On the replicability of Geometric Multivariate Analysis

Detecting cross-cultural differences in register variation across varieties of English

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

ICAME45 | 19 June 2024

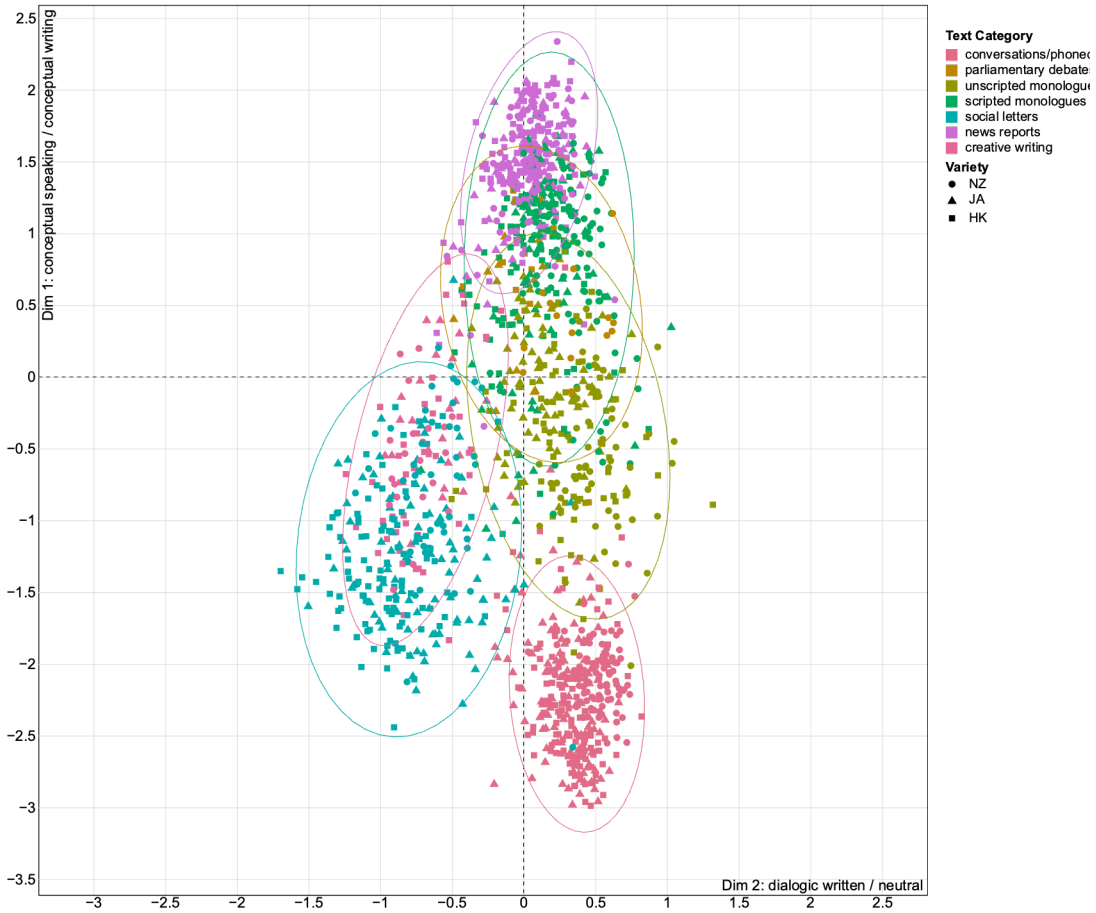
Funded by
DFG Deutsche Forschungsgemeinschaft German Research Foundation
FNSNF Swiss National Science Foundation
QuanTOR

Introduction

Background and Objective

- replicate and reproduce 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. informal spoken texts show differences by variety
    - » register-related patterns generally more pronounced
- extended data set: six additional components
  - → Canada, Great Britain, India, Ireland, Philippines, Singapore

- Is the register space and its interpretation stable?
  - → validate robustness of Geometric Multivariate Analysis (GMA)



Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]
FAU
RWTH AACHEN UNIVERSITY

Data set

Feature extraction

- consistent clean-up of component markup (Lehmann and Schneider 2012)
  - → esp. 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 outstanding

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

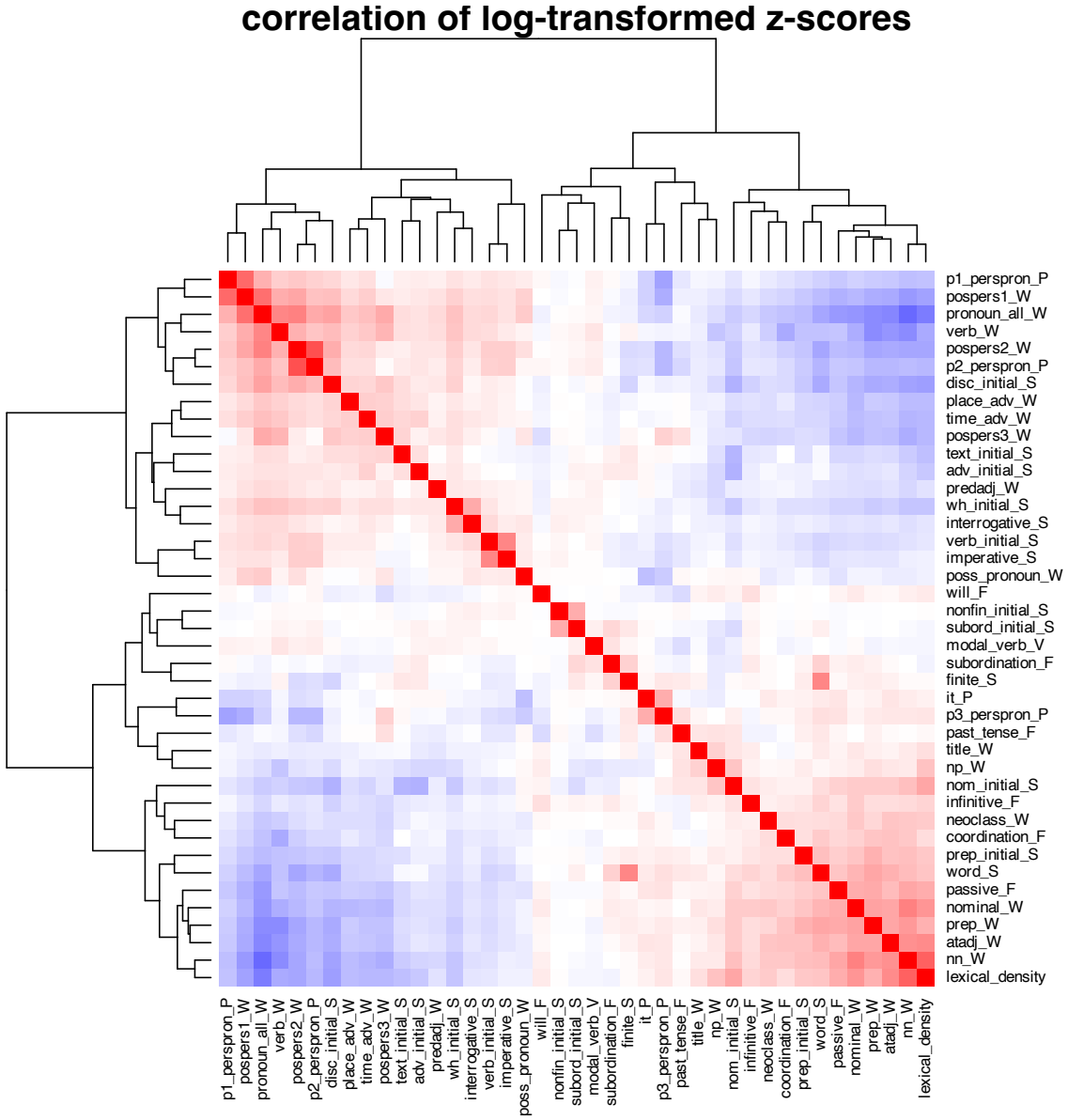
Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]
FAU
RWTH AACHEN UNIVERSITY

Data set

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 / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

RWTH AACHEN UNIVERSITY

Geometric Multivariate Analysis

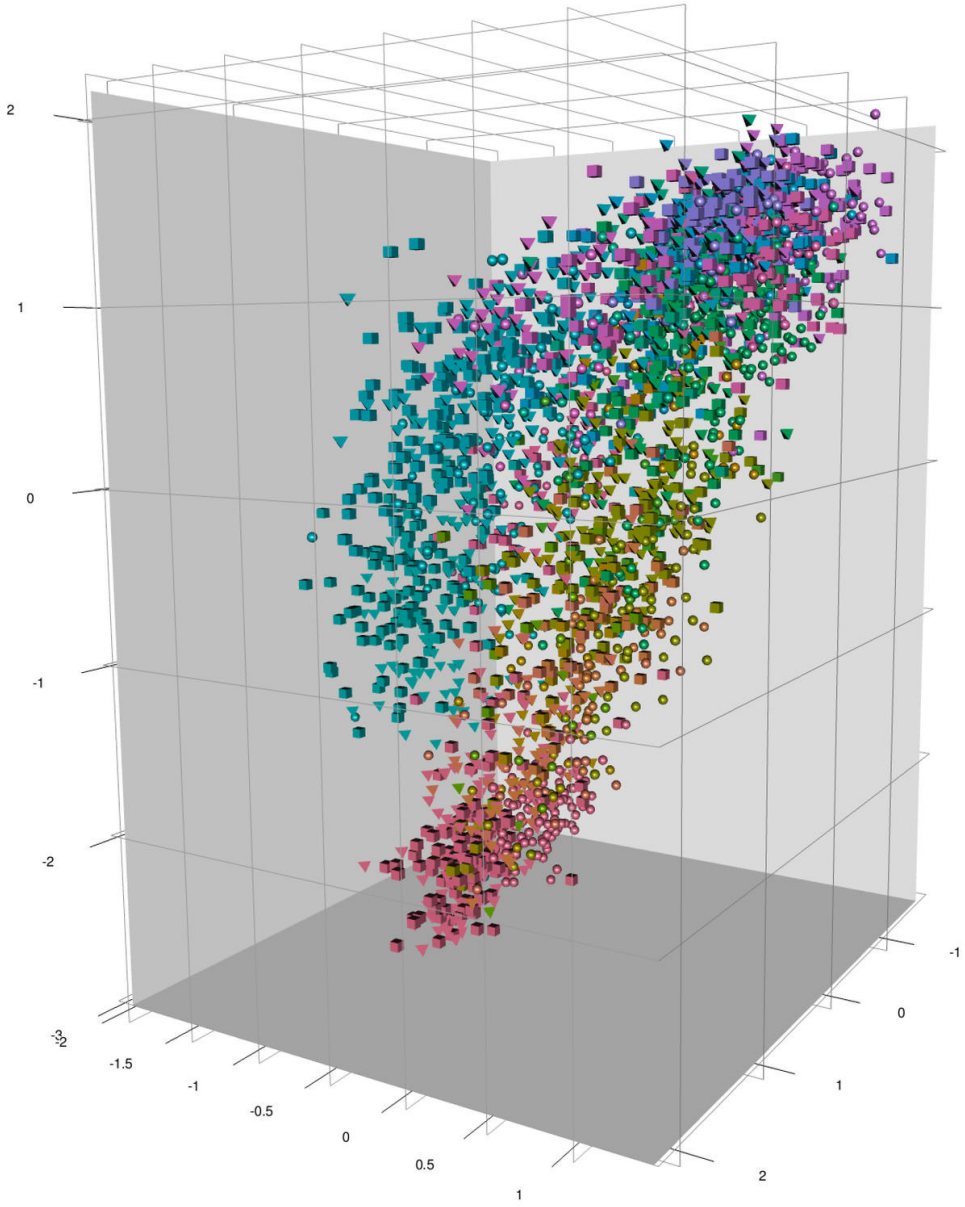
- Geometric multivariate analysis (GMA) developed in Diwersy et al. (2014) and Evert & Neumann (2017)
  → closely related to multidimensional analysis (MDA: Biber 1988, 1993, …), correspondence analysis, etc.

- Goal: geometric configurations in high-dimensional feature space → Euclidean distance as primary evidence
- Low-dimensional orthogonal projections preserve geometric structure (distances, angles, …)
- Decomposition into focus space (perspective) + orthogonal complement space (what is hidden from view)
- Central role of visualisation → intuitive interpretation complemented by feature weights
- Unsupervised perspective: principal component analysis (PCA), always used in complement space
- Minimally supervised perspective: linear discriminant analysis (LDA, also cf. Egbert & Biber 2018)

- Comes with user-friendly GMA tools in R (to be released this year) – see code examples at bottom of slides
- Further details in the reproduction materials at https://osf.io/8jzms/ (or use QR code on title slide)

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
- Projection into focus space spanned by first 4 LDA dimensions (in this talk: visualise only 3 dim's)
- Captures R^2 = 17.87% of total distance information



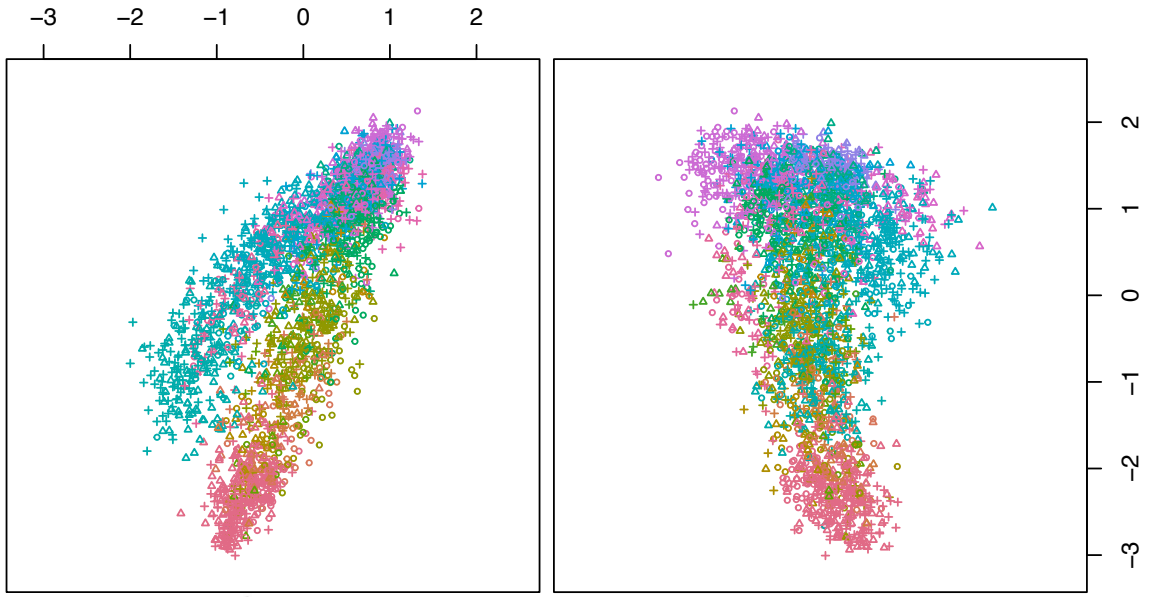

```
ICE3 <- GMA(ZL3)
GMA$add.discriminant(Meta3$textcat20, max.dim=4)
```

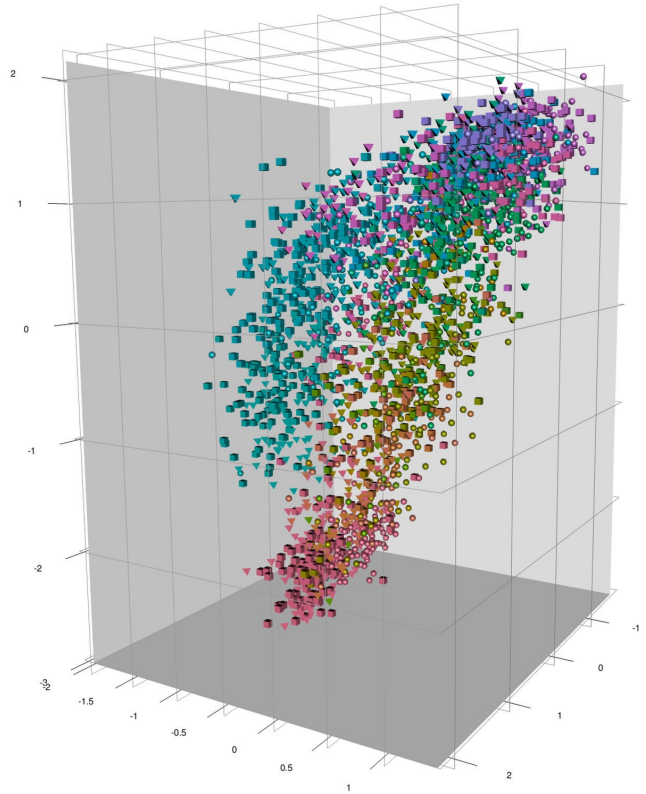

Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

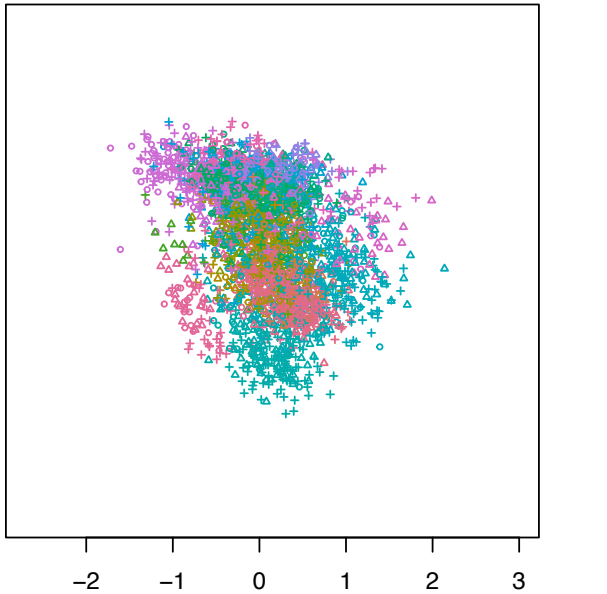
Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

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
- Projection into focus space spanned by first 4 LDA dimensions (in this talk: visualise only 3 dim's)
- Captures R^2 = 17.87% of total distance information
- Scatterplot matrix visualisation shows perspec views from the front, left and top of the cube
  - → generalises to $k > 3$ dimensions





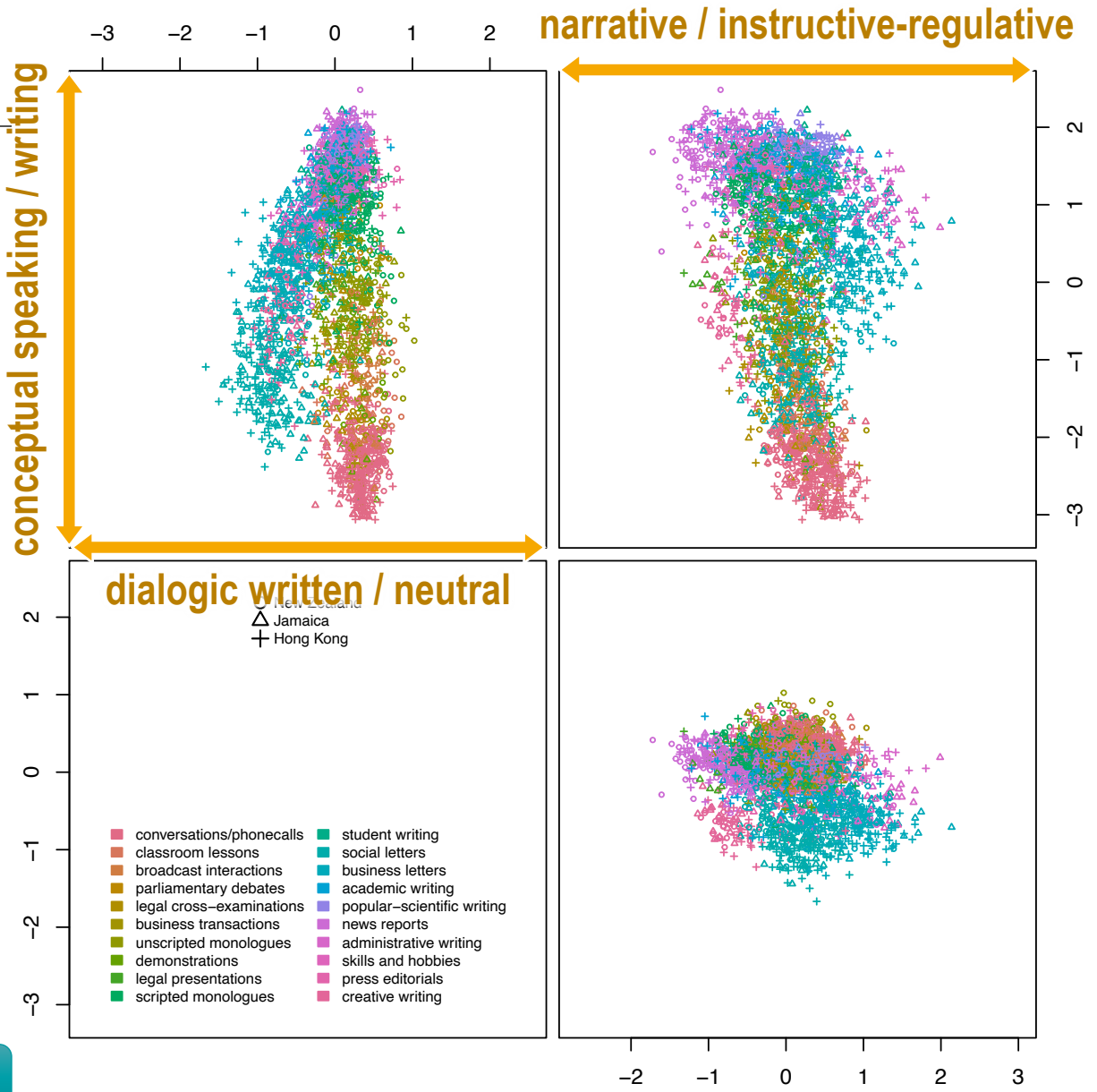



```
ICE3.X <- ICE3$projection()
gma.pairs(ICE3.X, dim=1:3, meta=Meta3,
          col=textcat20, pch=variety)
```

Reproducing Neumann & Evert (2021)

- Projection shows very clear structure, but the two “bananas” are not aligned with dimensions
- Factor analysis 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



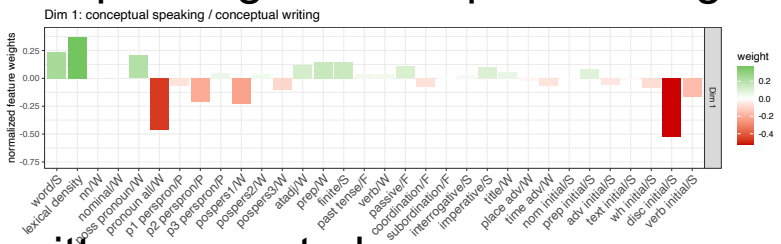

```
ICE3$rotation("pca", dim=1:2)
```


Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

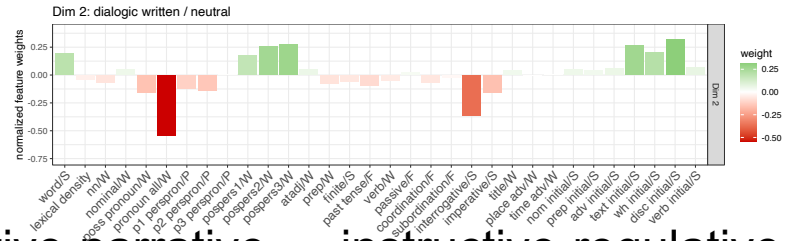
Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Reproducing Neumann & Evert (2021)

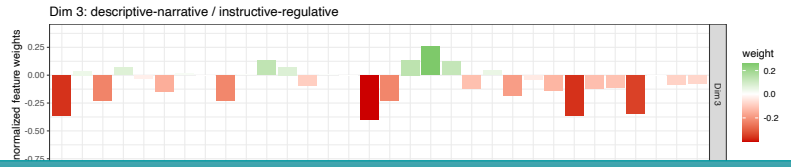
- Interpretation of dimensions based on topographic map created by ICE text categories, combined with feature weights of basis vectors (“loadings”)
- LD1: conceptual speaking — conceptual writing



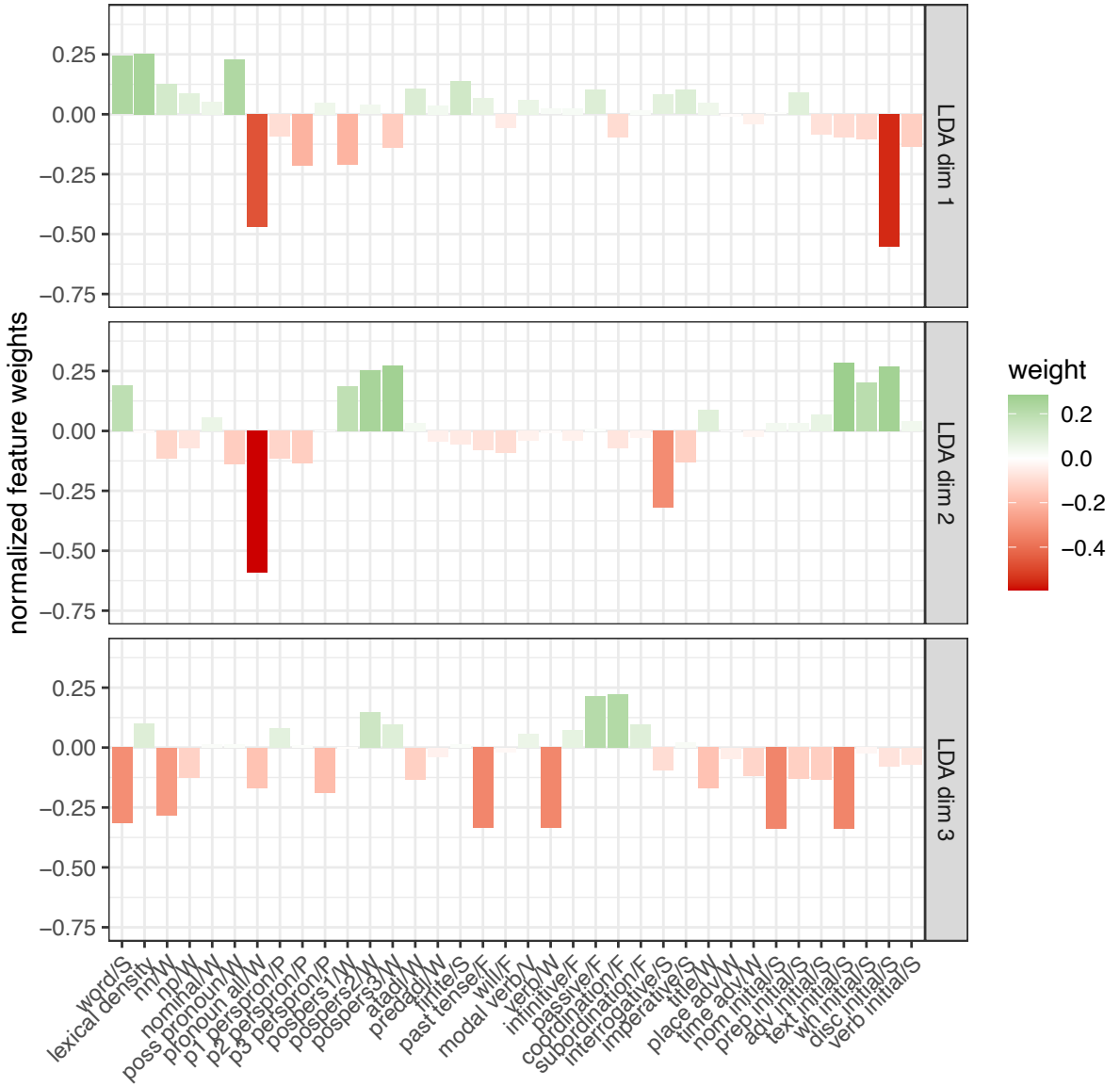
- LD2: dialogic written — neutral



- LD3: descriptive-narrative — instructive-regulative




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

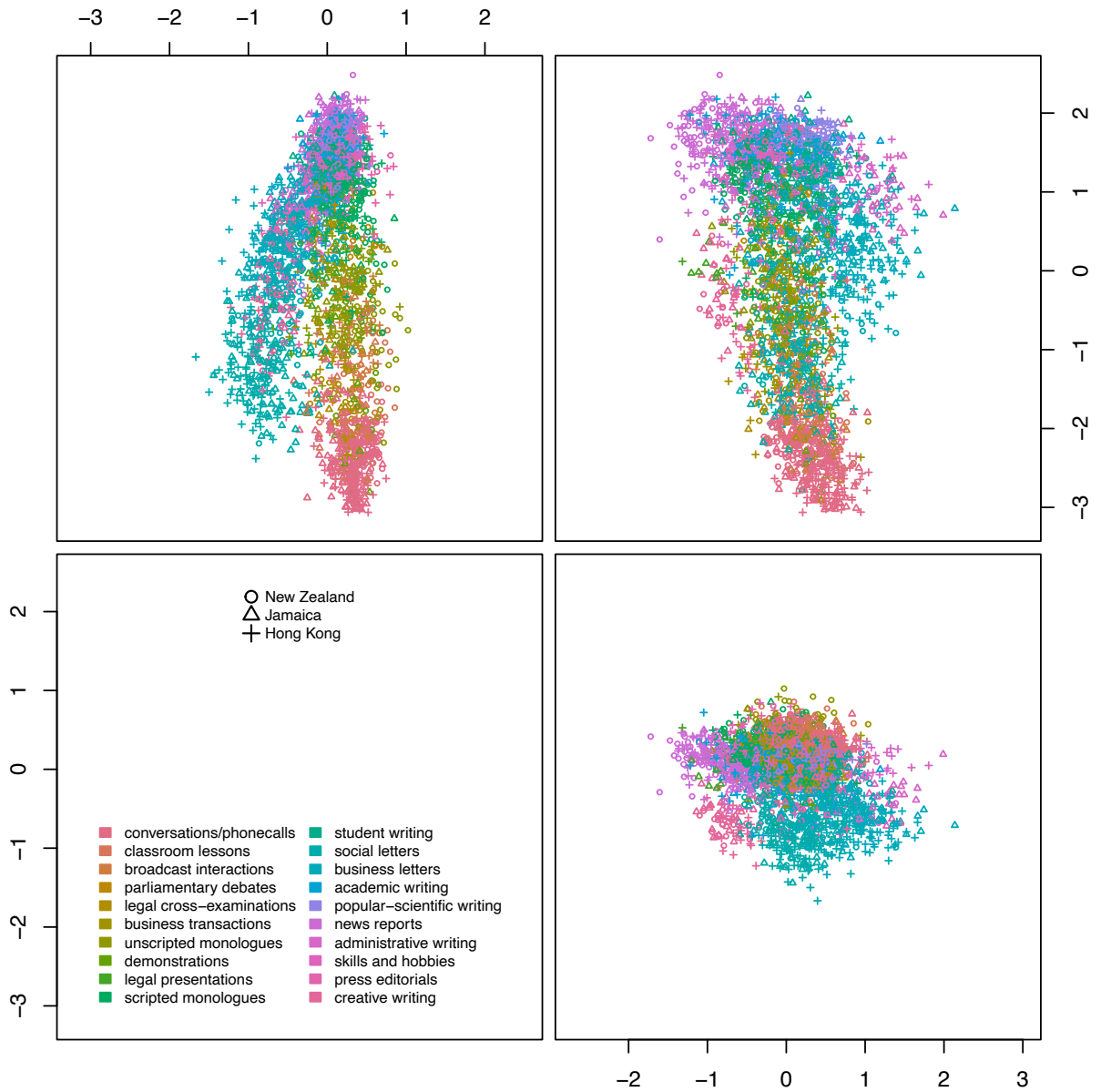


Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]
FAU
RWTH AACHEN UNIVERSITY

Reproducing Neumann & Evert (2021)

- Is there divergence between registers across the three language varieties?
  - → explored in terms of ICE text categories




```
gma.pairs(ICE3.X, 1:3, Meta=Meta3, pch=variety,
          col=textcat20)
```

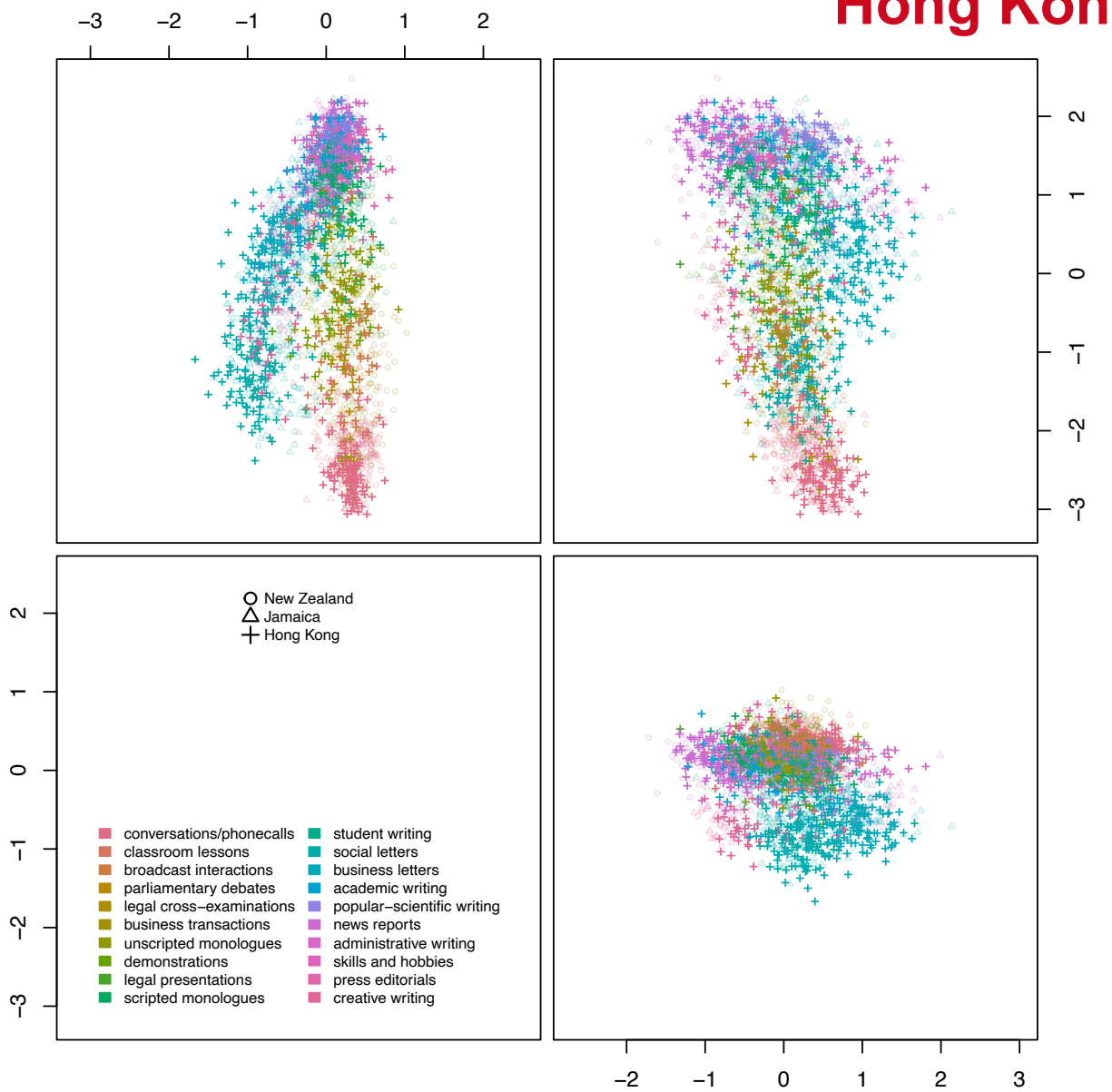

Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Reproducing Neumann & Evert (2021)

- Is there divergence between registers across the three language varieties?
  - → explored in terms of ICE text categories

```
gma.pairs(ICE3.X, 1:3, Meta=Meta3, pch=variety,
          col=textcat20, alpha.select=.2,
          select=(shortvar == "HK"))
```



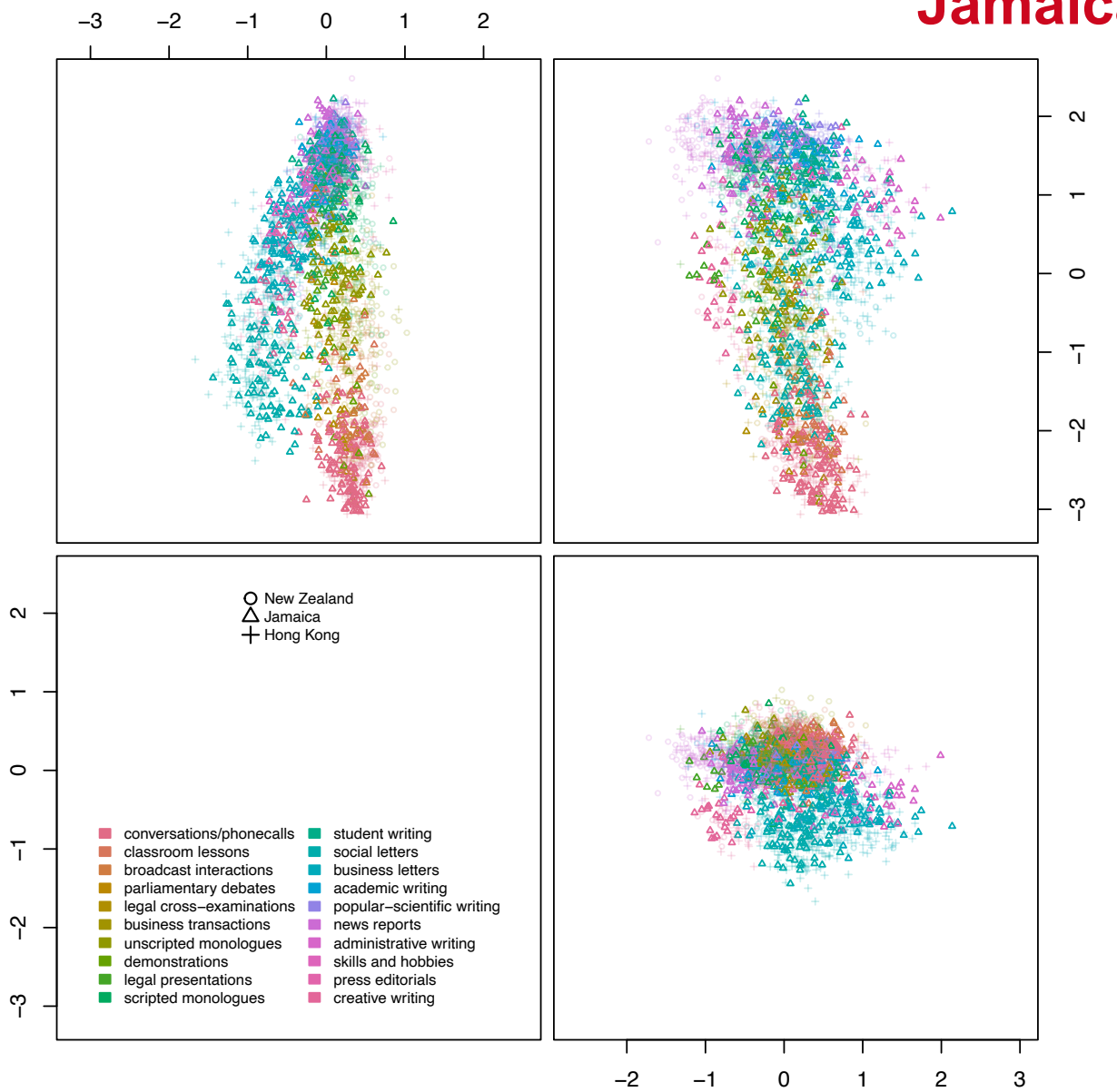
Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Reproducing Neumann & Evert (2021)

- Is there divergence between registers across the three language varieties?
  - → explored in terms of ICE text categories

```
gma.pairs(ICE3.X, 1:3, Meta=Meta3, pch=variety,
          col=textcat20, alpha.select=.2,
          select=(shortvar == "JAM"))
```

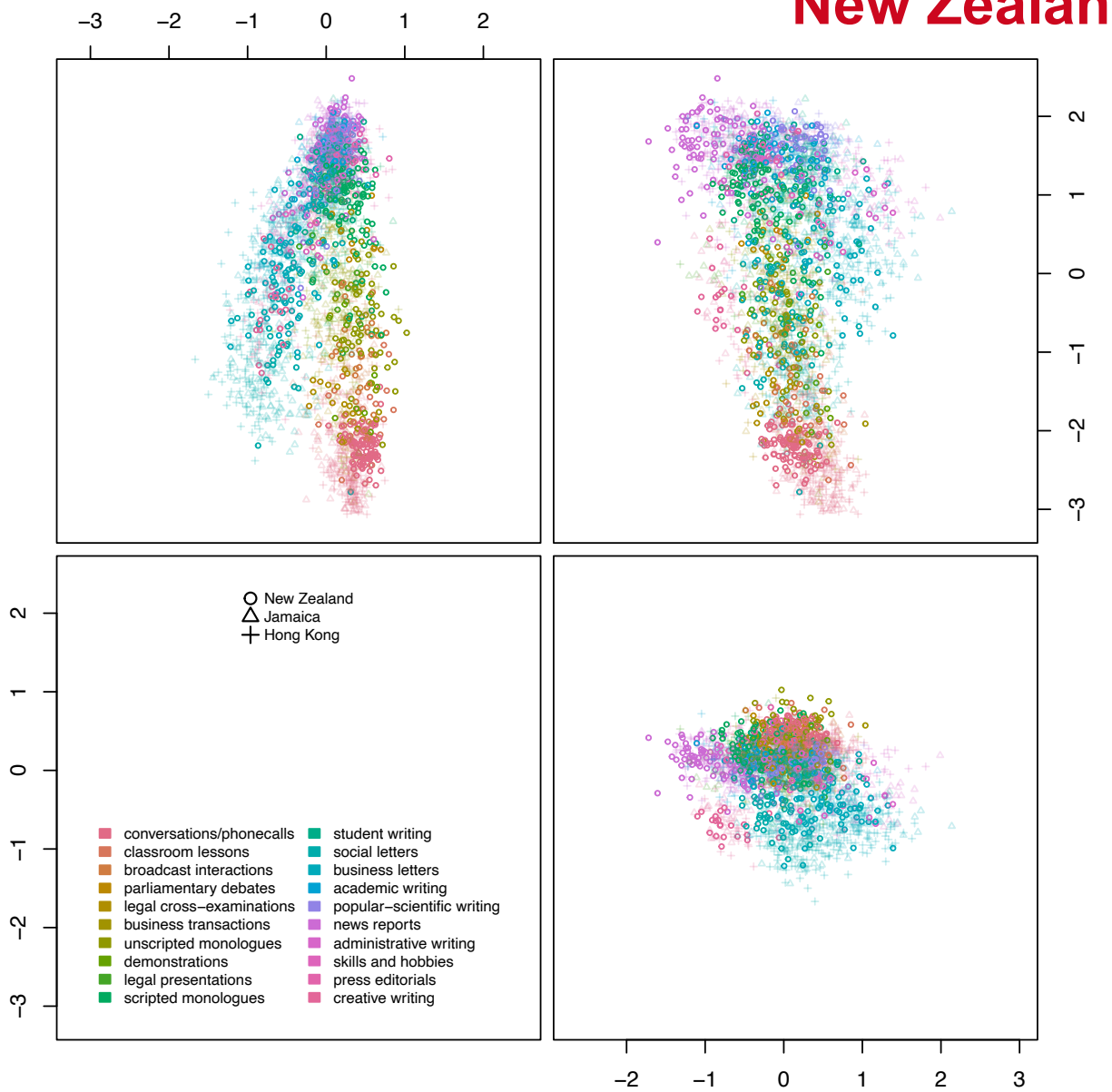


Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Reproducing Neumann & Evert (2021)

- Is there divergence between registers across the three language varieties?
  - → explored in terms of ICE text categories




```
gma.pairs(ICE3.X, 1:3, Meta=Meta3, pch=variety,
          col=textcat20, alpha.select=.2,
          select=(shortvar == "NZ"))
```

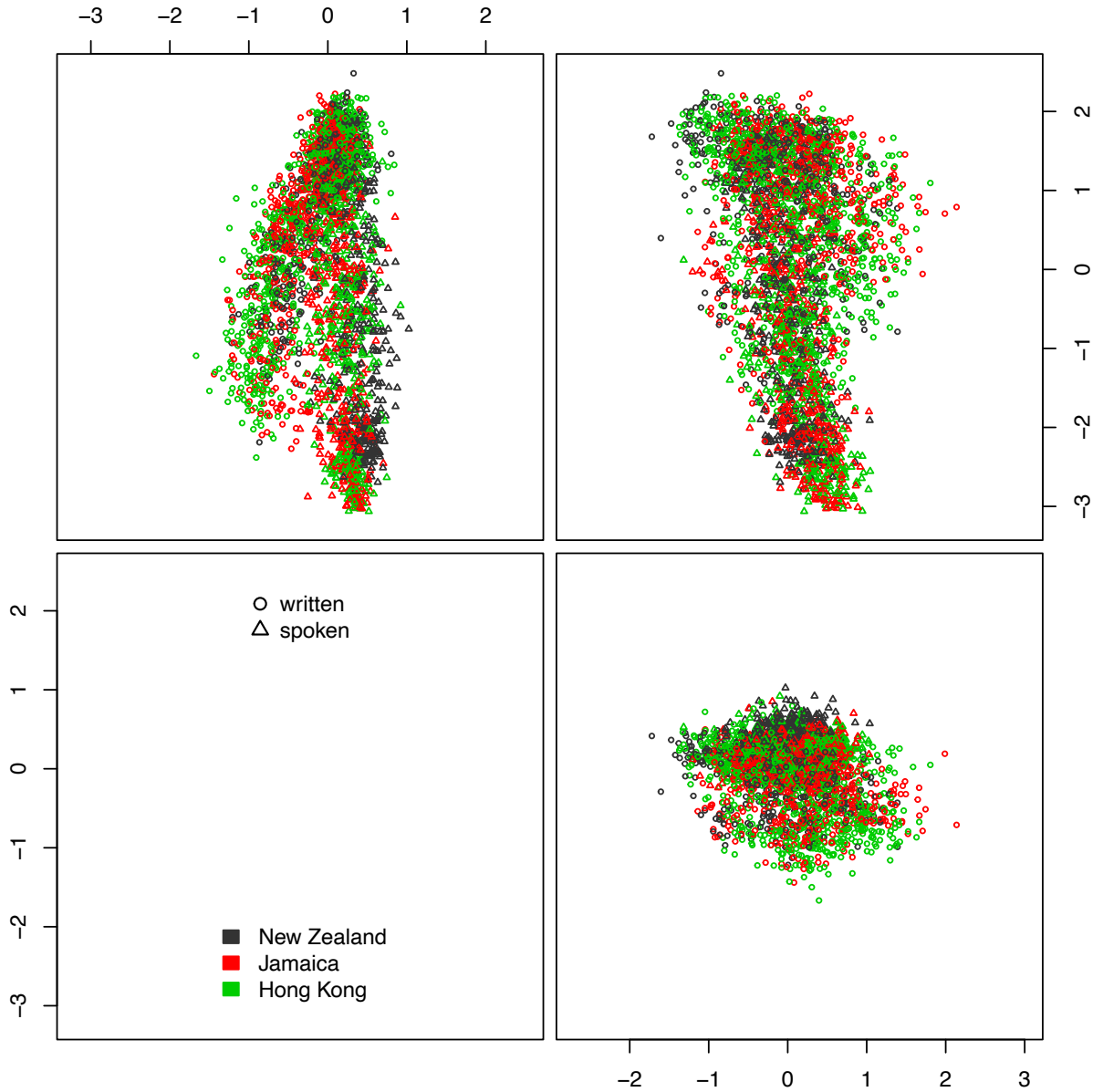

Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Replication experiment (v1)

- Replication of Neumann & Evert (2001):
  Will we come to the same conclusions if we carry out the analysis on a different data set?
- Here: extension to nine varieties of English (ICE9)

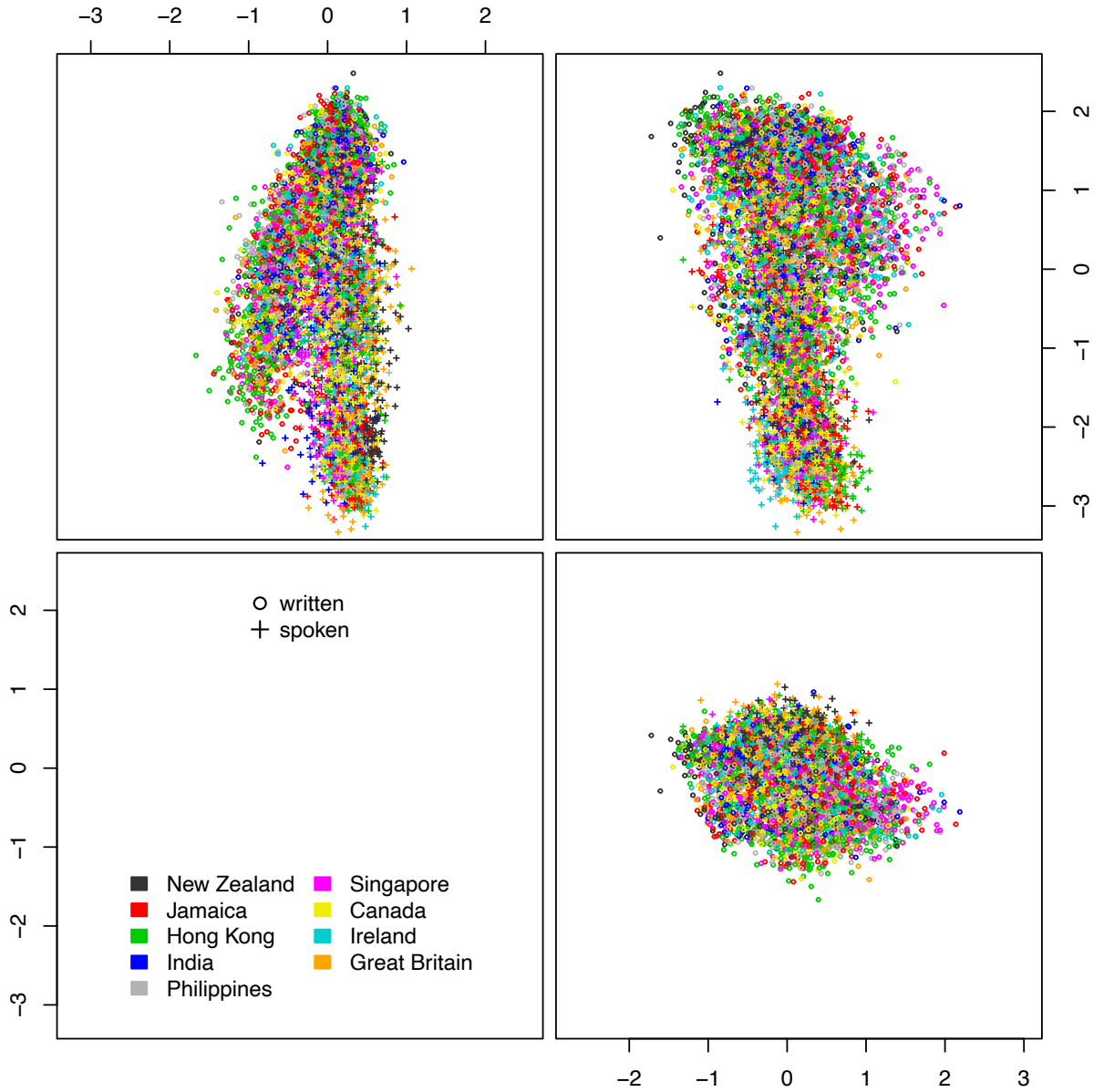
- Version 1: Do the other language varieties exhibit different patterns of register divergence, or do our observations from ICE3 remain valid?
- Project additional texts into the ICE3 focus space



Replication experiment (v1)

- Replication of Neumann & Evert (2001):
  Will we come to the same conclusions if we carry out the analysis on a different data set?
- Here: extension to nine varieties of English (ICE9)

- Version 1: Do the other language varieties exhibit different patterns of register divergence, or do our observations from ICE3 remain valid?
- Project additional texts into the ICE3 focus space




```
ICE3.X9 <- ICE3$projection("space", M=ZL)
```


Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

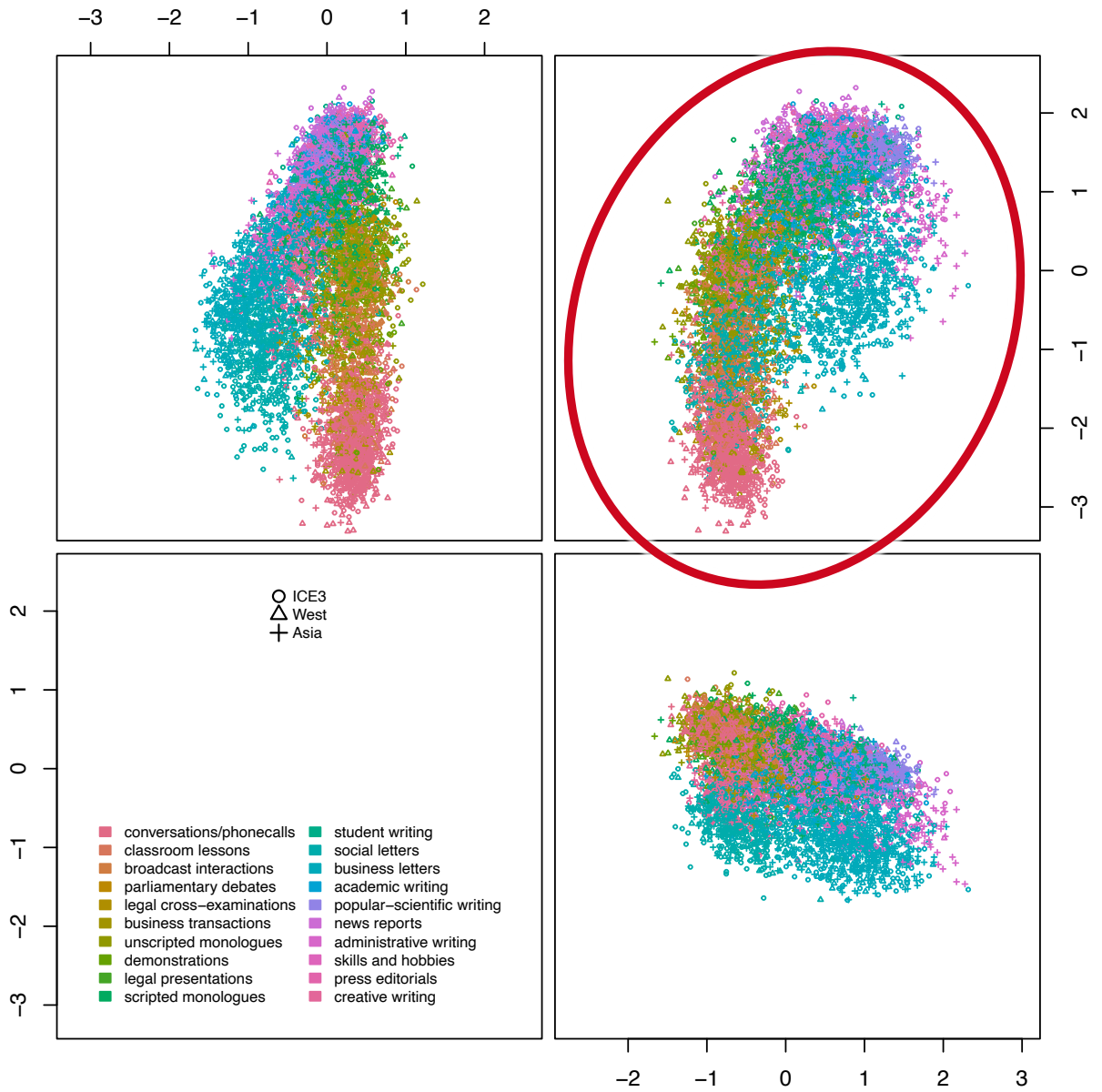
Replication experiment (v2)

- Replication of Neumann & Evert (2001):
  Will we come to the same conclusions if we carry out the analysis on a different data set?
- Here: extension to nine varieties of English (ICE9)

- Version 2: Do we obtain a comparable focus space if we carry out the LDA across all ICE9 varieties?
  Would we interpret dimensions in the same way?

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

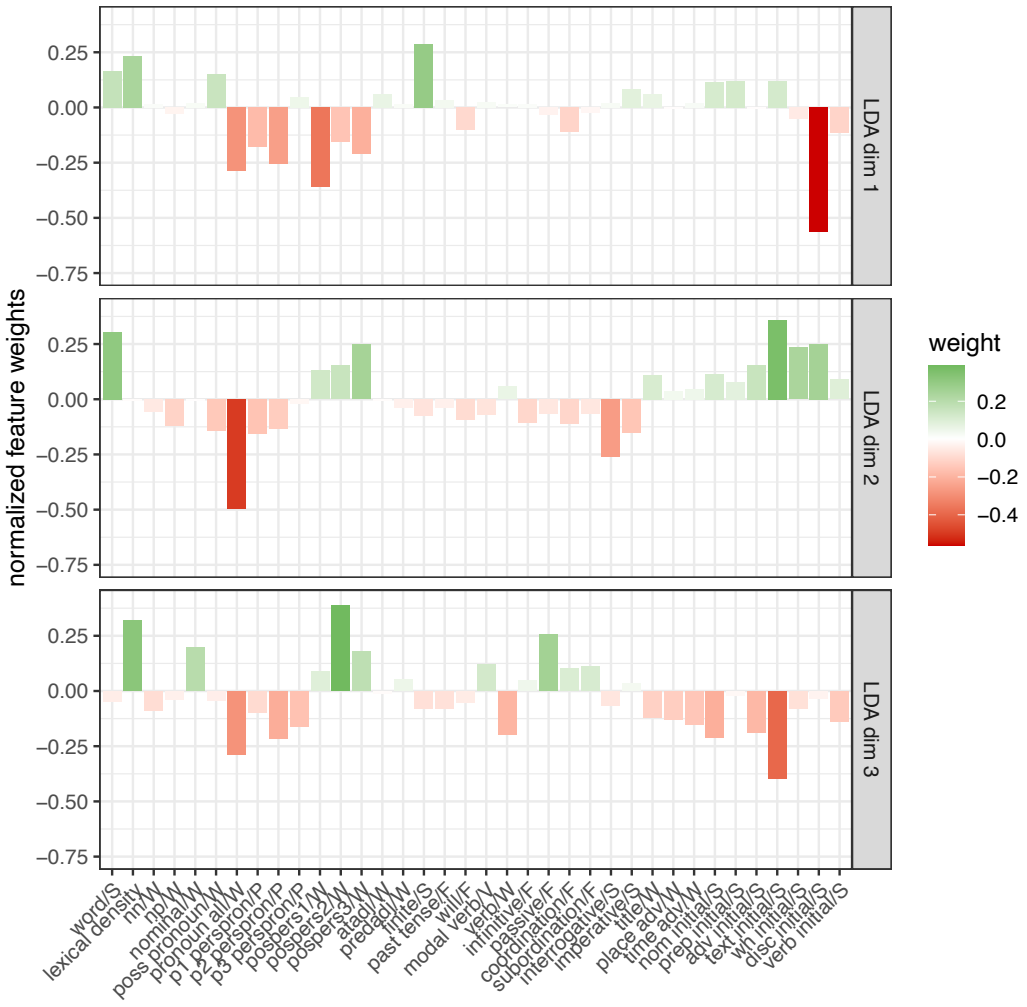
```
ICE9 <- GMA(ZL)
GMA$add.discriminant(Meta$textcat20, max.dim=4)
```

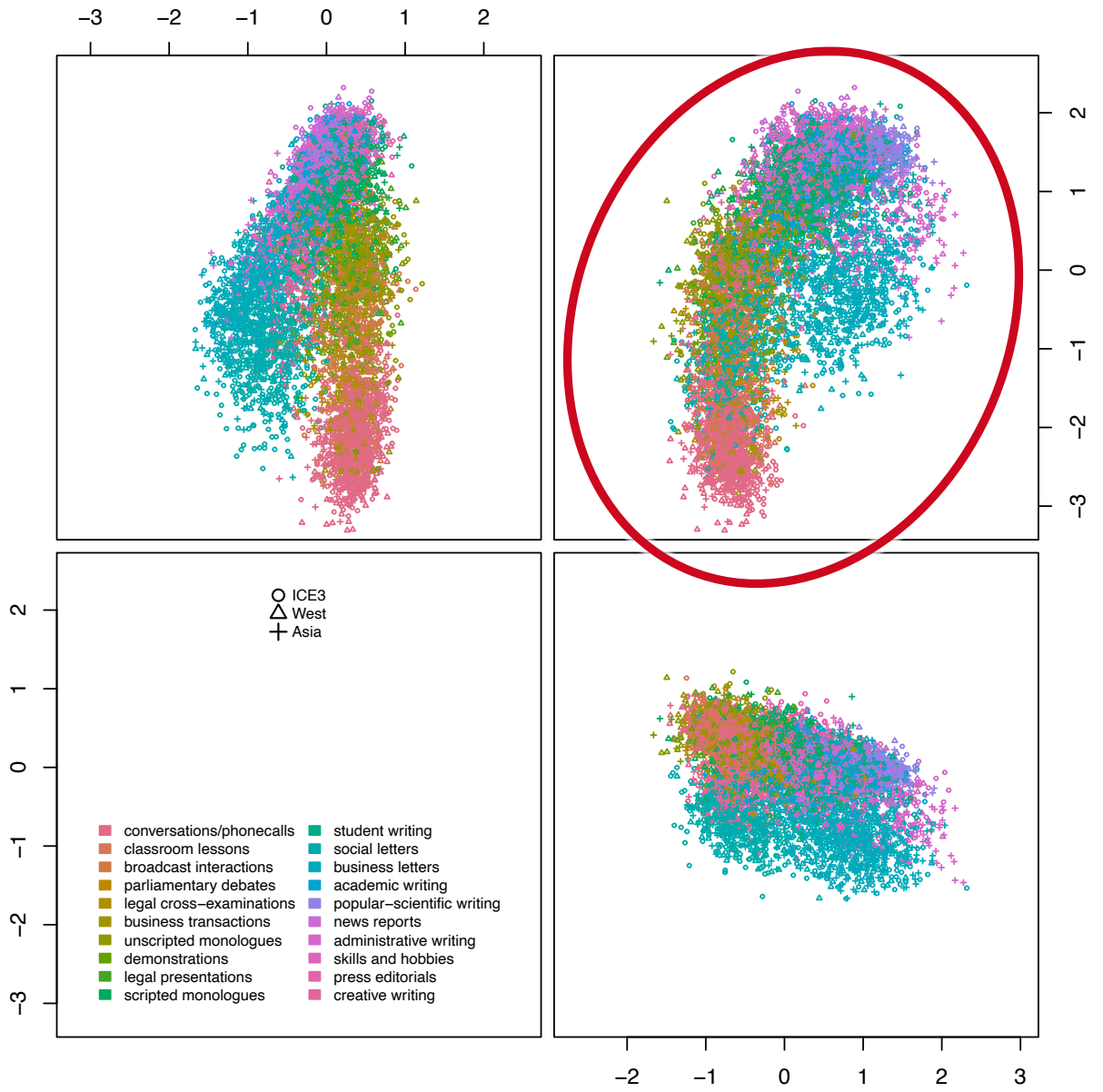


Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]
FAU
RWTH AACHEN UNIVERSITY

Replication experiment (v2)





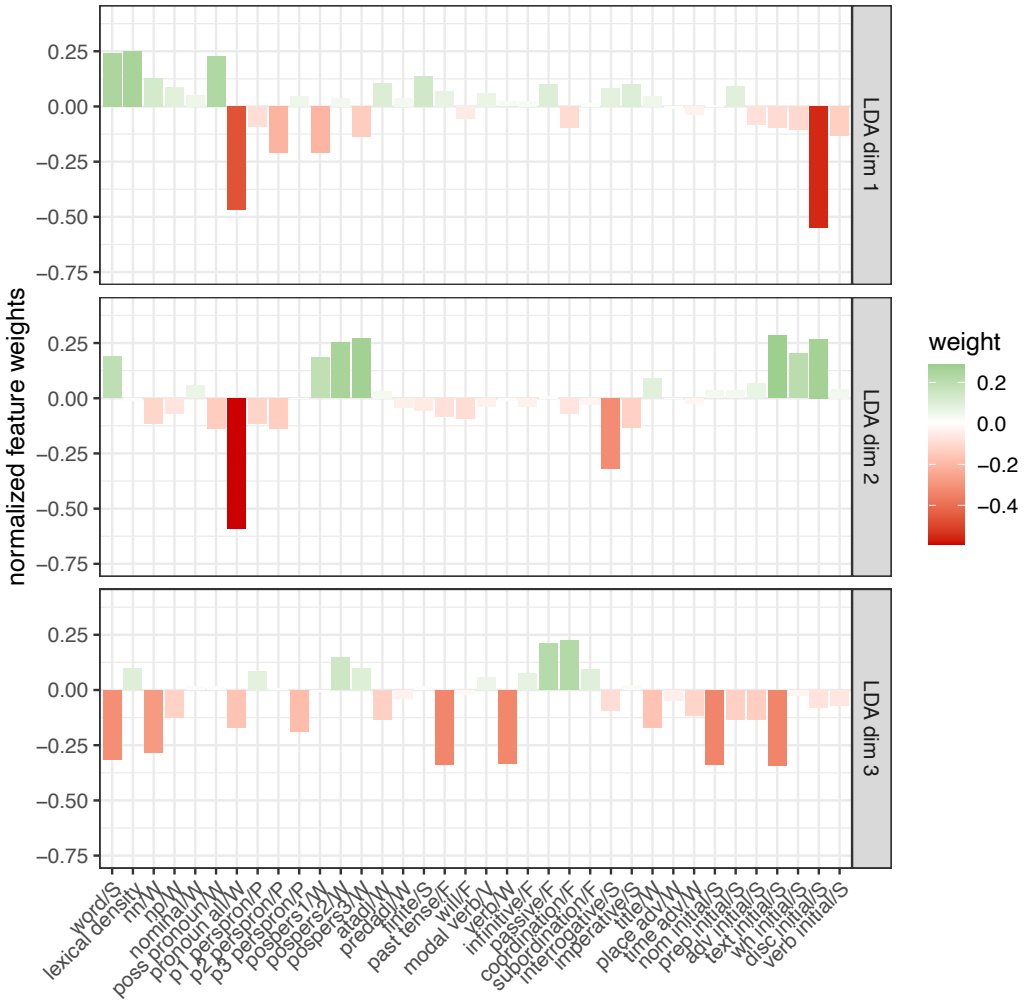
Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

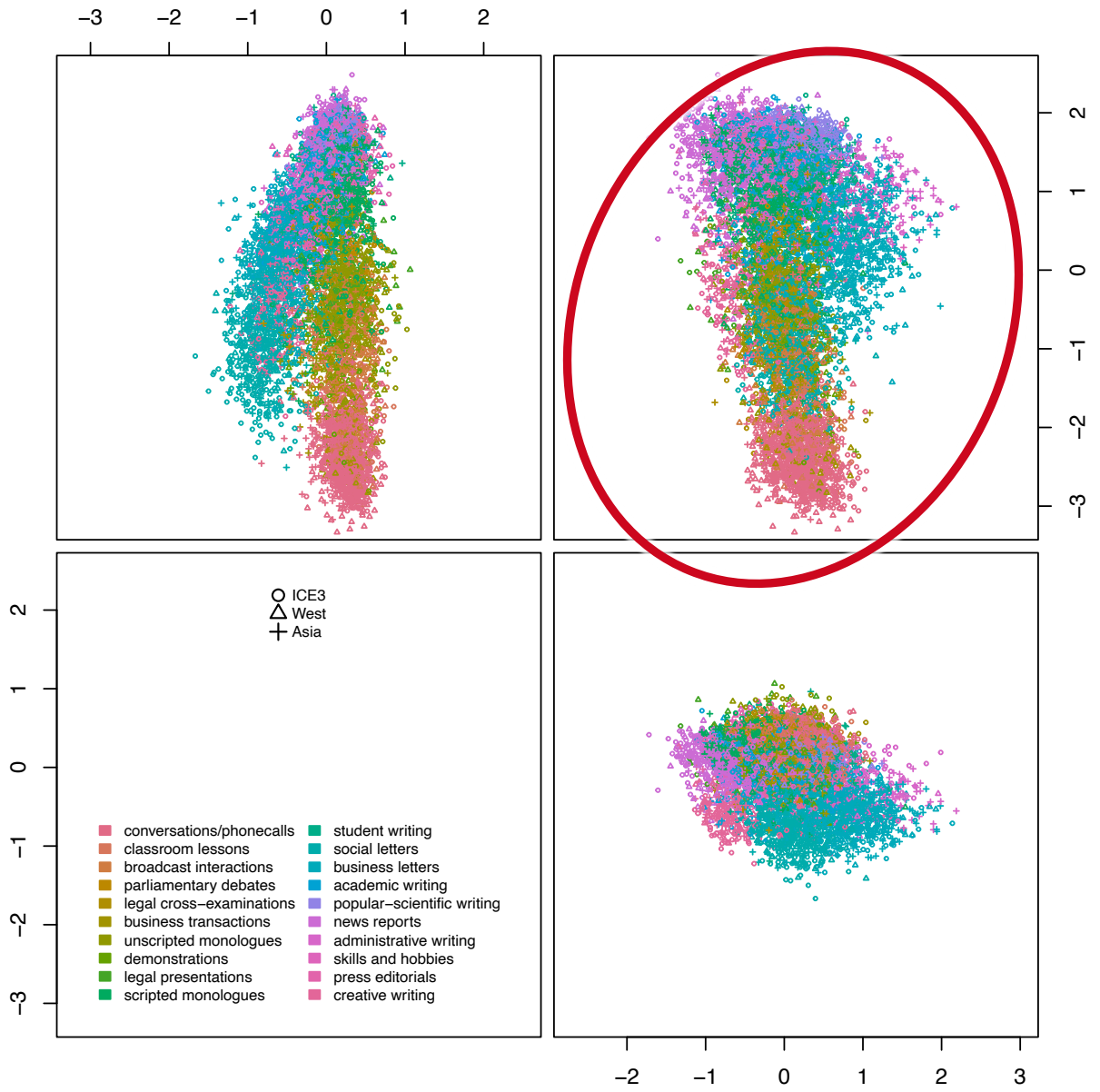
Universität Zürich[UZH]

FAU

RWTH AACHEN UNIVERSITY

Replication experiment (v2)





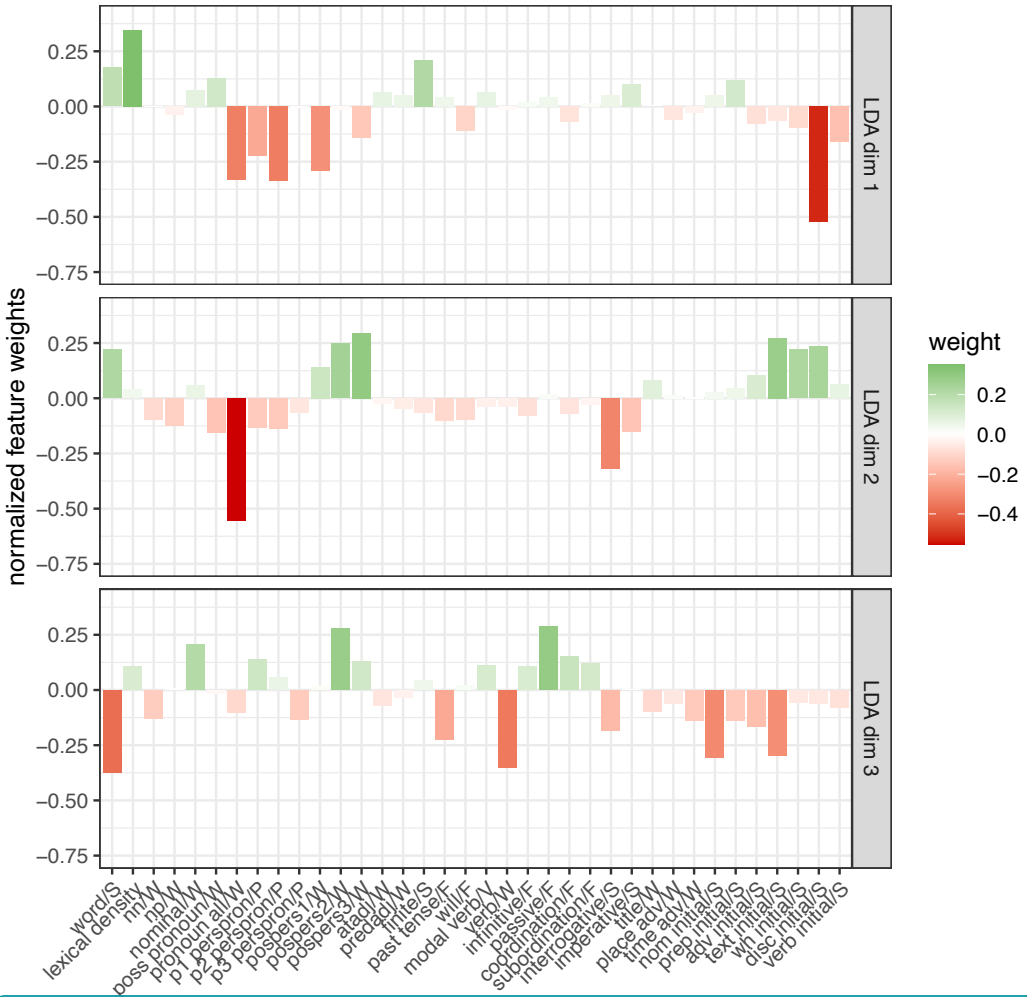
Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

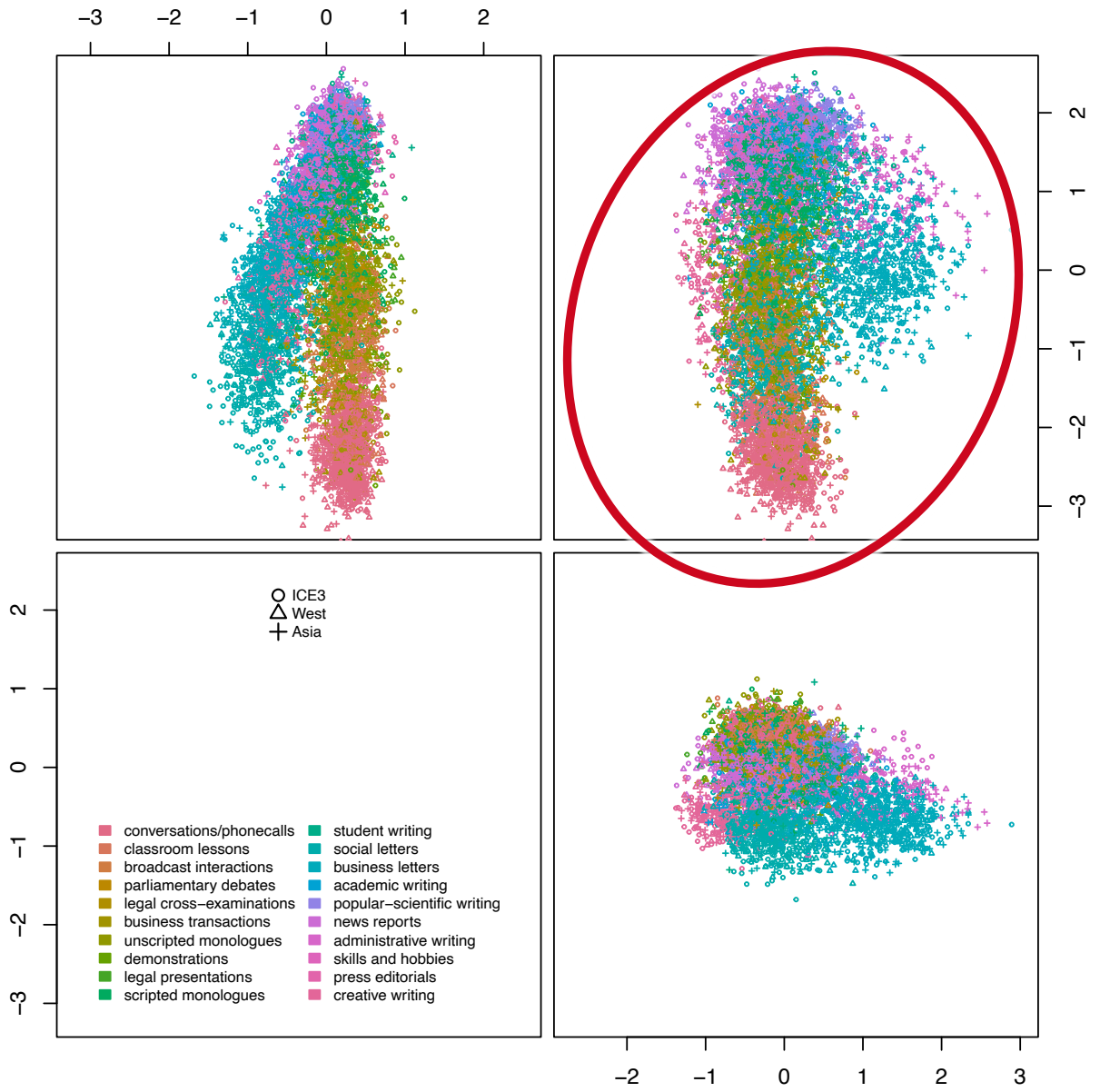
Universität Zürich[UZH]

FAU

RWTH AACHEN UNIVERSITY

Aligning the GMA focus spaces






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


Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

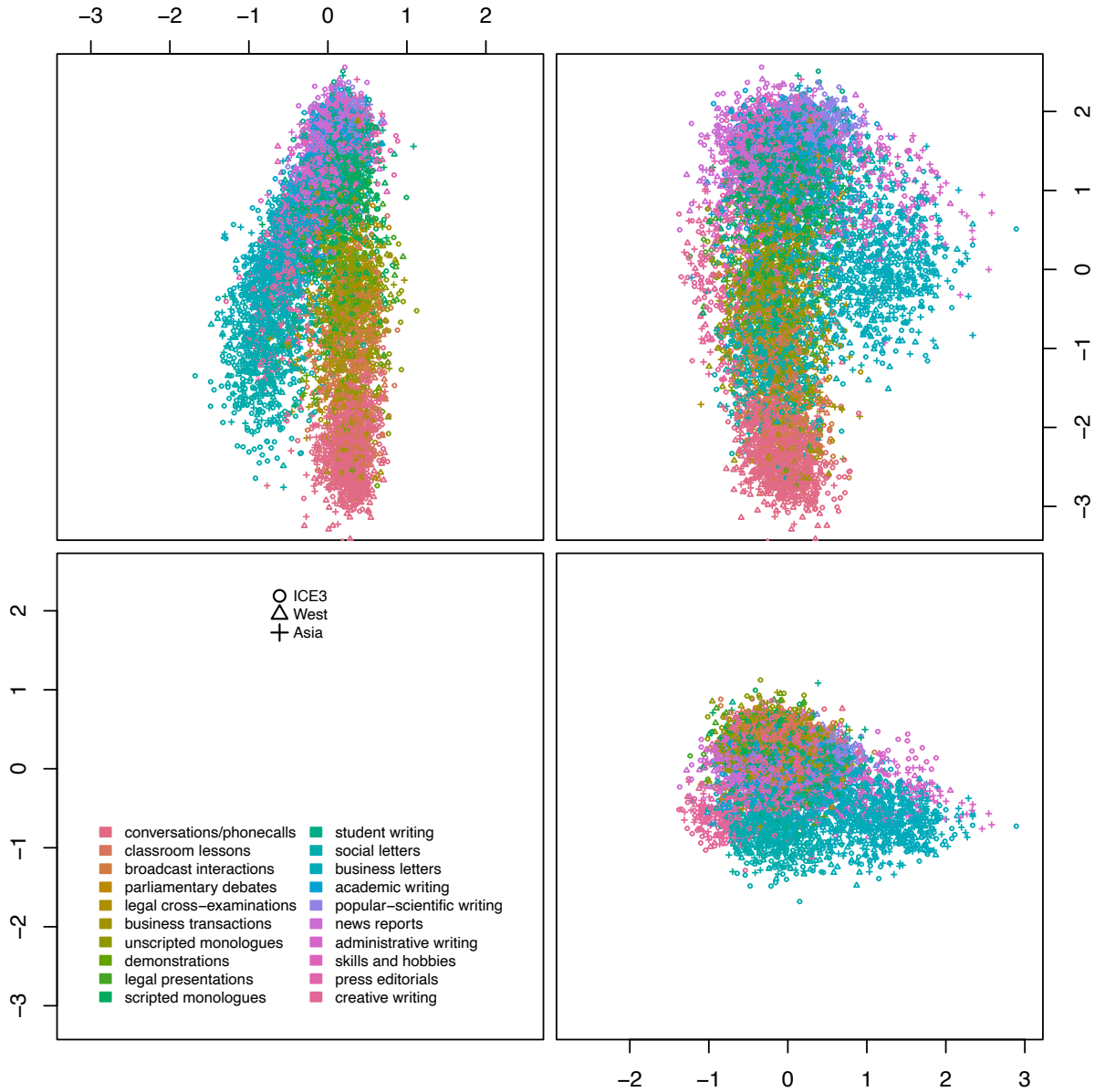
Universität Zürich[UZH]

FAU

RWTH AACHEN UNIVERSITY

A little bit of linguistics …

- ICAME talk needs to include a perfunctory linguistic analysis at the very least …
- 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)

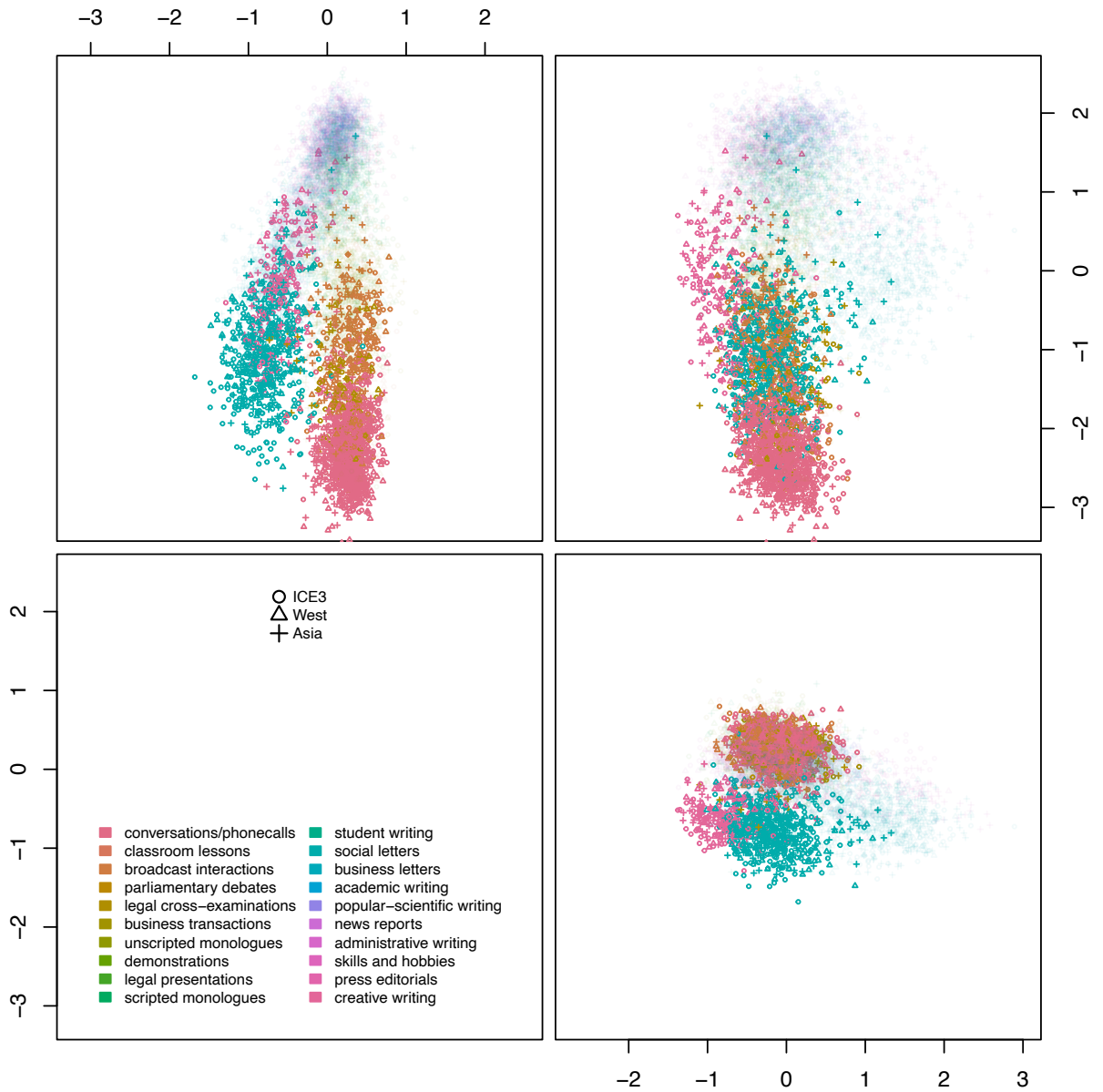


Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]
FAU
RWTH AACHEN UNIVERSITY

A little bit of linguistics …

- ICAME talk needs to include a perfunctory linguistic analysis at the very least …
- 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)



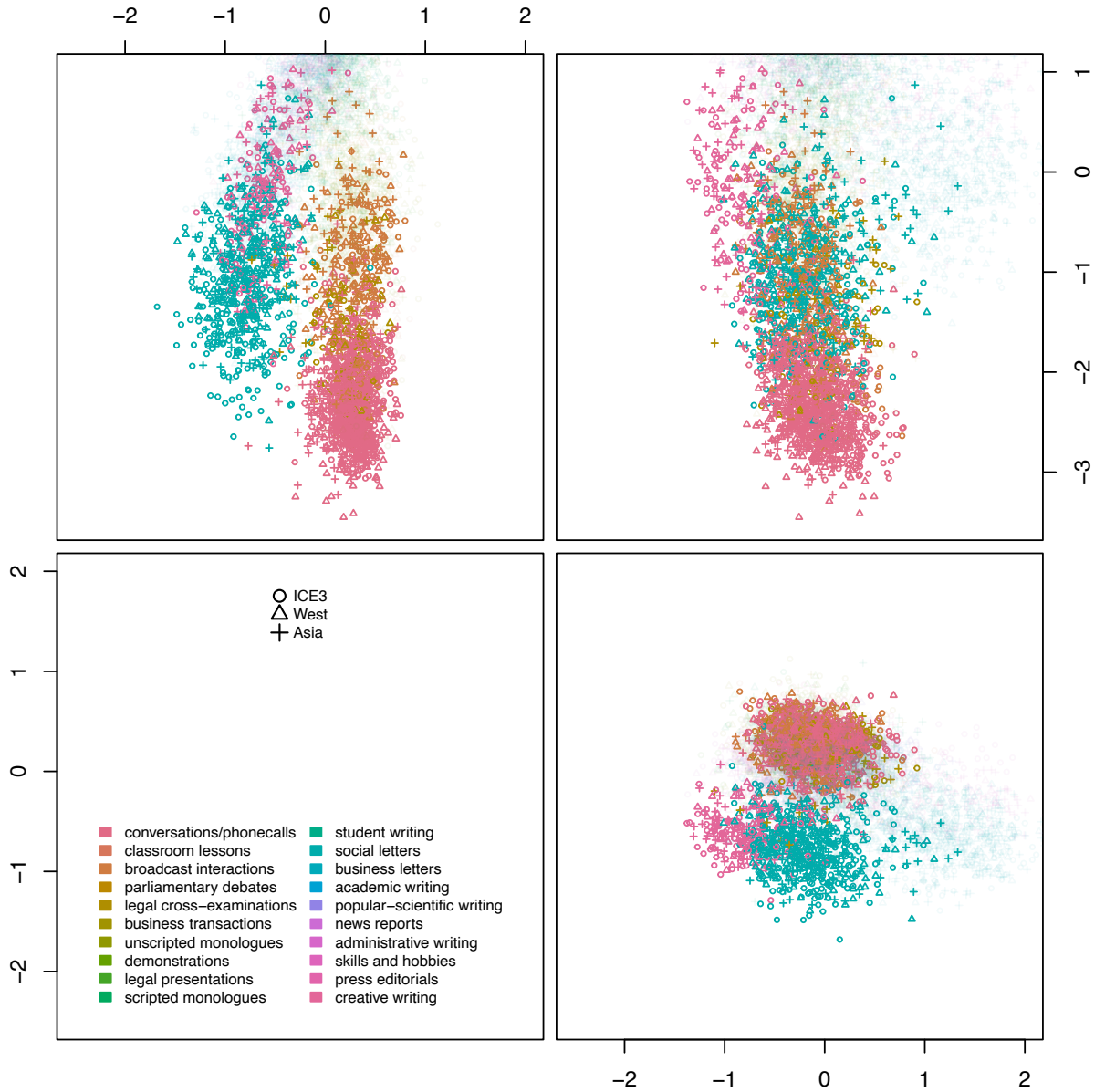

```
idx.sub <- Meta$short20 %in% qw("conv broadc …")
MetaSub <- droplevels(Meta[idx.sub, ])
ICE9.Sub <- ICE9.X[idx.sub, ]
```


Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

A little bit of linguistics …

- ICAME talk needs to include a perfunctory linguistic analysis at the very least …
- 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)



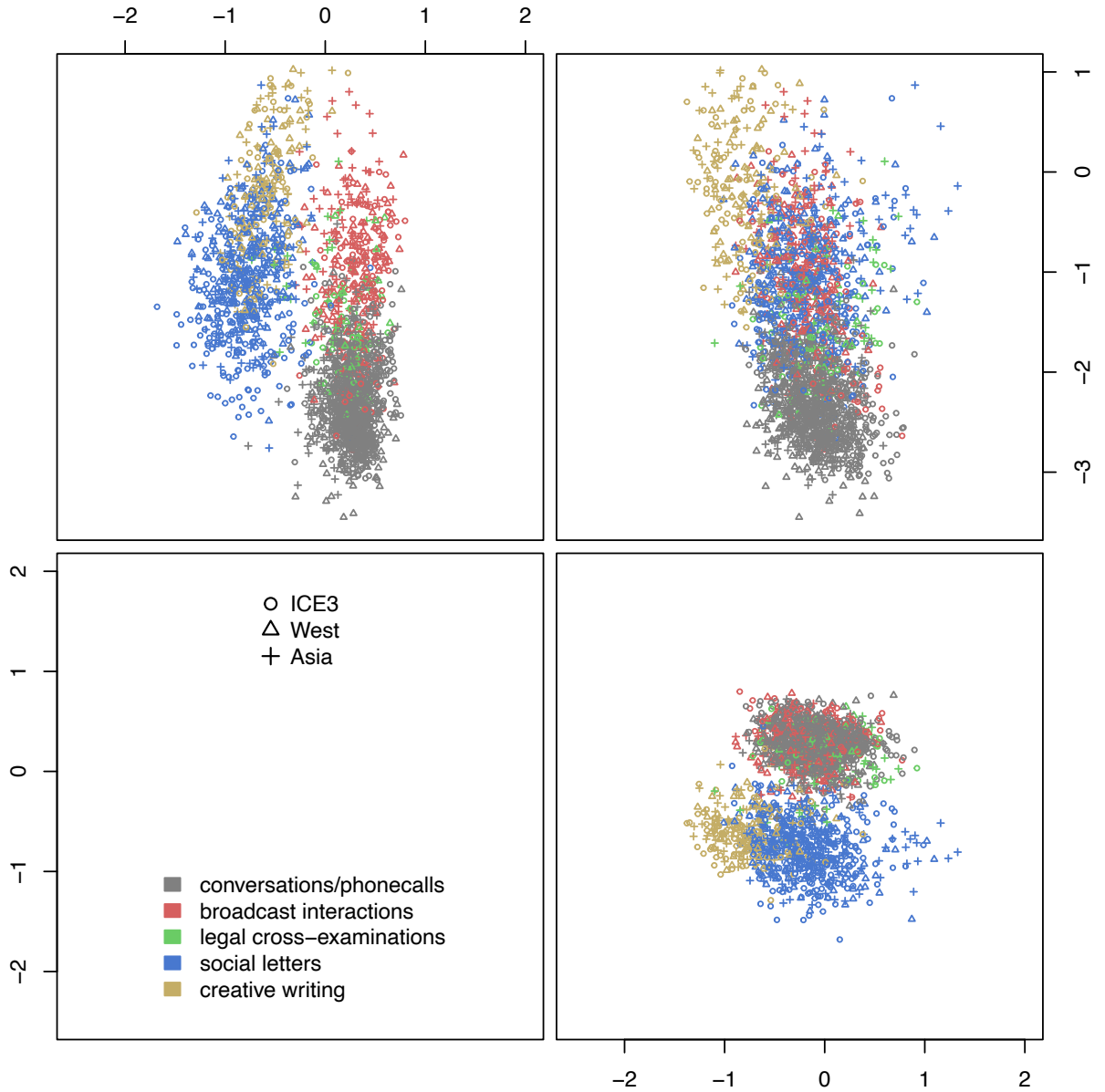

```
idx.sub <- Meta$short20 %in% qw("conv broadc …")
MetaSub <- droplevels(Meta[idx.sub, ])
ICE9.Sub <- ICE9.X[idx.sub, ]
```


Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

A little bit of linguistics …

- ICAME talk needs to include a perfunctory linguistic analysis at the very least …
- 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)
- New colour scheme helps to distinguish text categories more clearly

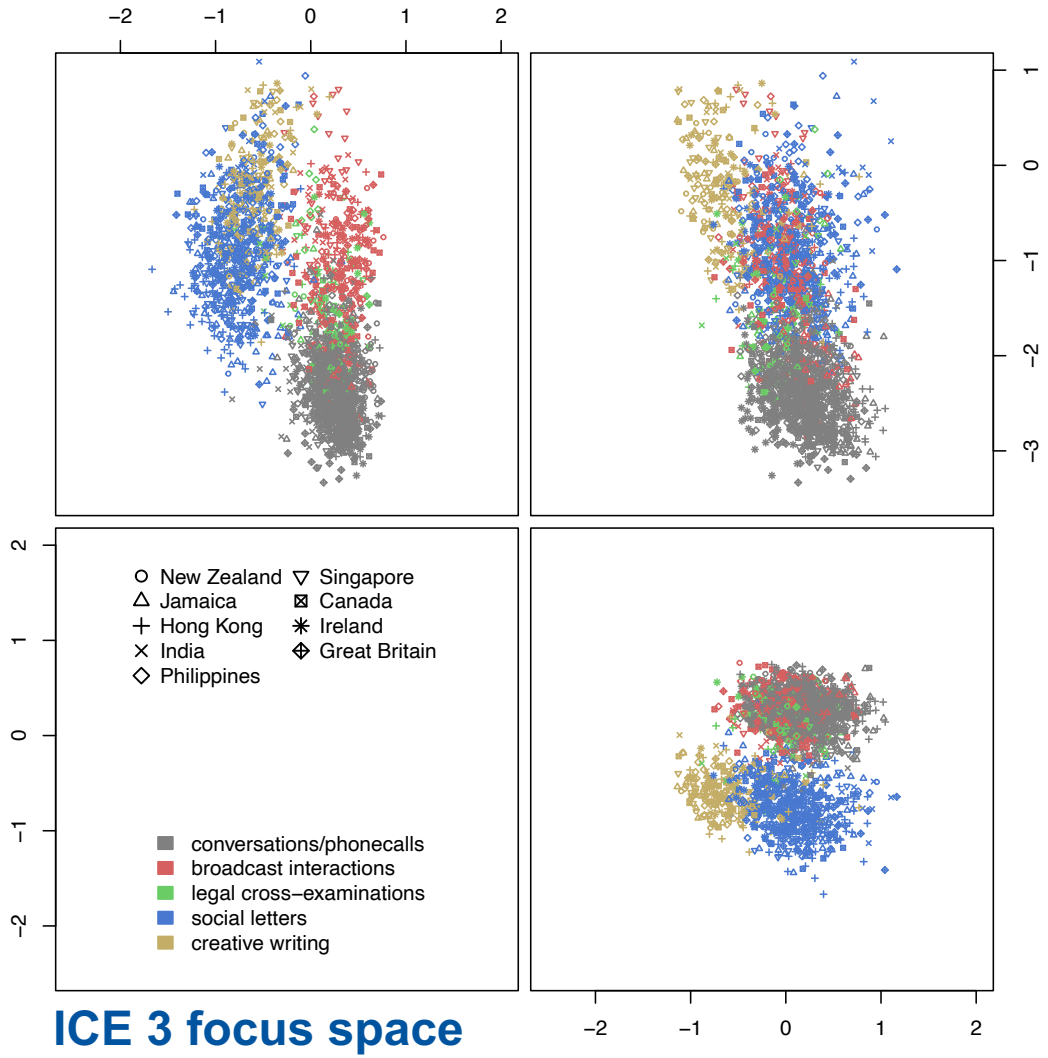



```
idx.sub <- Meta$short20 %in% qw("conv broadc …")
MetaSub <- droplevels(Meta[idx.sub, ])
ICE9.Sub <- ICE9.X[idx.sub, ]
```

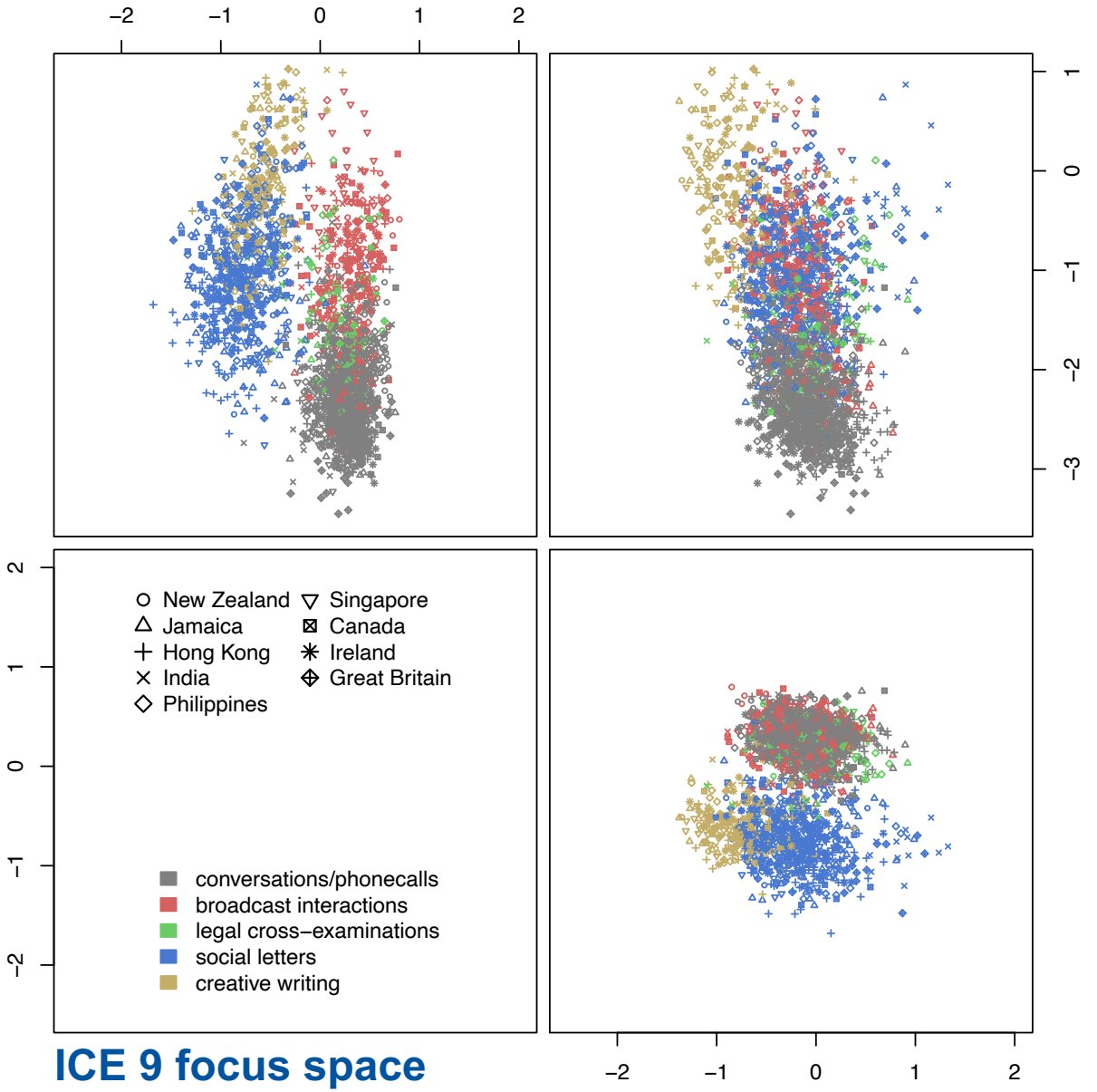

Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH] FAU RWTH AACHEN UNIVERSITY

Register divergence across varieties



ICE 3 focus space



ICE 9 focus space

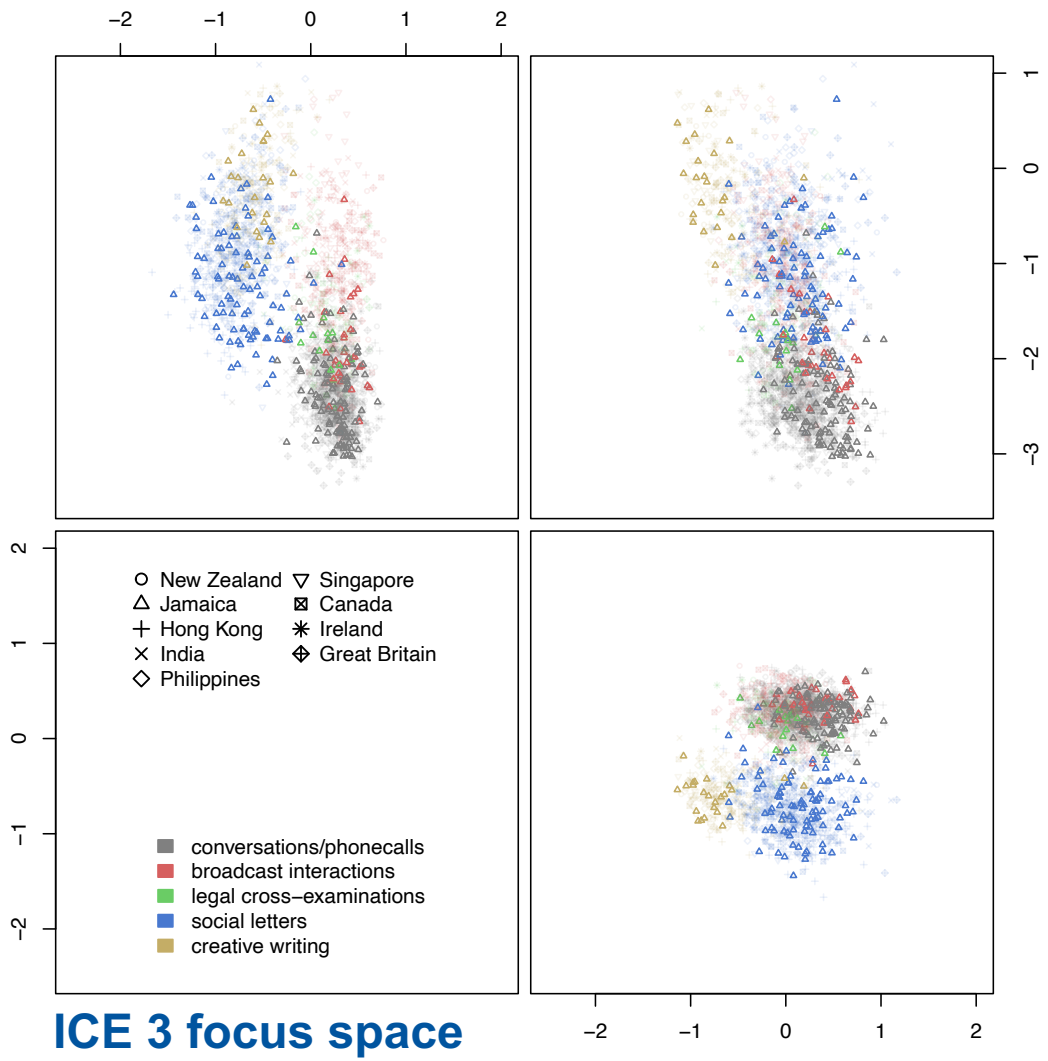
Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

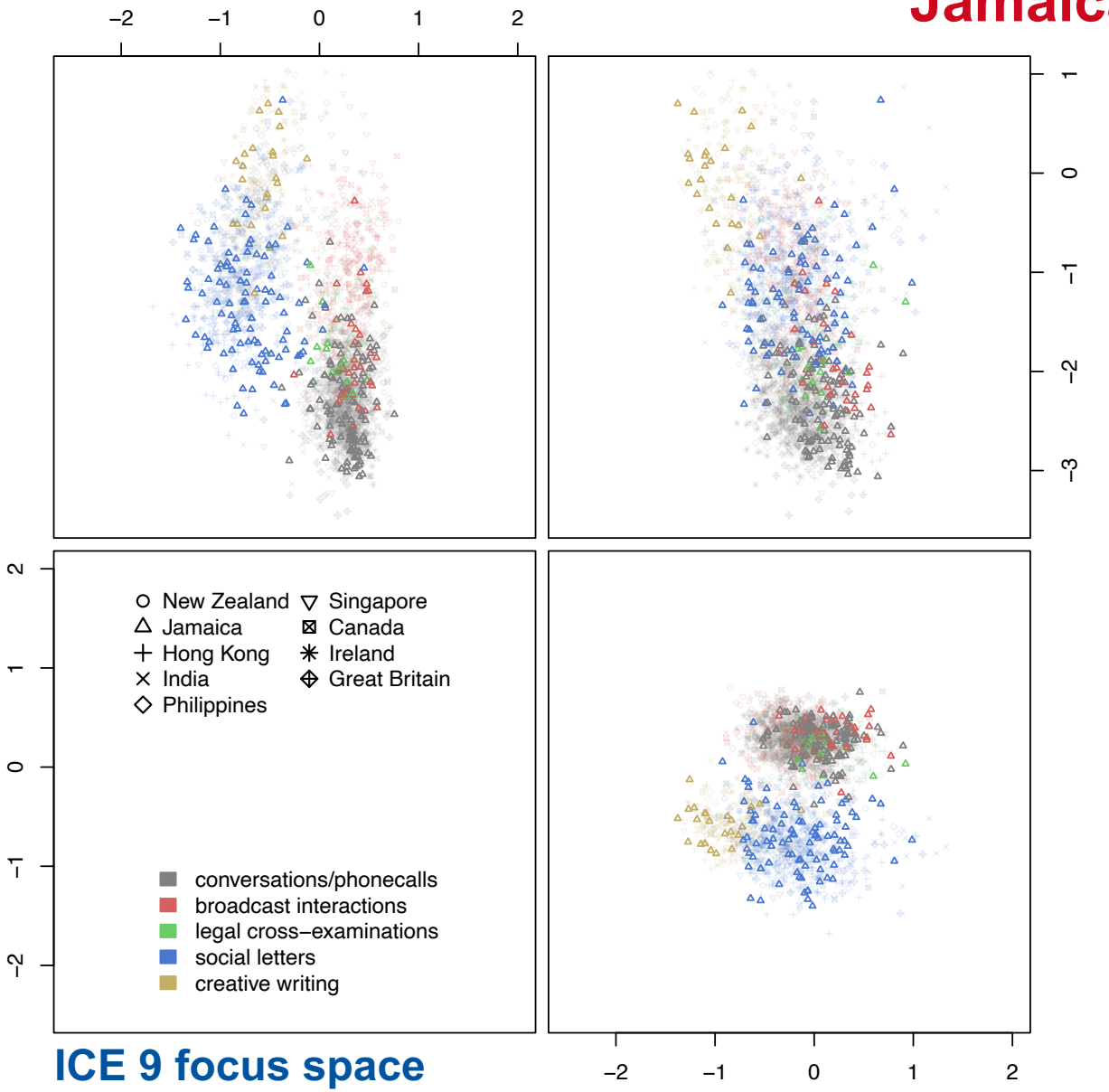
RWTH AACHEN UNIVERSITY

Register divergence across varieties



ICE 3 focus space

Jamaica



ICE 9 focus space

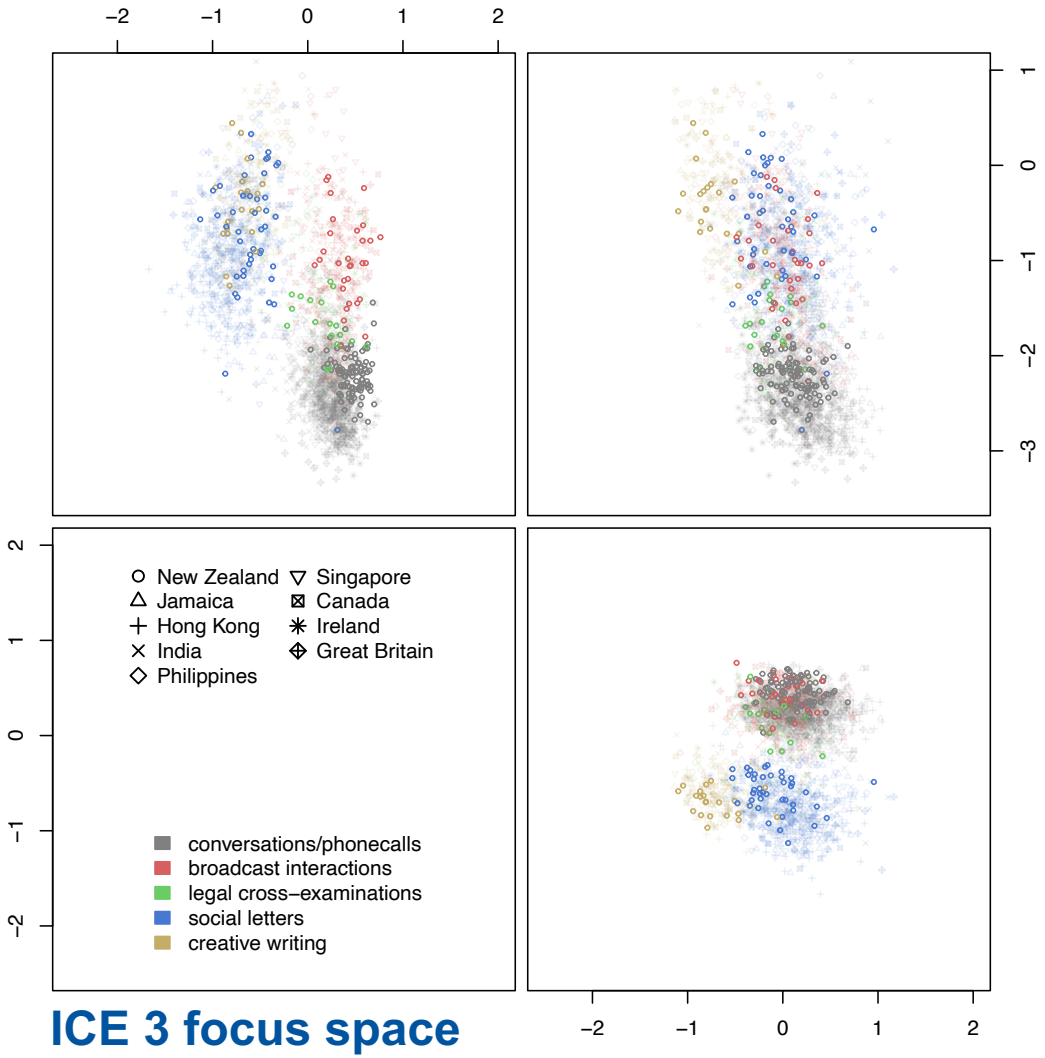
Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

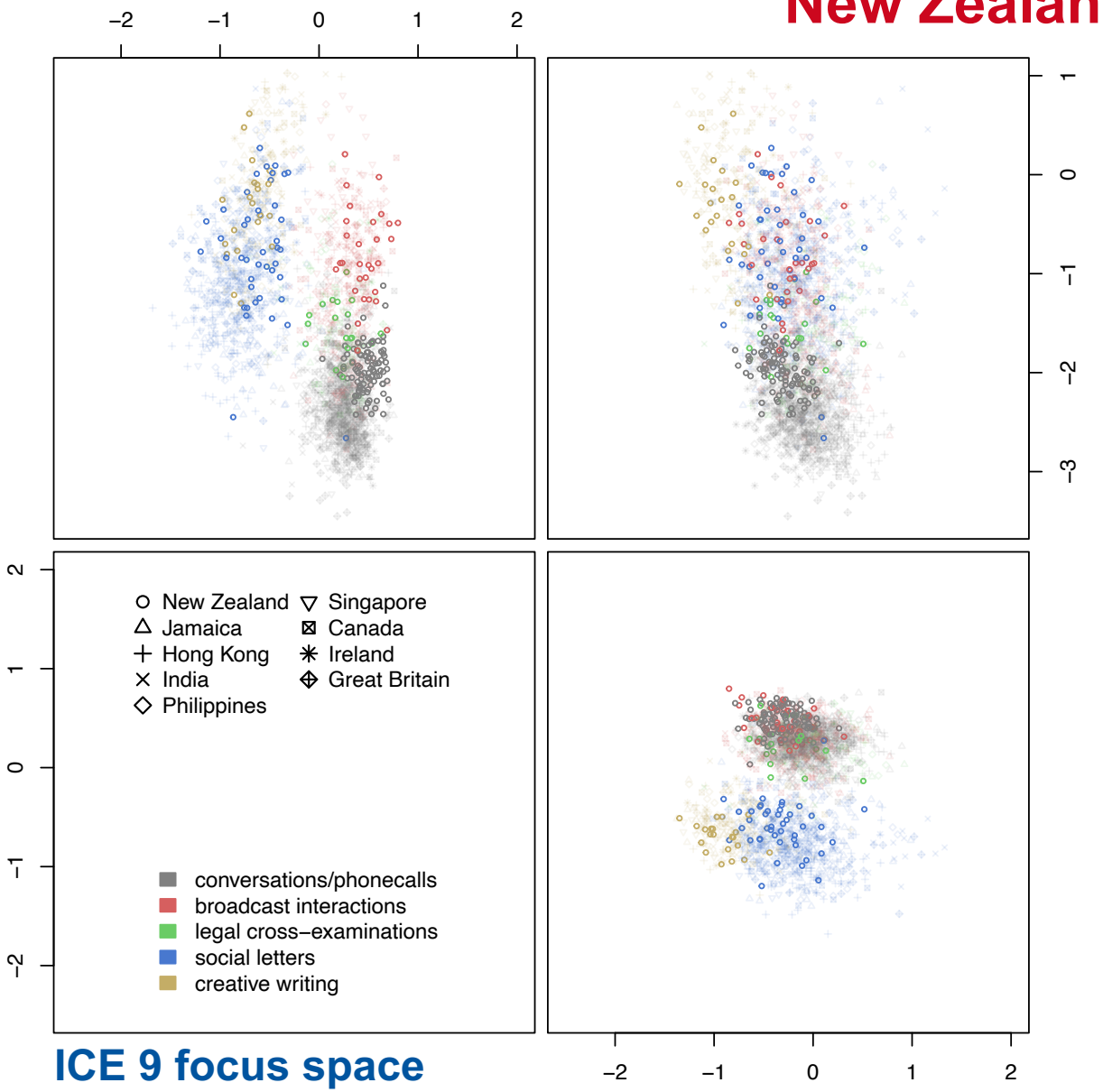
FAU

RWTH AACHEN UNIVERSITY

Register divergence across varieties



ICE 3 focus space



ICE 9 focus space

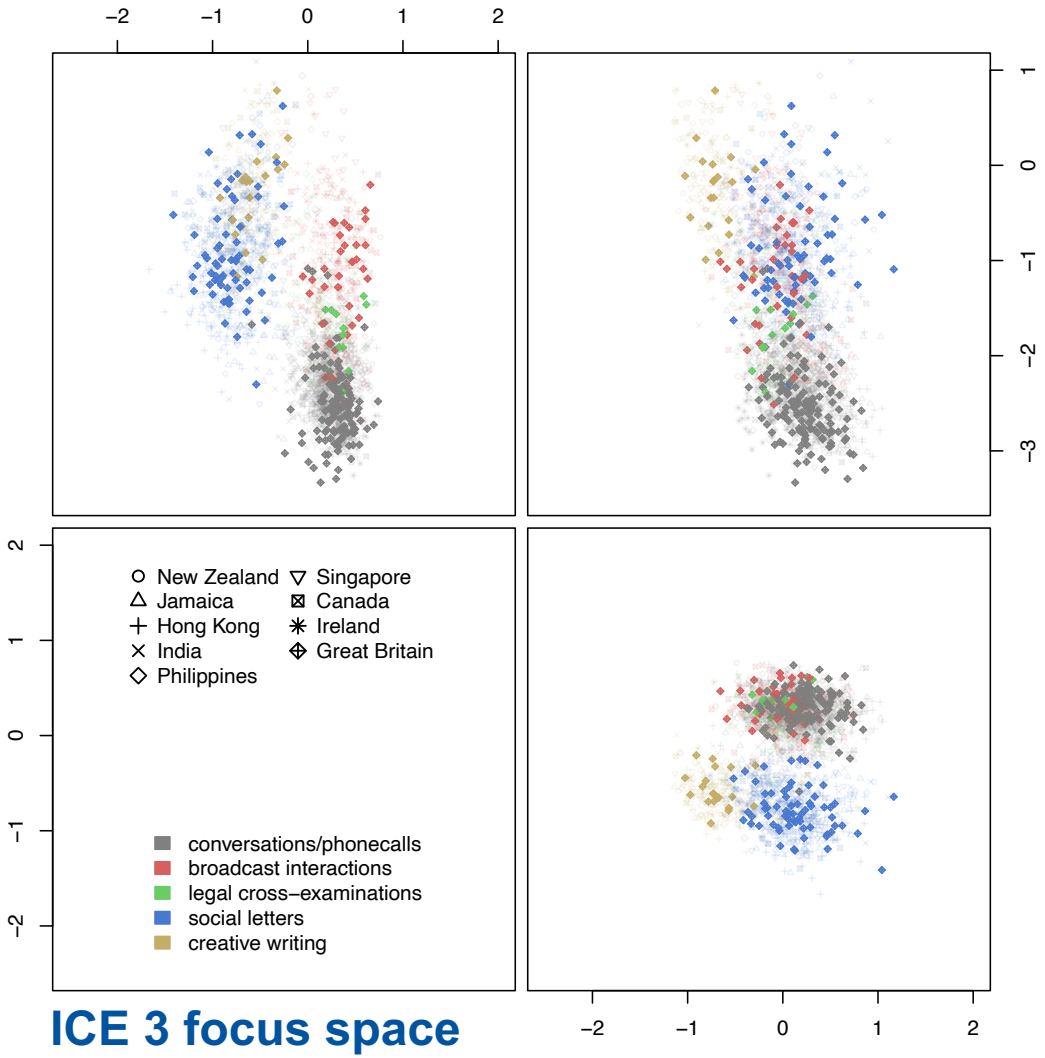
Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

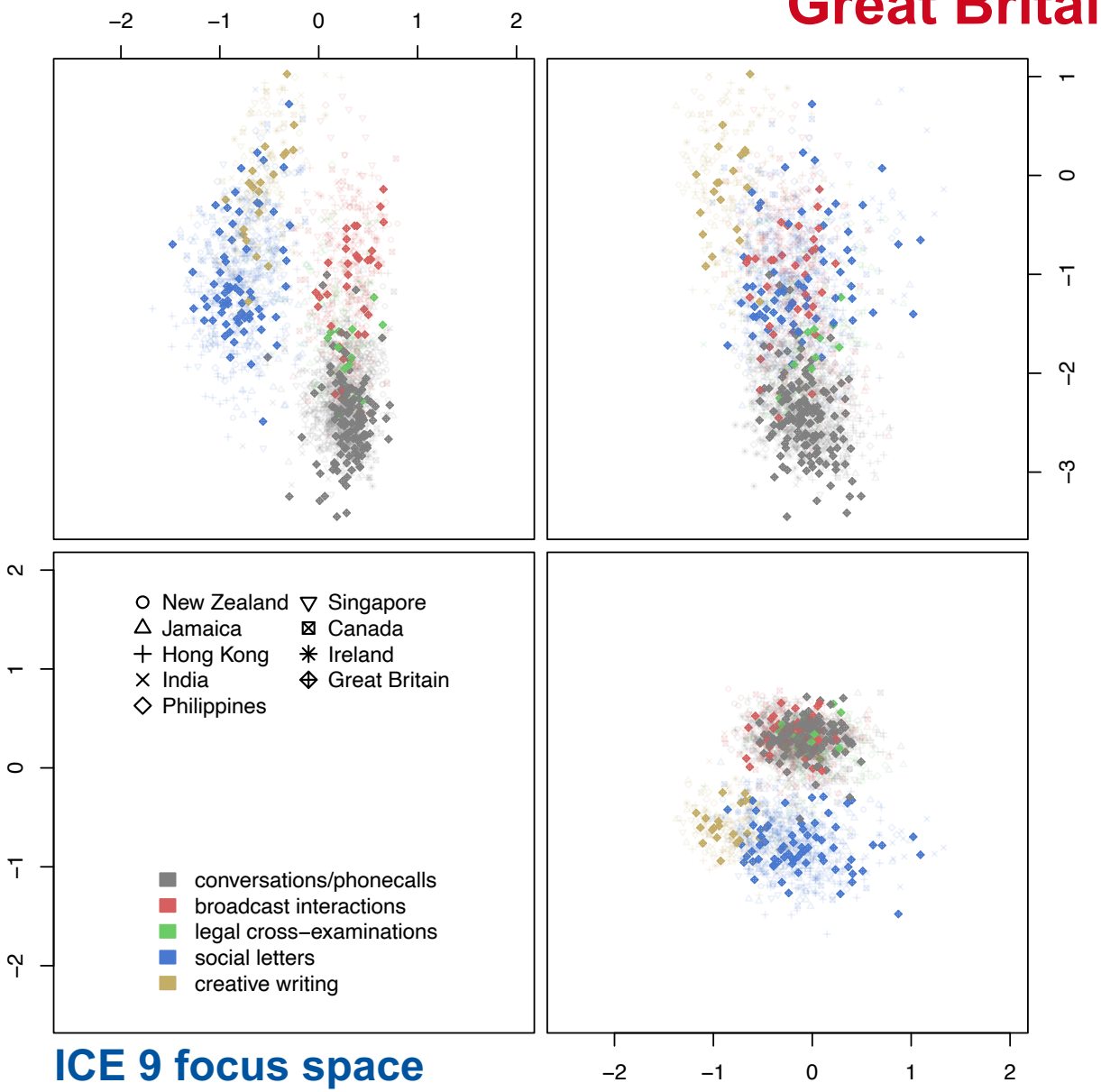
RWTH AACHEN UNIVERSITY

Register divergence across varieties



ICE 3 focus space

Great Britain



ICE 9 focus space

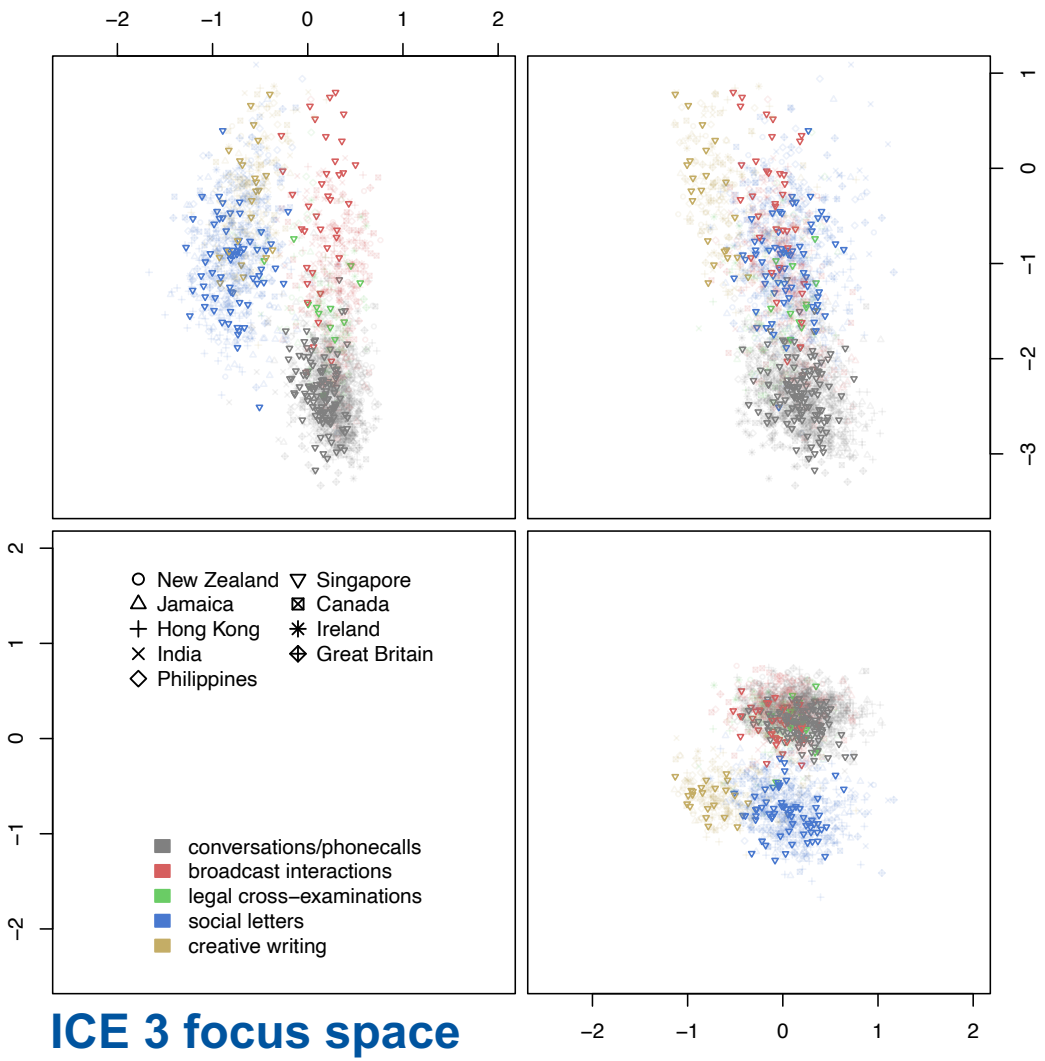
Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

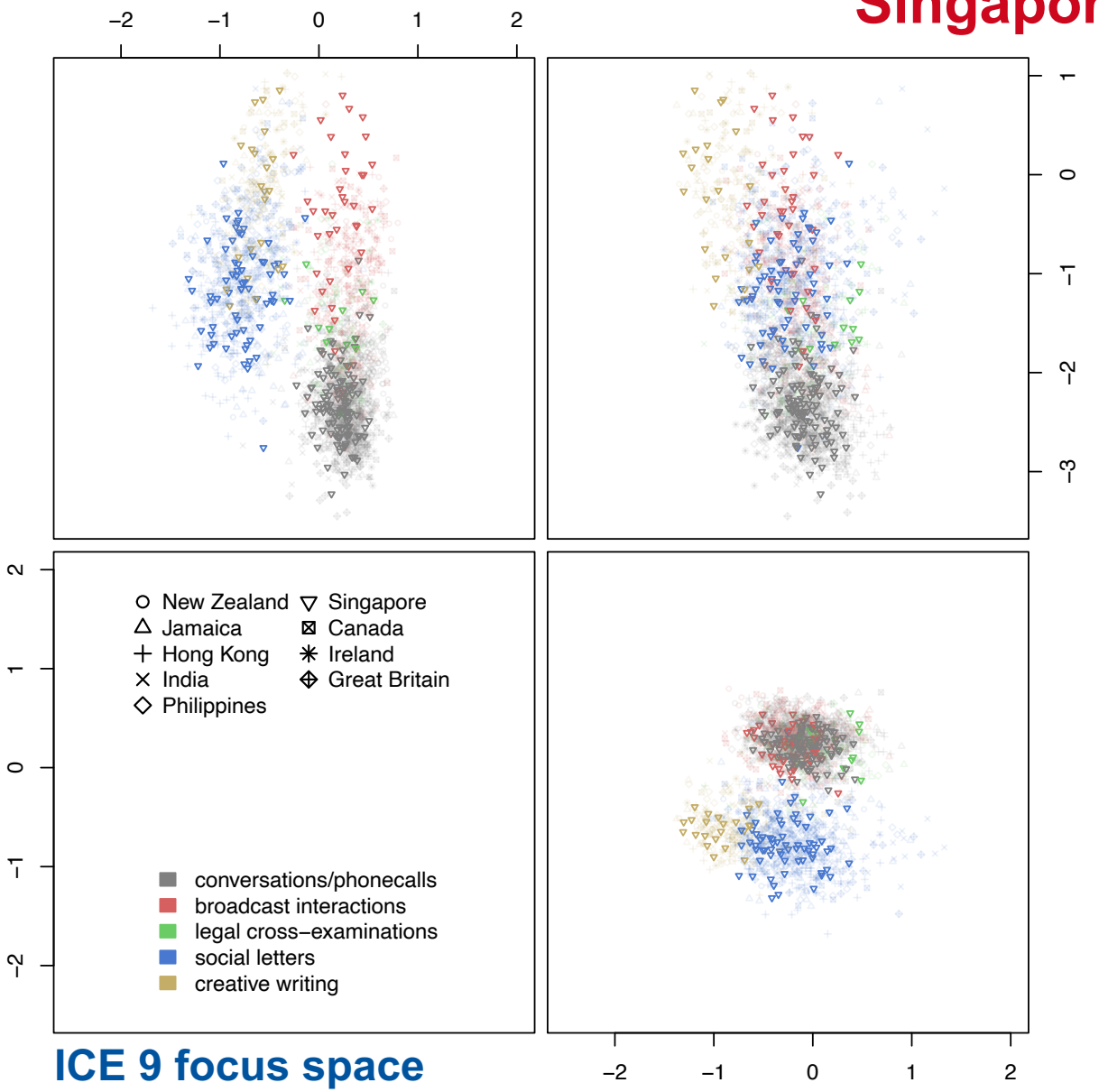
RWTH AACHEN UNIVERSITY

Register divergence across varieties



ICE 3 focus space

Singapore



ICE 9 focus space

Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

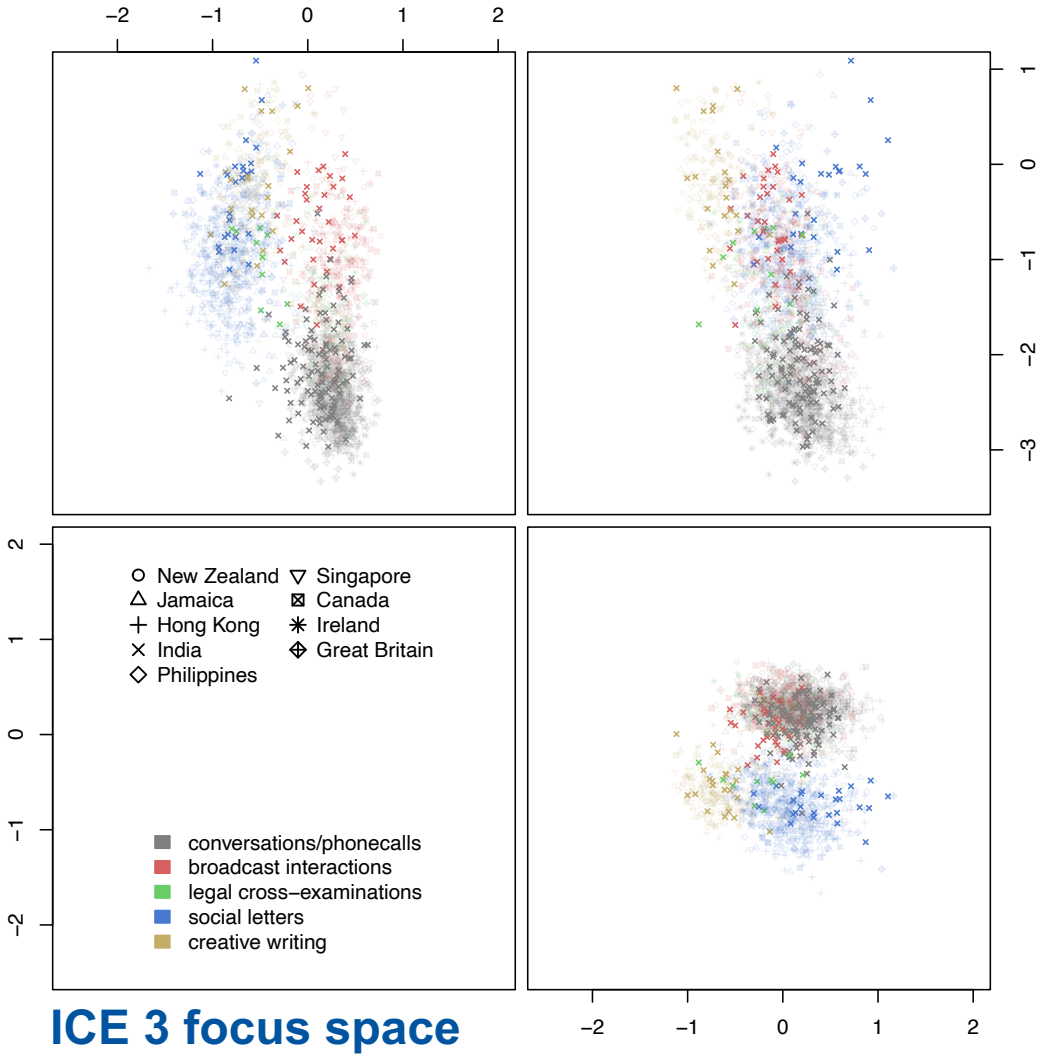
Universität Zürich[UZH]

FAU

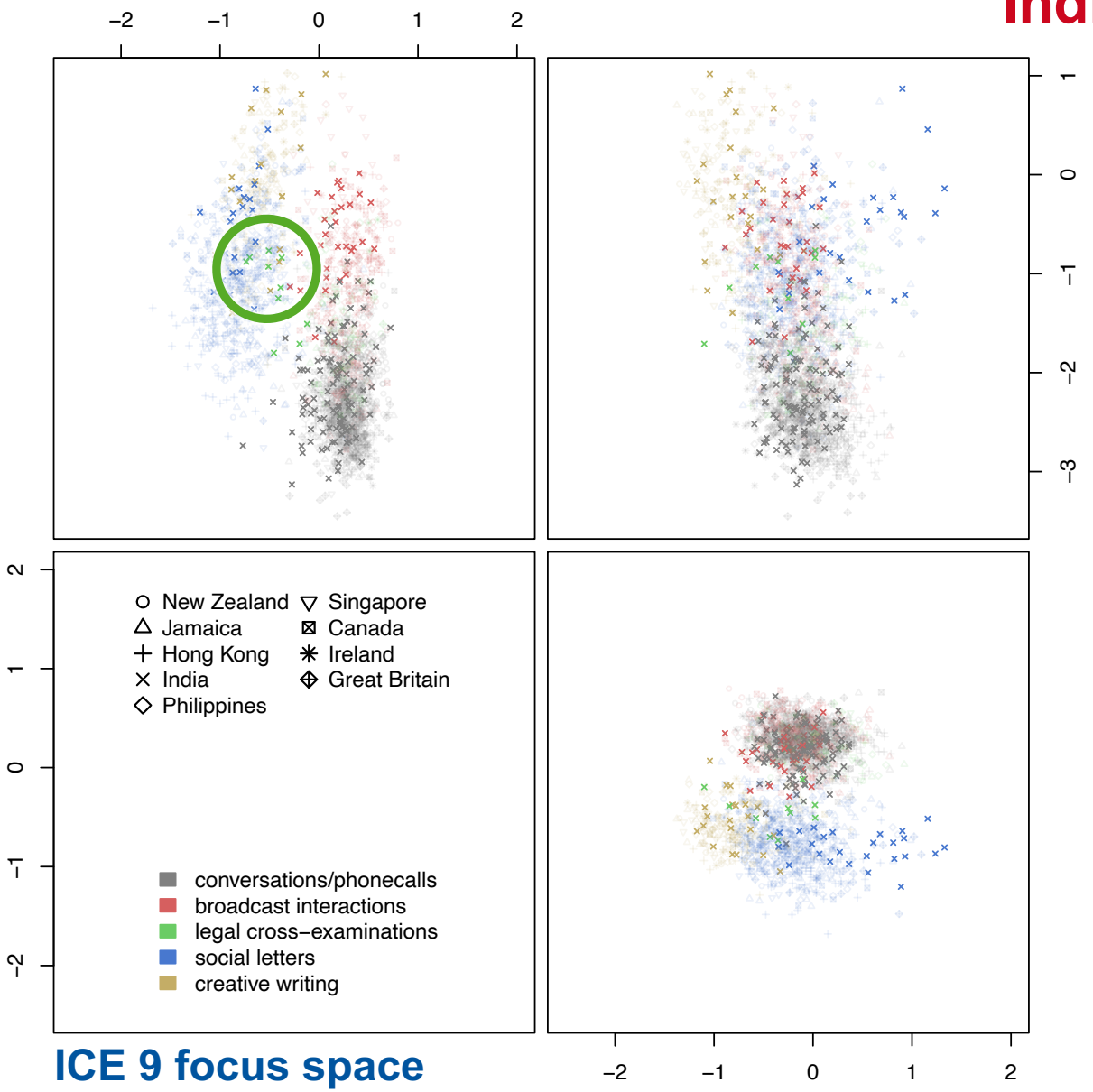
RWTH AACHEN UNIVERSITY

Register divergence across varieties

India



ICE 3 focus space



ICE 9 focus space

Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

RWTH AACHEN UNIVERSITY

Close Reading

- Complex models are hard to interpret. We minimally need to build some trust by close reading
  - → Look at the most prototypical=extreme cases (we will do that in Dim 1)
  - → Look at texts that appear in unexpected places in the space (we will do that in Dim 2)
- Dim 1 (*conceptual speaking* vs. *conceptual writing*): the ~10 most extreme *speaking* are all phone calls

```
```{r sub9identifyMostSpoken}
idx0 <- ICE9.Sub[, 1] < -3.2 ## Dim 1 most extreme left in *conceptual speaking vs. conceptual writing*
# focus in clear cases
# res0 <- subset(MetaSub, idx0 & short20 == "crossEx")
res0 <- subset(MetaSub, idx0)
print(res0)
```
```

Description: dt [6 × 18]

id <chr>	variety <fctr>	mode <fctr>	format <fctr>	short32 <fctr>	textcat32 <fctr>	code32 <fctr>	short20 <fctr>	textcat20 <fctr>	code20 <fctr>
icegb_s1a-095_2	Great Britain	spoken	dialogue	phone	phonecalls	S1A-091-100	conv	conversations/phonecalls	S1A
icegb_s1a-098_1	Great Britain	spoken	dialogue	phone	phonecalls	S1A-091-100	conv	conversations/phonecalls	S1A
icegb_s1a-098_2	Great Britain	spoken	dialogue	phone	phonecalls	S1A-091-100	conv	conversations/phonecalls	S1A
icegb_s1a-099_2	Great Britain	spoken	dialogue	phone	phonecalls	S1A-091-100	conv	conversations/phonecalls	S1A
icegb_s1a-100_2	Great Britain	spoken	dialogue	phone	phonecalls	S1A-091-100	conv	conversations/phonecalls	S1A
icesing_s1a-099_1	Singapore	spoken	dialogue	phone	phonecalls	S1A-091-100	conv	conversations/phonecalls	S1A

6 rows | 1-10 of 18 columns

Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Interpretation

Sentence length, lexical richness

Close Reading: Dim 1 (*conceptual speaking* vs. conceptual writing): e.g. icegb_s1a-095_2:

icegb_s1a-095_2_A	30041	62	Hallo
icegb_s1a-095_2_B	30103	54	Hallo
icegb_s1a-095_2_B	30157	96	How are you
icegb_s1a-095_2_A	30253	277	I'm very well thank you
icegb_s1a-095_2_B	30530	77	Jolly good
icegb_s1a-095_2_B	30607	179	How's work
icegb_s1a-095_2_A	30786	199	Uh oh not too bad
icegb_s1a-095_2_A	30985	182	not too bad
icegb_s1a-095_2_B	31167	75	Yeah
icegb_s1a-095_2_B	31242	330	Still all there
icegb_s1a-095_2_B	31572	280	Nobody else been thrown out
icegb_s1a-095_2_A	31852	157	Uh not yet no
icegb_s1a-095_2_B	32009	139	Not yet
icegb_s1a-095_2_B	32148	279	Is it in the offing or
icegb_s1a-095_2_A	32427	55	Well
icegb_s1a-095_2_A	32482	315	Could well be I think
icegb_s1a-095_2_B	32797	63	Yeah
icegb_s1a-095_2_A	32860	132	Uhm yeah
icegb_s1a-095_2_A	32992	217	Oh sorry the Gi Giles
icegb_s1a-095_2_A	33209	706	Giles's uh whatever they are lamb something or others are
icegb_s1a-095_2_A	33915	456	Sorry combined effort on the cooking going on there
icegb_s1a-095_2_B	34371	159	yeah
icegb_s1a-095_2_A	34530	91	Uh
icegb_s1a-095_2_B	34621	350	Yes so uh anything else interesting happening
icegb_s1a-095_2_A	34971	614	Uh well I don't quite know how to answer that one
icegb_s1a-095_2_B	35585	72	I see
icegb_s1a-095_2_B	35657	99	I see

Interpretation

Close Reading: Dim 1

- Dim 1 (*conceptual speaking* vs. *conceptual writing*): creative writing, social letters

```
```{r sub9identifyMostWritten}
idxx <- ICE9.Sub[, 1] > +1.1
# focus in clear cases
# res0 <- subset(MetaSub, idx0 & short20 == "crossEx")
resx <- subset(MetaSub, idxx)
print(resx)
```
```

Description: dt [6 × 18]

id <chr>	variety <fctr>	mode <fctr>	format <fctr>	short32 <fctr>	textcat32 <fctr>	code32 <fctr>	short20 <fctr>	textcat20 <fctr>	code20 <fctr>
icecan_w2f-013_1	Canada	written	printed	creat	novels and short stories	W2F-001-020	creat	creative writing	W2F
iceire_w2f-013_1	Ireland	written	printed	creat	novels and short stories	W2F-001-020	creat	creative writing	W2F
iceire_w2f-015_1	Ireland	written	printed	creat	novels and short stories	W2F-001-020	creat	creative writing	W2F
icephi_w1b-015_10	Philippines	written	non-printed	socLet	social letters	W1B-001-015	socLet	social letters	W1B1
icephi_w2f-003_1	Philippines	written	printed	creat	novels and short stories	W2F-001-020	creat	creative writing	W2F
icesing_w1b-013_1	Singapore	written	non-printed	socLet	social letters	W1B-001-015	socLet	social letters	W1B1

6 rows | 1-10 of 18 columns

Sentence length, lexical richness

icecan_w2f-013_1 11027 517 My father's peripatetic piety includes not only the missions but other churches and tabernacles , some far afield , in Elizabeth , in West New York , on Staten Island , all of which require hours of buses and ferries and buses again , several churches in Manhattan which are reached by subway , a place in Park Slope where a rene gade group from Elim calling themselves The Latter Rain anoint one another apostles and prophets and speak ecstatically of last things .

Interpretation

Close Reading: Dim 2 (*dialogic written* vs. *neutral*): e.g. iceind_s1b-062_1

- Some legal cross-examinations (most from ICE-IND) are located in the *dialogic written* range of LDA dim 2

```
```{r sub9identifyWeirdCrossEx}
idx1 <- ICE9.Sub[, 2] < -.2 # focus in clear cases
res1 <- subset(MetaSub, idx1 & short20 == "crossEx")
print(res1)
```
```

Description: dt [9 × 18]

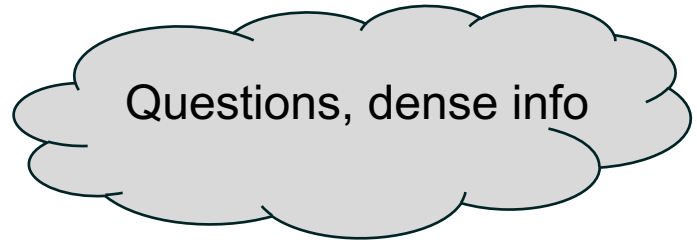
id <chr>	variety <fctr>	mode <fctr>	format <fctr>	short32 <fctr>	textcat32 <fctr>	code32 <fctr>	short20 <fctr>	textcat20 <fctr>
iceind_s1b-062_1	India	spoken	dialogue	crossEx	legal cross-examinations	S1B-060-070	crossEx	legal cross-examinations
iceind_s1b-063_1	India	spoken	dialogue	crossEx	legal cross-examinations	S1B-060-070	crossEx	legal cross-examinations
iceind_s1b-064_1	India	spoken	dialogue	crossEx	legal cross-examinations	S1B-060-070	crossEx	legal cross-examinations
iceind_s1b-066_1	India	spoken	dialogue	crossEx	legal cross-examinations	S1B-060-070	crossEx	legal cross-examinations
iceind_s1b-067_1	India	spoken	dialogue	crossEx	legal cross-examinations	S1B-060-070	crossEx	legal cross-examinations
iceind_s1b-068_1	India	spoken	dialogue	crossEx	legal cross-examinations	S1B-060-070	crossEx	legal cross-examinations
iceind_s1b-069_1	India	spoken	dialogue	crossEx	legal cross-examinations	S1B-060-070	crossEx	legal cross-examinations
iceind_s1b-070_1	India	spoken	dialogue	crossEx	legal cross-examinations	S1B-060-070	crossEx	legal cross-examinations
icesing_s1b-070_2	Singapore	spoken	dialogue	crossEx	legal cross-examinations	S1B-060-070	crossEx	legal cross-examinations

9 rows | 1-9 of 18 columns

Interpretation

Close Reading: Dim 2 (*dialogic written* vs. *neutral*): e.g. iceind_s1b-062_1

- legal cross-examinations in *dialogic written* range of dim 2



iceind_s1b-062_1_a	14	54	Mr Angale
iceind_s1b-062_1_b	75	37	Yes
iceind_s1b-062_1_a	119	109	in the year nineteen eighty-one what business you were doing ?
iceind_s1b-062_1_a	228	66	Were you in service or business ?
iceind_s1b-062_1_b	301	53	No I was in business
iceind_s1b-062_1_a	361	73	What business were you doing in eighty-one ?
iceind_s1b-062_1_b	441	130	That time I was an artist that those days and I was running my shows in Bombay
iceind_s1b-062_1_a	578	45	What shows ?
iceind_s1b-062_1_b	630	77	variety entertainment programme shows
iceind_s1b-062_1_a	714	100	I see not any business you were not in business I suppose
iceind_s1b-062_1_b	821	54	No not that time
iceind_s1b-062_1_a	882	97	I was not in business in the year nineteen eighty-one
iceind_s1b-062_1_c	986	84	However I was doing varieties of entertainment
iceind_s1b-062_1_a	1077	67	Varieties entertainment programme

Interpretation

Close Reading: Dim 2 (*dialogic written* vs. *neutral*): e.g. icejam_w1b-012_11

- Conversely, a small number of social letters are in the positive range of the dimension (“neutral”) that is normally reserved to spoken mode.

```
```{r sub9identifyWeirdSocLet}
idx2 <- ICE9.Sub[, 2] > 0 # focus in clear cases
res2 <- subset(MetaSub, idx2 & short20 == "socLet")
print(res2)
```
```

Description: dt [6 × 18]

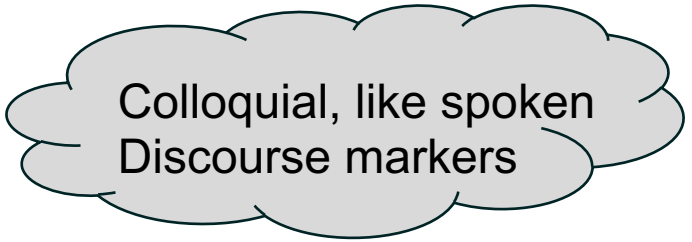
id <chr>	variety <fctr>	mode <fctr>	format <fctr>	short32 <fctr>	textcat32 <fctr>	code32 <fctr>	short20 <fctr>	textcat20 <fctr>	code20 <fctr>
icejam_w1b-012_11	Jamaica	written	non-printed	socLet	social letters	W1B-001-015	socLet	social letters	W1B1
icejam_w1b-012_3	Jamaica	written	non-printed	socLet	social letters	W1B-001-015	socLet	social letters	W1B1
icejam_w1b-013_5	Jamaica	written	non-printed	socLet	social letters	W1B-001-015	socLet	social letters	W1B1
icenz_w1b-010_6	New Zealand	written	non-printed	socLet	social letters	W1B-001-015	socLet	social letters	W1B1
icephi_w1b-015_10	Philippines	written	non-printed	socLet	social letters	W1B-001-015	socLet	social letters	W1B1
icesing_w1b-013_1	Singapore	written	non-printed	socLet	social letters	W1B-001-015	socLet	social letters	W1B1

6 rows | 1-10 of 18 columns

Interpretation

Close Reading: Dim 2 (*dialogic written* vs. *neutral*): e.g. icejam_w1b-012_11

- social letters are in (*neutral*) that is normally reserved to *spoken* mode.



icejam_w1b-012_11 9629 26 Hi Mark

icejam_w1b-012_11 9655 196 I am presently in Atlanta and looking jobs , the prospects seem good , I have an interview for tomorrow so hopefully I will be working soon (before I get broke) .

icejam_w1b-012_11 9851 184 Anyway in New York a lot of people keep saying it's tough but seriously I did not have a problem .

icejam_w1b-012_11 10035 109 The only problem finding work there is the screening , it takes a lot of time now since the 9/11 attack .

icejam_w1b-012_11 10144 30 But there are jobs there .

icejam_w1b-012_11 10174 22 I got many offers .

icejam_w1b-012_11 10196 104 They desperately need people in the home health aid and child care too .

icejam_w1b-012_11 10300 59 So if you have contacts there you could try it .

Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

The End is the Beginning

Conclusions

- Reproduction of Neumann & Evert (2001) with new preprocessing successful
- Replication with 9 ICE components: mostly successful, very similar observations
  - → but LDA projection into 4 dimensions produced a rotated coordinate system → not obvious in visualization
  - → may have led to different interpretation (esp. of LDA dim 3 and 4) without reference to ICE3 analysis
- The dimensions are meaningfully interpretable
- Robust register space created by LDA, differences between varieties are smaller than register variation
- Some variety-specific register differences emerge, but we need to zoom/interpret more →
- Future work
  - → Patterns along inner/outer circle (Kachru 1985) varieties; L1/L2
  - → Cultural differences, e.g. colloquial Jamaican / tough legal cross-examination in Indian?
  - → Choice of linguistic/stylistic features: e.g. connect to forensic linguistics, more syntactic features
  - → LDA partially factors out differences between varieties → need modified LDA algorithm
  - → Development of register & expectations & conventions throughout the document
  - → Detect/Reject cultural stereotypes/differences/conventions
  - → Profit from LLMs: BERT / GPT / Llama, also with a forensic approach
  - → Tove Larsson: less register variation in spoken ← more regional variation in spoken?

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.
- 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/.

References

- 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/.

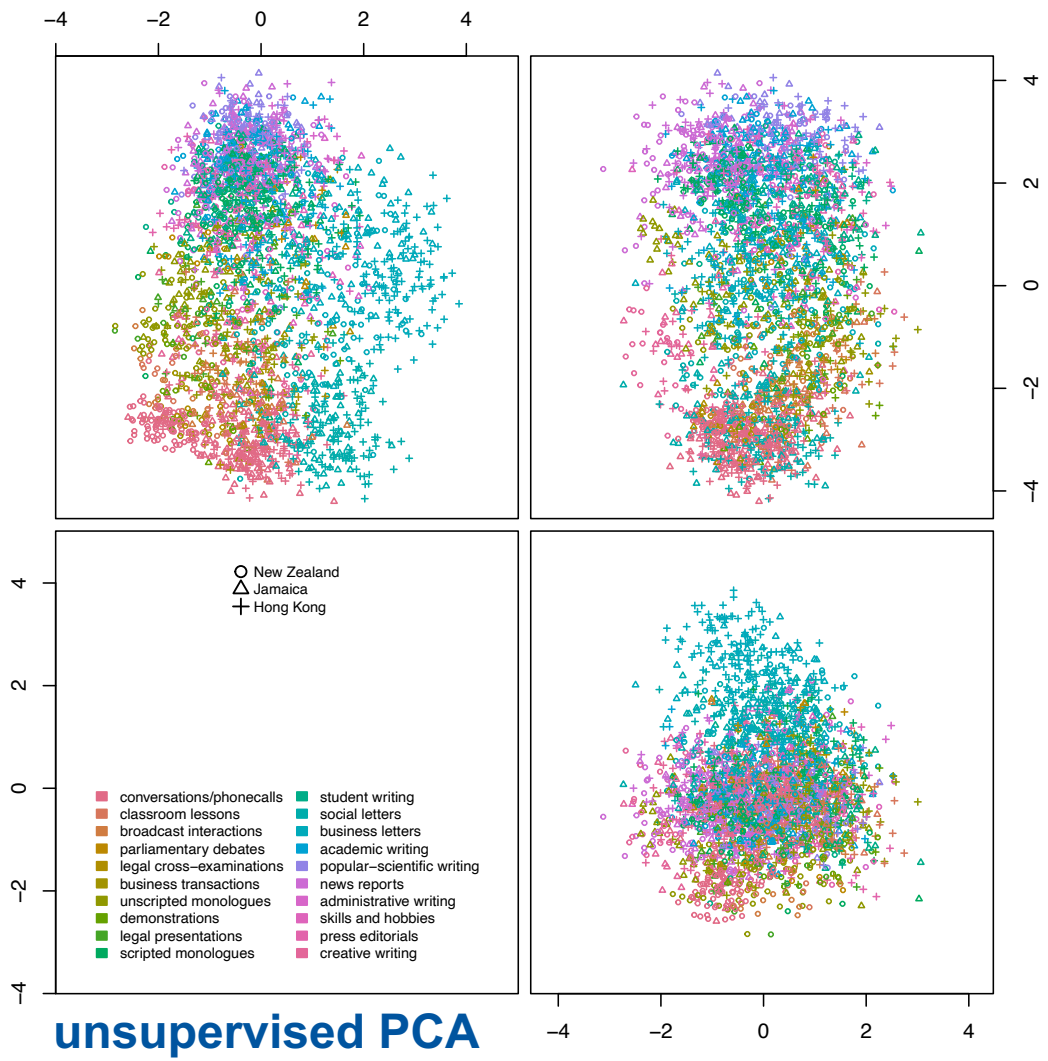
Appendix

Some additional visualisations and details

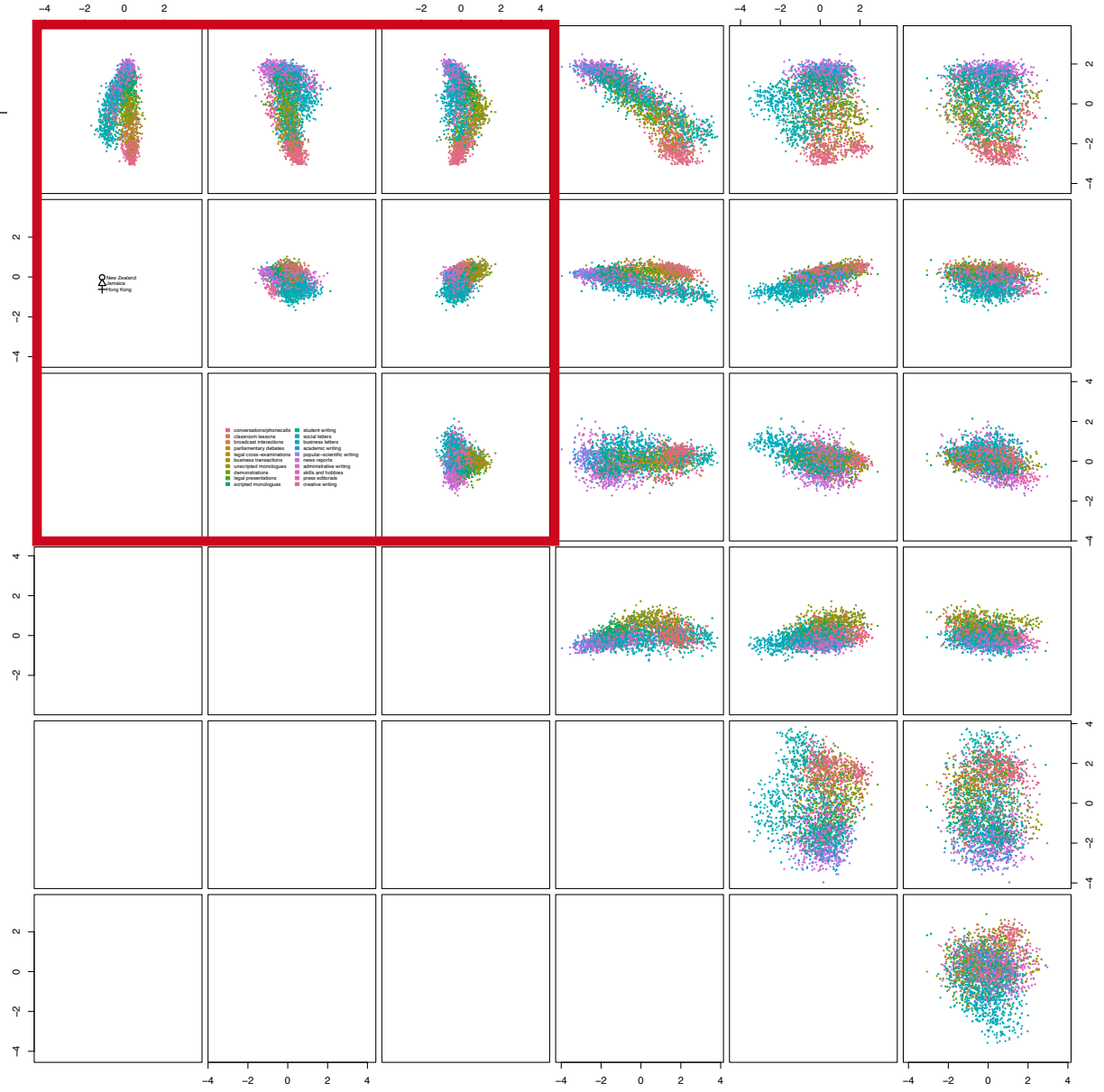
Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Universität Zürich[UZH]
FAU
RWTH AACHEN UNIVERSITY

Putting LDA dimensions into perspective



unsupervised PCA



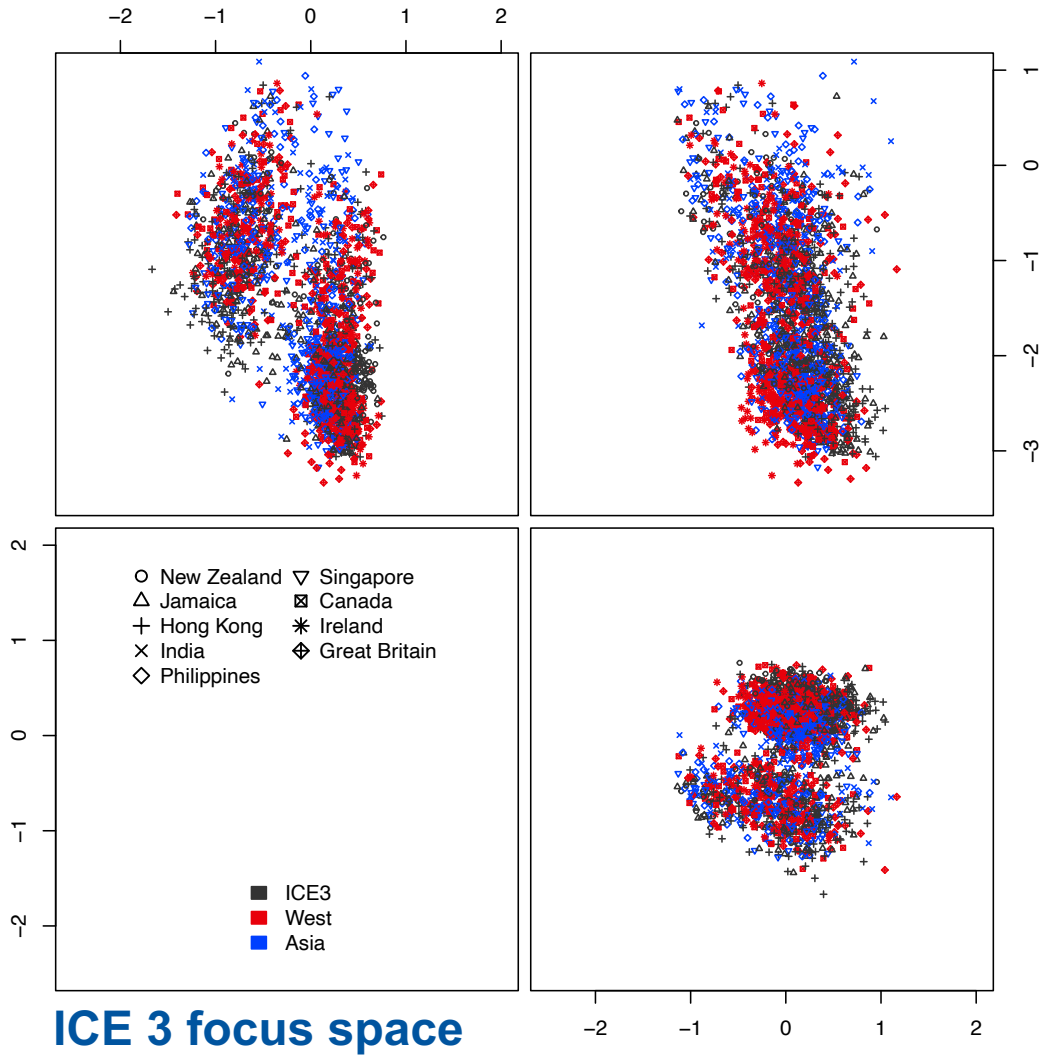
Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

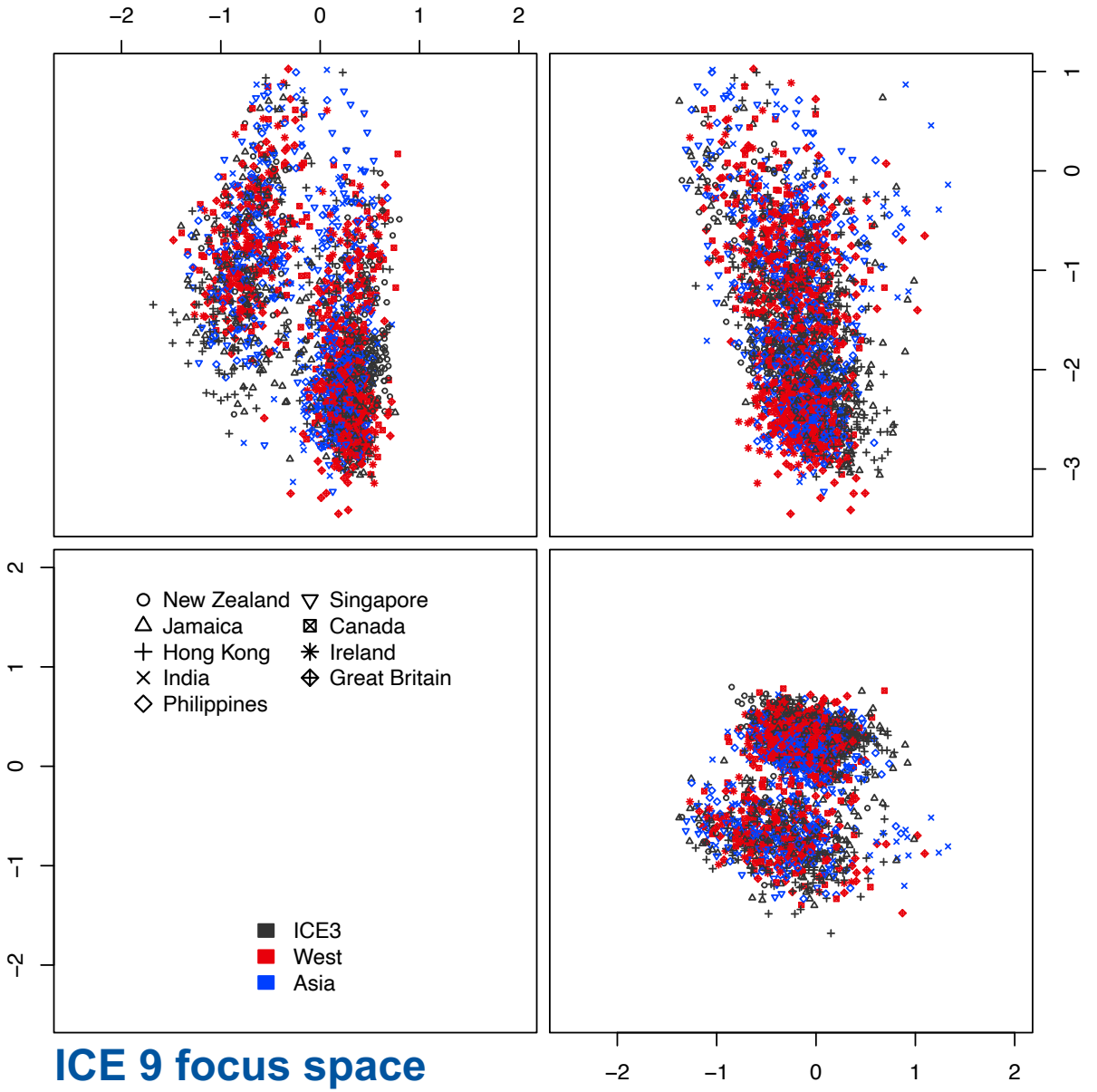
FAU

RWTH AACHEN UNIVERSITY

Divergence between varieties across registers



ICE 3 focus space



ICE 9 focus space

Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

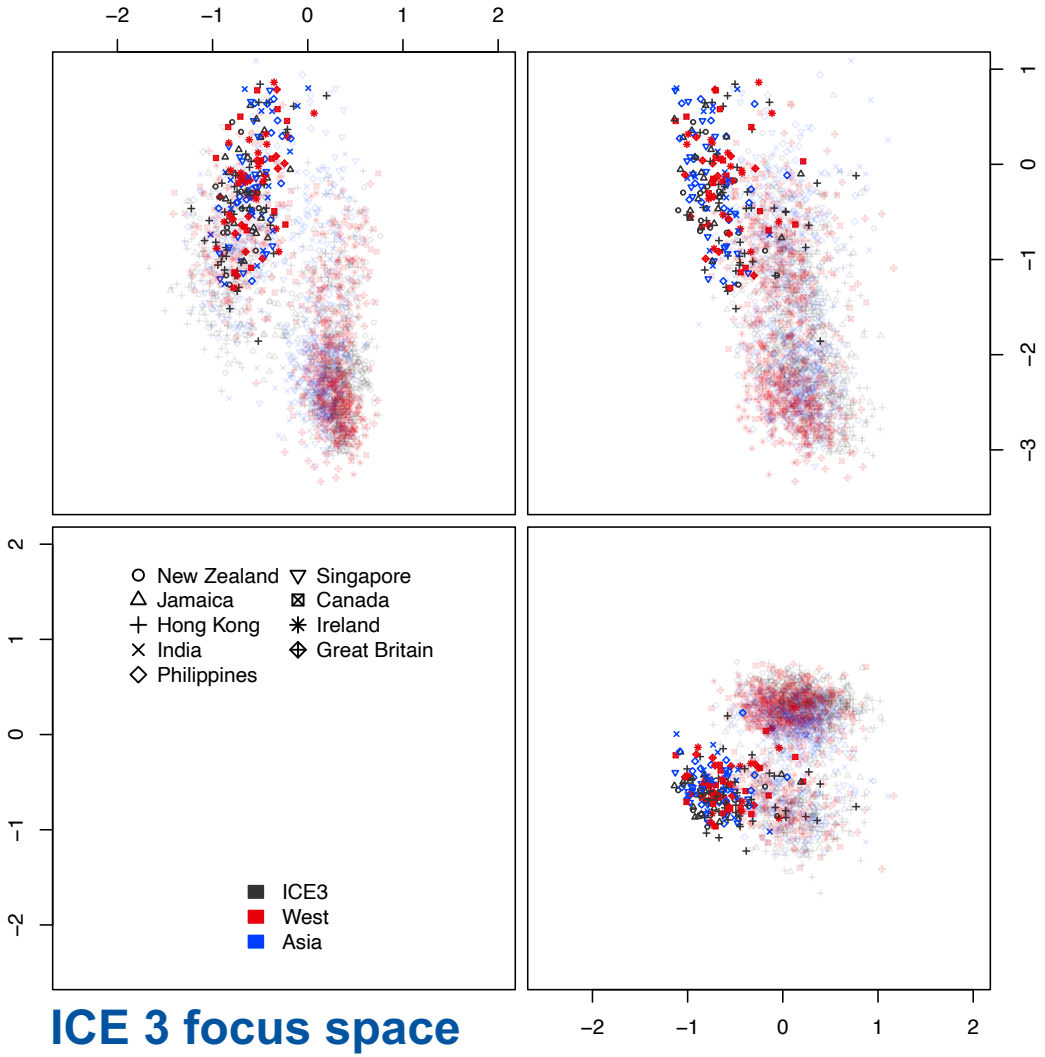
Universität Zürich[UZH]

FAU

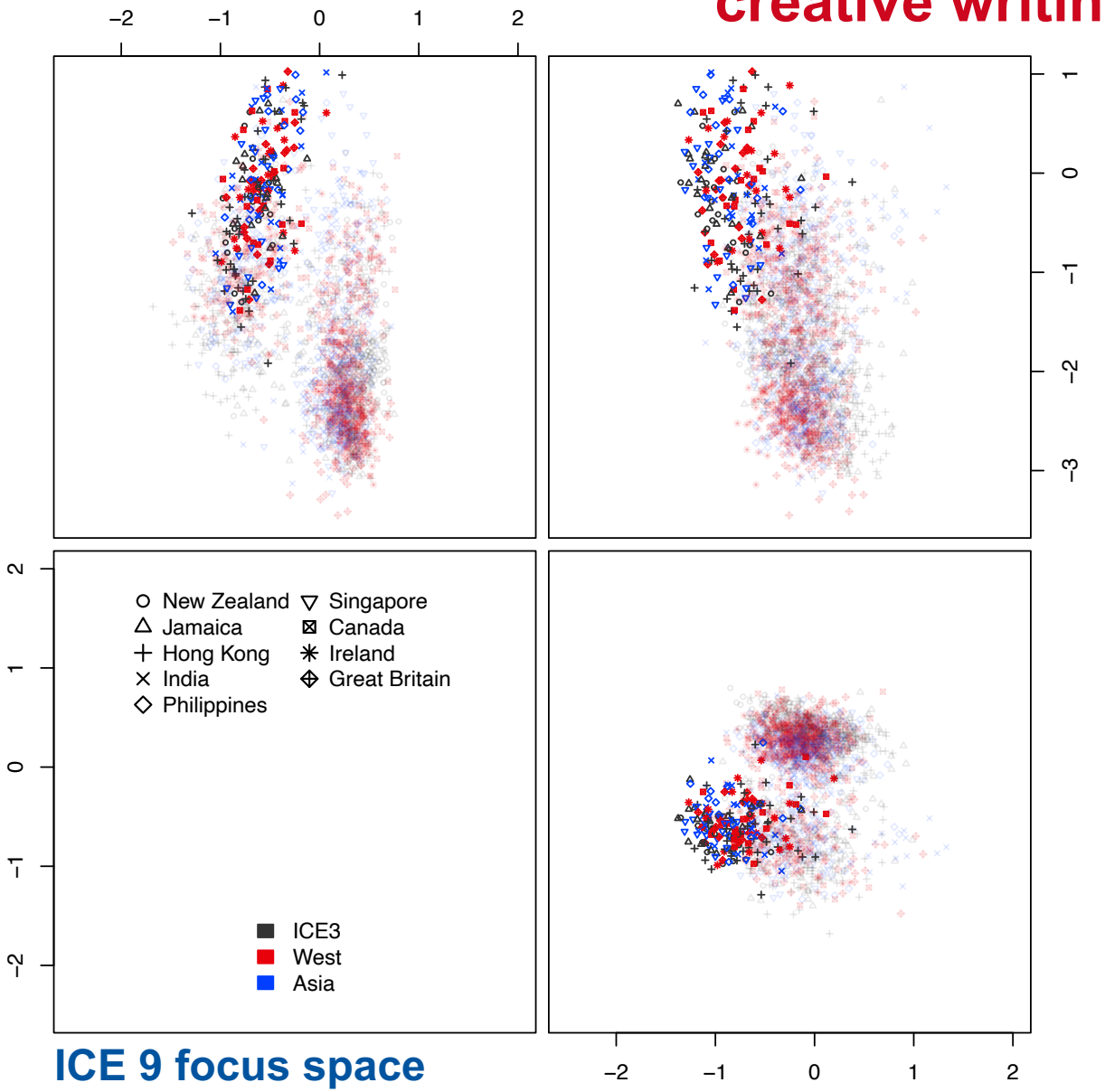
RWTH AACHEN UNIVERSITY

Divergence between varieties across registers

creative writing



ICE 3 focus space



ICE 9 focus space

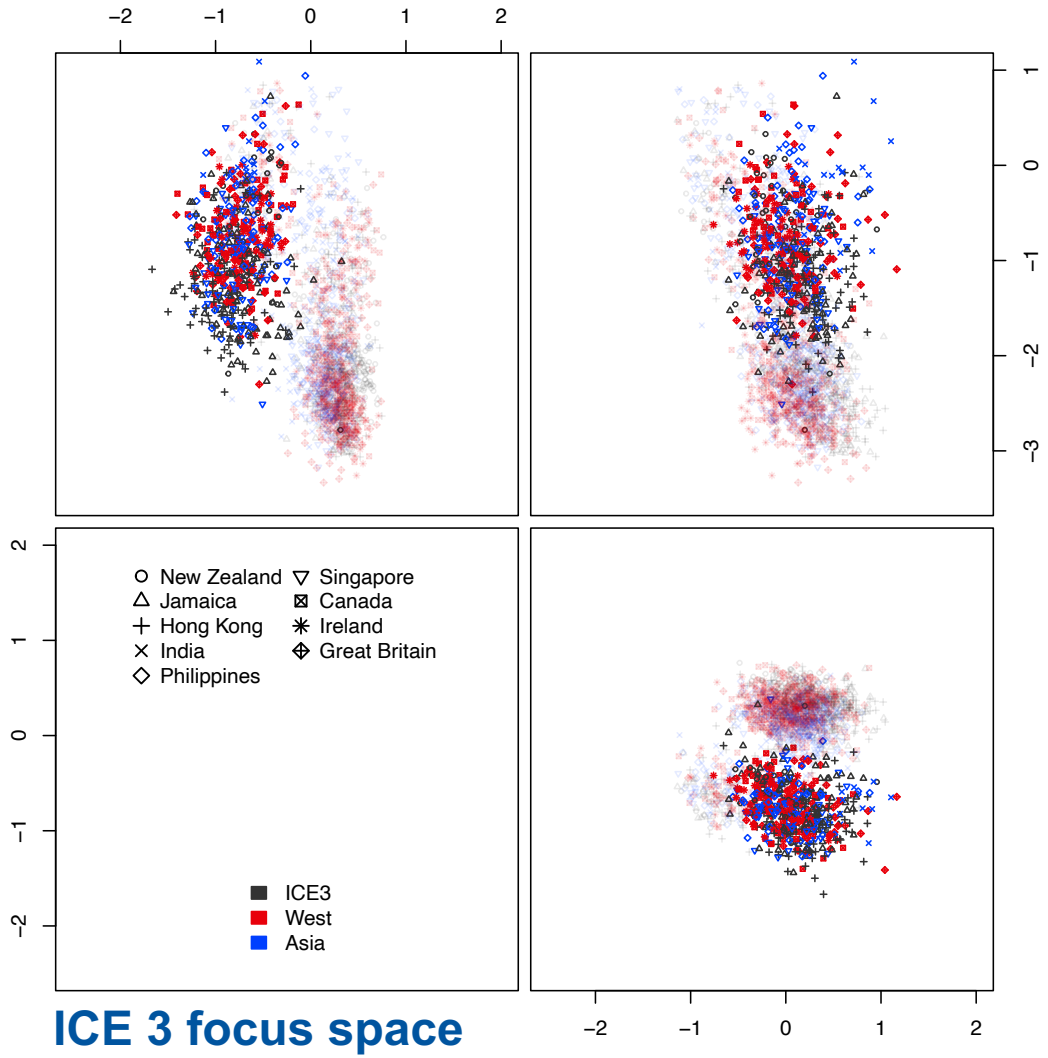
Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

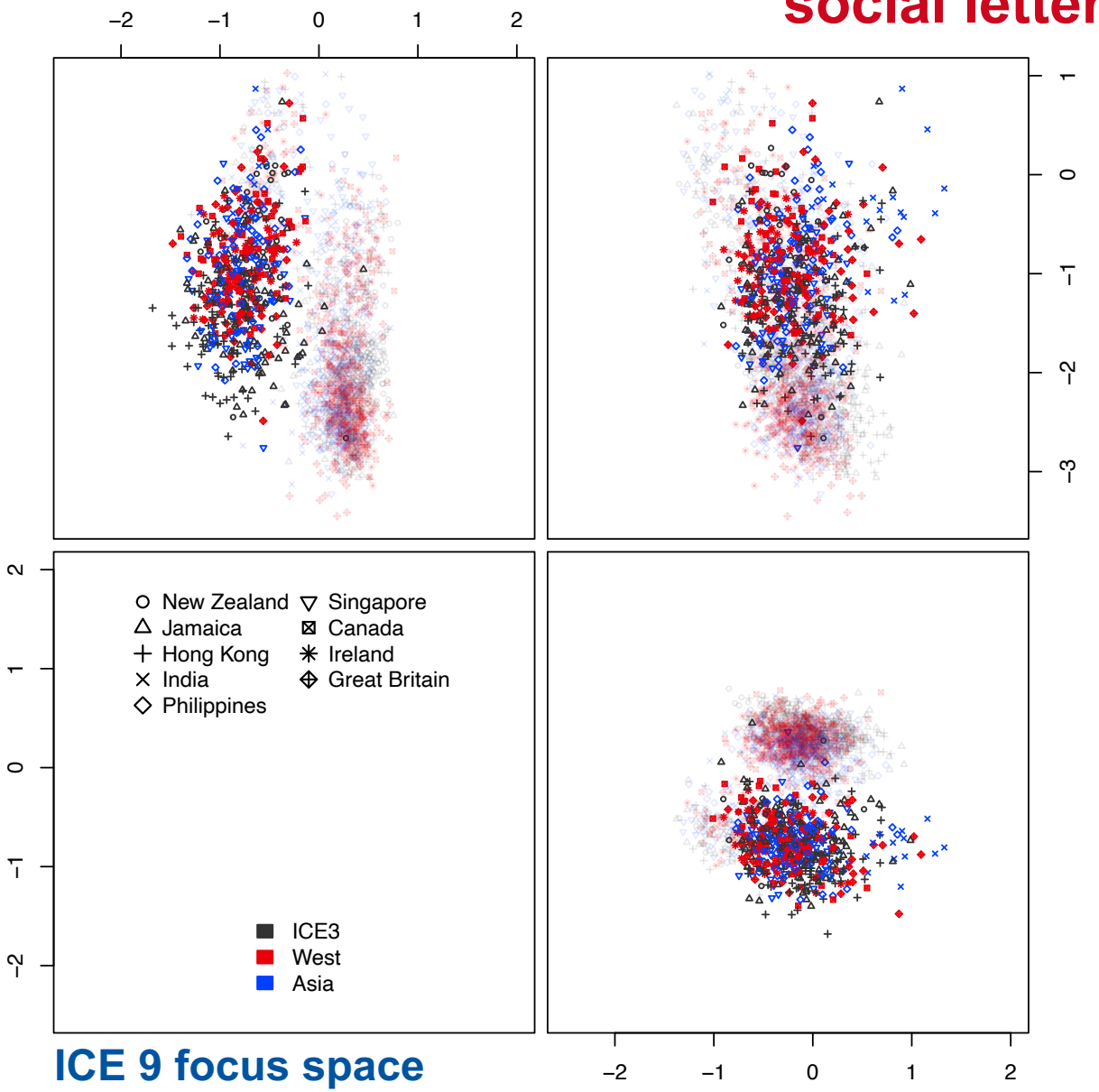
RWTH AACHEN UNIVERSITY

Divergence between varieties across registers



ICE 3 focus space

social letters



ICE 9 focus space

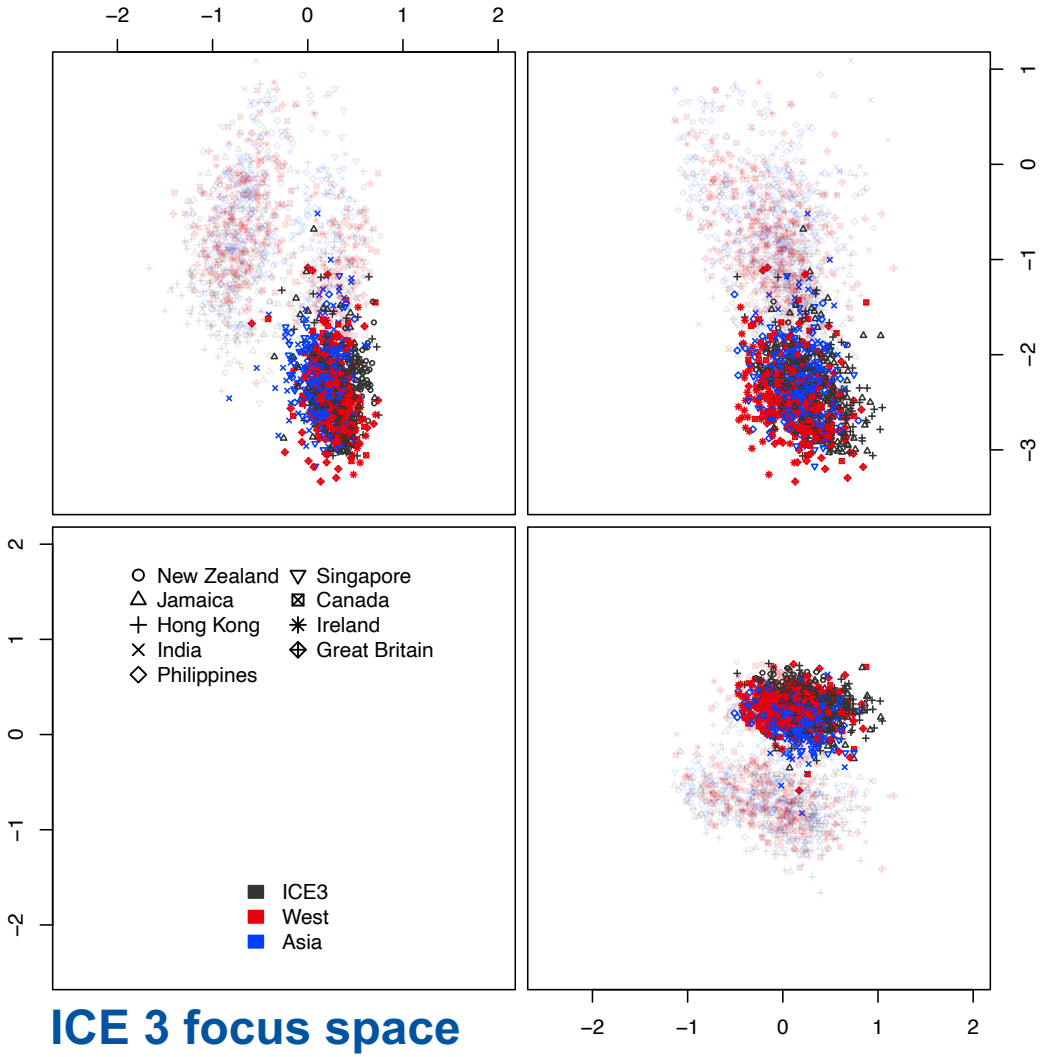
Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

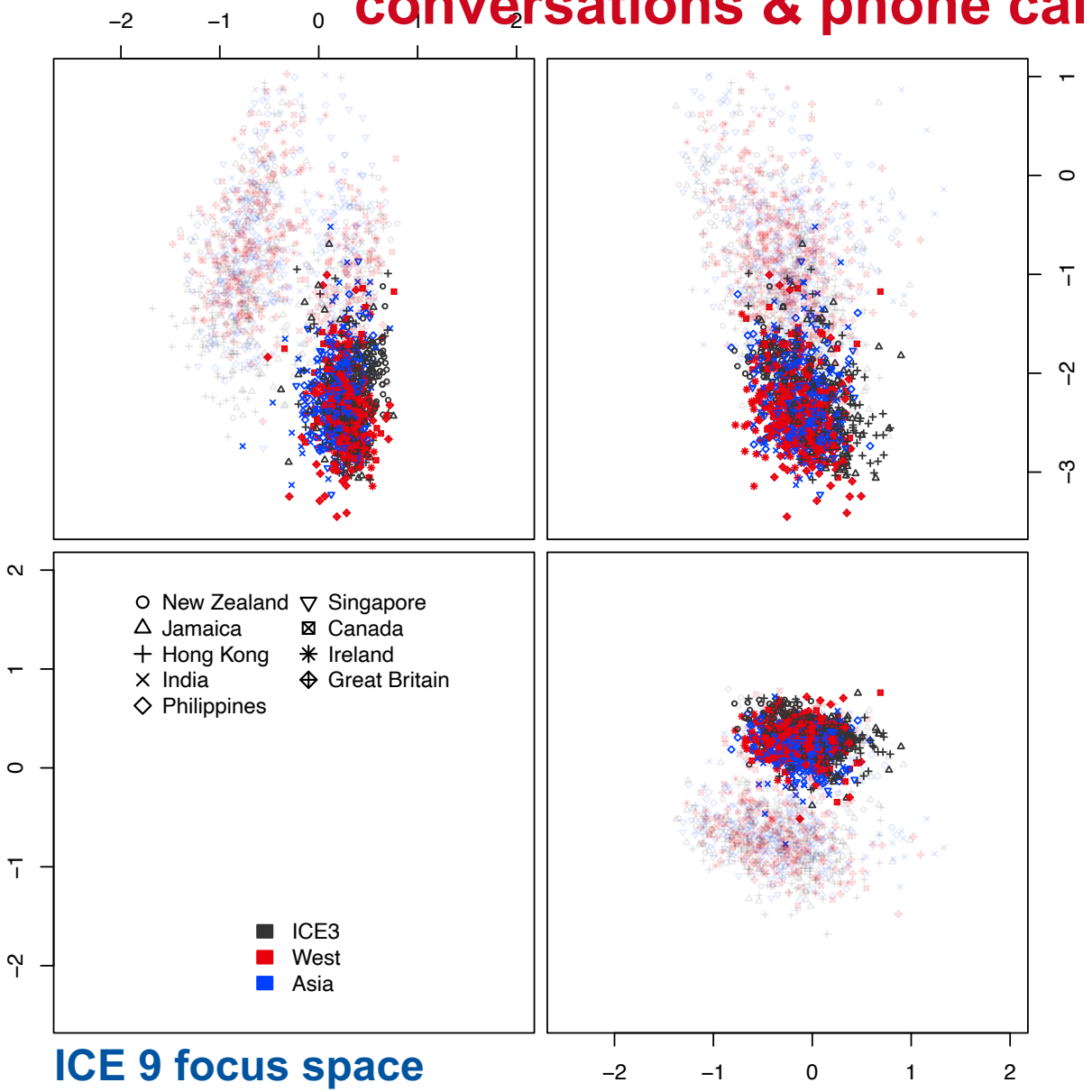
RWTH AACHEN UNIVERSITY

Divergence between varieties across registers



ICE 3 focus space

conversations & phone calls



ICE 9 focus space

Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

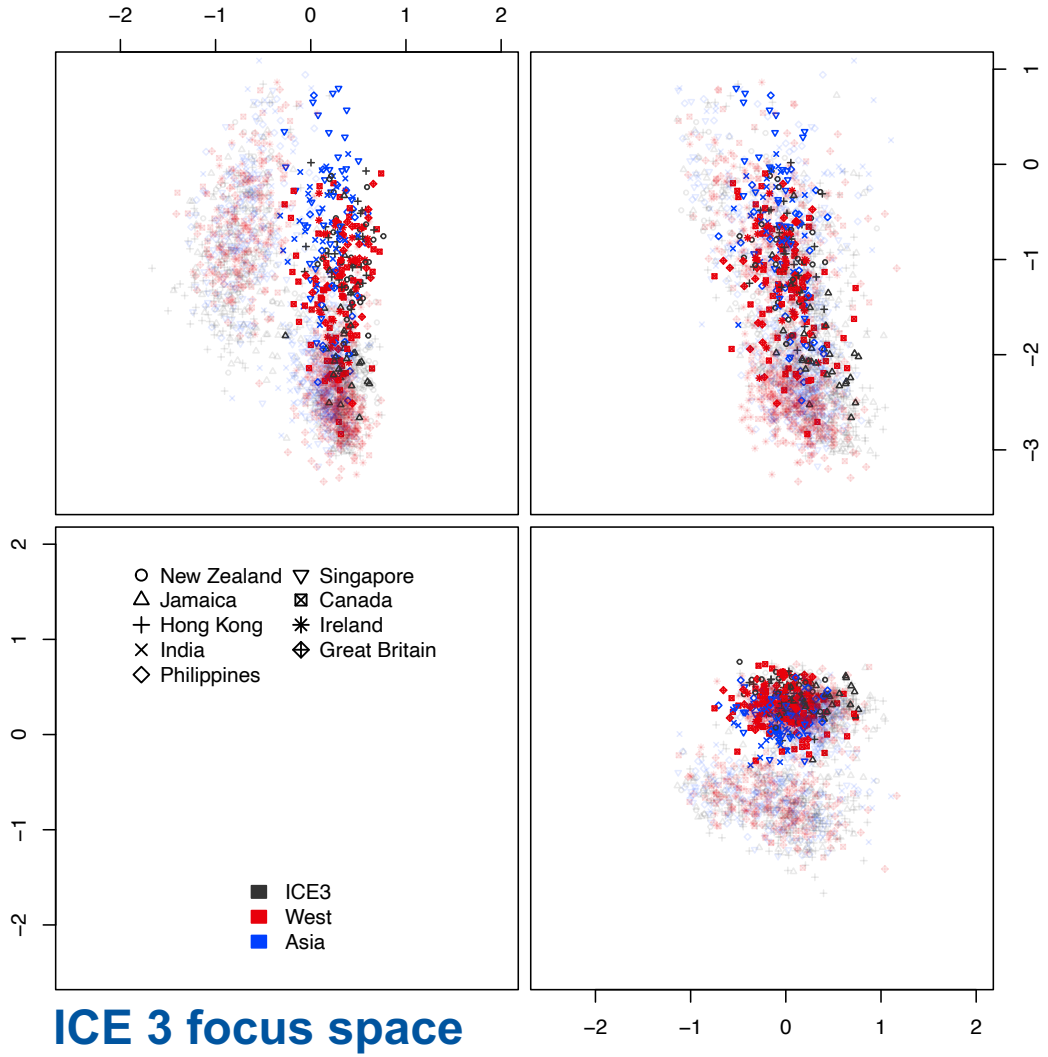
Universität Zürich[UZH]

FAU

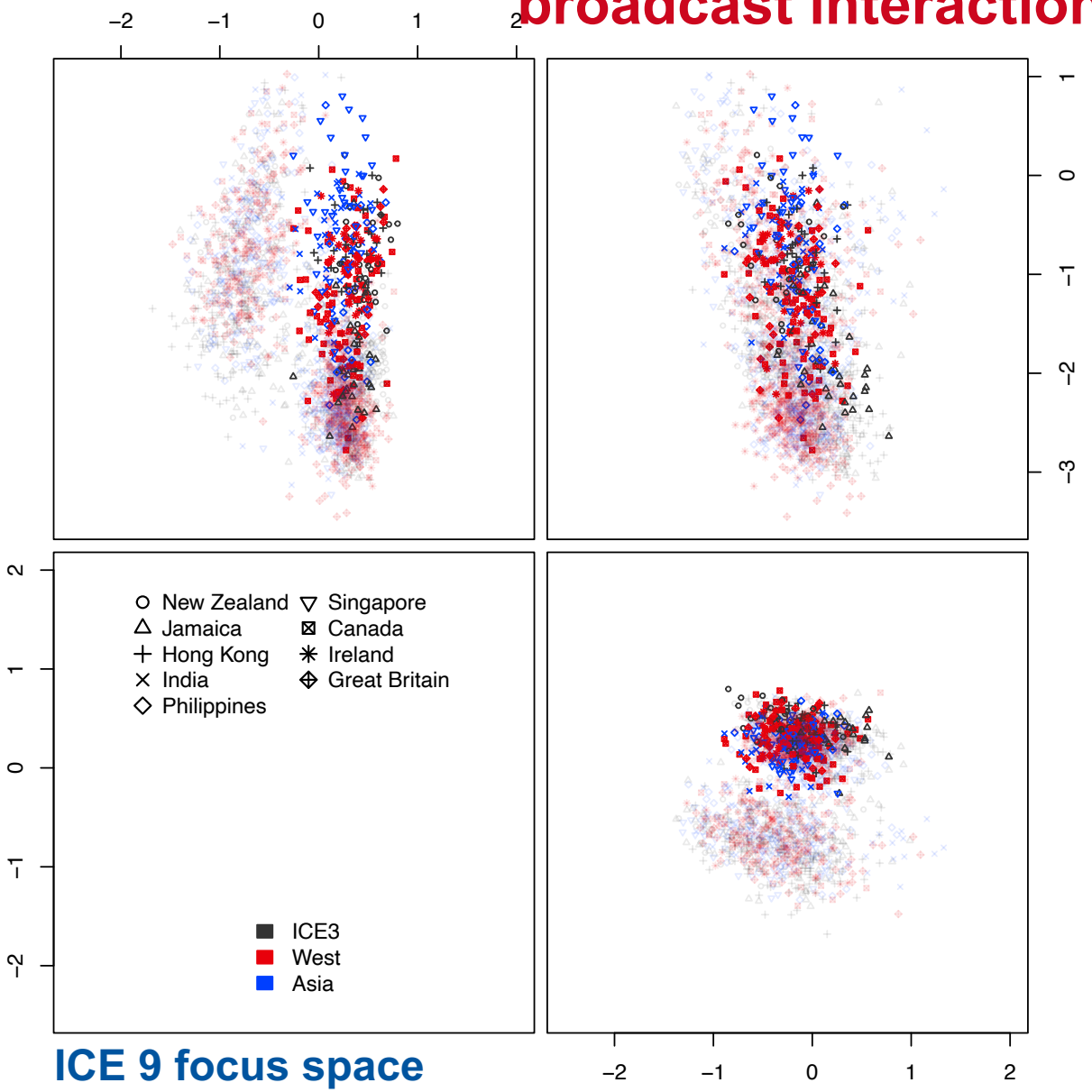
RWTH AACHEN UNIVERSITY

Divergence between varieties across registers

broadcast interactions



ICE 3 focus space



ICE 9 focus space

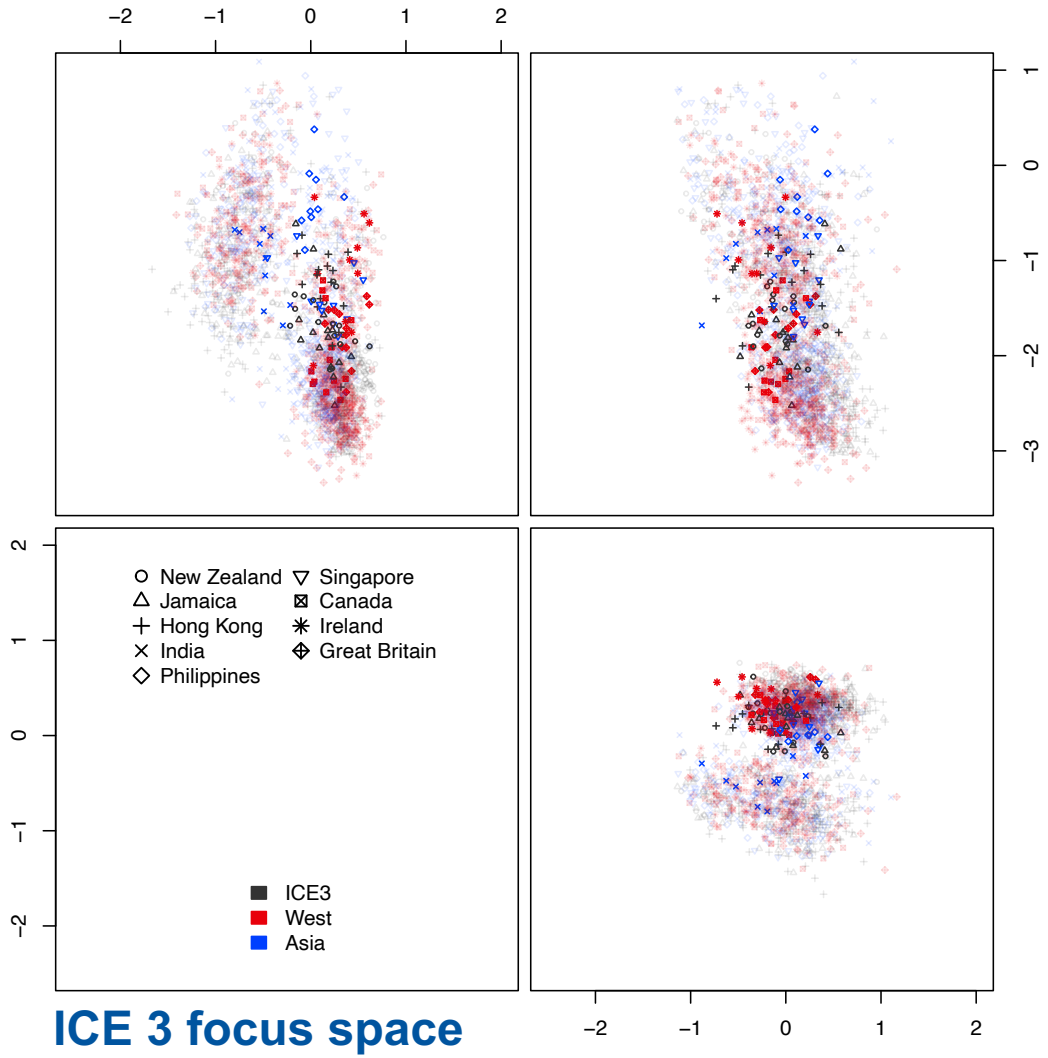
Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

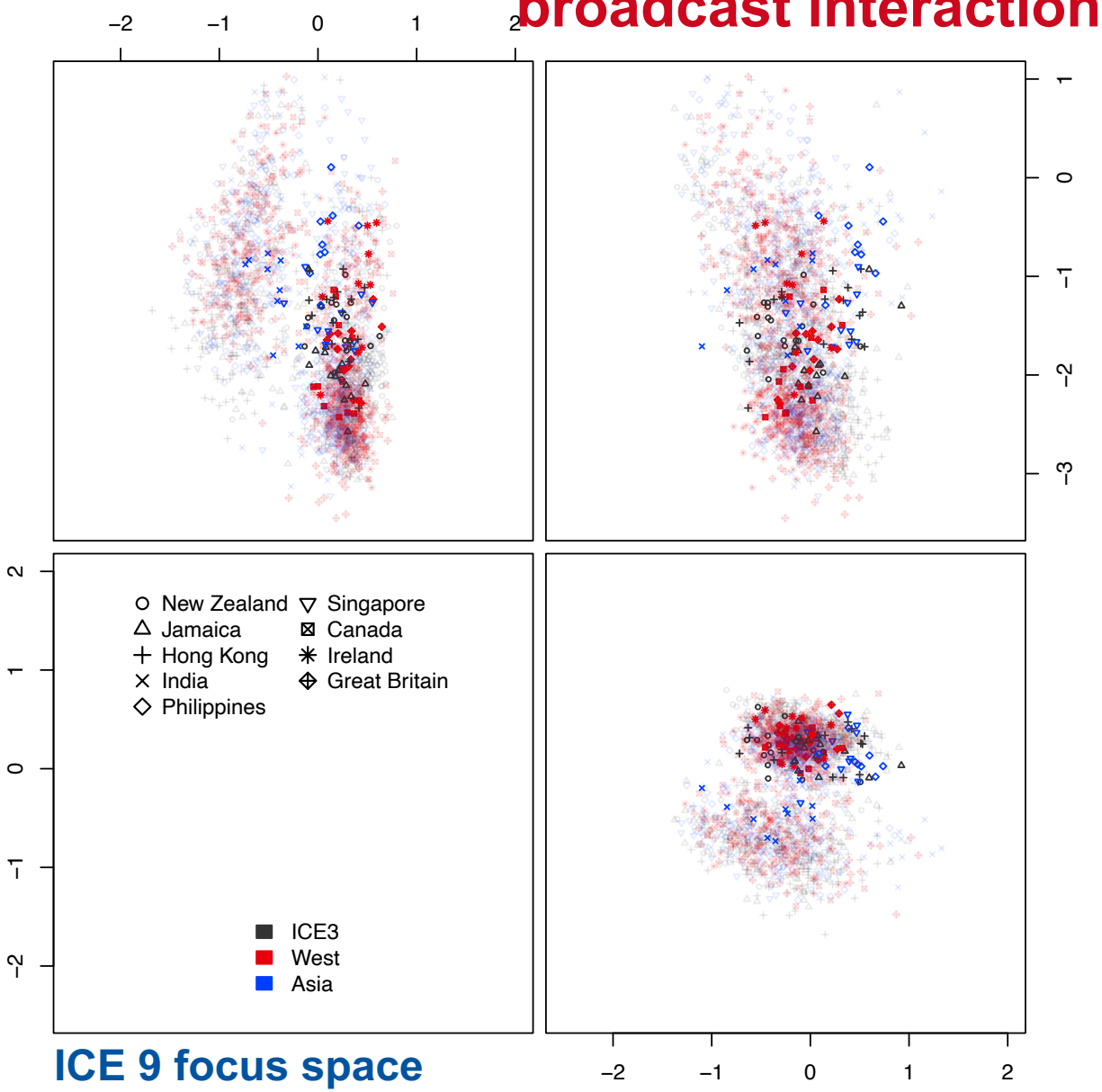
RWTH AACHEN UNIVERSITY

Divergence between varieties across registers



ICE 3 focus space

broadcast interactions



ICE 9 focus space

Evert / Frenken / Neumann / Schneider
ICAME45 | 19 June 2024
DFG Project No. 528467412
QuanTOR – Quantitative Analysis of Textual Organization across Registers

Universität Zürich[UZH]

FAU

RWTH AACHEN UNIVERSITY