Geometric Multivariate Analysis

Finding structure in multivariate data

Stephanie Evert[1] • **Stella Neumann**[2] • **Gerold Schneider**[3] • **Florian Frenken**[2]
[1]FAU Erlangen-Nürnberg • [2]RWTH Aachen • [3]Universität Zürich

Corpus Linguistics 2025 | 2 July 2025

https://quantor-project.github.io/

Funded by
DFG Deutsche Forschungsgemeinschaft German Research Foundation
FNSNF Swiss National Science Foundation
QuanTOR

Introduction

Multivariate analysis of linguistic variation

- Well-established, successful approach to studying linguistic variation based on observable lexico-grammatical features → most widely known multidimensional analysis (MDA: Biber 1988, 1993, …)
- But with limitations …
  - unsupervised FA good for major dimensions of variation, but cannot detect fine-grained patterns (e.g. original vs. translation, male vs. female speakers, language variety)
  - no good strategy for integrating multiple perspectives into a single quantitative analysis (in our case study today: register variation × language varieties)

Geometric multivariate analysis (GMA: Diwersy et al. 2014; Evert & Neumann 2017)

- Geometric perspective with orthogonal projections is intuitive & enables combination of dimensions
- Emphasis on visualisation → understand geometric structure first to obtain overview
- Minimally supervised perspective: linear discriminant analysis (LDA, also cf. Egbert & Biber 2018)

DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Introduction

Geometric multivariate analysis in R

- New: user-friendly R package `gmatools`
  → everything shown in this presentation was done with the R package
- Will be on CRAN soon – for now install directly from GitHub repository (see below)

- Code examples at bottom of slides – see RMarkdon notebook at https://osf.io/9p25y/ for details

```
devtools::install_github("schtepf/GMA/pkg/gmatools") # just once

library(gmatools)
```


Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

RWTH AACHEN UNIVERSITY

Goals & methodological issues

In this talk, I want to …

- … show how to find structure in multivariate data effectively with GMA.
- … convince you that it's easy & fun to apply GMA yourself with `gmatools`!
- … address some methodological issues that came up in recent studies.

Research questions:

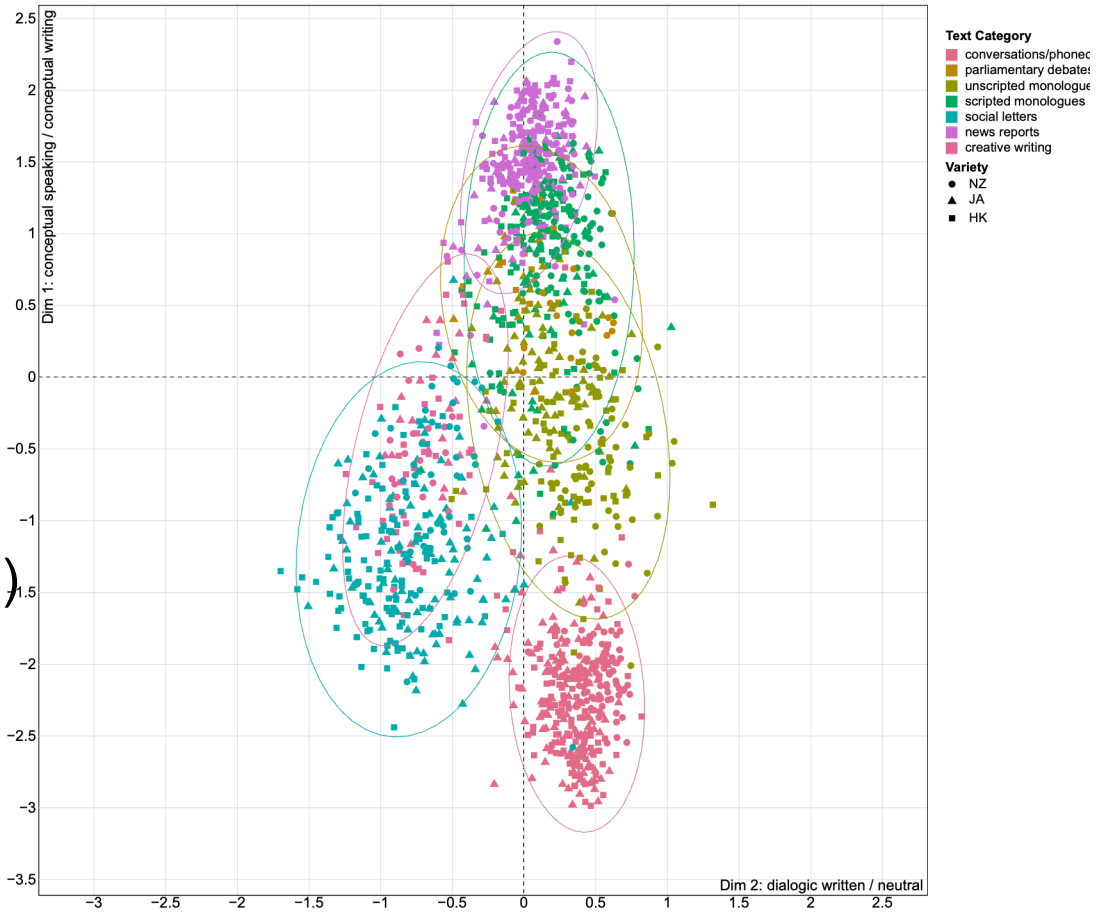
- Are GMA dimensions robust against modification of the data set?
- How stable are GMA dimensions identified by minimally supervised LDA?
- How much is the LDA analysis influenced by imbalance between proxy categories?
- What if there are confounding factors in the proxy categories?
  (especially if multiple distinctions of interest intersect, e.g. register and variety)

DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Case study

Background and objective

- Neumann and Evert (2021)
  - → International Corpus of English (ICE: Greenbaum 1996)
  - → Hong Kong, Jamaica, New Zealand
  - → linguistic variation across these three varieties of English, esp. with respect to register
- Replication on extended data set (Frenken et al. 2025)
  - → six additional ICE components: Canada, Great Britain, India, Ireland, Philippines, Singapore
- Now: expand on the replication experiment (ICE3 = original data | ICE9 = extended set)



Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Case study

Feature extraction

- Consistent clean-up of component markup (Lehmann and Schneider 2012; Conrad et al. *yesterday*)
  → including removal of extra-corpus material
- Fine-grained CLAWS part-of-speech annotation (Garside and Smith 1997)
- Extraction of 41 lexicogrammatical features
  → capture contextual dimensions of register variation
- Reproducible CQP query pipeline to be released separately (Evert et al. 2020)
- Detailed evaluation of precision and recall in progress

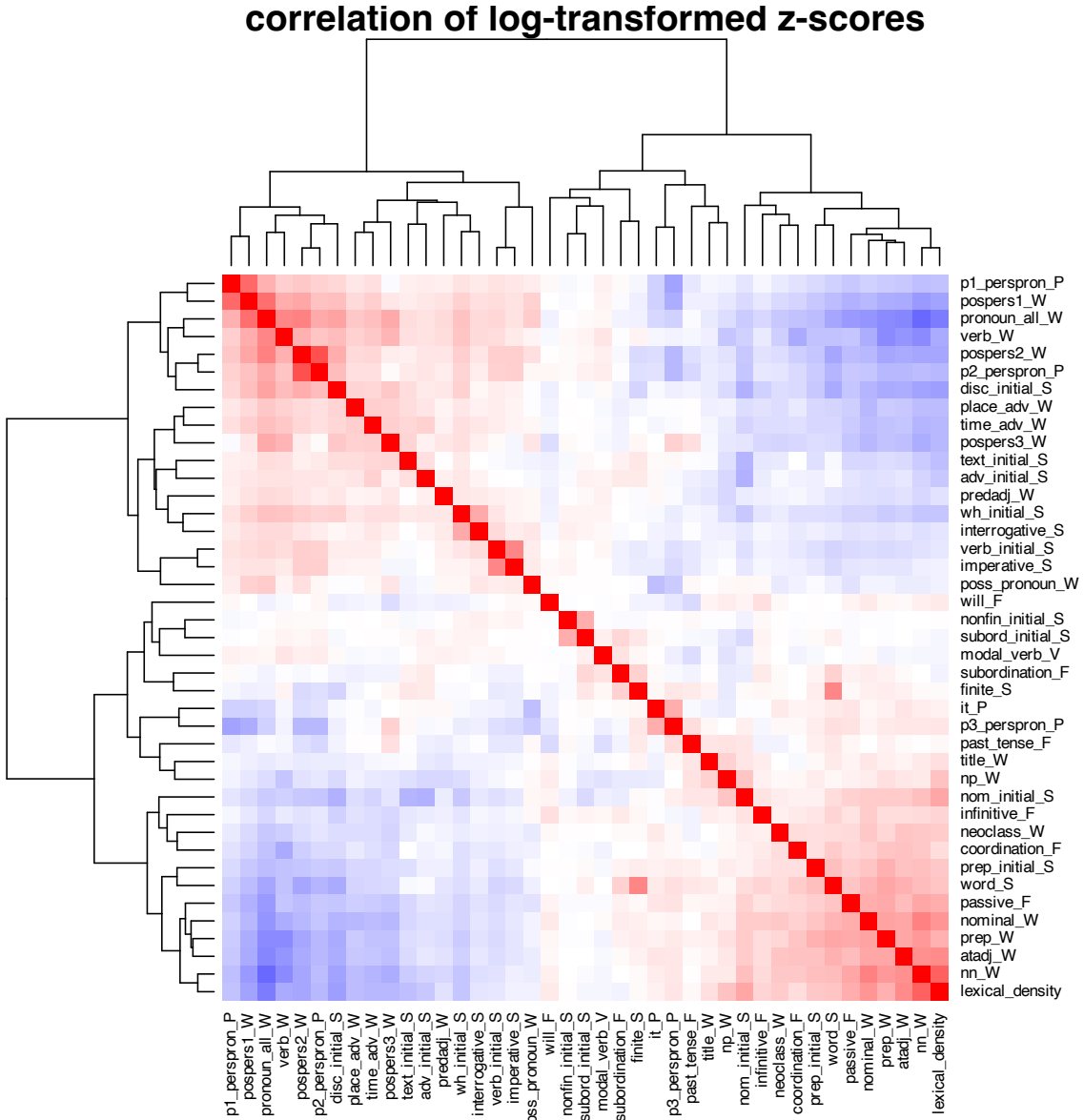
Features		
word/S	pospers2/W	interrogative/S
lexical density	pospers3/W	imperative/S
nn/W	atadj/W	title/W
np/W	predadj/W	place adv/W
nominal/W	prep/W	time adv/W
neoclass/W	finite/S	nom initial/S
poss pronoun/W	past tense/F	prep initial/S
pronoun all/W	will/F	adv initial/S
p1 perspron/P	modal verb/V	text initial/S
p2 perspron/P	verb/W	wh initial/S
p3 perspron/P	infinitive/F	disc initial/S
it/P	passive/F	nonfin initial/S
pospers1/W	coordination/F	subord initial/S
	subordination/F	verb initial/S

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Case study

Preprocessing

- relative frequencies wrt. sensible unit of measurement
- removal of (nearly) collinear features
- standardisation + signed log transformation of features
- exclusion of short texts (ca. 5%–14% of all texts)



Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

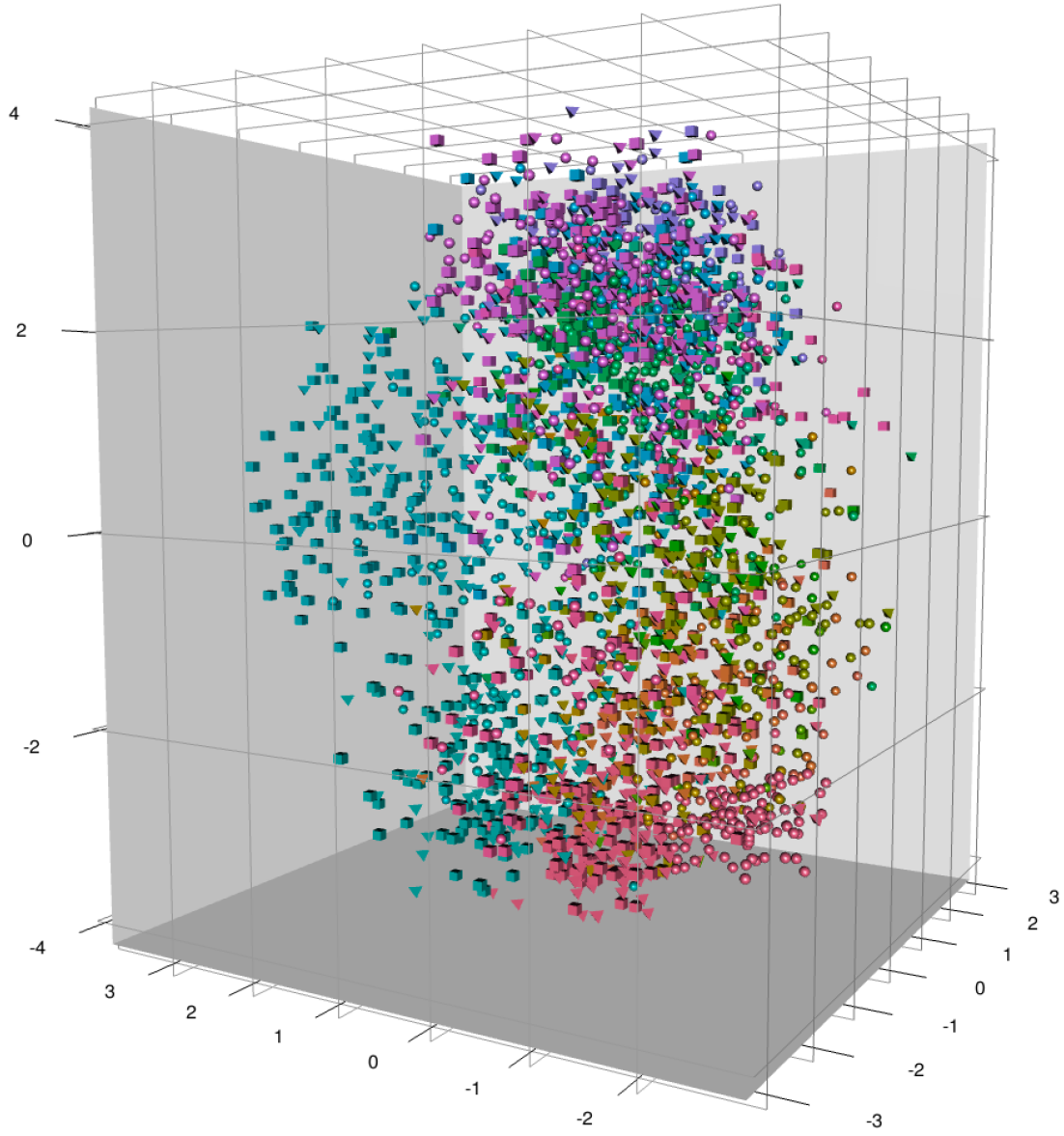
Universität Zürich[UZH]

FAU

RWTH AACHEN UNIVERSITY

Reproducing Neumann & Evert (2021)

- Unsupervised principal component analysis (PCA) of ICE3 data set
- Projection into first 3 PCA dimensions
  - →major dimensions of variation seem to align quite well with ICE text categories (colours)
- Captures R^2 = 45.83% of distance information



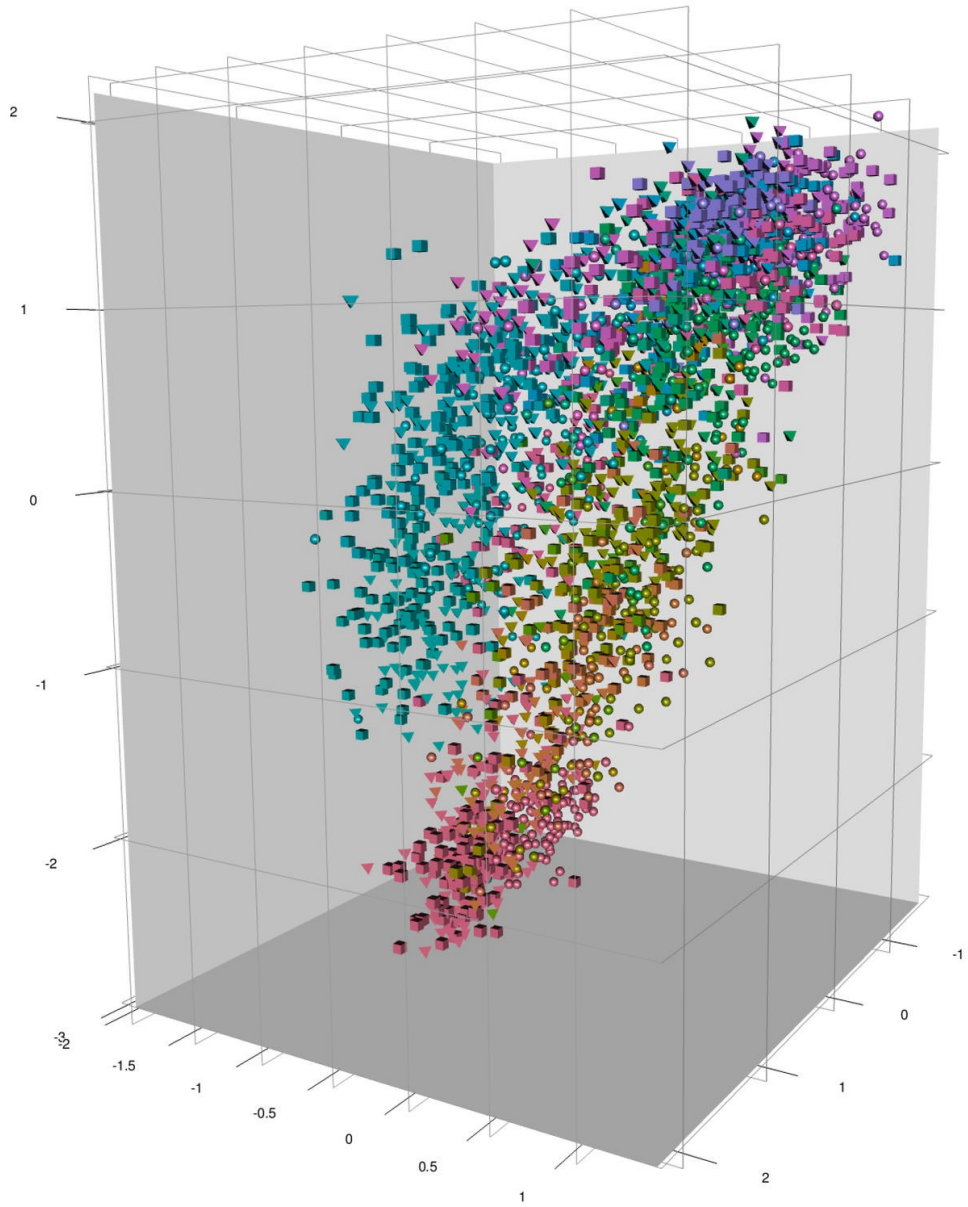

```
ICE3 <- GMA$new(ZL3)
ICE3.pca <- GMA$projection(space="both", dim=1:3)
```


Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Reproducing Neumann & Evert (2021)

- Linear discriminant analysis of ICE3 data set based on 20 mid-level text categories
  - →register space implied by ICE sampling frame
  - →clearer structure & visual interpretation than PCA
- Projection into focus space spanned by first 4 LDA dimensions (here: visualise only 3 dim's)
- Captures R^2 = 17.87% of distance information

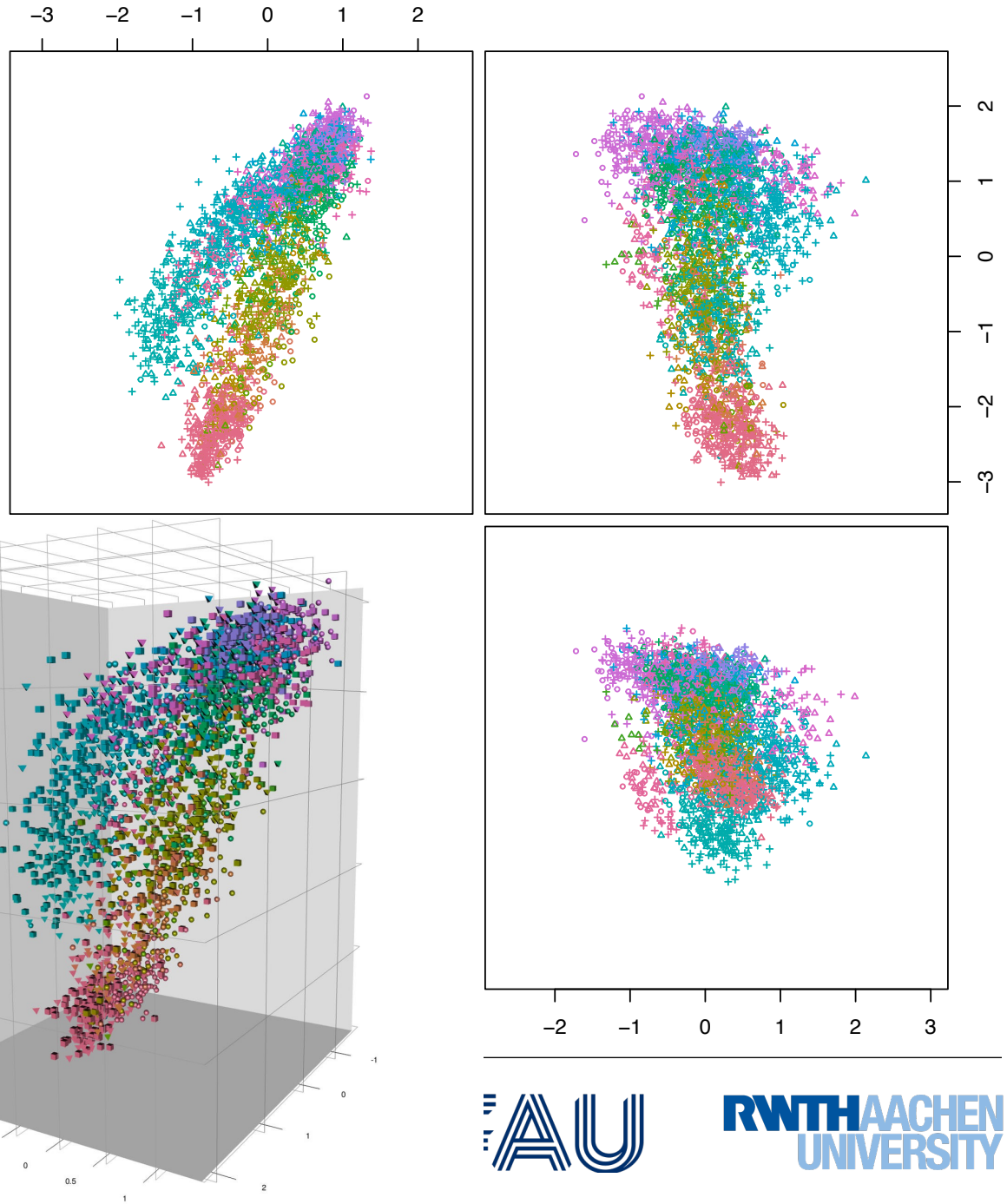



```
ICE3 <- GMA$new(ZL3)
GMA$add.discriminant(Meta3$textcat20, max.dim=4)
```


Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Reproducing Neumann & Evert (2021)

- Linear discriminant analysis of ICE3 data set based on 20 mid-level text categories
  - →register space implied by ICE sampling frame
  - →clearer structure & visual interpretation than PCA
- Projection into focus space spanned by first 4 LDA dimensions (here: visualise only 3 dim's)
- Captures R^2 = 17.87% of distance information
- Scatterplot matrix visualisation shows views from the front, left and top of the cube
  - →but generalises to *k* > 3 dimensions



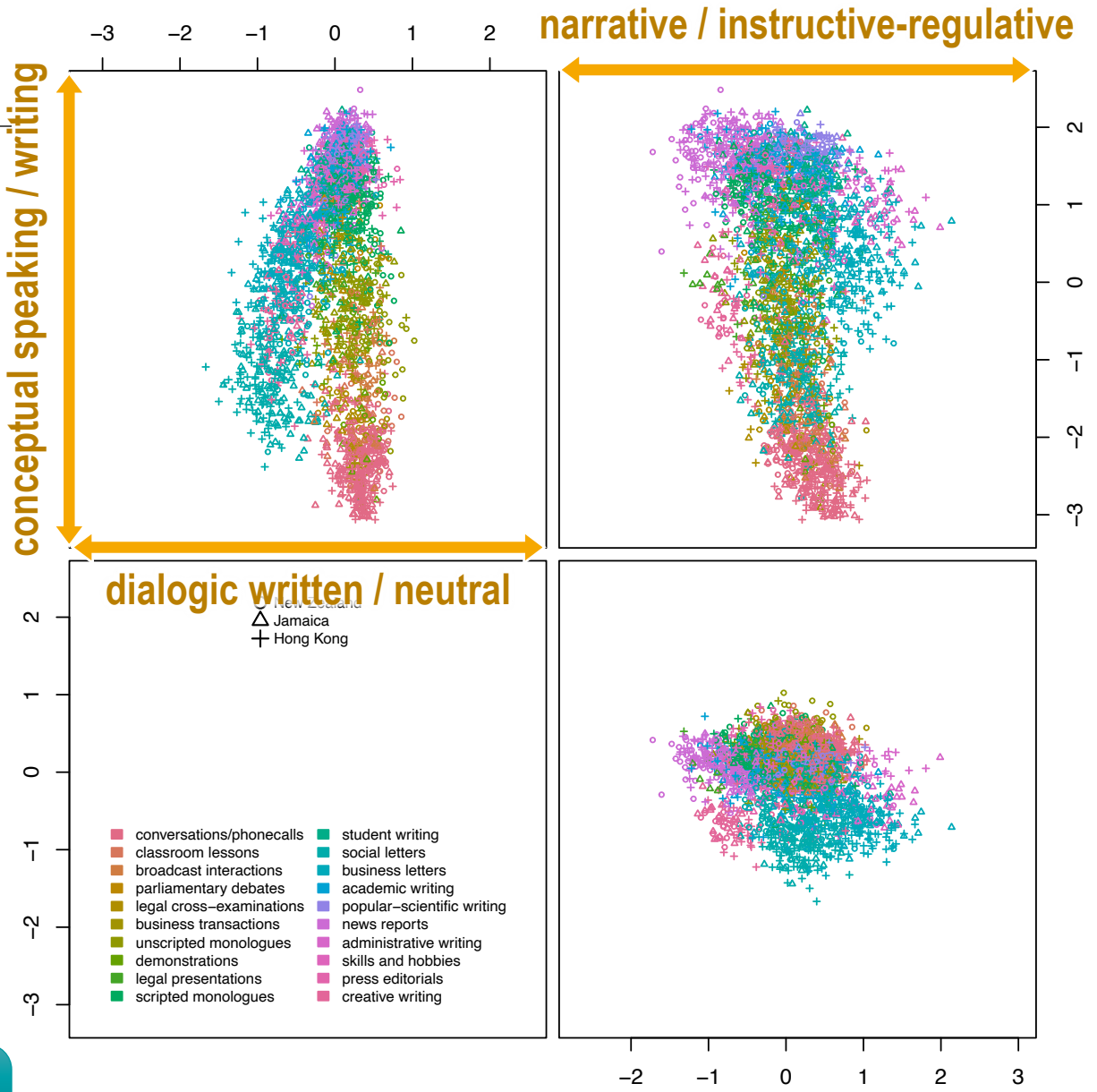

```
ICE3.X <- ICE3$projection()
gma.pairs(ICE3.X, dim=1:3, meta=Meta3,
          col=textcat20, pch=variety)
```


Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Reproducing Neumann & Evert (2021)

- Projection shows very clear structure, but the two “bananas” are not aligned with dimensions
- FA applies “rotations” in such cases
- GMA: actual rotation (isometric map) of first two basis vectors based on PCA of data set

  →similar to varimax rotation in factor analysis



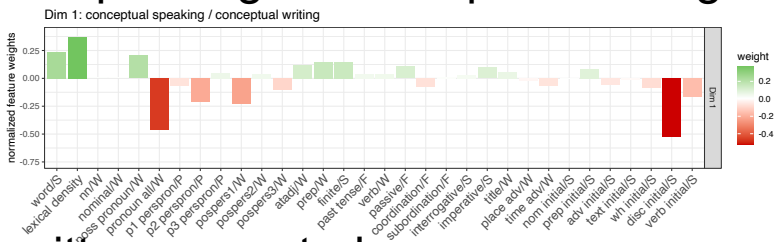

```
ICE3$rotation("pca", dim=1:2)
```


Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

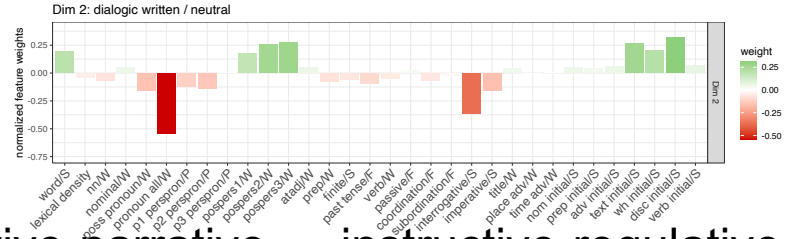
Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Reproducing Neumann & Evert (2021)

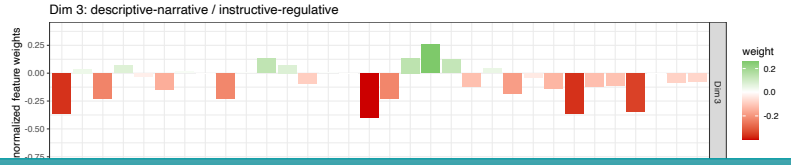
- Interpretation of dimensions based on topographic map created by ICE text categories, combined with feature weights of basis vectors (“loadings”)
- LD1: conceptual speaking — conceptual writing



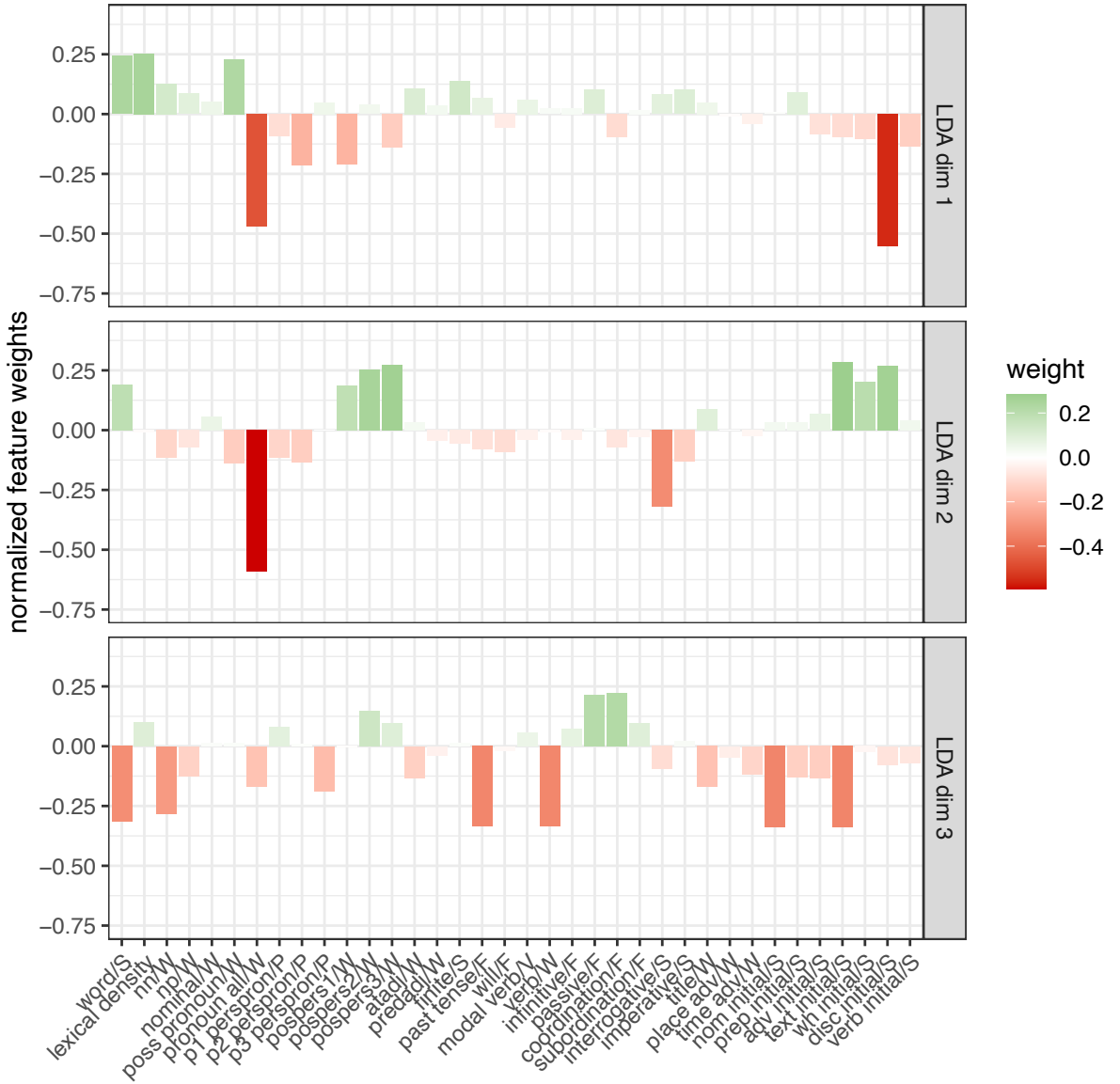
- LD2: dialogic written — neutral



- LD3: descriptive-narrative — instructive-regulative




```
gma.plot.weights(ICE3$basis(), dim=1:3)
```



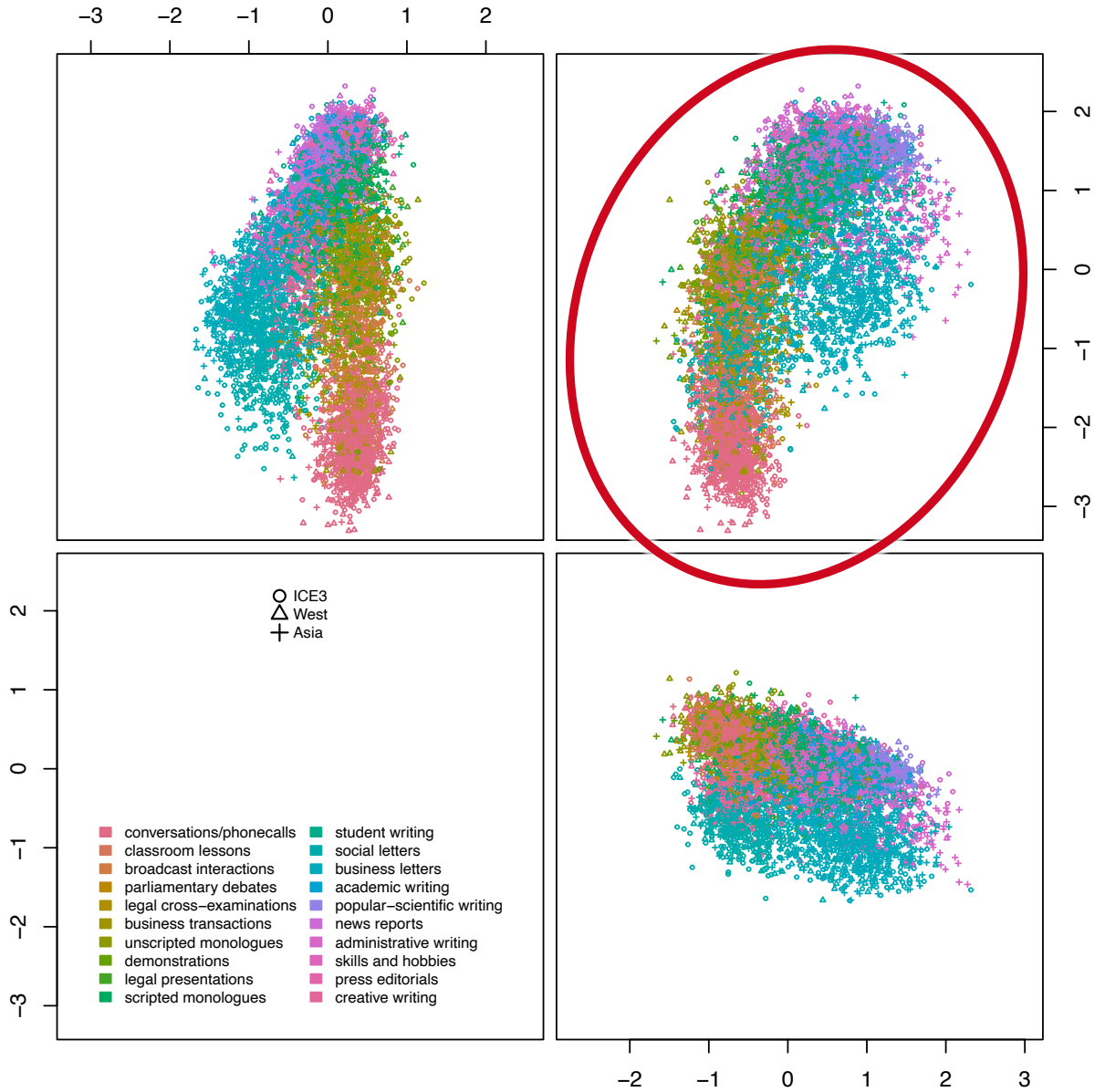
Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Robustness: Extended data set

- Replication by Frenken et al. (2025):
  Will we come to the same conclusions if we carry out the analysis on a different data set?
- Here: extend to nine varieties of English (ICE9)

- First two LDA dimensions look very similar at first sight, but third (and fourth) are markedly different!
- Feature weights also lead to different interpretation of the dimensions

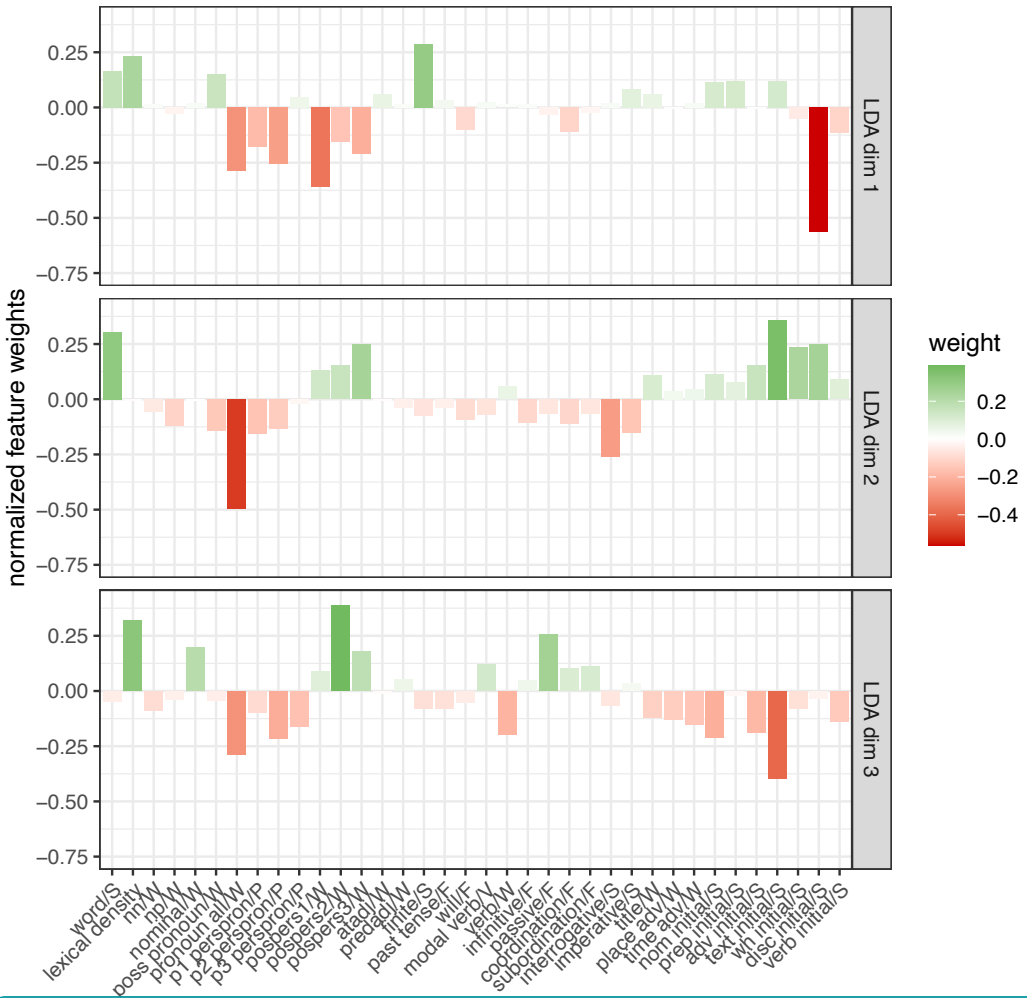


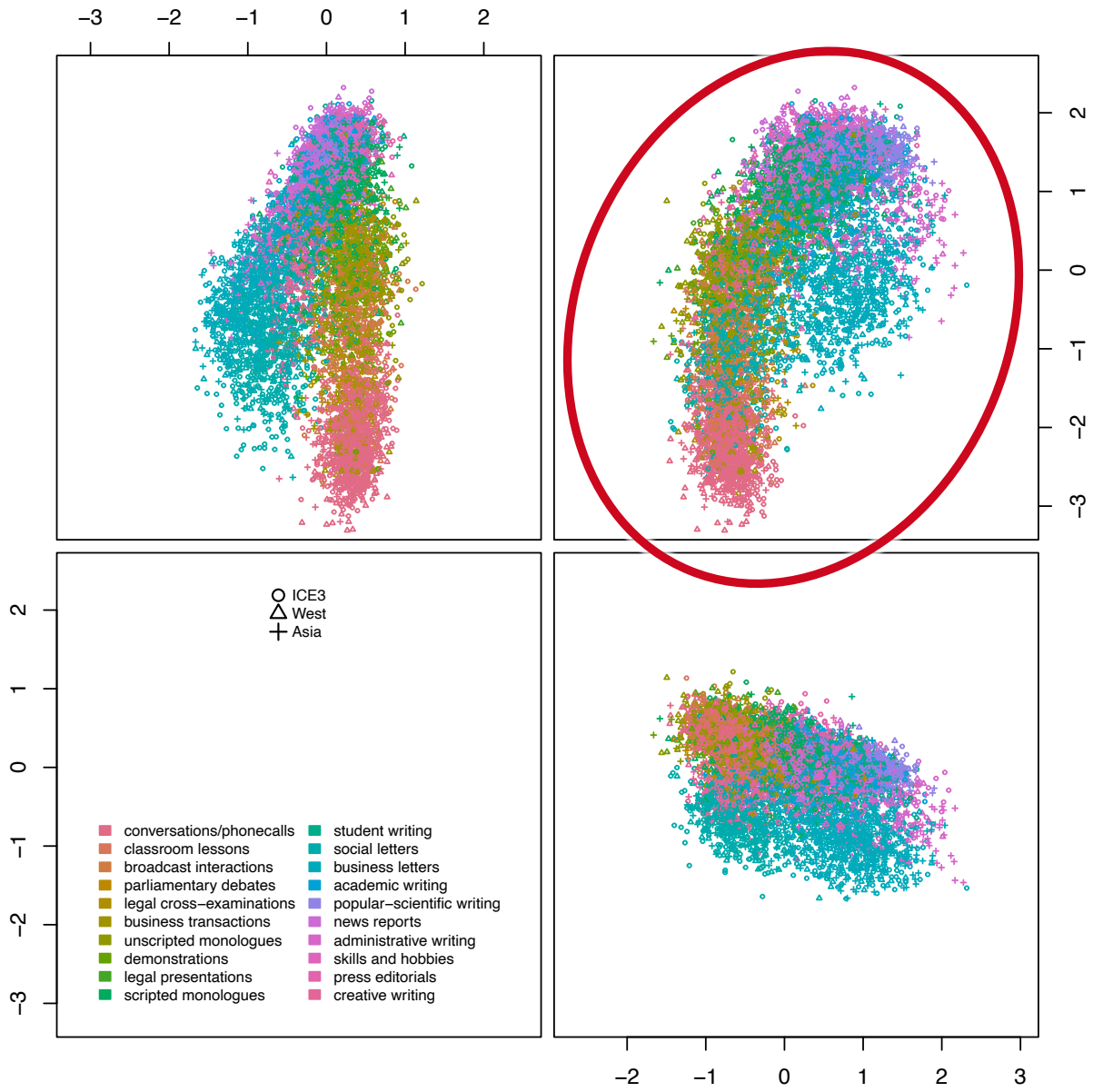

```
ICE9 <- GMA$new(ZL)
GMA$add.discriminant(Meta$textcat20, max.dim=4)
```


Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Robustness: Extended data set






```
ICE9$similarity(ICE3) # → 3.714 (out of 4)
```

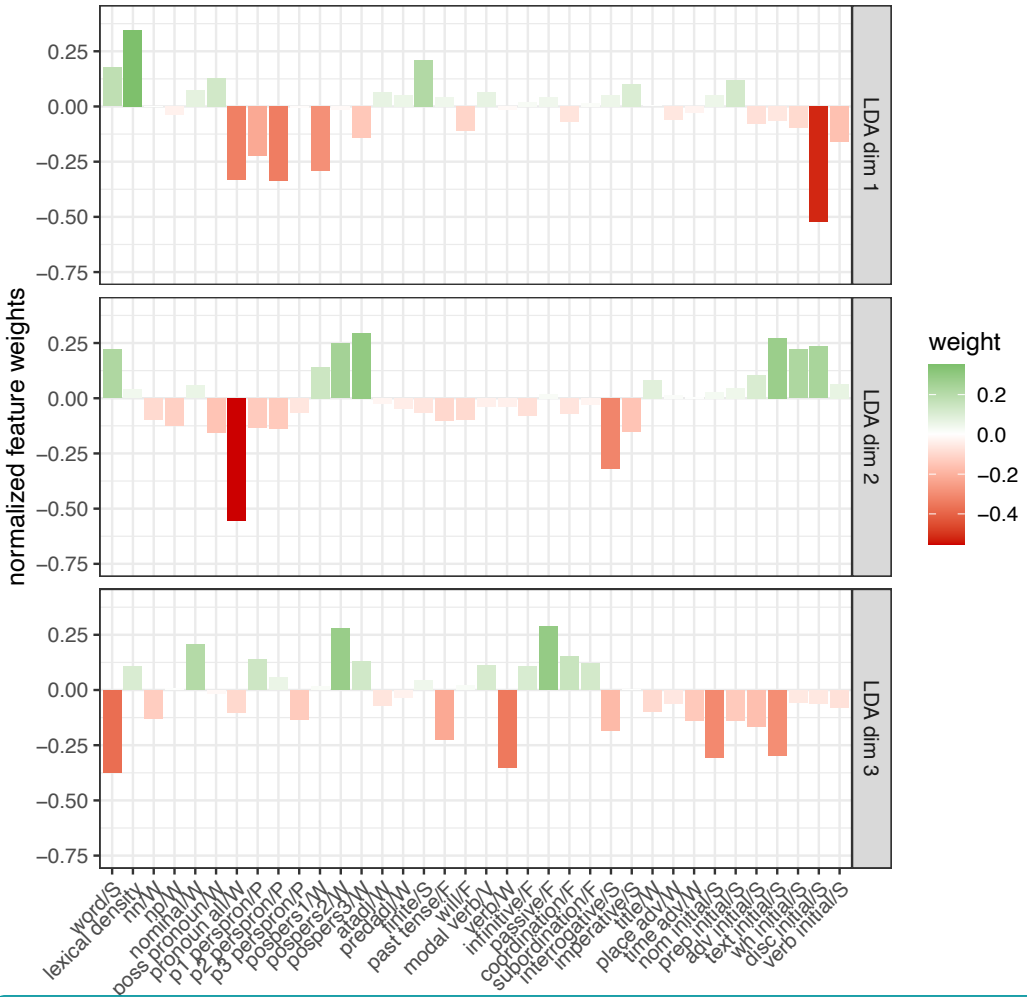

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

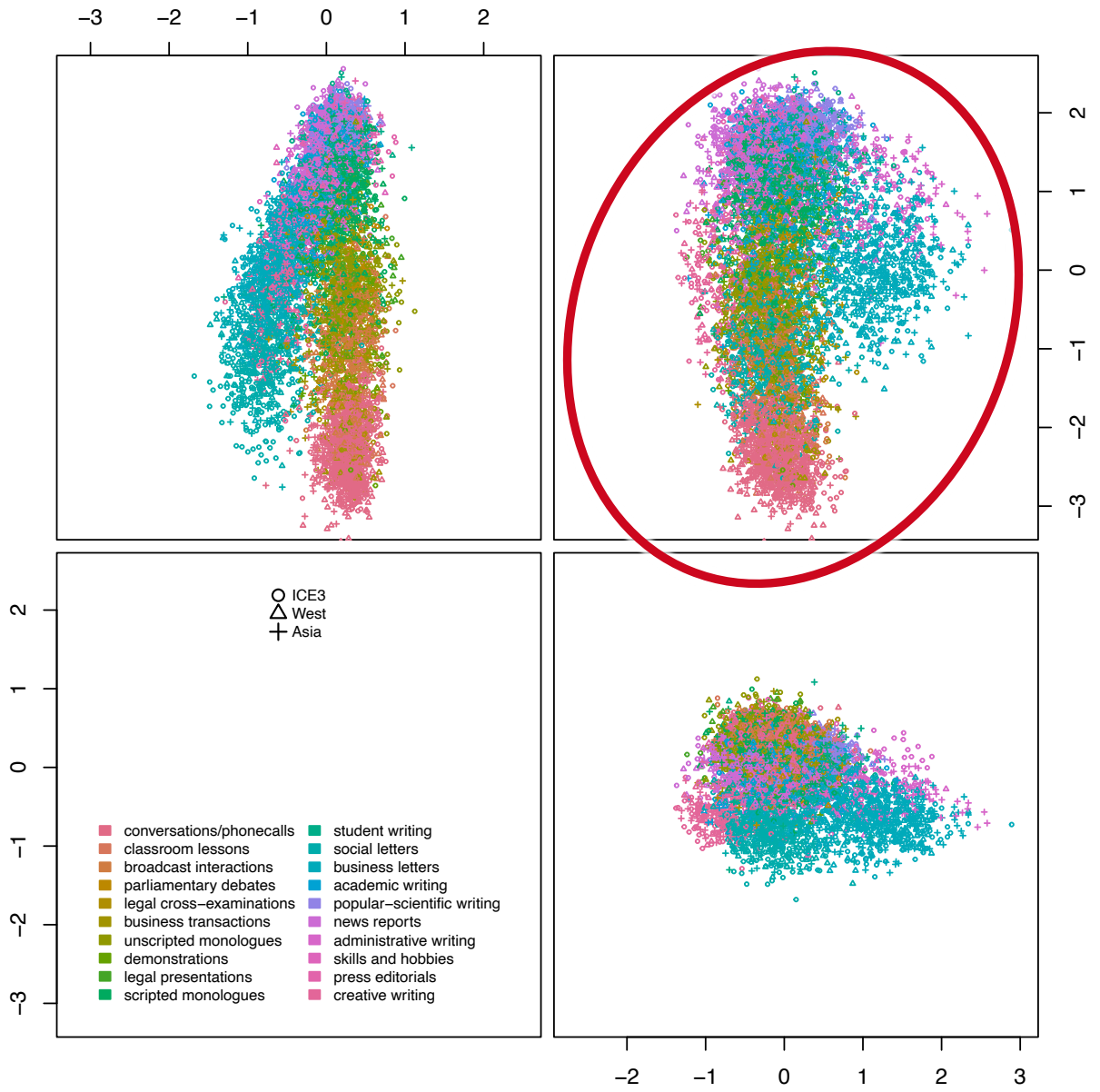
Universität Zürich[UZH]

FAU

RWTH AACHEN UNIVERSITY

Aligning the GMA focus spaces





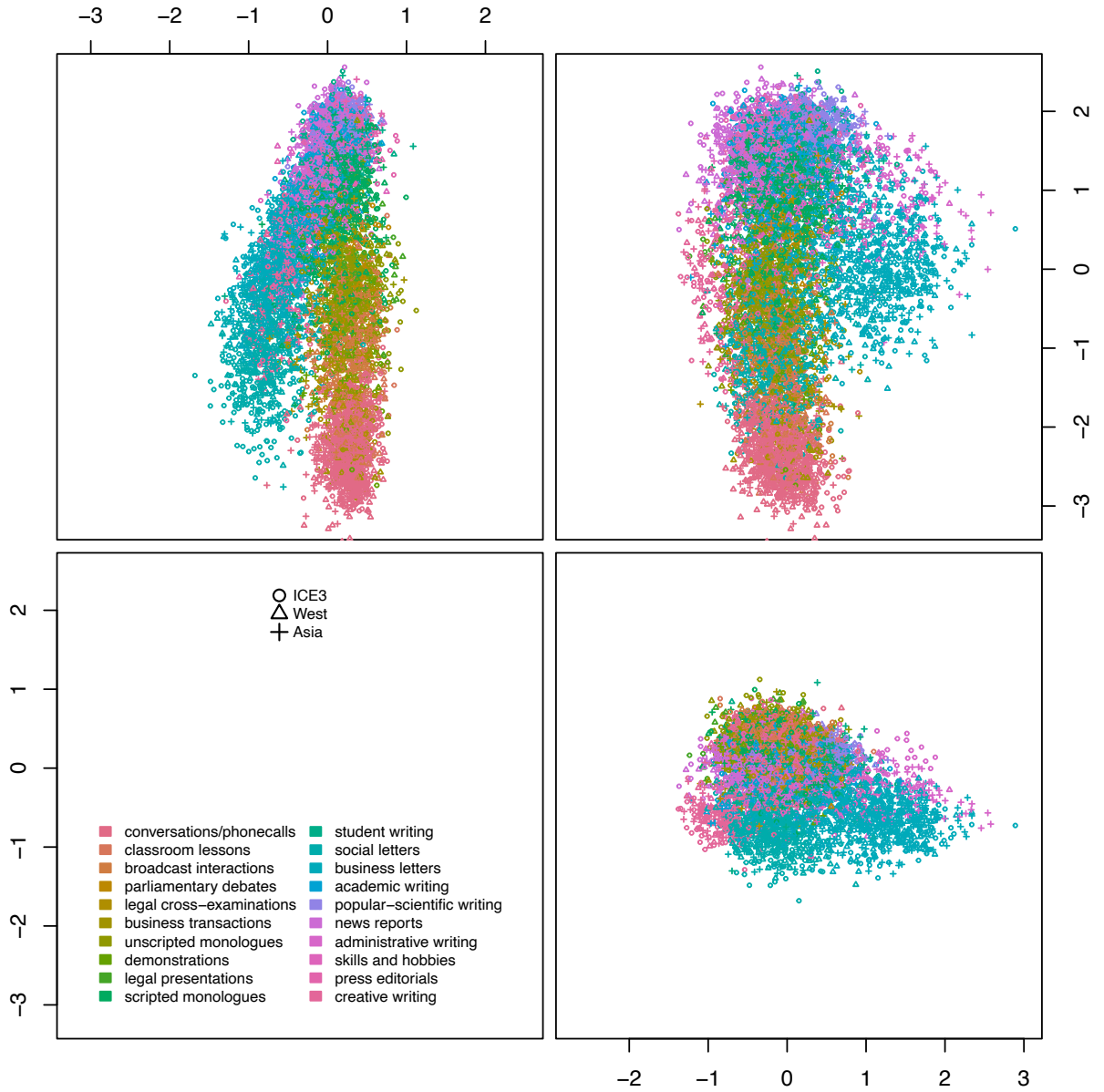

```
ICE9$rotation("manual", basis=ICE3, debug=TRUE)
```


Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Zooming in on language varieties

- Let us “zoom in” on five text categories in the conceptual speaking range of LD1
  - →spoken: conversations/phonecalls (1028), broadcast interactions (331), legal cross-examination (114)
  - →written: social letters (566), creative writing (213)
- Compare language varieties with respect to these text categories

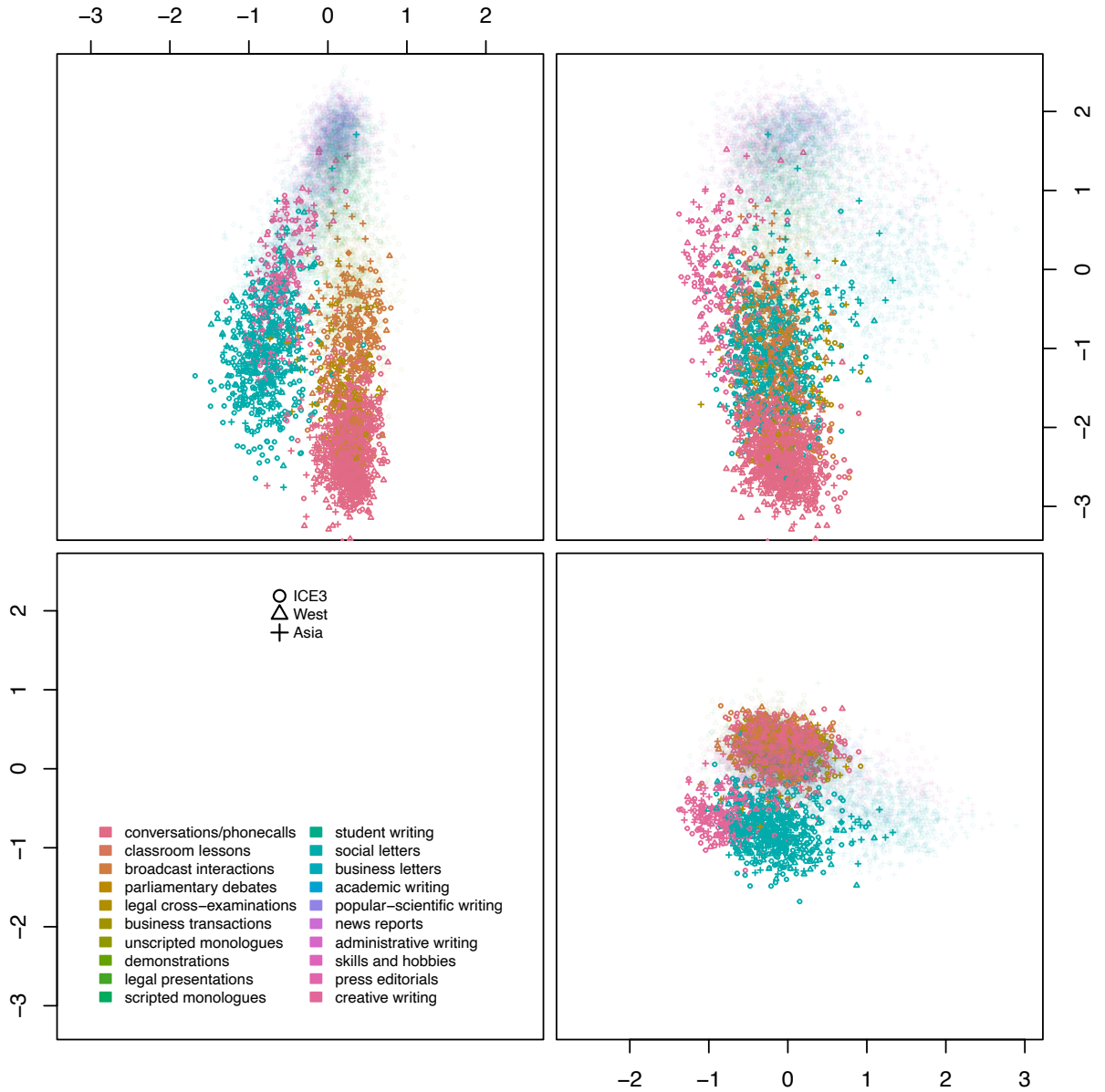


Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Zooming in on language varieties

- Let us “zoom in” on five text categories in the conceptual speaking range of LD1
  - →spoken: conversations/phonecalls (1028), broadcast interactions (331), legal cross-examination (114)
  - →written: social letters (566), creative writing (213)
- Compare language varieties with respect to these text categories



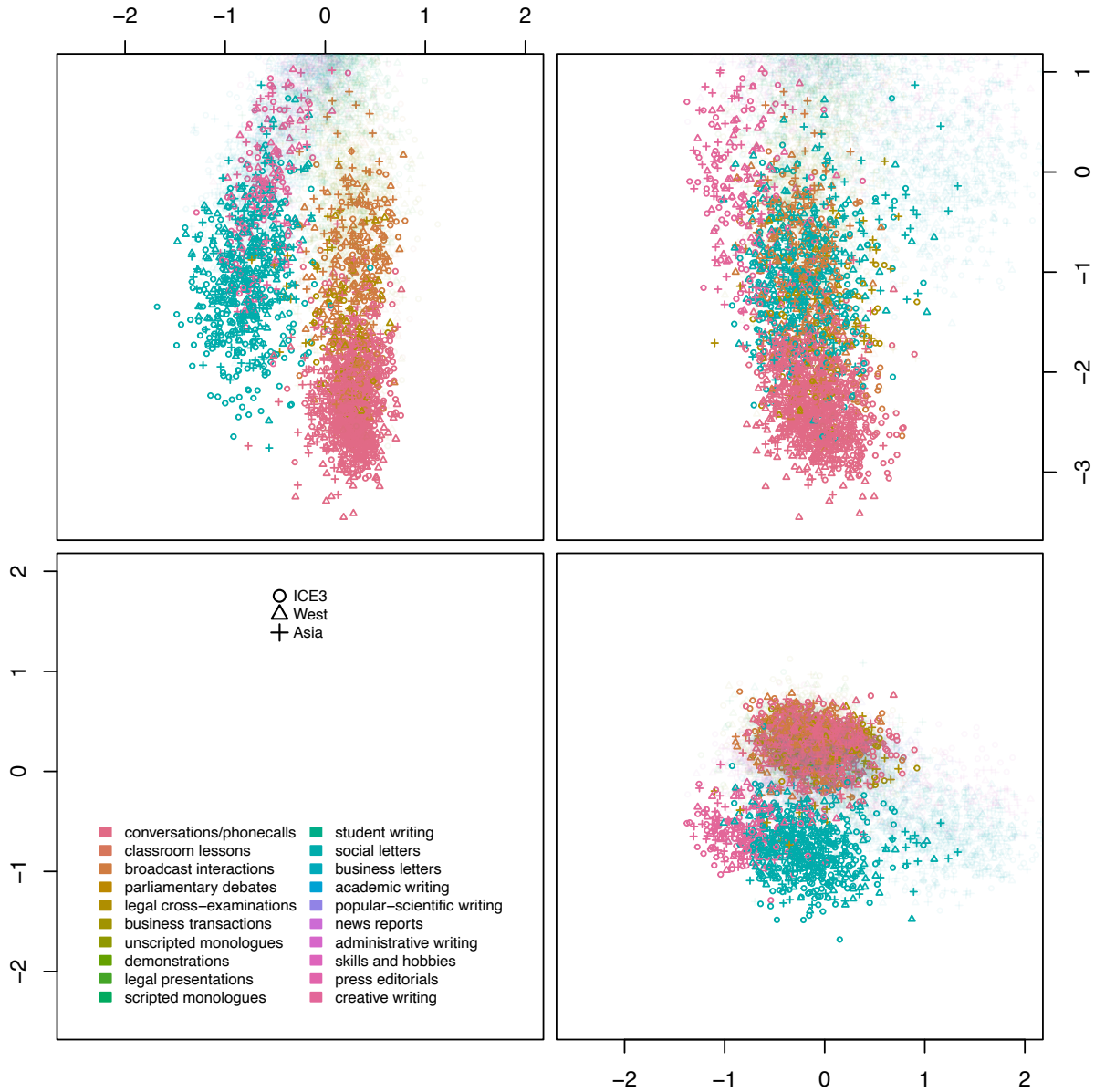

```
idx.sub <- Meta$short20 %in% qw("conv broadc …")
MetaSub <- droplevels(Meta[idx.sub, ])
ICE9.Sub <- ICE9.X[idx.sub, ]
```


Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Zooming in on language varieties

- Let us “zoom in” on five text categories in the conceptual speaking range of LD1
  - →spoken: conversations/phonecalls (1028), broadcast interactions (331), legal cross-examination (114)
  - →written: social letters (566), creative writing (213)
- Compare language varieties with respect to these text categories




```
idx.sub <- Meta$short20 %in% qw("conv broadc …")
MetaSub <- droplevels(Meta[idx.sub, ])
ICE9.Sub <- ICE9.X[idx.sub, ]
```

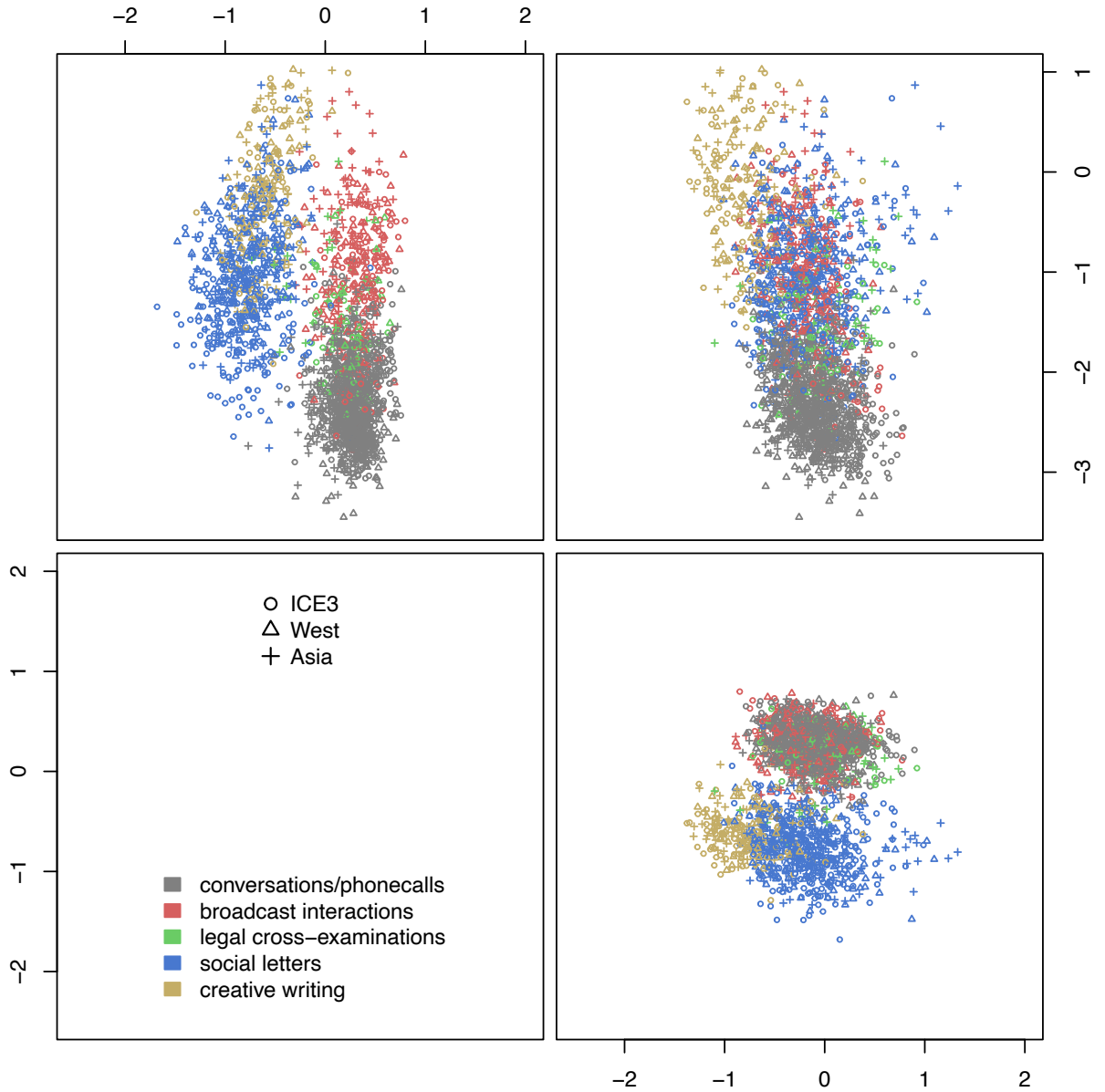

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Zooming in on language varieties

- Let us “zoom in” on five text categories in the conceptual speaking range of LD1
  - →spoken: conversations/phonecalls (1028), broadcast interactions (331), legal cross-examination (114)
  - →written: social letters (566), creative writing (213)
- Compare language varieties with respect to these text categories
  - →New colour scheme helps to distinguish text categories more clearly

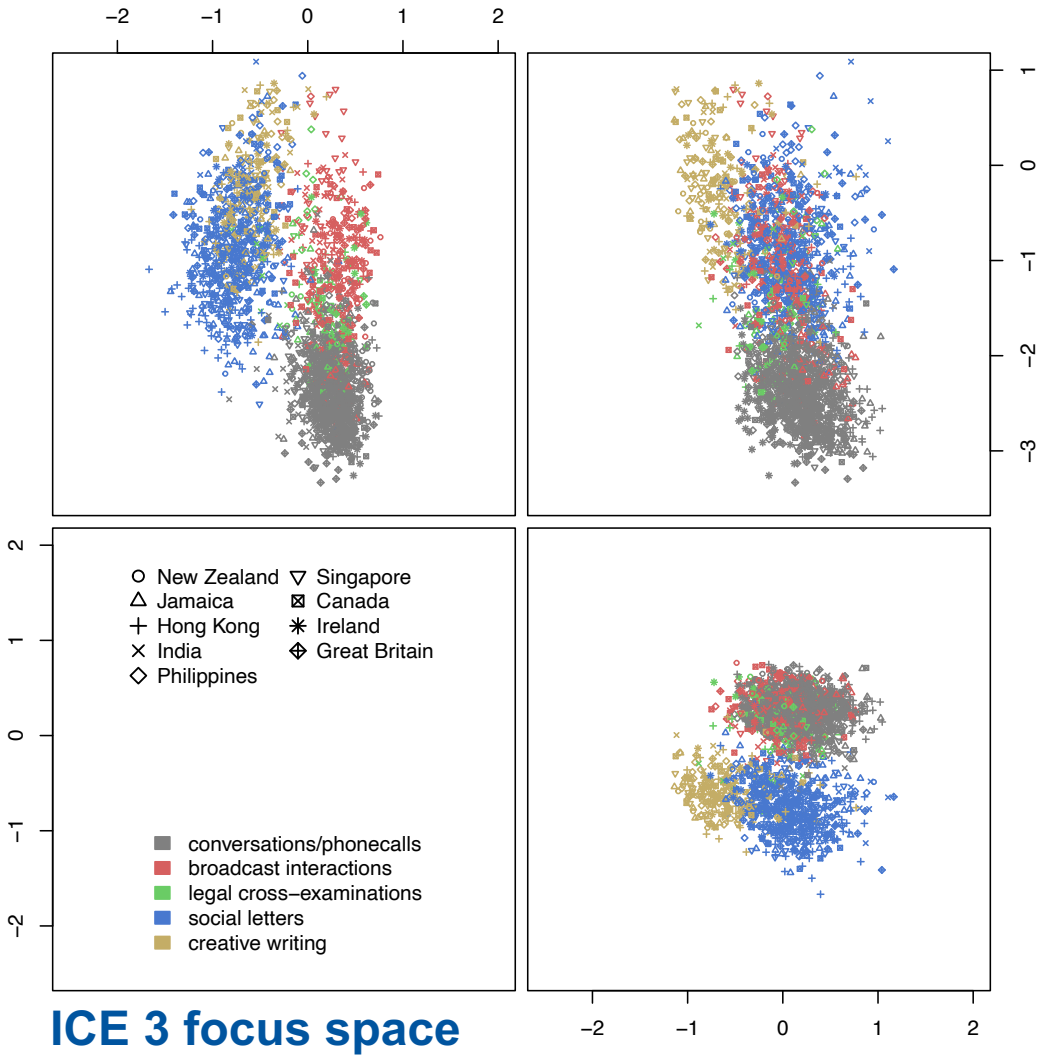
```
idx.sub <- Meta$short20 %in% qw("conv broadc …")
MetaSub <- droplevels(Meta[idx.sub, ])
ICE9.Sub <- ICE9.X[idx.sub, ]
```



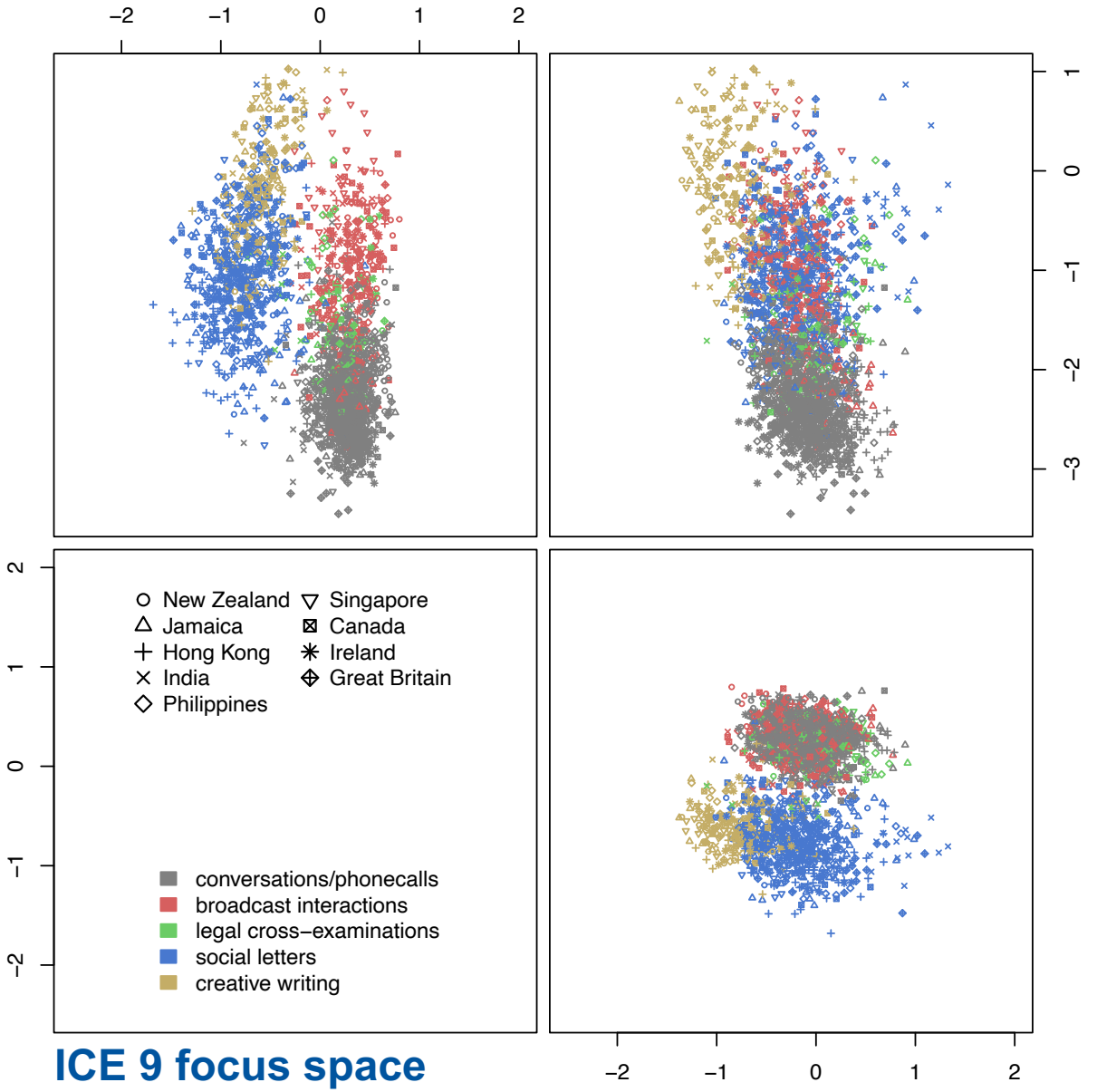
Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Register divergence across varieties



ICE 3 focus space



ICE 9 focus space

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

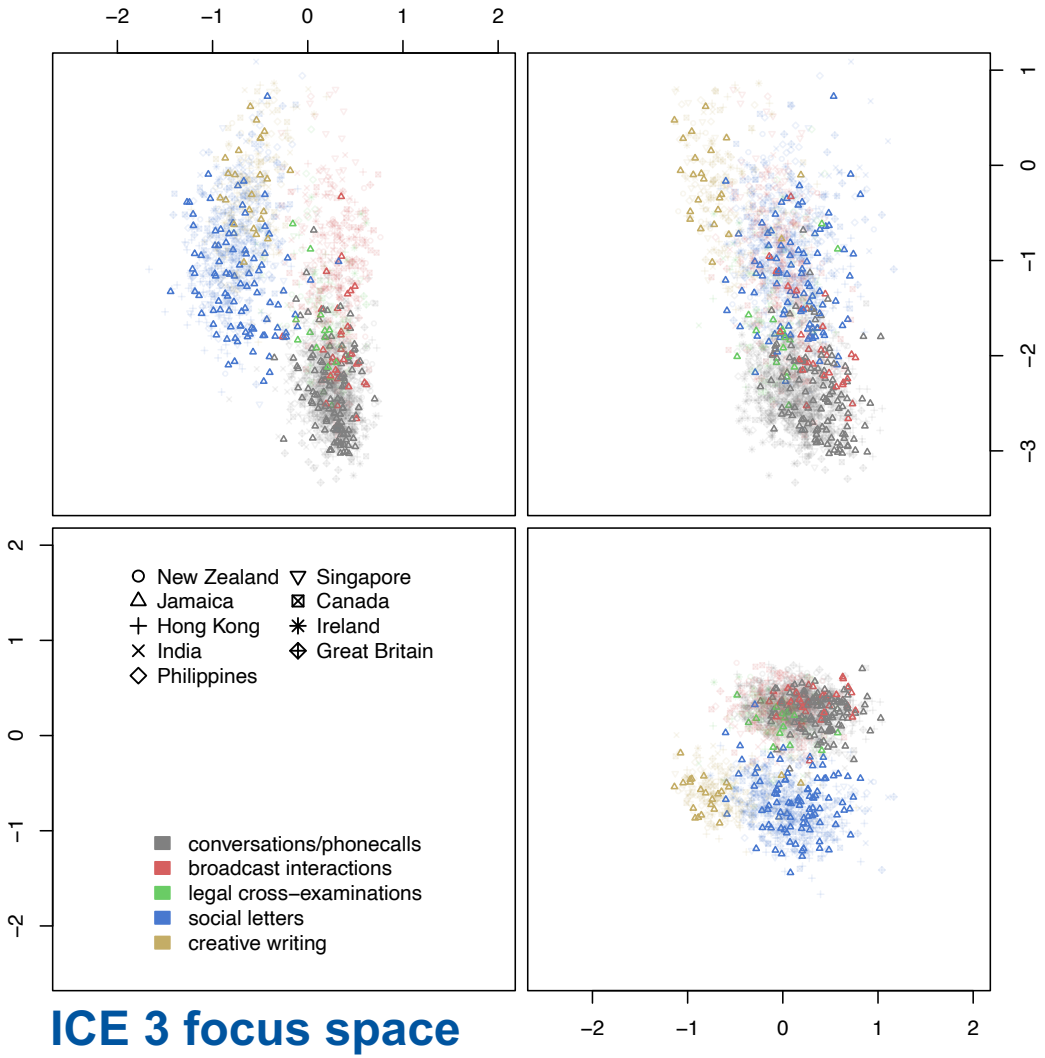
Universität Zürich[UZH]

FAU

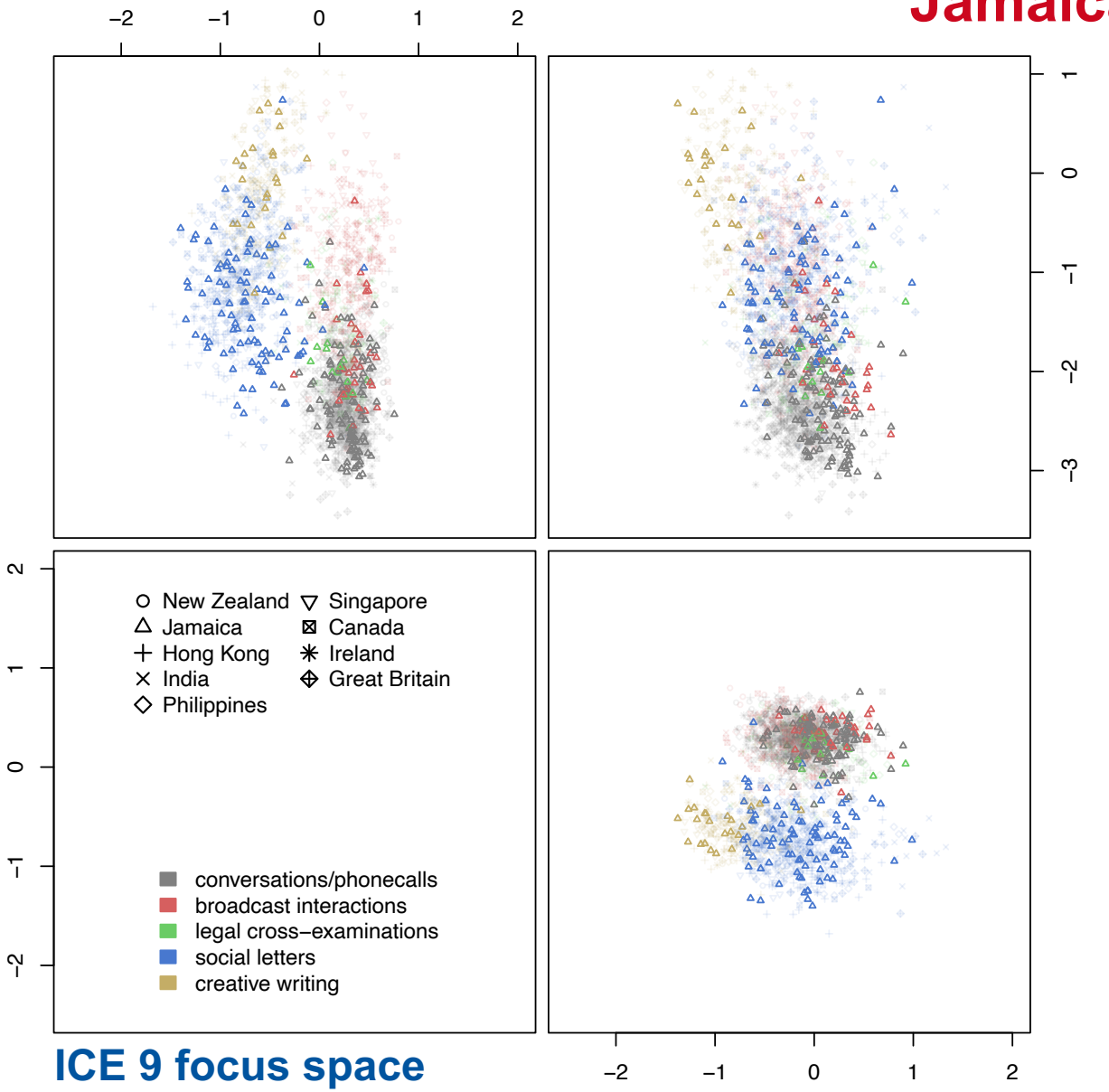
RWTH AACHEN UNIVERSITY

Register divergence across varieties

Jamaica



ICE 3 focus space



ICE 9 focus space

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

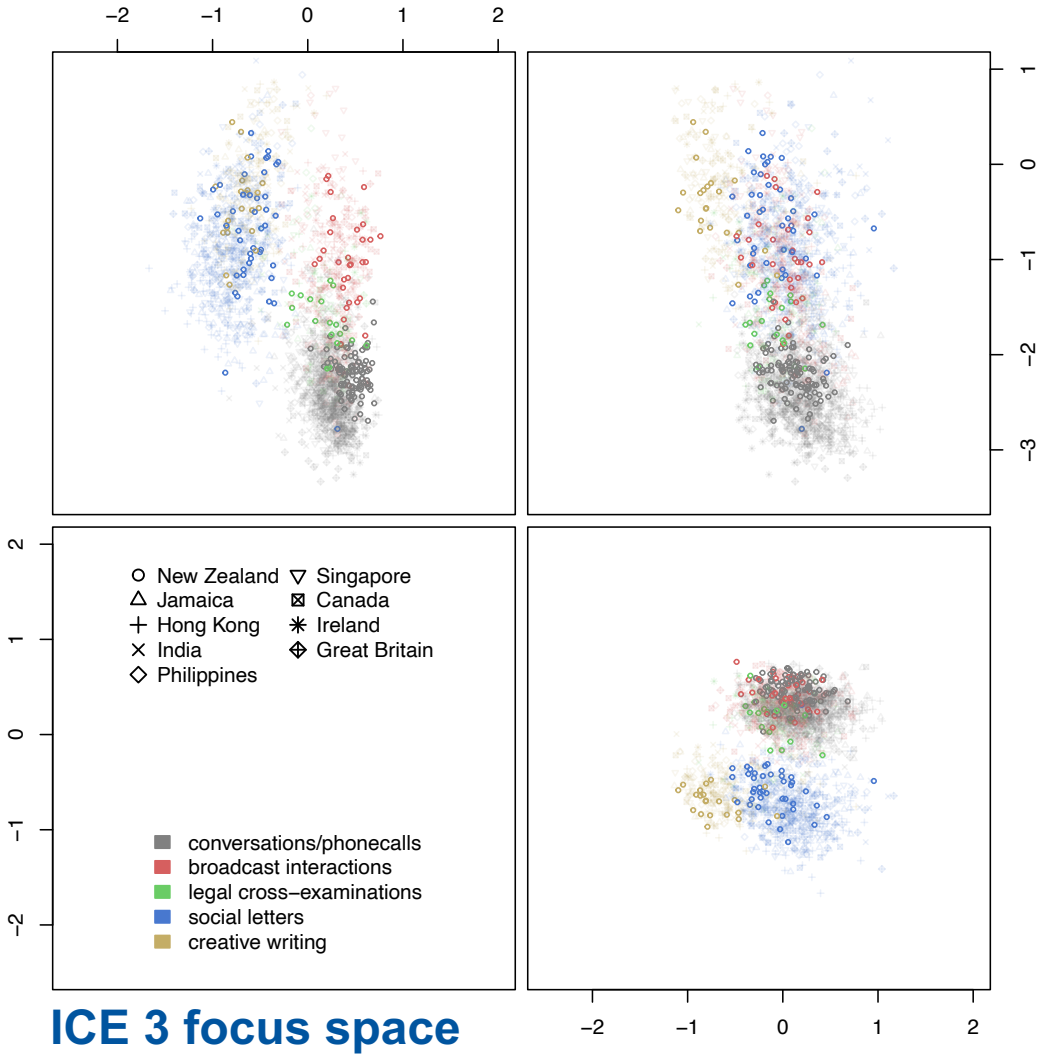
Universität Zürich[UZH]

FAU

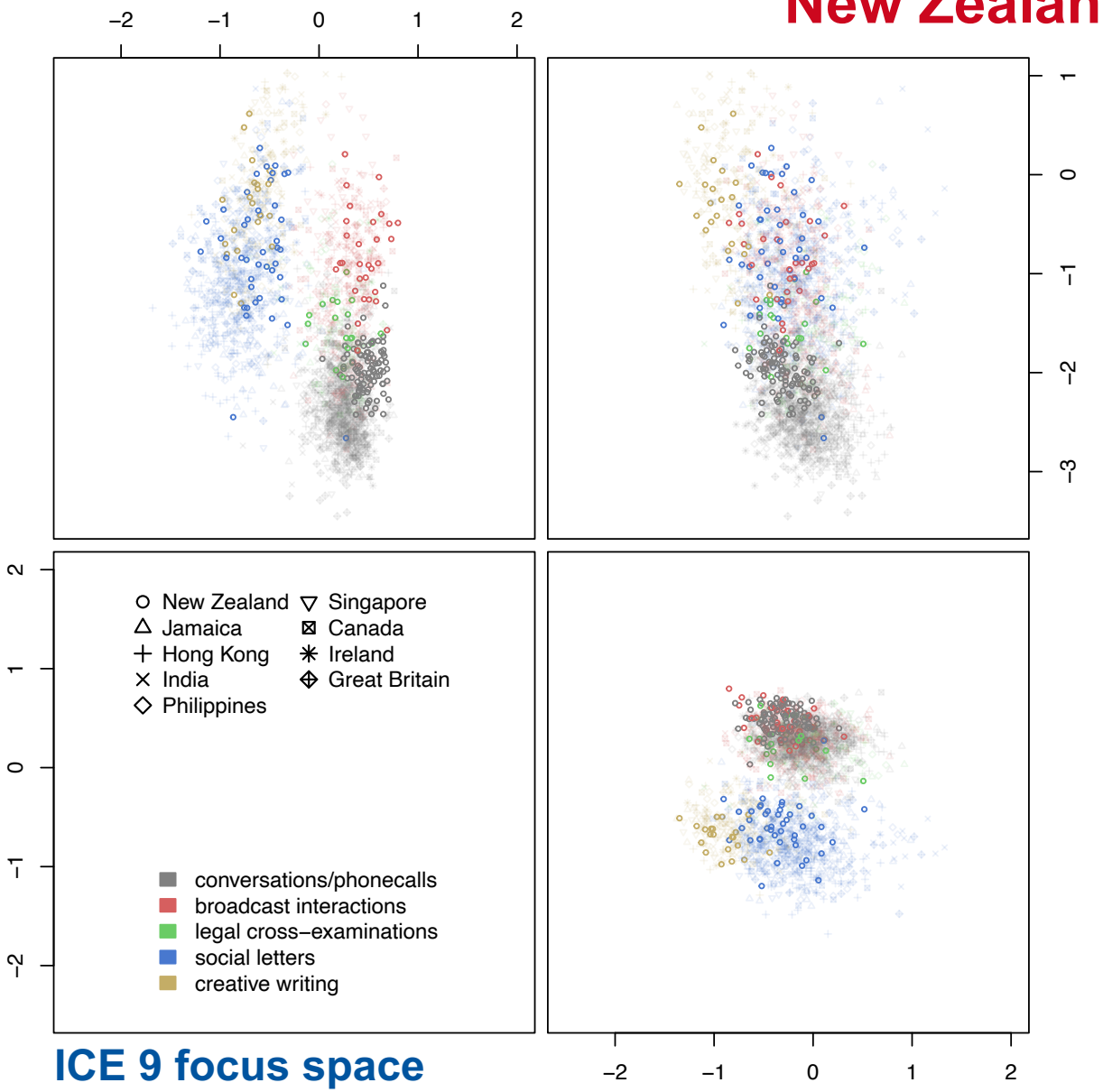
RWTH AACHEN UNIVERSITY

Register divergence across varieties

New Zealand



ICE 3 focus space



ICE 9 focus space

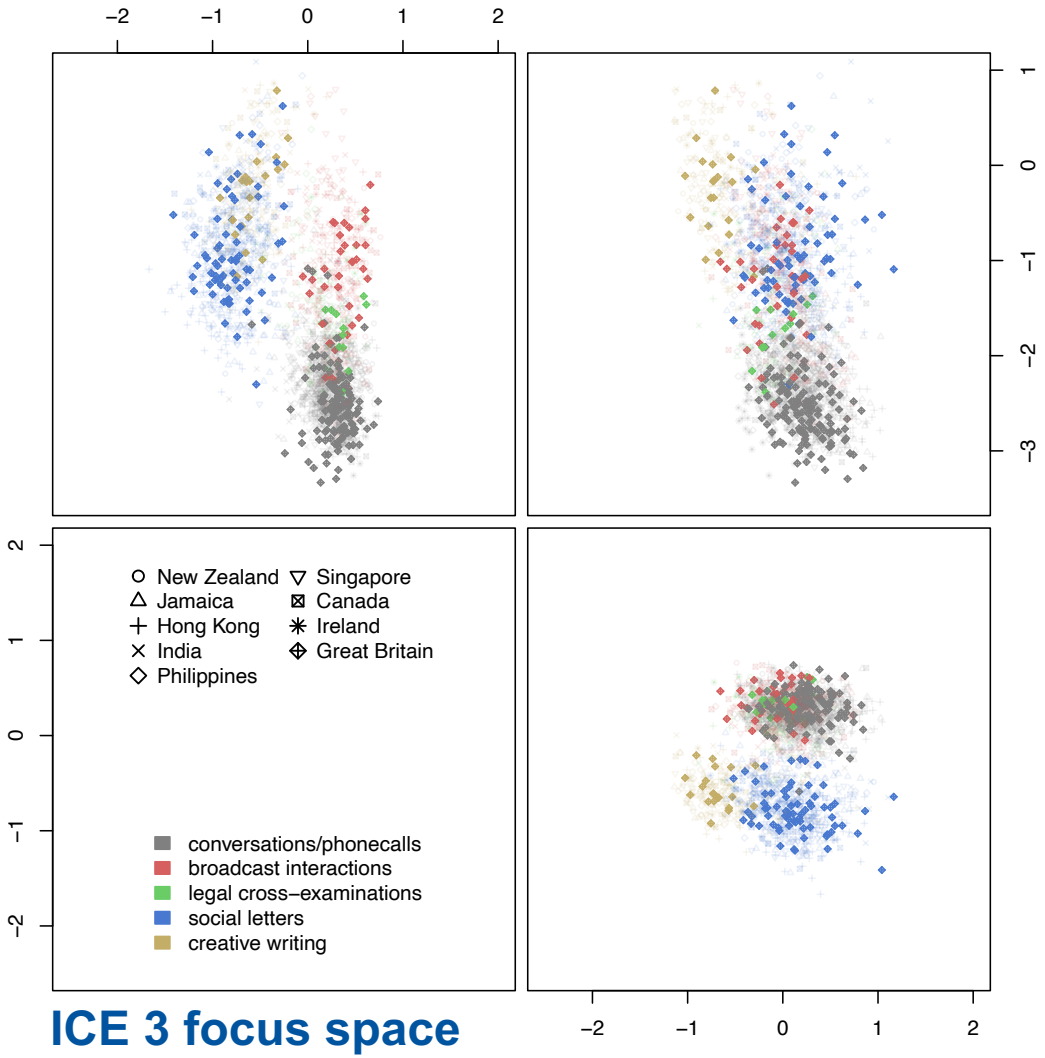
Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

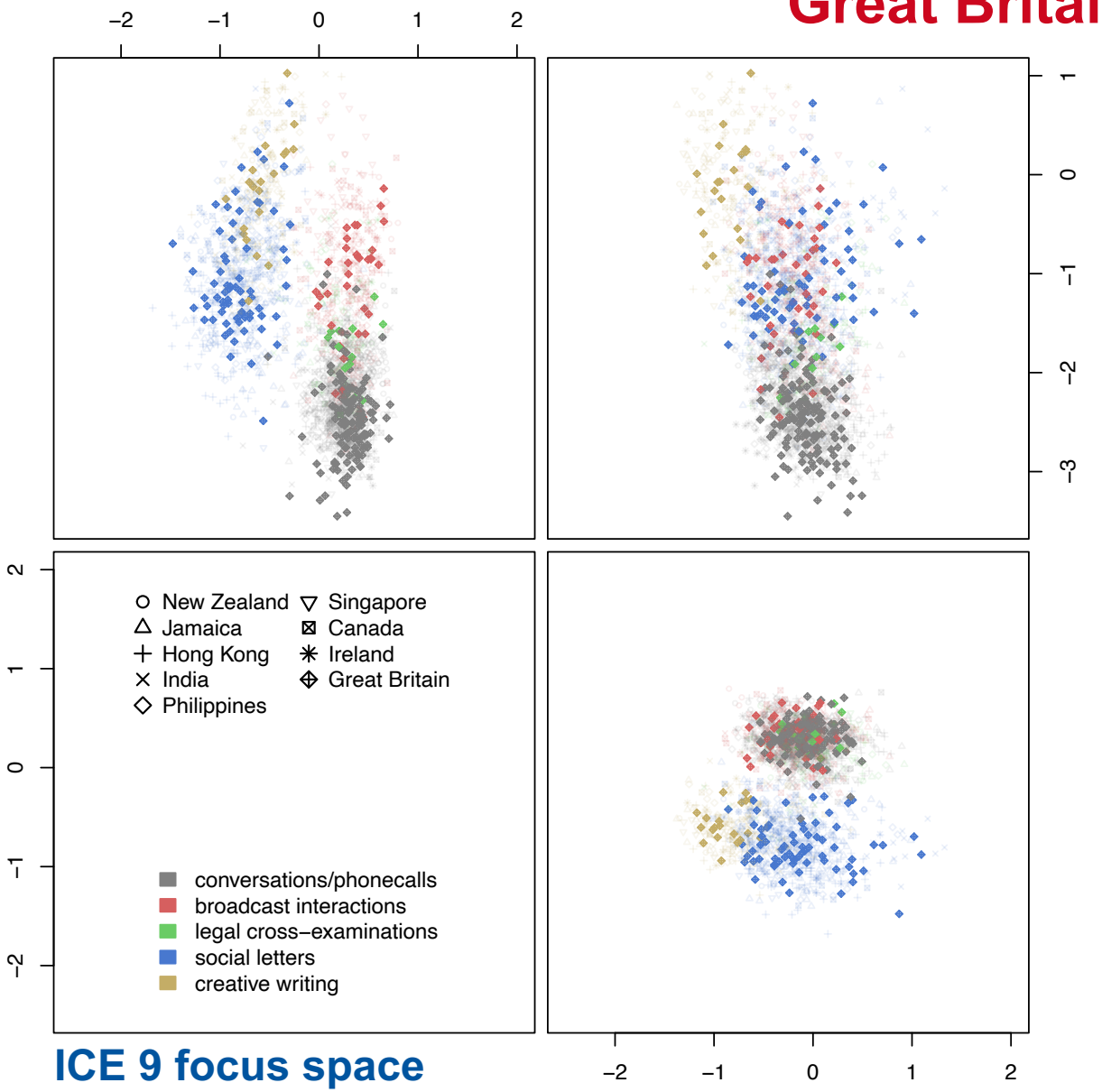
RWTH AACHEN UNIVERSITY

Register divergence across varieties



ICE 3 focus space

Great Britain



ICE 9 focus space

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

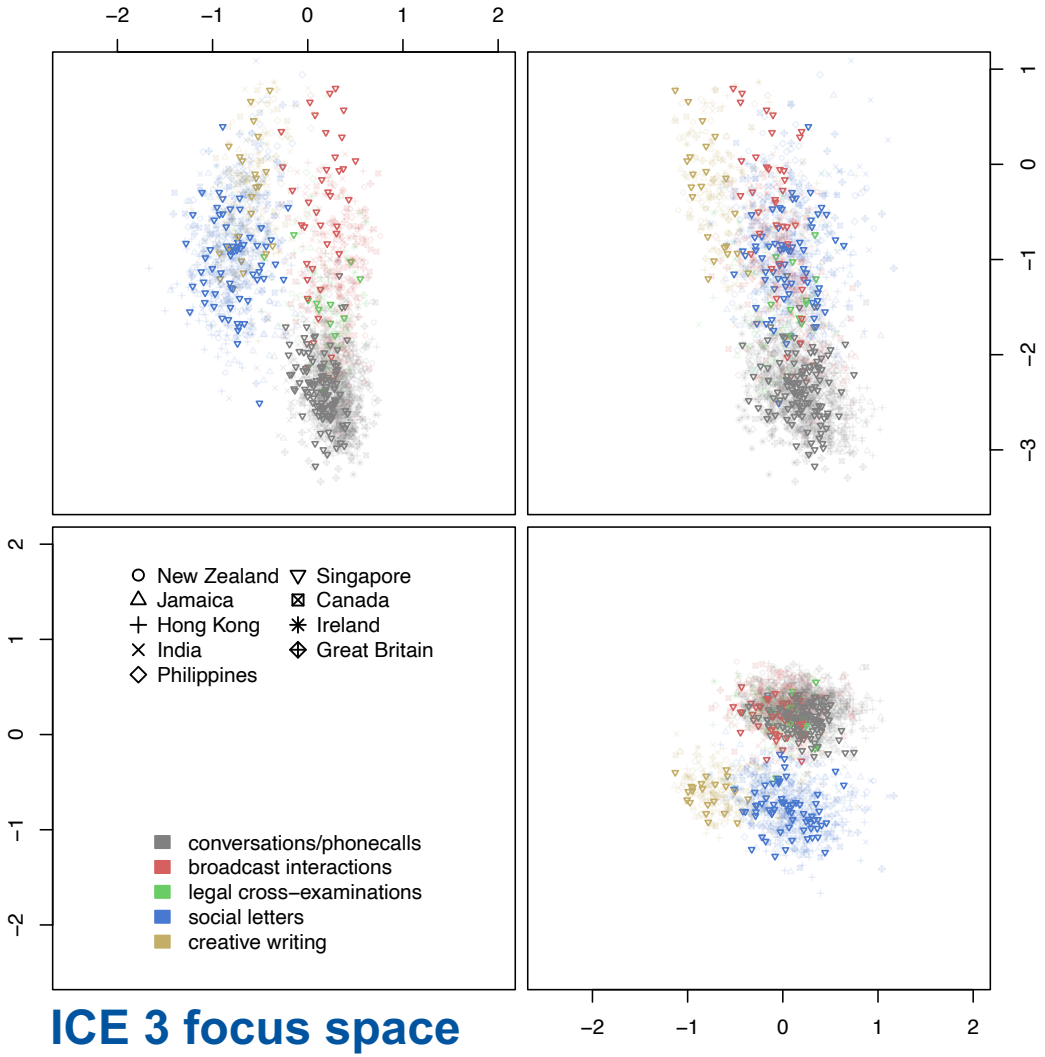
Universität Zürich[UZH]

FAU

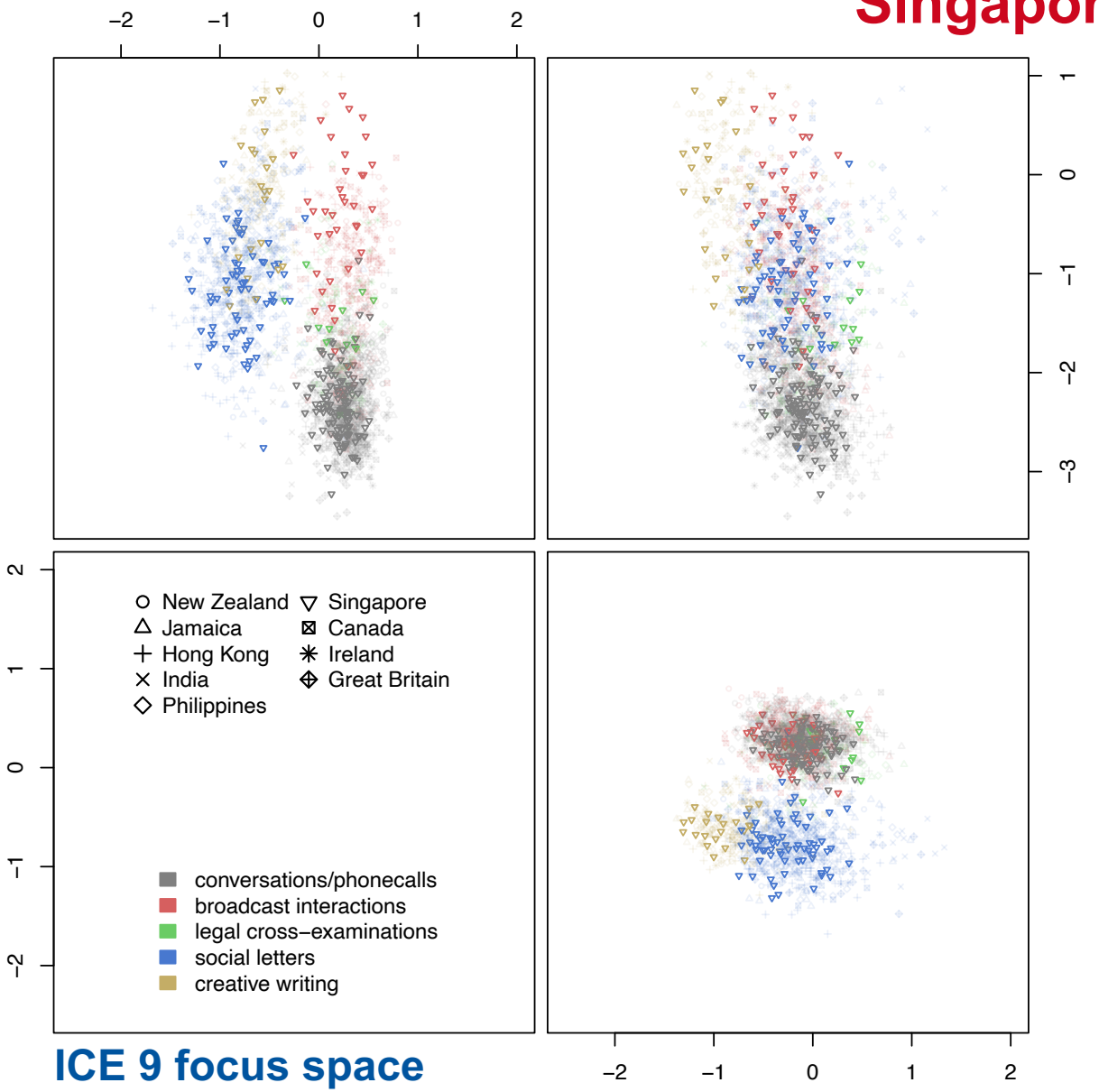
RWTH AACHEN UNIVERSITY

Register divergence across varieties

Singapore



ICE 3 focus space



ICE 9 focus space

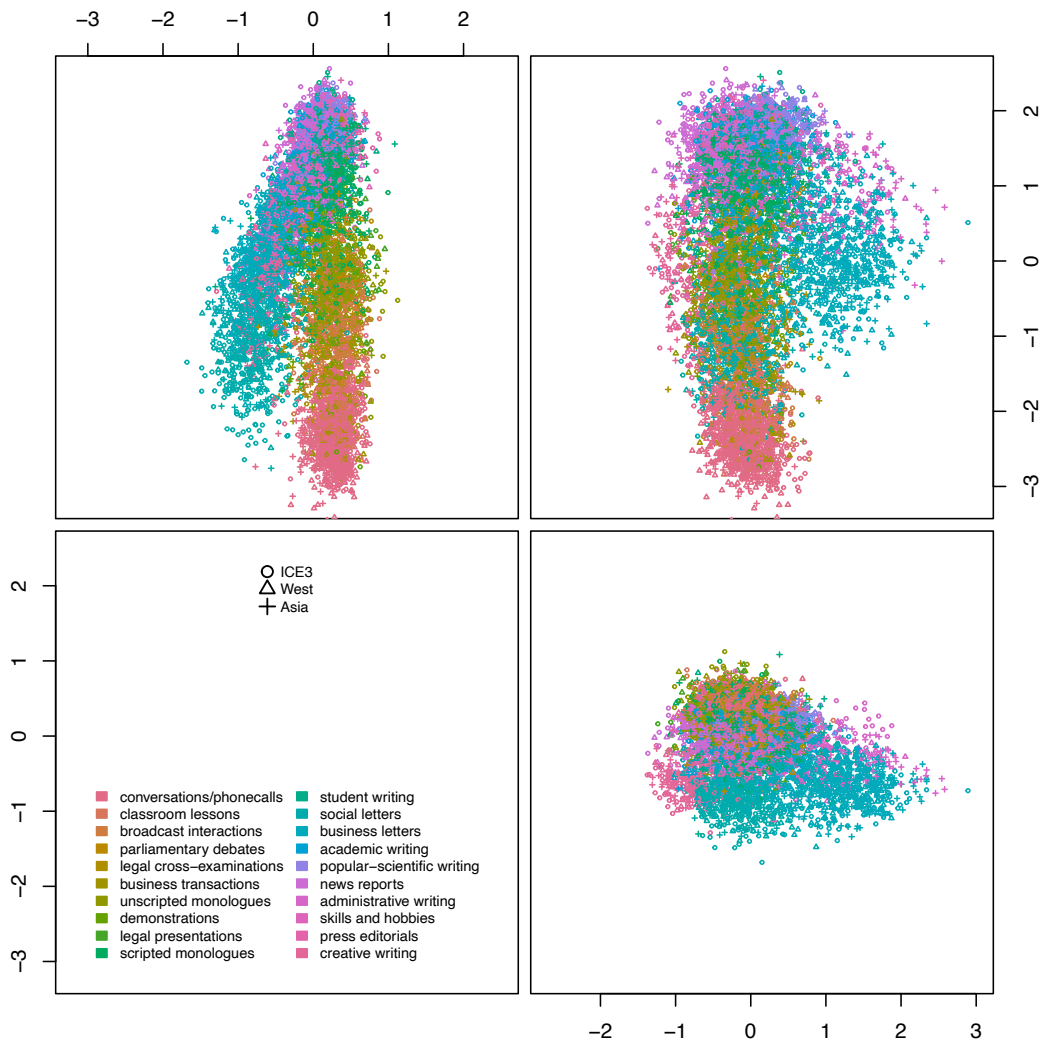
Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

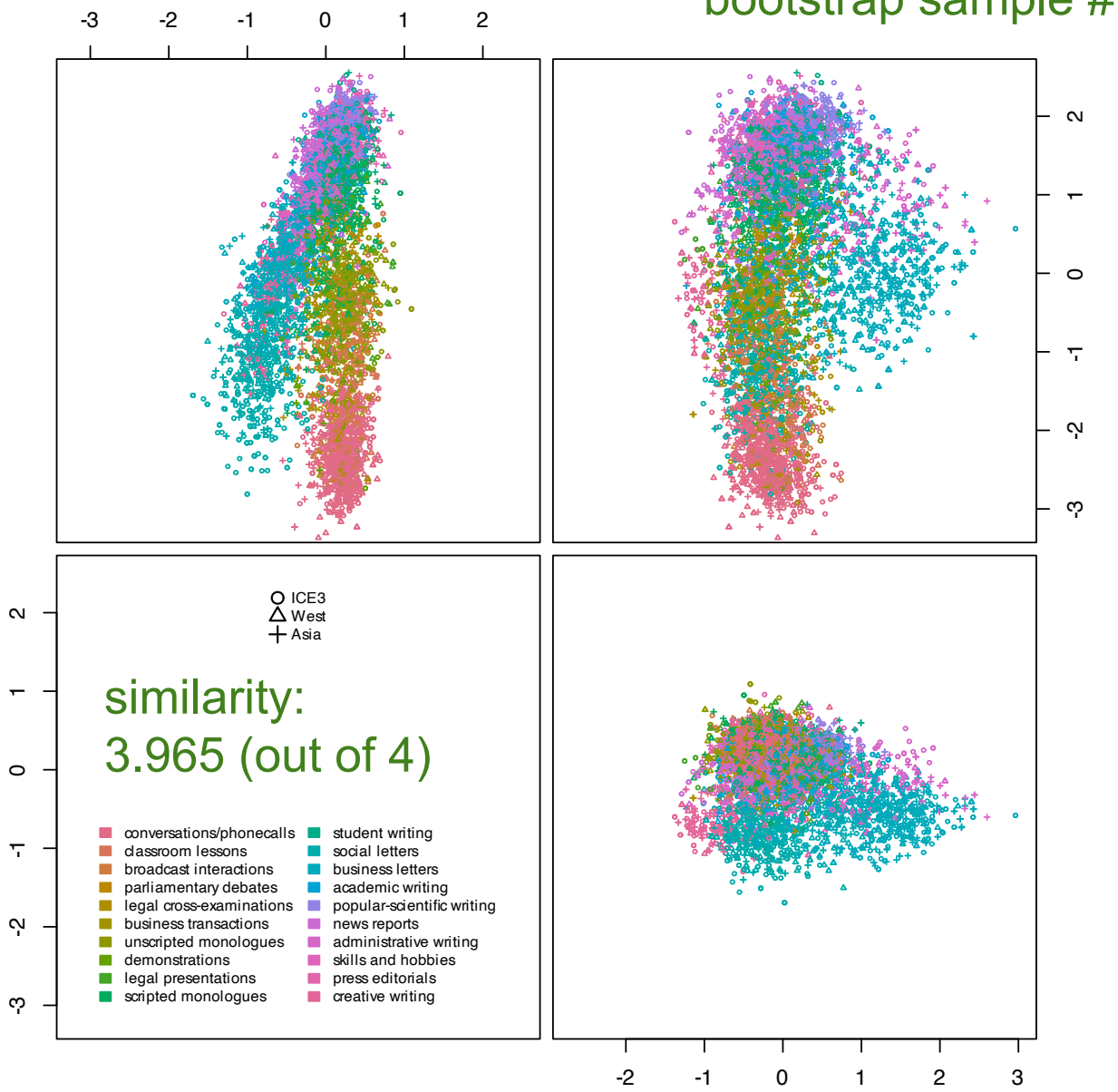
FAU

RWTH AACHEN UNIVERSITY

Stability: Bootstrapping (yay!)



bootstrap sample #1



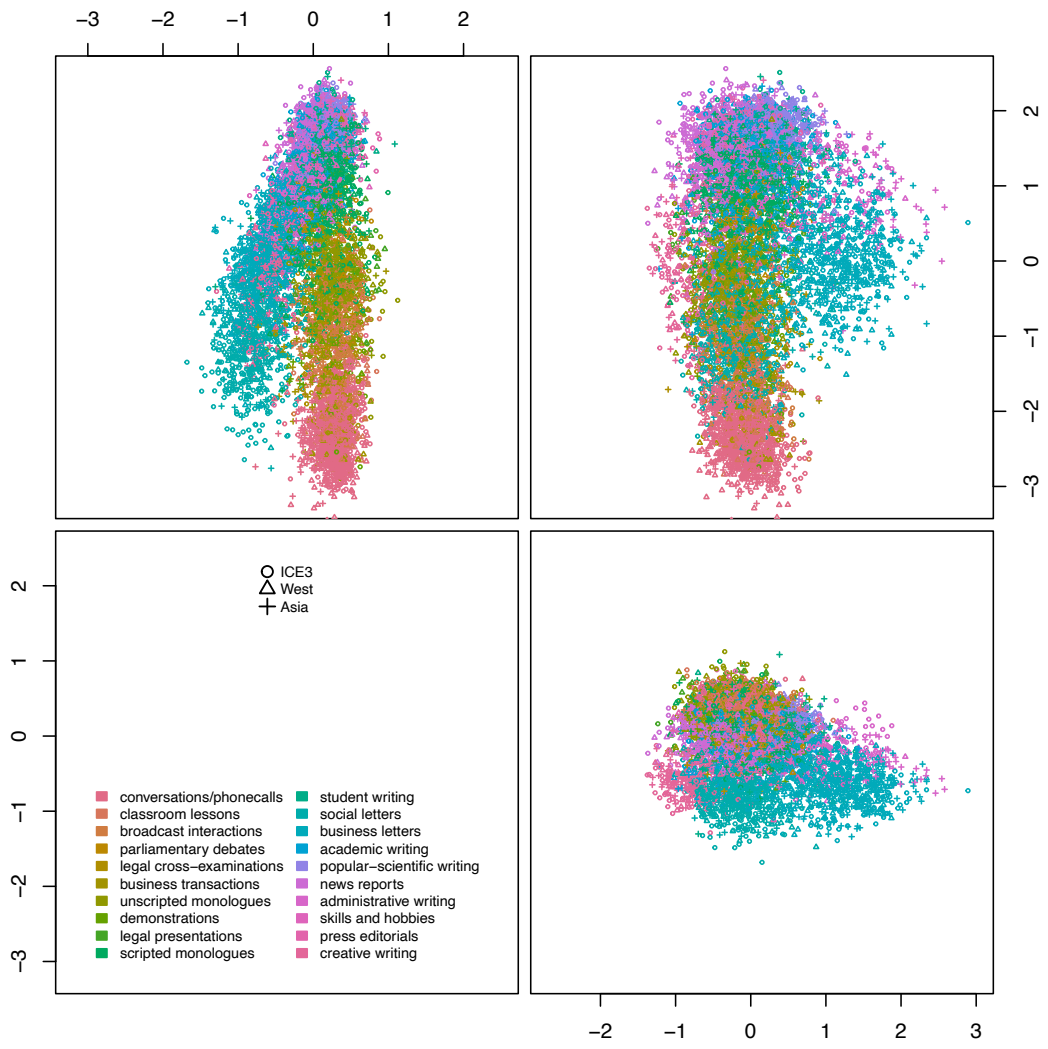
Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

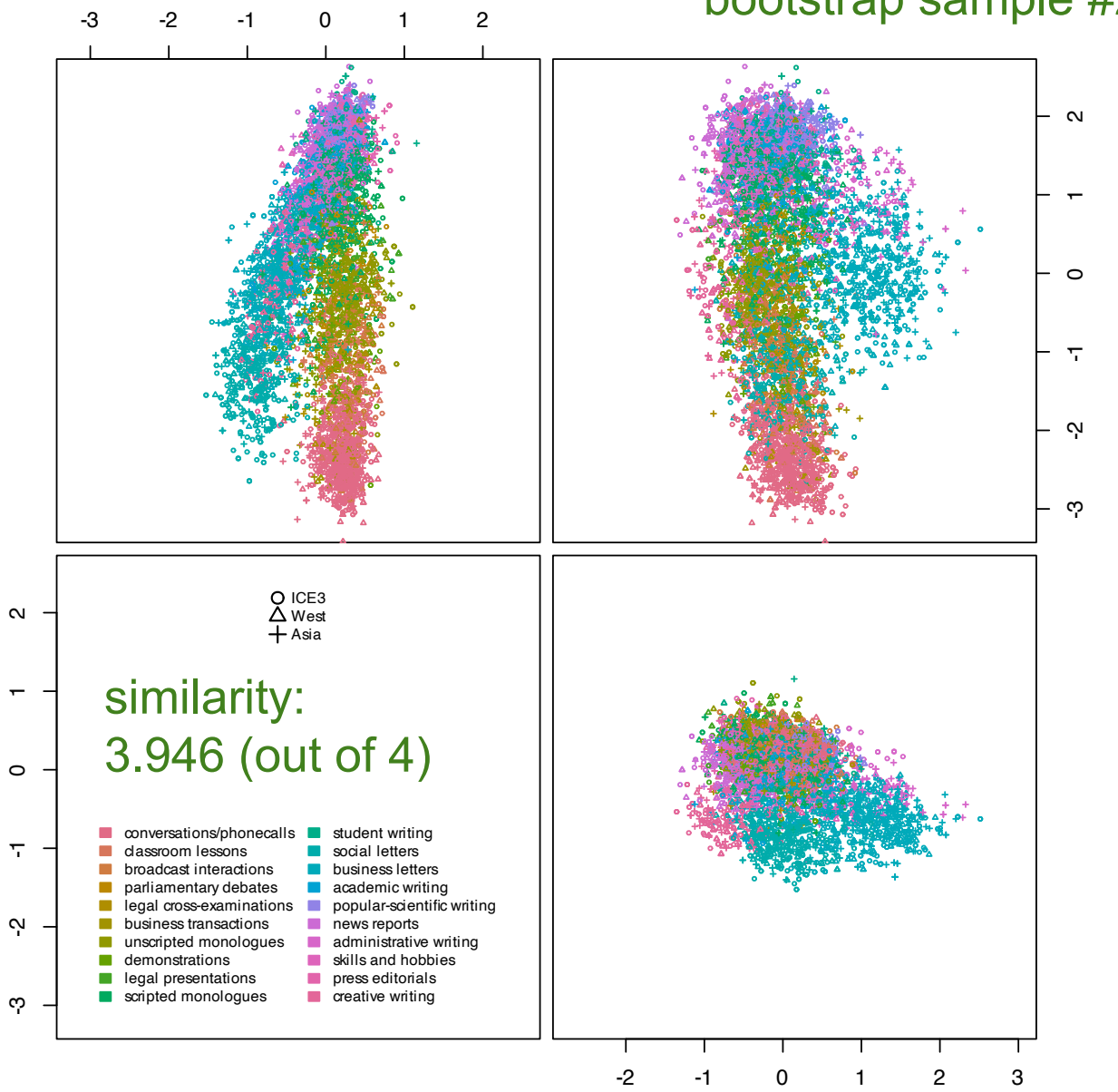
FAU

RWTH AACHEN UNIVERSITY

Stability: Bootstrapping (yay!)



bootstrap sample #2



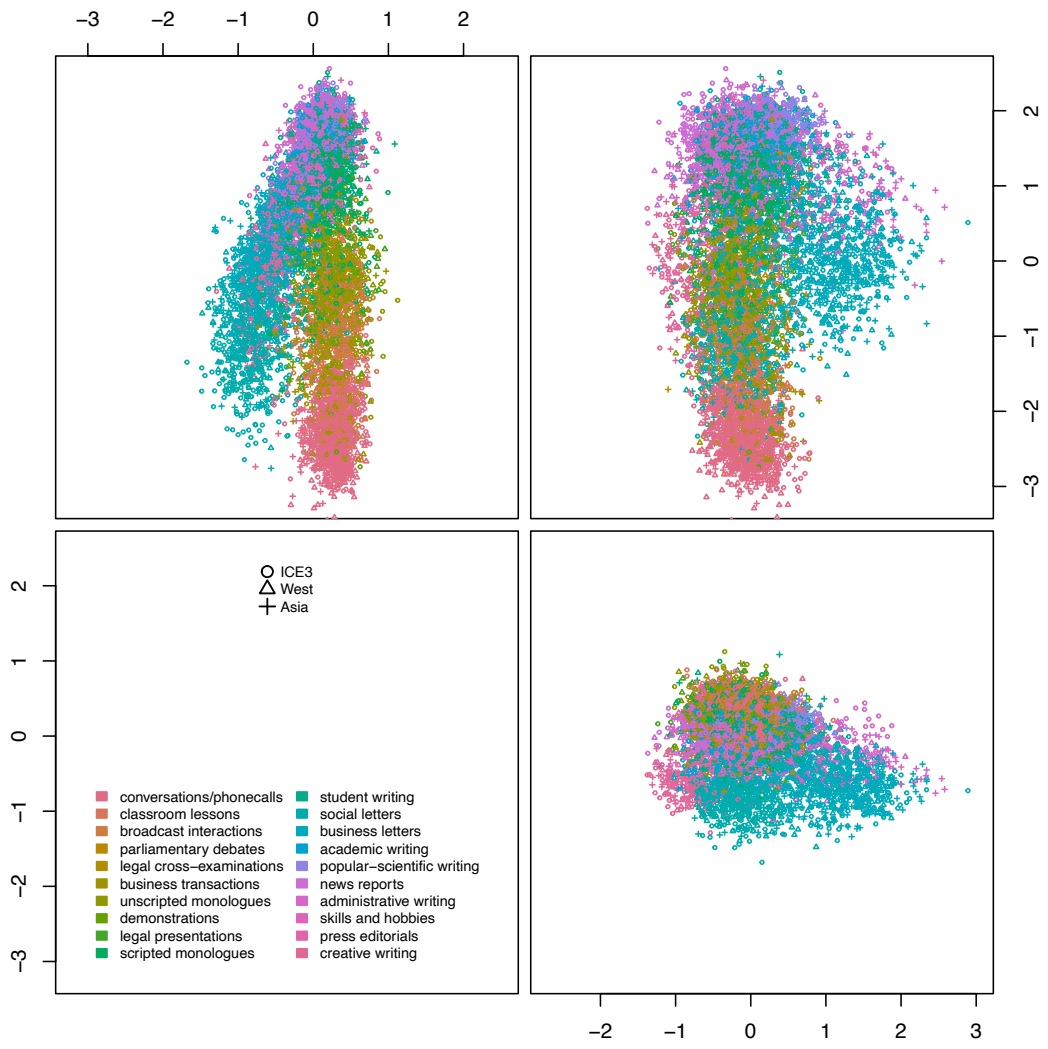
Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

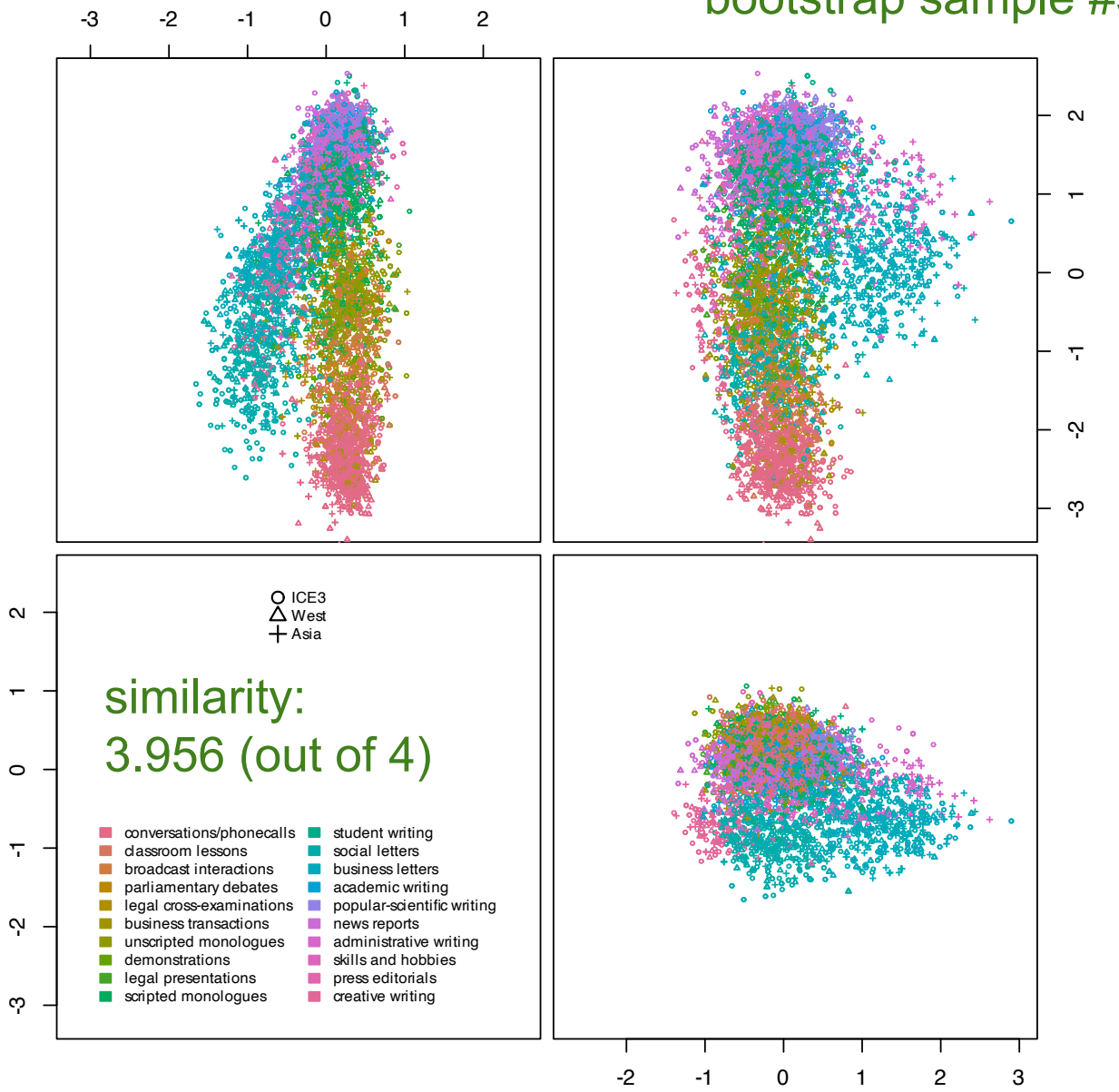
FAU

RWTH AACHEN UNIVERSITY

Stability: Bootstrapping (yay!)



bootstrap sample #3



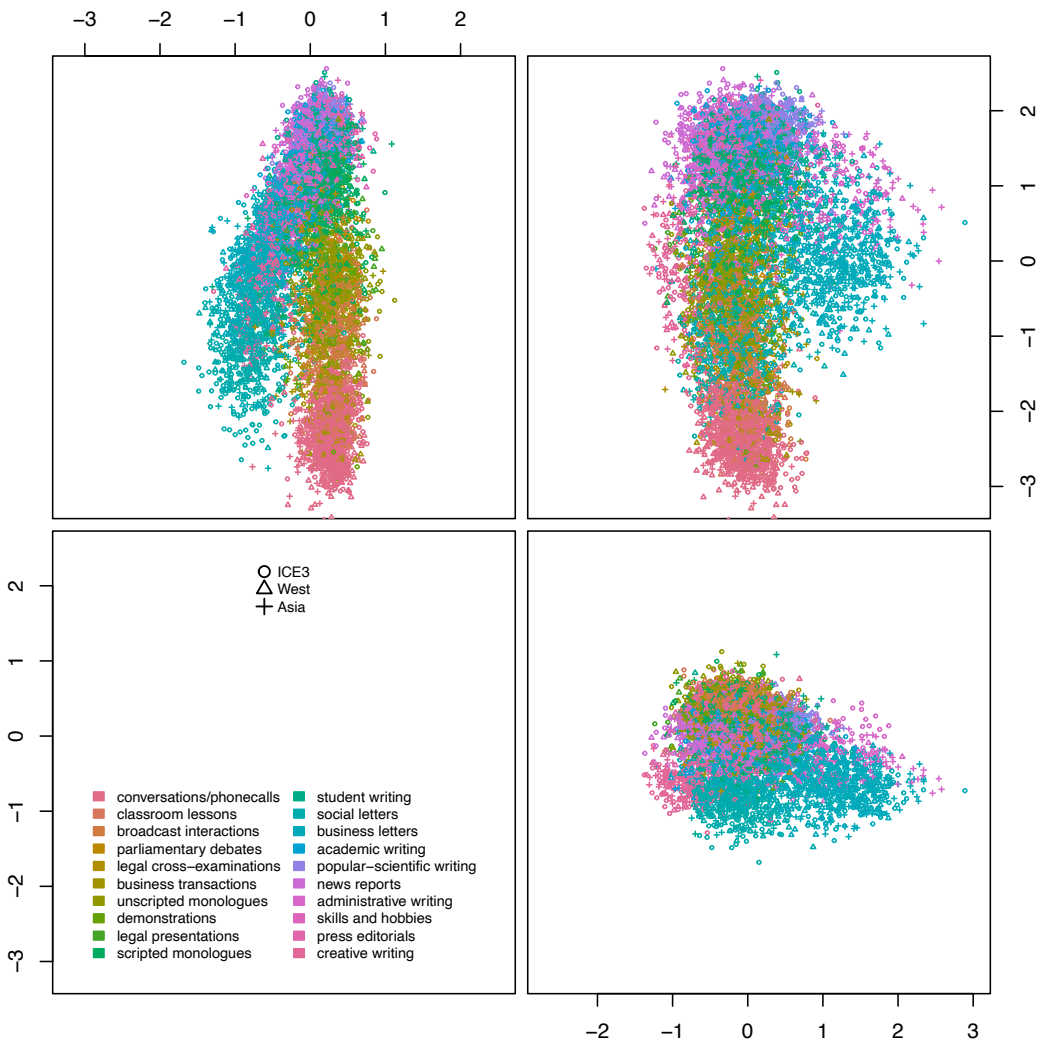
Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

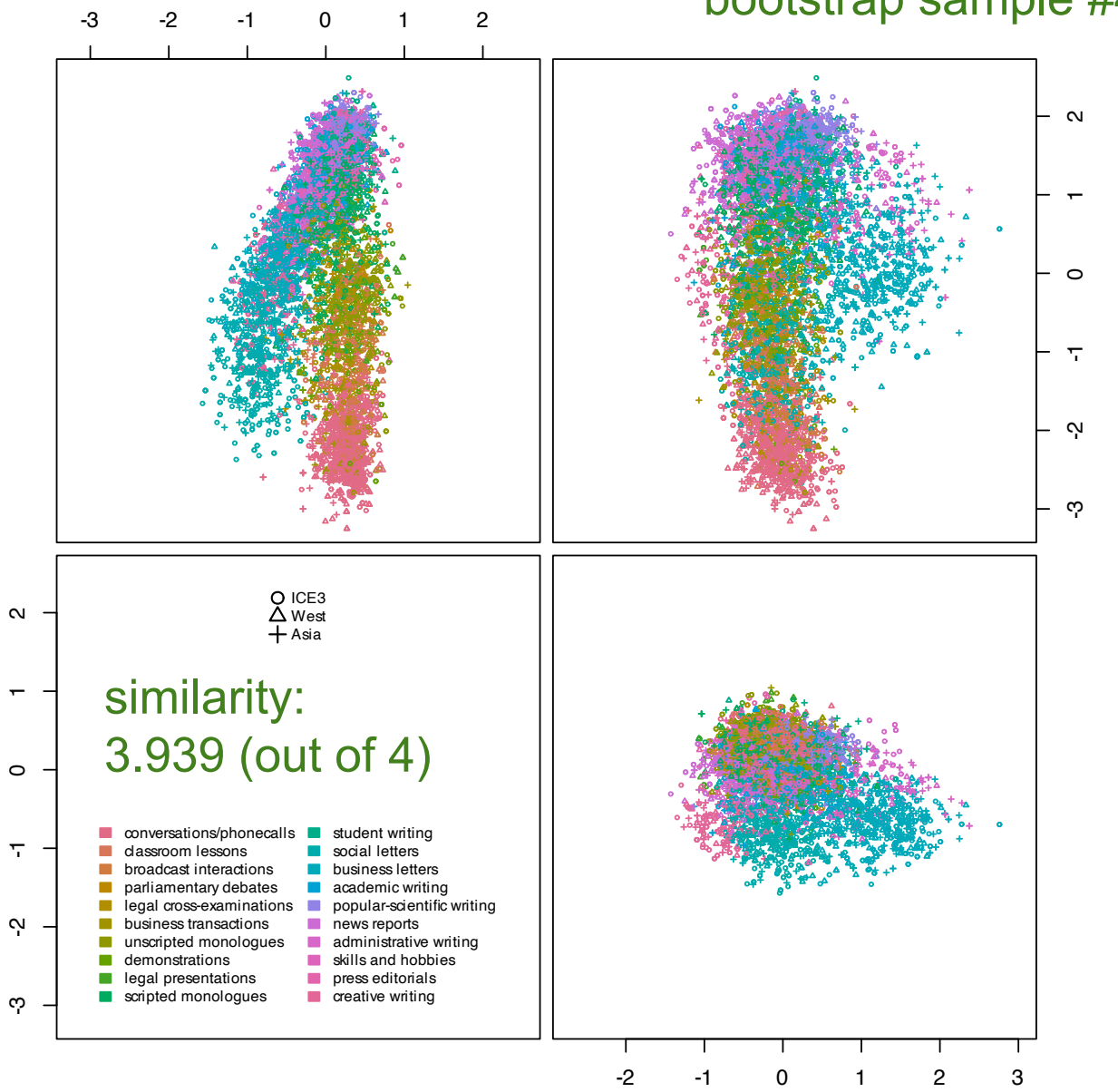
FAU

RWTH AACHEN UNIVERSITY

Stability: Bootstrapping (yay!)



bootstrap sample #4



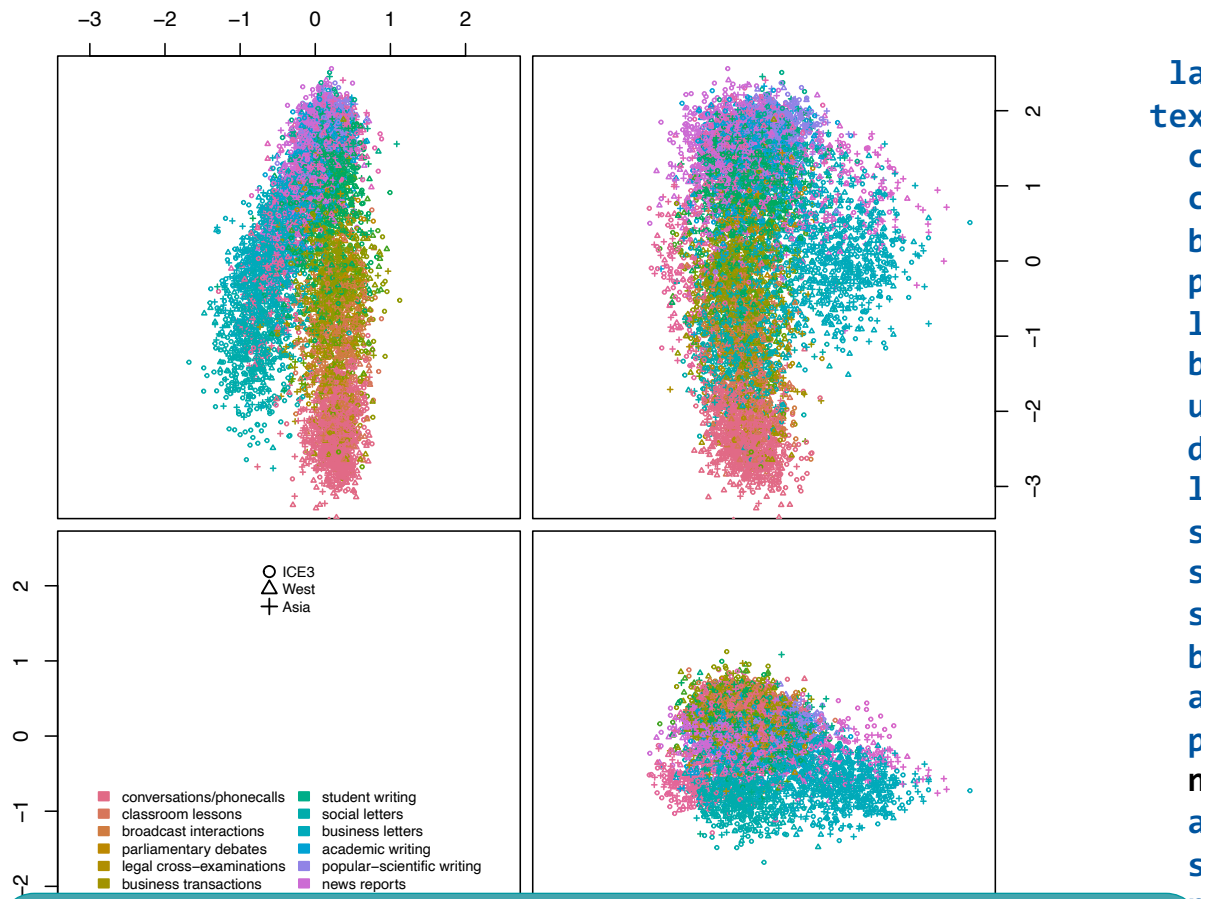
Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

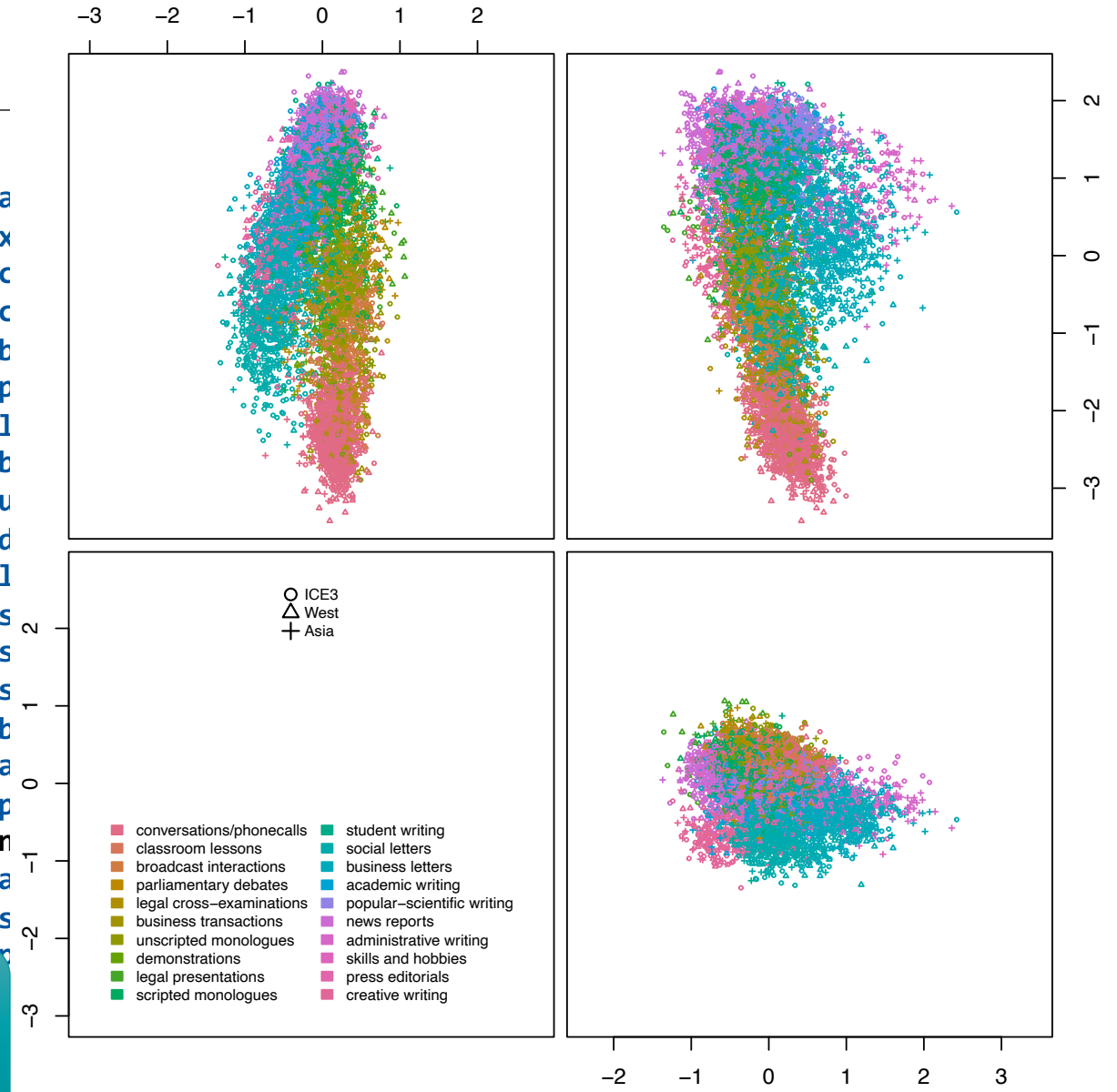
Universität Zürich[UZH]

FAU

RWTH AACHEN UNIVERSITY

Balance: between text categories






```
ICE9bal <- GMA$new(ZL)
ICE9bal$add.discriminant(Meta$textcat20,
  balanced=TRUE, max.dim=4)
```

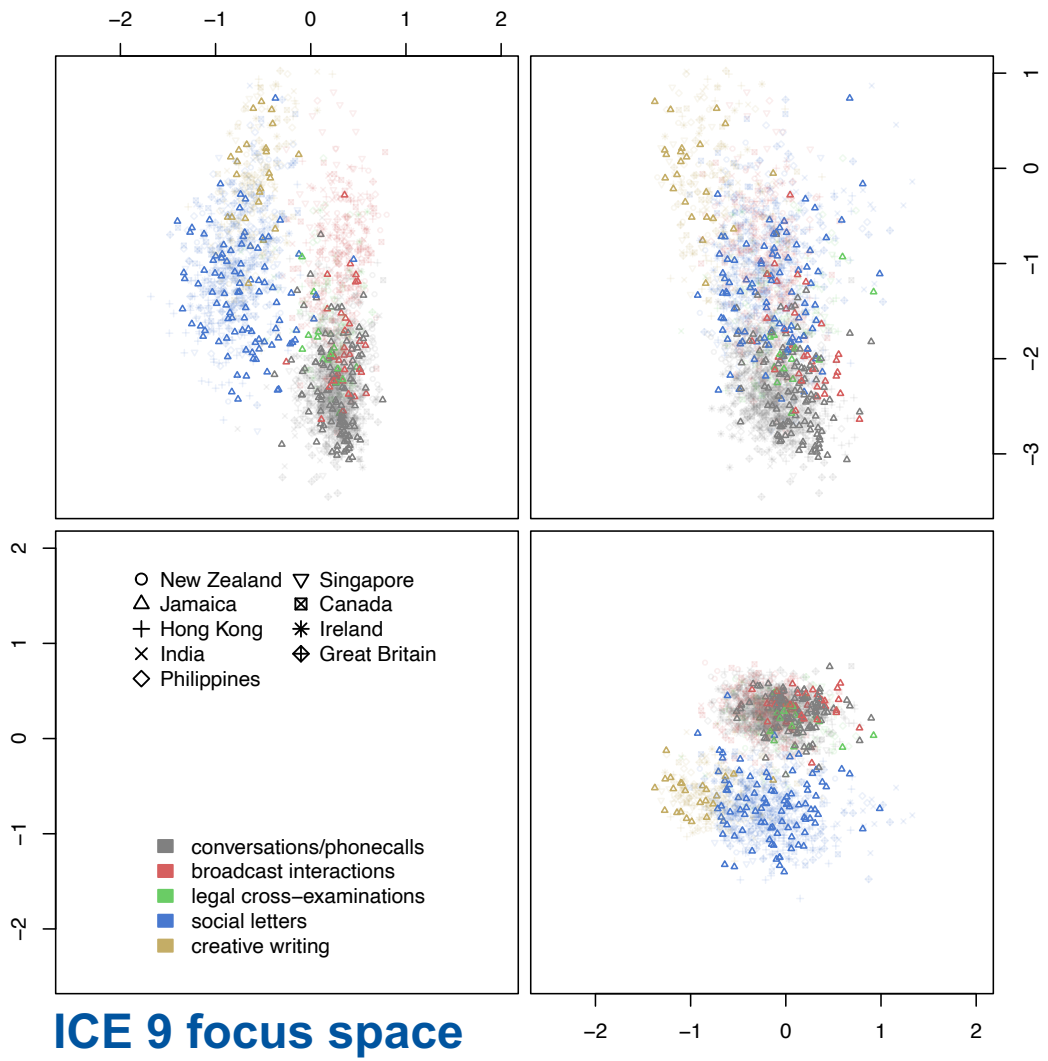

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

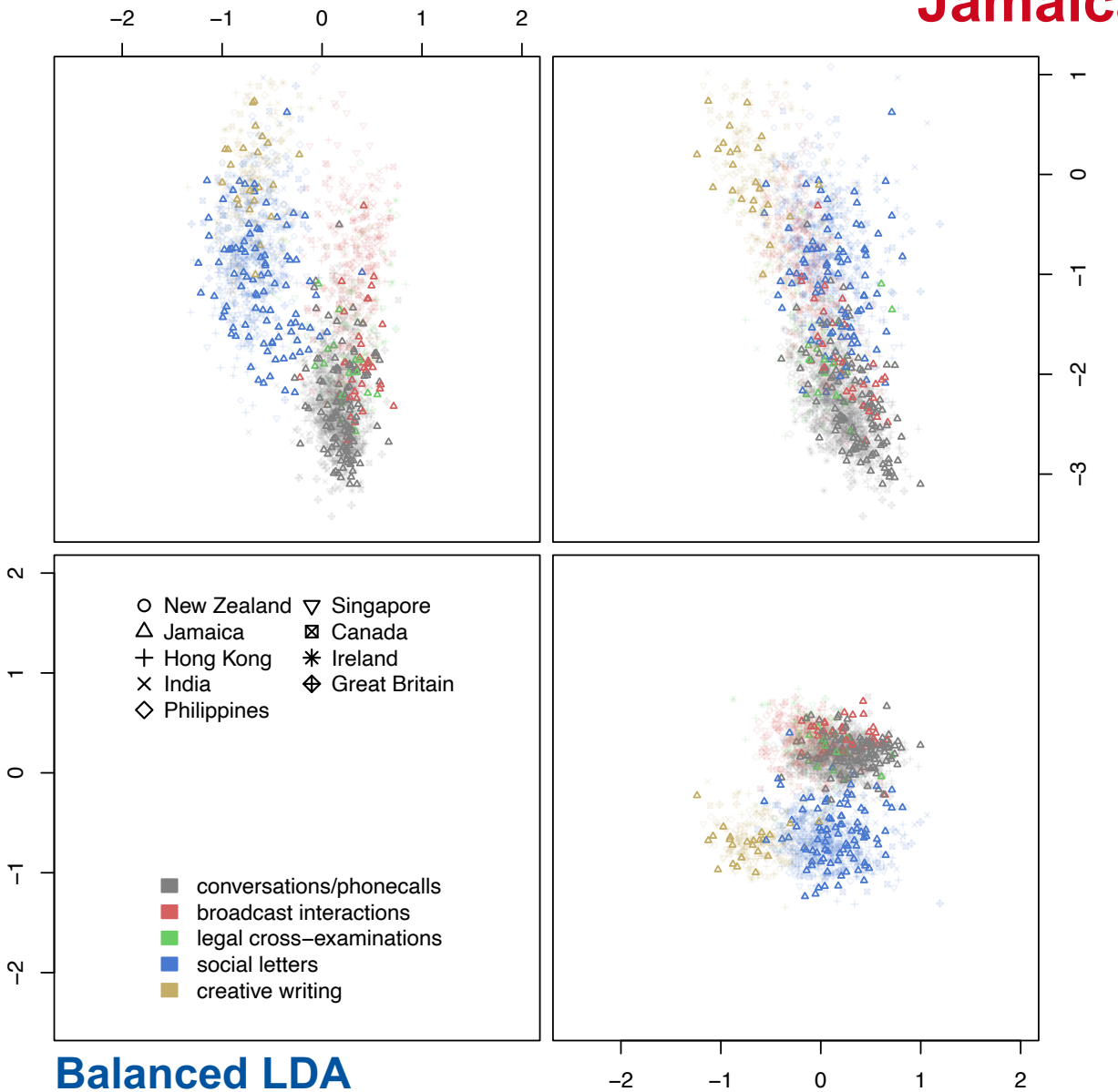
RWTH AACHEN UNIVERSITY

Balance: between text categories



ICE 9 focus space

Jamaica



Balanced LDA

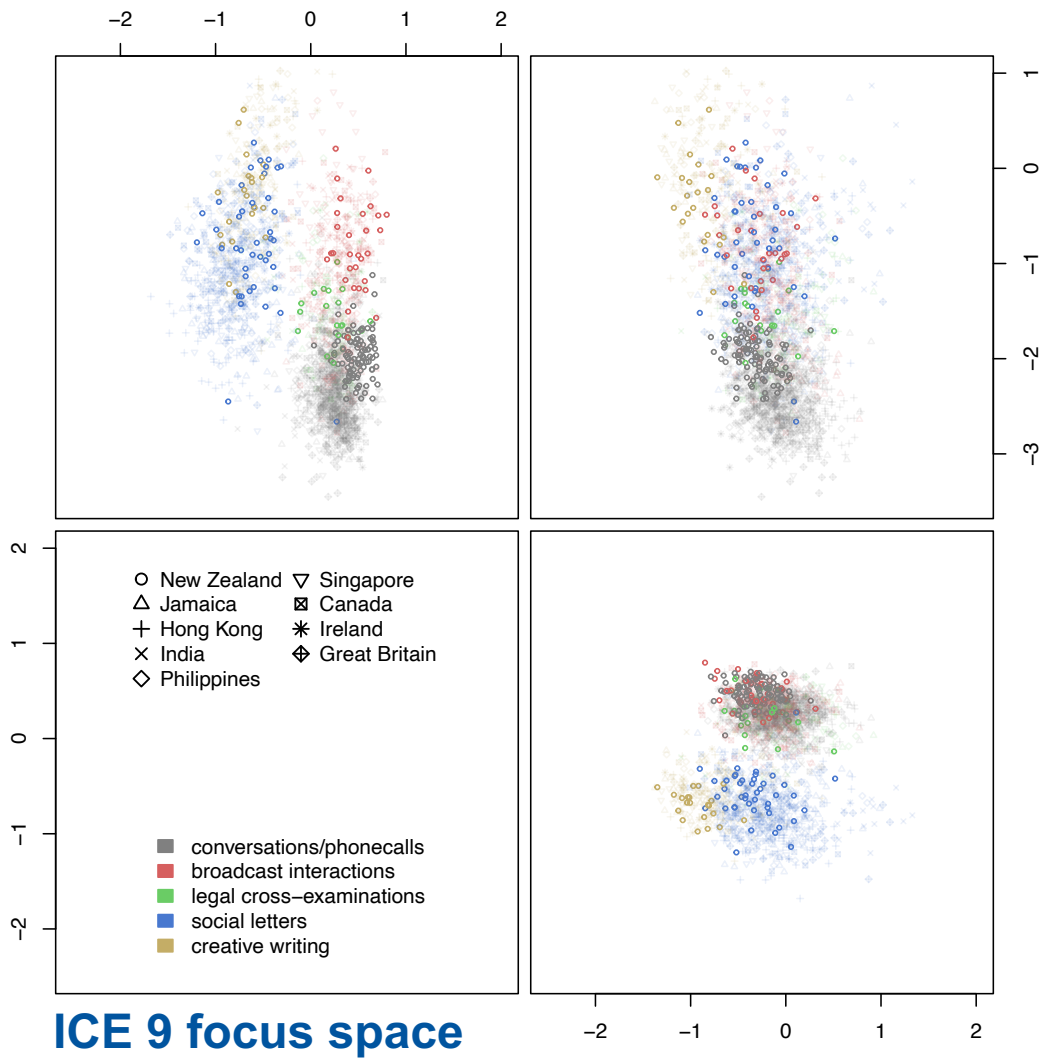
Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

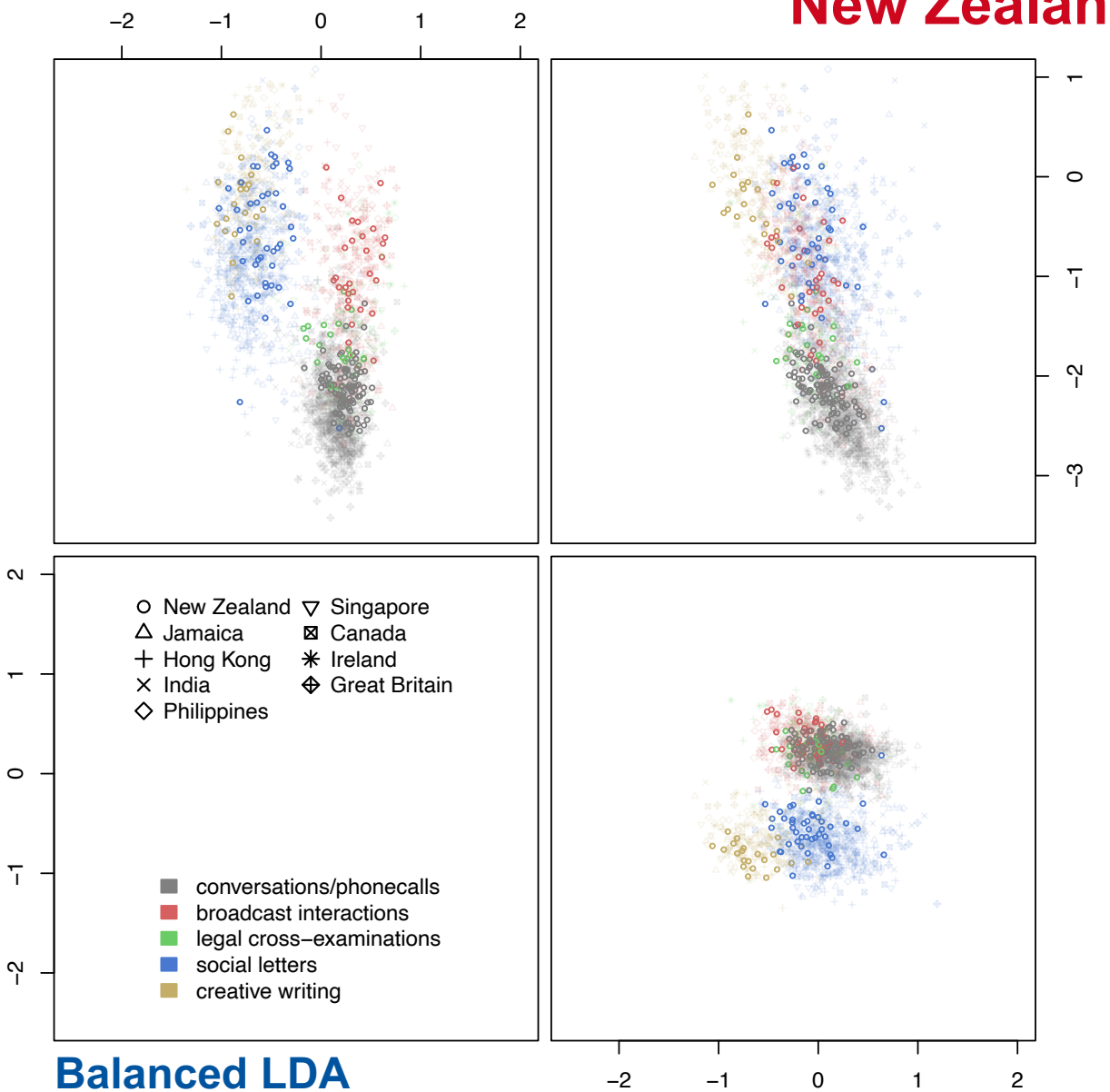
FAU

RWTH AACHEN UNIVERSITY

Balance: between text categories



ICE 9 focus space



Balanced LDA

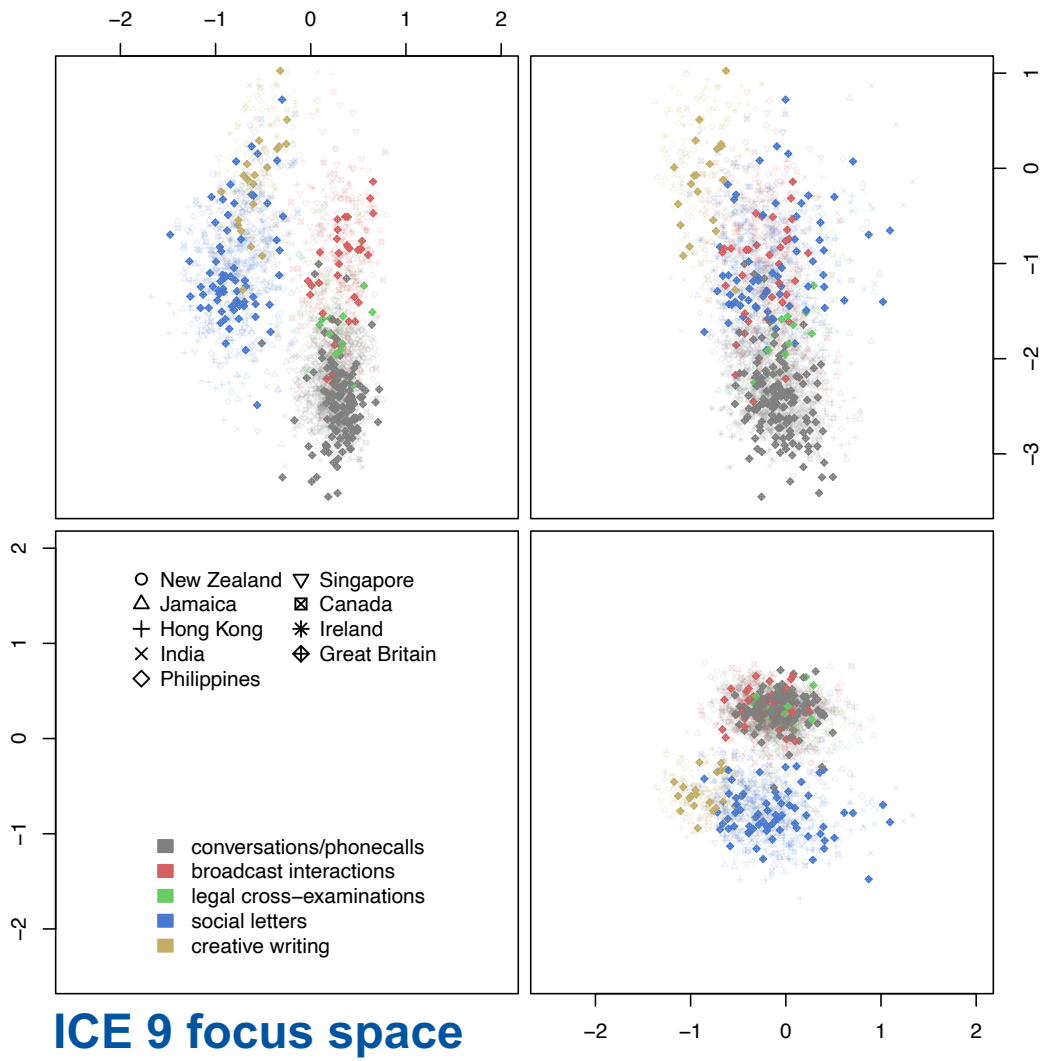
Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

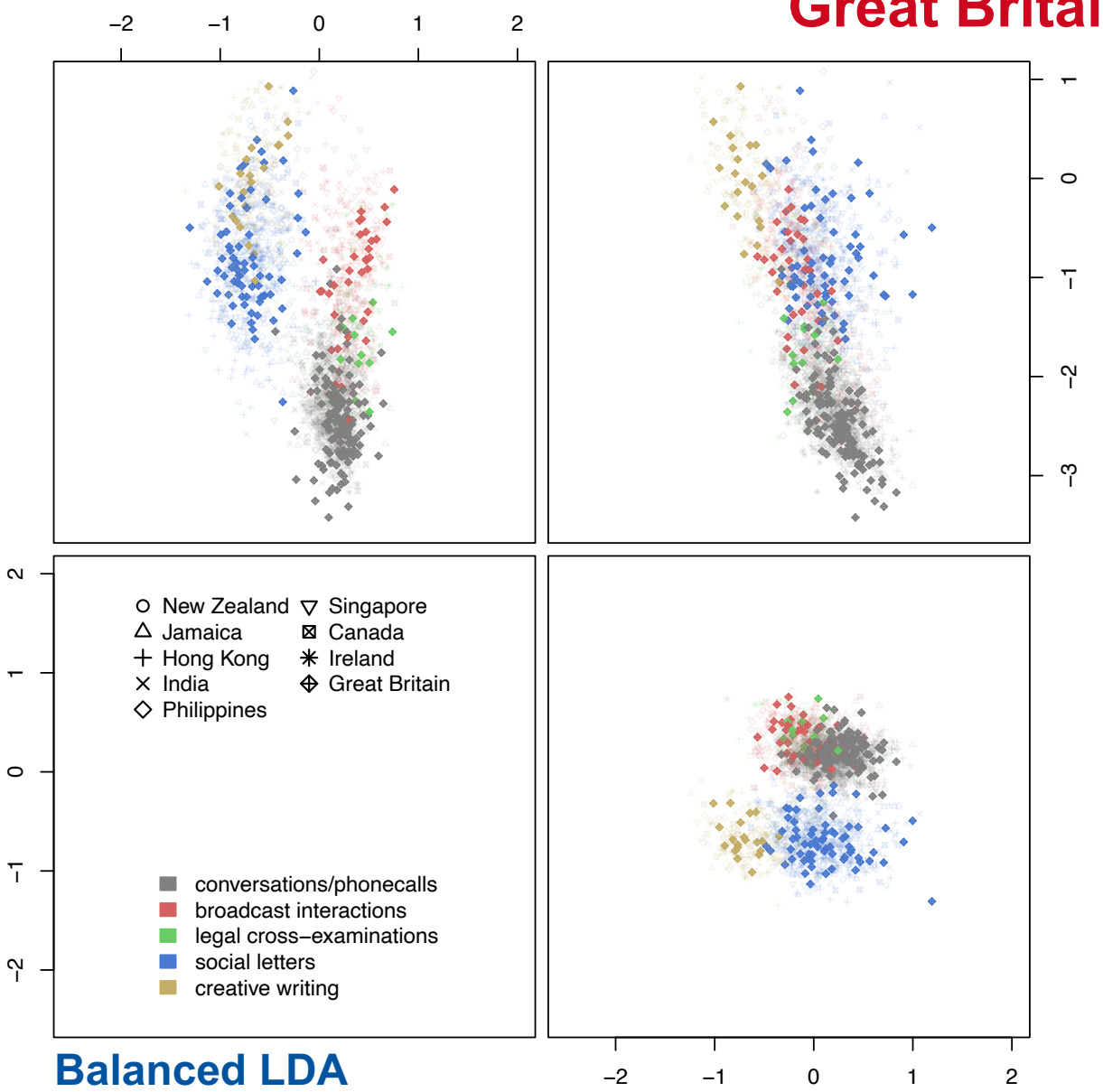
RWTH AACHEN UNIVERSITY

Balance: between text categories



ICE 9 focus space

Great Britain



Balanced LDA

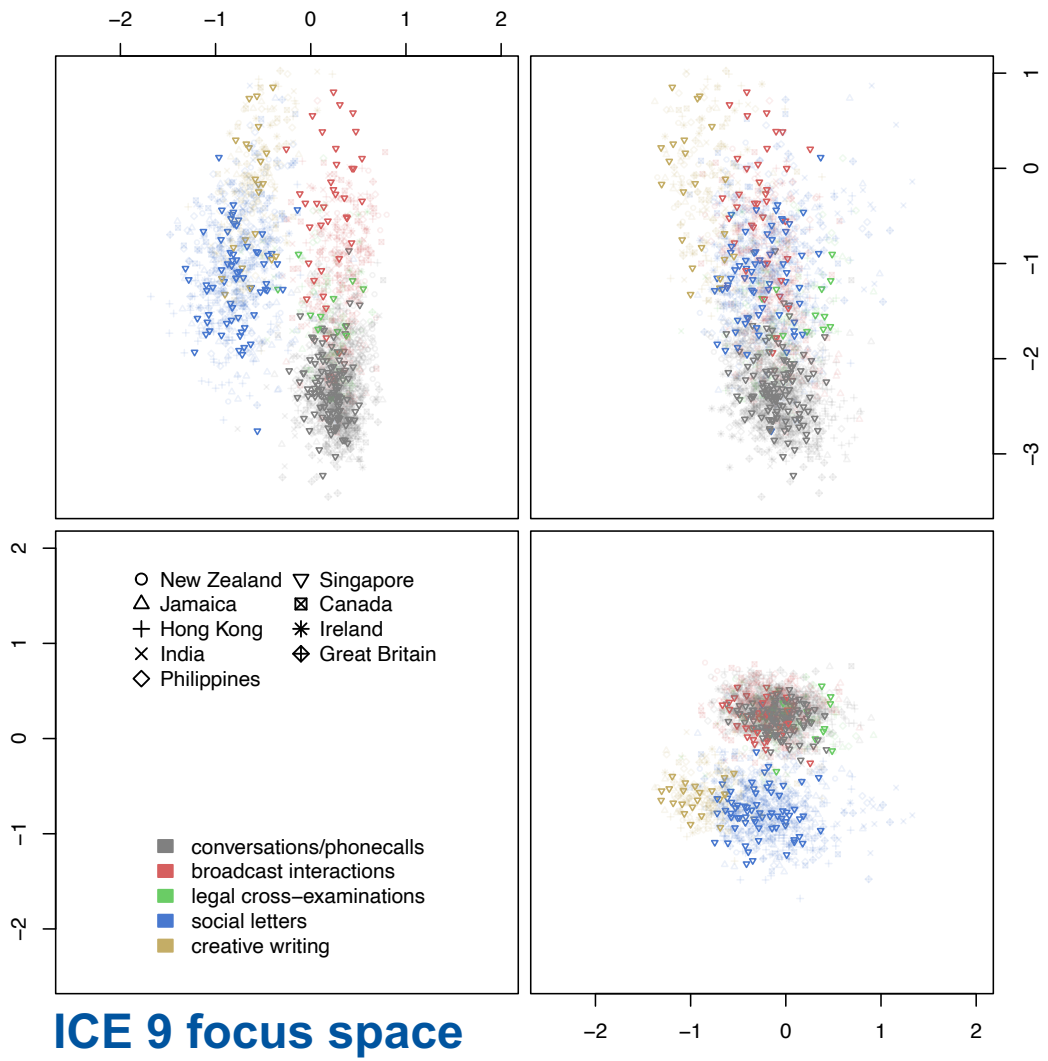
Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

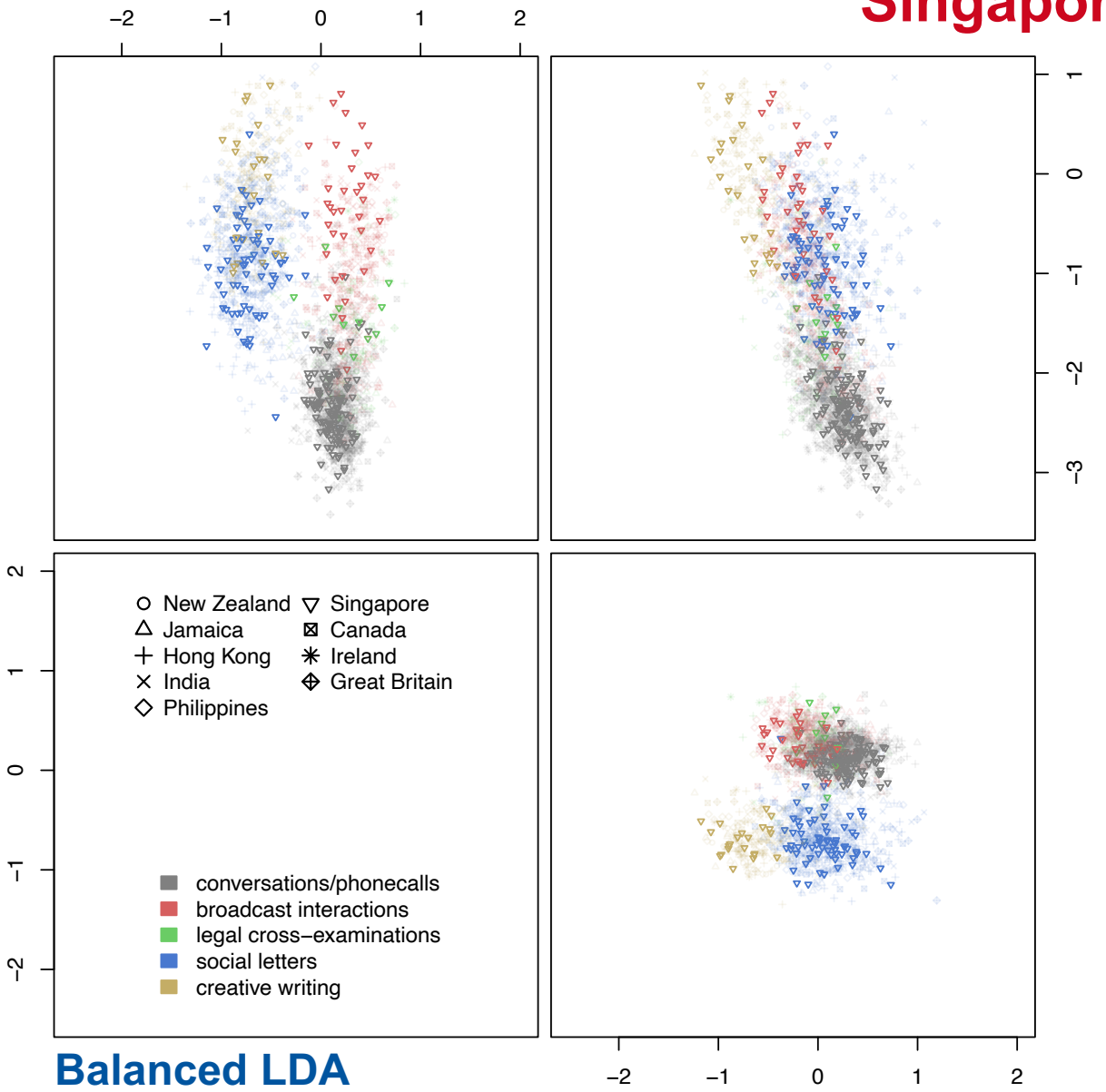
RWTH AACHEN UNIVERSITY

Balance: between text categories



ICE 9 focus space

Singapore



Balanced LDA

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

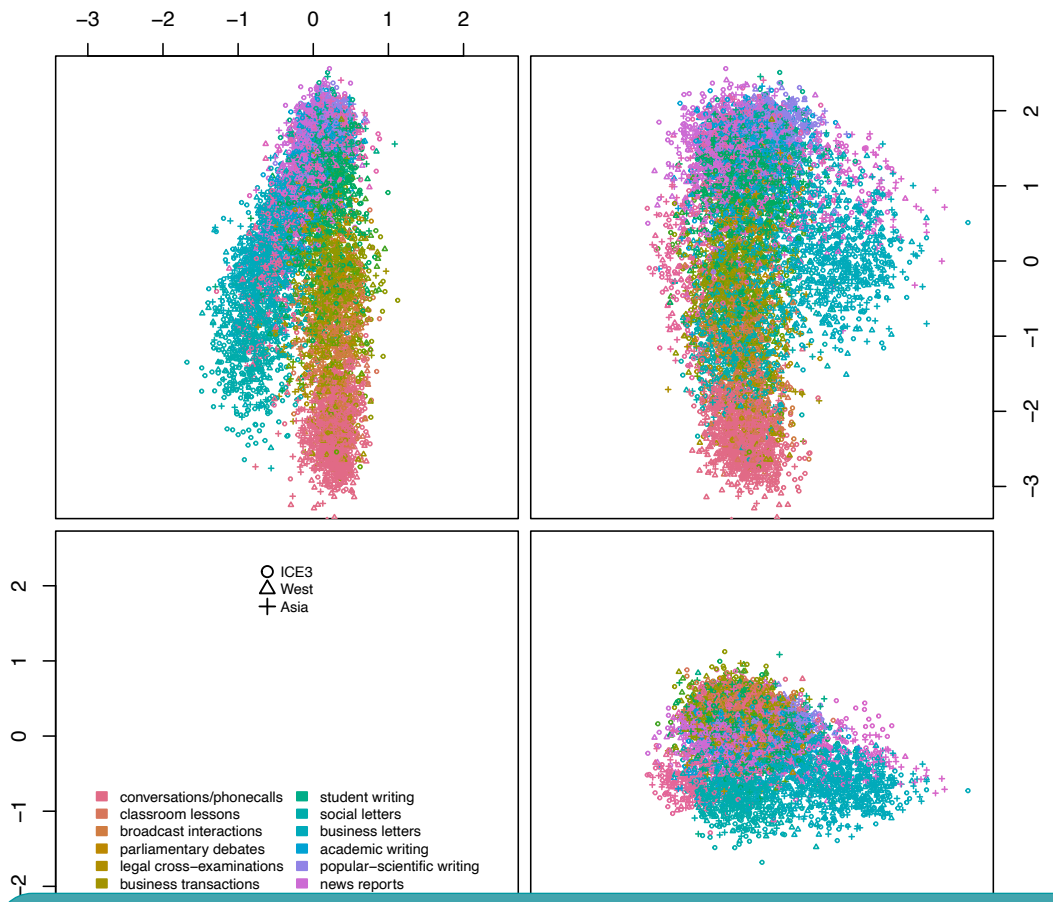
Universität Zürich[UZH]

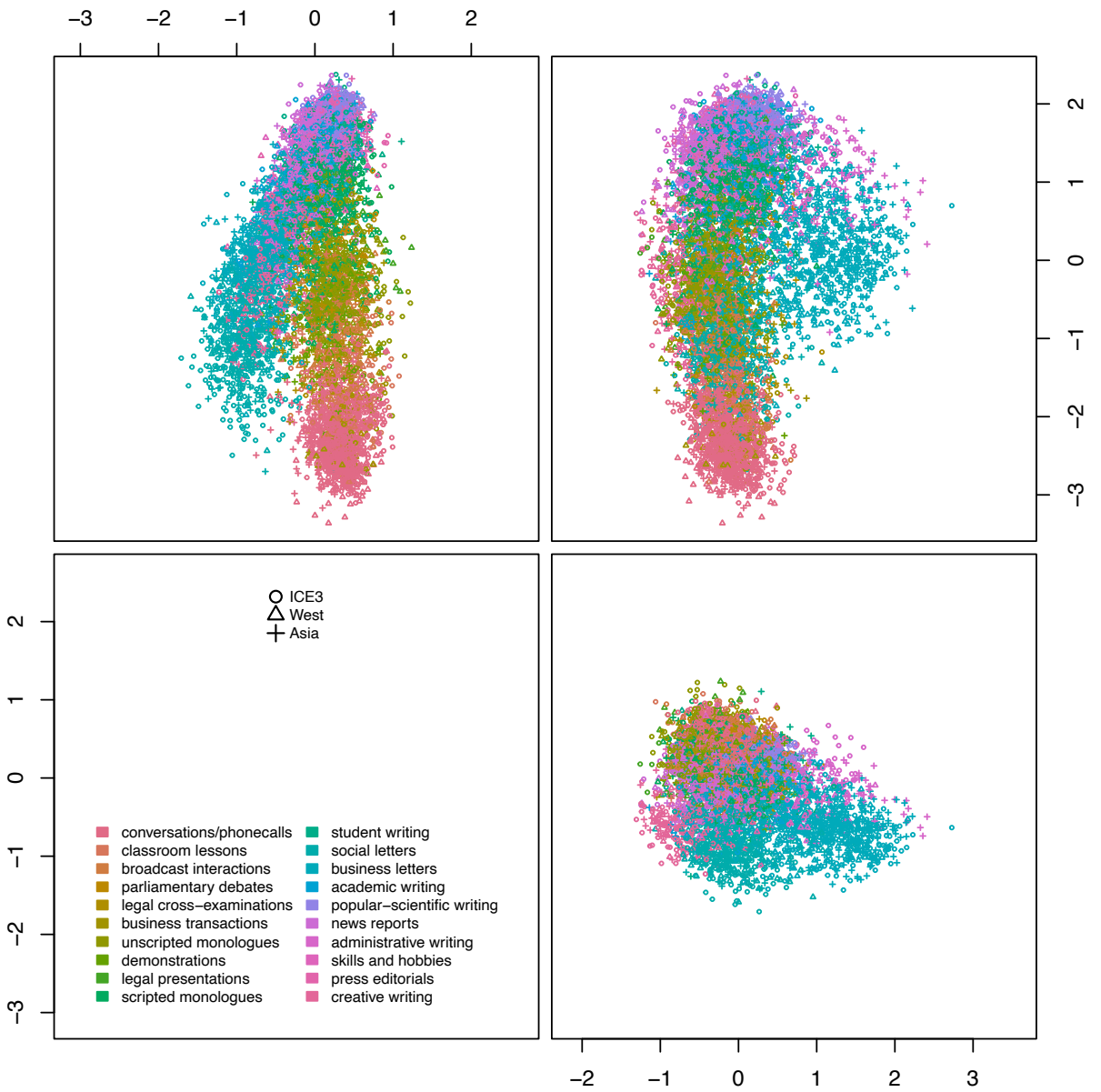
FAU

RWTH AACHEN UNIVERSITY

https://github.com/schtepf/GMA/blob/main/doc/gma_maths.pdf

Confounding proxies: repeated-measures LDA






```
ICE9rep <- GMA$new(ZL)
ICE9rep$add.discriminant(Meta$textcat20,
  cohorts=Meta$variety, max.dim=4)
```


Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

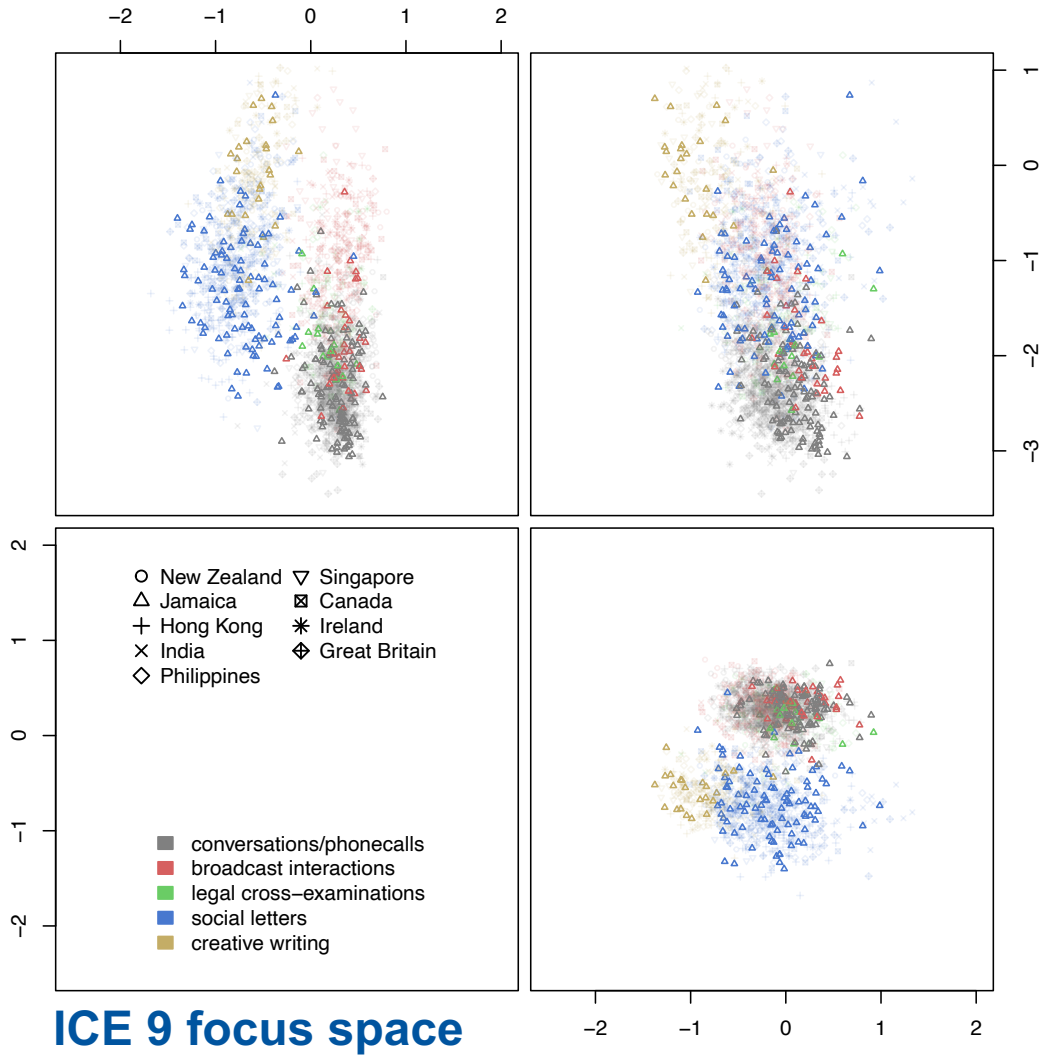
Universität Zürich[UZH]

FAU

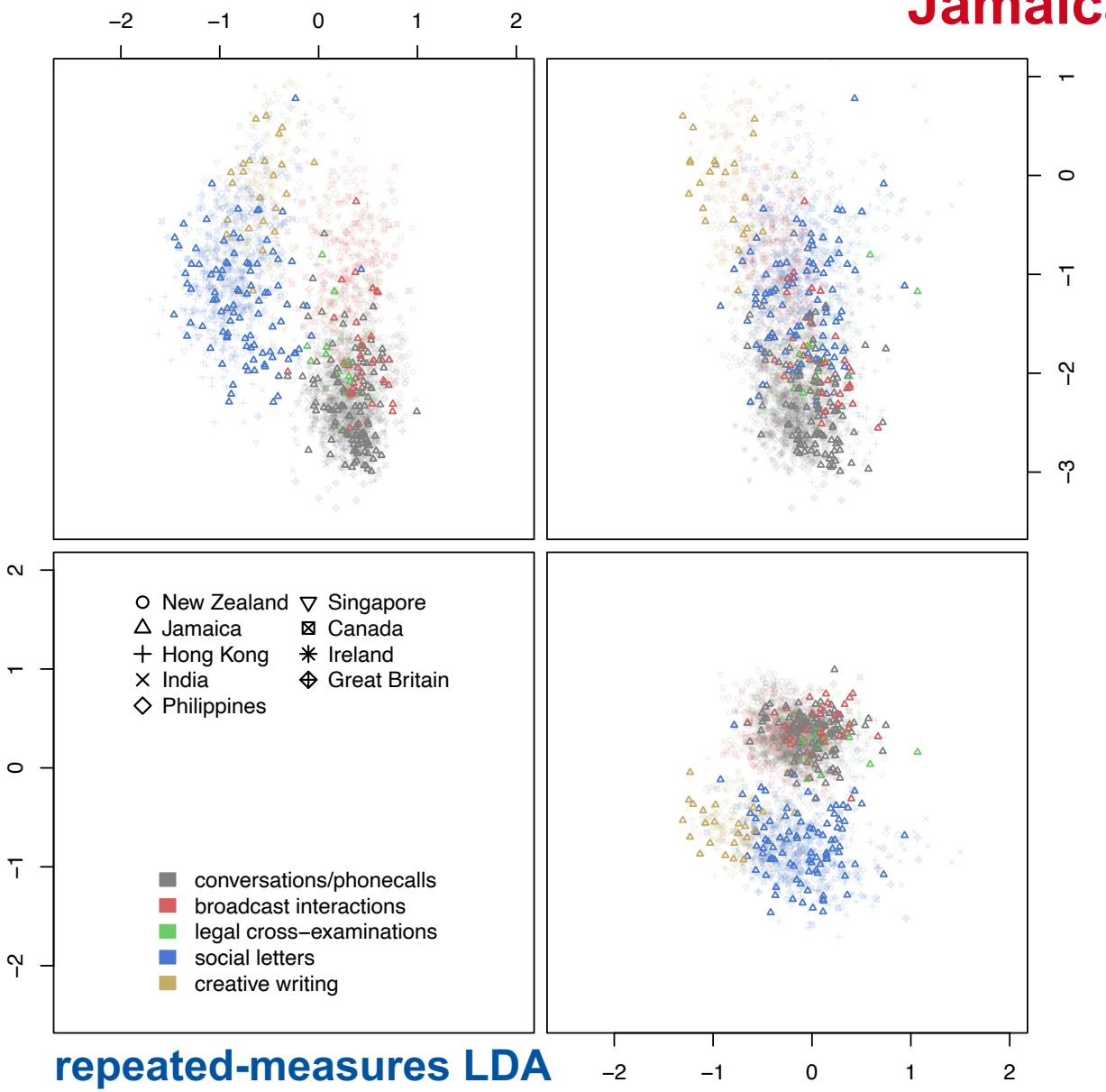
RWTH AACHEN UNIVERSITY

https://github.com/schtepf/GMA/blob/main/doc/gma_maths.pdf

Confounding proxies: repeated-measures LDA



ICE 9 focus space



repeated-measures LDA

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

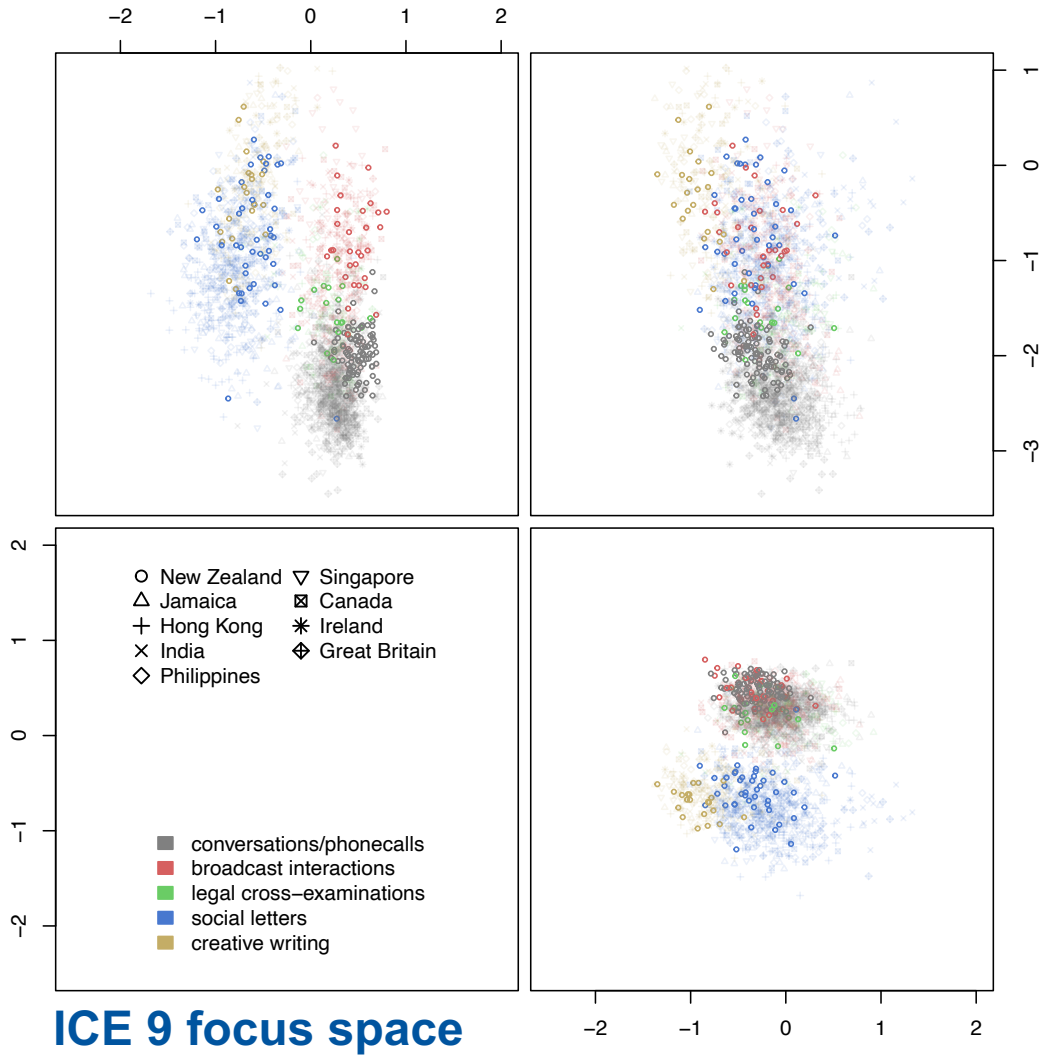
Universität Zürich[UZH]

FAU

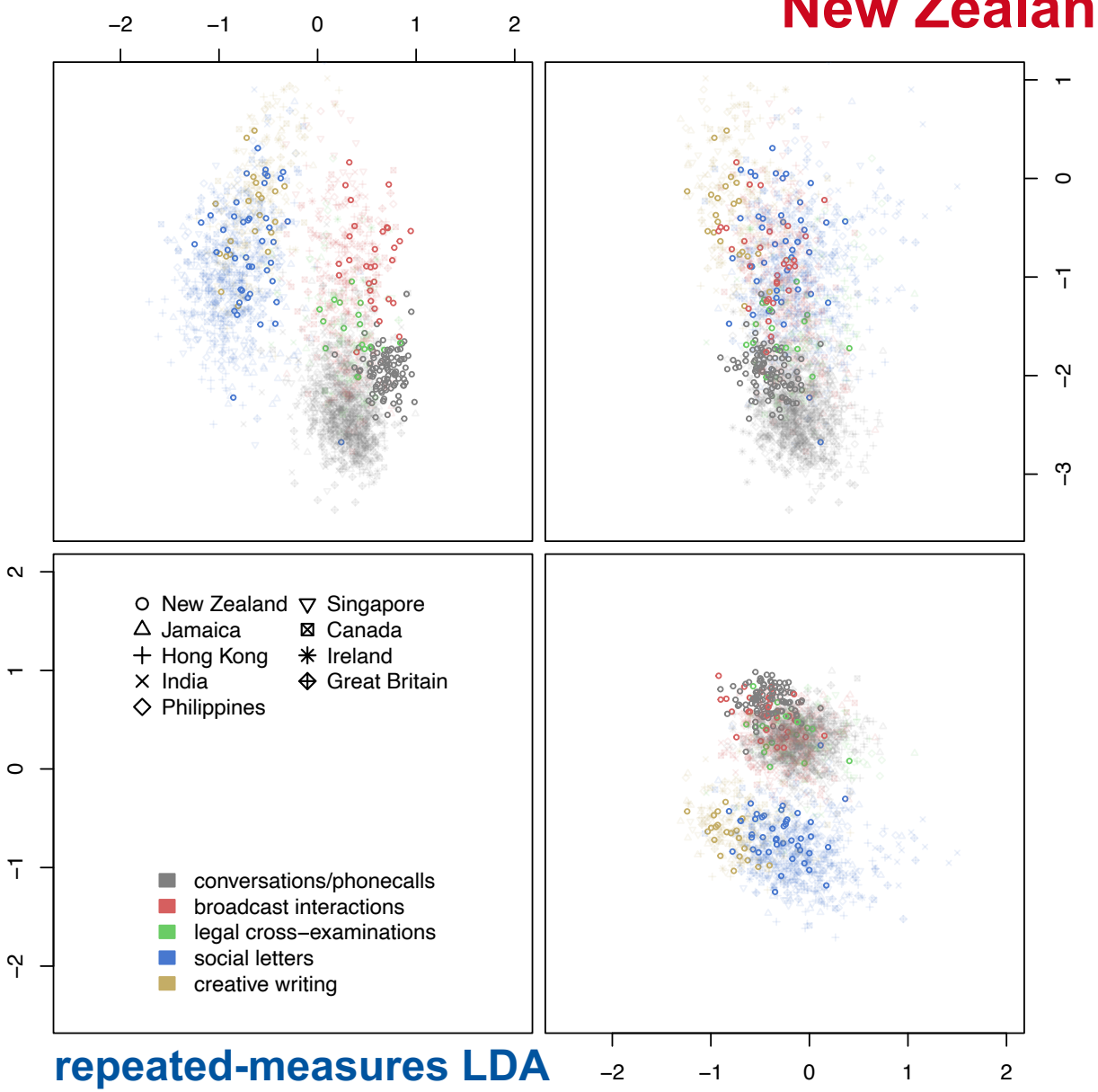
RWTH AACHEN UNIVERSITY

https://github.com/schtepf/GMA/blob/main/doc/gma_maths.pdf

Confounding proxies: repeated-measures LDA



ICE 9 focus space



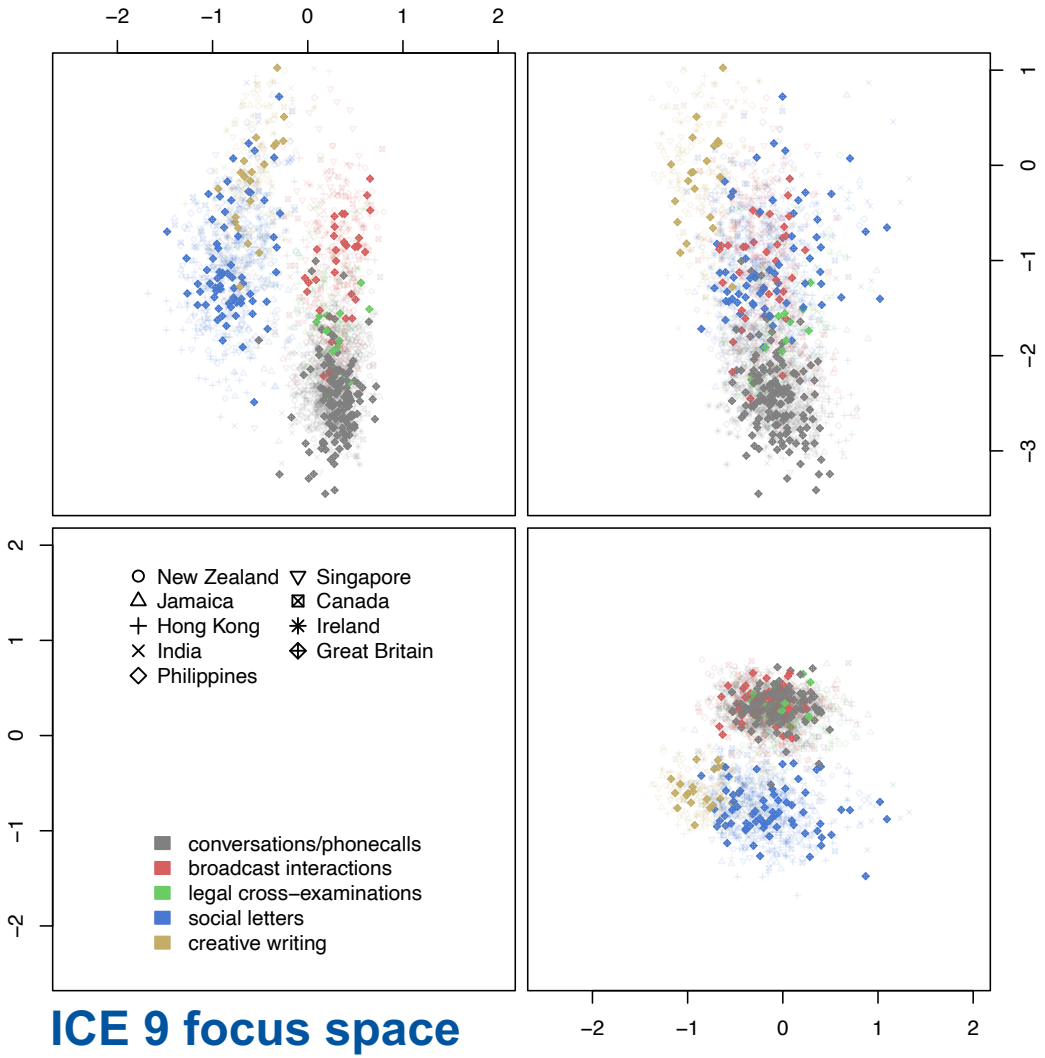
repeated-measures LDA

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

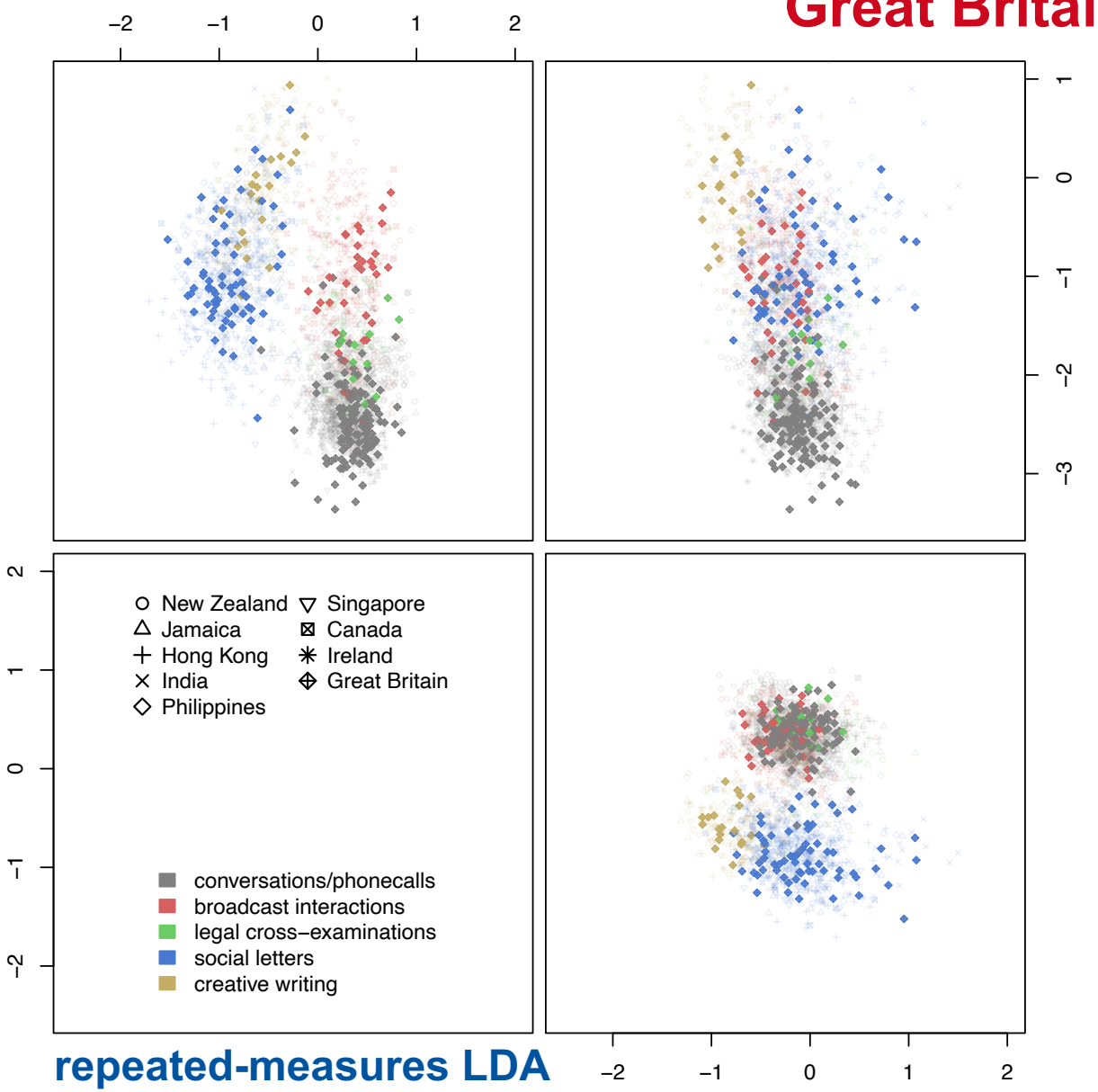
https://github.com/schtepf/GMA/blob/main/doc/gma_maths.pdf

Confounding proxies: repeated-measures LDA



ICE 9 focus space

Great Britain



repeated-measures LDA

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

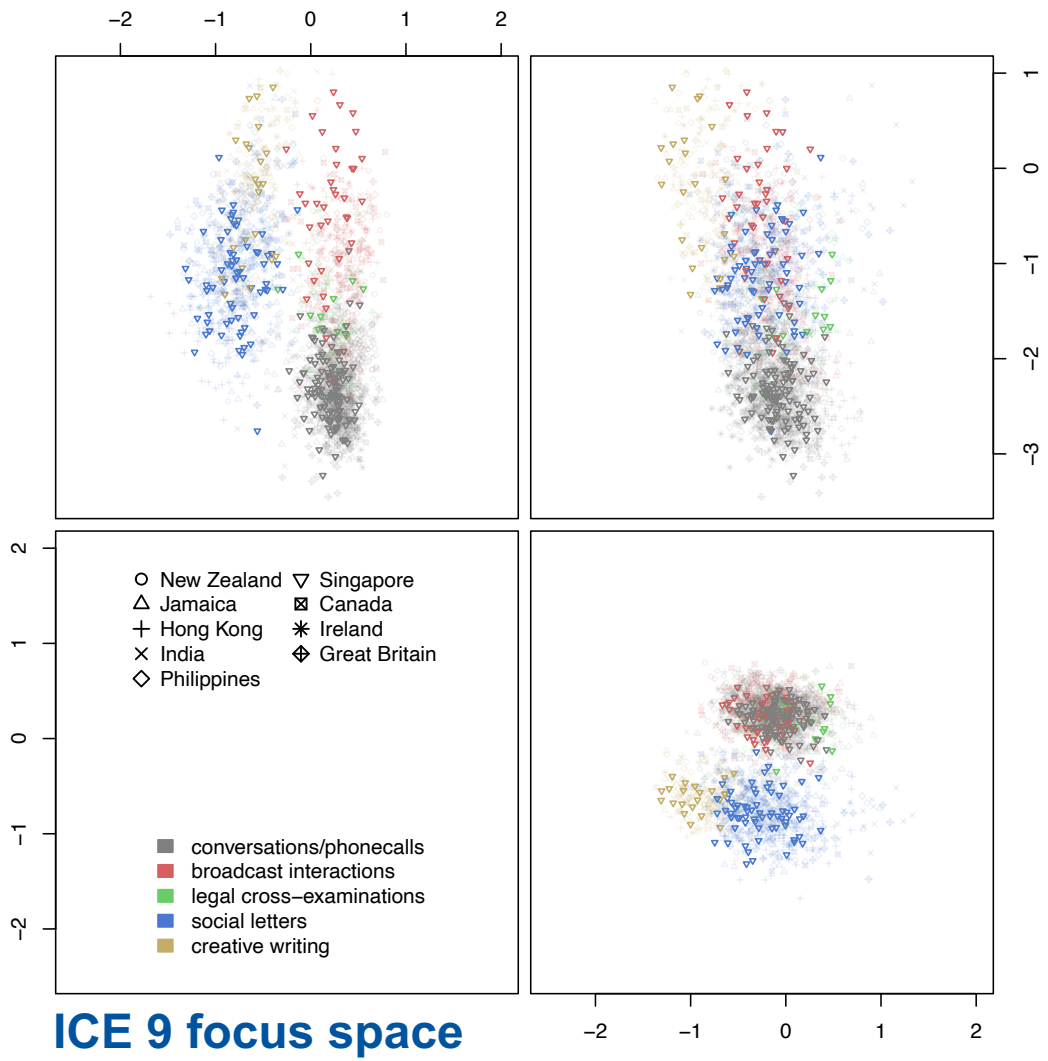
Universität Zürich[UZH]

FAU

RWTH AACHEN UNIVERSITY

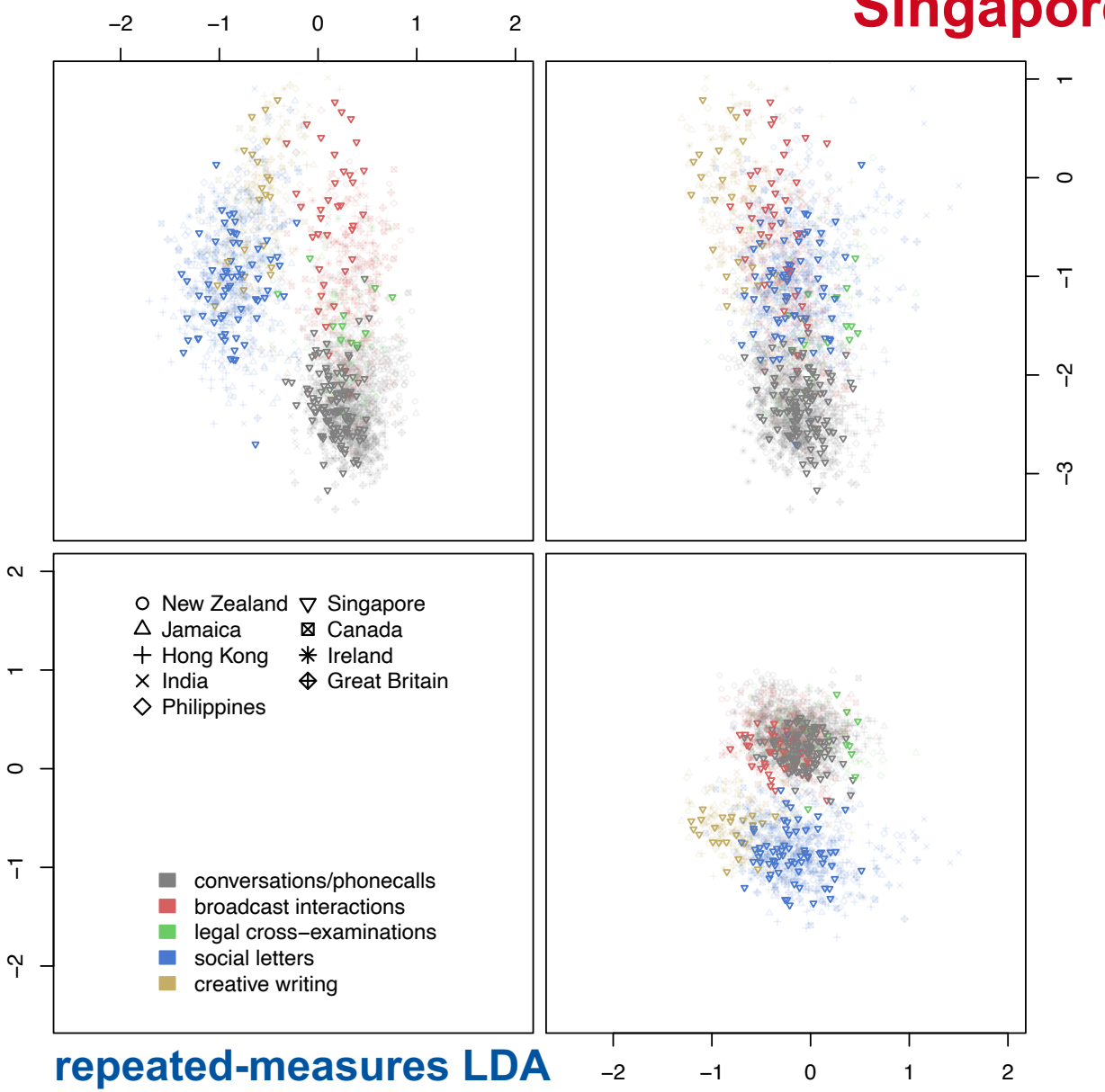
https://github.com/schtepf/GMA/blob/main/doc/gma_maths.pdf

Confounding proxies: repeated-measures LDA



ICE 9 focus space

Singapore



repeated-measures LDA

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

RWTH AACHEN UNIVERSITY

Conclusions

- GMA is an excellent approach for finding structure in multivariate data :-)
- Even better with `gmatools`!

Methodological conclusions:

- GMA appears to be fairly robust wrt. modification of the data set
- Excellent stability of LDA on our (fairly large) data set (→ bootstrapping)
- Imbalance between proxy categories w/o substantial effect in our case study (YMMV)
- Confounding proxies are often an issue → repeated-measures LDA is recommended

Evert / Neumann / Schneider / Frenken | Corpus Linguistics 2025 | 2 July 2025
DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Thank you
for your attention

https://osf.io/9p25y

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation
FNSNF Swiss National Science Foundation
QuanTOR
Universität Zürich[UZH]
FAU
RWTH AACHEN UNIVERSITY

References

- Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.
- Biber, D. (1993). The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, 26:331–345.
- Diwersy, S., Evert, S., & Neumann, S. (2014). A weakly supervised multivariate approach to the study of language variation. In Szmrecsanyi, B. and Wälchli, B., editors, *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*, pages 174–204. De Gruyter, Berlin, Boston.
- Egbert, J. & Biber, D. (2018). Do all roads lead to Rome?: Modeling register variation with factor analysis and discriminant analysis. *Corpus Linguistics and Linguistic Theory*, 14(2):233–273.
- Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics & Applied Probability. Chapman & Hall, CRC, Boca Raton.
- Evert, S., & The CWB Development Team (2020). The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial. CWB Version 3.5. https://cwb.sourceforge.io/documentation.php
- Evert, S. & Neumann, S. (2017). The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In De Sutter, G., Lefer, M.-A., and Delaere, I., editors, *Empirical Translation Studies. New Theoretical and Methodological Traditions*, TiLSM 300, pages 47–80. Mouton de Gruyter, Berlin. Online supplement: https://www.stephanie-evert.de/PUB/EvertNeumann2017/.

DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

References

- Frenken, F., Evert, S., Schneider, G., and Neumann, S. (2025). How stable are multivariate findings about register variation across varieties of English? On the replicability of geometric multivariate analysis. *ICAME Journal*, 49(1):23–45. https://github.com/quantor-project/gma-replication/
- Garside, R. & Smith, N. (1997). A Hybrid Grammatical Tagger: CLAWS4. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, edited by R. Garside, G. Leech, and A. McEnery, pages 102–121. London: Longman.
- Greenbaum, S. (ed.) (1996). *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.
- Kachru, B. (1985). Standards, codification and sociolinguistic realism: English language in the outer circle. In R. Quirk and H. Widowson (Eds.), *English in the world: Teaching and learning the language and literatures*, pages 11–36. Cambridge: Cambridge University Press.
- Lehmann, H. M. & Schneider, G. (2012). BNC Dependency Bank 1.0. In S. O. Ebeling, J. Ebeling, & H. Hasselgård (eds.), *Aspects of corpus linguistics: compilation, annotation, analysis*. Helsinki: VARIENG.
- Neumann, S. & Evert, S. (2021). A register variation perspective on varieties of English. In Seoane, E. and Biber, D., editors, *Corpus based approaches to register variation*, chapter 6, pages 143–178. Benjamins, Amsterdam. Online supplement: https://www.stephanie-evert.de/PUB/NeumannEvert2021/.

DFG project no. 528467412 | SNF grant no. 218836
QuanTOR – Quantitative Analysis of Textual Organization across Registers