Probabilistic variation in a comparative perspective: the grammar of varieties of English

Benedikt Szmrecsanyi, Jason Grafmiller, Benedikt Heller & Melanie Röthlisberger

KU Leuven
Quantitative Lexicology and Variational Linguistics

Introduction

A new project

- "Exploring probabilistic grammar(s) in varieties of English around the world" (5-year project, 2013–2018)
- synthesize disjoint lines of scholarship into one unifying project with a coherent empirical and theoretical focus
- main goal: understand the plasticity of probabilistic knowledge of English grammar, on the part of language users with diverse regional and cultural backgrounds

A new project

- "Exploring probabilistic grammar(s) in varieties of English around the world" (5-year project, 2013–2018)
- synthesize disjoint lines of scholarship into one unifying project with a coherent empirical and theoretical focus
- main goal: understand the plasticity of probabilistic knowledge of English grammar, on the part of language users with diverse regional and cultural backgrounds

- today: variation across three syntactic alternations $\times$ four international varieties of English

The "English World-Wide Paradigm"

- native varieties (e.g. British E), indigenized L2 varieties (e.g. Hong Kong E), shift varieties (e.g. Irish E), ...
- topics: scope, limits, parameters of variation; extent to which structural make-up of varieties of E can be predicted by communicative needs of colonizers/colonized
  (e.g. Kachru 1992; Schneider 2007; Mesthrie and Bhatt 2008)
- shortcoming: an often primarily descriptive interest in the variable presence/absence of features, or in usage frequencies of features

The Probabilistic Grammar framework

- rely on the variation-centered, usage- and experience-based probabilistic grammar framework developed by Joan Bresnan and collaborators

  (e.g. Bresnan 2007; Bresnan and Ford 2010; Wolk et al. 2013)

  1. syntactic variation – and change – is **subtle, gradient & probabilistic** rather than categorical in nature

     (Labov 1982; Bresnan and Hay 2008)
  2. linguistic knowledge includes **knowledge of probabilities**, and speakers have powerful predictive capacities

     (Gahl and Garnsey 2004; Gahl and Yu 2006)

Some research questions

- scope and limits of variation – do the varieties of English we study here share a core probabilistic grammar?
- dialect typology – does variety type (e.g. native versus non-native) predict probabilistic similarity between varieties of English?
- variation phenomena – do the alternations under study differ in terms of their probabilistic sensitivity to variety effects?

Method & Data

A methodological sketch of the project

1. tap into corpus data to explore 3 syntactic alternations across 9 varieties

A methodological sketch of the project

1. tap into corpus data to explore 3 syntactic alternations across 9 varieties
2. use the variationist method (Labov 1982) to create richly annotated corpus-derived datasets . . .

A methodological sketch of the project

1. tap into corpus data to explore 3 syntactic alternations across 9 varieties
2. use the variationist method (Labov 1982) to create richly annotated corpus-derived datasets . . .
3. . . . to study interplay of probabilistic factors constraining the alternations; check whether there are significant differences between varieties

A methodological sketch of the project

1. tap into corpus data to explore 3 syntactic alternations across 9 varieties
2. use the variationist method (Labov 1982) to create richly annotated corpus-derived datasets . . .
3. . . . to study interplay of probabilistic factors constraining the alternations; check whether there are significant differences between varieties

4. last stage of the project: significant & interesting differences according to corpus data
   ➪ conduct supplementary rating-task experiments

Varieties of English

- **British E**, **Canadian E**, **Indian E**, **Singapore E**, Irish E, New Zealand E, Hong Kong E, Jamaican E, Philippines E
- corpus database: 1.5m words of running text per variety, covering spoken written English (ICE), and (eventually) web-based language (GloWbE)

The genitive alternation

(1) a. [The Senator]$_{possessor}$'s [brother]$_{possessum}$
(the *s*-genitive)
b. [The brother]$_{possessum}$ of [the Senator]$_{possessor}$
(the *of*-genitive)

- variable context: identified *'s* & *of* occurrences; manually excluded e.g. partitive genitives and pronominal genitives

The dative alternation

(2) a. We sent [the president]$_{recipient}$ [a letter]$_{theme}$
(the ditransitive dative)
b. We sent [a letter]$_{theme}$ to [the president]$_{recipient}$
(the prepositional dative)

- variable context: used a list of dative verbs to identify occurrences; manually excluded e.g. passivized verbs, and elliptical structures

Particle placement

(3) a. The president looked$_{verb}$ [the word]$_{NP}$ up$_{particle}$
(V-DO-P – split pattern)
b. The president looked$_{verb}$ up$_{particle}$ [the word]$_{NP}$
(V-P-DO – unsplit pattern)

- variable context: transitive particle verbs involving one of the following 10 particles: *around*, *away*, *back*, *down*, *in*, *off*, *out*, *over*, *on*, *up*; manually excluded e.g. passive sentences and sentences with extracted direct objects

Findings

Some first findings

- three alternations × four varieties (BrE, CanE, IndE, and SgE)
- comparatively simple annotation
- two exploratory analysis techniques (conditional inference trees & random forests)
- non-web-based text types only

Annotation

- predictors across alternations:
  constituent length, constituent givenness, thematicity, TTR, overall frequency of head nouns, genre, variety
- alternation-specific predictors:
  e.g. presence of directional PPs after particle verb constructions, final sibilancy of genitive possessors, DO definiteness, NP expression type (common noun, proper noun, . . .)

Do the varieties of English we study here share a core probabilistic grammar?

- yes, in the sense that there clearly are variety-independent, qualitative generalizations
- the **effect directions** of factors are stable across varieties of English – but interesting differences with regard to **effect size**
- cross-variety differences only in contexts where neither alternate is more or less difficult to process

Particle placement: about length

(*look up* [*the difficult word*] vs *look* [*the difficult word*] *up*)

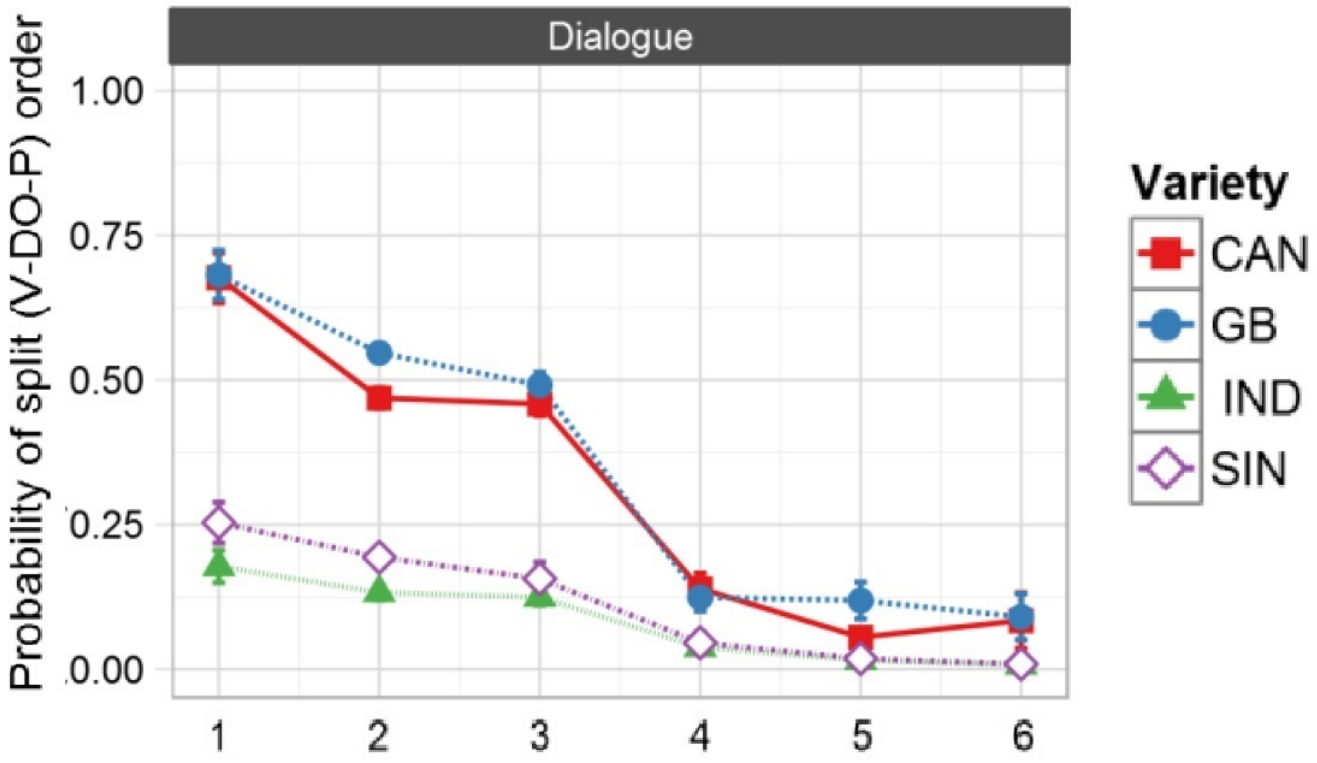


Figure: Predicted probabilities obtained from Conditional Random Forest model (with 95% confidence intervals)

Do we find a split between native and non-native varieties of English?

- in the particle placement alternation (and, to a lesser extent, the genitive alternation), varieties tend to pattern along native versus non-native lines
- in the dative alternation, IndE is set apart from the other varieties
- inconclusive

Dative alternation: conditional inference tree

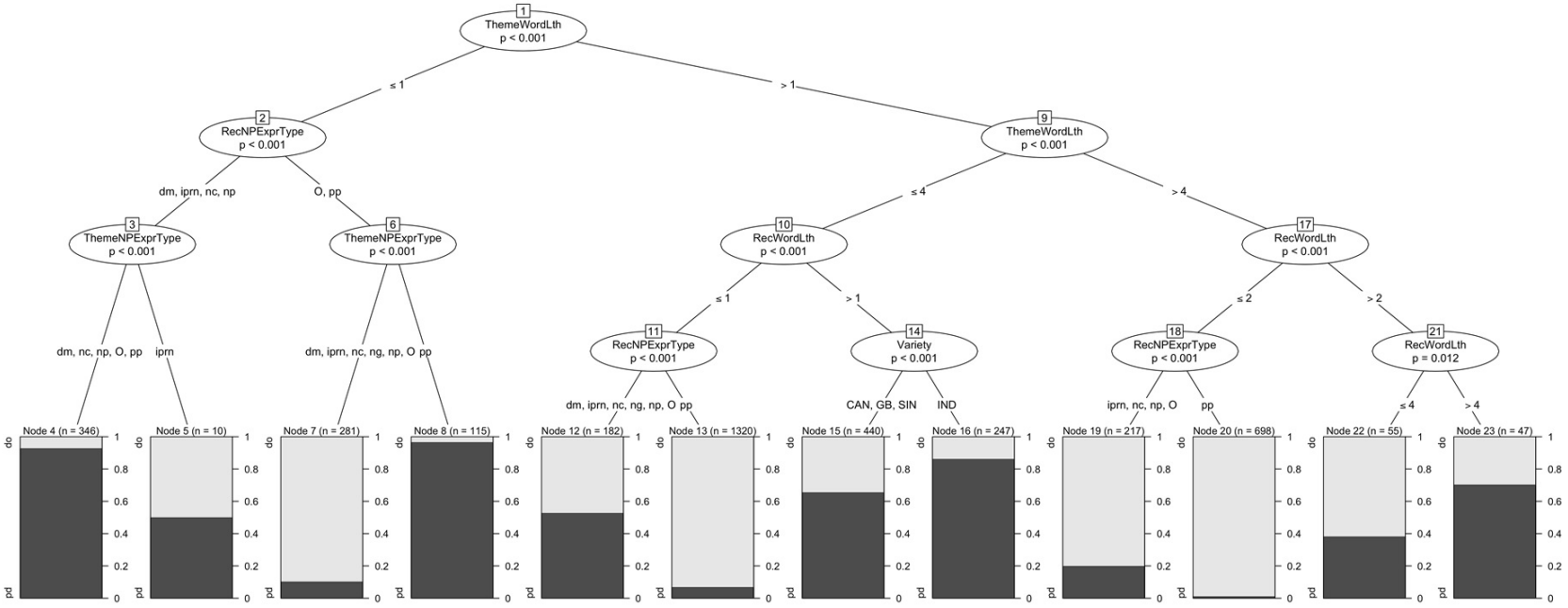
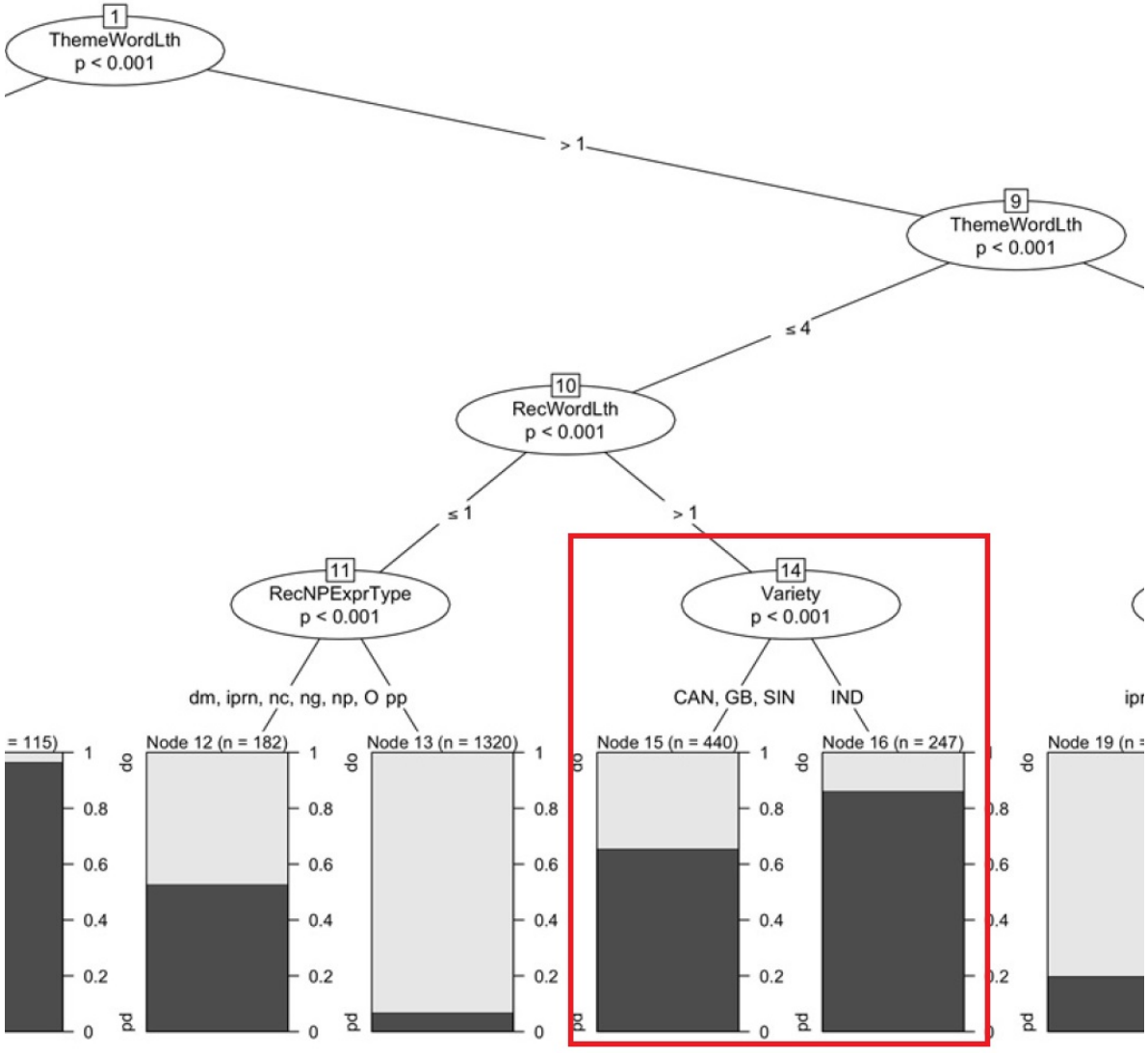


Figure: Conditional inference tree for dative choice

Accuracy: 87.1% (baseline: 68.2%); $C = 0.86$.

Dative alternation: conditional inference tree



Do the alternations under study differ in terms of their probabilistic sensitivity to variety effects?

amenability to "probabilistic indigenization":

- most amenable: particle placement
- least amenable: genitive alternation

Particle placement forest

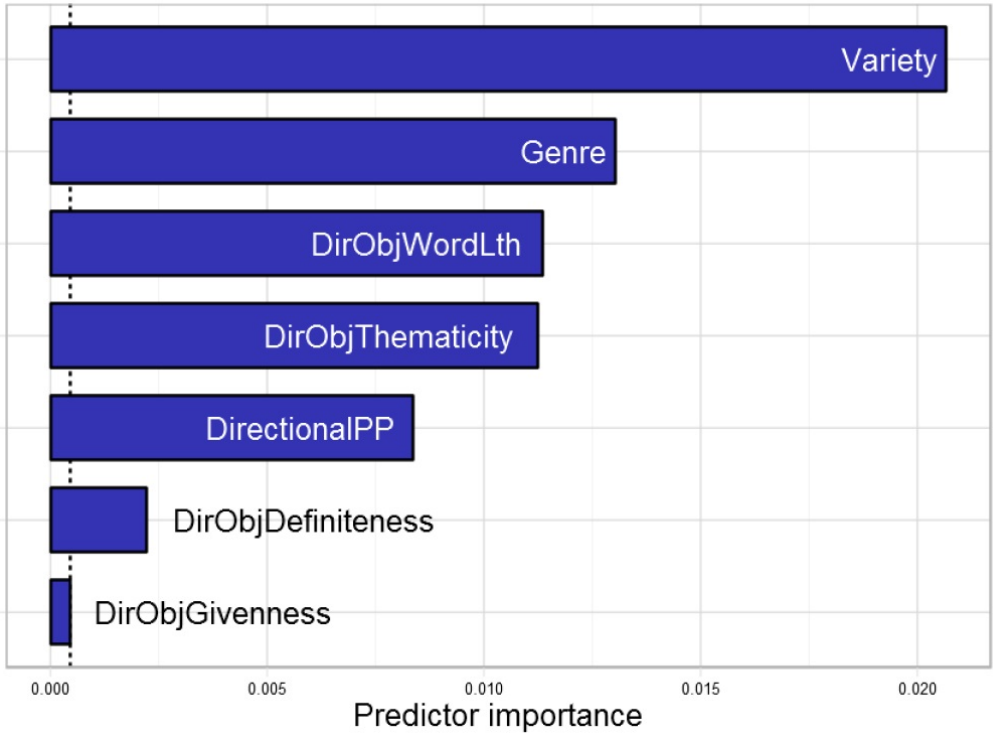


Figure: Predictor importance ranking for CRF analysis of particle placement. $C = 0.87$.

Genitive forest

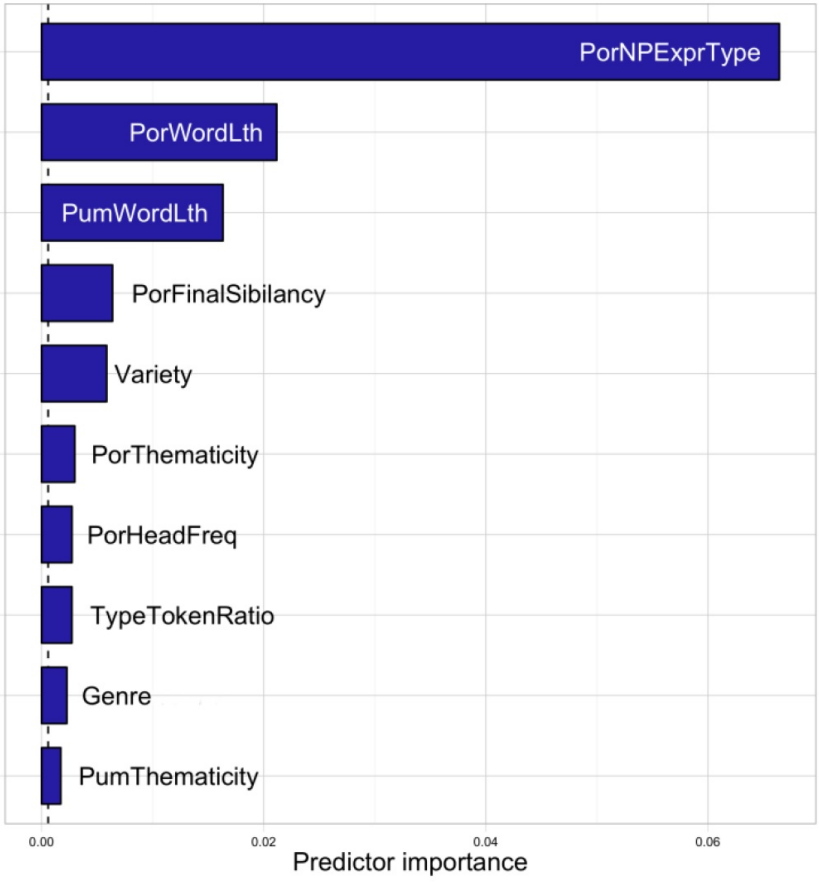


Figure: Predictor importance ranking for CRF analysis of genitive choice (displayed: 10 most important predictors). $C = 0.85$.

Do the alternations under study differ in terms of their probabilistic sensitivity to variety effects?

- Schneider (2003: 249): lexico-grammar is a prime target of early-stage indigenization
- tentative generalization: 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

  (e.g. Hoffmann 2014,Grafmiller forthcoming)

Concluding remarks

What's new?

- crossroads of research on English as a World Language, usage-based theoretical linguistics, variationist linguistics, and cognitive sociolinguistics

What's new?

- crossroads of research on English as a World Language, usage-based theoretical linguistics, variationist linguistics, and cognitive sociolinguistics
- interest in scope and limits of variation in a large-scale comparative perspective

What's new?

- crossroads of research on English as a World Language, usage-based theoretical linguistics, variationist linguistics, and cognitive sociolinguistics
- interest in scope and limits of variation in a large-scale comparative perspective
- assume that language users implicitly learn the probabilistic effects of constraints on variation by constantly (re-)assessing input of spoken and written discourses throughout their lifetimes

Team members

Jason Grafmiller
particle placement

Benedikt Heller
the genitive alternation

Melanie Röthlisberger
the dative alternation

Thank you!

benszm@kuleuven.be

http://wwwling.arts.kuleuven.be/qlvl/ProbGrammarEnglish.html

This presentation is based upon work supported by an Odysseus grant of the Research Foundation Flanders (FWO) (grant no. G.0C59.13N).

References I

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston and W. Sternefeld (Eds.), *Roots: Linguistics in Search of Its Evidential Base*, pp. 75–96. Berlin: Mouton de Gruyter.

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

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

Gahl, S. and S. Garnsey (2004). Knowledge of Grammar, Knowledge of Usage: Syntactic Probabilities Affect Pronunciation Variation. *Language 80*, 748–775.

Gahl, S. and A. C. Yu (2006). *Special theme issue: Exemplar-based models in linguistics*. The linguistic review. Mouton de Gruyter.

Grafmiller, J. Construction grammar goes global: Syntactic alternations, schematization, and collostructional diversity in world English(es).

Hoffmann, T. (2014). The cognitive evolution of Englishes: The role of constructions in the dynamic model. In S. Buschfeld, T. Hoffmann, M. Huber, and A. Kautzsch (Eds.), *Varieties of English Around the World*, pp. 160–180. Amsterdam: John Benjamins Publishing Company.

Kachru, B. B. (Ed.) (1992). *The Other tongue: English across cultures* (2nd ed ed.). English in the global context. Urbana: University of Illinois Press.

References II

Labov, W. (1982). Building on empirical foundations. In W. Lehmann and Y. Malkiel (Eds.), *Perspectives on Historical Linguistics*, pp. 17–92. Amsterdam, Philadelphia: Benjamins.

Mesthrie, R. and R. M. Bhatt (2008). *World Englishes: the study of new linguistic varieties*. Key topics in sociolinguistics. Cambridge, UK ; New York: Cambridge University Press.

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

Schneider, E. (2007). *Postcolonial English: Varieties Around the World*. Cambridge University Press.

Wolk, C., J. Bresnan, A. Rosenbach, and B. Szmrecsanyi (2013). Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica 30*(3), 382–419.