



UNIVERSITY OF
BIRMINGHAM



Universität
Zürich^{UZH}

Measuring the variational homogeneity of English as a world language: Probabilistic indigenization effects in four syntactic alternations

Iván Tamaredo, University of Santiago de Compostela

Benedikt Szmrecsanyi, KU Leuven

Jason Grafmiller, University of Birmingham

Benedikt Heller, Justus Liebig University of Giessen

Melanie Röthlisberger, University of Zurich

Theoretical background

Two research paradigms:

Probabilistic grammar
(e.g., Bresnan 2007)



World Englishes
(e.g., Schneider 2007)

1. grammatical knowledge is partially probabilistic
2. multiple probabilistic constraints influence the outcome of grammatical variation
3. grammatical knowledge is experience- and usage-based

structural characteristics and sociohistorical background of varieties of English

How similar or dissimilar is the probabilistic knowledge of English grammar on the part of speakers with different regional and cultural backgrounds?

Theoretical background

- Varieties share a common probabilistic grammar in that some language-internal constraints are largely shared (e.g., Bernaisch et al. 2014; Szmrecsanyi et al. 2016).
- BUT some constraints impact a speaker's choice differently in one variety compared to another.
 - Example: LENGTH OF CONSTITUENTS has a weaker impact in variety A than in variety B on the choice between *Mary gives John the apple* vs. *Mary gives the apple to John*
- Indigenization: “the emergence of locally characteristic linguistic patterns” (Schneider 2007: 6).
 - Lexical items in novel syntactic constructions: e.g., *visit with* in Philippine English instead of *visit*.

Theoretical background

- Probabilistic indigenization:

“the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties” (Szmrecsanyi et al. 2016: 133)
- Hypothesis: probabilistic indigenization effects arise as a function of the lexical specificity of alternations, with those that are strongly connected with specific lexical items being the most likely ones to exhibit cross-varietal indigenization effects.

Aims

- Measure the degree of alternation-internal homogeneity or heterogeneity across three varieties: British English, Indian English, and Singapore English.
- How speakers select a specific variant when they have a choice between “alternate ways of saying ‘the same’ thing” (Labov 1972: 188).
- Four syntactic alternations: **dative** ($N=3,012$; see, e.g., Bresnan & Hay 2008), **genitive** ($N=3,108$; see, e.g., Rosenbach 2014), **particle placement** ($N=2,480$, see, e.g., Gries 2003), and **subject pronoun omission** ($N=2,456$; see, e.g., Torres Cacoullos & Travis 2014)

Aims

(1) **a. The ditransitive dative variant**

That will give [the panel]_{recipient} [a chance]_{theme} to expand on what they've been saying. (ICE-GB:S1B-036)

b. The prepositional dative variant

[...] and that gives [a chance]_{theme} [to Bhupathy]_{recipient} to equalise the points at thirty all. (ICE-IND:S2A:019)

Aims

(2) a. The *s*-genitive

[Singapore]_{possessor}'s [small size]_{possessum} meant it could be quick to respond to changes in economic conditions (ICE-SIN:W2C-011)

b. The *of*-genitive

the [size]_{possessum} of [the eyes]_{possessor} is to help them at night. (ICE-GB:W2B-021)

Aims

(3) **a. Verb-object-particle order (or split order)**

you can just [cut]_{verb} [the tops]_{direct object} [off]_{particle} and leave them.
(ICE-GB:S1A-007)

b. Verb-particle-object order (or joined order)

[Cut]_{verb} [off]_{particle} [the flowers]_{direct object} as they fade. (ICE-CAN:W2B-023)

Aims

(4) **a. Overt subject pronoun**

The vision_i was not very clear. It_i was murky or rather uh foggy or misty. (ICE-IND:S1B-006)

b. Null subject pronoun

Oh, be4 I forget, “Chitra_i” sends you her love. \emptyset _i Has been asking about you since you left. (ICE-SIN:W1B-003)

Aims

The **greater** the degree of probabilistic indigenization (i.e. the smaller varieties' similarity)



The **greater** the impact of lexical specific constituents.

Data & methodology

- Relevant observations of the (a) and (b) variants of the four alternations retrieved from the British, Indian, and Singaporean components of the International Corpus of English (ICE).
- 5 most important predictors selected on the basis of conditional random forests fitted to the dataset of all three varieties.

Data & methodology

PREDICTORS DATIVE	LEVELS
Weight ratio	Recipient length in letters divided by theme length in letters (log value)
Recipient pronominality	Pronoun vs nominal
Recipient person	Local vs non-local
Theme complexity	Simple vs complex
Recipient head frequency	Global text frequency of recipient head (lemma)

PREDICTORS PART. PL.	LEVELS
Direct object length	Length of direct object in words
Semantics	Compositional vs non-compositional
Directional PP	Yes vs no
Verb surprisal	Predictability of the verb given the particle
Preposition surprisal	Predictability of the particle given the verb

PREDICTORS GENITIVE	LEVELS
Possessor animacy	Human & animal vs collective vs inanimate vs locative vs temporal
Possessor length	Length of possessor in letters
Possessum length	Length of possessum in letters
Possessor thematicity	Number of uses of the possessor head noun in a text divided by the total number of words in the text
Possessor final sibilancy	Yes vs no

PREDICTORS SUBJ. OM.	LEVELS
Text type	Spoken informal vs spoken formal vs written informal vs written formal
Coordination	Coordination vs no coordination
Clause type	Main vs embedded
Clause position	Initial vs non-initial
Pronoun-verb cooccurrence frequency	How many times the pronoun and the verb cooccur ¹²

Data & methodology

- Per variety binary mixed-effects logistic regression and conditional random forest analyses:
 - mixed-effects models included random intercepts for lexical items.
- Comparative sociolinguistics (Poplack & Tagliamonte 2001, *inter alii*):
 - compares and contrasts patterns of variability of linguistic features across different dialects or varieties using quantitative methods.
- Three lines of evidence:
 1. shared significant/non-significant predictors
 2. relative strength of predictors
 3. importance or rank of predictors

Data & methodology

- Three steps:
 1. Fit a mixed-effects model /conditional random forest per variety using the same model formula per alternation.
 2. Calculate a distance matrix:
 - a. **statistical significance:** number of shared significant and non-significant predictors (mixed-effects models)
 - b. **relative strength:** distance between coefficient estimates from models (mixed-effects models)
 - c. **constraint ranking:** Spearman's rank correlation coefficient between the constraint ranks as a distance measure (conditional random forests)
 3. Calculate the average similarity as a measure of overall stability.
- Three core grammar coefficients (0-1): the higher the value, the more homogeneous the alternation.

Data & methodology

- To gauge lexical effects: visualize the random slopes of the mixed-effects models.
- Lemmas of individual lexical items as random effects:
 - **verbs, recipients, and themes** in the dative alternation
 - **possessors** and **possessums** in the genitive alternation
 - **verbs, particles, and verb-particle** combinations in the particle placement alternation
 - **verbs** in the case of subject omission
- Variance accounted for by lexical effects in the random structure of the mixed-effects model:
`r.squaredGLMM()` in MuMIn package (Barton 2015).
- R^2 provides indication of model fit
 - Marginal R^2 = variance accounted for in model with fixef
 - Conditional R^2 = variance accounted for in model with ranef + fixef
 - variance accounted for by lexical effects (ranef) only: $cR^2 - mR^2$

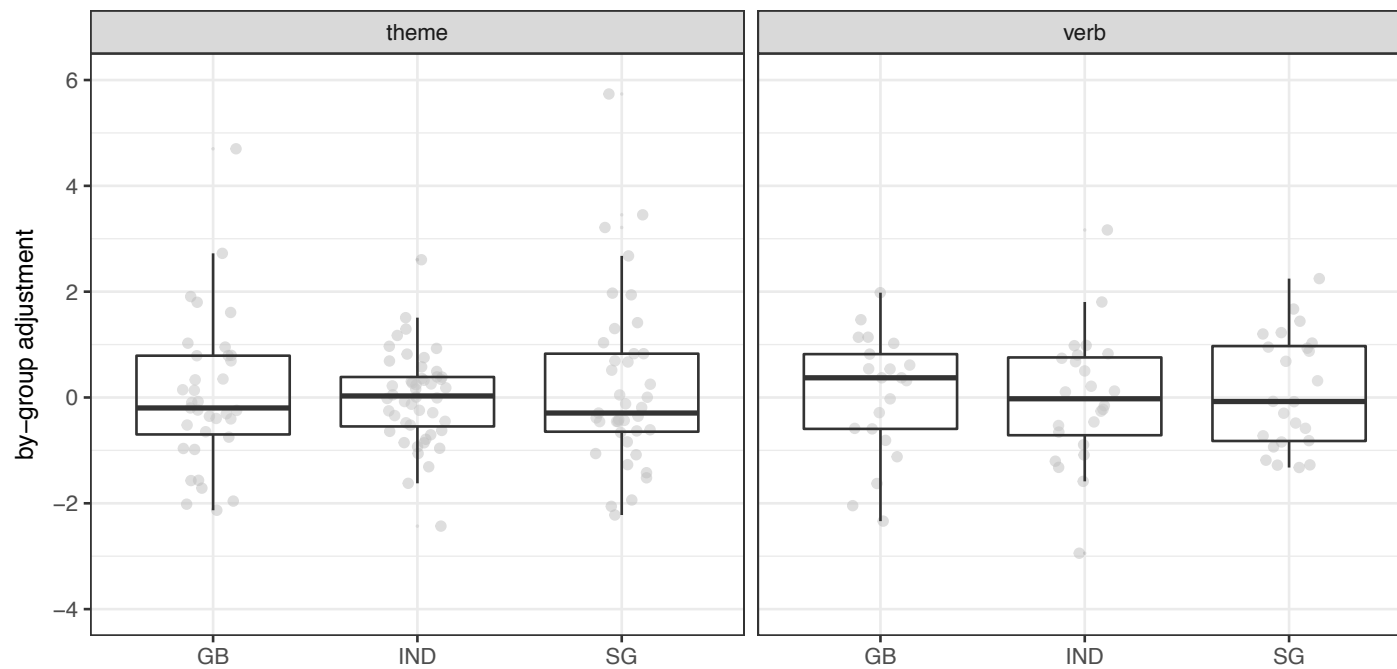
Results: core grammar coefficients

	Datives	Genitives	Particles	Subject o.	Mean
Significance	0.867	1.000	0.600	0.619	0.772
Effect strength	0.593	0.531	0.073	0.787	0.496
Constraint ranking	0.733	0.833	0.733	0.800	0.775
Mean	0.731	0.788	0.469	0.721	0.677

- Global mean: overall, varieties seem to share a core probabilistic grammar
- Mean per line: constraint ranking > significance > effect strength
- Mean per alternation: genitives > datives > subject omission > particle placement

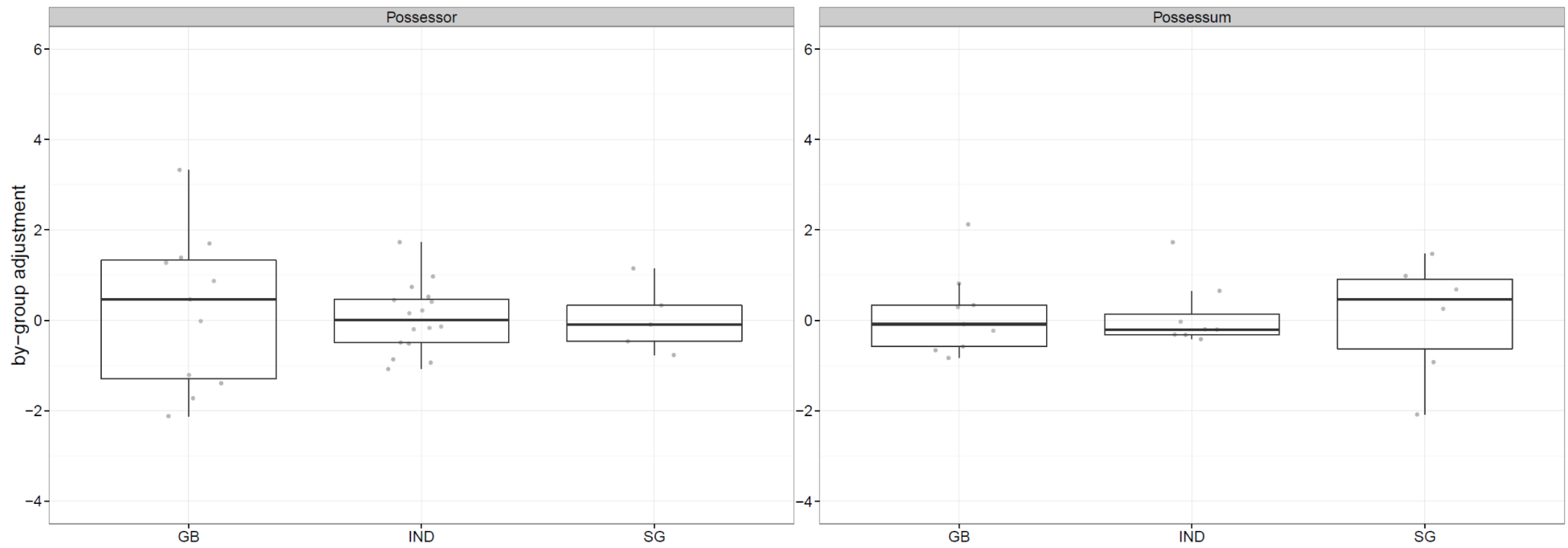
Results: lexical specificity

- Dative (recipient not shown as $SD=0$)



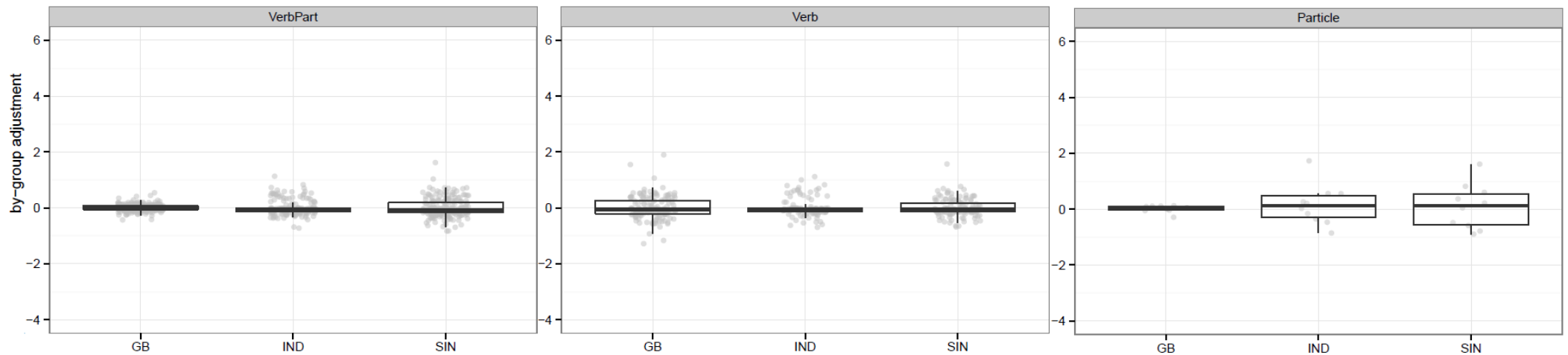
Results: lexical specificity

- Genitive



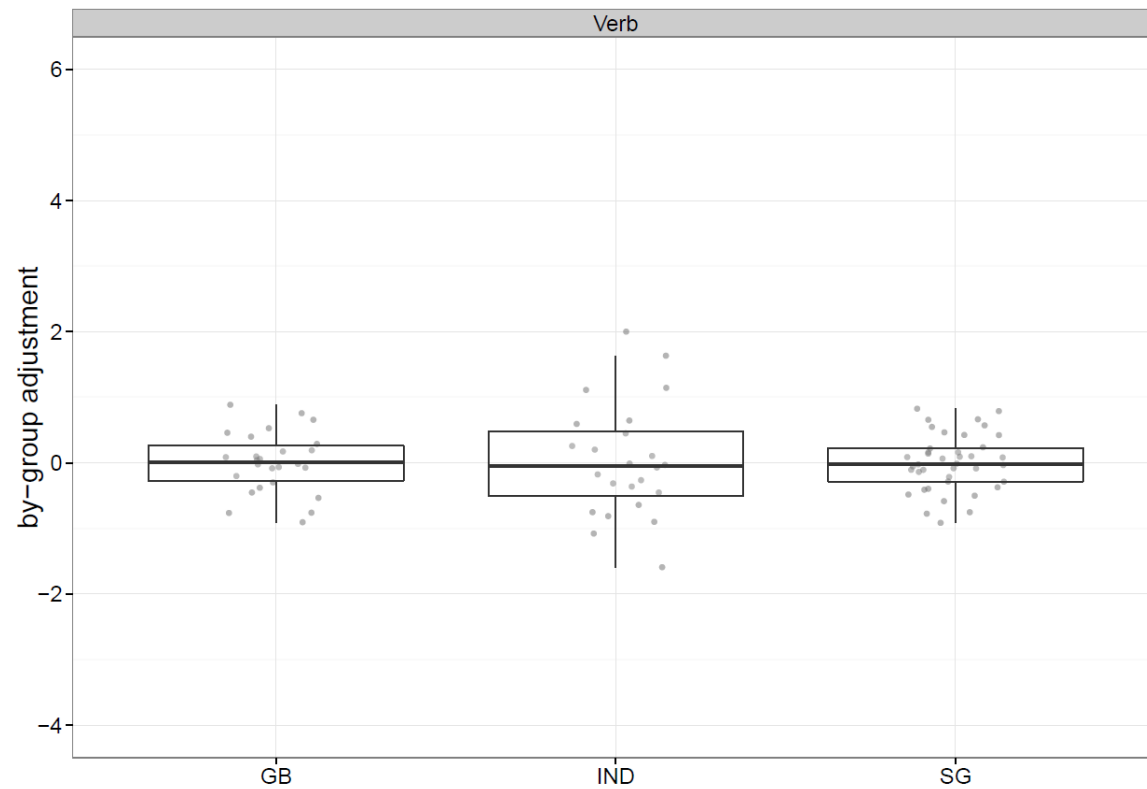
Results: lexical specificity

- Particle placement



Results: lexical specificity

- Subject omission



Results: lexical specificity

- Lexical specificity (from most to least):

➤ **genitives** > **datives/particles** > **subjects**

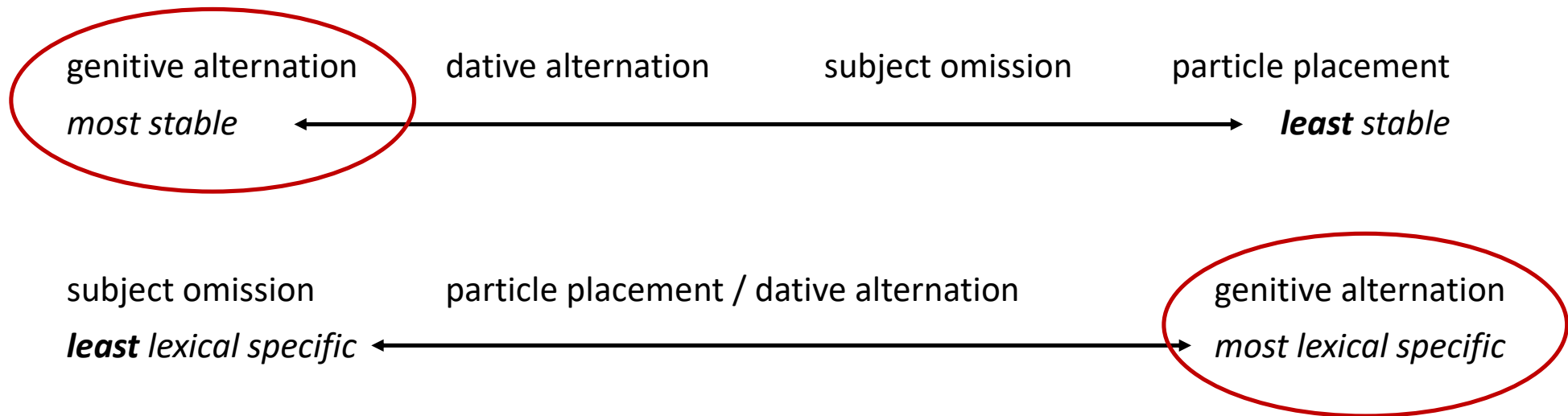
➤ no cross-varietal patterns

- Homogeneity:

genitives > **datives** > **subjects** > **particles**

Alternation	Marginal R ²	Conditional R ²	% of variance accounted for by random structure
Dative alternation:			
BrE	0.433	0.805	0.372
IndE	0.087	0.173	0.086
SinE	0.209	0.549	0.34
MEAN			0.266
Genitive alternation:			
BrE	0.293	0.762	0.469
IndE	0.409	0.666	0.257
SinE	0.431	0.717	0.286
MEAN			0.337
Particle placement:			
BrE	0.324	0.492	0.168
IndE	0.215	0.551	0.336
SinE	0.324	0.617	0.293
MEAN			0.266
Subject omission:			
BrE	0.749	0.782	0.033
IndE	0.596	0.709	0.113
SinE	0.568	0.618	0.05
MEAN			0.065

Summary

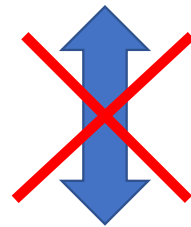


Conclusions

- Probabilistic indigenization can be observed to different degrees in the three varieties and across four alternations
- Various explanations have been offered (see e.g. Röthlisberger et al. 2017)
- The degree of alternation-internal homogeneity is not directly linked to an alternation's lexical specificity but seems to be reversed
 - genitive alternation exhibits the most stability in probabilistic constraints but is also the most lexical specific by variety
 - it seems not to be the case that the degree of probabilistic indigenization can be linked to an alternation's lexical specificity

Conclusions

The **greater** the degree of probabilistic indigenization (i.e. the smaller varieties' similarity)



The **greater** the impact of lexical specific constituents.

Next steps

- R^2 might not be a good heuristic to assess the lexical specificity of an alternation → collostructional analysis
- use other measures to compare/contrast varieties (e.g. AIC see Grafmiller & Szmrecsanyi under revision)

KIITOS

Slides can be downloaded from: www.melanie-roethlisberger.ch/research/publications

References

Primary sources:

- Barton, Kamil. 2015. *MuMIn: Multi-Model Inference*. R package version 1.13.4. <http://CRAN.R-project.org/package=MuMIn>
- Davies, Mark. 2013. *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE)*. <https://corpus.byu.edu/glowbe>
- International Corpus of English – the British Component*. 1998. Project coordinated by Bas Aarts at University College London, United Kingdom.
- International Corpus of English - the Indian Component*. 2002. Project coordinated by S. V. Shastri at Shivaji University, India, and Gerhard Leitner at Freie Universität Berlin, Germany.
- International Corpus of English - the Singaporean Component*. 2002. Project coordinated by Paroo Nihilani, Ni Yibin, Anne Pakir and Vincent Ooi at the National University of Singapore, Singapore.
- R Core Development Team. 2015. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <http://www.R-project.org>

Secondary sources:

- Bernaisch, Tobias, Stefan Th. Gries, and Joybrato Mukherjee. 2014. "The dative alternation in South Asian Englishes: Modelling predictors and predicting prototypes." *English World-Wide*, 35(1): 7–31. doi:10.1075/eww.35.1.02ber.
- Bresnan, Joan. 2007. "Is syntactic knowledge probabilistic? Experiments with the English dative alternation." In Sam Featherston, and Wolfgang Sternefeld, eds. *Roots: Linguistics in Search of Its Evidential Base*. Berlin: Mouton de Gruyter, 75–96.
- Bresnan, Joan, and Jennifer Hay. 2008. "Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English." *Lingua*, 118(2): 245–259. doi:10.1016/j.lingua.2007.02.007.
- Grafmiller, Jason, and Benedikt Szmrecsanyi. under revision. "Mapping out particle placement in Englishes around the world: A case study in comparative sociolinguistic analysis." *Language Variation and Change*.
- Gries, Stefan Th. 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement*. New York: Continuum Press.
- Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Philadelphia Press.
- Poplack, Shana, and Sali Tagliamonte. 2001. *African American English in the Diaspora*. Oxford: Basil Blackwell.
- Rosenbach, Anette. 2014. "English genitive variation – the state of the art." *English Language and Linguistics*, 18(02), 215–262. doi:10.1017/S1360674314000021.
- Röthlisberger, Melanie, Jason Grafmiller, and Benedikt Szmrecsanyi. 2017. "Cognitive indigenization effects in the English dative alternation." *Cognitive Linguistics*, 28(4): 673-710. doi: 10.1515/cog-2016-0051
- Schneider, Edgar. 2007. *Postcolonial English: Varieties Around the World*. Cambridge: Cambridge University Press.
- Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller, and Melanie Röthlisberger. 2016. "Around the world in three alternations: Modeling syntactic variation in varieties of English." *English World-Wide* 37 (2): 109–137. doi:10.1075/eww.37.2.01szm.
- Torres Cacoullos, Rena, and Catherine E. Travis. 2014. "Prosody, priming and particular constructions: The patterning of English first-person singular subject expression in conversation." *Journal of Pragmatics*, 63, 19–34. doi:10.1016/j.pragma.2013.08.003.