Variation-Based Distance and Similarity Modeling: Varieties of English and beyond

Benedikt Szmrecsanyi

KU LEUVEN

Introduction

Dialectometry: inter-speaker variation

Using atlas/survey classifications or frequency information from corpora to determine the aggregate similarity of varieties

(Séguy 1971; Goebl 1982; Nerbonne et al. 1999)

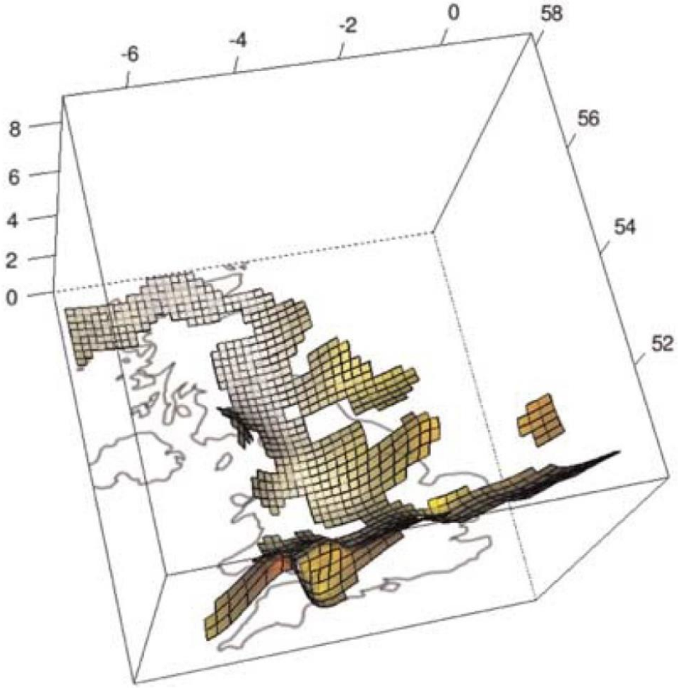


Figure 1. Frequency landscape for feature [33], multiple negation.

Wolk & Szmrecsanyi (2018: Fig 1)

Variationist (socio)linguistics: intra-speaker variability

How – that is, subject to which constraints – do language users choose between "alternate ways of saying `the same' thing" (Labov 1972: 188)?

Dialectometry meets variationist linguistics

Basic idea: quantify distance and similarity between lects (in our case study, nine international varieties of English) as a function of the (non-)correspondence of the ways in which language users choose between different ways of saying the same thing.

VADIS: VAriation-based DIstance and Similarity Modeling

- Inspired by work in comparative sociolinguistics and quantitative dialectometry
- Corpus-based
- Rigorously quantifies similarity/dissimilarity between lects as a function of the correspondence of the ways in which language users choose variants
- Use the output of variationist modeling as an input to dialectometric analysis ⇨ measure inter-speaker variation by assessing the structure of intra-speaker variation

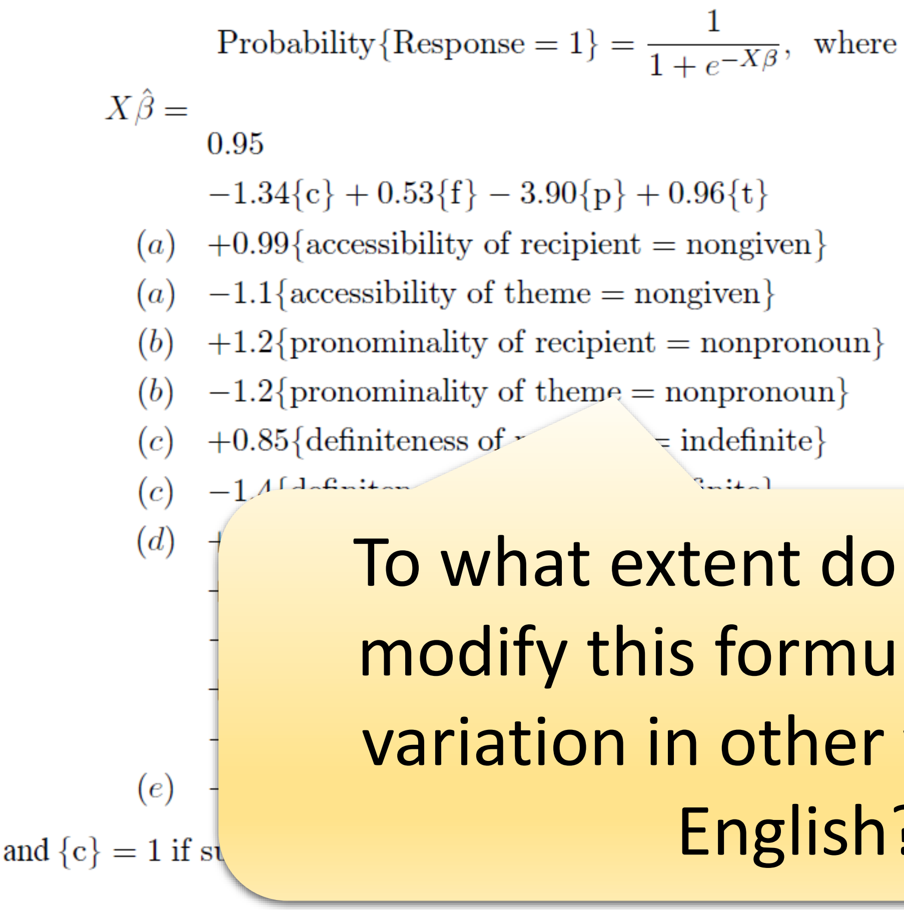
The dative alternation in English

(1) I've never even bought a gun myself. My dad’s **given it to me** or someone's **given me one**. So I'm probably real illegal, you know, carrying guns that aren't even mine.

(Switchboard US F/SM/67)

Inferring probabilistic grammars from corpus data on spoken US AmE

(Bresnan et al. 2007: Fig 4)



$$\text{Probability}\{\text{Response} = 1\} = \frac{1}{1 + e^{-X\beta}}, \quad \text{where}$$

$$X\hat{\beta} =$$

$$0.95$$

$$-1.34\{\text{c}\} + 0.53\{\text{f}\} - 3.90\{\text{p}\} + 0.96\{\text{t}\}$$

$$(a) \quad +0.99\{\text{accessibility of recipient} = \text{nongiven}\}$$

$$(a) \quad -1.1\{\text{accessibility of theme} = \text{nongiven}\}$$

$$(b) \quad +1.2\{\text{pronominality of recipient} = \text{nonpronoun}\}$$

$$(b) \quad -1.2\{\text{pronominality of theme} = \text{nonpronoun}\}$$

$$(c) \quad +0.85\{\text{definiteness of}$$

$$= \text{indefinite}\}$$

$$(c) \quad -1.4\{\text{definiten}$$

$$(d)$$

$$(e)$$

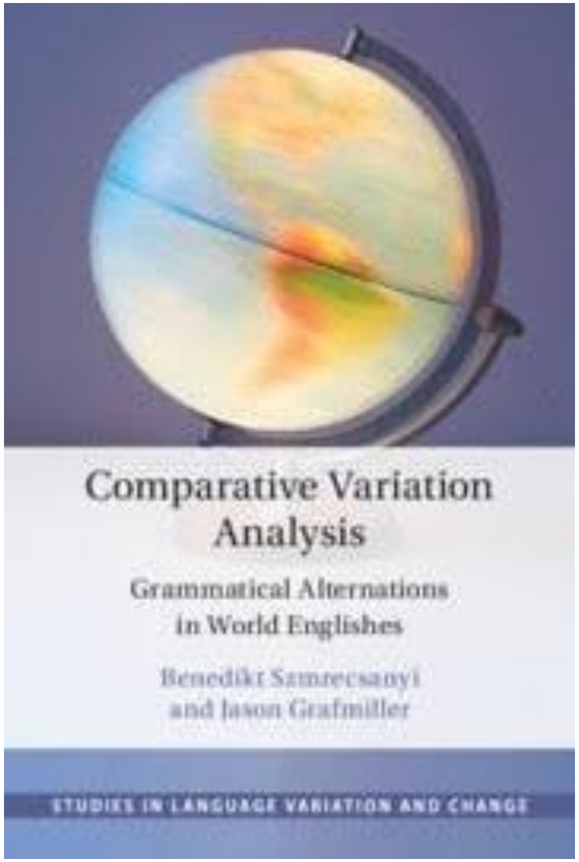
and $\{\text{c}\} = 1$ if su

To what extent do we have to modify this formula to model variation in other varieties of English?

Case study: data & alternations

The Leuven project

- **Data**: 9 international varieties of English x 3 syntactic alternations
- **Methods**: observational (corpus-based) & supplementary rating task experiments
- **Hot off the press**: Szmrecsanyi, Benedikt & Jason Grafmiller. 2023. *Comparative variation analysis: grammatical alternations in world Englishes*. Cambridge University Press.


Comparative Variation Analysis
Grammatical Alternations in World Englishes
Benedikt Szmrecsanyi and Jason Grafmiller
STUDIES IN LANGUAGE VARIATION AND CHANGE

Corpus track: nine varieties of English

British E

Canadian E

Irish E

New Zealand E

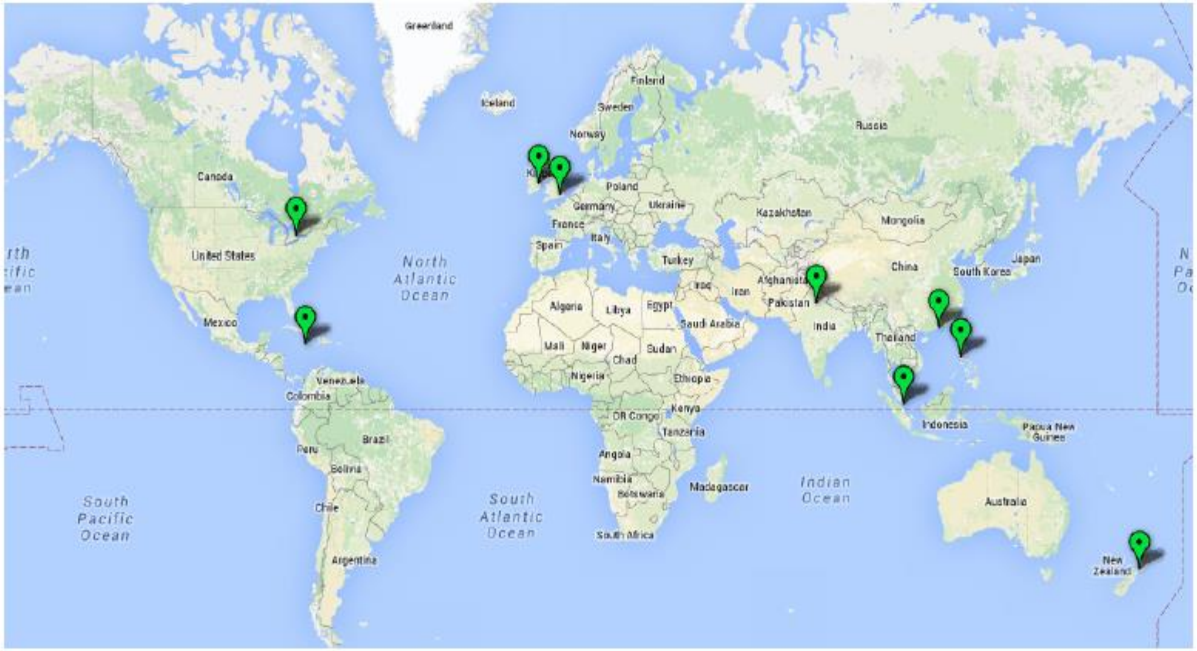
Hong Kong E

Indian E

Jamaican E

Philippine E

Singapore E



Corpora investigated

- The **International Corpus of English** (ICE; Greenbaum 1991): small & (comparatively) tidy, balanced design, classical off-line corpus, covers all sorts of classical text types (dialogues, monologues, written non-printed, written printed)
- The **Corpus of Global Web-based English** (GloWbE; Davies & Fuchs 2015): huge & a bit messy, covers automatically harvested blogs and websites

The dative alternation

(2) I’d given [Heidi]$_{\text{recipient}}$ [my T-Shirt]$_{\text{theme}}$
(the ditransitive dative variant)

(3) I’d given [the key]$_{\text{theme}}$ to [Helen]$_{\text{recipient}}$
(the prepositional dative variant)

Known language-internal probabilistic constraints: weight ratio between recipient and theme, recipient pronominality, theme complexity, theme head frequency, theme pronominality, theme definiteness, recipient givenness, recipient head frequency

The genitive alternation

(4) [the country]$_{\text{possessor}}$‘s [economic crisis]$_{\text{possesum}}$
(the *s*-genitive)

(5) [the economic growth]$_{\text{p’um}}$ of [the country]$_{\text{p’or}}$
(the *of*-genitive)

Known language-internal probabilistic constraints: possessor animacy, constituent length, possessor NP expression type, final sibilancy in possessor, priming, semantic relation, possessor head frequency

The particle placement alternation

(6) [cut]$_{\text{verb}}$ [the tops]$_{\text{object}}$ [off]$_{\text{particle}}$
(the split variant)

(7) [cut]$_{\text{verb}}$ [off]$_{\text{particle}}$ [the flowers]$_{\text{object}}$
(the continuous variant)

Known language-internal probabilistic constraints: length of the direct object, definiteness of the direct object, givenness of the direct object, concreteness of the direct object, thematicity of the direct object, presence of a directional modifier, semantics, surprisal of the particle

Method in a nutshell

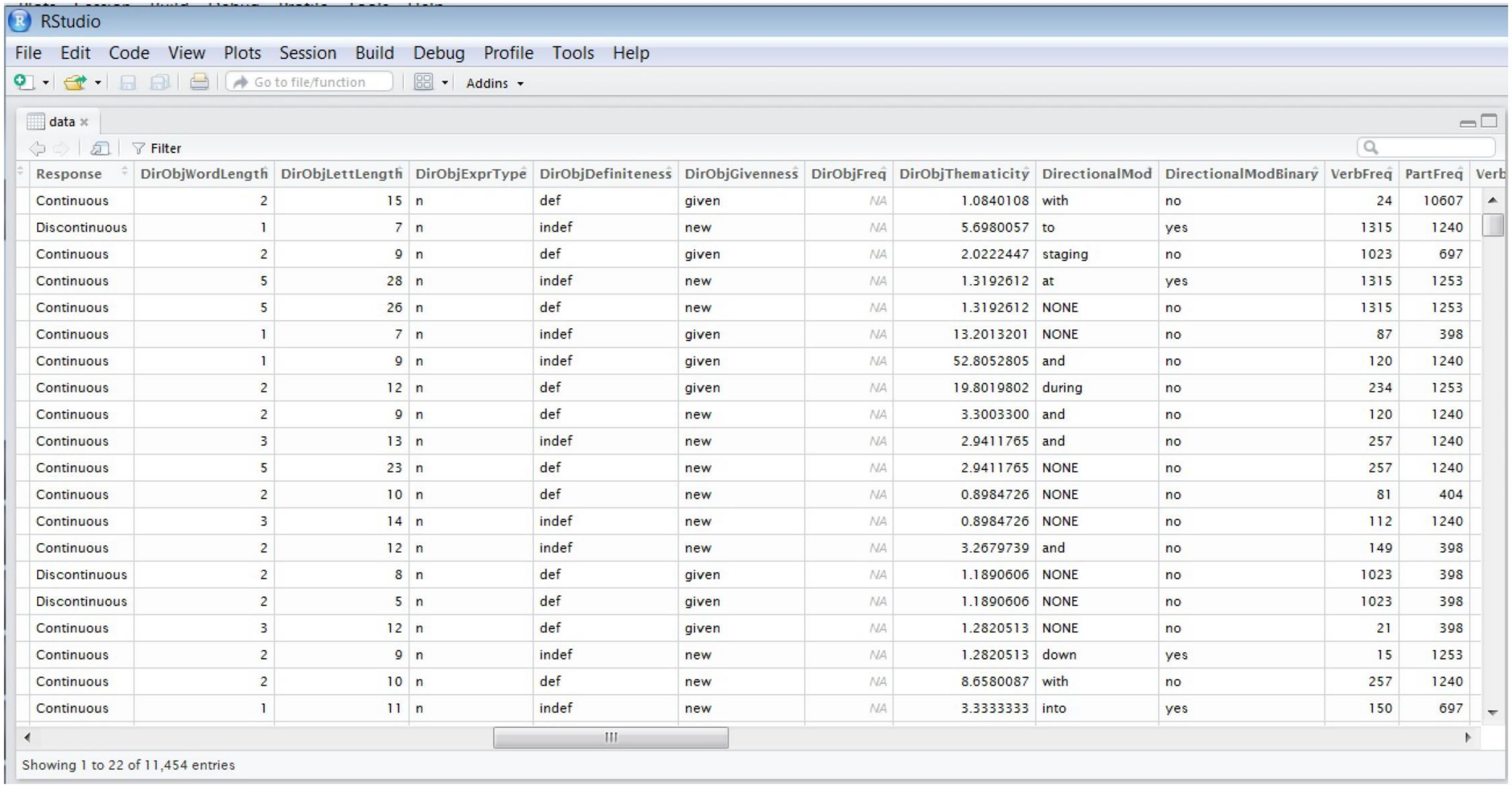
1. **Retrieve** interchangeable tokens (and interchangeable tokens only!) from the corpus database (hand-coding)
2. **Annotate** for the various known constraints (partly hand-coding)
3. **Analyze**.

4. Conduct supplementary **rating-task experiments** (not today’s topic)

Fairly large datasets

genitive alternation	*N* = 13,798
dative alternation	*N* = 13,171
particle placement alternation	*N* = 11,340

Corpus track: richly annotated variationist datasets


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
data
Filter

Response	DirObjWordLength	DirObjLettLength	DirObjExprType	DirObjDefiniteness	DirObjGivenness	DirObjFreq	DirObjThematicity	DirectionalMod	DirectionalModBinary	VerbFreq	PartFreq	Verb
Continuous	2	15	n	def	given	NA	1.0840108	with	no	24	10607	
Discontinuous	1	7	n	indef	new	NA	5.6980057	to	yes	1315	1240	
Continuous	2	9	n	def	given	NA	2.0222447	staging	no	1023	697	
Continuous	5	28	n	indef	new	NA	1.3192612	at	yes	1315	1253	
Continuous	5	26	n	def	new	NA	1.3192612	NONE	no	1315	1253	
Continuous	1	7	n	indef	given	NA	13.2013201	NONE	no	87	398	
Continuous	1	9	n	indef	given	NA	52.8052805	and	no	120	1240	
Continuous	2	12	n	def	given	NA	19.8019802	during	no	234	1253	
Continuous	2	9	n	def	new	NA	3.3003300	and	no	120	1240	
Continuous	3	13	n	indef	new	NA	2.9411765	and	no	257	1240	
Continuous	5	23	n	def	new	NA	2.9411765	NONE	no	257	1240	
Continuous	2	10	n	def	new	NA	0.8984726	NONE	no	81	404	
Continuous	3	14	n	indef	new	NA	0.8984726	NONE	no	112	1240	
Continuous	2	12	n	indef	new	NA	3.2679739	and	no	149	398	
Discontinuous	2	8	n	def	given	NA	1.1890606	NONE	no	1023	398	
Discontinuous	2	5	n	def	given	NA	1.1890606	NONE	no	1023	398	
Continuous	3	12	n	def	given	NA	1.2820513	NONE	no	21	398	
Continuous	2	9	n	indef	new	NA	1.2820513	down	yes	15	1253	
Continuous	2	10	n	def	new	NA	8.6580087	with	no	257	1240	
Continuous	1	11	n	indef	new	NA	3.3333333	into	yes	150	697	

Showing 1 to 22 of 11,454 entries

Key findings

1. Effect directions are stable.
2. Constraint strength is (fairly) variable.
3. All alternations are not equal.

More on VADIS

Comparative Sociolinguistics: exploring relatedness between dialects/varieties

Three lines of evidence in Comparative Sociolinguistics and VADIS
(e.g. Tagliamonte 2001):

1. Are the same constraints significant across varieties?

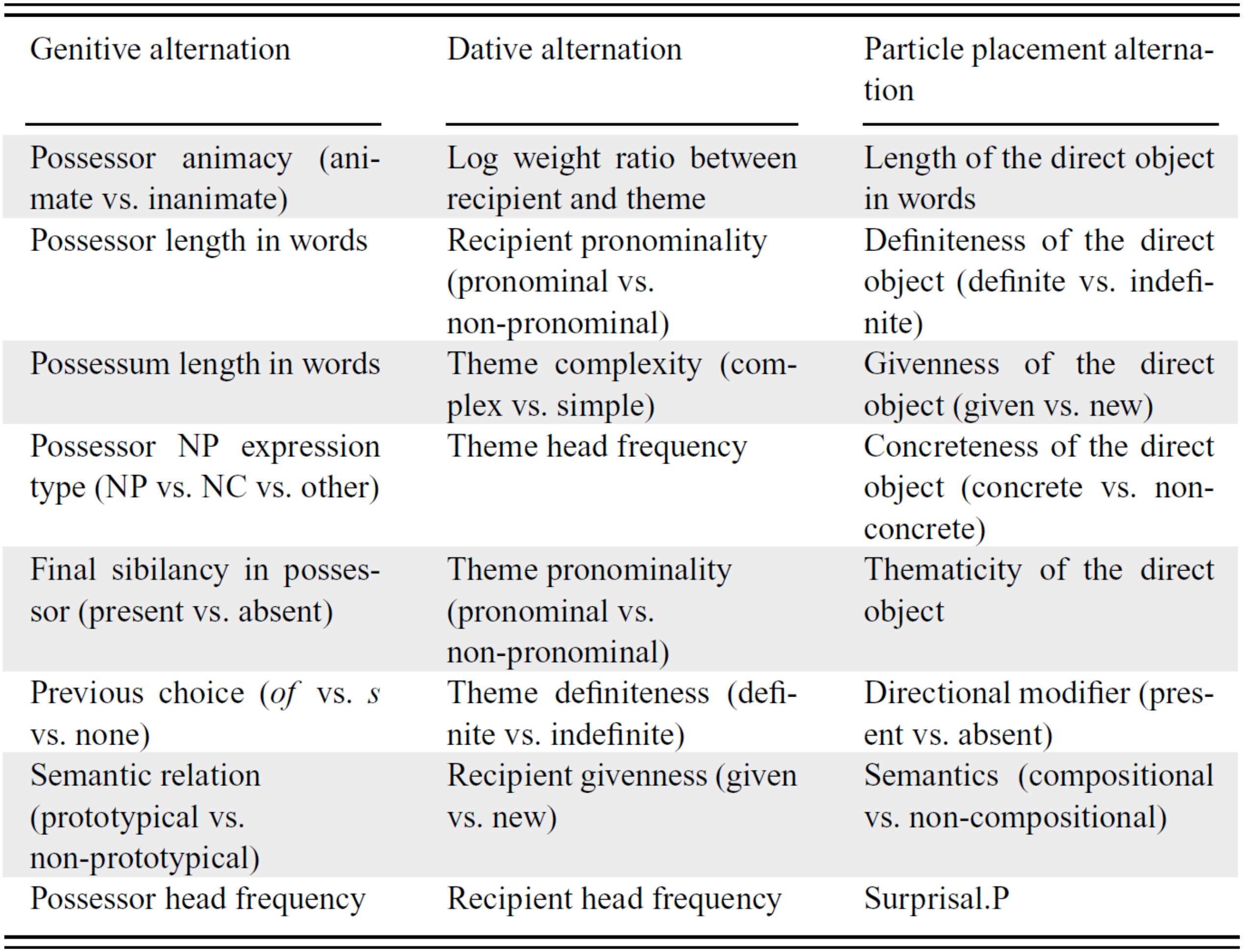
2. Do the constraints have the same strength across varieties?

3. Is the constraint hierarchy similar?

The VADIS pipeline: 7 steps

1. define, per alternation, the *p* most important constraints on variation
2. calculate a series of mixed-effects logistic regression models, one per variety and alternation
3. determine cross-variety distance based on predictor significance
4. determine cross-variety distance based on the magnitude of effects
5. fit a series of conditional random forest models (CRFs), one per variety and alternation
6. determine cross-variety distance basedon the importance rankings of the predictors
7. analyze

Probabilistic constraints under analysis
(Szmrecsanyi & Grafmiller 2023: Table 6.1)



Genitive alternation	Dative alternation	Particle placement alternation
Possessor animacy (animate vs. inanimate)	Log weight ratio between recipient and theme	Length of the direct object in words
Possessor length in words	Recipient pronominality (pronominal vs. non-pronominal)	Definiteness of the direct object (definite vs. indefinite)
Possessum length in words	Theme complexity (complex vs. simple)	Givenness of the direct object (given vs. new)
Possessor NP expression type (NP vs. NC vs. other)	Theme head frequency	Concreteness of the direct object (concrete vs. non-concrete)
Final sibilancy in possessor (present vs. absent)	Theme pronominality (pronominal vs. non-pronominal)	Thematicity of the direct object
Previous choice (*of* vs. *s* vs. none)	Theme definiteness (definite vs. indefinite)	Directional modifier (present vs. absent)
Semantic relation (prototypical vs. non-prototypical)	Recipient givenness (given vs. new)	Semantics (compositional vs. non-compositional)
Possessor head frequency	Recipient head frequency	Surprisal.P

VADIS outputs

- **Similarity coeffificients**: mean inverse distance scores to quantify overall similarity between the probabilistic grammars under investigation
- **Visualization**: pairwise distances yield distance matrices, which in turn serve as input to Multidimensional Scaling (MDS) and similar techniques

R package

An R package (under development) which performs all of the above calculations is available from https://github.com/jasongraf1/VADIS.

R scripts & datasets are available as supplementary materials to Szmrecsanyi & Grafmiller (2023).

Similarity coefficients

Calculating similarity coefficients

Similarity coefficients range between 0 and 1
(0: probabilistic grammars are totally different, 1: absolutely identical)

- 1st line of evidence (**significance**):
  similarity proportional to number of shared significance classifications (squared Euclidean distance)
- 2nd line of evidence (**effect strength**):
  similarity proportional to extent to which effect strengths are similar (Euclidean distance)
- 3rd line of evidence (**hierarchy**):
  similarity proportional to extent to which predictor rankings are similar (Spearman’s rho)

An example

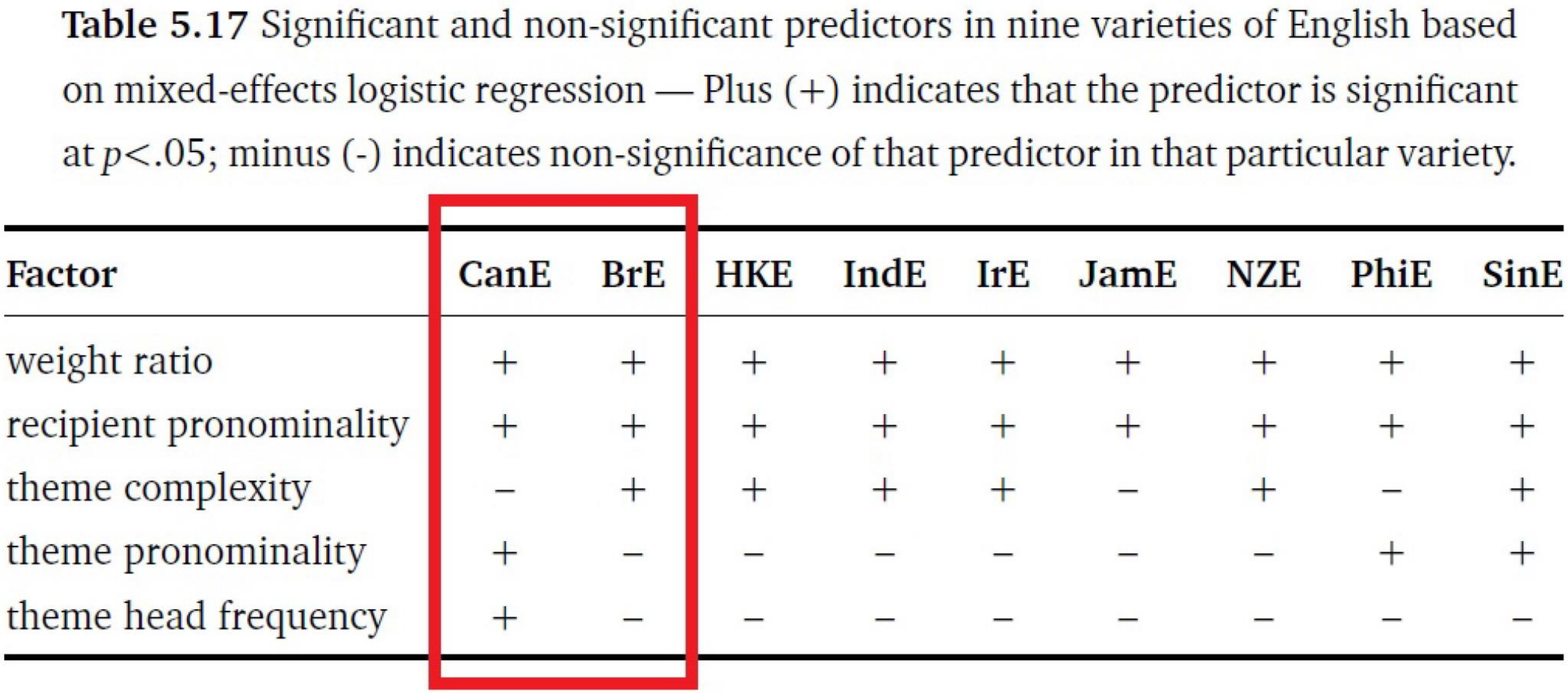


Table 5.17 Significant and non-significant predictors in nine varieties of English based on mixed-effects logistic regression — Plus (+) indicates that the predictor is significant at *p*<.05; minus (-) indicates non-significance of that predictor in that particular variety.

Factor	CanE	BrE	HKE	IndE	IrE	JamE	NZE	PhiE	SinE
weight ratio	+	+	+	+	+	+	+	+	+
recipient pronominality	+	+	+	+	+	+	+	+	+
theme complexity	–	+	+	+	+	–	+	–	+
theme pronominality	+	–	–	–	–	–	–	+	+
theme head frequency	+	–	–	–	–	–	–	–	–

(Röthlisberger 2018: Table 5.17)

Similarity coefficients – all data

(Szmrecsanyi & Grafmiller 2023: Table 6.5)

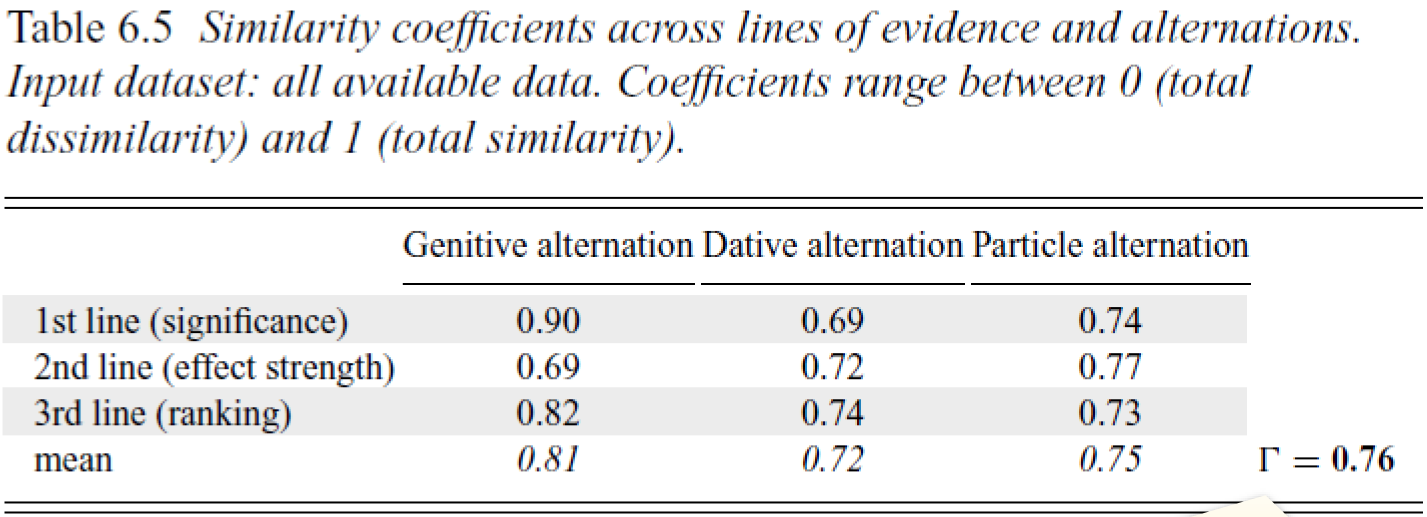


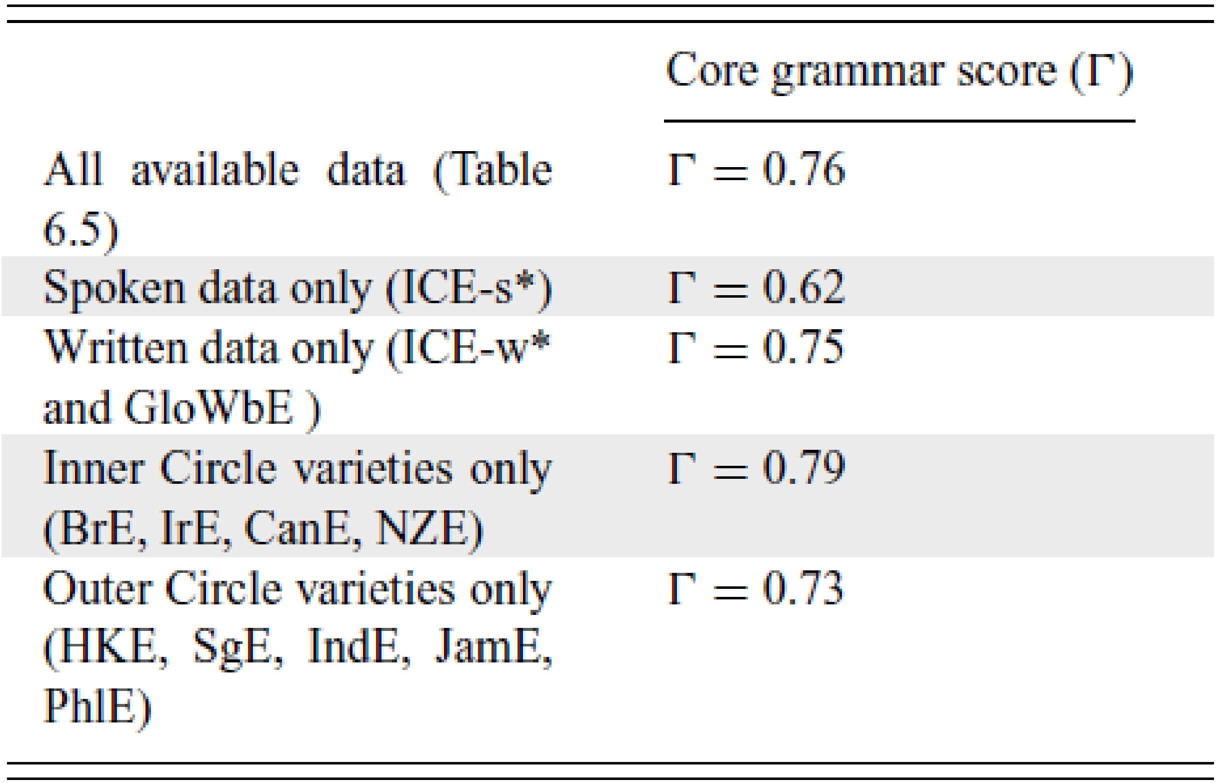
Table 6.5 *Similarity coefficients across lines of evidence and alternations. Input dataset: all available data. Coefficients range between 0 (total dissimilarity) and 1 (total similarity).*

	Genitive alternation	Dative alternation	Particle alternation	
1st line (significance)	0.90	0.69	0.74	
2nd line (effect strength)	0.69	0.72	0.77	
3rd line (ranking)	0.82	0.74	0.73	
mean	*0.81*	*0.72*	*0.75*	$\Gamma = \mathbf{0.76}$

core grammar score Γ

mean similarity coefficient across lines of evidence and across all alternations

Experimenting with sub-datasets
(Szmrecsanyi & Grafmiller 2023: Table 6.6)



	Core grammar score (Γ)
All available data (Table 6.5)	$\Gamma = 0.76$
Spoken data only (ICE-s*)	$\Gamma = 0.62$
Written data only (ICE-w* and GloWbE)	$\Gamma = 0.75$
Inner Circle varieties only (BrE, IrE, CanE, NZE)	$\Gamma = 0.79$
Outer Circle varieties only (HKE, SgE, IndE, JamE, PhlE)	$\Gamma = 0.73$

Interim summary

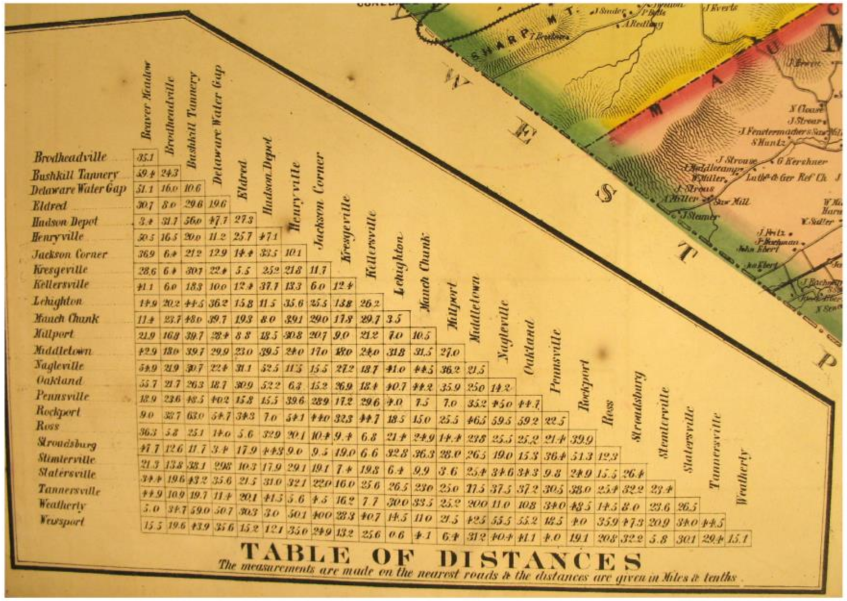
- Overall: substantial to very strong overlap
- Inner Circle varieties are more homogeneous than Outer Circle varieties
- Spoken production: more heterogeneity than written production
- All alternations are not equal

Mapping out (dis)similarity relationships between varieties

Pairwise similarity values ⇨ distance matrices

Customary input in classical Séguy-Goebl-Nerbonne-style dialectometry

(Séguy 1971; Goebl 1982; Nerbonne et al. 1999)



	Beaver Meadow	Brodheadville	Bushkill Tannery	Delaware Water Gap	Eldred	Hudson Depot	Henryville	Jackson Corner	Kresgeville	Kellersville	Lehighton	Mauch Chunk	Millport	Middletown	Nagleville	Oakland	Pennsville	Rockport	Ross	Stroudsburg	Stemlersville	Slatersville	Tannersville	Weatherly
Brodheadville	35.1																							
Bushkill Tannery	59.4	24.3																						
Delaware Water Gap	51.1	16.0	10.6																					
Eldred	30.7	8.0	29.6	19.6																				
Hudson Depot	3.4	31.1	56.0	47.7	27.3																			
Henryville	50.5	16.5	20.0	11.2	25.7	47.1																		
Jackson Corner	36.9	6.4	21.2	12.9	14.4	33.5	10.1																	
Kresgeville	28.6	6.4	30.1	22.4	5.5	25.2	21.8	11.7																
Kellersville	41.1	6.0	18.3	10.0	12.4	37.1	13.3	6.0	12.4															
Lehighton	14.9	20.2	44.5	36.2	15.8	11.5	35.6	25.5	13.8	26.2														
Mauch Chunk	11.4	23.7	48.0	39.7	19.3	8.0	39.1	29.0	17.8	29.1	3.5													
Millport	21.9	16.8	39.7	28.4	8.8	18.5	30.8	20.7	9.0	21.2	7.0	10.5												
Middletown	42.9	18.0	39.7	29.9	23.0	39.5	24.0	17.0	18.0	24.0	31.8	31.5	27.0											
Nagleville	54.9	21.9	30.7	22.4	31.1	52.5	11.5	15.5	27.2	18.7	41.0	44.5	36.2	21.5										
Oakland	55.7	21.7	26.3	18.7	30.9	52.2	6.3	15.2	26.9	18.4	40.7	44.2	35.9	25.0	14.2									
Pennsville	18.9	23.6	48.5	40.2	15.8	15.5	39.6	28.9	17.2	29.6	4.0	7.5	7.0	35.2	45.0	44.7								
Rockport	9.0	32.7	63.0	54.7	34.3	7.0	54.1	44.0	32.3	44.7	18.5	15.0	25.5	46.5	59.5	59.2	22.5							
Ross	36.3	5.8	25.1	14.0	5.6	32.9	20.1	10.4	9.4	6.8	21.4	24.9	14.4	23.8	25.5	25.2	21.4	39.9						
Stroudsburg	47.7	12.6	11.7	3.4	17.9	44.3	9.0	9.5	19.0	6.6	32.8	36.3	28.0	26.5	19.0	15.3	36.4	51.3	12.3					
Stemlersville	21.3	13.8	38.1	29.8	10.3	17.9	29.1	19.1	7.4	19.8	6.4	9.9	3.6	25.4	34.6	34.3	9.8	24.9	15.5	26.4				
Slatersville	34.4	19.6	43.2	35.6	21.5	31.0	32.1	22.0	16.0	25.6	26.5	23.0	25.0	11.5	37.5	37.2	30.5	38.0	25.4	32.2	23.4			
Tannersville	44.9	10.9	19.7	11.4	20.1	41.5	5.6	4.5	16.2	7.7	30.0	33.5	25.2	20.0	11.0	10.8	34.0	48.5	14.5	8.0	23.6	26.5		
Weatherly	5.0	34.7	59.0	50.7	30.3	3.0	50.1	40.0	28.3	40.7	14.5	11.0	21.5	42.5	55.5	55.2	18.5	4.0	35.9	47.3	20.9	34.0	44.5	
Weissport	15.5	19.6	43.9	35.6	15.2	12.1	35.0	24.9	13.2	25.6	0.6	4.1	6.4	31.2	40.4	41.1	4.0	19.1	20.8	32.2	5.8	30.1	29.4	15.1

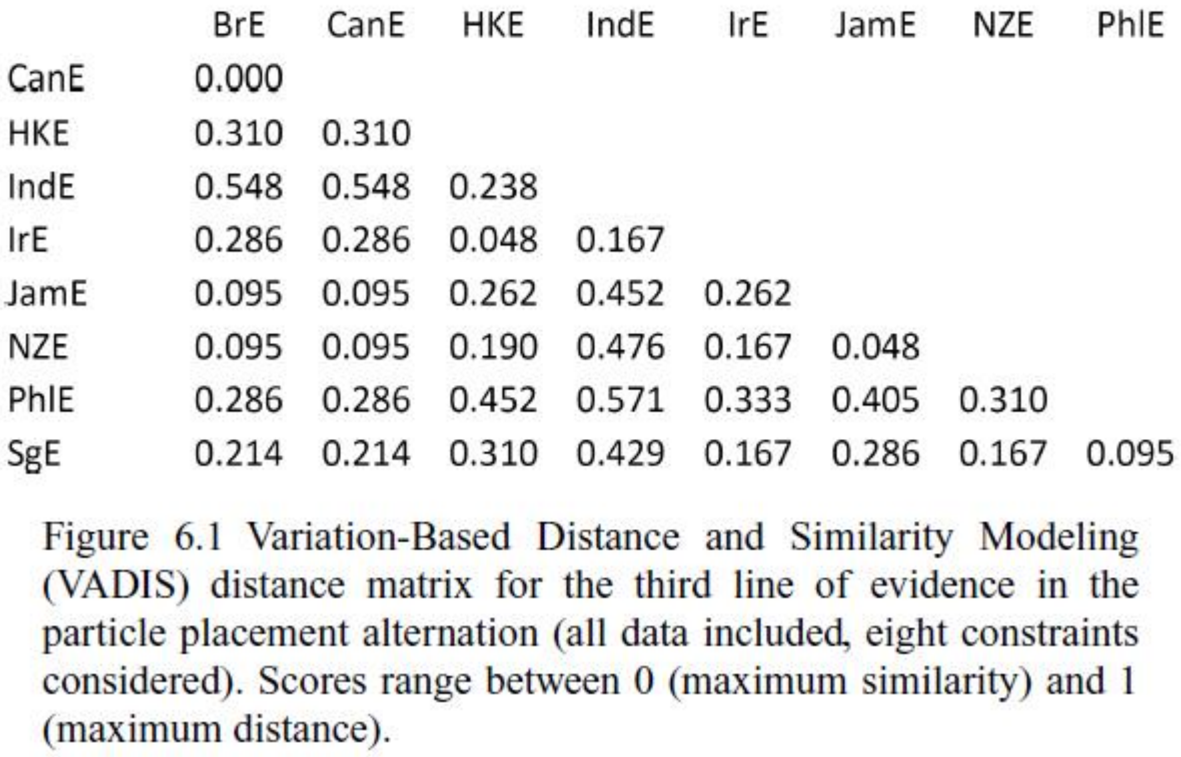
TABLE OF DISTANCES

The measurements are made on the nearest roads & the distances are given in Miles & tenths

https://ancestortracks.com/MonroeCo,1860/DistanceTable.jpg

Using inverse pairwise similarity coefficients as distance measure

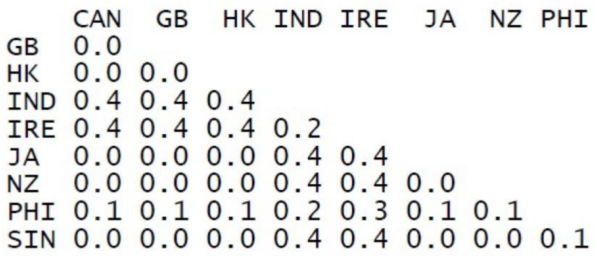
(Szmrecsanyi & Grafmiller 2023: Figure 6.1)



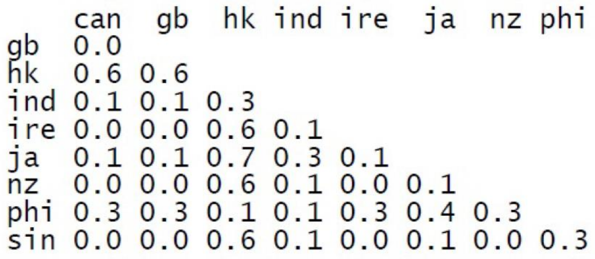
	BrE	CanE	HKE	IndE	IrE	JamE	NZE	PhlE
CanE	0.000							
HKE	0.310	0.310						
IndE	0.548	0.548	0.238					
IrE	0.286	0.286	0.048	0.167				
JamE	0.095	0.095	0.262	0.452	0.262			
NZE	0.095	0.095	0.190	0.476	0.167	0.048		
PhlE	0.286	0.286	0.452	0.571	0.333	0.405	0.310	
SgE	0.214	0.214	0.310	0.429	0.167	0.286	0.167	0.095

Figure 6.1 Variation-Based Distance and Similarity Modeling (VADIS) distance matrix for the third line of evidence in the particle placement alternation (all data included, eight constraints considered). Scores range between 0 (maximum similarity) and 1 (maximum distance).

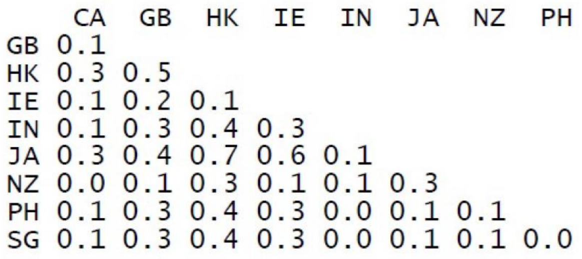
Merging across lines of evidence



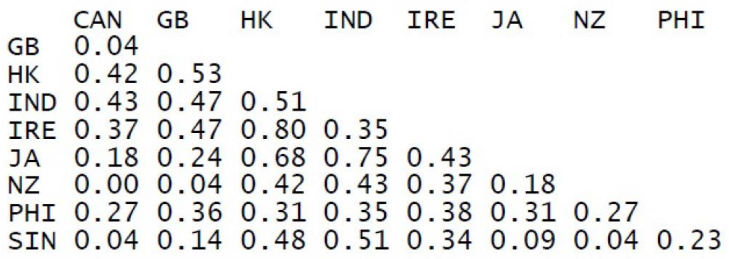
```
     CAN  GB   HK  IND IRE  JA   NZ  PHI
GB   0.0
HK   0.0 0.0
IND  0.4 0.4 0.4
IRE  0.4 0.4 0.4 0.2
JA   0.0 0.0 0.0 0.4 0.4
NZ   0.0 0.0 0.0 0.4 0.4 0.0
PHI  0.1 0.1 0.1 0.2 0.3 0.1 0.1
SIN  0.0 0.0 0.0 0.4 0.4 0.0 0.0 0.1
```



```
     can  gb   hk  ind ire  ja   nz  phi
gb   0.0
hk   0.6 0.6
ind  0.1 0.1 0.3
ire  0.0 0.0 0.6 0.1
ja   0.1 0.1 0.7 0.3 0.1
nz   0.0 0.0 0.6 0.1 0.0 0.1
phi  0.3 0.3 0.1 0.1 0.3 0.4 0.3
sin  0.0 0.0 0.6 0.1 0.0 0.1 0.0 0.3
```



```
    CA   GB   HK   IE   IN   JA   NZ   PH
GB 0.1
HK 0.3 0.5
IE 0.1 0.2 0.1
IN 0.1 0.3 0.4 0.3
JA 0.3 0.4 0.7 0.6 0.1
NZ 0.0 0.1 0.3 0.1 0.1 0.3
PH 0.1 0.3 0.4 0.3 0.0 0.1 0.1
SG 0.1 0.3 0.4 0.3 0.0 0.1 0.1 0.0
```



```
     CAN  GB   HK   IND  IRE  JA   NZ   PHI
GB   0.04
HK   0.42 0.53
IND  0.43 0.47 0.51
IRE  0.37 0.47 0.80 0.35
JA   0.18 0.24 0.68 0.75 0.43
NZ   0.00 0.04 0.42 0.43 0.37 0.18
PHI  0.27 0.36 0.31 0.35 0.38 0.31 0.27
SIN  0.04 0.14 0.48 0.51 0.34 0.09 0.04 0.23
```

Alternation-specific MDS plots
(Szmrecsanyi & Grafmiller 2023: Figure 6.3)

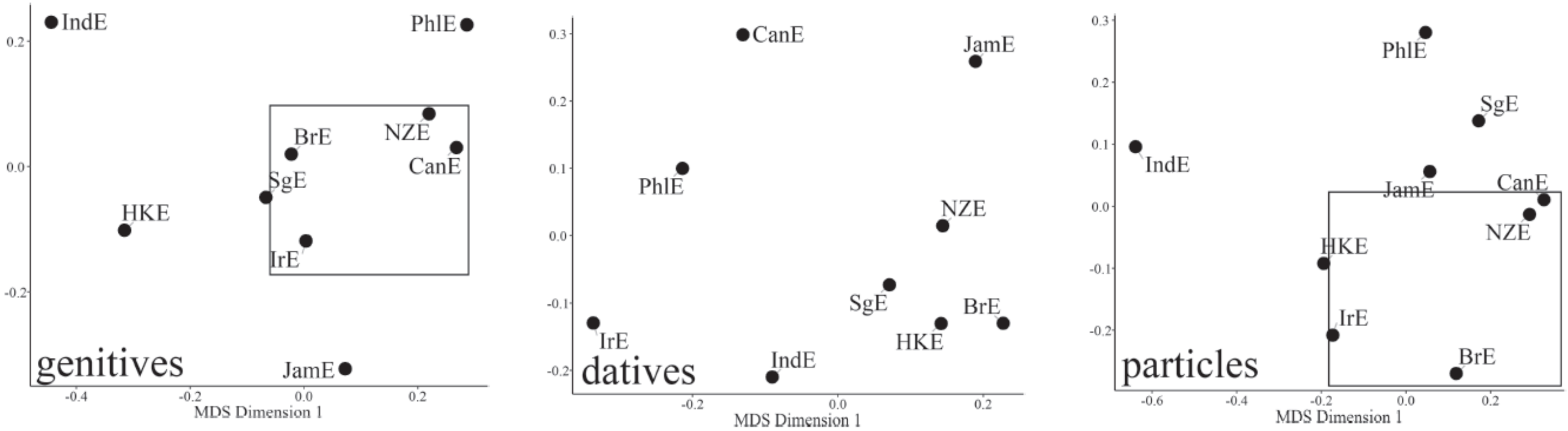


Figure 6.3 Multidimensional scaling representation of compromise distances per alternation: a) genitive alternation; b) dative alternation; c) particle placement alternation. Distances between data points in plots is proportional to probabilistic grammar distances between varieties. Boxes depict Inner Circle clusters.

NeighborNet
(Szmrecsanyi & Grafmiller 2023: Figure 6.6)

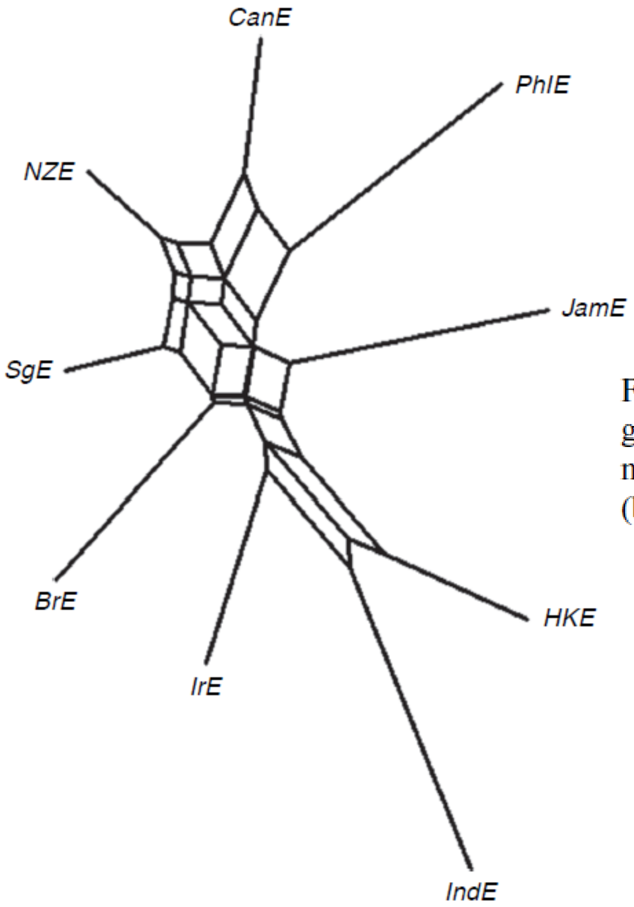


Figure 6.6 Visualizing aggregate similarities: NeighborNet diagram depicting the Γ-matrix (a single compromise distance matrix merged across all lines and alternations). Internode distances (branch lengths) are proportional to cophenetic linguistic distances.

Interim summary

- Using the dialectometrical toolbox, VADIS can map (dis)similarity relationships between varieties
- Big picture: split between Inner Circle varieties and Outer Circle varieties

Beyond English, and beyond geographical variation

Historical variation in English

Michiels, Jakob (2022). The Diachronic Perspective of VADIS. Unpublished MA thesis, KU Leuven

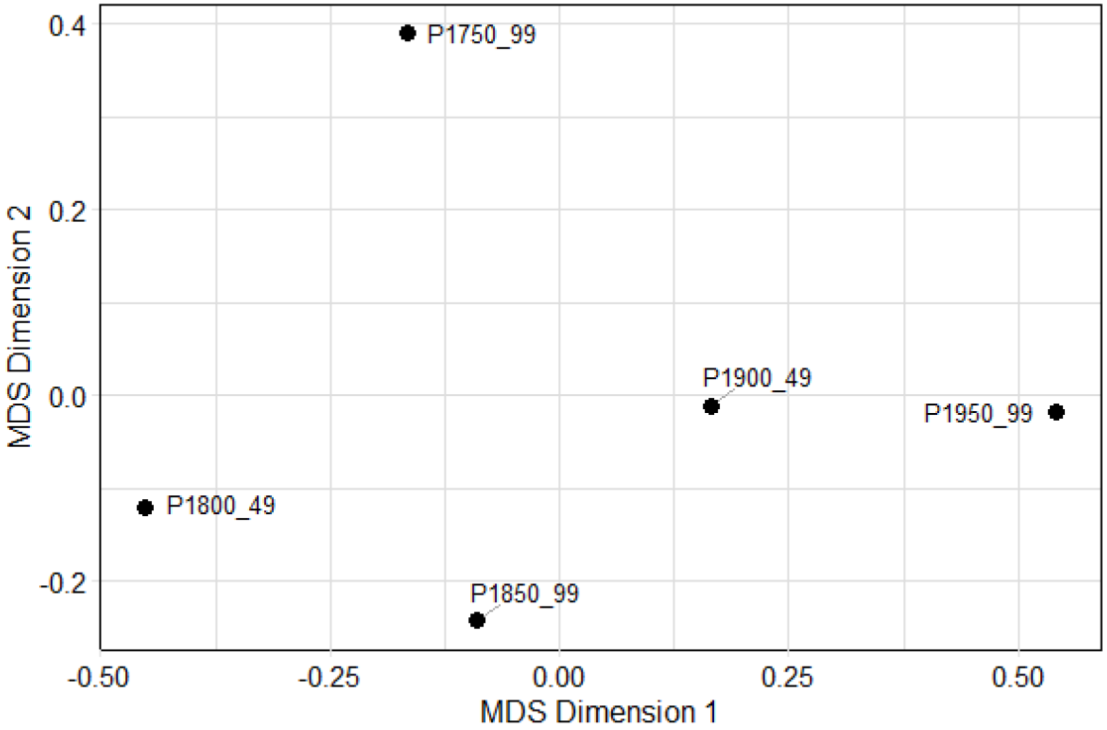


Figure 1 MDS representation of the compromise distances between five periods (1750-1799, 1800-1849, 1850-1899, 1900-1949, and 1950-1999) on the genitive alternation. Distances between data points in plot is proportional to probabilistic grammar distances between periods.

Register variation

Zhang, Xu (2023). A VADIS-based exploration of register variation. Unpublished MA thesis, KU Leuven (forthcoming in *Register Studies* pending revisions)

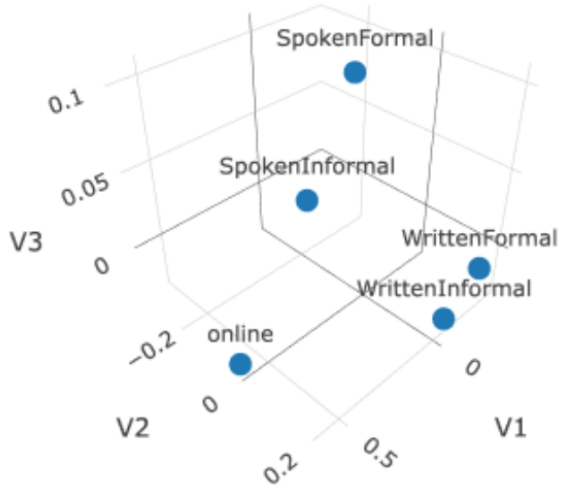


Figure1 MDS representation of the distance between five registers (Online, SpokenFormal, SpokenInformal, WrittenFormal and WrittenInformal) based particle placement alternation. Distance between data points in this plot is proportional to people's probabilistic grammar across the five registers.

Variation in Mandarin Chinese

Li, Yi, Benedikt Szmrecsanyi & Weiwei Zhang. 2024. Across time, space, and genres: measuring probabilistic grammar distances between varieties of Mandarin. *Linguistics Vanguard*. https://doi.org/10.1515/lingvan-2022-0134

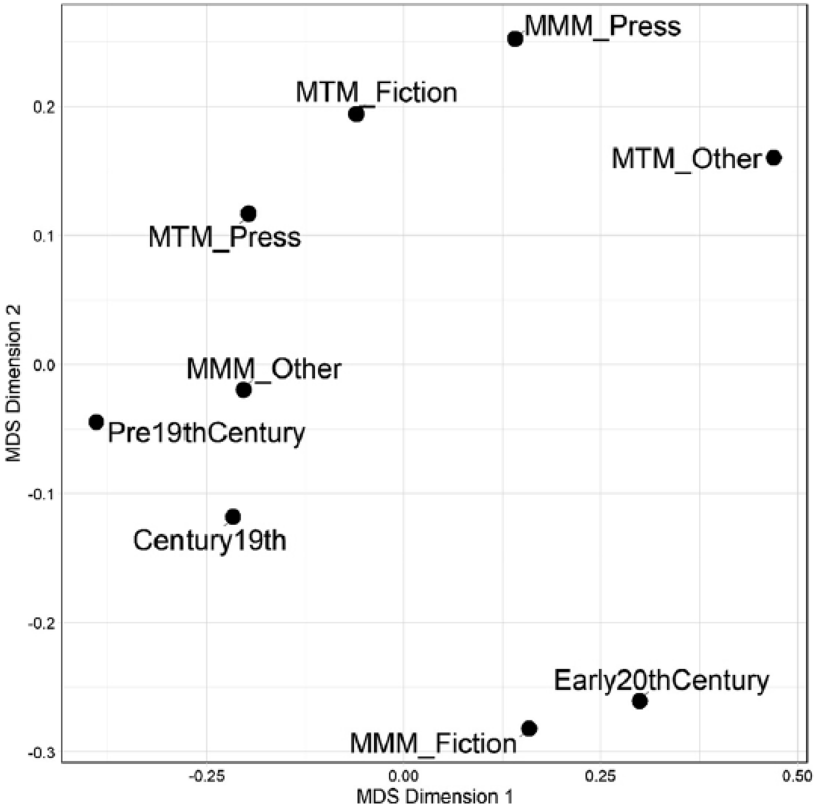


Figure 2: Multidimensional scaling plot representation of fused distances of the theme-recipient alternation. Distances between data points in the plot are proportional to probabilistic grammar distances between varieties. “Pre 19th century” refers to the fourteenth–eighteenth-century variety; “century 19th” to the nineteenth-century variety; “MMM” to Modern Mainland Mandarin; “MTM” to Modern Taiwan Mandarin.

Concluding remarks

Take-home message

- VADIS gauges the extent and structure of inter-speaker variation through assessing intra-speaker variation
- More usage-based bent than classical dialectometry, also more responsible cognitively
- Can pick up subtle differences even in cases where lects happen to have the same inventory of forms
- Scales up better to more varieties and more variation phenomena than classical comparative sociolinguistics
- Limitation: data-hungry

Thank you!

benszm@kuleuven.be
www.benszm.net