al.* 2007; Bresnan & Hay 2008; Bresnan & Ford 2010). Research in that spirit has shown that speakers are able to predict, with high accuracy, the odds of finding a particular linguistic variant in a particular context. This, in turn, entails that speakers' grammatical knowledge must necessarily include intuitions about the underlying probabilistic constraints governing linguistic behavior. Bresnan and colleagues further show that grammatical knowledge is gradient and subject to restructuration as a result of changes in speakers' experience with language, which is at least partly dependent on their sociocultural environment. The present study combines this probabilistic viewpoint with an interest in the connection between the structural characteristics of varieties of English and their sociohistorical background, in the spirit of the World Englishes framework (e.g. Schneider 2007; Mesthrie & Bhatt 2008). Crucial to the structural characteristics of varieties of English is the concept of nativization or indigenization. Nativization or indigenization refers to the process whereby speakers of postcolonial varieties make English their own,

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

expressing themselves by means of 'locally characteristic linguistic patterns' (Schneider 2007: 6). Indigenization processes have been claimed to exist mainly at the lexis–syntax interface (Schneider 2003: 249): rather than inventing novel syntactic patterns from scratch, these postcolonial varieties of English are characterized by innovative combinations of lexical items and existing syntactic constructions.

A growing body of literature has recently emerged from the incorporation of the principles of probabilistic grammar into the World Englishes paradigm with the aim of exploring and delimiting the extent to which the strength of probabilistic constraints fluctuates across varieties of English (e.g. Rosenbach 2002, 2003; Hinrichs & Szmrecsanyi 2007; Bresnan & Hay 2008; Szmrecsanyi & Hinrichs 2008; Bresnan & Ford 2010; Bernaisch *et al.* 2014; Szmrecsanyi *et al.* 2016; Heller *et al.* 2017; Röthlisberger *et al.* 2017; Szmrecsanyi *et al.* 2017; among others). Common to most of these studies is the observation that varieties share a fairly robust probabilistic grammar in that the constraints affecting a particular syntactic phenomenon are largely stable across varieties and fuel the same kind of syntactic choices. However, gradient regional differences seem to exist with respect to the strength with which such constraints impact speakers' constructional choices in each variety. For instance, American English (AmE) and BrE speakers differ in that speakers of AmE favor the *s*-genitive over the *of*-genitive more strongly with inanimate possessors and as the length of the possessum increases than BrE speakers (Hinrichs & Szmrecsanyi 2007; Szmrecsanyi & Hinrichs 2008). Similarly, Bresnan & Hay (2008) report that the animacy of the recipient impacts the choice of dative variant more strongly in New Zealand English (NZE) than in AmE, with inanimate recipients being more likely in the ditransitive dative variant in the former than in the latter variety.

In order to refer to these gradient regional differences, Szmrecsanyi *et al.* (2016: 133) extended the notion of indigenization from the World Englishes paradigm to the probabilistic domain and coined the term probabilistic indigenization, which they defined as 'the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties'. Probabilistic indigenization thus refers to a linguistic process that leads to statistical differences across varieties in the effects of probabilistic constraints. Szmrecsanyi *et al.* (2016) argue that divergences in the odds of finding a given syntactic variant in a given context across varieties, even if these patterns are not stable, are evidence of the existence of variety-specific grammars tied to unique sociolinguistic backgrounds. Comparing the effect of probabilistic constraints in three syntactic alternations – the genitive, dative and particle placement alternations – across four varieties (i.e. BrE, Canadian English (CanE), IndE and SgE), they show that the four varieties largely share a common probabilistic grammar, since the effect direction of constraints remains stable across varieties. Nonetheless, quantitative differences emerge with regard to the strength of these constraints. For instance, they observe that the influence of a directional prepositional phrase after the verb phrase on the particle placement alternation was weaker in IndE than in the other varieties they studied. Interestingly, the three alternations in Szmrecsanyi *et al*. (2016) turned out not to be equally sensitive to probabilistic indigenization effects, with the particle placement alternation exhibiting

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

stronger variety-specific patterns than the dative and genitive alternations. Szmrecsanyi *et al.* associate this difference in sensitivity with the lexical specificity of the alternation in question: they conclude their study by suggesting that probabilistic indigenization effects arise as a function of the lexical specificity of syntactic alternations, with those that are strongly connected to specific lexical items being the ones most likely to exhibit cross-varietal indigenization effects. Their argumentation finds support in previous work in World Englishes that has shown that cross-varietal differences mainly emerge at the lexis–syntax interface. Similar tendencies were found by Szmrecsanyi *et al.* (2017), a study of dative and genitive variation in spoken language in four native varieties of English, namely AmE, BrE, CanE and NZE. Their data also shows that syntactic alternations are not equally homogeneous across varieties, with the dative alternation displaying stronger variety effects than the genitive alternation.

In the present article, we extend our previous knowledge of the probabilistic grammars of varieties of English in two ways. First, we propose a corpus-based variationist method for calculating the extent to which syntactic alternations display probabilistic indigenization effects: what counts is not if and/or how often people use particular constructions, but how – that is, subject to which probabilistic constraints – they choose between 'alternate ways of saying "the same" thing' (Labov 1972: 188). Our approach is inspired by the Variation-Based Distance and Similarity (VADIS) method proposed in Szmrecsanyi *et al.* (MS), which assesses the degree of alternation-internal homogeneity across varieties of English along three lines of evidence, as proposed in the comparative sociolinguistics literature (see Poplack & Tagliamonte 2001: 92; Tagliamonte 2002: 731):

1. Statistical significance: Do the same constraints have a statistically significant effect across varieties?
2. Size and direction: Are probabilistic constraints similar with respect to the size and direction of their effects? Are there any constraints that have, e.g. a stronger effect, or an effect in the opposite direction in one variety as compared to rest?
3. Constraint ranking: Is the overall ranking of constraints homogenous across varieties? In other words, do the constraints have the same relative importance in all the varieties considered?

Second, we quantitatively test the hypothesis in Szmrecsanyi *et al.* (2016) according to which an alternation's degree of probabilistic indigenization is proportional to its lexical specificity, that is, the more lexically specific an alternation, the more indigenized it is.

3 Data and method

For the purposes of this article, we investigate variation in the choice of syntactic variants in three varieties of English, namely BrE, IndE and SgE.[2] BrE is an Inner Circle L1

[2] The varieties were ultimately selected for convenience: the subject pronoun omission database only contains instances from BrE, IndE and SgE so, for the purposes of the present study, we restricted our analyses to these three varieties in all alternations.

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

variety, while IndE and SgE are Outer Circle L2 varieties which are considered to have reached phase 4, endonormative stabilization, in Schneider's (2007) Dynamic Model. Despite their similarities regarding typological classification, IndE and SgE differ from each other in that the English in India has been described as being in a 'steady state' in which both progressive and conservative forces are at play (Mukherjee 2007: 158). Moreover, whereas the number of L1 speakers of SgE has been on the rise since the 1980s – in 2010 more than 32 percent of Singaporeans claimed that English was their dominant language (Leimgruber 2013: 9) – L1 users represent only about 0.25 percent of the total number of IndE speakers (Sharma 2012: 523). Therefore, the set of varieties of English studied, despite being restricted in number, represents very different variety types, evolutionary stages, and even more fine-grained distinctions as regards the varieties' social ecologies.

The data for the present study were extracted from the British (ICE-GB), Indian (ICE-IND) and Singaporean (ICE-SIN) components of the *International Corpus of English* (ICE). Each national component in ICE contains 500 texts, which amount to a total of approximately one million words: 600,000 words of spoken material and 400,000 of written material. A useful feature of ICE is that all its components follow the same design and annotation scheme, which makes them particularly appropriate for establishing comparisons between varieties. Interchangeable instances of the genitive, dative, particle placement, and subject pronoun omission alternations were retrieved as follows:

- In the case of the genitive alternation, being a relatively frequent syntactic phenomenon, 10 percent of the texts in ICE was enough for statistical analysis. A sample of texts containing text one, eleven, twenty-one and so on was created and then used to extract tokens of the genitive alternation in an automatic fashion. Appositive genitives, classifying genitives, double genitives, idiomatic genitives, partitive genitives and genitives involving indefinite possessums were excluded to make sure that both variants could have been used interchangeably (see Rosenbach 2002, 2014; Wolk *et al.* 2013). This process yielded 3,108 genitive tokens (for further details on the genitive database, see Heller *et al.* 2017; Heller 2018).
- Instances of the ditransitive and prepositional dative variants were retrieved using a list of dative verbs, shown in (5), adapted from previous literature (Levin 1993; Mukherjee & Hoffmann 2006; Bresnan *et al.* 2007; De Cuypere & Verbeke 2013; Wolk *et al.* 2013). To ensure interchangeability, we excluded tokens involving particle verbs, passivized verbs, elliptical structures, coordinated verbs, clausal or non-overt constituents, beneficiary constructions, constructions containing a spatial goal and idiomatic expressions. In addition, extremely long recipients (more than 18 words) and themes (more than 23 words) were eliminated from the prepositional and ditransitive dative variants respectively. This rendered a database of 3,012 tokens of the dative alternation (for further details on the dative database, see Röthlisberger *et al.* 2017; Röthlisberger 2018).

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

(5) *accord, advise, allocate, allot, allow, answer, appoint, ask, assign, assure, award, bequeath, bid, bring, call, carry, cause, cede, charge, concede, convey, cost, deal, deliver, demonstrate, deny, develop, drop, entrust, explain, extend, feed, flick, flip, forward, get, gift, give, grant, guarantee, hand, impart, inform, issue, keep, lease, leave, lend, loan, lose, mail, name, offer, owe, pass, pay, permit, play, pose, post, prescribe, present, promise, propose, provide, quote, read, recommend, refuse, render, sell, send, serve, set, show, sing, slip, submit, suggest, supply, take, teach, tell, throw, toss, vote, wish, write, yield*

– All verb-particle combinations including a transitive particle verb and one of the ten following particles were extracted from the corpus: *around*, *away*, *back*, *down*, *in*, *off*, *on*, *out*, *over* and *up*. Cases involving passive sentences, sentences with extracted direct objects, modified particles, names, titles, or other fixed expressions, and prepositional verbs were subsequently filtered out. In addition, instances involving pronominal direct objects and direct objects longer than six words were excluded. This process returned 2,480 tokens of the particle placement alternation (for further details on the particle placement database, see Grafmiller & Szmrecsanyi 2018).
– Due to the difficulty of automatically identifying all the relevant instances of omitted subject pronouns in the corpus, these had to be manually extracted. A balanced sample of 40 texts per ICE component was created by randomly selecting 10 spoken informal, 10 spoken formal, 10 written informal and 10 written formal texts, from which 1,229 interchangeable instances of omitted pronominal subjects were retrieved.[3] A random sample of 1,229 overt subject pronouns was then automatically obtained from the same texts, totaling 2,458 instances of both omitted and overt subject pronouns. Interchangeable tokens excluded non-referential omitted/overt subject pronouns, cases in which both the subject pronoun and the auxiliary verb of the clause were dropped, imperative sentences, and overt pronouns in tag questions (for further details on the subject pronoun omission database, see Tamaredo 2018).

The datasets were then annotated for several language-external and language-internal constraints although, for the purposes of the present article, we restricted our analysis to the five most important language-internal predictors of each syntactic alternation. This was a measure to ensure model convergence but, as the findings of Röthlisberger *et al.* (2017) suggest, the most prominent constraints of a syntactic alternation are also the ones most sensitive to probabilistic indigenization effects (see also Grafmiller 2014). Therefore, we can be relatively confident that our set of constraints capture most

[3] Spoken informal texts were selected from the S1A categories in ICE, that is, face-to-face conversations and telephone calls. Spoken formal texts belonged to the categories in S1B: classroom lessons, broadcast discussions, broadcast interviews, parliamentary debates, legal cross-examinations and business transactions. Written informal texts were extracted from the social letter category in W1B. Finally, written formal texts were obtained from the W2A (i.e. academic writing), W2B (i.e. non-academic writing), W2C (i.e. reportage), W2D (i.e. instructional writing) and W2E (i.e. persuasive writing) categories.

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

Table 1. *The five most important predictors of the genitive alternation*

Predictor	Levels
Possessor animacy	Human/animal versus collective versus inanimate versus locative versus temporal possessor phrase
Possessor final sibilancy	Presence versus absence of a sibilant consonant at the end of the possessor phrase
Possessor length	Number of orthographic characters in the possessor phrase
Possessum length	Number of orthographic characters in the possessum phrase
Possessor thematicity	Frequency of the possessor head noun in a text divided by the total number of words in the same text

potential differences between the varieties at hand. The five most important predictors were selected on the basis of per-alternation random forest analyses (see below) fitted to the whole dataset of all three varieties. Language-external constraints, such as register or medium of production, were excluded from this study because they did not consistently show up among the five most important predictors in all alternations. External factors have been shown to vary considerably across varieties of English, as they basically boil down to cultural and social differences (e.g. Heller *et al.* 2017), so including them would have potentially added extra heterogeneity to some alternations but not to others. Tables 1 to 4 display the five probabilistic constraints chosen for each syntactic alternation.

In addition, the data were annotated for the lemmas of the specific lexical items occurring in each of the instances: genitive tokens were annotated for possessor and possessum head nouns, datives for verb lemma, recipient and theme head nouns, tokens of the particle placement alternation for verb lemma, particles and verb-particle combinations, and subject pronouns for the following main verb lemmas.

The degree of alternation-internal homogeneity across the three varieties was estimated in three steps. First, we fitted a mixed-effects binary logistic regression model and a

Table 2. *The five most important predictors of the dative alternation*

Predictor	Levels
Recipient head frequency	Global text frequency of the recipient head, normalized as counts per million words
Recipient person	Local (i.e. first- and second-person pronouns) versus non-local (i.e. third-person pronouns and non-pronominal noun phrases) recipient
Recipient pronominality	Personal/impersonal pronouns versus all other nominal elements
Theme complexity	Simple (i.e. head without postmodification) versus complex (i.e. head with postmodification) theme
Weight ratio	Number of orthographic characters in the recipient phrase divided by the number of orthographic characters in the theme phrase (log transformed)

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

Table 3. *The five most important predictors of the particle placement alternation*

Predictor	Levels
Direct object definiteness	Definite versus non-definite direct object
Direct object length	Number of orthographic characters in the direct object
Particle surprisal	Predictability of the particle given the verb (i.e. the log inverse of the conditional probability of the particle given the verb)
Semantics	Compositional versus non-compositional meaning of the verb-particle combination
Verb surprisal	Predictability of the verb given the particle (i.e. the log inverse of the conditional probability of the verb given the particle)

random forest per variety using the same model formula per alternation. Statistical analyses were carried out in R (R Core Team 2017) using the glmer() function in the lme4 package (Bates *et al*. 2015) for mixed-effects models and the cforest() function in the party package (Hothorn *et al.* 2006; Strobl *et al.* 2007, 2008) for random forests. Both mixed-effects models and random forests seek to predict the choice between variants of a syntactic alternation (e.g. *s*-genitive vs *of*-genitive) given a set of predictors (here: restricted to five) and, as we detail below, they enable us to assess the strength, direction and relative predictive importance of our predictors. While mixed-effects models make their predictions on the basis of a mathematical equation, random forests establish the usefulness of a predictor through trial and error. Mixed-effects models are well-suited for analyzing corpus data of the kind used here because they allow us to take into account the non-independence of our observations via random effects adjustments for, e.g., lexical items or speakers sampled. Mixed models thus provide more reliable generalizations about broader patterns beyond the

Table 4. *The five most important predictors of the subject pronoun omission alternation*

Predictor	Levels
Clause position	Initial versus non-initial position of the subject pronoun in the clause
Clause type	Main versus embedded clause
Coordination	Coordination versus no coordination (i.e. whether the target pronoun is the subject of the second conjunct and coreferential with that of the first one or not)
Pronoun-verb frequency of co-occurrence	Co-occurrence frequency of the target pronoun and the following verb in *GloWbE* (Davies 2013), normalized as counts per million words
Verb class	Class of the verb following the target pronoun, that is, lexical versus non-modal auxiliary (i.e. *be*, *do* and *have*) versus modal auxiliary

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

specific lexical items or speakers observed in our datasets. Random forests, on the other hand, enable us to explore more idiosyncratic patterns within our datasets, e.g. non-linear effects and interactions. Random forests, as implemented in the most common packages, are not well suited to deal with so-called random effects, since they cannot handle categorical predictors with very large numbers of levels.[4] However, an advantage of random forests is that they are quite robust to common issues in linguistic analyses, such as data sparseness or predictor non-linearities (see also Tagliamonte & Baayen 2012: 158–61 for details). The method averages over a defined number of conditional inference trees using random subsampling and a permutation scheme (see Strobl *et al.* 2008 for details).

The computed mixed-effects models included the five most important language-internal predictors per alternation as fixed effects and the lemmas of lexical items as random effects (see Appendix). Due to the low number of some lexical items (and ensuing issues with model convergence), we grouped infrequent items together for the modelling process. For each alternation and random predictor, we identified the frequency value which separated lexical items into two groups: one containing 10 percent of the items, which occurred more often than the selected threshold, and another group containing the remainder 90 percent, which occurred less often than the specified frequency value. Lexical items in the low frequency group were subsequently bundled together, with the exception of particles in the particle placement alternation and verbs in the dative alternation which were not grouped due to their high frequency. No interaction terms were added. We are well aware of the fact that previous research on the alternations that we study often show robust interaction effects, although mostly between language-internal and -external predictors and hardly ever between language-internal predictors. Hence, we decided not to test for interaction effects in order to keep our models simple and because large models with many interactions often lead to serious fitting/convergence problems. Random forest model formulas comprised only the five most relevant constraints per alternation.

In a second step, we calculated the similarity between varieties along the 'three lines of evidence' (Tagliamonte 2002; see section 2) using the method proposed in Szmrecsanyi *et al.* (MS):

1. Statistical significance: the number of shared significant and non-significant constraints in per-variety mixed-effects models (at $p < 0.05$, following Szmrecsanyi *et al.* MS).
2. Relative strength: the inverse of the (Euclidean) distance between the coefficient estimates in per-variety mixed-effects models (calculated without the intercept).
3. Constraint ranking: Spearman's rank correlation coefficient between the predictors' variable importance values obtained from per-variety random forests using the varimpAUC() function in the party package (Strobl *et al.* 2008).

[4] There are some tools in development for computing mixed-effects random forests, but these methods are relatively new and untested (see, e.g., Hajjem *et al.* 2014; Speiser *et al.* 2019).

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

Table 5. *Similarity scores across alternations*

Line	Genitives	Datives	Particles	Subj. omission	Mean
Statistical significance	0.917	0.867	0.733	0.778	0.824
Relative strength	0.894	0.592	0.758	0.839	0.771
Constraint ranking	0.833	0.733	0.633	0.800	0.750
Mean	0.881	0.731	0.708	0.806	0.782

Note that when calculating Euclidean distances using coefficient estimates, a change in the reference and predicted level(s) of the constraints and dependent variable might lead to different results. To overcome this potential issue, we transformed all binary predictors to numbers (e.g. recipient animacy = {animate, inanimate} changed to = {0, 1}) and centered the values around the mean (following Gelman 2008). Furthermore, the reference level and predicted levels were set the same for each alternation. We chose Euclidean distance instead of Spearman's rank correlation to calculate similarity between varieties on the basis of coefficient estimates because the latter does not take into account patterns across the sizes of the predictors' effects, only their relative (absolute) size. Coefficient estimates with the same values but opposite signs would thus be maximally similar using Spearman's rank correlation, while Euclidean distance would recognize the distance between them.

For each of the three lines, we obtained one similarity score for each alternation separately by variety (see tables 6 to 9). On this basis, we calculated a mean similarity score for each line and each alternation averaging across all varieties (see table 5). And lastly, we calculated the mean similarity between the varieties for each alternation as a measure of its overall stability by averaging across all three lines, which we interpret here as reflecting the alternations' degree of probabilistic indigenization across BrE, IndE, and SgE: the lower the value, the more heterogeneous the alternation and thus the greater its degree of probabilistic indigenization.

Finally, the alternations' lexical specificity was operationalized on the basis of the concordance index *C*, which represents how well the model discriminates between the levels of the response variable. In order to tease out the lexical effects from the random structure in the mixed-effects models, we additionally computed fixed-effects only models per variety and alternation by means of the glm() function in R (R Core Team

Table 6. *Per-variety similarity scores of the genitive alternation for each line*

Variety	Statistical significance	Relative strength	Constraint ranking	Mean
BrE	0.938	0.921	0.900	0.920
IndE	0.875	0.861	0.800	0.845
SgE	0.938	0.901	0.800	0.880

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

Table 7. *Per-variety similarity scores of the dative alternation for each line*

Variety	Statistical significance	Relative strength	Constraint ranking	Mean
BrE	0.900	0.545	0.800	0.748
IndE	0.900	0.558	0.600	0.686
SgE	0.800	0.673	0.800	0.758

Table 8. *Per-variety similarity scores of the particle placement alternation for each line*

Variety	Statistical significance	Relative strength	Constraint ranking	Mean
BrE	0.700	0.777	0.650	0.709
IndE	0.800	0.736	0.650	0.729
SgE	0.700	0.759	0.600	0.686

Table 9. *Per-variety similarity scores of the subject pronoun omission alternation for each line*

Variety	Statistical significance	Relative strength	Constraint ranking	Mean
BrE	0.833	0.813	0.850	0.832
IndE	0.833	0.875	0.700	0.803
SgE	0.667	0.829	0.850	0.782

2017). After computing the *C* statistic for both the mixed-effects and fixed-effects model, we subtracted the *C* statistic of the fixed-effect models from the *C* index obtained from mixed-effects models. To calculate *C*, we made use of the somers2() function in the Hmisc package (Harrell 2014). The resulting value indicates the increase in discriminative power from a fixed-effects only model to a model comprising both fixed effects and lexical items as random effects, thus signaling the importance of lexically specific constituents. We also considered an alternative heuristic to quantify lexical specificity by making use of R^2 values. R^2 is a goodness-of-fit statistic which is usually equated to the proportion of variance accounted for by the model: an R^2 value of 1 would correspond to 100 percent of the variance accounted for by the model. An alternation's degree of lexical specificity could hypothetically be operationalized as the increase in R^2 values from a fixed-effects only model to a model with both fixed effects and lexical items as random effects. This would in theory reflect the importance of random effects in the model and, therefore, indicate how strongly associated each alternation is with specific lexical items. However, we refrained from using R^2 as a measure of lexical specificity since its interpretation is not as clear as in linear

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

regression models, where it accounts for the proportion of variance in the response variable that is explained by the predictors (see Levshina 2015: 259). Furthermore, R^2 values are usually lower in logistic regression than in linear regression models, even when they are equivalent in terms of goodness of fit. This is why the concordance index C is commonly reported in logistic regression analysis instead of R^2 (e.g. Hosmer & Lemeshow 2000: 162), and why we chose to rely on lexical specificity values calculated on the basis of the former statistic.

4 Results

Before zooming in on the variety-specific similarity scores per line and alternation (tables 6 to 9), we first take a cross-varietal aggregate perspective. Table 5 displays the values for the averaged similarities across all varieties per alternation and by line of evidence. Means for each alternation across all three lines of evidence are given in the last row, means of each line of evidence are provided in the last column, and a global mean in the bottom right cell. All values range from 0 to 1, with 0 indicating no similarity between varieties and 1 indicating complete overlap. Overall, the numbers suggest that there is a great deal of grammatical homogeneity across the varieties at hand. This is noticeable in the global mean across alternations and lines of evidence (i.e. 0.782), as well as in the individual means for each alternation and line, which all range above 0.700. The proposed similarity between varieties on the probabilistic level is striking: speakers' choices between competing variants seem to be influenced by language-internal constraints that behave very similarly – within each alternation – across varieties irrespective of regional distinctions. Differences do exist, however, between the alternations: looking at the overall mean across all three lines of evidence (last row), the genitive alternation exhibits the highest mean homogeneity (0.881) of the four syntactic phenomena, followed, in increasing order of heterogeneity, by subject pronoun omission (0.806), the dative (0.731) and the particle placement (0.708) alternations. This means that particle placement is the most probabilistically indigenized alternation across the set of varieties studied, closely followed by the dative alternation. On a global level, probabilistic indigenization is mostly driven by the relative importance of predictors, i.e. the constraint ranking, as indicated by the mean value of 0.750, and by relative strength (mean of 0.771 across all alternations). In contrast, statistical significance, i.e. whether or not a predictor is significant in variety A and variety B, adds less to the global heterogeneity across all alternations (mean value of 0.824). Note also that the genitive and dative alternations follow the global pattern in that statistical significance is mostly similar across varieties, while particle placement and subject pronoun omission are most cross-regionally homogeneous with regard to relative strength of the predictors.

Leaving the aggregate perspective in table 5, we now turn to the similarity values for each variety separately across all three lines of evidence (see tables 6 to 9) to provide us with a more fine-grained perspective on alternation-internal differences between varieties.

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

The genitive alternation (table 6) is overall highly homogeneous, with values over 0.800 in all varieties and each line of evidence. The last column of table 6, which contains the mean values per variety across the three lines, reveals that the genitive alternation is more homogeneous in BrE, followed by SgE and, lastly, IndE, where it exhibits the greatest degree of heterogeneity.

In the dative alternation (table 7), the lowest values are found with regard to relative strength (BrE: 0.545, IndE: 0.558, SgE: 0.673) and, particularly in IndE, in the constraint ranking scores (0.600). The mean values across the three lines (last column) reflect a cline of varieties in which the dative alternation is more homogeneous in SgE than in BrE with IndE exhibiting the least homogeneity across all three lines (0.686).

Moving on to particle placement alternation (table 8), the similarity scores indicate more heterogeneity than in the dative or genitive alternation with most values ranging between 0.600 and 0.800. The lowest values, 0.600 and 0.650, are found in constraint ranking. SgE displays the most heterogeneity across all three lines of evidence while IndE is the most homogeneous variety, with BrE occupying an intermediate position.

Finally, the subject pronoun omission alternation is again overall highly homogeneous, with most values exceeding 0.800, except for statistical significance (0.667) and constraint ranking (0.700) scores in SgE and IndE respectively (see table 9). Regarding the varieties' alternation-internal homogeneity, subject pronoun omission is more homogeneous in BrE, followed by IndE and, lastly, SgE.

Next, we averaged the varieties' mean values across all four alternations to calculate the mean cross-alternation homogeneity per variety. Results show that the alternations are most homogeneous in BrE (0.802), with SgE (0.777) and, particularly, IndE (0.766) exhibiting a greater degree of cross-alternation heterogeneity. This finding is consistent with the Inner Circle/Outer Circle and L1/L2 statuses of the varieties: we would expect Outer Circle/L2 varieties to display more probabilistic indigenization effects than Inner Circle/L1 varieties as suggested by the literature (e.g. Grafmiller & Szmrecsanyi 2018, and references therein), and this is in fact what our results seem to indicate.

To follow up on the second main objective of the present study, we next examined the extent to which the degree of probabilistic indigenization reflects an alternation's lexical specificity. If, as Szmrecsanyi *et al.* (2016) suggest, the degree of probabilistic indigenization of a given alternation is proportional to its lexical specificity, we should observe a correspondence between the cline of alternations regarding their homogeneity and the amount of variance accounted for by the lexical effects per alternation from the random effects structure in the per-variety mixed-effects models. Recall that we calculated the lexical specificity of an alternation as the difference in *C*-statistic between those mixed-effects models and fitted fixed-effects models using the same model formula for the fixed effects. To this end, we subtracted the *C* values obtained from fixed-effects-only models from the *C* values of the mixed-effects models. The larger the value, and thus the larger the difference between fixed- and mixed-effects models, the more the random effect structure contributes to the model's discriminative power. Results are shown in table 10: particle placement emerges as the

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

Table 10. *Lexical specificity across alternations and varieties – values indicate difference in C statistic between mixed-effects and fixed-effects models*

Variety	Genitives	Datives	Particles	Subj. omission	Mean
BrE	0.050	0.062	0.066	0.017	0.049
IndE	0.037	0.046	0.059	0.025	0.042
SgE	0.055	0.072	0.069	0.022	0.055
Mean	0.047	0.060	0.065	0.021	

most lexically specific alternation (0.065), followed by the dative (0.060), genitive (0.047), and subject pronoun omission (0.021) alternations.[5]

The distribution of the three varieties as to the alternations' level of lexical embedding according to the models' *C* values is summarized in (6) from the most to the least lexically specific variety.

(6) Genitive alternation: SgE > BrE > IndE
Dative alternation: SgE > BrE > IndE
Particle placement alternation: SgE > BrE > IndE
Subject pronoun omission alternation: IndE > SgE > BrE

The genitive, dative and particle placement alternations are more lexically specific in SgE than in BrE and IndE, while subject pronoun omission is more tightly associated with individual lexical items in IndE than in SgE and BrE. The mean values across alternations (right-most column in table 10) reveal that, overall, the alternations are more lexically specific in SgE (0.055), surpassing both BrE (0.049) and IndE (0.042).

The varieties' cline in probabilistic indigenization (from most to least indigenized) and their cline in lexical specificity (from most to least lexically specific) are shown in (7). The order obtained from the *C*-statistic (7b) almost mirrors the one based on the degree of an alternations' probabilistic indigenization across varieties (7a) with the exception of the genitive and subject pronoun omission alternations whose order is reversed.

(7) (a) Varieties' cline in probabilistic indigenization
particle > dative > pronoun omission > genitive
(b) Varieties cline in lexical specificity according to *C*-statistic
particle > dative > genitive > pronoun omission

The comparison between (7a) and (7b) shows a high degree of overlap between the two clines and thus provides preliminary support for our initial hypothesis, namely that an alternations' degree of probabilistic indigenization is proportional to its lexical

[5] As one reviewer rightly points out, variation as to the increase in *C*-values from fixed-effects to mixed-effects models may be due to differences in sample size or in the type frequency of the lexical items included in the random-effects structures of each variety and alternation. To account for this possibility, we ran a linear regression model with the lexical specificity values in table 10 as dependent variable and sample size and relativized type frequency as predictors. However, neither predictor turned out to be statistically significant, thus suggesting that lexical specificity does not vary as a function of sample size or type frequency.

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

specificity. The implications of these and the rest of the findings described in this section are discussed next.

5 Discussion

The present study has investigated the extent to which an alternation's degree of probabilistic indigenization is proportional to its lexical specificity in a comparison of three varieties of English using a novel approach that applies comparative sociolinguistic methods to compare probabilistic grammars quantitatively. Results show that, overall, the varieties investigated are very homogenous in their alternation-specific probabilistic grammar. Our findings thus support previous claims in the literature that varieties of English are overall grammatically similar, since the same probabilistic constraints tend to influence speakers' constructional choices across varieties. Our results further highlight that this grammatical stability persists across a wide range of syntactic alternations, suggesting that English is indeed syntactically very stable regardless of differences in the regional backgrounds of its speakers and irrespective of whether it is spoken as a first or second language. We should add at this point that including a larger number of varieties and probabilistic constraints – we considered only three varieties and five predictors per alternation – could increase the degree of grammatical heterogeneity observed so far (Szmrecsanyi *et al.* MS). This limitation certainly warrants further investigations in the future.

Despite the overall similarities observed, the four alternations, in line with previous studies, are not equally prone to exhibit probabilistic indigenization effects. Particle placement seems to be more sensitive to probabilistic indigenization effects than the dative alternation, which in turn is more sensitive than the genitive alternation (see, e.g., Szmrecsanyi *et al.* 2016; Szmrecsanyi *et al.* 2017). Subject pronoun omission emerged from our analysis as being a highly homogeneous syntactic alternation, situated between genitives and datives in terms of its cross-varietal stability. This alternation-internal homogeneity is surprising considering that the main substrate languages of IndE and SgE, namely Hindi and Mandarin Chinese respectively, allow the omission of pronouns in subject (and other) positions more frequently and in a wider range of contexts than is commonly assumed to be the case in Standard English (e.g. Kachru 2006: 258–9, for Hindi; Li & Thompson 1989: 657–62, for Mandarin Chinese). The influence of the substrate languages could, hypothetically, have resulted in IndE and SgE manifesting a preference for the omitted pronominal subject variant vis-à-vis BrE. Since the effect size of probabilistic constraints has been shown to be sensitive to language contact in the form of substrate influence (Rosenbach 2017) or second language acquisition effects (Heller *et al.* 2017), we could have observed a weaker effect of predictors favoring the overt variant and a stronger effect of those selecting the omitted variant, or even a change in the direction of the effect of certain constraints in favor of omitted pronouns. However, no such substrate effects were discerned in the present data on subject pronoun omission. To the contrary, the mean similarity score for subject pronoun omission is the second highest (with 0.806) in the comparison.

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

Our results also showed that most differences between the four alternations arise as a function of relative strength and constraint ranking. In other words, alternations do not generally differ with regard to which constraints influence speakers' syntactic choices across the three varieties but more with respect to (a) the extent to which the constraints have an effect on the choice between the variants and (b) the constraints' relative importance. For instance, direct object length is a significant predictor in all three varieties in particle placement, with longer direct objects disfavoring the verb–object–particle order. However, the strength of this effect fluctuates across the varieties as indicated by the mixed-effects models: IndE, with a direct object length coefficient estimate of –3.608, disfavors the discontinuous particle–verb order variant more strongly with each one-letter increase in the length of the direct object phrase than SgE (–2.508) and BrE (–2.091) (see Appendix). Similarly, constraint ranking emerges as a prominent locus of variation in the particle placement alternation. As shown in table 11, all five predictors get a different rank in at least one of the varieties. Moreover, the rankings of three predictors – direct object definiteness, semantics and verb surprisal – are never constant across the three varieties.

Regarding variety-specific patterns, our results indicate that the two Outer Circle/L2 varieties are less homogeneous and more probabilistically indigenized than BrE. Furthermore, IndE displays a lower mean alternation-internal homogeneity score than SgE. A variety's degree of probabilistic indigenization thus seems to correspond directly to its variety type: broadly speaking, L1 varieties are less indigenized than L2 varieties. Note, however, that any generalizations obtained on the basis of only three varieties have to be taken with a pinch of salt and need further substantiation by future studies aggregating over a larger number of varieties and alternation phenomena.

Following Szmrecsanyi *et al.* (2016), we hypothesized to find a correspondence between the extent to which a syntactic alternation exhibits cross-varietal probabilistic indigenization effects – measured as its degree of internal homogeneity across varieties – and its lexical specificity, that is, how strongly associated the alternation is with concrete representations containing specific lexical items. Lexical specificity was operationalized in the present article on the basis of the concordance index *C*, by computing fixed-effects only models and subtracting the *C* statistic from the *C* index obtained from mixed-effects models. The order of alternations as to their lexical specificity almost perfectly matches our hypothesis: particle placement is more

Table 11. *Constraint ranking of five predictors in the particle placement alternation*

Predictor	BrE ranking	IndE ranking	SgE ranking
Direct object definiteness	4	3	5
Direct object length	1	1	2
Particle surprisal	2	2	1
Semantics	3	5	4
Verb surprisal	5	4	3

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

lexically specific than the dative alternation, which in turn is more specific than the genitive and subject pronoun omission alternations. Only the specificity values of the latter two alternations are somewhat inconsistent with the hypothesis in that genitives emerged from the analysis as being slightly more strongly connected with particular lexical items than subject pronoun omission. Therefore, and despite the existence of small inconsistencies, the four syntactic alternations behave largely as we had hypothesized with respect to their sensitivity to probabilistic indigenization effects and their lexical specificity.

With respect to the alternations' degree of lexical specificity in each of the three varieties, there are no discernible patterns owing to Inner Circle/Outer Circle and L1/L2 distinctions. SgE is the variety where alternations are overall most lexically specified, followed by BrE and then IndE. It has been suggested that L2 varieties rely on concrete instantiations of syntactic constructions involving specific lexical items more strongly than other varieties (e.g. Hoffmann 2014: 175–6; Röthlisberger *et al.* 2017), but this does not seem to hold in our data: even though the alternations are indeed more lexically specific in SgE than BrE, this is not the case in IndE compared to BrE. This finding is surprising considering that IndE emerged as being highly lexically specific in those studies and, in particular, as being more specific than SgE and BrE. Note, however, that especially Röthlisberger *et al.* focused on recipients in the dative alternation, while our study includes all lexical items that instantiate a construction and averages across their discriminative power in syntactic choice making. Another crucial difference is the method used to measure the level of lexical embedding of alternations. Whereas we calculated lexical specificity as the increase in discriminative power from a fixed-effects only to a mixed-effects model, Hoffmann (2014) and Röthlisberger *et al.* (2017) did so on the basis of the degree of collostructional strength between lexical items and particular constructions. By way of a somewhat ad hoc explanation, we would like to suggest that different heuristics may provide diverging results as they seem to measure different aspects of lexicality. Also, lexical specificity seems to depend on the construction investigated as shown in our results and in the comparison to Hoffmann (2014). Comparing methodologically mismatching studies, then, can strengthen our understanding of the limitations of measuring lexical specificity across varieties.

6 Conclusion

The aims of the present article were twofold. First, we sought to estimate the extent to which four alternations – the genitive, dative, particle placement and subject pronoun omission alternations – exhibited probabilistic indigenization effects, that is, the occurrence of locally characteristic stochastic patterns of syntactic variation, across three varieties of English, namely BrE, IndE and SgE. We did this by delineating a corpus-based variationist method for quantifying differences in the underlying probabilistic constraints that regulate the choice between competing syntactic variants across varieties. Three lines of evidence, as proposed in the comparative

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

sociolinguistics literature, were considered: the statistical significance of predictors, their relative strength, and the order of constraints as to their relative importance in the alternation-internal grammars of the varieties. The results obtained from the application of this methodology to our data allowed us to arrive at two important conclusions. First, English is on the whole highly syntactically stable as a world language, since there is a great deal of alternation-internal grammatical homogeneity across varieties regardless of regional differences between speakers. Second, probabilistic indigenization effects can be observed to different degrees across syntactic alternations: in our study, particle placement emerged as the most heterogenous alternation, followed, in increasing order of homogeneity, by the dative, subject pronoun omission and genitive alternations. This order coincides with the findings of previous studies, in which particle placement also surpassed datives and genitives in terms of grammatical instability across varieties, thus providing independent validation for the method proposed here.

A second aim of this study was to assess the lexical specificity of the four syntactic phenomena investigated, that is, the strength of the association of each alternation with concrete representations of more abstract schemas involving specific lexical items. To this end, we employed a procedure to quantify an alternation's degree of lexical specificity which relied on the *C* goodness-of-fit statistic and reflected the importance of individual lexical items in order to account for the variance observed in the data. The order of alternations as to their lexical specificity across varieties was (almost) a mirror image of the cline based on their grammatical homogeneity: particle placement turned out to be the most lexically specified alternation, followed, in decreasing order of specificity, by the dative, genitive and subject pronoun omission alternations. Even though further research is still needed to ascertain the most appropriate way of measuring the role of individual lexical items in the choice between competing syntactic variants, our study provides empirical evidence supporting the connection between an alternation's sensitivity to cross-varietal probabilistic indigenization effects and its degree of lexical specificity.

The VADIS-method employed is (still) in an experimental stage and will need further applications to other alternations and datasets, preferably also more complex syntactic phenomena such as verb complementation, alternations with more than two variants (see Gerwin & Röthlisberger, to appear) and alternations on other levels of the grammar, e.g. lexical variation, semantic variation or pragmatic variation. Aggregating then over multiple analyses that capture different parts of speakers' grammar would enable us to paint a more complete picture of variation in probabilistic indigenization. Especially the calculation of Euclidean distances on the basis of coefficient estimates from regression modeling needs further testing regarding concept validity and reliability (see also Heller 2018: 199–204, who tested concept validity, and Röthlisberger 2018: 175, 215–16, who used a bootstrapping procedure to assess reliability). As one reviewer rightly pointed out, the way we tested lexical specificity here ignores the fact that the same character strings might express different meanings depending on the other lexical items used in the variant (e.g. *give back to the*

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

community is different in idiomatic meaning from *give back to my mother*). At the moment, we only measured lexical specificity by focusing on individual lexical items neglecting the wider context of usage. Other useful additional heuristics to assess the importance of lexical constituents would need to be considered in future work, e.g. collostructional analysis (Stefanowitsch & Gries 2003), as applied, e.g. in Röthlisberger *et al.* (2017) to the dative alternation. Such additional methods can provide more data to help us validate the results obtained here and thus to overcome the limitations of the present study with regard to the number of varieties and alternations studied. Furthermore, and despite the fact that our models had an outstanding predictive capacity, it would be desirable to include more than five predictors per alternation to reach a more representative description of the phenomena at hand. Lastly, other methods could be used to compare varieties of English as to their degree of grammatical homogeneity, such as the *Akaike Information Criterion* (Grafmiller & Szmrecsanyi 2018). These and other measures would enable us to be in a better position to delimit the scope of variation within and across varieties of English around the world.

Authors' addresses:

Department of English and German
University of Santiago de Compostela
Avenida de Castelao (Campus Norte)
15782 Santiago de Compostela
Spain
ivan.tamaredo@usc.es

English Department
University of Zurich
Plattenstrasse 47
8032 Zürich
Switzerland
melanie.roethlisberger@es.uzh.ch

Department of English Language and Linguistics
University of Birmingham
Edgbaston
Birmingham B15 2TT
United Kingdom
j.grafmiller@bham.ac.uk

Continental AG
mail@benedikt-heller.de

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

Sources

Davies, Mark. 2013. *Corpus of Global Web-Based English: 1.9 Billion Words from Speakers in 20 Countries (GloWbE)*. https//corpus.byu.edu/glowbe (accessed 11 April 2018).
ICE-GB: *International Corpus of English – The British Component*. www.ice-corpora.net/ice/download (accessed 11 April 2018).
ICE-IND: *International Corpus of English – The Indian Component*. www.ice-corpora.net/ice/download (accessed 11 April 2018).
ICE-SIN: *International Corpus of English – The Singaporean Component*. www.ice-corpora.net/ice/download (accessed 11 April 2018).

References

Bates, Douglas, Martin Maechler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.
Bernaisch, Tobias, Stefan Th. Gries & Joybrato Mukherjee. 2014. The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide* 35(1), 7–31.
Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternfeld (eds.), *Roots: Linguistics in search of its evidential base*, 75–96. Berlin: Mouton de Gruyter.
Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.
Bresnan, Joan & Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118(2), 245–59.
Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1), 168–213.
De Cuypere, Ludovic & Saartje Verbeke. 2013. Dative alternation in Indian English: A corpus-based analysis. *World Englishes* 32(2), 169–84.
Gelman, Andrew. 2008. Scaling regression inputs by two standard deviations. *Statistics in Medicine* 27(15), 2865–73.
Gerwin, Johanna & Melanie Röthlisberger. To appear. Dialectal ditransitive patterns in British English: Weighing sociolinguistic factors against language-internal constraints. In Melanie Röthlisberger, Eva Zehentner & Timothy Colleman (eds.), *Ditransitive constructions in Germanic languages: Diachronic and synchronic aspects* (Studies in Germanic Linguistics). Amsterdam and Philadelphia: John Benjamins.
Grafmiller, Jason. 2014. Variation in English genitives across modality and genres. *English Language and Linguistics* 18(3), 471–96.
Grafmiller, Jason & Benedikt Szmrecsanyi. 2018. Mapping out particle placement in Englishes around the world: A case study in comparative sociolinguistic analysis. *Language Variation and Change* 30(3), 385–412.
Gries, Stefan Th. 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement*. New York: Continuum.
Hajjem, Ahlem, François Bellavance & Denis Larocque. 2014. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation* 84(6), 1313–28.
Harrell Jr, Frank E. 2014. *Hmisc: Harrell miscellaneous*. R Package Version 3.14-6. http://CRAN.R-project.org/package=Hmisc (accessed 17 September 2018).
Heller, Benedikt. 2018. Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. PhD dissertation, KU Leuven.

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

Heller, Benedikt, Benedikt Szmrecsanyi & Jason Grafmiller. 2017. Stability and fluidity in syntactic variation world-wide: The genitive alternation across varieties of English. *Journal of English Linguistics* 45(1), 3–27.
Hinrichs, Lars & 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(3), 437–74.
Hoffmann, Thomas. 2014. The cognitive evolution of Englishes: The role of constructions in the Dynamic Model. In Sarah Buschfeld, Thomas Hoffmann, Magnus Huber & Alexander Kautzsch (eds.), *The evolution of Englishes: The Dynamic Model and beyond*, 160–80. Amsterdam and Philadelphia: John Benjamins.
Hosmer, David W. & Stanley Lemeshow. 2000. *Applied logistic regression*. New York: Wiley.
Hothorn, Torsten, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro & Mark Van Der Laan. 2006. Survival ensembles. *Biostatistics* 7(3), 355–73.
Hundt, Marianne, Melanie Röthlisberger & Elena Seoane. To appear. Predicting voice alternation across academic Englishes. *Corpus Linguistics and Linguistic Theory.*
Kachru, Yamuna. 2006. *Hindi*. Amsterdam and Philadelphia: John Benjamins.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Leimgruber, Jakob R. E. 2013. *Singapore English: Structure, variation, and usage*. Cambridge and New York: Cambridge University Press.
Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.
Levshina, Natalia. 2015. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam and Philadelphia: John Benjamins.
Li, Charles N. & Sandra A. Thompson. 1989. *Mandarin Chinese: A functional reference grammar*. Berkeley, Los Angeles and London: University of California Press.
Mesthrie, Rajend & Rakesh M. Bhatt. 2008. *World Englishes: The study of new linguistic varieties*. Cambridge and New York: Cambridge University Press.
Mukherjee, Joybrato. 2007. Steady states in the evolution of New Englishes: Present-day Indian English as an equilibrium. *Journal of English Linguistics* 35(2), 157–87.
Mukherjee, Joybrato & Sebastian Hoffmann. 2006. Describing verb-complementational profiles of New Englishes: A pilot study of Indian English. *English World-Wide* 27(2), 147–73.
Poplack, Shana & Sali A. Tagliamonte. 2001. *African American English in the diaspora*. Oxford: Blackwell.
R Core Team. 2017. *R: A Language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
Rosenbach, Anette. 2002. *Genitive variation in English: Conceptual factors in synchronic and diachronic studies*. Berlin: Mouton de Gruyter.
Rosenbach, Anette. 2003. Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In Günter Rohdenburg & Britta Mondorf (eds.), *Determinants of grammatical variation in English*, 379–711. Berlin: Mouton de Gruyter.
Rosenbach, Anette. 2014. English genitive variation – The state of the art. *English Language and Linguistics* 18(2), 215–62.
Rosenbach, Anette. 2017. Constraints in contact: Animacy in English and Afrikaans genitive variation – A cross-linguistic perspective. *Glossa: A Journal of General Linguistics* 2(1), 72. 1–21.
Röthlisberger, Melanie. 2018. Regional variation in probabilistic grammars: A multifactorial study of the English dative alternation. PhD dissertation, KU Leuven.
Röthlisberger, Melanie, Jason Grafmiller & Benedikt Szmrecsanyi. 2017. Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics* 28(4), 673–710.
Schneider, Edgar W. 2003. The dynamics of New Englishes: From identity construction to dialect birth. *Language* 79(2), 233–81.

Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the world*. Cambridge: Cambridge University Press.
Sharma, Devyani. 2012. Indian English. In Bernd Kortmann & Kerstin Lunkenheimer (eds.), *The Mouton world atlas of variation in English*, 523–30. Berlin and Boston: Mouton de Gruyter.
Speiser, Jaime Lynn, Bethany J. Wolf, Dongjun Chung, Constantine J. Karvellas, David G. Koch & Valerie L. Durkalski. 2019. BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomes. *Chemometrics and Intelligent Laboratory Systems* 185. doi: 10.1016/j.chemolab.2019.01.002
Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2), 209–43.
Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis & Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8 (25), http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-25 (accessed 17 September 2018).
Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9(307). http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-307 (accessed 17 September 2018).
Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller & Melanie Röthlisberger. 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide* 37(2), 109–37.
Szmrecsanyi, Benedikt, Jason Grafmiller, Joan Bresnan, Anette Rosenbach, Sali A. Tagliamonte & Simon Todd. 2017. Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa: A Journal of General Linguistics* 2(1), 86. 1–27.
Szmrecsanyi, Benedikt, Jason Grafmiller & Laura Rosseel. MS. Variation-based distance and similarity modeling: A case study in World Englishes. Unpublished manuscript.
Szmrecsanyi, Benedikt & Lars Hinrichs. 2008. Probabilistic determinants of genitive variation in spoken and written English: A multivariate comparison across time, space, and genres. In Terttu Nevalainen, Irma Taavisainen, Paivi Pahta & Minna Korhonen (eds.), *The dynamics of linguistic variation: Corpus evidence on English past and present*, 291–309. Amsterdam: John Benjamins.
Tagliamonte, Sali A. 2002. Comparative sociolinguistics. In J. K. Chambers & Natalie Schilling (eds.), *The handbook of language variation and change*, 729–63. Oxford: Wiley-Blackwell.
Tagliamonte, Sali A. & R. Harald 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.
Tamaredo, Iván. 2018. Processing grammatical structures: Morphosyntactic complexity and efficiency in varieties of English around the world, with special reference to pronoun omission. PhD dissertation, University of Santiago de Compostela.
Torres Cacoullos, Rena & Catherine E. Travis. 2014. Prosody, priming and particular constructions: The patterning of English first-person singular subject expression in conversation. *Journal of Pragmatics* 63, 19–34.
Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3), 382–419.

Appendix

Table A1. *Per-variety mixed-effect models of the genitive alternation*

	BrE		IndE		SgE	
Fixed effects	Estimate		Estimate		Estimate	
Intercept	−6.516***		−4.824***		−5.484***	
Possessor animacy: human/animal	5.017***		4.242***		4.657***	
Possessor animacy: collective	3.751***		2.834***		4.046***	
Possessor animacy: locative	1.678**		0.404		2.076***	
Possessor animacy: temporal	3.775***		1.791**		3.768***	
Possessor length	−9.472***		−7.439***		−9.357***	
Possessum length	1.235**		1.150**		3.811***	
Possessor thematicity	0.201		−0.462		0.457	
Possessor final sibilancy	−1.202***		−1.815***		−2.056***	
Random effects	Variance	SD	Variance	SD	Variance	SD
Possessor head	7.551	2.748	2.280	1.510	3.789	1.946
Possessum head	0.892	0.944	0.795	0.891	2.048	1.431
Goodness-of-fit statistics						
C-index	0.957		0.949		0.964	
Proportion of correct predictions	89.2%		91.03%		91.3%	

Legend for tables: * = 0.05, ** = 0.01, *** = 0.001

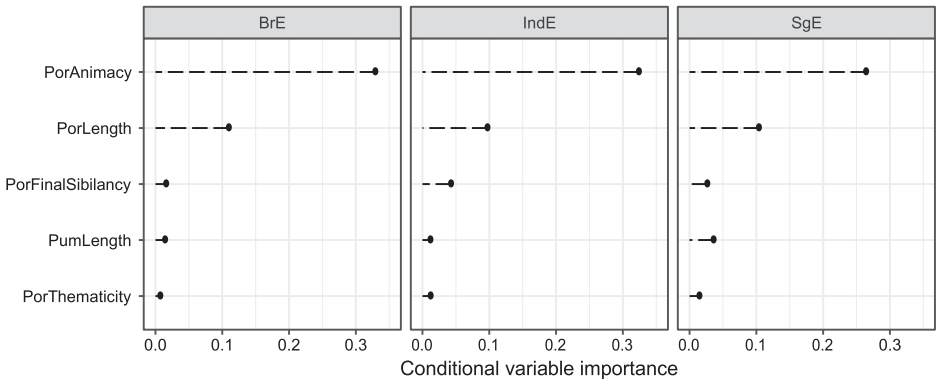

Figure A1. Per-variety random forests of the genitive alternation

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

Table A2. *Goodness-of-fit statistics for per-variety random forests of the genitive alternation*

Goodness-of-fit statistics	BrE	IndE	SgE
C-index	0.943	0.948	0.947
Proportion of correct predictions	87.6%	89.7%	87.7%

Table A3. *Per-variety mixed-effects models of the dative alternation*

	BrE		IndE		SgE	
Fixed effects	Estimate		Estimate		Estimate	
Intercept	−1.936***		0.284		−1.713***	
Weight Ratio	3.302***		2.025***		2.552***	
Recipient pronominality: nominal	1.441**		3.235***		2.174***	
Theme complexity: simple	1.108*		1.567***		1.744***	
Recipient person: non-local	1.480*		1.082*		0.836	
Recipient head frequency	−0.335		0.062		−0.304	
Random effects	Variance	SD	Variance	SD	Variance	SD
Theme head	3.375	1.837	1.553	1.246	3.416	1.848
Recipient head	0.000	0.000	0.000	0.000	0.000	0.000
Verb	3.542	1.882	5.517	2.349	3.980	1.995
Goodness-of-fit statistics						
C-index	0.972		0.972		0.975	
Proportion of correct predictions	93.5%		92.0%		92.5%	

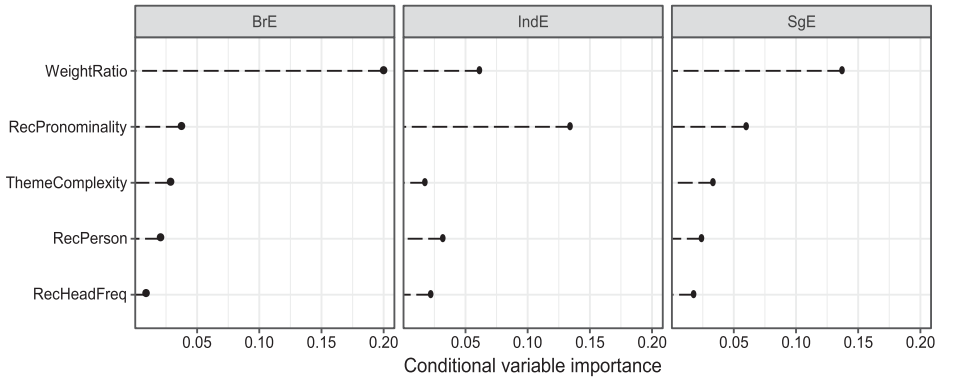

Figure A2. Per-variety random forests of the dative alternation

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

Table A4. *Goodness-of-fit statistics for per-variety random forests of the dative alternation*

Goodness-of-fit statistics	BrE	IndE	SgE
C-index	0.918	0.940	0.914
Proportion of correct predictions	85.5%	86.6%	85.1%

Table A5. *Per-variety mixed-effects models of the particle placement alternation*

	BrE		IndE		SgE	
Fixed effects	Estimate		Estimate		Estimate	
Intercept	−0.950**		−3.080***		−1.667**	
Direct object definiteness	−0.911***		−1.072**		−0.741**	
Direct object length	−2.091***		−3.608***		−2.508***	
Particle surprisal	1.220***		1.087**		1.036**	
Semantics	0.625**		0.060		0.452	
Verb surprisal	0.061		0.189		1.445**	
Random effects	Variance	SD	Variance	SD	Variance	SD
Verb	1.008	1.004	0	0	0.711	0.843
Particle	0.160	0.400	1.167	1.080	1.606	1.267
Verb-particle	0.461	0.679	0.328	0.573	0.672	0.820
Goodness-of-fit statistics						
C-index	0.862		0.913		0.906	
Proportion of correct predictions	77.7%		91.4%		87.4%	

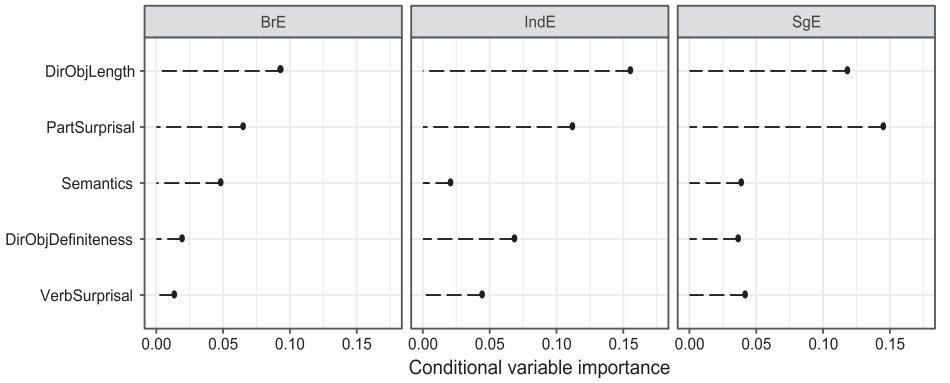

Figure A3. Per-variety random forests of the particle placement alternation

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

Table A6. *Goodness-of-fit statistics for per-variety random forests of the particle placement alternation*

Goodness-of-fit statistics	BrE	IndE	SgE
C-index	0.875	0.927	0.918
Proportion of correct predictions	79.2%	90.2%	85.5%

Table A7. *Per-variety mixed-effects models of the subject pronoun omission alternation*

	BrE		IndE		SgE	
Fixed effects	Estimate		Estimate		Estimate	
Intercept	0.605*		0.429		0.247	
Verb class: non-modal	−0.913**		−1.135**		−0.351	
Verb class: modal	−0.196		−0.690		0.113	
Frequency of co-occurrence	−0.566		−0.001		−0.597*	
Clause type	−1.158**		−0.851*		−1.551***	
Clause position	3.431***		1.773***		1.598***	
Coordination	5.036***		4.409***		3.466***	
Random effects	Variance	SD	Variance	SD	Variance	SD
Verb	0.750	0.866	1.026	1.013	0.475	0.690
Goodness-of-fit statistics						
C-index	0.959		0.948		0.891	
Proportion of correct predictions	88.7%		87.7%		80.4%	

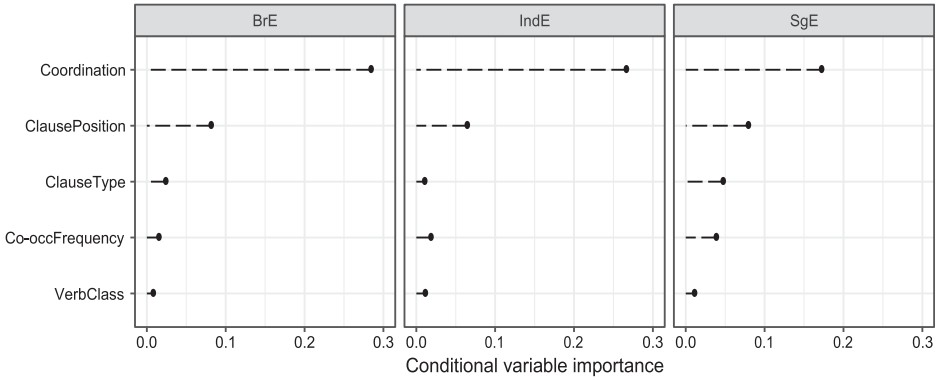

Figure A4. Per-variety random forests of the subject pronoun omission alternation

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press

Table A8. *Goodness-of-fit statistics for per-variety random forests of the subject pronoun omission alternation*

Goodness-of-fit statistics	BrE	IndE	SgE
C-index	0.956	0.938	0.896
Proportion of correct predictions	86.8%	86.6%	80%

https://doi.org/10.1017/S1360674319000133 Published online by Cambridge University Press