Mapping out particle placement in Englishes around the world. A study in comparative sociolinguistic analysis

1st Author: Jason Grafmiller

Affiliation: University of Birmingham

3 Elms Road

Birmingham, B15 2TT

United Kingdom

Email: j.grafmiller@bham.ac.uk

Telephone: +4401214147017

2nd Author: Benedikt Szmrecsanyi

Affiliation: KU Leuven

Short title: Mapping particle placement out

ABSTRACT

This study explores variability in particle placement across nine varieties of English around the globe, utilizing data from The International Corpus of English and the Global corpus of Web-based English. We introduce a quantitative approach for comparative sociolinguistics that integrates linguistic distance metrics and predictive modelling, and use these methods to examine the development of regional patterns in grammatical constraints on particle placement in World Englishes. We find a high degree of uniformity among the conditioning factors influencing particle placement in native varieties, (e.g., British, Canadian, and New Zealand English), while English as a second language varieties, (e.g., Indian and Singaporean English), exhibit a high degree of dissimilarity with the native varieties and with each other. We attribute the greater heterogeneity among second language varieties to the interaction between general L2 acquisition processes and the varying sociolinguistic contexts of the individual regions. We argue that the similarities in constraint effects represent compelling evidence for the existence of a shared variable grammar, and variation among grammatical systems is more appropriately analyzed and interpreted as a continuum rather than multiple distinct grammars.

ACKNOWLEDGEMENTS

We thank Melanie Röthlisberger, Benedikt Heller, Jack Grieve, and members of the Quantitative Lexicography and Variational Linguistics research unit at KU Leuven for valuable comments and feedback. This research was supported by an Odysseus grant (PI: second author) from the Research Foundation Flanders (FWO, grant no. G.0C59.13N). The usual disclaimers apply.

INTRODUCTION

This study focuses on geographical variation in grammatical constraints on the English particle verb (PV) alternation, exemplified in (1) and (2), which hinges here on the variation in the ordering of the direct object, *the book*, and the post-verbal particle *up*.[1]

(1) the continuous variant/order (V-P-O):

My 6yo [sic] old actually **picked up** the book and started reading it! <GloWbE:NZ>

(2) the split variant (V-O-P):

I have not **picked** the book **up** in 2 years. <GloWbE:CA>

In this paper, we investigate the degree of (in)stability in the factors, such as the length and information status of the direct object, which probabilistically influence the choice between these variants among different varieties of English around the world. Particle placement has been a popular research topic in recent years, especially in usage-based studies of variation (e.g., Cappelle 2009; Goldberg 2016; Gries, 2003; Lohse, Hawkins, & Wasow, 2004; Szmrecsanyi, Grafmiller, Heller, & Röthlisberger, 2016), as well as in diachronic (e.g., Elenbaas, 2013; Rodríguez-Puente, 2016; Thim, 2012), and L1 and L2 acquisition research (e.g., Diessel & Tomasello, 2005; Gilquin, 2014; Gries, 2011). Bringing together insights from these various domains, the present study contributes to a growing body of work investigating the nature and limits of grammatical variation among so-called 'World Englishes'[2], in order to identify cross-varietal differences in the factors conditioning syntactic variables (see also Bernaisch, Gries, & Mukherjee, 2014; Bresnan & Ford, 2010; Bresnan & Hay, 2008). In doing so, we present an innovative technique for comparative sociolinguistic research (Poplack & Tagliamonte, 2001; Tagliamonte, 2013) that incorporates both predictive modelling and linguistic distance measures

to examine the degree to which the grammatical system that underlies particle placement has been, and continues to be, evolving and diversifying in varieties across the globe.

No single feature or set of features shapes particle placement in a categorical, deterministic way, and numerous influential factors have been identified over the years. These include the length, information status, and conceptual accessibility of the direct object, as well as the semantics of the verb and the presence of a post-modifying prepositional phrase (*pick the rock up off the ground*) (see Cappelle [2009] for review). But what do we know about regional variability in particle placement? The answer turns out to be: surprisingly little. A few studies have hinted at possible regional differences within the UK (e.g., Elenbaas, 2007) and the US (Grieve, 2016:70-71), but until recently evidence for differences in the influence of the linguistic constraints on particle placement has been thin on the ground. Haddican and Johnson (2012) investigated regional variation in the influence of the length and information status of the direct object within the US and UK, using judgment surveys and Twitter data, and found no evidence of regional variation within either of the two countries in the overall rate of use of PV variants, or in the influence of internal constraints on the choice of variant; however, they did nonetheless find a consistent difference in the overall rate of variant use *between* the two countries (see also Cappelle, 2009:177-178). Notably, no significant interactions of country and linguistic features were found, suggesting that while UK and US speakers may vary in their overall use of PV variants, their choices are governed by a shared set of contextual constraints to roughly the same degree. That is, they share the same variable grammar.

If we zoom out to a wider perspective on the global English-speaking diaspora, even less is understood about PV variation, whether in other native, or 'Inner Circle' (Kachru, 1992) varieties such as Canadian or Australian English, or more established 'Outer Circle' varieties

such as Indian or Philippines English. To the extent that PVs have been investigated in this domain, researchers have focused mainly on overall frequencies of PVs across varieties and/or genres (e.g., Zipp & Bernaisch, 2012), but have said little about variability in the choice between the two PV variants. Still, PVs are notoriously challenging for L2 learners, especially for speakers whose first language lacks particle constructions (e.g., González, 2010; Liao & Fukuya, 2004; Siyanova & Schmitt, 2007), but there is some evidence that proficient L2 users can achieve native-like performance with respect to the influence of certain features (e.g., Blais & Gonnerman, 2013; Gilquin, 2014). In light of this, we believe PVs offer a potentially fruitful test case for examining variation across World Englishes, and for studying the ways in which variable grammars evolve and diversify.

For example, Szmrecsanyi, Grafmiller, Heller, and Röthlisberger (2016) investigated particle placement in spoken and written corpus data from four national varieties of English. They found that the split variant is used significantly less in the two Outer Circle varieties (Singapore and Indian English) than in the Inner Circle varieties (British and Canadian English), but also that the effects of some internal constraints, such as the length of the direct object, are substantially weaker in the Outer Circle varieties. They argue that in contexts where neither variant is substantially more difficult to process than the other(s), other forces—whether social, semantic, lexical, or just plain random fluctuation—are more likely to affect changes to the variable grammar. Szmrecsanyi et al. (2016:133) characterized the interplay among these forces and their statistical associations with different grammatical variants, which shapes the grammar(s) of new varieties of English as the process of 'probabilistic indigenization'.

Following Szmrecsanyi et al. (2016), we conceptualize probabilistic indigenization in variable grammars as so: To the extent that the conditioning constraints on a variable in a new variety *A*, for example, 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 variation grammar of *A*. These patterns may not be stable in the earlier stages of a variety's development, but they may provide indications of changes in progress and imply the emergence of a unique variable grammar. This approach assumes a probabilistic, usage-based model of grammar that is committed to the notion that grammar is the "cognitive organization of one's experience with language" (Bybee, 2006:711). Probabilistic grammars are usage-based in that quantitative patterns derived from experience are associated not only with surface forms or lexical items (as in pure exemplar models), but with abstract features or constraints. Many of these features may reflect inherent, possibly universal aspects of linguistic structure, such as the tendency to map animate referents to more prominent structures or positions (e.g., Branigan, Pickering, & Tanaka, 2008). A desirable feature of probabilistic grammar models is that they can account for gradient, experience-driven variability within the context of general constraints on the range of possible variation. From this perspective, inter- and intra-systemic variation arises from the interplay between biases in language production and comprehension and acquired form-meaning associations, and this interplay leads to variability in the distribution of forms which speakers implicitly learn (e.g., Bresnan & Ford, 2010; Bresnan & Hay, 2008). Over time, the aggregation of quantitative differences in such distributions give rise to distinct variable grammars.

In mapping out a landscape of PV grammars across World Englishes, we confront questions central to research in the corpus-based variationist and comparative sociolinguistics traditions (e.g., Bresnan & Hay, 2008; Szmrecsanyi et al., 2016; Tagliamonte, 2013). To what extent can the effects of internal factors (e.g., the length or definiteness of a PV direct object), vary across regional varieties in their strength, direction, or relative importance? Methodologically, how can we compare variable grammars across multiple varieties in an

empirically sound way? How does the variability we do (or do not) observe inform the notion of a "common core" (Quirk, Greenbaum, Leech, & Svartvik, 1985:16) of English grammar? What role might external factors play in explaining this (lack of) variability? It is these questions that motivate our study.

DATA SOURCES, VARIABLE EXTRACTION, AND ANNOTATION

Data sources

Data for the study were taken from nine components of the International Corpus of English (ICE; Greenbaum, 1996), and the corpus of Global Web-based English (GloWbE; Davies & Fuchs, 2015): Great Britain (GB), Canada (CA), New Zealand (NZ), Ireland (IE), Jamaica (JA), Singapore (SG), Hong Kong (HK), India (IN), and Philippines (PH). ICE is an ongoing project, initiated in 1990, which was designed to create a set of parallel, balanced corpora representative of language usage across a wide range of (standard) national varieties. Each ICE component contains 500 texts of ~2000 words each, sampled from 12 spoken and written genres/registers, totaling ~1 million words. The structure of ICE components is shown in Table A.1 in the appendix.[3] Despite being published at different times over the past 20 years, all ICE components included here contain data from the early 1990s, with some also containing data collected as late as the early 2000s.

In addition, we included data from web-based language as represented in GloWbE with the aim of exploring the extent to which particle placement in an online medium might differ from language use in more "traditional" mediums. GloWbE contains data collected from 1.8 million English language websites—both blogs and general web pages—from 20 different countries (~1.8 billion words in all). To keep our dataset to a manageable size, we randomly

sampled texts from each of the nine varieties, totaling ~500k words per variety. This resulted in a sample web-based corpus approximately half the size of our ICE corpus.

Extracting tokens and defining the variable context

To extract potential PV tokens, we used simple regular expression searches of the part-of-speech (POS) tagged versions of the two corpora, which looked for any verb followed within a 10-word window by one of the ten particles in (3).

(3) around, away, back, down, in, off, on, out, over, up

These are the ten most frequent particles according to Gries' (2003:203-210) list of transitive PVs from several dictionaries of English phrasal verbs (Gries, 2003:67).

Following the automatic extraction, we winnowed down the set of interchangeable PVs such as (4) by manually removing non-interchangeable tokens (5).

(4) You do have to be firm with lawyers to not send [endess] letters back and forth and **run up** the bill. (interchangeable)

(5) The lift was out of order, so he decided to **run up** the stairs (NOT interchangeable)

Such non-interchangeable tokens included intransitive (*they **called in** at Port Ross*) or passive PVs (*a […] mosaic **brought back** by crusaders*), and any transitive PVs in which the direct object was a *wh-* form or a relative pronoun (***What** did Nasir **bring back** from the moon? The beliefs about god **that** this survey **throws up***). We also removed any token in which the particle/preposition occurred with a complement (***get** the ball **in** the net*). In addition to excluding these clear-cut cases, we employed a number of tests to distinguish PVs from non-interchangeable 'prepositional' verbs, as in (6).

(6) He **called on** the U.S. to end its economic embargo of Cuba.

These tests are discussed further in our annotation manual.[4]

After inspection of the dataset, we further narrowed the envelope of variation according to two additional criteria: the length and category (pronominal vs. non-pronominal) of the direct object NP (see also Cappelle, 2009). In our data, direct objects greater than 6 words in length categorically favor the continuous variant (n = 947/953, 99.4%), while the continuous order is almost never used when the direct object is a personal pronoun (n = 7/3245, 0.2%). Thus, tokens containing pronominal direct objects or direct objects longer than 6 words were excluded from the analysis in order to restrict the scope of the study to truly variable contexts (D'Arcy & Tagliamonte, 2015). The final dataset consisted of n = 11340 PV tokens.

Annotation of conditioning factors

Each interchangeable PV token was annotated for numerous features hypothesized to influence the choice of particle placement.

Length of the direct object

Longer objects increasingly favoring the continuous variant. The general tendency for longer, or "heavier", elements to be placed at the end of a phrasal unit in English is an example of the well-known Principle of End Weight (e.g., Wasow, 2002), and this effect has been widely documented in the PV literature. We use a measure of direct object length based on the number of words (*median* = 2, *median absolute deviation* = 1.5) rather than the number of letters, though the two measures are naturally very strongly correlated (r = 0.84).

Accessibility of the direct object

Accessibility-based theories of production preferences hold that elements that are more active in working memory are more likely to be uttered sooner (e.g., Bock & Warren, 1985; Branigan et al., 2008; MacDonald, 2013). With this in mind, four features related to the conceptual

accessibility of the direct object were annotated for: information status (**Givenness**), **Definiteness**, **Concreteness**, and degree of topicality or **'Thematicity'** (Osselton, 1988). The first three factors have been consistently shown to influence particle placement, with less newsworthy, that is, discourse given, definite and/or concrete direct objects favoring the split variant (e.g., Gries, 2003; Haddican & Johnson, 2012). Givenness of the direct object was determined automatically via string matching for any instance of the direct object head lemma within the first 100 words of text preceding the token in which it occurred. If the head was found within this window, it was coded as 'given', otherwise it was coded as 'new'. Definiteness was coded as a binary feature ('definite' versus 'indefinite') following guidelines developed by Garretson (2004). Concreteness ('concrete' versus 'nonconcrete') was determined based on whether the referent of the direct object was visible and/or physically manipulable (Gries, 2003). Thematicity was operationalized as the normalized text frequency (Hinrichs & Szmrecsanyi, 2007) of the direct object head noun. More discourse thematic direct objects are assumed to be more active in working memory, hence accessible, therefore we predict that the split variant should be more likely the greater the thematicity of the direct object.

Presence of a post-modifying directional PP

The presence of a postmodifying directional PP, as in (7), is also known to influence particle placement in present day English (Fraser, 1976; Gries, 2003; Rodríguez-Puente, 2016), thus we included a binary factor indicating the presence of a directional PP modifier following the verb phrase as in the examples below (directional PP in italics).

(7) Well, as we have said, applying VAT on government contracts is **taking** money **away** *from one government pocket* and putting it in another. <ICE:PH-W2E-002>

According to Gries (2003), the split variant foregrounds the spatial meaning of the PV, thus the presence of additional material spelling out the direction or endpoint of spatial movement is to be expected with this variant.

Structural persistence/priming

The tendency for language users to reuse previously encountered grammatical structures is variously referred to as structural or syntactic "persistence" or "priming" (see Gries and Kootstra [2017] for a recent review). The basic idea is that recent use of a particular structure, for example, *I picked the book up* should make the use of that same structure, for example, *I put the book down* more likely than competing structures, *I put down the book*, in upcoming choice contexts. For this study, we focus only on priming from one choice context to the next. For each token, we coded for the PV variant that had been used most recently within a 100-word span preceding the token in question. If no interchangeable PV occurred within that span, the prime type was coded as 'none'.

CV alternation

To investigate potential phonological/phonetic effects, we coded for the type (C or V) of the final segment of the verbform and the type of the initial segment of the particle: 'CC', 'VV', 'CV', 'VC' (Browman, 1986; Gries, 2011).

(8) It was an hour long telephone interview where Essa **laid down** his law. (CC) <GloWbE:CA>

(9) His job was to **throw** people **out** for non-payment of rent. (VV) <ICE:IE:S1B-045>

(10) It took fire fighters an hour to **put out** the blaze. (CV) <ICE:PH:S2B-002>

(11) Letting me **lie** my head **down** for a few days in your Penthouse was great. (VC) <ICE:GB:W2F-012>

Gries (2011) found a *horror aequi* effect in children's PV usage, where the split variant was preferred when the two segments were of the same type (CC, VV), however, our data points mainly to a binary distinction between consonant "clusters" (CC) and the other patterns, therefore we coded this variable as 'CC' vs. 'Other'.

Rhythm

In natural usage, the tendency of English speakers to alternate stressed and unstressed syllables where permitted by the grammar has been shown to influence change and variation at multiple levels of grammatical structure (see Shih [2017] for review). To calculate the rhythmicity of a PV token we counted the number of unstressed syllables at the boundary between the verb and particle in the continuous variant, and the verb and direct object in the split variant. This resulted in a measure of 'eurhythmy distance' (ED) for each variant, from which we calculated a measure of comparative rhythm by subtracting the split ED from the continuous ED.

(12) Split preference (Rhythm < 0)

a. **hand**ing **out** the stockings [ED = 1]

b. **hand**ing the **stock**ings out [ED = 2]

(13) Continuous preference (Rhythm > 0)

a. **throw** a**way** books [ED = 1]

b. **throw books** away [ED = 0]

(14) No preference (Rhythm = 0)

a. **pick**ing **up** gorgeous Samsons [ED = 1]

b. **pick**ing **gor**geous Samsons up [ED = 1]

More positive measures indicate that the split variant is more rhythmic than the continuous variant, and more negative scores indicate that the continuous variant is more rhythmic than the split one. Scores of 0 reflect cases where neither is more rhythmically ideal than the other.

Semantic compositionality of the PV

To annotate for semantic compositionality, we used a binary coding which captured the difference between PV tokens whose meaning is entirely predictable from that of their parts, and those tokens whose meaning is not fully predictable. For annotation, we followed the heuristic of Lohse et al. (2004:244–246), who employed two entailment tests to determine semantic compositionality. If [X V (P) NP (P)] entails both [X V NP] (verb entailment) *and* [NP PredV P] (particle entailment)—where PredV represents a predication verb of the type *be, become, come, go, stay*—then the PV was coded as 'compositional'. If either entailment test failed, the token was coded as 'non-compositional'. Semantic annotation was conducted by the first author, after which a random sample of 200 tokens was coded independently by a trained research assistant and the ratings were compared. Agreement between coders was reasonably good given the task (90%, Cohen's κ = .75).

Information content of verb and particle

As an additional measure of verb-particle dependency, we calculated the informativity, or surprisal, of the verb and the particle as components of a PV, where surprisal is defined in the information-theoretic sense (Shannon, 1948) as the log inverse of the conditional probability of an item given a particular context. For a given verb-particle pair, we calculated the surprisal of the particle given the verb (**Surprisal.P**), and the surprisal of the verb given the particle (**Surprisal.V**). The formula in (18) illustrates how we calculate the surprisal of the particle *up* given the verb *pick* (**Surprisal.P**).

$$\text{Surprisal}(up|pick) = -\log_2 P(up|pick) = -\log_2 \frac{Freq(pick\ up)}{Freq(pick)}$$

Informally, the less predictable a word is in a given context, the more surprising or informative it is. We interpret the surprisal measures in much the same fashion as Lohse et al.'s

(2004) entailment tests, that is, as measures of the relative syntactic/semantic (in)dependence of the verb and particle. From an information-theoretic perspective, we associate higher informativity (lower predictability) with greater independence, since more informative verbs/particles are those that occur more often in contexts without their respective PV component. PV types with especially low **Surprisal.P** scores are those for which, once we know the verb, we can predict the particle with a high degree of confidence. We hypothesize such PVs will strongly favor the continuous variant.

External predictors

Due to the lack of sufficient sociolinguistic metadata for some of our corpora, we limit our external predictors mainly to **Variety** and **Genre**. Looking at the distribution of PV variants across varieties (Figure 1), we find the biggest difference lies in the ICE data in which the discontinuous variant is clearly more frequent in the four Inner Circle varieties (British, Canadian, New Zealand, and Irish English), than the remaining varieties. This contrast is not found in the GloWbE data, however.

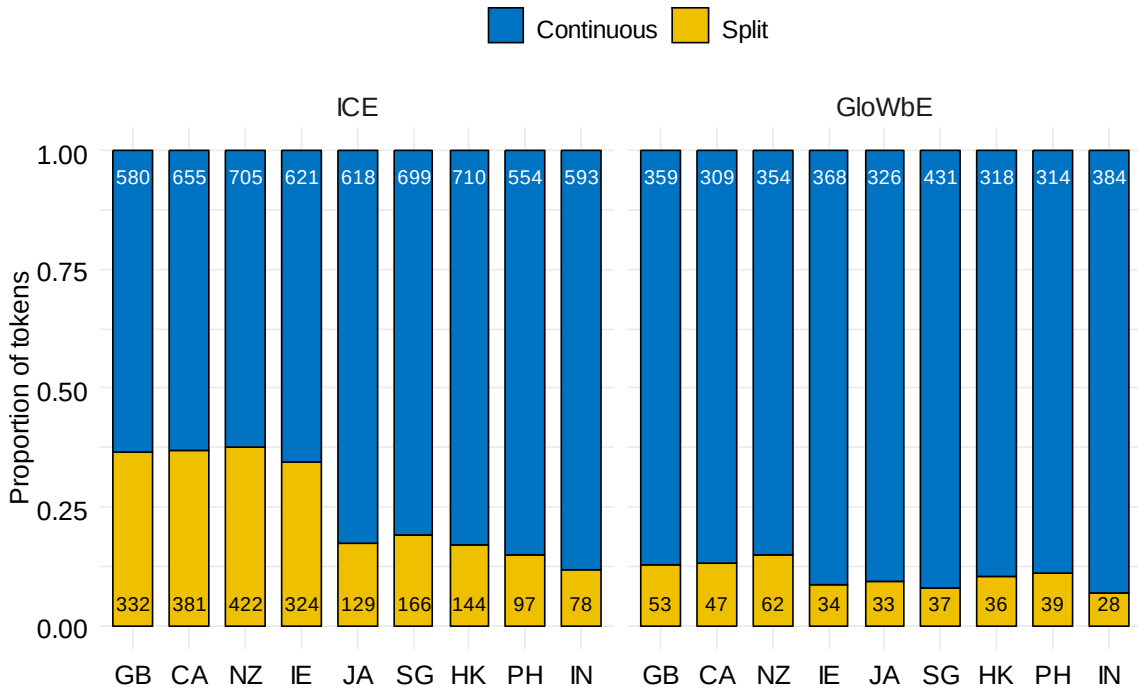

FIGURE 1. Distribution of PV variants by Variety and Corpus. The length of bars reflects the proportion of variants and the values represent the number of tokens of a given variant.

Turning to the distribution of PV variants across genres (Figure 2), the split variant is much more common in more informal/colloquial speech and writing (e.g., letters and creative fiction), than in other genres, and this tendency is more pronounced in the Inner Circle varieties. On the other hand, PV usage in GloWbE is on par with that of other written genres such as academic and newspaper writing, which heavily favor the continuous variant. Upon closer inspection, we find that many of the texts from GloWbE are in fact structurally and/or thematically similar to some of the non-fiction ICE texts. Many of these GloWbE texts are either news reports or non-fiction writing with an informational rather than interpersonal (Biber, 1988) focus. Thus, it appears that formality and medium may have less of an effect on PV use than more finely-tuned functional or thematic factors (see also our discussion constraint rankings in IN below).

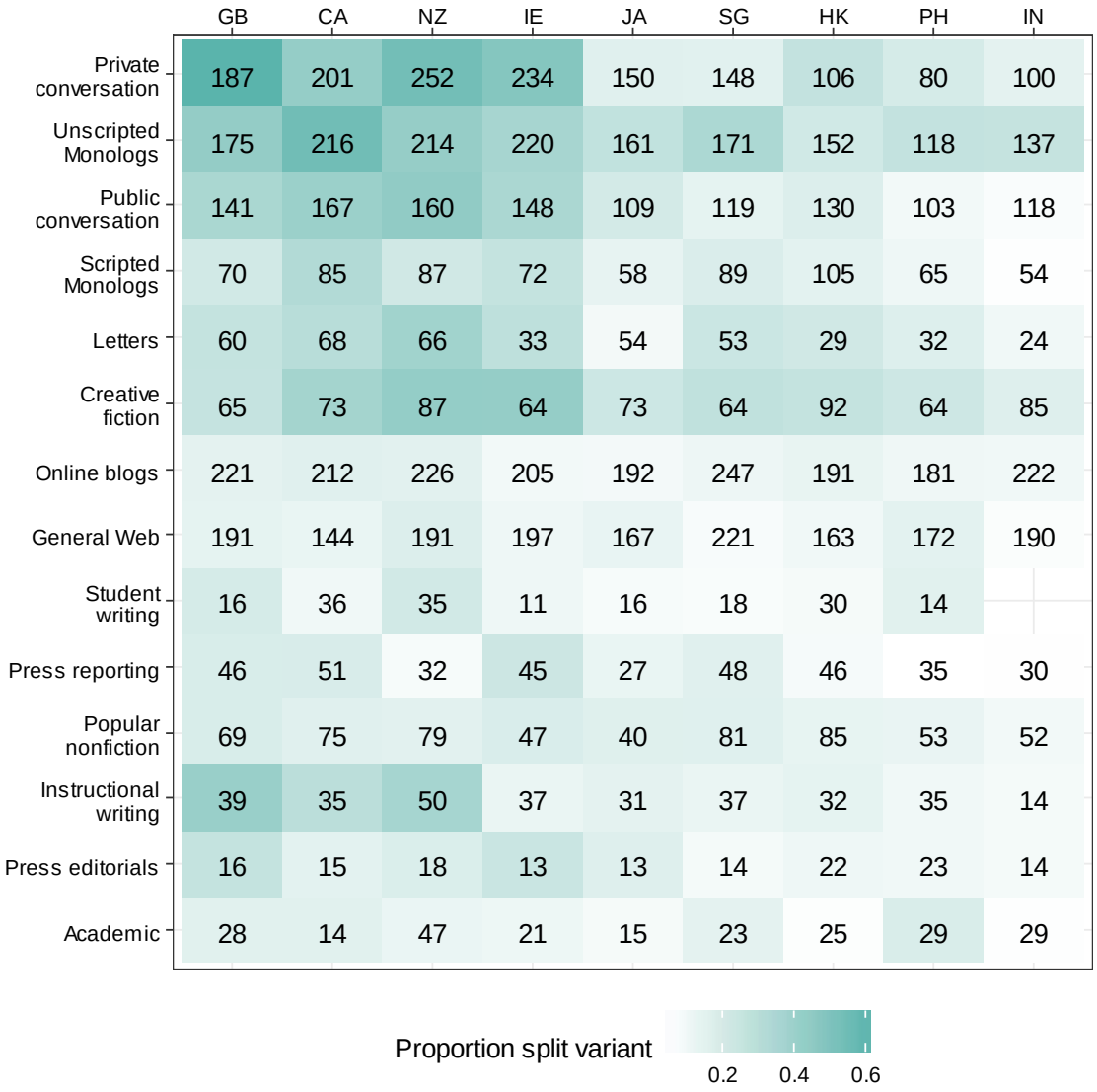

FIGURE 2. Heat map showing distribution of PV variants by Genre in ICE and GloWbE ('Online blogs' and 'General Web'). Numbers represent the total number of tokens (split and continuous), and darker areas represent greater proportions of the split variant.

Lastly, we also coded a simplified 5-level predictor of **Register** which collapsed genres according to formality and modality: spoken informal texts (genres: Private dialogues, Unscripted monologues) versus spoken formal texts (Public dialogues, Scripted monologues) versus written informal texts (Student writing, Letters, Press editorials, Creative fiction) versus written formal

texts (Academic writing, Press reportage, Popular nonfiction, Instructional writing) versus online texts (Blogs, General websites).

QUANTITATIVE ANALYSIS

To explore the differences among our varieties' PV grammars, we draw inspiration from the methodological framework of Comparative Sociolinguistics (Poplack & Tagliamonte, 2001; Tagliamonte, 2013). The comparative approach is characterized by its use of three "lines of evidence" (Poplack & Tagliamonte, 2001:92) derived from the statistical modelling of a linguistic variable across two or more datasets: the ranking, relative strength, and statistical significance of conditioning constraints. Examining variable patterns from multiple quantitative angles in this way helps us to determine how much, and in what ways, certain varieties have diverged or may be diverging from one another as well as from their mother variety. In short, similarities and/or differences in these lines of evidence can "provide a microscopic view of the underlying grammatical system[s]" (Tagliamonte, 2013:130), among different varieties of English.

Comparative analysis: Constraint ranking and probabilistic distances

For the comparative analysis we utilize generalized linear mixed models (GLMMs), specifically binomial logistic mixed models (e.g., Bresnan, Cueni, Nikitina, & Harald, 2007).[5] We fit GLMMs of each variety dataset individually, which incorporated reasonable, though non-maximal random structures (see below). Fitting individual by-variety models is an essential component of the Comparative Sociolinguistics as it offers us direct measures of effect size, direction, and statistical significance for individual predictors. A potential drawback is that we

cannot directly assess whether a particular cross-varietal difference is statistically significant; however, considering that researchers may have few *a priori* hypotheses about *which* internal constraints should vary in *which* varieties, even comparative analyses using regression models with cross-varietal interaction terms (e.g., Szmrecsanyi, Grafmiller, Bresnan, Rosenbach, Tagliamonte, & Todd, 2017) are still mainly exploratory in nature (see Paolillo, 2013:114). We therefore believe that comparative analysis is better served by a variegated statistical methodology that deploys multiple techniques, including significance tests where possible, but is interpreted with a more conservative exploratory slant.

As mentioned above, the Comparative Sociolinguistic method involves the evaluation of three 'lines of evidence' derived from separate quantitative models fit to different datasets (Poplack & Tagliamonte, 2001; Tagliamonte, 2013). A first line of evidence involves a simple comparison of constraints that are determined by a statistical model to be significantly correlated—under multivariate control—with a linguistic variable: Do the varieties in question share the same set of significant constraints (statistical significance)? A second line of evidence involves ranking the different (sets of) constraints, or factor groups, according to the strength of their influence on the variable: Do the constraints have the same relative impact or explanatory power on the variable in each variety (relative strength/magnitude)? The third line of evidence consists of comparing the magnitude and direction of the significant constraints across datasets: Are the directions of the constraints the same; are the orders of the levels within a constraint the same (constraint hierarchy)? Are the effects of some constraints stronger in one variety/dataset than others?

Evaluation of these lines of evidence typically proceeds by visually inspecting the statistical results across each of the datasets (see e.g., Tagliamonte, 2013). We extend this approach by developing a method for quantifying the variability within these lines of evidence

and visualizing the degree of (dis)similarity among varieties. In the first step, we fit mixed-effects regression models to each of our variety datasets separately, which are checked against standard diagnostics. For comparing relative constraint importance in each variety, we use a permutation-based importance measure to determine the change in model *AICc*, an information theoretic measure of how much the in-/exclusion of a given constraint affects our ability to model variation in particle placement. To assess possible differences in constraint hierarchies across varieties, we examine the coefficient estimates for our predictors returned by the models. Coefficients provide information about possible differences in the direction of a given predictor's effect and/or differences in the relative strength across multiple levels of a predictor. In the second step, we use the results from our models as input data for analysis of linguistic distances between varieties. As in traditional comparative sociolinguistic studies, our approach incorporates information from both significant and non-significant constraints; however, it moves beyond simple visual inspection toward more quantitatively rigorous methods of detecting relationships among datasets that may be difficult to perceive otherwise. Further details of the methods are provided in the supplementary materials.

Model fitting

The model structure for all by-variety regression models is shown in (15).

(15)
```
Response ~ Register + DirObjLength + Semantics +
DirObjConcreteness + DirObjGivenness + DirObjDefiniteness +
DirObjThematicity + DirectionalPP + CV.binary + Surprisal.P +
Surprisal.V + Rhythm + PrimeType + (1|Verb) + (1|Particle) +
(1|VerbPart) + (1|Genre)
```

All models fit the datasets quite well, with very little evidence of data multicollinearity or predictor intercorrelation (see supplementary materials for details). Figure 3 shows the individual constraint rankings obtained from each model. The length of the lines reflects the relative importance of a given predictor—the longer the line, the more important the predictor in that variety. From this figure, it is immediately clear that the PV variation grammars differ substantially across the varieties in terms of the relative importance of the various constraints affecting particle placement.

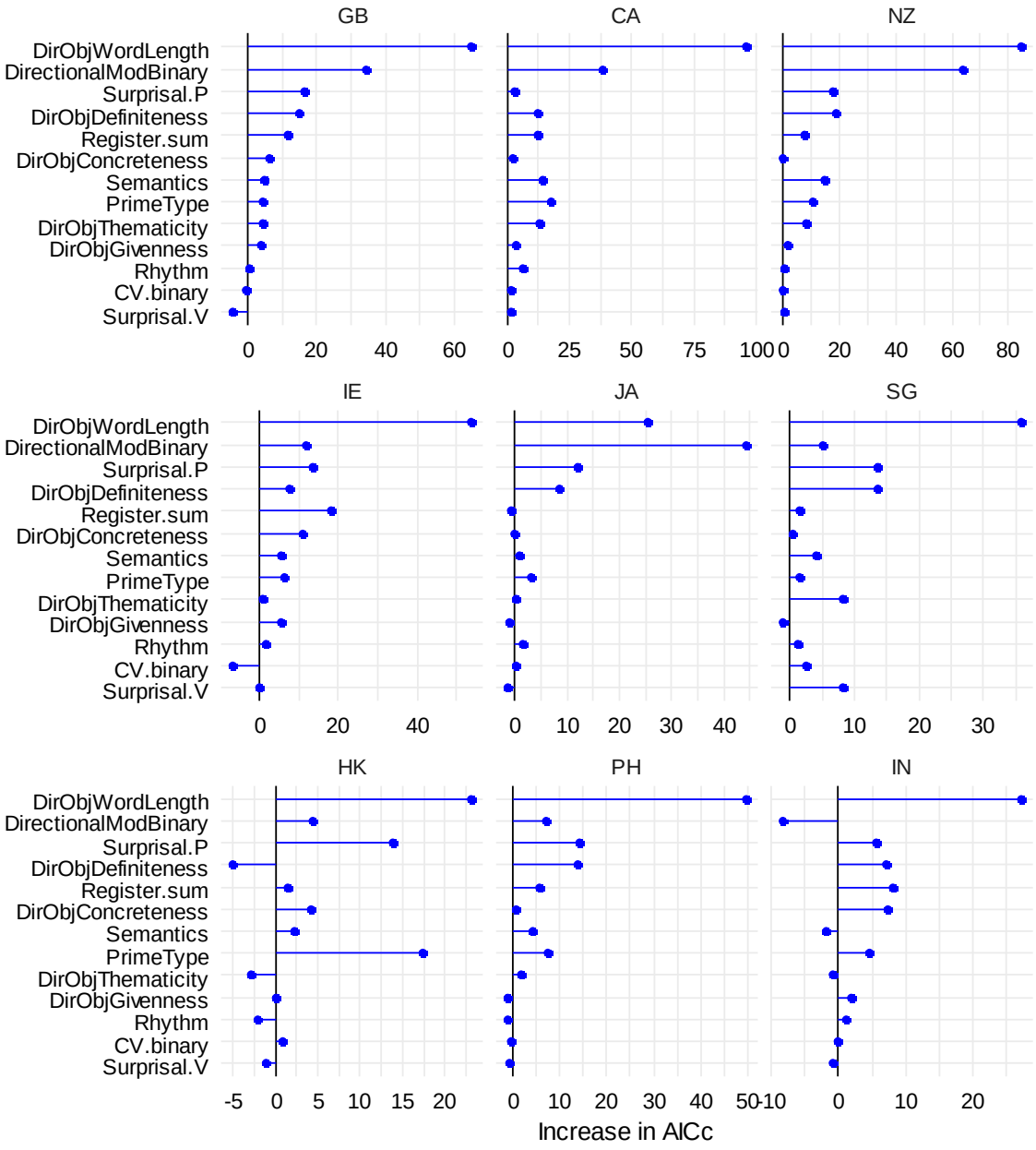

FIGURE 3. AICc based importance rankings by predictor (fixed-effects only) for variety-specific GLMMs. For comparison, predictors in all varieties are sorted according to their ranking in the GB model. Scores reflect the increase in AICc (decrease in model fit) compared to the original model for a model in which the values of a given predictor are randomly reshuffled.

Several constraints are stable in their relative ranking. In particular, **DirObjLength** ranks quite high in all varieties, while information structural factors (**DirObjGivenness** and **DirObjThematicity**) and phonological factors (**CV.binary** and **Rhythm**) rank consistently

among the least important constraints in our datasets. The consistently strong influence of **DirObjLength** is compatible with theories of variation in production that attribute end weight effects to users' drive to reduce online processing effort (e.g., Hawkins, 2004; MacDonald, 2013). Other factors differ considerably across the varieties. For instance, direct object concreteness ranks quite high in Indian while priming plays a substantial role in the Hong Kong English dataset, but both predictors rank rather low in the other varieties. At the same time, the presence of a directional modifier ranks high in GB, CA, NZ, and especially in Jamaican English, but is relatively unimportant in the other varieties. More generally, we observe a clear similarity among the native varieties in the overall trajectory of their predictor rankings.

From a variety-centric perspective, there are some suggestive patterns within these rankings worth pointing out. For one, the effect of **Register.sum** plays a relatively less important role in Outer Circle varieties than in the Inner Circle varieties, where it is always one of the very highest-ranking predictors. Notable exceptions here are Irish and Indian English, where **Register.sum** also shows relatively high ranking compared to other predictors. Upon closer inspection, it appears this effect in IN is due to the high preponderance of split tokens in the unscripted spoken monologue subset of the data (25.2% compared to an average of 7.8% split tokens across the remaining genres). Notably, many of the texts in this genre come from yoga or exercise instructions, which frequently involve directions for positioning body parts and thus involve compositional PVs with concrete, definite direct objects, for example, *turn the left foot out, take your shoulder up, lift your hands up*. It turns out that almost one third (31.8%) of the split PV tokens in our entire IN dataset come from this one genre. What effect this may have on our models is hard to say at this point, though we note that the same model also reports an unusually high ranking for **DirObjConcreteness**.

We can further measure correlations between varieties' constraint rankings quantitatively (Table 1) using a rank correlation measure, specifically Spearman's ρ, in which scores range from -1 (perfect negative correlation) to 1 (perfect positive correlation, i.e. identical rankings). British, Canadian, New Zealand, and to a lesser extent Irish datasets tend to correlate more strongly with one another, although we also find strong correlations of GB and NZ with the Outer Circle variety PH. To the extent that shared rankings imply a shared grammar, we interpret this as evidence for a "core" PV variation grammar shared among Inner Circle Englishes. The remaining five varieties' rankings waver considerably in their correlations, both with the native varieties as well as among themselves. Of these, Indian English especially stands out, as it exhibits very weak correlations with most of the other varieties, with the exception of GB and IE. This is not necessarily true of the others, which exhibit varying degrees of similarity with different native and non-native varieties.

TABLE 1. *Spearman rank correlations of the variable importance rankings between each of the nine varieties*

	GB	CA	NZ	IE	JA	SG	HK	PH
CA	0.60							
NZ	0.79	0.77						
IE	0.91	0.48	0.59					
JA	0.72	0.64	0.82	0.49				
SG	0.45	0.23	0.66	0.21	0.54			
HK	0.55	0.41	0.37	0.62	0.45	0.14		
PH	0.85	0.58	0.82	0.71	0.76	0.69	0.55	
IN	0.42	-0.01	0.06	0.65	0.03	0.02	0.28	0.40

Table 2 shows the effects of each predictor level on the log odds scale as estimated by the individual GLMMs. Positive values indicate an increase, and negative values indicate a decrease in the probability of the split variant. A number of predictors such as **DirObjLength**, **Surprisal.P**, and **DirectionalPP**, behave consistently across all our varieties, as indicated by the

fact that the sign of their coefficients does not change. Other predictors are less consistent, though the differences in effect direction that we find mostly involve coefficients close to 0, and would probably fail to reach significance in many cases.

TABLE 2. *Estimates for fixed effects coefficients and random effects standard deviations in per-variety GLMMs. (* = p-values at < .05; ** = p-values at Bonferroni corrected < .005)*

Fixed effects	GB	CA	NZ	IE	JA	SG	HK	PH	IN
Intercept	-1.18*	-0.69	-1.18*	-1.37**	-2.86**	-1.04	-2.13**	-2.69**	-2.09**
Spok.Inf	1.15**	0.66**	1.07**	0.51**	0.09	0.42	0.25	0.62*	1.02**
Spok.For	0.25	0.57**	0.26	0.64**	0.17	-0.07	-0.01	-0.08	-0.56
Writ.Inf	-0.36	-0.04	-0.26	0.38	-0.24	0.40	0.21	-0.09	-0.07
Writ.For	-0.34	-0.64*	-0.48	-0.42	0.23	0.02	-0.17	-0.56	-0.12
DirObjLength	-1.76**	-2.06**	-1.91**	-1.52**	-1.50**	-1.69**	-1.37**	-2.57**	-1.72**
Semantics = non-comp.	-0.47*	-0.80**	-0.81**	-0.63**	-0.43	-0.56*	-0.49	-0.75*	0.30
DirObj = non-concrete	-0.48*	-0.29	-0.17	-0.61**	-0.15	-0.31	-0.52*	-0.22	-0.81*
DirObj = new	-0.42*	-0.37	-0.28	-0.50*	-0.04	0.15	-0.31	0.27	0.58
DirObj = indef	-0.91**	-0.72**	-0.84**	-0.57**	-0.83**	-0.91**	-0.54*	-1.06**	-0.84*
DirObj Thematicity	0.40*	0.67**	0.53**	0.16	0.21	0.64**	-0.01	0.31	0.20
DirectionalPP	1.91**	1.76**	2.43**	1.22**	2.26**	1.45**	0.97**	1.08**	0.02
CV.binary = other	-0.15	-0.70	-0.28	-0.30	0.33	-1.38	-0.47	-0.05	-0.70
Surprisal.P	1.23**	0.45	1.27**	1.05**	1.23**	1.20**	1.32**	1.54**	0.83*
Surprisal.V	-0.08	0.39	0.28	0.09	0.22	1.00*	0.06	0.09	0.55
Rhythm	-0.11	-0.53*	-0.18	-0.32	-0.31	-0.45	-0.42	-0.41	-0.42
PrimeType = Continuous	-0.26	-0.81*	-0.35	-0.41	0.04	-0.28	-0.50	-0.98*	-0.55
PrimeType = split	0.70*	0.90**	0.89*	0.55*	0.71*	0.52	1.67**	1.33**	1.35*

Random effect ***SD*s**									
VerbPart	0.38	0.68	0.76	0.44	0.86	0.96	0.85	0.00	0.68
Verb	0.97	0.90	1.07	1.03	0.80	0.70	1.03	1.19	0.00
Genre	0.29	0.00	0.42	0.00	0.00	0.17	0.00	0.00	0.14
Particle	0.61	0.64	0.48	0.40	0.56	1.02	0.41	0.30	0.74

In terms of our random effects, we find a considerable degree of lexical bias in our datasets, as we expected. The various verbs and verb-particle combinations are not distributed evenly across the split and continuous variants, and we again refer interested readers to our supplementary materials for complete details of the random effects parameters.

Variation grammars in multidimensional space

Having established that our models are satisfactory, we move on to the multivariate comparative analysis. For various reasons, we ignore the role of statistical significance in assessing similarities among varieties (see also Tagliamonte, 2013:152 n.5). Our aim rather is to provide a "bird's eye view" of probabilistic indigenization in particle placement. To do this we use NEIGHBOR-NET diagrams (Bryant & Moulton, 2004) to explore latent cross-varietal patterns in the PV variation grammars based on the predictor rankings and effect sizes.

The first step in this process involves calculation of pairwise distances between the varieties, similar to matrices of distances between cities found on geographical maps. Using the tables of predictor rankings and coefficient estimates, we calculate two sets of pairwise distances between varieties based on their constraint rankings and effect sizes. Tables 3 & 4 display the matrices: lower scores reflect greater similarity, and 0 indicates identical variable grammars. The correlation between these two distance matrices is moderate (Mantel test: $r = .58$, $p < .01$), suggesting that the two lines of evidence assessed here (constraint rankings and effect size

hierarchies) are converging toward the same underlying pattern, though still independent of one another.

TABLE 3. *Dissimilarity matrix derived from pairwise spearman correlations of predictor rankings in per-variety GLMMs*

	GB	CA	NZ	IE	JA	SG	HK	PH
CA	0.40							
NZ	0.21	0.23						
IE	0.09	0.52	0.41					
JA	0.28	0.36	0.18	0.51				
SG	0.55	0.77	0.34	0.79	0.46			
HK	0.45	0.59	0.63	0.38	0.55	0.86		
PH	0.15	0.42	0.18	0.29	0.24	0.31	0.45	
IN	0.58	1.01	0.94	0.35	0.97	0.98	0.72	0.60

TABLE 4. *Euclidean distances between varieties based on fixed effects coefficient estimates in per-variety GLMMs*

	GB	CA	NZ	IE	JA	SG	HK	PH
CA	1.57							
NZ	1.07	1.47						
IE	1.33	1.63	1.78					
JA	1.91	2.55	1.82	2.26				
SG	2.43	2.25	2.47	2.29	2.59			
HK	2.21	2.43	2.34	1.96	2.05	2.33		
PH	2.49	2.64	2.50	2.62	2.43	2.91	2.03	
IN	3.07	3.19	3.43	3.06	2.99	2.39	2.16	2.62

In the next step, we use the matrices to generate NeighborNet diagrams. Originally developed in bioinformatics to represent uncertainty in phylogenies and reticulate effects such as genetic recombination, NeighborNets are increasingly popular in linguistics (see, e.g., Wälchli, 2014), and are related to well-known hierarchical agglomerative cluster analysis methods (Aldenderfer & Blashfield, 1984). But unlike classical cluster analysis, NeighborNet diagrams can depict conflicting signals in the underlying distance matrices. This means in practice that when the data (or a segment of the data) can be adequately represented as a hierarchical tree, the

NeighborNet diagram (or the relevant segment in the diagram) will be tree-shaped; when the distances provide a conflicting signal, reticulations (which indicate the co-existence of several possible tree structures) will be drawn. The resulting boxy shapes are in biology often interpreted as being indicative of horizontal gene transfer, and in linguistics as suggesting language contact. We skip further technicalities and refer the reader to the introduction in Szmrecsanyi and Wolk (2011:574–577). Suffice it to say that we present NeighborNet diagrams without insisting on a strictly phylogenetic interpretation. Instead, we use the method essentially as an advanced visualization technique to highlight general trends in complex datasets and to summarize the similarities between the varieties under study.

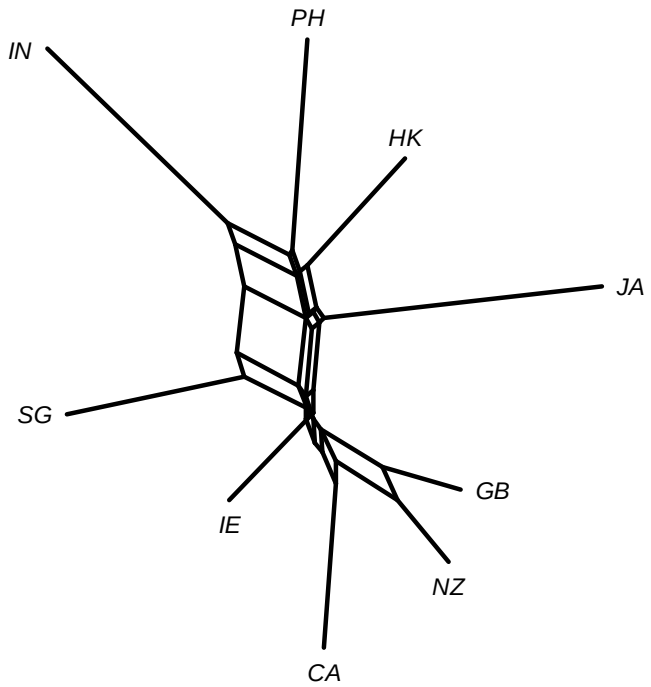

FIGURE 4. Visualizing variety relationships: NeighborNet diagram based on coefficient estimate distances. Internode distances (branch lengths) are proportional to linguistic distances.

Figure 4 depicts a NeighborNet network generated from the effect size-based distance matrix (Table 4). In NeighborNet diagrams, each line represents a way of splitting the total set of

varieties into exactly two groups. The longer a given line or set of lines, the greater the difference between the groups. For example, the comparatively large, almost vertical set of lines directly above the point where Irish English (IE) joins the network divides the varieties into the following two groups: one group that consists of Philippines, Jamaican, Indian, Hong Kong, and Singapore English (all Outer Circle varieties); and another cluster consisting of the other varieties in the bottom of the diagram (all the Inner Circle varieties in the sample). This split is fairly straightforward and represents, in fact, the most important groupings in the diagram. When divisions cannot be represented as strictly hierarchical, the algorithm draws reticulations which result in boxy shapes. Against this backdrop, the diagram indicates that variety relationships within the two big clusters are not strictly hierarchical. We conclude that the two clusters in Figure 1 can be interpreted primarily in terms of variety type (Inner vs. Outer Circle). The length of the lines additionally suggests that the Outer Circle cluster is internally less homogeneous than the Inner Circle cluster.

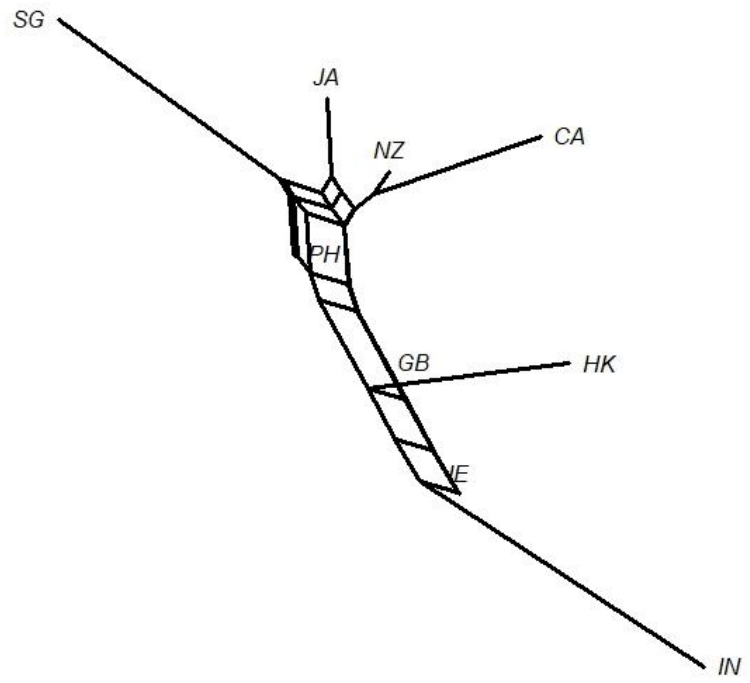

FIGURE 5. Visualizing variety relationships: NeighborNet diagram based on constraint ranking distances. Internode distances (branch lengths) are proportional to linguistic distances.

The NeighborNet in Figure 5 visually summarizes the distance matrix based on variable importance ranking (Table 3). Compared to the NeighborNet in Figure 4, the diagram in Figure 5 has somewhat fewer reticulations, which is another way of saying that the underlying distance matrix can be split up in a more hierarchical, tree-like fashion. The main split in the dataset distinguishes between two clusters: a five-member group in the upper half of the diagram joining Philippines, Singapore, Jamaican, New Zealand, and Canadian English (with New Zealand and Canadian English, the two L1 varieties in this cluster, being rather close); and a four-member cluster at the bottom of the diagram consisting of British, Hong Kong, Irish, and Indian English (with Indian English, as in Figure 1, being fairly distant from the other varieties). These two clusters are fairly mixed bags: neither variety type nor geography seem to be decisive. But by

way of an ad hoc explanation, we note that the upper cluster has a rather North American bent, while the bottom cluster seems to favor varieties located in the British Isles or with an orientation towards British English (such as Indian English).

Discussion

Our goals in this study were two-fold. One was to investigate particle placement across several World English varieties—do the same constraints influence the choice between PV variants for speakers of these varieties, and if not, how and where do the underlying variable grammars differ? We view these questions about the global scope of particle placement primarily in terms of its degree of probabilistic indigenization, which we characterize as the process whereby numerous social, cognitive, and/or linguistic forces shape the development of region-specific quantitative patterns in different varieties' PV grammars. Our second goal was to expand the analytical toolkit for Comparative Sociolinguistics by applying distance metrics to the outputs of predictive models used in variationist analysis. By taking advantage of complementary methods from variationist and corpus linguistic traditions, we were able to show that there is in fact a high degree of stability among Inner Circle varieties in terms of the strength, direction, and ranking of the linguistic constraints on particle placement, while the patterns among different constraints were more variable among Outer Circle varieties. Moreover, our NeighborNet analysis suggests that the direction (and possibly the pace) of probabilistic indigenization in particle placement is far from uniform. How might we make sense of these patterns?

As with a number of recent studies of syntactic alternations across English varieties (e.g., Deshors & Gries 2016; Heller, Szmrecsanyi & Grafmiller, 2017; Röthlisberger, Grafmiller, & Szmrecsanyi, 2017), the present study is couched within a theoretical framework that assumes an

experience-based, probabilistic model of language. Our choice of predictors was motivated not only by what had been found in prior work on particle placement, but also by psycholinguistic theories of language production (e.g., Bock & Warren, 1985; Hawkins, 2004; MacDonald, 2013), giving us a reasonable starting point for analyzing particle placement in the lesser studied Outer Circle varieties. Our results are largely consistent with functional accounts of language production, in which word order variation is thought to be governed by language users' implicit drive to minimize processing load during online production planning. For example, the effect of **DirObjLength**, that is, "end weight"—often considered a classic manifestation of such processing constraints in human language (see, e.g., Futrell, Mahowald, & Gibson, 2015)—is one of the strongest and highest-ranking predictors in every variety, and it exerts exactly the influence we expected: the longer the direct object, the less likely the split variant. Similarly motivated constraints such as **Surprisal.P** also exhibit considerable stability across our datasets, while others related to information status (**DirObjGivenness)** and discourse topicality **(DirObjThematicity**) appear to be reliably unimportant—a rather unexpected finding given previous research (e.g., Dehé, 2002; Haddican & Johnson, 2012). Notably, the relative ranking of all constraints is most varied among Outer Circle varieties, and most consistent among Inner Circle varieties. To the extent that certain constraints exhibit consistently strong or weak effects across varieties, we believe these effects form part of a "common core" PV grammar that is likely to be found in all varieties of English.

These findings raise a number of thorny questions, to which we suggest answers here. Perhaps the most obvious question involves the contrast between Inner and Outer Circle varieties in our analyses. Why are native varieties so similar to each other, while the Outer Circle varieties are so dissimilar, both to the native varieties and to one another? Conventional wisdom in World Englishes research maintains that varieties at more advanced stages of nativization tend to be

more differentiated from modern British English (e.g., Mukherjee & Gries, 2009); however, recent research on cross-varietal syntactic alternations suggests a more complex picture (e.g., Deshors & Gries, 2016; Heller et al., 2017; Röthlisberger et al., 2017). For example, the Dynamic Model (Schneider, 2007) places post-colonial varieties along 5 phases (5 being the final phase 'differentiation') according to characteristic correlations between the emergence of localized linguistic patterns and the changing sociocultural dynamics of the region. Sociolinguistically, the phases form a cline reflecting, among other things, the prevalence of bilingualism within the population, the degree of English integration into the local linguistic repertoire, the establishment of local norms or standards, and the extent of intra-regional variation. Linguistically, the phases are manifested in an increasing deployment of region-specific vocabulary, collocations, and especially, novel morphosyntactic frames (Schneider, 2007:78-85). For example, Schneider and Zipp (2013) report on a number of innovative uses of PVs in Fijian English and Indian English, including forms such as *rid off* (16) and *involve into* (17).

(16) One of the side effects of alcohol is that it **rids** our body **off** nutrients […] (from Schneider & Zipp, 2013: ex.6)

(17) Some of these are; women **involving** themselves **into** prostitution […] (from Schneider & Zipp, 2013: ex.8)

Among the varieties included here, CA, NZ, and IE fall squarely into phase 5, SG and JA are considered advanced stage 4 ('endonormative stabilization'), and HK, IN, and PH are grouped into phase 3 ('nativization'), though the latter two may be considered transitional between 3 and 4 (Schneider, 2007). Our results corroborate those of other recent alternation studies: less advanced varieties (phases 3-4) tend to be more different from their mother variety, that is, British English in most cases (Deshors & Gries, 2016; Heller et al., 2017; Röthlisberger et al., 2017).[6] Generally speaking, the Dynamic Model predicts greater drift away from the mother

variety as a new variety proceeds along the cycle, yet this does not seem to apply to languages at the level of the abstract probabilistic grammar.

But we can speculate further about why we might find a split among varieties along the lines of "nativeness" when we look at regional variation in the probabilistic conditioning of this particular variable. Simplifying somewhat, the varieties that we have been characterizing as Outer Circle are alike in that they are—or were at some prior stage—predominantly acquired and used as a second language. It thus seems plausible that the patterns we see in our data are at least partly attributable to biases in L2 acquisition. Despite being quite common and productive in English, PVs are notoriously challenging for L2 learners, especially those whose first language lacks phrasal verbs (Blais & Gonnerman, 2013; Gardner & Davies, 2007; Schmitt & Redwood 2011), and learners are known to avoid using PVs when they can (Liao & Fukuya, 2004; Siyanova & Scmidtt, 2007). When learners do use PVs, they tend to over-use the continuous variant in contexts where native speakers might not use it (Gilquin, 2014). Crucially however, the successful acquisition of a fully productive syntactic alternation requires a sufficient degree of type diversity among the lexical fillers in both competing variants (Hoffmann, 2014; Perek & Goldberg, 2015; Wonnacott, 2011). In essence, "the more verbs that are witnessed alternating between two [variants], the more learners are willing to assume that other verbs can alternate as well" (Perek & Goldberg, 2015:123). What we may be seeing in our data then is the result of a decades, even centuries-long process in which widespread general avoidance of particle verbs among L2 learners, coupled with a tendency toward regularization of the continuous variant, led to a substantial reduction in the diversity of fully interchangeable verbs in these varieties. The net effect of this process is a high degree of variational asymmetry among individual verbs where one variant, in this case the split variant, becomes strongly associated with a much narrower range of verbs, leading to an overall dispreference for the split variant in the developing variety.

Changes in the uses and associations of specific PVs in specific variants will likely lead to changes in the probabilistic associations among those variants and the higher-level constraints operative in speakers' PV grammar. Many of these changes may be difficult to predict in the earliest phases and highly variable across varieties, yet we may be able to detect broad trends at latter stages, just as we do in our data. While we cannot possibly explore every aspect of the sociolinguistic history of our nine varieties here, we further note that the kind of structured heterogeneity in particle placement we observe at the global level fits in with a growing body of research highlighting the interrelated roles of L2 acquisition and socio-demographic factors in shaping language evolution (e.g., Bentz & Winter, 2013; Dale & Lupyan, 2014). Overall, it appears that change at the level of abstract probabilistic conditioning is relatively conservative in typical contexts of first language transmission (see Labov, 2007), but it is particularly sensitive to disruptions in this process, for example, when the language passes through a kind of second language acquisition "filter".

The scenario just described is commensurate with probabilistic experience-based approaches to grammar, yet it does not fully explain our findings. While we do find some PV variability across our varieties, the scope of this variability pales in comparison to the variability in global Englishes observed at other levels of structure, most notably lexis and phonetics/phonology. Once again, this relative stability in variable grammars has also been demonstrated for numerous syntactic phenomena, for example, datives, genitives, and *to* versus -*ing* complementation (Bernaisch et al., 2014; Deshors & Gries, 2016; Heller et al., 2017; Röthlisberger et al., 2017). We see two plausible explanations for why this should be: shared history and shared humanity.

The shared history explanation appeals to the common ancestry of our varieties. Because these varieties have developed from the same mother variety, they have inherited many of the

same syntactic forms, and even the same underlying grammatical systems. The fact that the "native" varieties NZ, CA, and IE exhibit the highest degree of similarity with British English would seem to lend support to this explanation, as Canadian and New Zealand English evolved from settlement colonies in which native speakers were directly transplanted from the home country. Shared history is also manifest in the length and depth of continued contact with Great Britain as these countries have developed post-independence, hence the similarities among British and Irish English.

Shared humanity, on the other hand, refers to the influence of universal, putatively inherent biases in production and comprehension that are thought to influence language structure (e.g., Culbertson, Smolensky, & Legendre, 2012; Hawkins, 2004; MacDonald, 2013; Tamminga, MacKenzie, & Embick, 2016). Theories come in many flavors, but they all converge on the same basic idea: the cumulative effect of universal biases in language processing gives rise to the typological patterns we observe across languages and phenomena. Since all language users are sensitive to these biases, we expect similar general tendencies to predominate as structures that satisfy users' biases become more frequent. At the same time, the periodic emergence of novel (uses of) lexical items and or semantic/pragmatic functions will lead to fluctuations in the quantitative associations between specific variants and features of the linguistic context which users implicitly learn (Perek & Goldberg, 2015; Wonnacott, 2011). The cognitive biases and learned lexical/functional associations are themselves probabilistic, hence variable, which leads to the development of the constrained variation in probabilistic indigenization that we observe in our study. Our results can thus be taken as evidence for a model of grammar which is shaped both by distributional patterns in the input as well as deep-seated cognitive biases. While purely distributional learning may allow for any possible pattern to emerge, for example, new before given, long before short, (probabilistic) biases in language acquisition and processing will over

time cause certain general tendencies to prevail, while still leaving room for some variability in the degree to which these tendencies are realized in the grammar, that is, their effect size. This is just the pattern of probabilistic indigenization that we find.

Finally, when positing possible explanations for novel ESL patterns, a usual suspect that we have not discussed thus far is the influence of substrate effects. To the extent that the substrate L1 plays a role, its strongest influence may lie in the difficulty imbued by different L1s on the acquisition of particle verbs, though we cannot rule out more direct L1 to L2 transfer of probabilistic cues (e.g., MacWhinney, 1997). In general, we note that more detailed investigation of individual Outer Circle varieties is sorely needed. In particular, diachronic studies of particle placement in the Outer Circle would provide far better resolution of the historical picture, as well as offer testing ground for explanations of probabilistic indigenization in L2 varieties. Of course, such topics require far more detailed language-specific investigation than we can provide here.

CONCLUDING REMARKS

Our study charts the regional variability in the particle placement alternation in English around the globe, conducting a comparative analysis of variation in particle placement in a corpus of spoken and written language data from numerous registers and genres across nine regional varieties. Using several statistical techniques, we adapted and extended methodologies developed in the context of Comparative Sociolinguistics to examine the emergence of region-specific patterns in the underlying grammatical system of particle placement in these nine varieties, exemplifying a process we call *probabilistic indigenization*. Because probabilistic indigenization is shaped by numerous competing social and linguistic forces, it proceeds in various places at variable paces, yet we were able to identify several trends in our analyses by focusing on the

strength, direction, and ranking of probabilistic constraints as key dimensions for distinguishing among the varieties' PV grammars. The most notable finding was the high degree of homogeneity in the grammars of the native language varieties in our study, namely British, Canadian, New Zealand, and Irish English. Greater heterogeneity was found among Outer Circle varieties, and none of these varieties were found to be very similar to the native varieties. We (speculatively) attribute this pattern among Outer Circle varieties to the complex interaction between the simplification processes due to second language acquisition and the varying cultural, political, and sociolinguistic ecologies of the individual regions. Our findings also largely accord with a probabilistic grammar approach, which assumes that grammatical knowledge includes implicit knowledge of quantitative patterns over constraints.

On a final note, we want to highlight the potential of our methodology. One of the chief advantages of our method lies in its relative ease of interpretation. Compared to large tables of numbers, techniques such as NeighborNet diagrams provide visualizations of the relationships between variation grammars that are immediately intuitive. Communicating the main findings thus requires relatively little exegesis. A second advantage is that the method can be applied to any kind of sociolinguistic variable—lexical, phonetic/phonological, etc.—and it can be scaled up to comparisons of larger and larger numbers of datasets. Another innovation here is the focus on exploratory statistical methods rather than confirmatory analysis, that is, significance tests, as a primary metric for comparison. If the aim of comparative studies is to discern meaningful differences between varieties' grammars based on quantitative distributions among variable constraints, we need a theoretical model of intra-community (or speaker) variability and inter-community effect size in order for significance tests to be truly meaningful. With this in mind, we argue that the substantial similarities in constraint effects, especially among Inner Circle varieties, nonetheless represent strong evidence for the existence of a shared PV grammar, and

more important, that the similarities among variable grammars are more appropriately analyzed and interpreted as a continuum rather than distinct grammars.

NOTES

1. Example tokens come from two sources: The International Corpus of English (ICE), and the Corpus of Global Web-Based English (GloWbE).
2. These are also sometimes referred to as ‘Post-colonial’ or ‘New’ Englishes (e.g., Kachru, 1992; Schneider, 2007).
3. We point out that the varieties included here were selected primarily for convenience—they were the complete ICE components available to us at the onset of the project.
4. See *PV_annotation_manual.pdf* at https://osf.io/x8vyw/
5. All statistical analyses were conducted in R. Code and data are available at https://osf.io/ev2ck/
6. The one exception here is the Philippines, which was an American rather than British colony.

APPENDIX

TABLE A.1. *Structure of the ICE corpus. Numbers refer to the number of texts comprising each (sub)category*

Spoken (300)	Dialogues (180)	Private (100)	face-to-face phone calls
		Public (80)	classroom lessons broadcast discussion broadcast interviews parliamentary debates legal cross-examinations business transactions
	Monologues (120)	Unscripted (70)	spontaneous commentaries unscripted speeches demonstrations legal presentations
		Scripted (50)	broadcast news broadcast talks non-broadcast talks
Written (200)	Non-printed (50)	Student writing (20)	essays examinations
		Letters (30)	social letters business letters
	Printed (150)	Academic (40)	humanities social sciences natural sciences technology
		Popular nonfiction (40)	humanities social sciences natural sciences technology
		Press reporting (20)	press news
		Instructional (20)	administrative skills/hobbies
		Press editorials (10)	press editorials
		Creative fiction (20)	novels & short stories

REFERENCES

Aldenderfer, Mark & Blashfield, Roger. (1984). *Cluster analysis*. Thousand Oaks, CA: SAGE Publications, Inc.

Bentz, Christian & Winter, Bodo. (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change* 3(1):1–27.

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

Biber, Douglas. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Blais, Mary-Jane & Gonnerman, Laura M. (2013). Explicit and implicit semantic processing of verb–particle constructions by French–English bilinguals. *Bilingualism: Language and cognition* 16(4):829–846.

Bock, J. Kathryn & Warren, Richard K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition* 21(1):47–67.

Branigan, Holly P., Pickering, Martin J., & Tanaka, Mikihiro. (2008). Contributions of animacy to grammatical function assignment and word order during production. *Lingua* 118(2):172–189.

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

Bresnan, Joan & Ford, Marilyn. (2010). Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1):168–213.

Bresnan, Joan & Hay, Jennifer. (2008). Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118(2):245–259.

Browman, Catherine P. (1986). The hunting of the quark: The particle in English. *Language and Speech* 29(4):311–334.

Bryant, D. & Moulton, Vincent. (2004). Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2):255–265.

Bybee, Joan. (2006). From usage to grammar: The mind's response to repetition. *Language* 82(4):711–733.

Cappelle, Bert. (2009). Contextual cues for particle placement. In A. Bergs & G. Diewald (eds.), *Contexts and constructions*. Amsterdam: John Benjamins. 145–192.

Culbertson, Jennifer, Smolensky, Paul, & Legendre, Géraldine. (2012). Learning biases predict a word order universal. *Cognition* 122(3):306–329.

Dale, Rick & Lupyan, Gary. (2012). Understanding the origins of morphological diversity: the linguistic niche hypothesis. *Advances in Complex Systems* 15(03n04):1150017.

D'Arcy, Alexandra & Tagliamonte, Sali A. (2015). Not always variable: Probing the vernacular grammar. *Language Variation and Change* 27(3):255–285.

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

Dehé, Nicole. (2002). Particle verbs in English: Syntax, information structure, and intonation. Amsterdam: John Benjamins.

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

Diessel, Holger & Tomasello, Michael. (2005). Particle placement in early child language: A multifactorial analysis. *Corpus Linguistics and Linguistic Theory* 1:89–112.

Elenbaas, Marion. (2007). *The synchronic and diachronic syntax of the English verb-particle combination*. Doctoral dissertation, Radboud University.

Elenbaas, Marion. (2013). Motivations for particle verb word order in Middle and Early Modern English. *English Language and Linguistics* 17(3):489–511.

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

Futrell, Richard, Mahowald, Kyle & Gibson, Edward. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences* 112(33):10336-41.

Gardner, Dee & Davies, Mark. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly* 41(2):339–359.

Garretson, Gregory. (2004). Coding practices used in the project Optimality Typology of Determiner Phrases. Unpublished manuscript, Boston University, Boston, MA. http://npcorpus.bu.edu/documentation/BUNPCorpus_coding_practices.pdf

Gilquin, Gaëtanelle. (2014). The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory* 11(1):51–88.

Goldberg, Adele E. (2016). Tuning in to the verb-particle construction in English. In L. Nash & P. Samvelian (eds.), *Approaches to complex predicates*. Boston, MA: Brill. 110-141.

González, Rafael Alejo. (2010). Making sense of phrasal verbs: A cognitive linguistic account of L2 learning. *AILA Review* 23(1):50–71.

Greenbaum, Sidney. (1996). *Comparing English worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Gries, Stefan Th. (2003). *Multifactorial analysis in corpus linguistics: A study of particle placement*. New York: Continuum Press.

Gries, Stefan Th. (2011). Acquiring particle placement in English: A corpus-based perspective. In P.G. Medina (ed.), *Morphosyntactic alternations in English: functional and cognitive perspectives*. London & Oakville, CT: Equinox. 236–263.

Gries, Stefan Th. & Kootstra, Gerrit Jan. (2017). Structural priming within and across languages: a corpus-based perspective. *Bilingualism: Language and Cognition* 20(2):235–250.

Grieve, Jack. (2016). *Regional variation in written American English*. Cambridge: Cambridge University Press.

Haddican, Bill & Johnson, Daniel Ezra. (2012). Effects on the particle verb alternation across English dialects. *University of Pennsylvania working papers in linguistics 18*. University of Pennsylvania. 31–40.

Hawkins, John A. (2004). *Efficiency and complexity in grammars*. Oxford University Press.

Heller, Benedikt, Szmrecsanyi, Benedikt, & Grafmiller, Jason. (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 & Szmrecsanyi, Benedikt. (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.

Hoffmann, Thomas. (2014). The cognitive evolution of Englishes: The role of constructions in the Dynamic Model. In S. Buschfeld, T. Hoffmann, M. Huber & A. Kautzsch (eds.), *Varieties of English around the world*, vol. G49. Amsterdam: John Benjamins Publishing Company. 160–180.

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

Labov, William. (2007). Transmission and diffusion. *Language* 83(2):344–387.

Liao, Yan & Fukuya, Yoshinori J. (2004). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning* 54(2):193–226.

Lohse, Barbara, Hawkins, John A., & Wasow, Thomas. (2004). Domain minimization in English verb-particle constructions. *Language* 80(2):238–261.

MacDonald, Maryellen C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology* 4:1–16.

MacWhinney, Brian. (1997). Second language acquisition and the competition model. In J.F. Kroll & A.M.B. de Groot (eds.), *Tutorials in bilingualism.* Mahwah, N.J: Lawrence Erlbaum. 113–142.

Mukherjee, Joybrato & Gries, Stefan Th. (2009). Collostructional nativisation in new Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide* 30(1):27–51.

Osselton, Noel E. (1988). Thematic genitives. In G. Nixon & J. Honey (eds.), *An historic tongue: Studies in English linguistics in memory of Barbara Strang*. London: Routledge.

Paolillo, John C. (2013). Individual effects in variation analysis: Model, software, and research design. *Language Variation and Change* 25(1):89–118.

Perek, Florent & Goldberg, Adele E. (2015). Generalizing beyond the input: The functions of the constructions matter. *Journal of Memory and Language* 84:108–127.

Poplack, Shana & Tagliamonte, Sali. (2001). *African American English in the diaspora*. Malden, MA: Blackwell.

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

Rodríguez-Puente, Paula. (2016). Particle placement in Late Modern English and Twentieth-century English: Morpho-syntactic variables. *Folia Linguistica* 37(1):145-175.

Röthlisberger, Melanie, Grafmiller, Jason, & Szmrecsanyi, Benedikt. (2017). Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics* 28(4):673-710.

Schmitt, Norbert & Redwood, Stephen. 2011. Learner knowledge of phrasal verbs: A corpus-informed study. In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (eds.), *A taste for corpora: In honour of Sylviane Granger*. Amsterdam and Philadelphia: John Benjamins. 173–207.

Schneider, Edgar. (2007). *Postcolonial English: Varieties around the world*. New York: Cambridge University Press.

Schneider, Gerold & Zipp, Lena. (2013). Discovering new verb-preposition combinations in New Englishes. *Studies in Variation, Contacts and Change in English* 13:online.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27(3):379–423.

Shih, Stephanie S. (2017). Phonological Influences in syntactic alternations. In V. Gribanova & S.S. Shih (eds.), *The morphosyntax-phonology connection*. Oxford University Press. 223–252.

Siyanova, Anna & Schmitt, Norbert. (2007). Native and nonnative use of multi-word vs. one-word verbs. *IRAL - International Review of Applied Linguistics in Language Teaching* 45(2):119-39.

Szmrecsanyi, Benedikt, Grafmiller, Jason, Bresnan, Joan, Rosenbach, Anette, Tagliamonte, Sali, & Todd, Simon. (2017). Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa: a journal of general linguistics* 2(1):online.

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

Szmrecsanyi, Benedikt & Wolk, Christoph. (2011). Holistic corpus-based dialectology. *Revista Brasileira de Linguística Aplicada* 11(2):561–592.

Tagliamonte, Sali. (2013). Comparative Sociolinguistics. In J.K. Chambers & N. Schilling (eds.), *Handbook of language variation and change*, 2nd ed. Chichester, West Sussex, United Kingdom: John Wiley & Sons Inc. 130–156.

Tamminga, Meredith, MacKenzie, Laurel & Embick, David. (2016). The dynamics of variation in individuals. *Linguistic Variation* 16(2):300–336.

Thim, Stefan. (2012). *Phrasal verbs: The English verb-particle construction and its history*. Berlin: Mouton de Gruyter.

Wälchli, Bernhard. (2014). Algorithmic typology and going from known to similar unknown categories within and across languages. In B. Szmrecsanyi & B. Wälchli (eds.), *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*. Berlin: De Gruyter. 355–393.

Wasow, Thomas. (2002). *Postverbal behavior*. Stanford: CSLI Publications.

Wonnacott, Elizabeth. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language* 65(1):1–14.

Zipp, Lena & Bernaisch, Tobias. (2012). Particle verbs across first and second language varieties of English. In M. Hundt & U. Gut (eds.), *Mapping unity and diversity world-wide: corpus-based studies of New Englishes.* Amsterdam: John Benjamins. 167–196.