# Representing grammatical similarity in comparative variationist analysis

UNIVERSITY OF BIRMINGHAM

Jason Grafmiller

## Background

Existing techniques for variationist modelling are capable of capturing differences with respect to individual linguistic constraints across varieties, e.g. as interaction effects, yet they offer limited insight into the relative degree of similarity across *systems* of constraints in aggregate. Drawing inspiration from methods in Comparative Sociolinguistics[1] and corpus-based dialectometry[2], we introduce a new method for assessing the grammatical similarity between language varieties, **Variation-based Distance and Similarity Modeling (VADIS)**[3,4], which applies distance-based visualization techniques to the outputs of the standard tools used for quantitative analysis of sociolinguistic variables.

## Comparative Sociolinguistics

3 'lines of evidence' for determining relatedness among varieties vis-à-vis a given variable:

**Line 1** — **Statistical Significance**

Are the same constraints significant across varieties?

**Line 2** — **Effect Strength**

Do the constraints have the same strength (and direction) across varieties?

**Line 3** — **Constraint ranking**

Is the relative predictive importance of constraints the same across varieties?

- Early uses simply visually compared these lines of evidence
- Modern approaches tend to use single models with multiple interactions, but complex models with many interactions are difficult to interpret, and often have problems converging
- No existing way of assessing and evaluating different lines of evidence comprehensively

## Variation-based Distance and Similarity Modelling

VADIS fits models on individual datasets and assesses the overall (dis)similarity among the outputs of these models using common tools for dimension reduction and phylogenetic analysis.

### Key Steps

1. Fit identical regression model to each variety dataset
2. **Line 1:** Calculate distances among varieties based on **statistical significance of constraints**
3. **Line 2:** Calculate distances based on **constraint coefficient values**
4. Fit identical random forest models to each variety dataset
5. **Line 3:** Calculate distances based on **constraint importance rankings**

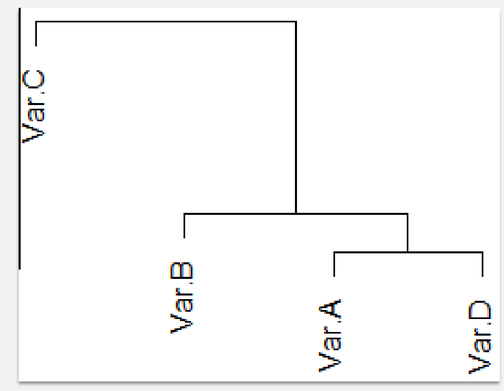
## Visualizing relatedness

| | Var.A | Var.B | Var.C | Var.D |
|---|---|---|---|---|
| Factor 1 | 1.5 | 2.0 | 0.6 | 1.3 |
| Factor 2 | 0.2 | 0.1 | 1.2 | 0.4 |
| Factor 3 | -0.5 | 0.1 | -0.1 | -0.5 |
| Factor 4 | -1.0 | -1.3 | -0.4 | -0.9 |

Table of regression coefficient values (or significance values (Y/N) or random forest rankings) of each constraint calculated from models on individual variety datasets

| | Var.A | Var.B | Var.C |
|---|---|---|---|
| Var.B | 0.84 | | |
| Var.C | 4.17 | 4.73 | |
| Var.D | 0.30 | 1.05 | 3.98 |

Compute pairwise distance matrix based on table of out from models. Larger values reflect greater dissimilarity




Visualize relatedness among varieties via e.g. cluster analysis, neighbour-nets[5] or multidimensional scaling

## Case study: Genre variation in American English genitives

**English genitive alternation:**

(1) **the best interest of both governments** [*of* genitive]

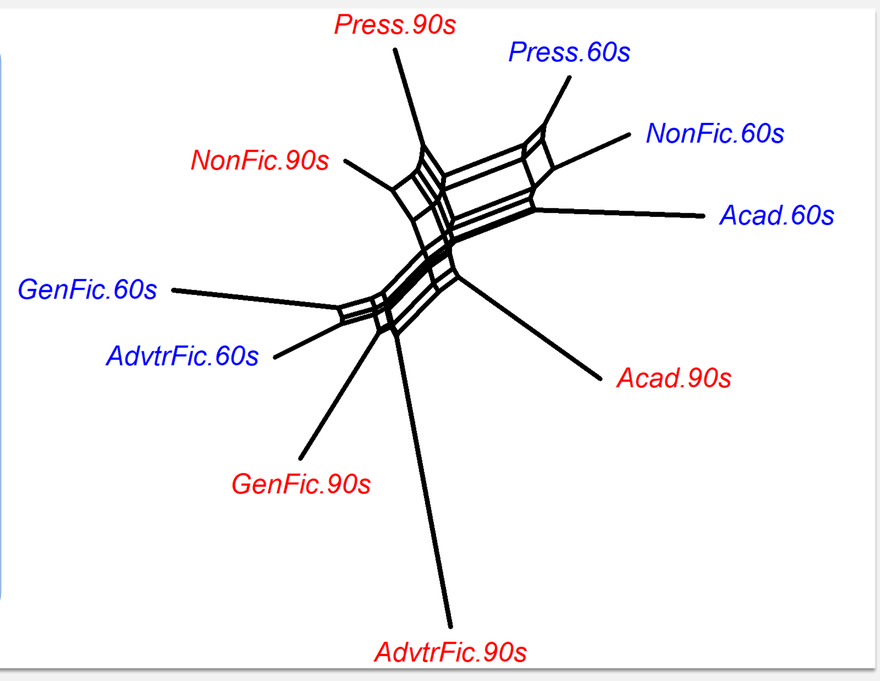
(2) **both governments' best interest** [*s* genitive]

- Examine genitive alternation in 5 genres sampled at 2 time periods (**1960s** and **1990s**):
  - *Press*, *Academic*, Popular non-fiction (*NonFic*), General fiction (*GenFic*), Adventure fiction (*AdvtrFic*)
- For each of the 10 datasets, model genitive choice based on well-known constraints, e.g.
  - Possessor animacy, length, and topicality; semantic relation; presence of a sibilant

Research Question: **How much do genres differ with respect to the factors shaping genitive choice?**

## Findings

- Close affinity between similar genres at similar times
- Large differences in genitive use between fiction and non-fiction genres
- Genitive use in non-fiction genres is changing more markedly than in fiction genres
- Academic genitives are diverging from those in other non-fiction genres




Neighbor-net derived from individual regression model coefficients for each dataset (**Line 2**)

## Advantages of VADIS

- **Macroscopic:** Offers a view of relatedness based on how variable grammars function as a system (not just differences in individual constraints)
- **Versatile:** Can be applied to phonological, lexical or grammatical variables
- **Big data potential:** Can be scaled up to large numbers of varieties and/or individuals, and multiple variables (in progress)

Code and data available on Open Science Framework
R package (in development) available on GitHub

```
devtools::install_github("jasongraf1/VADIS")
```

https://osf.io/cp9dx/

## References

1. Tagliamonte, S. 2013. Comparative Sociolinguistics. In J. K. Chambers & N. Schilling (eds.), *Handbook of Language Variation and Change*, 130–156.
2. Szmrecsanyi, B. & B. Kortmann. 2009. The morphosyntax of varieties of English worldwide: A quantitative perspective. *Lingua* 119(11). 1643–1663.
3. Grafmiller, J. & B. Szmrecsanyi. 2018. Mapping out particle placement in Englishes around the world: A study in comparative sociolinguistic analysis. *LVC* 30(3). 385–412.
4. Szmrecsanyi, B., J. Grafmiller & L. Rosseel. To appear. Variation-based distance and similarity modeling: A case study in World Englishes. *Frontiers*.
5. Bryant, D. & V. Moulton. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2). 255–265.

CONTACT j.grafmiller@bham.ac.uk