Zeitschrift für
SPRACHVARIATION
& SOZIOLINGUISTIK
JOURNAL OF LANGUAGE
VARIATION AND SOCIOLINGUISTICS

https://doi.org/10.5282/jlvs/5

Exploring probabilistic grammar(s) in varieties of English around the world

Benedikt Szmrecsanyi[1], Jason Grafmiller[2], Laura Rosseel[3] and Benedikt Stemmler[4]

[1] Department of Linguistics, KU Leuven, Leuven, Belgium
[2] Department of Linguistics and Communication, University of Birmingham, Birmingham, UK
[3] Brussels Centre for Language Studies, Vrije Universiteit Brussel, Brussels, Belgium
[4] Continental Automotive Technologies GmbH, Frankfurt am Main, Germany

E-mail: benedikt.szmrecsanyi@kuleuven.be

Received 28.09.2024
Accepted for publication 15.01.2025
Published 21.03.2025

Abstract

This paper sketches a range of state-of-the-art quantitative methods to responsibly and rigorously analyze variationist datasets from a comparative perspective. In so doing, we will explore intersections between variationist linguistics and related subfields, such as dialectology and dialect typology, comparative linguistics, probabilistic linguistics, usage-based theoretical linguistics, psycholinguistics, and research on English as a world language. As a case study, we explore three grammatical alternations in nine international varieties of English. Analysis is mostly based on observational corpus data, with supplementary rating task experiments. Key findings include the fact that the probabilistic grammars constraining linguistic variation are overall remarkably homogeneous. With that being said, we often see a split between L1 varieties of English (e.g. British English) and indigenized L2 varieties of English (e.g. Indian English).

Keywords: variation, varieties of English, grammar, syntax, probabilistic grammar, dialectology, dialectometry, psycholinguistics, sociolinguistics

1 Introduction

This paper is a condensed summary (see Szmrecsanyi & Grafmiller 2023 for a less condensed, book-length summary) of a (by now completed) five-year research project entitled “Exploring probabilistic grammar(s) in varieties of English around the world” conducted at KU Leuven about the scope and limits of grammatical variation in a global language such as English. In the present paper (and in the project on which it builds), we adopt the variationist methodology and take a particular interest in how people choose between “alternate ways of saying ‘the same’ thing” (Labov 1972: 188), subject to various probabilistic constraints. In so doing, we break new ground by marrying the spirit of Probabilistic Grammar research (according to which grammatical knowledge is experience-based and partially probabilistic; see Grafmiller et al. 2018) to research along the lines of the English worldwide paradigm (which is concerned with the dialectology and sociolinguistics of postcolonial English-speaking communities around the world – see Schneider 2007).

Our goal in this paper is to explore the plasticity of probabilistic knowledge of English grammar, on the part of language users with diverse regional and cultural backgrounds: how different are the ways a speaker of, say, Canadian English chooses between different ways of saying the same thing (e.g. *he sent the president a letter* vs. *he sent a letter to the president*) from how a speaker of, say, Indian English chooses? For example, we know that long theme constituents (as in *he sent the president* [*a beautifully written and very detailed letter*] as opposed to *he sent the president* [*a letter*]) normally favor the ditransitive variant, but it is conceivable that the exact strength of this effect varies across varieties of English. In other words, we are not primarily interested in frequencies or variant rates (as customary in corpus linguistics) or in feature inventories and/or per cent usage of (non-)standard variants (as customary in classical dialectology and dialectometry), but in the probabilistic conditioning of linguistic variation – and in the extent to which the probabilistic conditioning of variation is different across World Englishes.

To address this issue, we investigate the three grammatical alternations in (1) to (3):

(1) The genitive alternation
  a. *the country's economic crisis*
     (the *s*-genitive variant)
  b. *the economic growth of the country*
     (the *of*-genitive variant)

(2) The dative alternation
  a. *I'd given Heidi my T-Shirt*
     (the ditransitive dative variant)
  b. *I'd given the key to Helen*
     (the prepositional dative variant)

(3) The particle placement alternation
  a. *just cut the tops off*
     (the 'split' verb-object-particle variant)
  b. *cut off the flowers*
     (the 'continuous' verb-particle-object variant)

The variables in (1) to (3) are first and foremost so-called 'permutation alternations' (see Szmrecsanyi & Grafmiller 2023: 18 for discussion): by switching between variants, language users can change the order of possessor and possessum constituents (genitive alternation), of recipient and theme constituents (dative alternation), or of direct object and particle (particle placement alternation). We are thus specifically talking about syntactic alternations. We chose to investigate these specific syntactic alternations because the language-internal constraints governing each of them are well-known (for instance, all three alternations are subject to end-weight effects à la Behaghel 1909 – language users tend to place heavier constituents after shorter constituents), and some of the constraints have been shown to exhibit regional variation in previous small-scale studies (e.g. Bresnan & Hay 2008).

As to diatopic variation, we consider the following nine varieties of English around the world:

1. British English (abbreviated BrE)
2. Canadian English (CanE)
3. Irish English (IrE)
4. New Zealand English (NZE)
5. Hong Kong English (HKE)
6. Indian English (IndE)
7. Jamaican English (JamE)
8. Philippines English (PhlE)
9. Singapore English (SgE)

This selection of varieties is best described as a convenience sample. While we sought to cover a geographically and dialect-typologically wide range of varieties, at the time of study design suitable corpora covering, for example, African varieties of English or US-American English were not available. That said, the above varieties fall into two important dialect-typological groups:

- English as Native Language (ENL) varieties: British, Canadian, Irish, and New Zealand English
- Indigenized English as Second Language (ESL) varieties: Hong Kong, Indian, Jamaican, Philippines, and Singapore English.

This classification is a customary one adopted in well-known reference works (e.g. Kortmann & Lunkenheimer 2013). ENL varieties are "classical" anglophone language varieties. Indigenized ESL varieties, by contrast, are not typically people's first language, but do have important cultural, political, and educational functions in the speech communities in question. We add that the ENL-ESL distinction also translates into the terminology of customary World Englishes models such as Kachru's (1985, 1992) Three Circle Model (where ENL corresponds to the 'Inner Circle' and ESL to the 'Outer Circle'). In the remainder of this paper, we will be mostly talking about the difference between Inner Circle and Outer Circle varieties of English.

On the methodological plane, our analysis will be mostly based upon observational corpus data but will be supplemented behaviorally by rating task experiments.

This paper is structured as follows. Section 2 summarizes our methodology. Section 3 highlights key findings emerging from a classical variationist analysis of the three alternations, one by one, in corpus data. Section 4 engages in variation-based distance and similarity modeling of the corpus datasets. Section 5 reports on the supplementary rating task experiments. Section 6 offers some concluding remarks.

2 Methods and data

In this section, we sketch our methods and data sources. More specific information will be given in the results sections below.

2.1 Observational data

In terms of corpus analysis, this study explores the genitive alternation dataset originally investigated by Heller (2018), the dative dataset examined by Röthlisberger (2018a, 2018b), and the particle placement dataset explored by Grafmiller and Szmrecsanyi (2018) (see Szmrecsanyi & Grafmiller 2023 for a synthesis). Observations were retrieved from the International Corpus of English (ICE) (Greenbaum 1991) and the Corpus of Global Web-based English (GloWbE) (Davies & Fuchs 2015). The ICE and GloWbE materials from which the datasets under study here were generated are register-diversified, with registers that share some situational parameters related to e.g. mode, channel, relationships between participants, interactivity, and processing circumstances. ICE covers spoken dialogues and monologues, as well as written printed and non-printed registers. GloWbE, on the other hand, covers informal blogs and other web-based materials such as newspapers and company websites (see https://www.english-corpora.org/glowbe/). And, crucially, the corpora under analysis here cover nine international varieties of English, as discussed in the Introduction: British English, Irish English, Canadian English, New Zealand English, Jamaican English, Indian English, Hong Kong English, Singapore English, and Phillipine English (with each variety being fairly equally represented in the datasets).

Following best practice in variationist (socio)-linguistics, the datasets contain interchangeable tokens (and interchangeable tokens only). This means that the datasets include hand-coded genitive, dative and particle placement tokens which can be paraphrased by the competing variant with no semantic change. So, for example, (4a) can be paraphrased by (4b), and so would be included in the dataset, but (5a) cannot be paraphrased by (5b), and so would not be included in the dataset.

(4) a. *the speech of the president*
b. *the president's speech*

(5) a. *three liters of wine*
b. ? *wine's three liters*

For reasons of space, we are unable to provide extensive descriptions of the variable contexts in this section, and refer readers instead to the extensive documentation in Heller (2018), Röthlisberger (2018a), and Grafmiller and Szmrecsanyi (2018). What follows is a brief summary:

- The genitive alternation dataset includes all *'s* and *of*-constructions with two NP constituents that do not fall into one of the following categories: appositive genitives, classifying genitives, double genitives, idiomatic/fixed genitives, partitive genitives, and genitives with indefinite possessums. The dataset also does not include *of*-genitives that do not contain a definite possessum.
- The dative alternation dataset includes tokens that were retrieved from the corpus material using a list of dative verbs adapted from previous literature. Constructions that are not inter-

changeable were then weeded out manually – for example, beneficiary constructions (as in *Tom bakes Mary a cake*) were discarded.

- The particle placement dataset was compiled by considering ten particles (*around, away, back, down, in, off, on, out, over, up*) and retrieving instances of these particles co-occurring with a transitive particle verb and a direct object. Non-interchangeable tokens were then discarded. These included transitive particle verbs in which the direct object was a *wh*-form or a relative pronoun, or tokens in which the particle/ preposition occurred with a complement.

After all non-interchangeable tokens were removed from the materials, retaining only truly variable tokens (genitive alternation: N=13,798; dative alternation: N=13,241; particle placement alternation: N=11,340), each observation was annotated, manually or (semi-)-automatically, for a range of known probabilistic constraints on syntactic variation. In the analyses to be presented in this paper, we take into consideration up to eight language-internal constraints, which are listed in Table 1. Again, we refer the reader to Heller (2018), Röthlisberger (2018a), and Grafmiller and Szmrecsanyi (2018) for detailed codebooks. Suffice it to say that the constraints in Table 1 include the "usual suspects" in the literature on the probabilistic conditioning of the alternations in question, such as weight effects, animacy effects, and so on.

2.2 Experimentation

To check the psycholinguistic plausibility of the patterns uncovered in the observational track of the project, we conducted supplementary rating task experiments to tap into language user's introspective preferences. We specifically conducted experiments across four varieties – British English, New Zealand English, Indian English, and Singapore English – in which we set out to replicate contrasts we observed in corpus analysis from these same varieties. The experimental design we followed was the work pioneered by Bresnan (2007) who used the "100 split task" to assess participants' preferences for one or the other variant in a given alternation. In this design, participants are presented with observations of a given alternation sampled directly from corpus datasets, and are asked to distribute 100 points between

Table 1: Lists of language-internal probabilistic constraints under consideration in the present study.

Genitive alternation	Dative alternation	Particle placement alternation
Possessor animacy (Rosenbach 2008)	Log weight ratio between recipient and theme (Röthlisberger 2018a)	Length of the direct object in words (Biber et al. 1999: 932–933)
Possessor length in words (Rosenbach 2014)	Recipient pronominality (Szmrecsanyi et al. 2017)	Definiteness of the direct object (Grafmiller & Szmrecsanyi 2018)
Possessum length in words (Rosenbach 2014)	Theme complexity (Röthlisberger 2018a)	Givenness of the direct object (Chen 1986)
Possessor NP expression type (Heller 2018)	Theme head frequency (Röthlisberger 2018a)	Concreteness of the direct object (Gries 2003)
Final sibilancy in possessor (Zwicky 1987)	Theme pronominality (Röthlisberger 2018a)	Thematicity of the direct object (Grafmiller & Szmrecsanyi 2018)
Previous choice / priming (Hinrichs & Szmrecsanyi 2007)	Theme definiteness (Bresnan & Ford 2010)	Directional modifier (Fraser 1976)
Semantic relation (Rosenbach 2014)	Recipient givenness (Bresnan et al. 2007)	Semantics (Gries 2003)
Possessor head frequency (Heller 2018)	Recipient head frequency (Röthlisberger 2018a)	Surprisal (Grafmiller & Szmrecsanyi 2018)

two competing variants. The test alternations are presented as part of their surrounding context in order to assess the influence of contextual factors and to increase the overall ecological validity of the stimulus items.

Participants were recruited using several different methods. British participants were recruited via the Prolific online recruitment platform (https://prolific.com), but we were unable to recruit sufficient participants from the other three regions with this platform. We therefore used the recruitment services offered by Qualtrics to recruit participants from Singapore and New Zealand. Indian English speaking participants, finally, were recruited by word of mouth through colleagues in contact with large populations of IndE speakers. For each country, we initially recruited 100 participants at random, with the expectation that an unknown percentage would be filtered out based on certain criteria. To ensure that participants were representative of their respective varieties, we included several post-test demographic questions about where participants grew up and lived most of their lives, whether English was their first language, and whether they had taken a linguistics course before. For India and Singapore participants, we also asked additional questions about their use and proficiency in English.

3 Traditional variationist analysis: alternations in corpus data

With a view towards qualitative generalization, this section explores what traditional, corpus-based variationist modeling of the alternations under analysis can tell us about the extent to which users of different varieties of English employ different probabilistic grammars to choose between syntactic variants (see also Szmrecsanyi & Grafmiller 2023: Chap. 5). As to the technicalities, we will take the following measurements:

- Variable importance: What are the most important constraints on variation? Are some constraints more or less important in particular varieties of English? To address these questions, we will use Conditional Random Forest (CRF) modeling (see e.g. Tagliamonte & Baayen 2012). CRF modeling can be used to rank individual constraints per variety and per alternation according to the constraints' overall explanatory importance.
- Effect directions and effect sizes: To determine the extent to which effect directions and effect sizes are stable (i.e. invariant) or unstable (i.e. fluctuating) across varieties of English, we turn to mixed-effects binary logistic regression analysis (Gelman & Hill 2007; Zuur et al. 2009). As the workhorse multivariate analysis tool in variation studies, the technique estimates models that quantify the contributions of individual conditioning factors (by themselves or in interaction). In considering random (i.e. non-repeatable) effects, these quantifications take repeated measures into account and are thus more generalizable than simple fixed-effect models.

In what follows, we will discuss some key take-aways that emerge from the analysis of the above measures.

3.1 Effect directions are stable

There is actually a considerable amount of cross-varietal homogeneity: wherever we look in the data, effect directions are stable. Hence, if a particular factor favors a particular syntactic variant in some variety, it will also favor that same variant in all other varieties. For example, pronominal dative recipients favor the ditransitive dative construction wherever we look, the presence of directional PPs favors the split particle placement variant across the board, and animate genitive possessors favor the *s*-genitive variant throughout.

Consider Figure 1, which illustrates this type of homogeneity based on one particular constraint – possessor animacy – on genitive variation. Figure 1 was generated from a regression model predicting genitive choice across varieties of English. As a partial effects plot, Figure 1 summarizes information from the regression model by systematically altering the values of the constraints under scrutiny while holding the values of all other predictors in the model constant at their default levels. In other words, the plot illustrates probabilistic grammar differences by showing how users of different varieties of English react differently – as measured by variant selection probabilities – to what we may call stimuli, such as animate possessors (as in *<u>Tom</u>'s speech*) as opposed to inanimate possessors (as in *<u>inflation</u>'s* *consequences*). In this spirit, Figure 1 plots the strength of the possessor animacy effect: for each variety, the larger the distance between the light blue dots (representing the probability of selecting the *s*-genitive) and the dark blue dots (representing the probability of selecting the *s*-genitive when the possessor is inanimate), the stronger the possessor animacy effect. In short, the effect size is reflected in the vertical distance between the dots. The point though is that in Figure 1, the light blue dots are consistently located *above* the dark blue dots. This means that no matter what variety of English, language users are more likely to select the *s*-genitive when the possessor animate compared to when it is inanimate (see Heller 2018; Heller, Szmrecsanyi & Grafmiller 2017 for more discussion). This kind of qualitative stability is the typical pattern we see in the data, and can be interpreted as evidence for a solid, supra-regional, qualitative "common core" (Quirk et al. 1985: 16) of the probabilistic grammar of English.

3.2 Effect strength and variable importance differs

While as we saw in the previous section probabilistic grammars are qualitatively very similar, we do find quantitative differences with regard to the effect size and the variable importance of constraints on variation. As to effect sizes, consider again Figure 1: in CanE, animate possessors come with a predicted probability >40% for the *s*-genitive (vis-à-vis <10% predicted probability with inanimate possessors), but in PhlE, animate possessors are associated with a predicted probability of approx. only 20% for the *s*-genitive (vis-à-vis <10% predicted probability with inanimate possessors). This is just one example of the effect size-differentials we often see in the data.

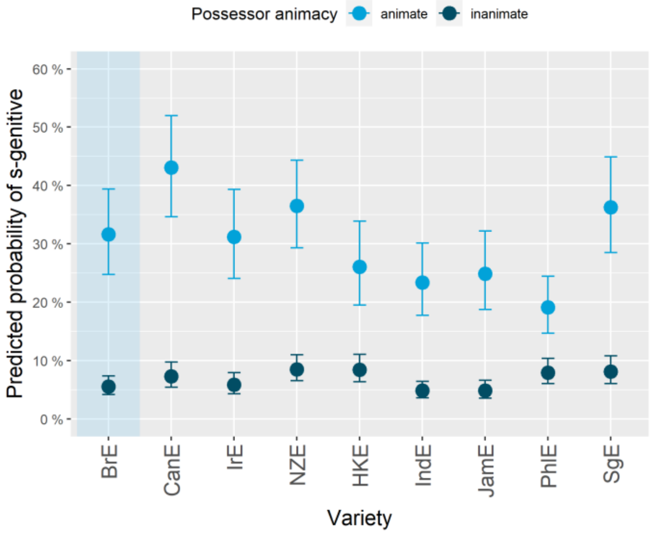

Figure 1: Partial effects plot about the effect of possessor animacy across varieties of English in logistic regression. Predicted probabilities (vertical axis, expressed in percent) are for the *s*-genitive. Vertical distance between dots is proportional to effect size. Varieties to the left are Inner Circle varieties, varieties to the right are Outer Circle varieties. (Adapted from Szmrecsanyi & Grafmiller 2023: Fig. 5.3)

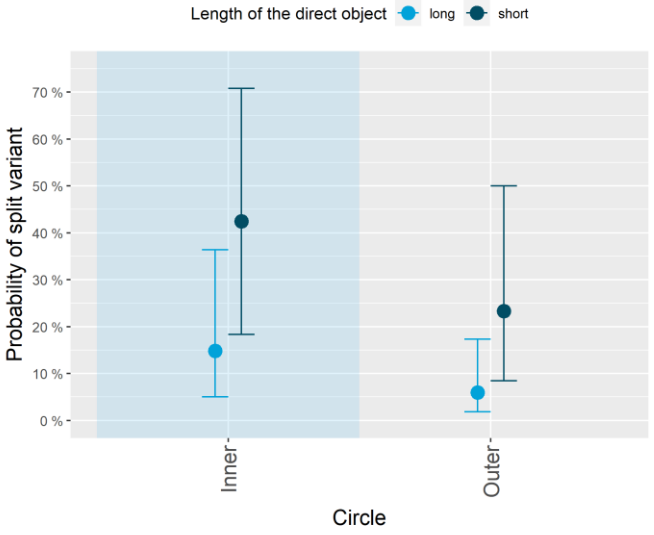

Figure 2: Partial effects plot about the effect of direct object length across varieties of English in logistic regression. Predicted probabilities (vertical axis, expressed in percent) are for the split variant. Vertical distance between dots is proportional to effect size. (Adapted from Szmrecsanyi & Grafmiller 2023: Fig. 5.13)

The language-internal constraints involved in these differentials cover all conceivable domains: noun class (animacy – one of the rather "usual" suspects for regional variability), expression type (pronominality), phonetic form, semantics – and even end weight effects. The Principle of End Weight (Behaghel 1909) predicts that in VO languages such as English, language users have a preference for placing longer, heavier constituents after shorter, lighter constituents. Take the particle placement alternation: according to the literature (see e.g. Gries 2003), long direct objects disfavor the split variant (as in *cut* *the flowers* *off*) and favor the continuous variant (as in *cut off* *the beautiful red flowers*) because the continuous variant will place the long direct object after the short particle. Now, the partial effects plot in Figure 2 shows that indeed, long direct objects disfavor the split variant across the board, both in "Inner Circle" (parlance of Kachru 1992) L1 varieties of English, such as BrE, and in "Outer Circle" indigenized L2 varieties of English, such as IndE. However, Figure 2 also demonstrates that the end weight effect is somewhat less powerful in Outer Circle varieties of English than in Inner Circle varieties of English. That end weight effects are variable regionally is interesting because they are sometimes argued (e.g. Hawkins 1994) to be rooted in the architecture of the human speech processing system, and should against this backdrop not be particularly variable across varieties. In short, there does not seem to be a pattern. Effect sizes are regionally malleable, regardless of the constraints involved.

Given that effect strengths are variable, it is perhaps not surprising that variable importance also differs across varieties of English. Figure 3, for example, ranks constraints (for descriptions of the language-internal constraints, see Table 1) on the dative alternation according to variable importance. There is some variance

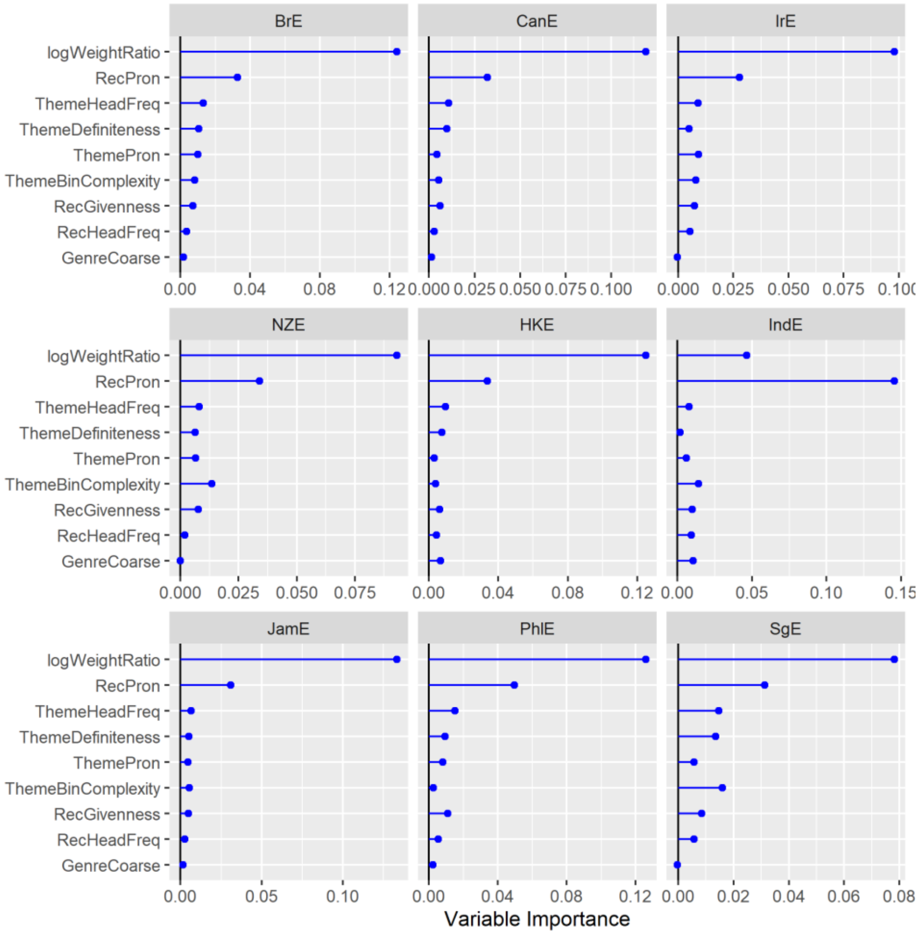

Figure 3: CRF permutation variable importance ranking of constraints on the dative alternation by variety of English. (Adapted from Szmrecsanyi & Grafmiller 2023: Fig. 5.6)

here. Specifically, the four Inner Circle varieties (BrE, CanE, IrE, and NZE) are fairly homogeneous: logWeightRatio (i.e. end weight) is consistently ranked as the most important constraint, and RecPron (i.e. recipient pronominality) as the second most important constraint. Genre differences (label: GenreCoarse) do not play an important role in Inner Circle varieties. More often than not logWeightRatio is also the highest-ranked constraint in the Outer Circle varieties under study, with the exception of IndE where RecPron is substantially more important than logWeightRatio. Another difference among the Outer Circle varieties concerns GenreCoarse: as in the Inner Circle varieties, genre differences are negligible in JamE, PhlE, and SgE, though they turn out to be a bit more important in HKE and IndE.

3.3 All alternations are not equal

Lastly, it is important to note that the three alternations subject to study differ as to how amenable they are to regional differences, or to "probabilistic indigenization", which we have defined elsewhere as follows:

> […] the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties. To the extent that patterns of variation in a new variety A, e.g. the probability of item x in context y, can be shown to differ from those of the mother variety, we can say that the new pattern represents a novel, if gradient, development in the grammar of A. These patterns need not be consistent or stable (especially in the early stages of nativization), but they nonetheless reflect the emergence of a unique, region-specific grammar. (Szmrecsanyi et al. 2016: 133)[1]

This idea of probabilistic indigenization draws inspiration from the observation that lexico-grammar is a prime target of early-stage indigenization (Schneider 2003: 249), based on which Szmrecsanyi et al. (2016: 133) formulate the following prediction: "the more tightly associated a given syntactic alternation is with concrete instantiations involving specific lexical items […] the more likely it is to exhibit cross-varietal indigenization effects." (Szmrecsanyi et al. 2016: 133)

Of the three alternations under study here, the particle placement alternation is the one that is lexically most specific, as each token involves a lexically specific particle verb plus a lexically specific particle. The genitive alternation, by contrast, is abstract and non-specific lexically. The dative alternation takes the middle road, being a constituent order alternation as well but involving different, lexically specific dative verb lemmas. Analysis of the importance of the predictor 'variety of English' in overall CRF models predicting variant choice shows that the particle placement alternation is indeed more amenable to probabilistic indigenization effects than the other alternations. We will return to this issue in the next section.

4 From alternations to distances and similarities

In the previous section, we adopted what we would like to call a jeweler's eye perspective to engage in a fairly fine-grained analysis of the corpus data, with the aim of drawing qualitative generalizations. In this section, we adopt a bird's eye-perspective designed to complement the jeweler's eye perspective. Specifically, we will utilize Variation-based Distance and Similarity Modeling (VADIS for short) (see Szmrecsanyi et al. 2019; Szmrecsanyi & Grafmiller 2023: Chap. 6 for introductory work) to calculate distances and similarities between dialects and varieties. The basic idea behind VADIS is that distances between lects are proportional to the extent to which probabilistic grammars regulating linguistic choice-making are different. In short, then, VADIS marries the variationist modeling of linguistic variation in the spirit of e.g. Labov (1966) to dialectometric distance calculation à la Goebl (1982) and Nerbonne et al. (1999).

Let us illustrate with the dative alternation: *Tom gave me flowers* (the ditransitive dative variant) versus *Tom gave flowers to me* (the prepositional dative variant). An important study by Bresnan et al. (2007) shows that according to regression analysis, the dative alternation in English is conditioned by more than ten language-internal probabilistic constraints. What follows is a simplified version of the dative alternation model formula ("Model A") in Bresnan et al. (2007: Fig. 4):

Probability{prepositional dative} =

[…]

+0.99{accessibility of recipient = nongiven}

−1.1{accessibility of theme = nongiven}

+1.2{pronominality of recipient = nonpronoun}

−1.2{pronominality of theme = nonpronoun}

+0.85{definiteness of recipient = indefinite}

−1.4{definiteness of theme = indefinite}

+2.5{animacy of recipient = inanimate}

+0.48{person of recipient = nonlocal}

−0.03{number of recipient = plural}

+0.5{number of theme = plural}

−0.46{concreteness of theme = nonconcrete}

−1.1{parallelism = 1}

−1.2 length difference (log scale)

[…]

The numbers in the formula are regression coefficients: positive coefficients indicate constraints that favor the prepositional dative variant, negative coefficients indicate constraints that disfavor the prepositional dative construction (and so favor the ditransitive dative variant). For example, nongiven themes (i.e. themes not mentioned in the recent discourse) favor the prepositional variant (+0.99), while nongiven recipients disfavor the prepositional variant (-1.1). The size of the coefficients is proportional to effect size: for example, the effect of inanimate recipients (+2.5) has roughly twice the effect size of nonpronominal recipients (+1.2). In short, the formula above is a blueprint of the probabilistic grammar that regulates the dative alternation in Bresnan et al.'s dataset.

Crucially, however, Bresnan et al. (2007) calculated the above formula based on data from the Switchboard Corpus of spoken US English (Godfrey et al. 1992). We are thus possibly dealing with a variety-specific formula. The question then arises: What is the extent to which we have to adapt the formula as we switch from US American English to, say, New Zealand English? This is precisely how VADIS measures distances between varieties and dialects: variety differences are defined as being proportional to probabilistic grammar differences. In other words, VADIS draws on the variationist methodology to quantify (dis)similarities between lects.

4.1 Technicalities

Technically speaking, VADIS builds on methods developed in comparative sociolinguistics (Tagliamonte 2012: 162–173), which is a sub-discipline in variationist sociolinguistics that evaluates the relatedness between varieties and dialects based on how similar the conditioning of variation is in these varieties. Comparative sociolinguists investigate three "lines of evidence" to determine relatedness:

1. Are the same constraints significant across varieties?
2. Do the constraints have the same strength across varieties?
3. Is the constraint hierarchy similar?

Similarity according to these lines of evidence is often interpreted as historical and genetic relatedness. VADIS draws inspiration from this literature and adapts the comparative sociolinguistics method so that it can be scaled up to the study of more than a couple of lects, and to more than one variable phenomenon at a time.

Figure 4 sketches the different steps necessary to conduct a VADIS analysis. For more details and discussion, we refer the reader to Szmrecsanyi et al. (2019) and to Szmrecsanyi and Grafmiller (2023: Chap. 6). Suffice it to say that the analysis presented in this section considers the various constraints listed in Table 1 and submits these to multivariate modeling, both regression and CRF. The output of this multivariate modeling is subsequently used as input to distance/similarity calculations. Let us very briefly illustrate the basic idea. In Step 3 of the VADIS workflow (see Figure 4), we determine – based on output from regression analysis – cross-variety similarity according to predictor significance. We specifically define the probabilistic distance between two varieties as being proportional to the extent to which the varieties do not overlap with regard to which constraints significantly regulate variant choice. Consider two hypothetical lects A and B and five constraints a–e which regulate some variation phenomenon, as shown in Table 2.

Table 2: Hypothetical configuration of Lects and Constraints.

	Lect A	Lect B
Constraint a	significant	significant
Constraint b	significant	not significant
Constraint c	not significant	significant
Constraint d	not significant	not significant
Constraint e	significant	significant

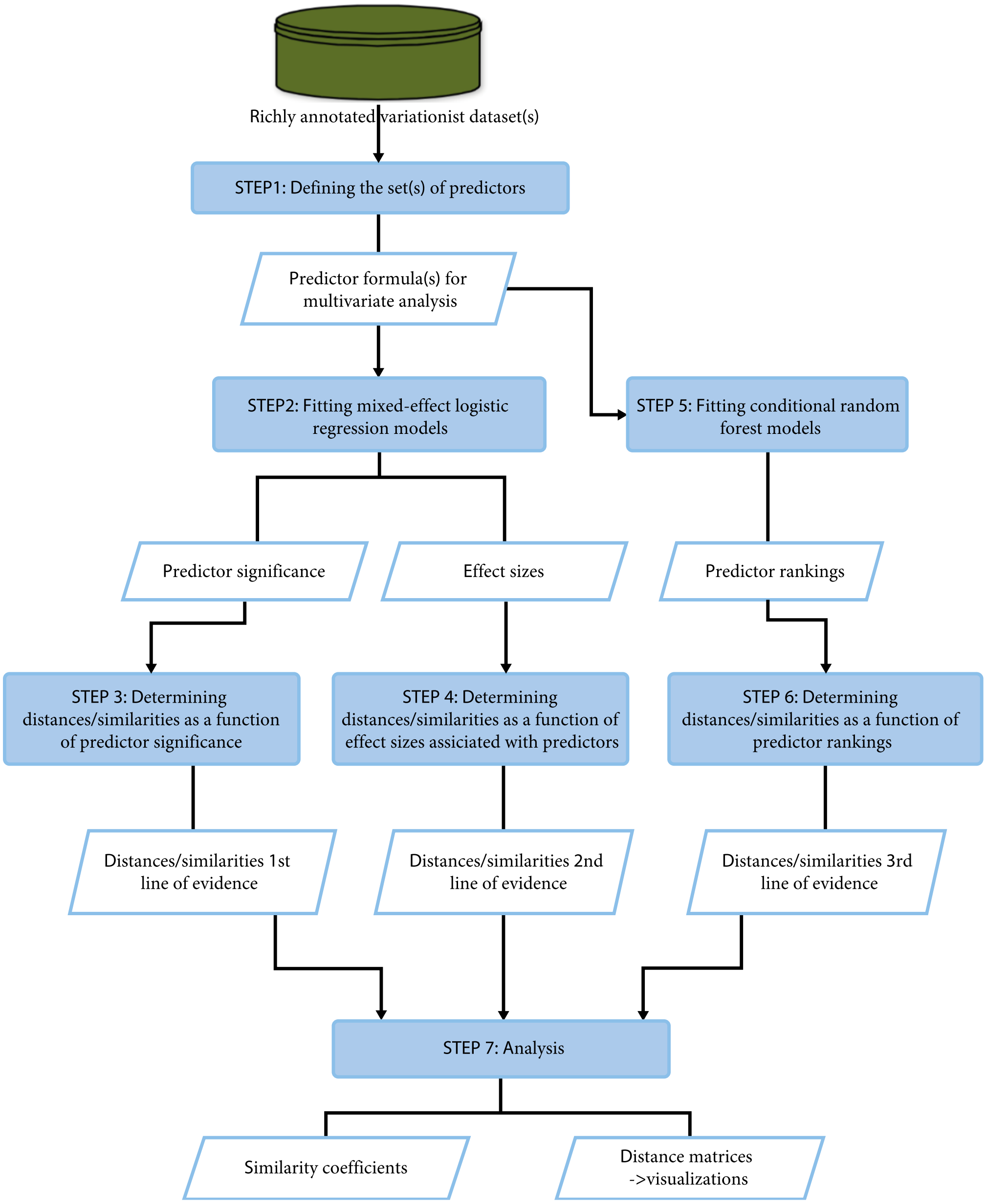
Richly annotated variationist dataset(s)
STEP1: Defining the set(s) of predictors
Predictor formula(s) for multivariate analysis
STEP2: Fitting mixed-effect logistic regression models
STEP 5: Fitting conditional random forest models
Predictor significance
Effect sizes
Predictor rankings
STEP 3: Determining distances/similarities as a function of predictor significance
STEP 4: Determining distances/similarities as a function of effect sizes assiciated with predictors
STEP 6: Determining distances/similarities as a function of predictor rankings
Distances/similarities 1st line of evidence
Distances/similarities 2nd line of evidence
Distances/similarities 3rd line of evidence
STEP 7: Analysis
Similarity coefficients
Distance matrices ->visualizations

Figure 4: The VADIS workflow. (Reprinted, with permission from John Benjamins Publishing Company, from Zhang & Szmrecsanyi 2024: Fig. 1)

Assume that as per regression modeling, lect A and B “agree” on the significance of three constraints (a, d, e), and disagree with regard to two constraints (b, c). Using the squared Euclidiean distance measure, the distance between the two lects is therefore two out of five squared Euclidean points, or 0.4 with a corresponding similarity value of 1 − 0.4 = 0.6. Distance/similarity calculation works in a similar fashion for the other lines of evidence.

An R package (under development) which performs all of the calculations necessary for a VADIS analysis is available from https://github.com/jasongraf1/VADIS. For more discussion of the technicalities, see Szmrecsanyi and Grafmiller (2023: Chap. 6). The book comes with commented scripts and datasets to facilitate replication.

4.2. Similarity coefficients

One way in which VADIS can explore relationships between varieties consists of calculating ‘similarity coefficients’. These quantify the similarity between varieties through coefficients that range between 0 and 1, where

Table 3: Similarity coefficients across lines of evidence and alternations. Input dataset: all available data. Similarity coefficients range between 0 (total dissimilarity) and 1 (total similarity).

	Genitive alternation	Dative alternation	Particle placement alternation	
1st line (significance)	0.9	0.69	0.74	
2nd line (effect strength)	0.69	0.72	0.77	
3rd line (ranking)	0.82	0.74	0.73	
mean	*0.82*	*0.72*	*0.75*	**Γ=0.76**

0 indicates total dissimilarity and 1 indicates total similarity. Similarity coefficients are calculated as follows: for every variable phenomenon under study, we obtain *n* x (*n*-1) / 2 unique pairwise distance values for each line of evidence (steps 3, 4, and 6; see Figure 4), where *n* is the number of varieties under analysis. For example, if we study, say, the dative alternation in 9 varieties, then we obtain 9 x 8 / 2 = 36 unique pairwise distance values for each of the three lines of evidence. Next, we turn these distance values into similarity values by subtracting them from 1, such that a distance value of, e.g., .3 is converted into similarity value of 1 - .3 = .7. We can then go ahead and calculate one mean similarity coefficient per line of evidence.

Table 3 displays similarity coefficients per line of evidence and alternation. The coefficients range between 0.69 (second line, genitive alternation) and 0.90 (first line, genitive alternation). The last row displays mean similarity coefficients per alternation across lines of evidence: the mean similarity coefficient for the genitive alternation is 0.81; for the dative alternation it is 0.72; and for the particle placement alternation it is 0.75. This tells us that the genitive alternation is the most stable alternation across varieties (because the genitive alternation has the highest mean similarity coefficient), and the dative alternation is least stable (because it has the lowest mean coefficient); the particle placement alternation takes the middle road. We interpret this as evidence that the alternations under study are differentially sensitive to “probabilistic indigenization” (see Szmrecsanyi et al. 2016: 133 and the discussion in the previous section), in that more abstract (i.e. more syntactic) alternations, such as the genitive alternation, are less hospitable to probabilistic indigenization than other alternations.

The value in the bottom-right corner of Table 3 is the so-called ‘core grammar score’ Γ: this is the mean similarity coefficient across lines of evidence and across all alternations subject to study. Γ thus, abstracts away from particular alternations and lines of evidence: The higher Γ, the more similar the varieties. The dataset studied here (three grammatical alternations x nine varieties of English) yields a core grammar score of Γ = 0.76. Relying on customary schemes for interpreting (correlation) coefficients (e.g. De Vaus 2002: 272), we are dealing with “very strong” similarities between the varieties under study. How does this score compare to other VADIS research?

- Zhang and Szmrecsanyi (2024) investigate the same datasets under study in this paper, but with a primary interest in register differentiation rather than geographic variability. They report a core grammar score of Γ=0.73.
- Bartels and Szmrecsanyi (2024) explore the future temporal reference (FTR) alternation across nine geographical varieties of English, and report a similarity score of Γ=0.41.
- Li et al. (2024) explore the theme-recipient alternation in Mandarin Chinese and report, for various lectal dimensions, mean similarity scores ranging between Γ=0.62 and Γ=0.67.
- La Peruta (2022) investigates the mandative subjunctive alternation in British English, American English, and Canadian English and reports a similarity score of Γ=0.76.

The upshot is that given our core grammar score of Γ=0.76, the probabilistic grammars of nine international varieties of English investigated here are remarkably homogeneous.

Table 4: Core grammar scores for subsets of the data.

	Core grammar score Γ
All available data (Table 3)	0.76
Spoken data only	0.62
Written data only	0.62
Inner Circle varieties only (BrE, IrE, CanE, NZE)	0.79
Outer Circle varieties only (HKE, SgE, IndE, JamE, PhlE)	0.73

Of course, Table 3's core grammar score of Γ=0.76 was calculated based on an analysis of all available data – written and spoken English, and including both Inner Circle and Outer Circle varieties. Experimentation with various subsets of the data yields the core grammar scores reported in Table 4. The largest core grammar score is obtained when attention is restricted to Inner Circle varieties, indicating that these varieties are particularly homogeneous. Outer Circle varieties are substantially less homogeneous, probably thanks to variety-specific substrate effects. As to the difference that medium makes, written varieties are somewhat more homogeneous than spoken varieties. This is a bit surprising given a widely held suspicion that the production of spoken language is subject to universal (and thus potentially homogenizing) processing and production constraints and biases (e.g. Hawkins 1994; MacDonald 2013) in a way that the production of written language is perhaps not. On the other hand, we know that while especially vernacular speech is "the style in which the minimum attention is given to the monitoring of speech" (Labov 1972: 208), written language is more "governed by prescription" (D'Arcy & Tagliamonte 2015: 255), a fact that may level out regional differences.

4.3 Mapping out (dis)similarity relationships

We have seen thus far that there is a lot of homogeneity in the data, but there is also some heterogeneity. It is this heterogeneity that we will move on to map out next. The basic idea is the following. Pairwise distance measurements between varieties may be arranged in 'distance matrices'. Distance matrices are the customary input in fields such as dialectometry (e.g. Nerbonne et al. 1999) and function essentially like distance tables in road atlases, which indicate geographic distances between locations.

Table 5 illustrates by displaying one of the distances matrices that we generated, i.e. the distance matrix for the third line of evidence (constraint ranking) in the particle placement alternation (this is a random choice –

Table 5: Variation-Based Distance and Similarity Modeling (VADIS) distance matrix for the third line of evidence in the particle placement alternation (all data included, eight constraints considered; see Table 1). Scores range between 0 (maximum similarity) and 1 (maximum distance).

	BrE	CanE	HKE	IndE	IrE	JamE	NZE	PhlE	SgE
BrE	0	0.00	0.31	0.55	0.29	0.10	0.10	0.29	0.21
CanE	0.00	0	0.31	0.55	0.29	0.10	0.10	0.29	0.21
HKE	0.31	0.31	0	0.24	0.05	0.26	0.19	0.45	0.31
IndE	0.55	0.55	0.24	0	0.17	0.45	0.48	0.57	0.43
IrE	0.29	0.29	0.05	0.17	0	0.26	0.17	0.33	0.17
JamE	0.10	0.10	0.26	0.45	0.26	0	0.05	0.40	0.29
NZE	0.10	0.10	0.19	0.48	0.17	0.05	0	0.31	0.17
PhlE	0.29	0.29	0.45	0.57	0.33	0.40	0.31	0	0.10
SgE	0.21	0.21	0.31	0.43	0.17	0.29	0.17	0.10	0

for illustration, we could have picked any of the other distance matrices). All distances are scaled between 0 (no distance) and 1 (maximum distance). Take the pairing between BrE and NZE, which is associated with a comparatively small distance value of 0.10. Thus, BrE and NZE are very similar in terms of the constraint ranking in the particle placement alternation. By contrast, the distance between BrE and IndE is 0.55, which is considerably larger. Note that distance matrices as in Table 5 may also be fused at various levels. Initially, we obtain three different distance matrices per alternation (one per line of evidence). For one thing, line-of-evidence-specific distance matrices may be fused at the level of individual alternations, thus arriving at line-merged but alternation-specific distance matrices. This step leaves us with one distance matrix per alternation. We may then take a further aggregation step for the sake of raising the analysis of distance relationships to an even higher level of generalization. This we can accomplish by fusing the three alternation-specific distance matrices into a single compromise distance matrix merged across all lines and alternations, called Γ-matrix for short.

Distance matrices can be visually depicted using various techniques from the dialectometric toolbox. In this spirit, Figure 5 shows a Multidimensional Scaling (MDS) (Kruskal & Wish 1978) visualization of our Γ-matrix. The three Outer Circle varieties JamE, IndE, and PhlE can be considered outliers, due to their peripheral position in the plot. In the upper right-hand quadrant of the plot we find the four Inner Circle varieties BrE, IrE, NZE, and CanE. This comparatively tight cluster of Inner Circle varieties also echoes our earlier finding based on the analysis of core grammar coefficients (see Table 4) that Inner Circle varieties form a tighter (i.e. internally more homogeneous) cluster than Outer Circle varieties. Also located in this quadrant is SgE, a variety that we consider an Outer Circle variety a priori. However, at the risk of indulging in an untestable post-hoc explanation, we note that according to some analysts (see e.g. Leimgruber 2013: 122) SgE is an Outer Circle variety in the process of becoming an Inner Circle variety (essentially because more and more Singaporeans learn English as their first language). In summary, Figure 5 demonstrates a fairly robust pattern of Inner versus Outer Circle varieties.

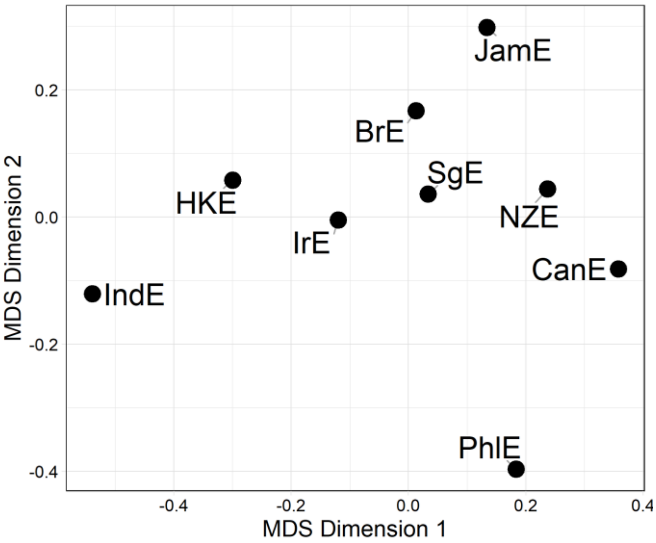

Figure 5: Multidimensional scaling representation of the Γ-matrix (a single compromise distance matrix merged across all lines and alternations). Distances between data points in plot is proportional to probabilistic grammar distances between varieties. (Adapted from Szmrecsanyi & Grafmiller 2023: Fig. 6.4)

5 Experimentation

Our experimental investigation focused on particle placement in four varieties – BrE, NZE, IndE, and SgE – and examined the degree to which one of the alternation's most dominant factors, the length of the direct object (see Figure 2), might influence participants' judgments in a preference rating task. The study was guided by three related research questions. First, when participants are provided with the same contextual information as a corpus model, do they exhibit a similar pattern of graded preferences for the use of a particular variant? Second, when changes in contextual features (e.g. length) correlate with changes in the probability of a variant in the corpus model, do those contextual changes correlate in the same way with changes to participants' ratings? Third, to what extent do cross-varietal differences in the effect of a contextual feature on participants' ratings, such as may exist, correlate with cross-varietal differences in the probabilistic associations of contextual features in a corpus model?

Our hypothesis is that language production and preference ratings are tapping into similar experience-based aspects of linguistic knowledge, therefore we should see positive correlations between ratings and corpus-model predicted probabilities. We should also see a negative effect of the direct object length on participant ratings, and furthermore, we expect that effect to be stronger on average among the Inner Circle participants (BrE and NZE), in line with patterns observed in the corpus data.

5.1 Materials and Procedure

Our design is modeled on Bresnan's (2007) "100 split task", in which participants would be asked to distribute 100 points for each of the two particle placement variants (continuous or split order) according to which they found better/worse in this context. Thus if participants feel that both variants are equally good, they should suggest a 50:50 split; if they feel that then first variant is much better than the second variant, they are supposed to suggest e.g. a 90:10 split; and so on. An example is shown in (6).

(6) On the other hand, I got a letter from a regular BBC correspondent who said he always *[turned the radio off / turned off the radio]* immediately if it was my turn on the programme, but he would like to take issue with something I had said last week.

________ *turned the radio off* [split]
________ *turned off the radio* [continuous]

Points can be directly correlated with corpus-model predictions. For this token, the corpus model assigns a 0.68 probability of the split variant *turned the radio off*, and we expect that participants, on average, should assign points for this same token in a similar range (60–70 points). Such parallelism provides evidence that similar forces are guiding both the off-line preference ratings and online language production.

We created thirty stimulus items, each consisting of an edited excerpt from the BrE component of the ICE corpus particle placement dataset collected by Grafmiller and Szmrecsanyi (2018). As in (6), both variants were inserted into the text and highlighted for participants, with the order randomized across items. We sampled items from across the range of probabilities predicted by a model of the corpus data (Grafmiller & Szmrecsanyi 2018). Six items were selected from each of five probability bins across the distribution of model predictions, and within each bin items varied in the length of the direct object. We also created fifteen filler items, which presented comparable lexical or grammatical alternations, e.g. the dative alternation and *that* complementizer omission. The full list of items can be found at https://osf.io/5hvtw.

Surveys were delivered online via Qualtrics, where participants were shown a welcome page with instructions and an example item illustrating the task. Participants were instructed that their task was to read each passage carefully and rate how natural each of the options sounded. Instead of assigning points, participants were presented with a slider bar with the two variants on either end, and asked to move the slider towards the variant that they felt sounded more natural in the context provided. We created two blocks of 15 test questions together with the 15 filler items. For the first block, three test items were randomly selected from each probability bin, and the remaining items were used in the second block. Each participant saw only one block. Pilot tests estimated the task to take approximately 20 minutes, and participants who took less than 10 minutes or longer than 1 hour were excluded. Five comprehension questions were also included, and participants who answered more than two comprehension questions incorrectly were excluded. In total, 260 participants were included in the analysis (BrE = 60, NZE = 81, IndE = 55, SgE = 64). See Szmrecsanyi and Grafmiller (2023: Chap. 7) for complete details.

5.2 Results

We fit a linear mixed model to predict ratings (centered at 50) with VARIETY, and the DIRECT OBJECT LENGTH and CORPUS PREDICTION for each token, as well as two-way interactions of the latter two factors with Variety. Variety was custom coded to test for three comparisons across and within Circles: Inner (BrE, NZE) vs. Outer Circle (IndE, SgE); BrE vs. NZE; IndE vs. SgE. Random effects included by-item and by-participant intercepts, as well as by-participant slopes for the effects of Direct Object Length and Corpus Prediction, and a by-item slope for the effect of Variety. The model's total explanatory power is substantial (N=3900; conditional R^2=0.38; marginal R^2=0.24). Again, see Szmrecsanyi and Grafmiller (2023: Chap. 7.2) for complete details of the statistical model.

In our data, Inner Circle participants rated the split variants significantly higher on average than Outer Circle participants (*β* = 11.69, *t* = 4.55, *p* < .001). Within-circle comparisons (BrE – NZE and IndE – SgE) did not reach significance. Thus, BrE participants did not rate the split variant higher or lower on average than NZE participants, and likewise IndE participants did not rate the split variant higher or lower on average than SgE participants (Figure 6).

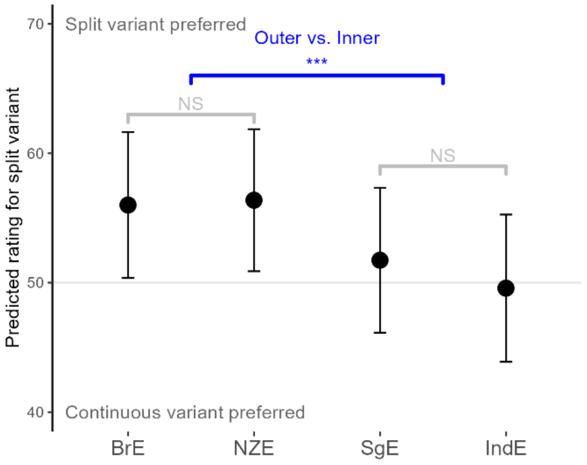

Figure 6: BrE and NZE participants rated the split variant (*turn the radio off*) significantly higher on average than IndE and SgE participants.

Participant ratings were significantly positively correlated with Corpus Prediction (*β* = 33.27, *t* = 6.35, *p* < .001), but we found little evidence that this trend varies meaningfully across varieties. We found a significant negative correlation of ratings with Direct Object Length (*β* = –13.01, *t* = –2.49, *p* = .013), which corroborates findings from our corpus analysis, which found a strong influence of this factor in all four varieties (Szmrecsanyi & Grafmiller 2023: Chap. 5). Participants' preferences on average aligned very well with the corpus model predictions, and furthermore their ratings were sensitive to the length of the direct object, even after adjusting for the additional correlation with corpus model predictions – which are based on numerous other contextual factors. We did find slight evidence of cross-varietal differences in the effect of Direct Object Length on our ratings, both across the Inner and Outer Circles (*β* = –12.03, *t* = –2.70, *p* = .007), and between IndE and SgE (*β* = –7.35, *t* = 2.29, p = .023). The effect of Direct Object Length was slightly stronger among BrE and NZE participants compared to IndE and SgE participants (Figure 7), and the difference was largest with very short direct objects but diminished somewhat as the direct object gets longer.

A similar pattern was found in the analysis of the similar ICE corpus data in Szmrecsanyi et al. (2016), where

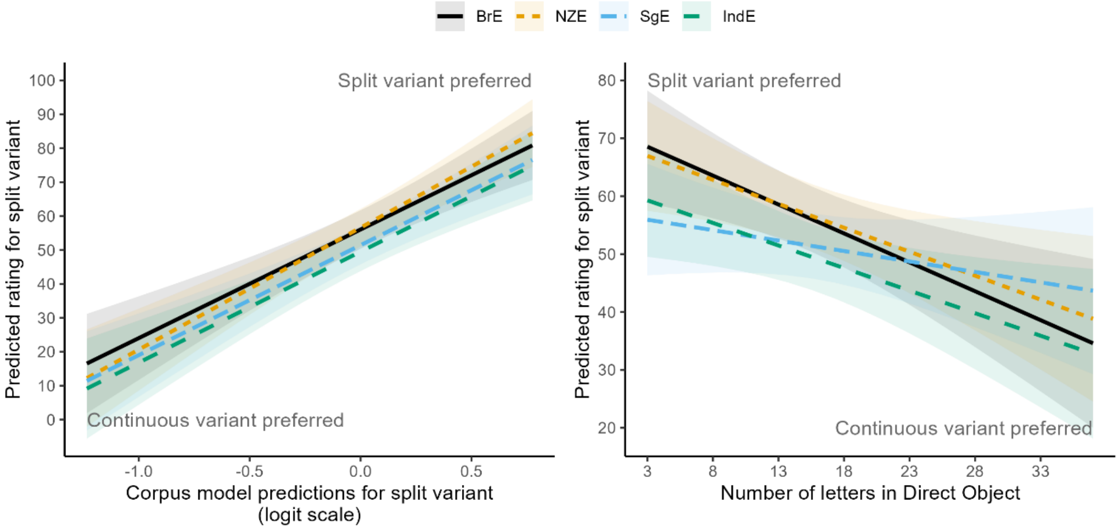

Figure 7: Ratings are positively correlated with corpus model predictions (Left) across all varieties. Ratings are negatively correlated with direct object length (Right) in all varieties, but correlations are significantly weaker across in Outer Circle (IndE and SgE) compared to Inner Circle (BrE and NZE) varieties, and in SgE compared to IndE.

the association between Direct Object Length and probability of the split variant is noticeably attenuated in IndE and SgE compared to BrE and Canadian English (CanE).

Assuming that length effects are driven by processing-related pressures to minimize long distance dependencies (e.g. Hawkins 1994), we expect that cross-varietal differences should emerge only in those contexts where the processing load is relatively light, i.e. the direct object is short. Both the corpus data and the experimental data are compatible with this expectation. In all then, the parallels between patterns in our corpus analysis and our experiment ratings provide compelling evidence that our corpus-based statistical models are capable of modeling cross-lectal variability in probabilistic linguistic knowledge (see also Bresnan & Ford 2010; Engel et al. 2022).

6 Concluding remarks

This paper would seem to have suggested that theorizing in variationist (socio)linguistics can profit substantially from cross-pollination with related subfields, such as probabilistic grammar research, psycholinguistics, and dialect typology. The point is that we can learn a lot from investigating how language-internal constraints (as opposed to the usual social constraints we see in variationist work) differ, or do not differ, across varieties of the same language.

We specifically explored probabilistic grammar variation – as manifested in three grammatical alternations – across nine L1 and L2 varieties of English. Taking a variationist interest in modeling how language users choose between “structurally and/or lexically different ways to say functionally very similar things” (Gries 2017: 7), our project builds methodologically on established quantitative methods in comparative sociolinguistics while expanding the analytical toolkit to include methods common in dialectology/dialectometry and in psycholinguistics. This inter-subdisciplinarity has allowed us to address questions such as the following: for a given alternation, how consistent are the probabilistic effects of the variable grammar’s constraints across varieties? Do some alternations vary more than others with respect to their probabilistic conditioning? Are there some (types of) constraints that are more variable than others? To what extent can the patterns we observe in corpus data be replicated in rating task experiments? Do the cross-varietal patterns we find align with our current understanding of typological variation among varieties of English? These are questions that, we believe, may add new theoretical twists to dialectological scholarship, on varieties of English and beyond.

Some key findings that we discussed in more detail in the preceding sections include – short and to the point – the following. First, probabilistic grammars across World Englishes are overall surprisingly stable: on a scale between 0 and 1, where 0 indicates total dissimilarity and 1 indicated total identity, the overall similarity of the alternation phenomena under study calculates as approximately 0.7. Second, effect directions are stable across varieties. If a particular language-internal constraint (consider, e.g. end-weight effects) favors a particular grammatical outcome in a given variety, it will also do so in the other varieties. Third, what is variable is the strength of effects. For example, constituent animacy may have strong effects on grammatical outcomes in variety A, but comparatively weaker effects in variety B. Fourth, different alternations are differentially hospitable to what we call ‘probabilistic indigenization’: for example, the particle placement alternation is (probably in function of its comparatively strong lexical anchoring) particularly malleable. Fifth, we often see a dialect-typological split between Inner Circle (ENL) and outer Circle (ESL) varieties. Sixth, experiments and corpus analysis converge largely but not entirely.

This is the general picture. However, what about particularities? Why is e.g. PhlE an outlier in Figure 5? Any explanations for indigenization and variation that may be proposed (see Szmrecsanyi & Grafmiller 2023: Chap. 8 for a fuller discussion) would be rather speculative. This is perhaps inevitable in a research program of this scope, where we cannot possibly delve into the complex histories of each variety and its (socio)linguistic details. We have tried to step back and draw connections at macro-varietal scale between our work and the tremendous body of research on these nine varieties of English (and many others) around the globe, yet we fully recognize the need for more empirical work on the finer specifics of the genitive, dative, and particle placement alternations in each variety. Such work would include not only more thorough investigation of the historical development of these varieties (subject to the availability of historical corpus data) but also a deeper look into how these alternations are acquired and how their use develops among both L1 and L2 users (see e.g. de Marneffe et al. 2012; Dubois et al. 2023).

Needless to say, the research summarized in this paper can and should be extended in many ways. More alternations need to be considered; we possibly need to consider other linguistic levels than grammar (e.g. phonological variation); properly sociolinguistic predictors (age, gender, and so on) ought to be included if the data source permits; and we need to cover more and other types of varieties (such as pidgin and creoles). Crucially, it is necessary to broaden the scope beyond English, to varieties of e.g. Spanish, French, Dutch, or German around the world.

Acknowledgements

Funding by the Research Foundation Flanders (grant #: G.0C59.13N) is gratefully acknowledged. We would also like to express our gratitude to the FWO-funded scientific research networks “Re-examining dialect syntax” (REEDS) (grant #: W002320N) and “Broadening the theoretical and methodological scope of translation and interpreting studies: towards an interdisciplinary language-contact framework” (INTERACT) (grant #: 3W0-02220W) for feedback and discussion. Two anonymous referees provided very constructive suggestions for improvement of this paper. The usual disclaimers apply.

Author Statement

All authors have consented to the submission of the manuscript to the journal, reviewed all the results and approved the final version of the manuscript. The authors jointly conducted the research reported in this paper. Szmrecsanyi was responsible for manuscript preparation.

Data availability

The annotated corpus-based datasets as well as the dataset of experimental ratings that we will investigate in the remainder of this paper are available at https://osf.io/5hvtw/. All statistical analyses were conducted using R statistical software (https://www.r-project.org/), and the R scripts necessary to replicate our analysis are likewise available from the repository.

Endnotes

[1] See Roethlisberger et al. (2017: 677) for adding a cognitive dimension to this process.

References

Bartels, Birgit & Benedikt Szmrecsanyi. 2024. Future temporal reference in spoken world Englishes. *World Englishes*. https://doi.org/10.1111/weng.12686

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

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in Search of Its Evidential Base*, 75–96. Berlin: Mouton de Gruyter.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald R. Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Kraemer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213. https://doi.org/10/cb3tn2

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). 245–259. https://doi.org/10/fjfpks

Chen, Ping. 1986. Discourse and particle movement in English. *Studies in Language* 10(1). 79–95. https://doi.org/10.1075/sl.10.1.05che

D’Arcy, Alexandra & Sali A. Tagliamonte. 2015. Not always variable: Probing the vernacular grammar. *Language Variation and Change* 27(3). 255–285. https://doi.org/10/gf7nqt

Davies, Mark & Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide. A Journal of Varieties of English* 36(1). 1–28. https://doi.org/10.1075/eww.36.1.01dav

De Vaus, David A. 2002. *Analyzing social science data*. London: SAGE.

Dubois, Tanguy, Magali Paquot & Benedikt Szmrecsanyi. 2023. Alternation phenomena and language proficiency: the genitive alternation in the spoken language of EFL learners. *Corpus Linguistics and Linguistic Theory* 19(3). 427–450. https://doi.org/10.1515/cllt-2021-0078

Engel, Alexandra, Jason Grafmiller, Laura Rosseel & Benedikt Szmrecsanyi. 2022. Assessing the complexity of lectal competence: the register-specificity of the dative alternation after GIVE. *Cognitive Linguistics*. https://doi.org/10.1515/cog-2021-0107

Fraser, Bruce. 1976. *The verb-particle combination in English*. New York: Academic Press.

Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Godfrey, John J, Edward C Holliman & Jane McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *IEEE International Conference on Speech, and Signal Processing, ICASSP-92*, vol. 1, 517–520.

Goebl, Hans. 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Österreichische Akademie der Wissenschaften.

Grafmiller, Jason & Benedikt Szmrecsanyi. 2018. Mapping out particle placement in Englishes around the world. A study in comparative sociolinguistic analysis. *Language Variation and Change* 30(3). 385–412. https://doi.org/10.1017/S0954394518000170

Grafmiller, Jason, Benedikt Szmrecsanyi, Melanie Röthlisberger & Benedikt Heller. 2018. General introduction: A comparative perspective on probabilistic variation in grammar. *Glossa: a journal of general linguistics* 3(1). 94. https://doi.org/10/gd3zrp

Greenbaum, Sidney. 1991. ICE: the International Corpus of English. *English Today* 7(4). 3–7. https://doi.org/10.1017/S0266078400005836

Gries, Stefan Th. 2017. Syntactic alternation research: Taking stock and some suggestions for the future. *Belgian Journal of Linguistics* 31. 8–29. https://doi.org/10.1075/bjl.00001.gri

Gries, Stefan Th. 2003. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. New York: Continuum Press.

Hawkins, John A. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Heller, Benedikt. 2018. *Stability and Fluidity in Syntactic Variation World-Wide: The Genitive Alternation Across Varieties of English*. Leuven: KU Leuven PhD dissertation.

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. https://doi.org/10/gf7nv8

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). 437–474. https://doi.org/10.1017/S1360674307002341

Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: the English language in the outer circle. In Randolph Quirk & Henry G. Widdowson (eds.), *English in the World: Teaching and Learning the Language and Literatures*, 11–30. Cambridge: Cambridge University Press.

Kachru, Braj B. (ed.). 1992. *The Other tongue: English across cultures*. 2nd ed. Urbana: University of Illinois Press.

Kortmann, Bernd & Kerstin Lunkenheimer (eds.). 2013. *eWAVE*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://ewave-atlas.org/

Kruskal, Joseph B. & Myron Wish. 1978. *Multidimensional Scaling*. Newbury Park, London, New Delhi: Sage Publications.

La Peruta, Roberta. 2022. Using VADIS to weigh competing epicentral influence. *World Englishes* 41(3). 400–413. https://doi.org/10.1111/weng.12585

Labov, William. 1966. *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

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

Leimgruber, Jakob R. E. 2013. *Singapore English: Structure, Variation, and Usage*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139225755

Li, Yi, Benedikt Szmrecsanyi & Weiwei Zhang. 2024. Across time, space, and genres: measuring probabilistic grammar distances between varieties of Mandarin. *Linguistics Vanguard* 10(1). 427–437. https://doi.org/10.1515/lingvan-2022-0134

MacDonald, Maryellen C. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology* 4. 1–16. https://doi.org/10/gbfpt3

Marneffe, Marie-Catherine de, Scott Grimm, Inbal Arnon, Susannah Kirby & Joan Bresnan. 2012. A statistical model of the grammatical choices in child production of dative sentences. *Language and Cognitive Processes* 27(1). 25–61. https://doi.org/10/d58h58

Nerbonne, John, Wilbert Heeringa & Peter Kleiweg. 1999. Edit Distance and Dialect Proximity. In David Sankoff & Joseph Kruskal (eds.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, v–xv. Stanford: CSLI Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London, New York: Longman.

Rosenbach, Anette. 2008. Animacy and grammatical variation—Findings from English genitive variation. *Lingua* 118(2). 151–171. https://doi.org/10.1016/j.lingua.2007.02.002

Rosenbach, Anette. 2014. English genitive variation – the state of the art. *English Language and Linguistics* 18(2). 215–262. https://doi.org/10.1017/S1360674314000021

Röthlisberger, Melanie. 2018a. *Regional variation in probabilistic grammars: A multifactorial study of the English dative alternation*. Leuven: KU Leuven PhD dissertation. https://lirias.kuleuven.be/handle/123456789/602938

Röthlisberger, Melanie. 2018b. *The dative dataset of World Englishes*. KU Leuven. https://doi.org/10.5281/zenodo.2553357

Röthlisberger, Melanie, Jason Grafmiller & Benedikt Szmrecsanyi. 2017. Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics* 28(4). 673–710. https://doi.org/10/gddnmm

Schneider, Edgar. 2003. The Dynamics of New Englishes: From Identity Construction to Dialect Birth. *Language* 79(2). 233–281. https://doi.org/10/c3b23n

Schneider, Edgar. 2007. *Postcolonial English: Varieties Around the World*. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt & Jason Grafmiller. 2023. *Comparative variation analysis: grammatical alternations in world Englishes*. Cambridge, New York: Cambridge University Press. https://doi.org/10.1017/9781108863742

Szmrecsanyi, Benedikt, Jason Grafmiller, Joan Bresnan, Anette Rosenbach, Sali 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. https://doi.org/10.5334/gjgl.310

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). 109–137. https://doi.org/10/gf7ntq

Szmrecsanyi, Benedikt, Jason Grafmiller & Laura Rosseel. 2019. Variation-Based Distance and Similarity Modeling: A Case Study in World Englishes. *Frontiers in Artificial Intelligence* 2. 23. https://doi.org/10.3389/frai.2019.00023

Tagliamonte, Sali. 2012. *Variationist sociolinguistics: change, observation, interpretation*. Malden, MA: Wiley-Blackwell.

Tagliamonte, Sali & 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. 135–178. https://doi.org/10/gf7c9n

Zhang, Xu & Benedikt Szmrecsanyi. 2024. Variation-based Distance and Similarity Modeling: A new way of measuring distances between registers. *Register Studies* 6(1). 31–59. https://doi.org/10.1075/rs.23011.zha

Zuur, Alain F., Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev & Graham M. Smith. 2009. *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer.

Zwicky, Arnold M. 1987. Suppressing the Zs. *Journal of Linguistics* 23. 133–148. https://doi.org/10/dmr4xj