

Syntactic variation, probabilistic indigenization, and World Englishes

Melanie Röthlisberger, Jason Grafmiller, Benedikt Heller,
Benedikt Szmrecsanyi

KU Leuven
Quantitative Lexicology and Variational Linguistics

Linguists' Day 2016 of the Linguistic Society of Belgium

13 May 2016, Louvain-la-Neuve



Introduction

Introduction

- ▶ “Exploring probabilistic grammar(s) in varieties of English around the world” (5-year project, 2013–2018)
- ▶ syntactic choices within and across varieties of a given language are governed by language-internal forces that can exhibit subtle degrees of variability across regions
- ▶ qualitative stability (in effect direction) vs. probabilistic indigenization (in effect size)

- ▶ Syntactic variation,
- ▶ probabilistic indigenization,
- ▶ and World Englishes

- ▶ Syntactic variation,
 - ▶ dative alternation
 - ▶ genitive alternation
 - ▶ particle placement
- ▶ probabilistic indigenization,

- ▶ and World Englishes

- ▶ Syntactic variation,
 - ▶ dative alternation
 - ▶ genitive alternation
 - ▶ particle placement
- ▶ probabilistic indigenization,
 - ▶ syntactic variation is constrained probabilistically and speakers of diverse regional backgrounds reinterpret / indigenize the effects of these constraints
- ▶ and World Englishes

- ▶ Syntactic variation,
 - ▶ dative alternation
 - ▶ genitive alternation
 - ▶ particle placement
- ▶ probabilistic indigenization,
 - ▶ syntactic variation is constrained probabilistically and speakers of diverse regional backgrounds reinterpret / indigenize the effects of these constraints
- ▶ and World Englishes
 - ▶ large-scale comparative perspective

Research questions

- ▶ Do the varieties of English we study share a core probabilistic grammar?
- ▶ What are the constraints on variation that are particularly likely to be indigenized?

Research questions

- ▶ Do the varieties of English we study share a core probabilistic grammar?
- ▶ What are the constraints on variation that are particularly likely to be indigenized? → **end-weight**

End-weight effects

Behaghel's Gesetz der wachsenden Glieder: constituents tend to occur in order of increasing size or complexity (Behaghel 1910)

End-weight effects

Behaghel's Gesetz der wachsenden Glieder: constituents tend to occur in order of increasing size or complexity (Behaghel 1910)

- (1) In my laboratory we use it as an easily studied instance of mental grammar, allowing us to document

End-weight effects

Behaghel's Gesetz der wachsenden Glieder: constituents tend to occur in order of increasing size or complexity (Behaghel 1910)

- (1) In my laboratory we use it as an easily studied instance of mental grammar, allowing us to document in great detail

End-weight effects

Behaghel's Gesetz der wachsenden Glieder: constituents tend to occur in order of increasing size or complexity (Behaghel 1910)

- (1) In my laboratory we use it as an easily studied instance of mental grammar, allowing us to document in great detail the psychology of linguistic rules

End-weight effects

Behaghel's Gesetz der wachsenden Glieder: constituents tend to occur in order of increasing size or complexity (Behaghel 1910)

- (1) In my laboratory we use it as an easily studied instance of mental grammar, allowing us to document in great detail the psychology of linguistic rules from infancy to old age

End-weight effects

Behaghel's Gesetz der wachsenden Glieder: constituents tend to occur in order of increasing size or complexity (Behaghel 1910)

- (1) In my laboratory we use it as an easily studied instance of mental grammar, allowing us to document in great detail the psychology of linguistic rules from infancy to old age in both normal and neurologically impaired people, in much the same way that biologists focus on the fruit fly *Drosophila* to study the machinery of genes.

(Wasow, 1997: 81)

Why focus on end-weight?

- ▶ operative in many phenomena
- ▶ motivated by processing demands
- ▶ typologically robust & putatively universal

(e.g. Hawkins 1994)

Why focus on end-weight?

- ▶ operative in many phenomena
- ▶ motivated by processing demands
- ▶ typologically robust & putatively universal
(e.g. Hawkins 1994)

- ▶ evidence for **instability** across time and space
(e.g. Wolk et al. 2013, Bresnan and Ford 2010)

Why focus on end-weight?

- ▶ operative in many phenomena
- ▶ motivated by processing demands
- ▶ typologically robust & putatively universal
(e.g. Hawkins 1994)

- ▶ evidence for **instability** across time and space
(e.g. Wolk et al. 2013, Bresnan and Ford 2010)
→ **to what extent are end-weight effects cross-lectally variable?**

Method & Data

A methodological sketch

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

A methodological sketch

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

A methodological sketch

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

A methodological sketch

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

Varieties of English



Corpus

- ▶ International Corpus of English = ICE
- ▶ 500 texts à 2,000 words = 1 mio words of text per component
- ▶ 60% spoken, 40% written English
- ▶ 12 different registers
 - ▶ face-to-face conversations
 - ▶ broadcast discussions
 - ▶ unscripted speeches
 - ▶ exam scripts
 - ▶ academic and popular writing
 - ▶ ...

Genitive alternation

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

$N=10,594$

Dative alternation

- (3) 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)

$N=8,549$

Particle placement

10 particles: *around, away, back, down, in, off, out, over, on, up*

- (4) a. The president looked_{verb} [the word]_{NP} up_{particle}
(split order: V-Obj-P)
- b. The president looked_{verb} up_{particle} [the word]_{NP}
(joined order: V-P-Obj)

$N=8,072$



Predictors across alternations

- ▶ **constituent length**: in characters
- ▶ **constituent animacy**: ‘animate’ vs. ‘inanimate’
- ▶ **constituent givenness**: ‘given’ vs. ‘new’
- ▶ **constituent definiteness**: ‘def’ vs. ‘indef’
- ▶ **thematicity**: normalized text frequency
- ▶ **register**: spoken formal, spoken informal, written formal, written informal
- ▶ **variety**: CanE, BrE, HKE, IrE, IndE, ...

Mixed-effects logistic regression

- ▶ treatment coding: GB as reference level
- ▶ random effects included to account for idiosyncracies of speakers and corpus structure (Gries 2015)
 - ▶ corpus metadata
 - ▶ verbs, constituents
- ▶ bootstrap validation (Baayen 2008: 283)

Findings

Genitive alternation

Table: Interactions with Variety

Variety	P'or animacy	End-weight	Final sibilancy
HKE	–	+	+
NZE	–		
PhiE	–	+	
CanE		+	
IrE		+	
SinE		+	
IndE			+

(Predicted outcome: s-gen; reference level: GB)



Genitive alternation

Table: Interactions with Variety

Variety	P'or animacy	End-weight	Final sibilancy
HKE	—	+	+
NZE	—		
PhiE	—	+	
CanE		+	
IrE		+	
SinE		+	
IndE			+

(Predicted outcome: s-gen; reference level: GB)

Genitive alternation

Table: Interactions with Variety

examples

[Hong Kong]'s [fiscal reserves of seventy-six billion Hong Kong dollars] <ICE-HK:s1b-001>

[mud-caked face that had been human flesh] of [a Iraqi soldier] <ICE-GB:s1b-031>

IrE	+	
SinE	+	
IndE		+

(Predicted outcome: s-gen; reference level: GB)

Dative alternation

Table: Interactions with Variety

Variety	End-weight	RecPron
CanE		+
IndE		+
JamE	+	

(Predicted outcome: prepositional dative; reference level: GB)

Dative alternation

Table: Interactions with Variety

Variety	End-weight	RecPron
CanE		+
IndE		+
JamE	+	

(Predicted outcome: prepositional dative; reference level: GB)

Dative alternation

examples

offered [advice] to [other motorists and drivers of buggies]
<ICE-JA:w2b-032>

giving [people in high-valued property] [a subsidy]
<ICE-GB:s1b-034>

JAME +

(Predicted outcome: prepositional dative; reference level: GB)

Particle placement

Table: Interactions with Variety

Variety	End-weight	Idiom.	Concreteness	Givenness	Modpp
IndE				-	-
NZ		+	-		
PhiE	-		-		
SinE				-	

(Predicted outcome: V-Obj-P; reference level: GB)

Particle placement

Table: Interactions with Variety

Variety	End-weight	Idiom.	Concreteness	Givenness	Modpp
IndE				-	-
NZ		+	-		
PhiE	-		-		
SinE				-	

(Predicted outcome: V-Obj-P; reference level: GB)

Particle placement

examples

—	get	[back]	[your money]	<ICE-PHI:s1b-035:>
Val				
—	put	[stories of humanity passion and theatricality]	[back]	
ll				<ICE-GB:s1b-050>
INZ			+	—
PhiE	—			—
SinE				—

(Predicted outcome: V-Obj-P; reference level: GB)

Particle placement

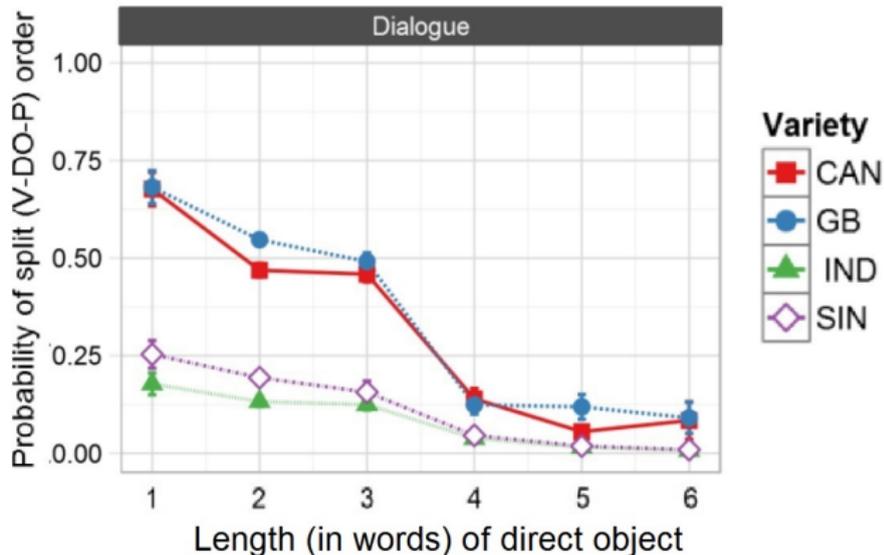


Figure: Predicted probabilities obtained from Conditional Random Forest model on corpus data (with 95% confidence intervals); from Szmrecsanyi et al. 2016

Discussion

Probabilistic indigenization of end-weight



Probabilistic indigenization of end-weight

- ▶ varieties do share a core probabilistic grammar: **effect directions** of factors are stable across varieties - but differences with regard to **effect size**

Probabilistic indigenization of end-weight

- ▶ varieties do share a core probabilistic grammar: **effect directions** of factors are stable across varieties - but differences with regard to **effect size**
- ▶ the probabilistic indigenization of end-weight effects
 - ▶ **genitive alternation**: effect size of length is stronger in CanE, HKE, IrE, PhiE and SinE
 - ▶ **dative alternation**: effect size of length is stronger in JamE
 - ▶ **particle placement**: effect size of length is weaker in PhiE

Stability vs. indigenization

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

Stability vs. indigenization

- ▶ Do the varieties of English we study share a core probabilistic grammar? → YES

Stability vs. indigenization

- ▶ Do the varieties of English we study share a core probabilistic grammar? → YES
- ▶ What are the constraints on variation that are particularly likely to be indigenized?

Stability vs. indigenization

- ▶ Do the varieties of English we study share a core probabilistic grammar? → YES
- ▶ What are the constraints on variation that are particularly likely to be indigenized? → most important predictors = highest cue validity = unstable

Stability vs. indigenization

- ▶ **limits on variation** – stability in effect direction vs. variability in effect strength

Stability vs. indigenization

- ▶ **limits on variation** – stability in effect direction vs. variability in effect strength
- ▶ synchronic and **diachronic** effects

Stability vs. indigenization

- ▶ **limits on variation** – stability in effect direction vs. variability in effect strength
- ▶ synchronic and **diachronic** effects
- ▶ probabilistic indigenization (of end-weight effects) due to shifting usage frequencies in linguistic material

Stability vs. indigenization

- ▶ **limits on variation** – stability in effect direction vs. variability in effect strength
- ▶ synchronic and **diachronic** effects
- ▶ probabilistic indigenization (of end-weight effects) due to shifting usage frequencies in linguistic material → **causes?**

Stability vs. indigenization

- ▶ **limits on variation** – stability in effect direction vs. variability in effect strength
- ▶ synchronic and **diachronic** effects
- ▶ probabilistic indigenization (of end-weight effects) due to shifting usage frequencies in linguistic material → **causes?**
 - ▶ **second language acquisition** – transfer of cue strength & preferences of the semantically more transparent option (PD, of-gen, joined)

Stability vs. indigenization

- ▶ **limits on variation** – stability in effect direction vs. variability in effect strength
- ▶ synchronic and **diachronic** effects
- ▶ probabilistic indigenization (of end-weight effects) due to shifting usage frequencies in linguistic material → **causes?**
 - ▶ **second language acquisition** – transfer of cue strength & preferences of the semantically more transparent option (PD, of-gen, joined)
 - ▶ **language contact** – substrate influence & structural nativization

Stability vs. indigenization

- ▶ **limits on variation** – stability in effect direction vs. variability in effect strength
- ▶ synchronic and **diachronic** effects
- ▶ probabilistic indigenization (of end-weight effects) due to shifting usage frequencies in linguistic material → **causes?**
 - ▶ **second language acquisition** – transfer of cue strength & preferences of the semantically more transparent option (PD, of-gen, joined)
 - ▶ **language contact** – substrate influence & structural nativization
 - ▶ **constructional/semantic changes** – lexical preferences

Concluding remarks

Conclusion

- ▶ The extent to which syntactic variation is constrained in postcolonial Englishes is influenced both by qualitative stability and probabilistic indigenization.

What's next?

- ▶ validating corpus results with rating task experiments

What's next?

- ▶ validating corpus results with rating task experiments
- ▶ extending the dataset to include web-based language (Corpus of Global web-based English)

What's next?

- ▶ validating corpus results with rating task experiments
- ▶ extending the dataset to include web-based language (Corpus of Global web-based English)
- ▶ extending the analysis to memory-based learning (TiMBL), NDL, etc.

Thank you!

`melanie.rothlisberger@kuleuven.be`

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

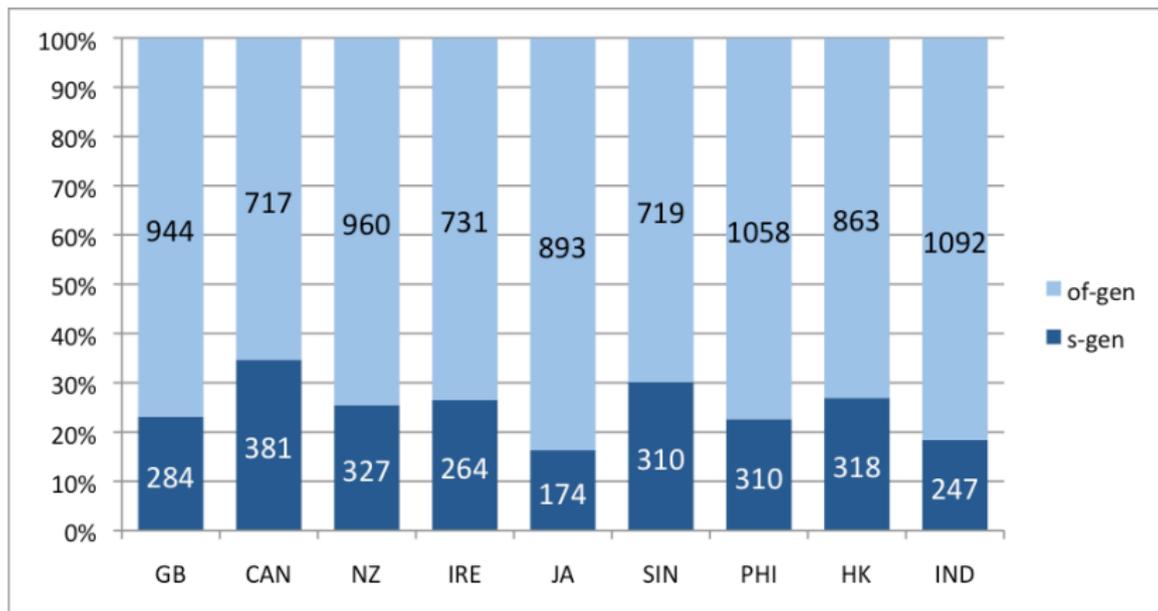
This presentation is based upon work supported by an
Odysseus grant of the Research Foundation Flanders (FWO)
(grant no. G.0C59.13N).



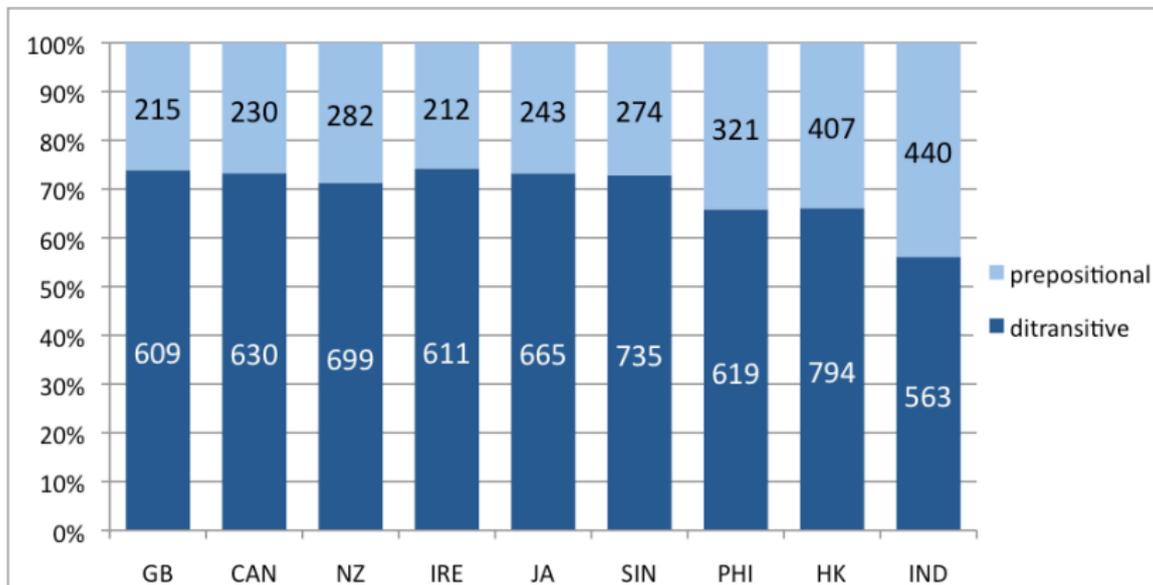
References I

- Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Behaghel, O. (1909/1910). Beziehungen zwischen umfang und reihenfolge von satzgliedern. *Indogermanische Forschungen* 25, 110–142.
- Bresnan, J. and M. Ford (2010). Predicting Syntax: Processing dative constructions in American and Australian Varieties of English. *Language* 86(1), 168–213.
- Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1), 95–125.
- Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.
- Labov, W. (1982). Building on empirical foundations. In W. Lehmann and Y. Malkiel (Eds.), *Perspectives on Historical Linguistics*, pp. 17–92. Amsterdam, Philadelphia: Benjamins.
- Szmrecsanyi, B., J. Grafmiller, B. Heller, and M. Röthlisberger (2016). Around the world in three alternations: modeling syntactic variation in varieties of English. *English World-Wide* 37(2).
- Wasow, T. (1997). Remarks on grammatical weight. *Language Variation and Change* 9, 81–105.
- Wolk, C., J. Bresnan, A. Rosenbach, and B. Szmrecsanyi (2013, jan). Dative and genitive variability in Late Modern English: Exploring cross-constructural variation and change. *Diachronica* 30(3), 382–419.

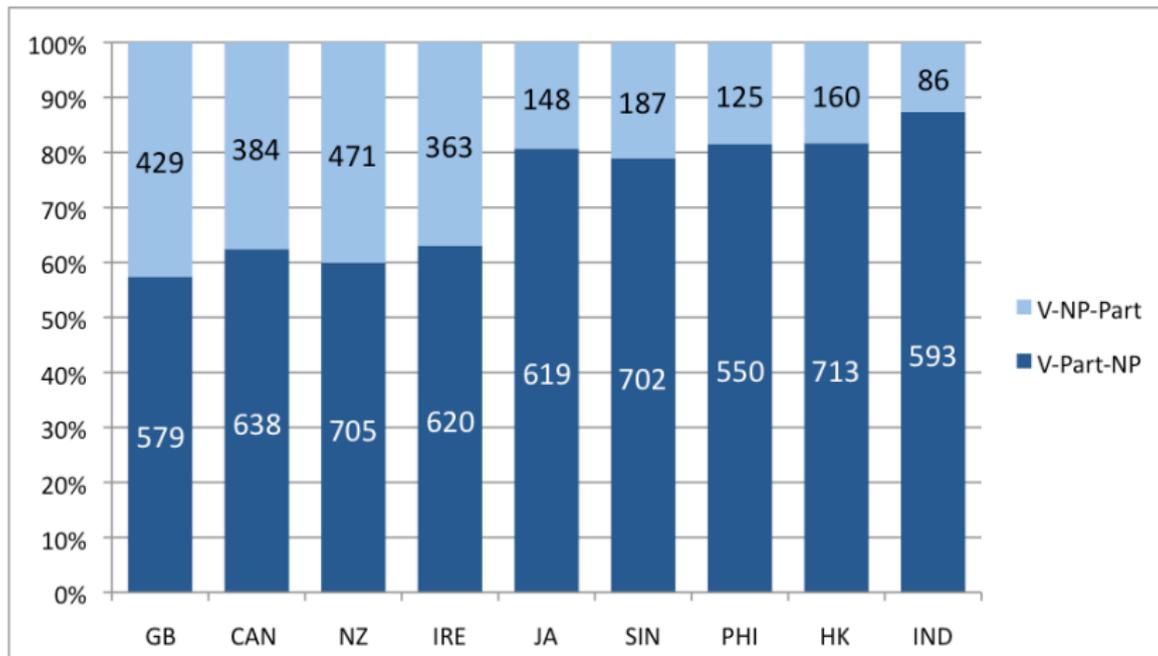
proportion of genitives per variety



proportion of datives per variety



proportion of particle verbs per variety



team members



Benedikt Szmrecsanyi
Principle investigator



Jason Grafmiller
Ph.D., 2013, Stanford University
particle placement



Benedikt Heller
MA, 2013, University of Giessen
the genitive alternation



Melanie Röthlisberger
MA, 2011, University of Zurich
the dative alternation

Summary stats

Table: Summary statistics of all three models

	C-value	% accuracy (% baseline)
GEN	0.98	94.1% (75.3%)
DAT	0.97	92.2% (69%)
PART	0.92	80.4% (70.8%)