Syntactic variation and probabilistic indigenization in World Englishes

Benedikt Heller, Benedikt Szmrecsanyi, Jason Grafmiller, Melanie Röthlisberger

KU Leuven
Quantitative Lexicology and Variational Linguistics

ICAME 37

29 May 2016, Hong Kong

Introduction

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

- Syntactic variation

- World Englishes

- Syntactic variation
  - dative alternation
  - genitive alternation
  - particle placement
- World Englishes

- Syntactic variation
  - dative alternation
  - genitive alternation
  - particle placement
- World Englishes
  - large-scale comparative perspective

- Syntactic variation
  - dative alternation
  - genitive alternation
  - particle placement
- World Englishes
  - large-scale comparative perspective
- ‘probabilistic indigenization’
  - process in which speakers of different varieties reinterpret / indigenize the probabilistic effects of constraints governing syntactic variation

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? $\rightarrow$ end-weight

End-weight effects

Behaghel's Gesetz der wachsenden Glieder: constituents tend to occur in order of increasing size or complexity (Behaghel 1909)

(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 1909)

(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 1909)

(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 1909)

(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 1909)

(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 is end-weight interesting?

- 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 & Ford 2010)

Why is end-weight interesting?

- 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 & Ford 2010)

  $\rightarrow$ to what extent are end-weight effects cross-lectally variable?

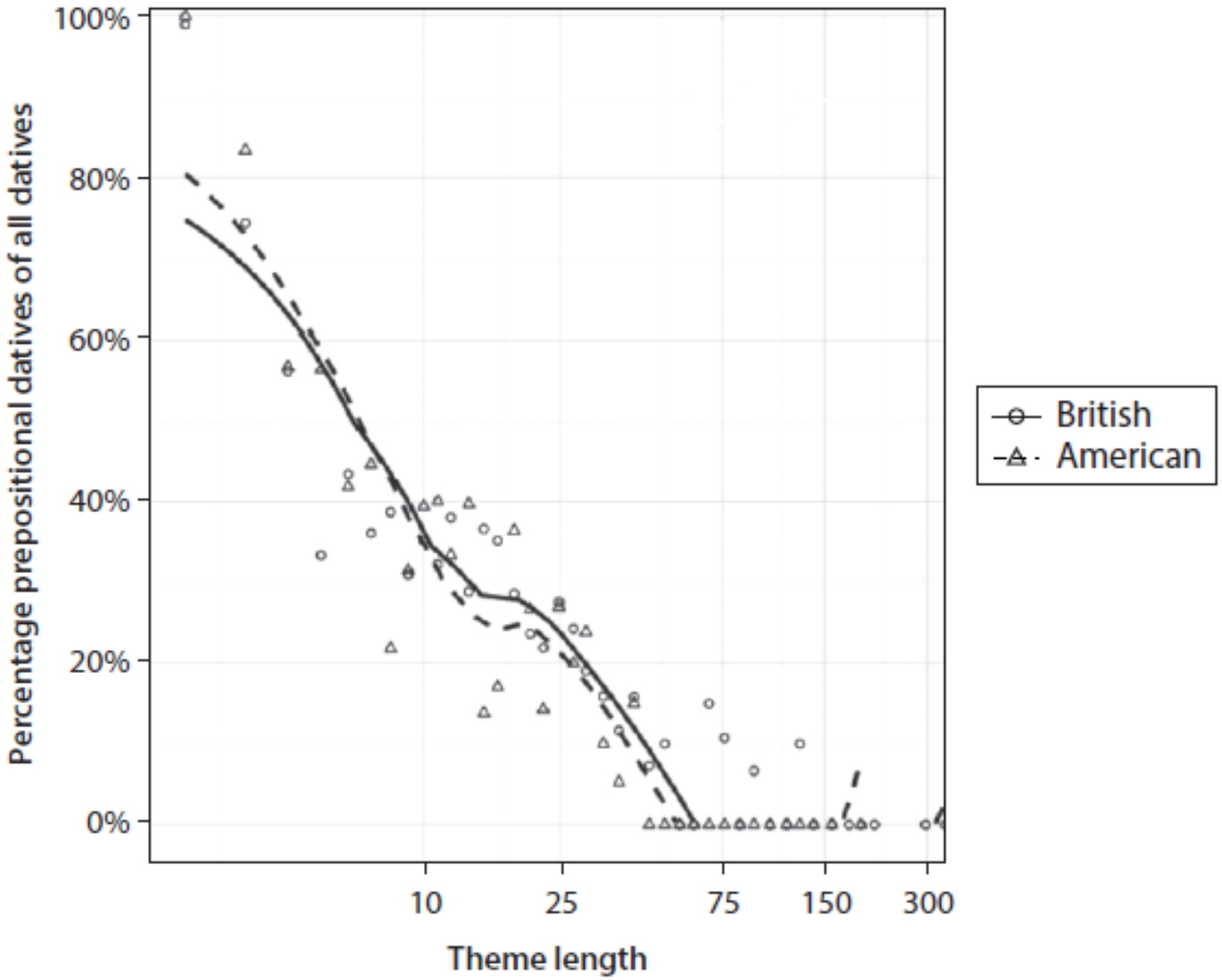


Figure: Percentage prepositional datives as function of theme length; from Wolk et al. 2013:407

Methods & Data

A methodological sketch

1. explore 3 syntactic alternations across 9 varieties of English

A methodological sketch

1. 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. 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. 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]$_{possessum}$
(*s*-genitive)

b. [the brother]$_{possessum}$ of [the Senator]$_{possessor}$
(*of*-genitive)

N = 10,594

Dative alternation

83 dative verbs: *give, send, offer, bring, carry, show, tell, . . .*

(3) a. We sent [the president]$_{recipient}$ [a letter]$_{theme}$
(ditransitive dative)

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

N = 8,549

Particle placement

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

(4) a. The student looked$_{verb}$ [the word]$_{direct\ object}$ up$_{particle}$
(split order: V-Obj-P)

b. The student looked$_{verb}$ up$_{particle}$ [the word]$_{direct\ object}$
(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			
HK: [*Britain*]'s [*youngest university graduate*]			<ICE-HK:s2b-001>
GB: [*gods and customs*] of [*Italy*]			<ICE-GB:w2a-001>
PhilE	−	+	
CanE		+	
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	
JA: *send* [*them*] [*their itinerary*]	<ICE-JA:s1a-043>
GB: *sent* [*their children*] *to* [*them*]	<ICE-GB:s2a-021>

IndE		+
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				−	
JamE			−		+
CanE		+			

(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				−	
JamE			−		+
CanE		+			

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

Particle placement

Table: Interactions with Variety

Va	examples				
	PHI: *put on* [*a cheerful front*]				<ICE-PHI:w2f-011>
I	GB: *put* [*a good front*] *on*				<ICE-GB:s1b-041>
NZ		+	–		
PhiE	+		–	–	
SinE				–	
JamE			–		+
CanE		+			

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

Discussion

Indigenization effects

Indigenization effects

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

Indigenization effects

- varieties do share a core probabilistic grammar: effect directions of constraints 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 stronger in PhiE and IndE

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? $\rightarrow$ YES

Stability vs. indigenization

- Do the varieties of English we study share a core probabilistic grammar? $\rightarrow$ 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? $\rightarrow$ YES
- What are the constraints on variation that are particularly likely to be indigenized? $\rightarrow$ 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 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 due to shifting usage frequencies in linguistic material $\rightarrow$ causes?

Stability vs. indigenization

- limits on variation – stability in effect direction vs. variability in effect strength
- synchronic and diachronic effects
- probabilistic indigenization due to shifting usage frequencies in linguistic material $\rightarrow$ causes?
  - second language acquisition – transfer of cue strength & preferences for more explicit / transparent / frequent option (*of*-gen, PD, V-P-Obj)

Stability vs. indigenization

- limits on variation – stability in effect direction vs. variability in effect strength
- synchronic and diachronic effects
- probabilistic indigenization due to shifting usage frequencies in linguistic material $\rightarrow$ causes?
  - second language acquisition – transfer of cue strength & preferences for more explicit / transparent / frequent option (*of*-gen, PD, V-P-Obj)
  - language contact – substrate influence & dialect contact

Stability vs. indigenization

- limits on variation – stability in effect direction vs. variability in effect strength
- synchronic and diachronic effects
- probabilistic indigenization due to shifting usage frequencies in linguistic material → causes?
  - second language acquisition – transfer of cue strength & preferences for more explicit / transparent / frequent option (*of*-gen, PD, V-P-Obj)
  - language contact – substrate influence & dialect contact
  - constructional / semantic changes – shifts in lexical preferences

Concluding remarks

Conclusion

- end weight effects consistently weaker in British English

Conclusion

- end weight effects consistently weaker in British English
- **Big picture:** syntactic variation in postcolonial Englishes is characterized both by qualitative stability and probabilistic indigenization of gradient constraints.

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)

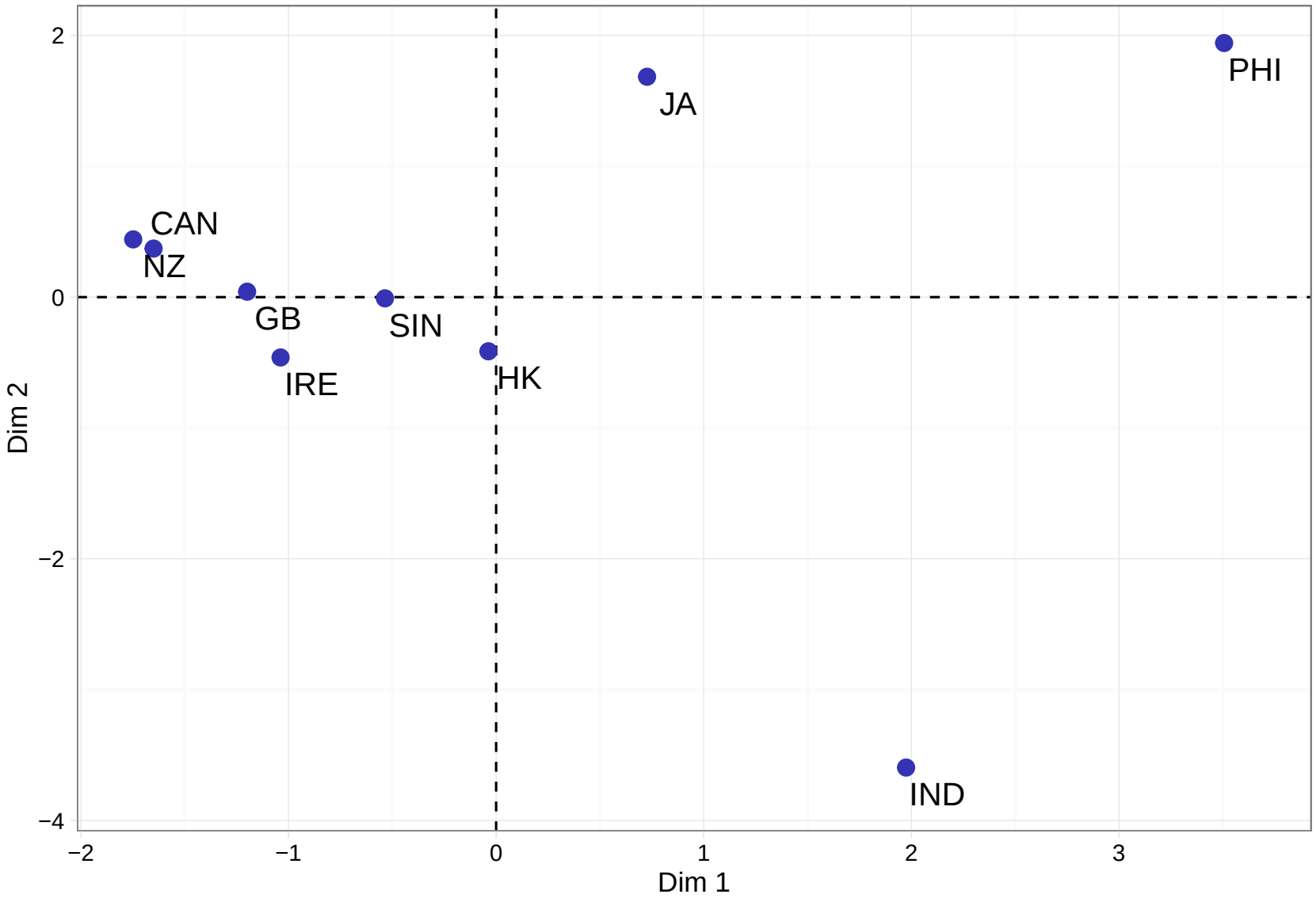
What's next?

- 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.

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.
- assessing overall cross-varietal similarity in probabilistic grammar(s)

Probabilistic similarity in particle placement



Thank you!

benedikt.heller@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, R Harald. 2008. Analyzing linguistic data: a practical introduction to statistics using R. Cambridge, New York: Cambridge University Press.

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

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

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. Language 86(1). 168–213. doi: 10.1353/lan.0.0189.

Gries, Stefan Th. 2015. The most under-used method in corpus linguistics: multi-level (and mixed-effects) models. Corpora 10(1). 95–125. doi: 10.3366/cor.2015.0068.

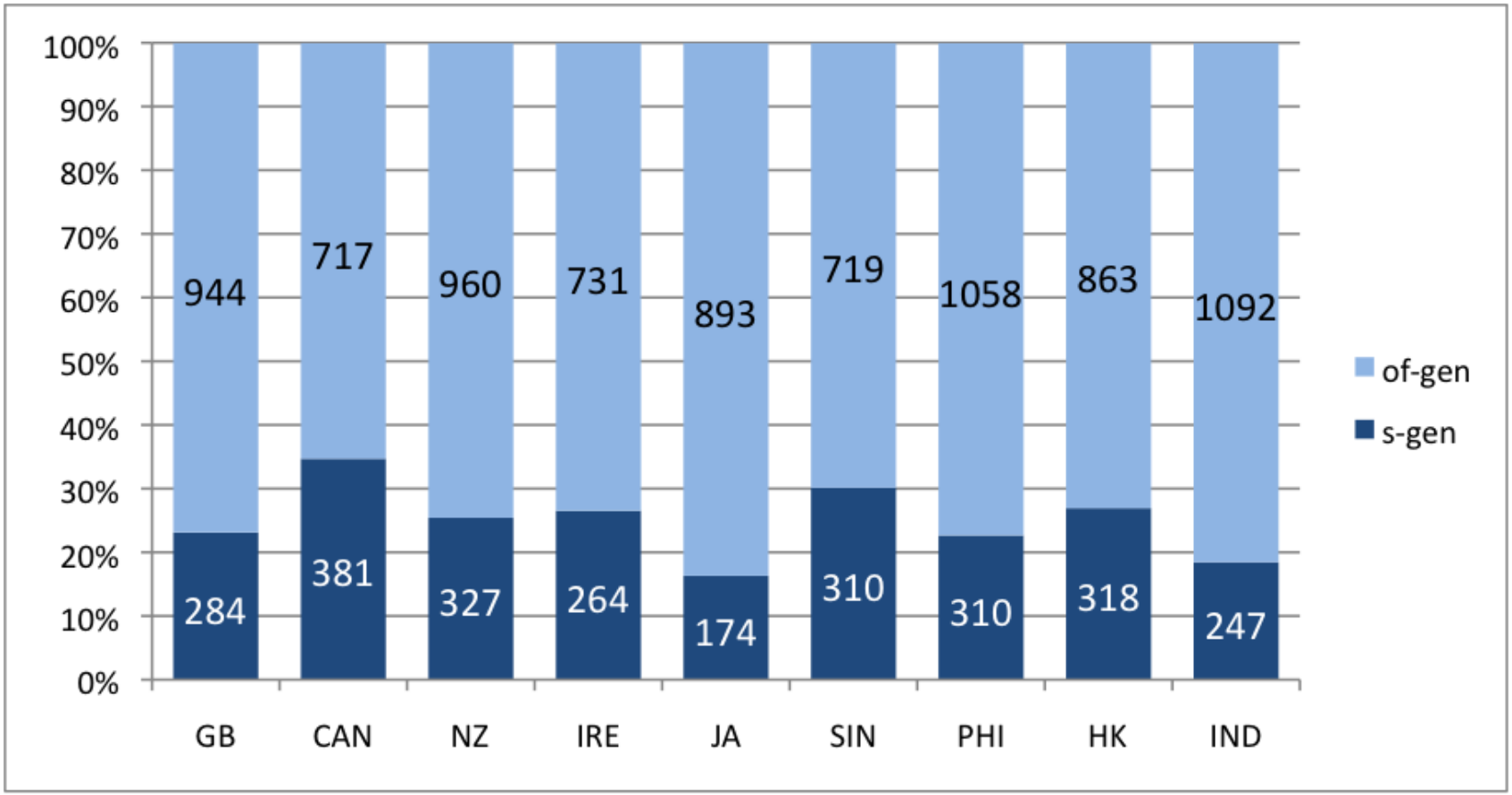
Hawkins, John A. 1994. A performance theory of order and constituency. Cambridge; New York: Cambridge University Press.

Labov, William. 1982. Building on empirical foundations. In Winfred Lehmann & Yakov Malkiel (eds.), Perspectives on Historical Linguistics, 17–92. Amsterdam, Philadelphia: Benjamins.

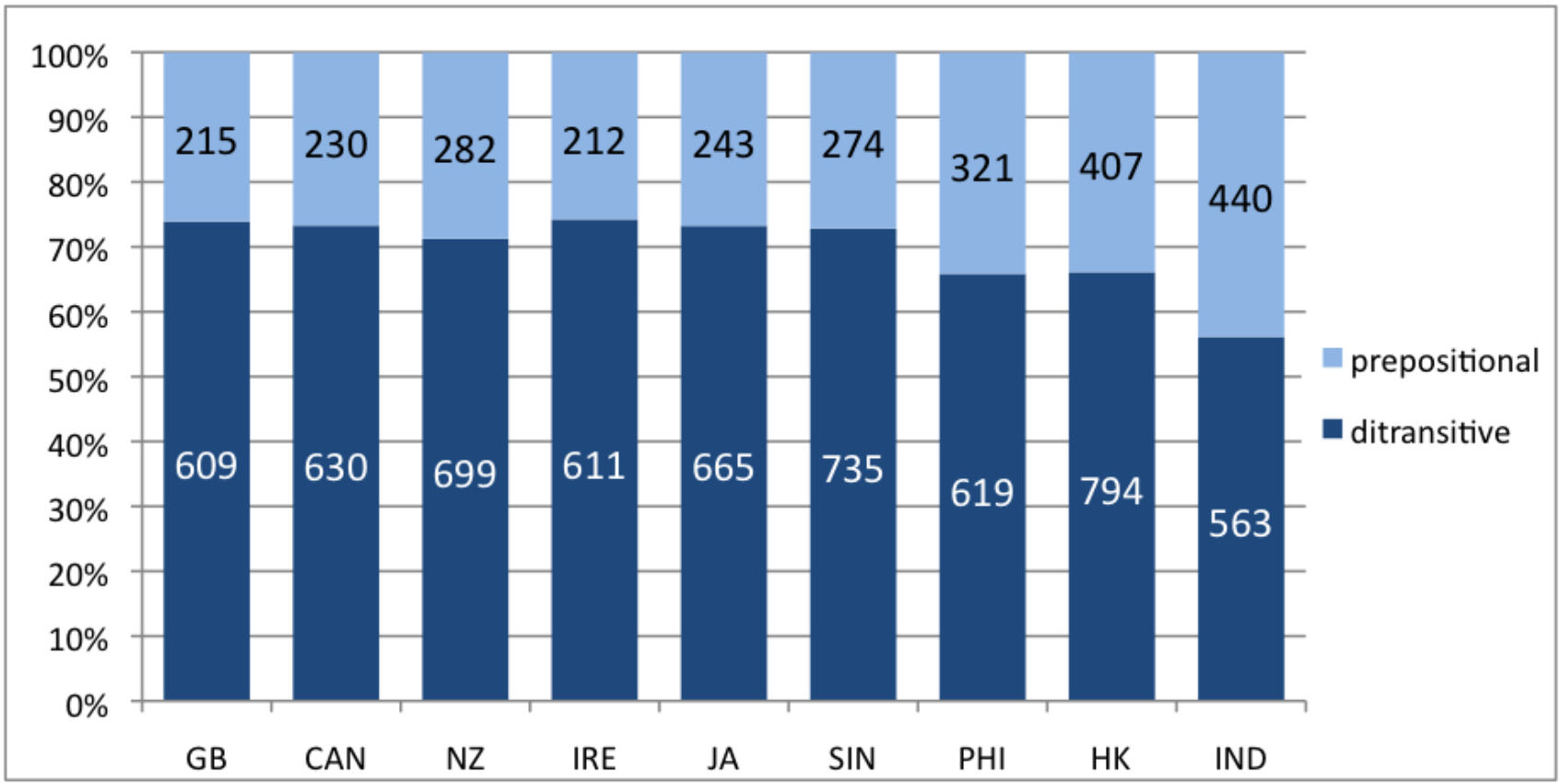
Wasow, Thomas. 1997. End-weight from the speaker's perspective. Journal of Psycholinguistic Research 26. 347–361.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. Diachronica 30(3). 382–419. doi: 10.1075/dia.30.3.04wol.

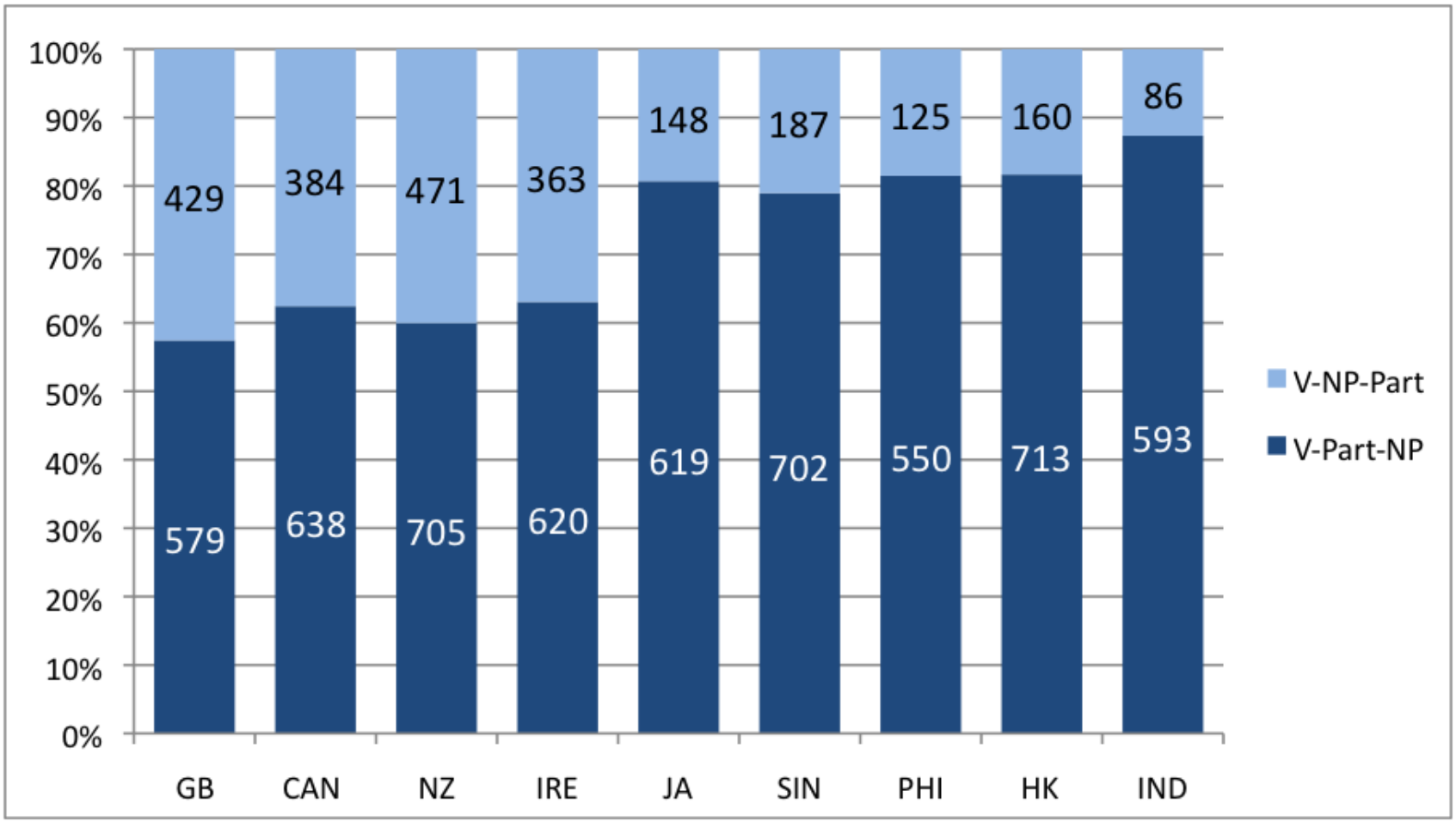
proportion of genitives per variety



proportion of datives per variety



proportion of particle verbs per variety



Supplementary experiments

- Bresnan (2007): regression models match probabilistic intuitions
- converging evidence for psychological reality of the experience-based probabilistic grammars?
- replicate Experiment 1 in Bresnan (2007:76-84):
  - recruit native-speaker subjects from different VoE backgrounds
  - subjects rate randomly sampled observations from the corpus database
  - do subjects' ratings match probabilities predicted by the corpus models?

The 100-split task: an example

I'm in college, and I'm only twenty-one but I had a speech class last semester, and there was a girl in my class who did a speech on home care of the elderly. And I was so surprised to hear how many people, you know, the older people, are like, fastened to their beds so they can't get out just because, you know, they wander the halls. And they get the wrong medicine, just because, you know, the aides or whatever

The 100-split task: an example

I'm in college, and I'm only twenty-one but I had a speech class last semester, and there was a girl in my class who did a speech on home care of the elderly. And I was so surprised to hear how many people, you know, the older people, are like, fastened to their beds so they can't get out just because, you know, they wander the halls. And they get the wrong medicine, just because, you know, the aides or whatever

(1) just give them the wrong medicine

(2) just give the wrong medicine to them

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%)