Unit 7: A multivariate approach to linguistic variation

Statistics for Linguists with R – A SIGIL Course

Stephanie Evert

Computational Corpus Linguistics Group
FAU Erlangen-Nürnberg

Linguistic variation

Variation of a quantitative linguistic feature

- frequency of passive, past perfect, split infinitive, …
- frequency of expression, semantic field, topic, …
- association strength, lexical density, productivity, …

across

- languages and language varieties
- regions & social strata
- time (diachronic change)
- individual speakers & discourses

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Studying linguistic variation

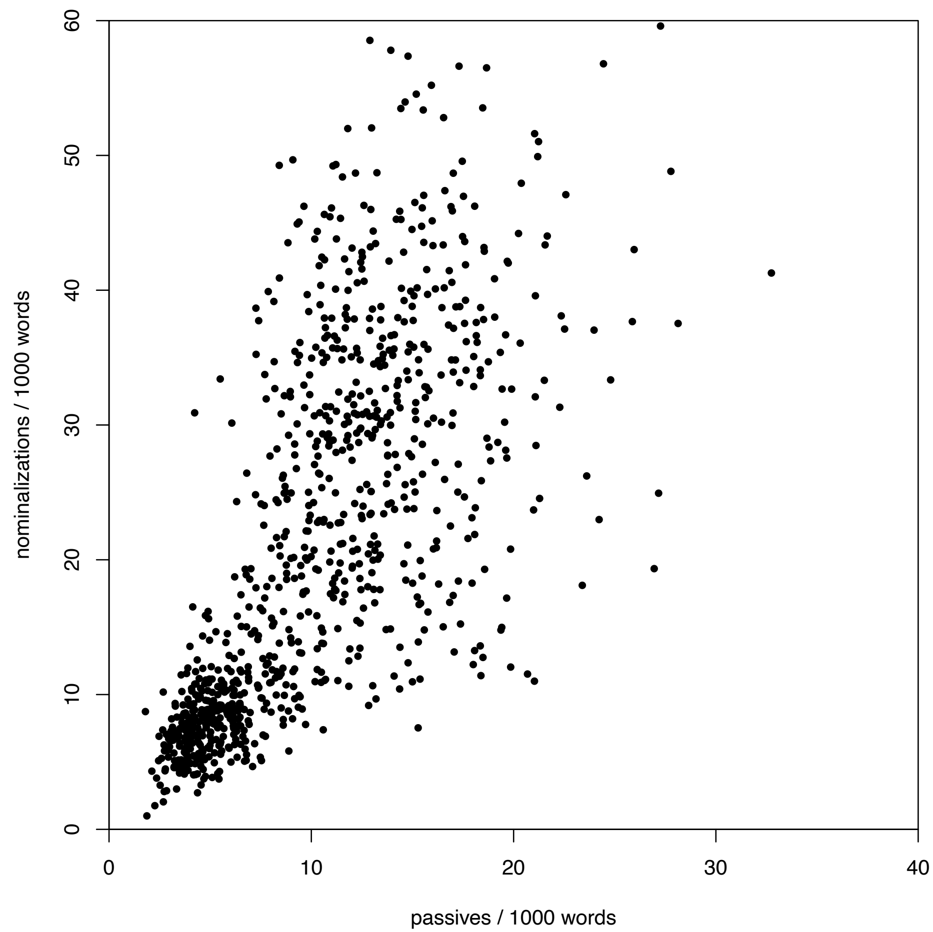
- Univariate approach
  - compare single feature across two or more conditions
  - e.g. AmE vs. BrE vs. IndE vs. … / male vs. female / etc.
  - corpus frequency comparison
- Regression approach
  - predict single quantity from multiple explanatory factors
- Multivariate approach
  - identify common patterns of variation across multiple different features ➞ correlation analysis
  - inductive techniques don't require pre-defined conditions

Variation as a nuisance parameter

- Many aspects of linguistic variation are nuisance parameters in corpus linguistics
  - e.g. difference in frequency of passives between AmE and BrE, as well as development from 1960s to 1990s (Unit #2)
  - ignore other dimensions such as genre/register variation by pooling frequency data from all texts of each corpus
  - corpus is analyzed as a random sample of VP tokens
- Consequences
  - variation → non-randomness → overestimate significance
  - discussed in much more detail in Unit #8

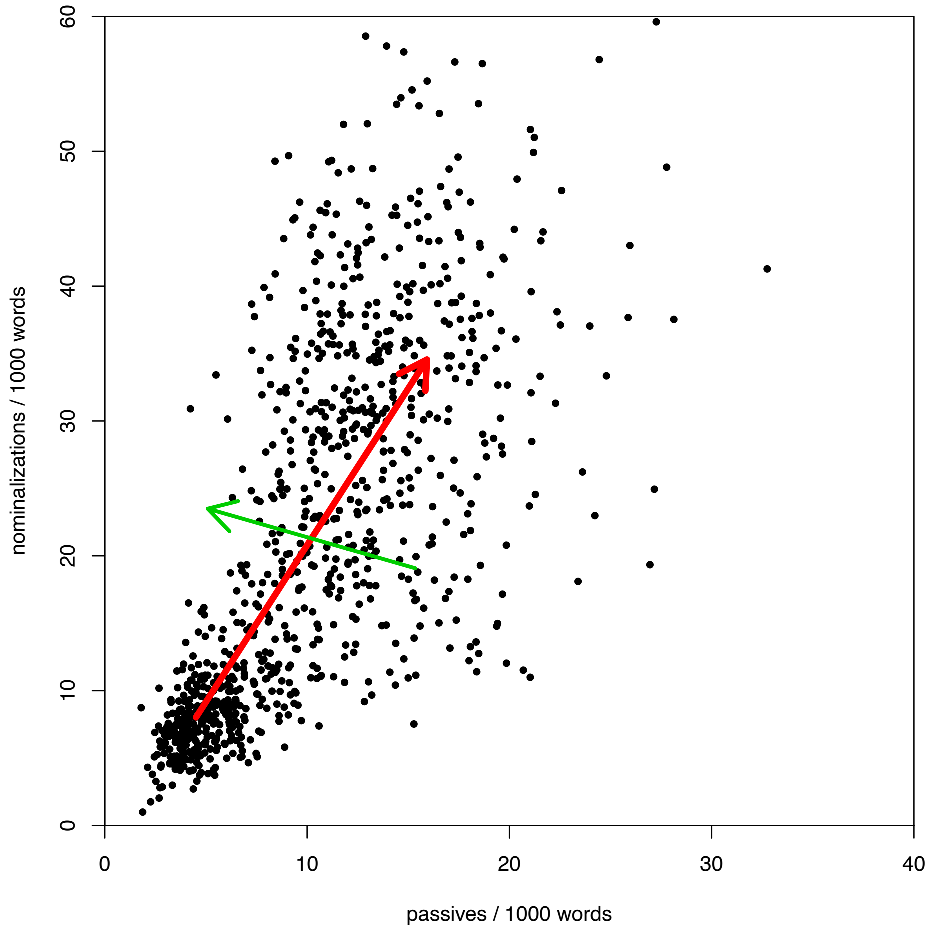
The multivariate approach

- Different linguistic features often show similar patterns of variation
- E.g. passives and nominalizations

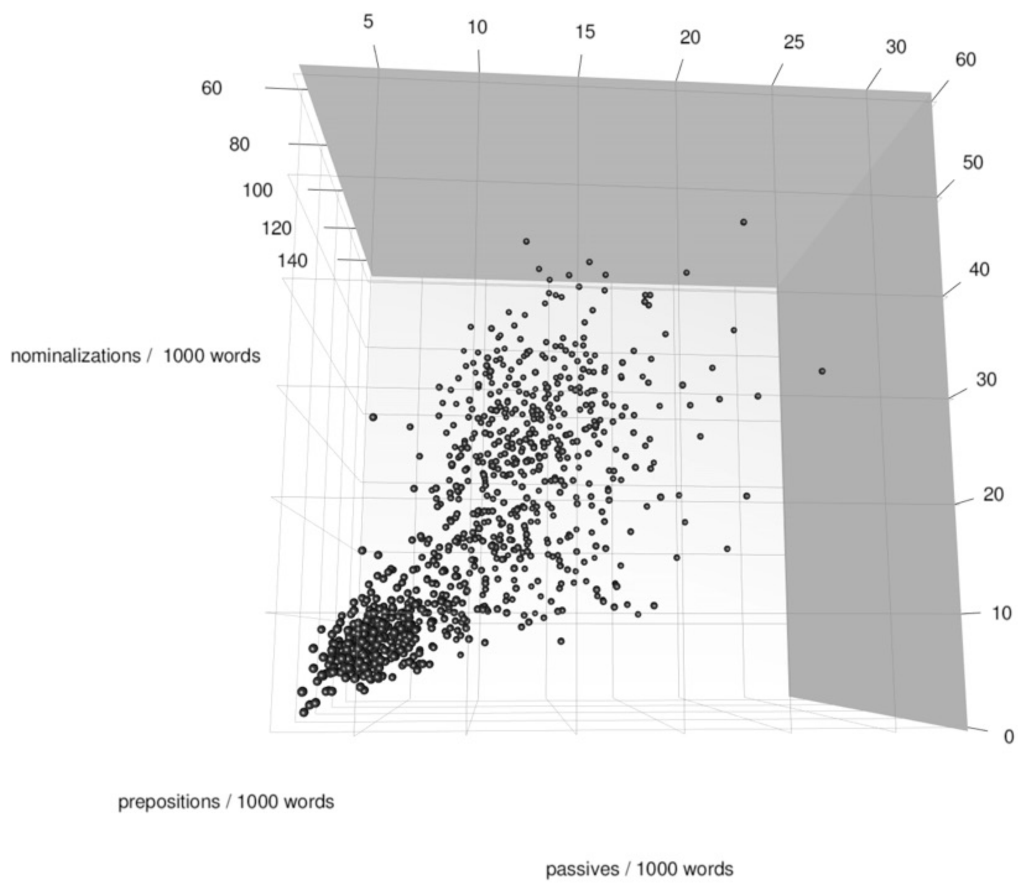


The multivariate approach

- Different linguistic features often show similar patterns of variation
- E.g. passives and nominalizations
- Such correlations can be exploited to determine major dimensions of var.



The multivariate approach



The multivariate approach

- Multivariate analysis exploits correlations between features in order to determine latent dimensions
  - interpreted as underlying "causes" of variation
- An inductive, data-driven approach
  - no theoretical assumptions about linguistic variation and categories / sub-corpora to be compared
- Pioneering work by Doug Biber (1988, 1993, 1995, …)
  - "multidimensional analysis" of register variation
- Related approaches: correspondence analysis, distributional semantics, topic modelling, …

Biber's multidimensional analysis (MDA)

Table 5.7 *Linguistic features used in the analysis of English*

A. Tense and aspect markers
- 1 Past tense
- 2 Perfect aspect
- 3 Present tense

B. Place and time adverbials
- 4 Place adverbials (e.g., *above, beside, outdoors*)
- 5 Time adverbials (e.g., *early, instantly, soon*)

C. Pronouns and pro-verbs
- 6 First-person pronouns
- 7 Second-person pronouns
- 8 Third-person personal pronouns (excluding *it*)
- 9 Pronoun *it*
- 10 Demonstrative pronouns (*that, this, these, those* as pronouns)
- 11 Indefinite pronouns (e.g., *anybody, nothing, someone*)
- 12 Pro-verb *do*

D. Questions
- 13 Direct WH questions

E. Nominal forms
- 14 Nominalizations (ending in *-tion, -ment, -ness, -ity*)
- 15 Gerunds (participial forms functioning as nouns)
- 16 Total other nouns

F. Passives
- 17 Agentless passives
- 18 *by*-passives

G. Stative forms
- 19 *be* as main verb
- 20 Existential *there*

H. Subordination features
- 21 *that* verb complements (e.g., *I said that he went.*)
- 22 *that* adjective complements (e.g., *I'm glad that you like it.*)
- 23 WH-clauses (e.g., *I believed what he told me.*)
- 24 Infinitives
- 25 Present participial adverbial clauses (e.g., *Stuffing his mouth with cookies, Joe ran out the door.*)
- 26 Past participial adverbial clauses (e.g., *Built in a single week, the house would stand for fifty years.*)
- 27 Past participial postnominal (reduced relative) clauses (e.g., *the solution produced by this process*)
- 28 Present participial postnominal (reduced relative) clauses (e.g., *The event causing this decline was . . .*)
- 29 *that* relative clauses on subject position (e.g., *the dog that bit me*)
- 30 *that* relative clauses on object position (e.g., *the dog that I saw*)
- 31 WH relatives on subject position (e.g., *the man who likes popcorn*)
- 32 WH relatives on object position (e.g., *the man who Sally likes*)
- 33 Pied-piping relative clauses (e.g., *the manner in which he was told*)

Table 5.7 (*cont.*)

- 34 Sentence relatives (e.g., *Bob likes fried mangoes, which is the most disgusting thing I've ever heard of.*)
- 35 Causative adverbial subordinator (*because*)
- 36 Concessive adverbial subordinators (*although, though*)
- 37 Conditional adverbial subordinators (*if, unless*)
- 38 Other adverbial subordinators (e.g., *since, while, whereas*)

I. Prepositional phrases, adjectives, and adverbs
- 39 Total prepositional phrases
- 40 Attributive adjectives (e.g., *the big horse*)
- 41 Predicative adjectives (e.g., *The horse is big.*)
- 42 Total adverbs

J. Lexical specificity
- 43 Type–token ratio
- 44 Mean word length

K. Lexical classes
- 45 Conjuncts (e.g., *consequently, furthermore, however*)
- 46 Downtoners (e.g., *barely, nearly, slightly*)
- 47 Hedges (e.g., *at about, something like, almost*)
- 48 Amplifiers (e.g., *absolutely, extremely, perfectly*)
- 49 Emphatics (e.g., *a lot, for sure, really*)
- 50 Discourse particles (e.g., sentence-initial *well, now, anyway*)
- 51 Demonstratives

L. Modals
- 52 Possibility modals (*can, may, might, could*)
- 53 Necessity modals (*ought, should, must*)
- 54 Predictive modals (*will, would, shall*)

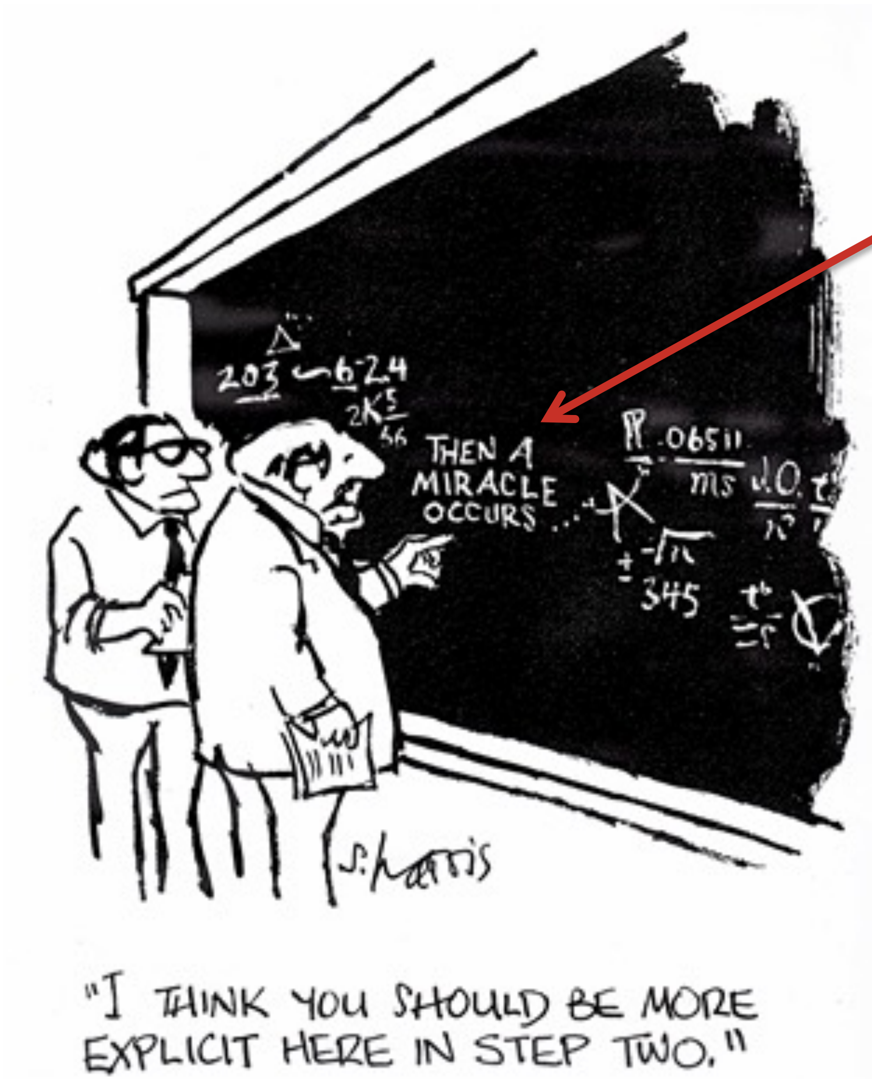
M. Specialized verb classes
- 55 Public verbs (e.g., *assert, declare, mention*)
- 56 Private verbs (e.g., *assume, believe, doubt, know*)
- 57 Suasive verbs (e.g., *command, insist, propose*)
- 58 *seem* and *appear*

N. Reduced forms and dispreferred structures
- 59 Contractions
- 60 Subordinator *that* deletion (e.g., *I think* [*that*] *he went.*)
- 61 Stranded prepositions (e.g., *the candidate that I was thinking of*)
- 62 Split infinitives (e.g., *He wants to convincingly prove that . . .*)
- 63 Split auxiliaries (e.g., *They were apparently shown to . . .*)

O. Co-ordination
- 64 Phrasal co-ordination (NOUN *and* NOUN; ADJ; *and* ADJ; VERB *and* VERB; ADV *and* ADV)
- 65 Independent clause co-ordination (clause-initial *and*)

P. Negation
- 66 Synthetic negation (e.g., *No answer is good enough for Jones.*)
- 67 Analytic negation (e.g., *That's not likely*)

Biber's MDA



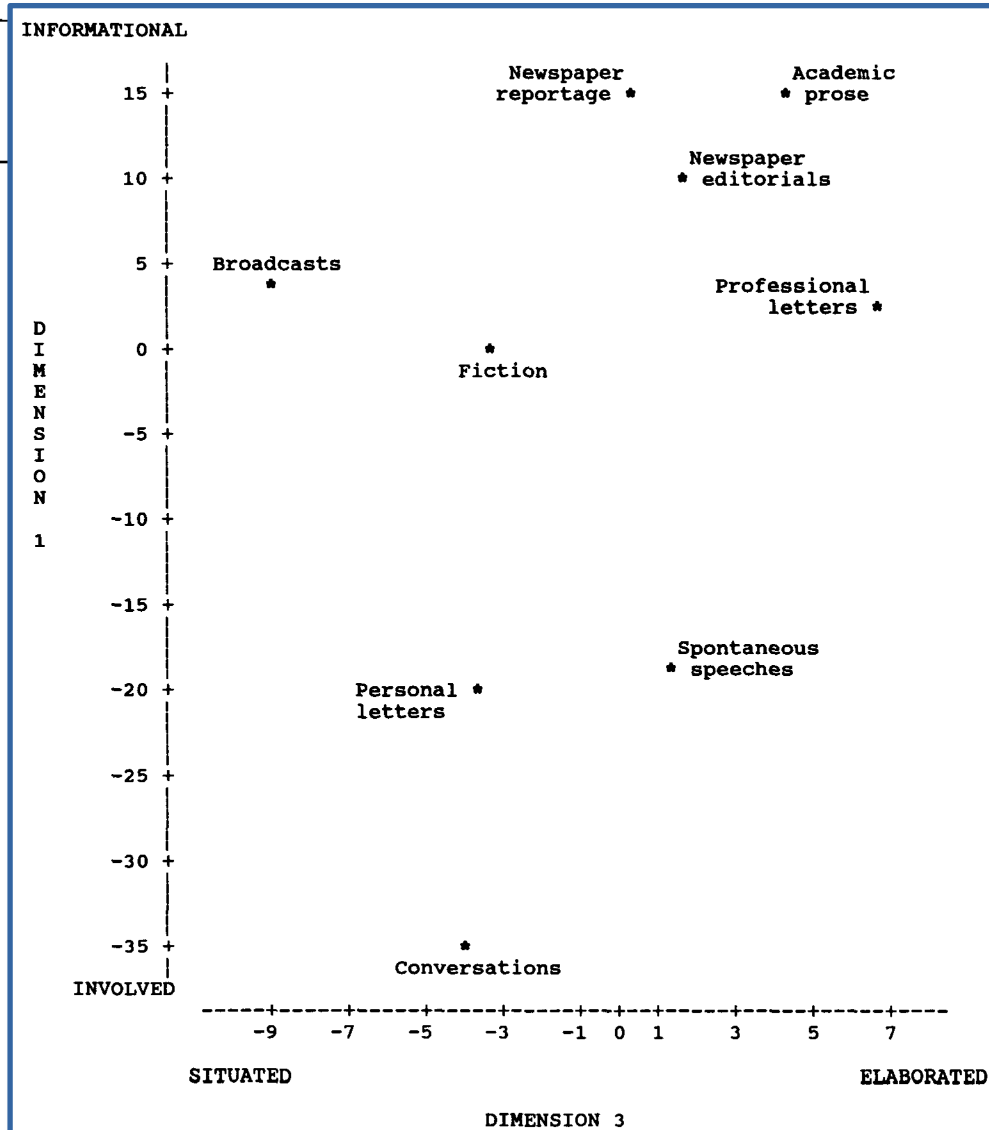
factor analysis (FA)

Biber's MDA

TABLE 2

Summary of the co-occurrence patterns underlying five major dimensions of English.

DIMENSION 1 (Informational vs. Involved)		DIMENSION 2 (Narrative versus Non-Narrative)	
nouns	0.80	past tense verbs	0.90
word length	0.58	third person pronouns	0.73
prepositional phrases	0.54	perfect aspect verbs	0.48
type / token ratio	0.54	public verbs	0.43
attributive adjs.	0.47	synthetic negation	0.40
		present participial clauses	0.39
private verbs	−0.96	present tense verbs	−0.47
that deletions	−0.91	attributive adjs.	−0.41
contractions	−0.90		
present tense verbs	−0.86		
2nd person pronouns	−0.86		
do as pro-verb	−0.82		
analytic negation	−0.78		
demonstrative pronouns	−0.76		
general emphatics	−0.74		
first person pronouns	−0.74		
pronoun *it*	−0.71		
be as main verb	−0.71		
causative subordination	−0.66		
discourse particles	−0.66		
indefinite pronouns	−0.62		
general hedges	−0.58		
amplifiers	−0.56		
sentence relatives	−0.55		
WH questions	−0.52		
possibility modals	−0.50		
non-phrasal coordination	−0.48		
WH clauses	−0.47		
final prepositions	−0.43		



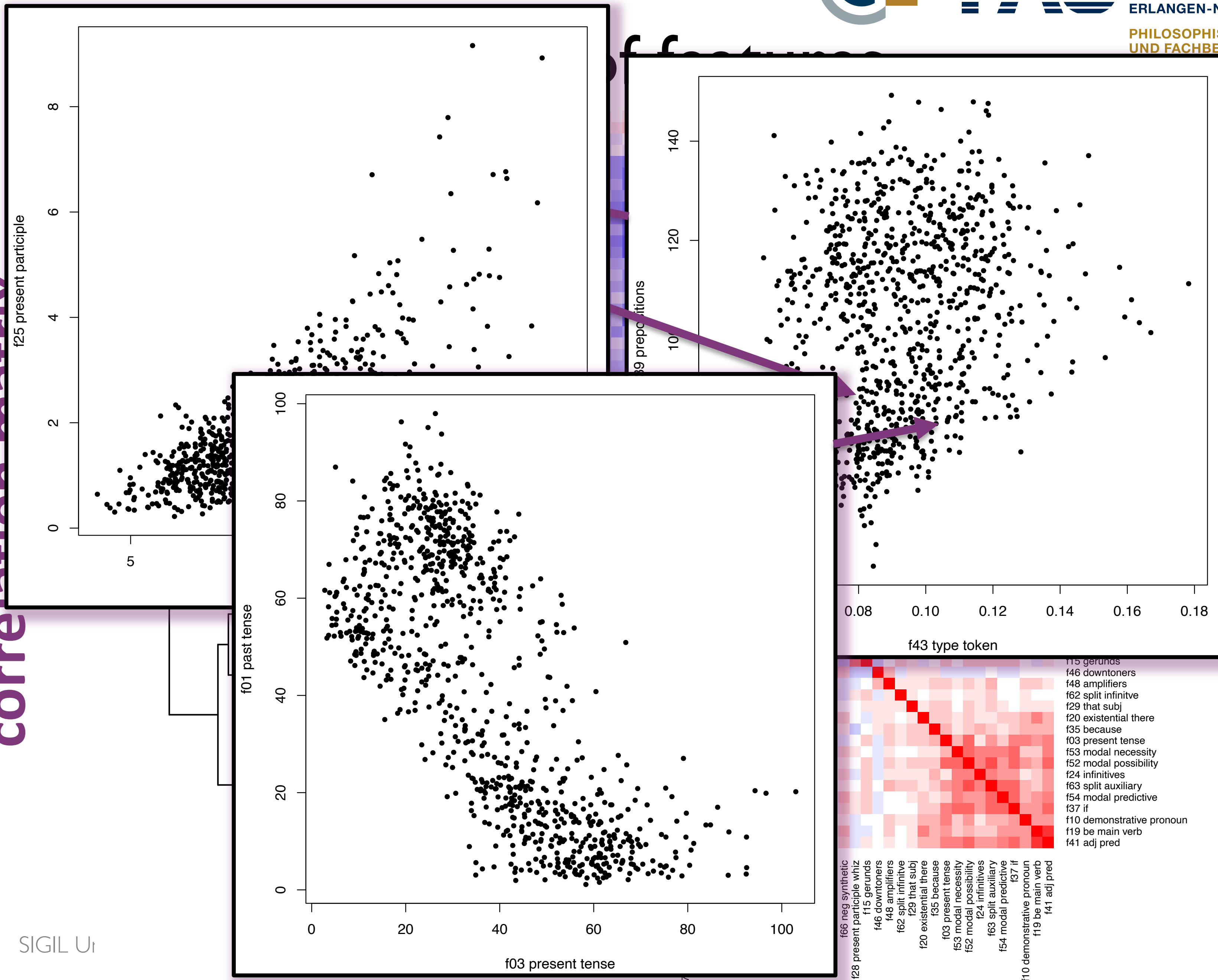
Pitfalls

- Design bias: choice of quantitative features
- Design bias: selection of text samples
- Involves a miracle
  - not clear what quantitative patterns are captured by FA
  - magic number: how many factor dimensions?
- Interpretation bias
  - arbitrary cutoff for feature weights ("loadings")
  - risk of reading one's own expectations into features
- More subtle patterns of variation invisible
- Significance & reproducibility of results?

Reproducing Biber's dimensions

- Sample of 923 medium-length published texts from written part of British National Corpus (BNC)
- Covers 4 different text types + male/female authors
  - academic writing, non-academic prose, fiction, misc.
- Biber features extracted automatically with Python script (Gasthaus 2007)
  - all frequencies normalized per 1000 words
  - data available in R package `corpora` (`BNCbiber`)
- Factor analysis with 4 latent dimensions + varimax
  - seems to yield the most clearly structured analysis

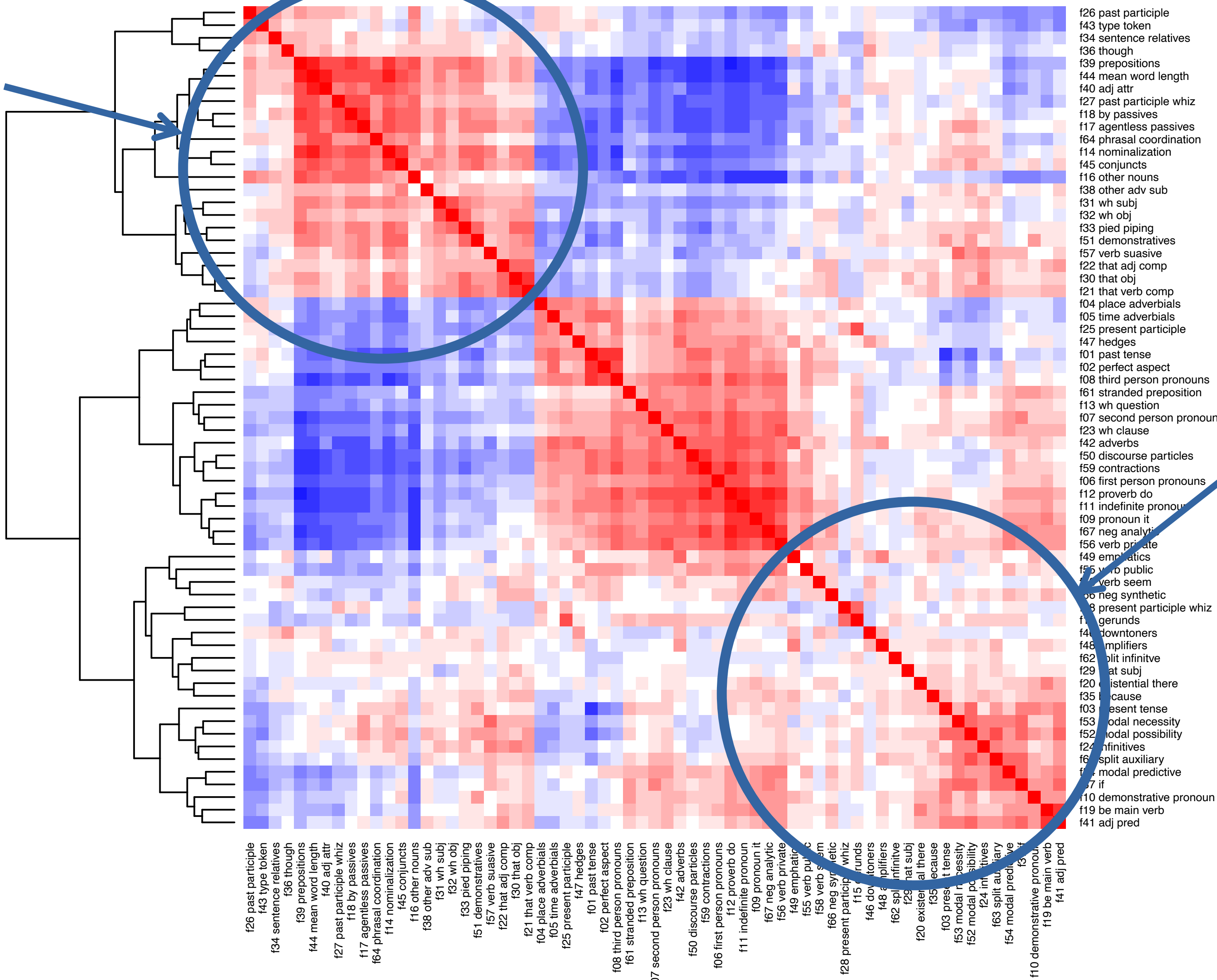
correlation matrix



Design bias: choice of features

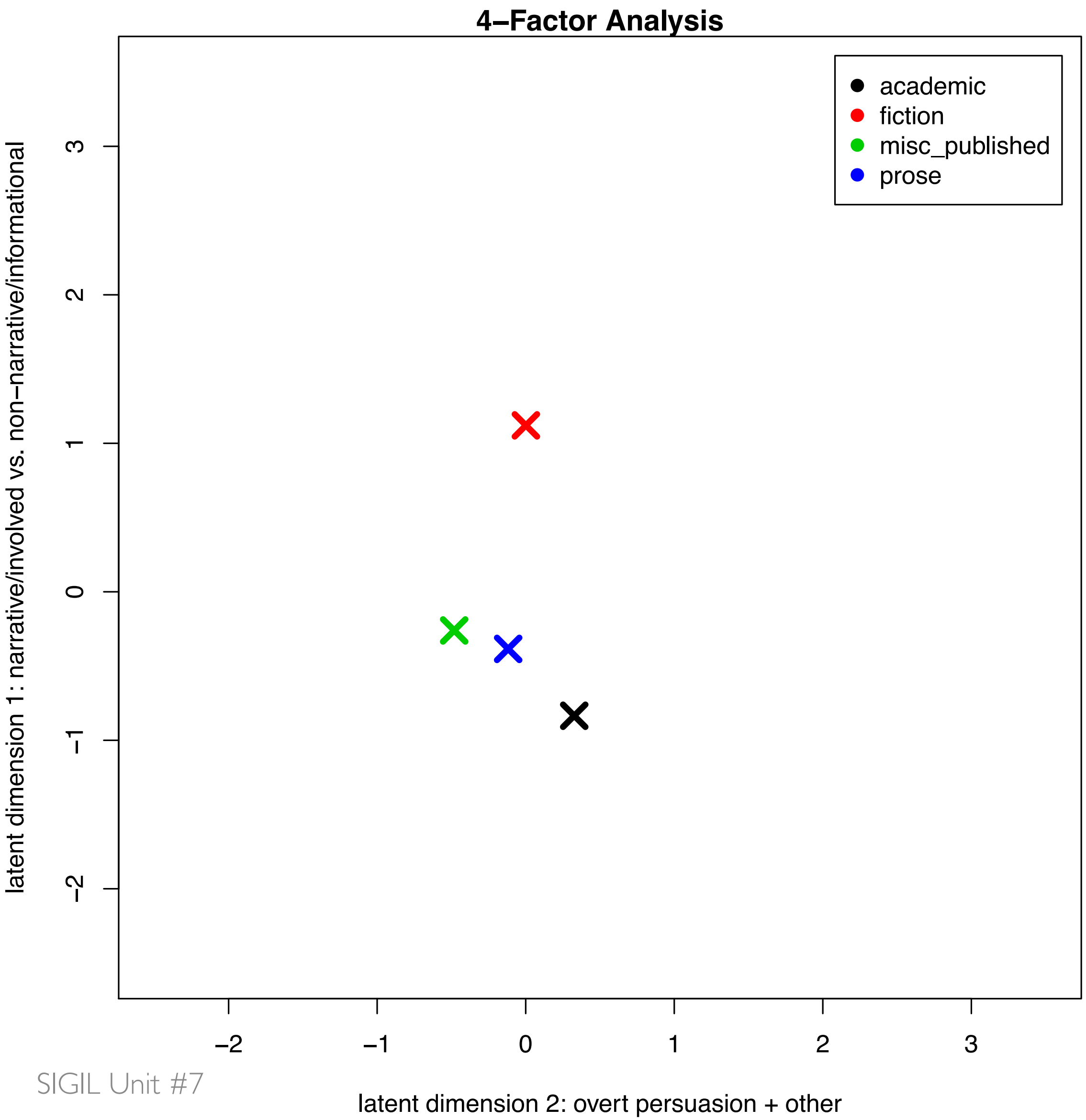
correlated with noun frequency

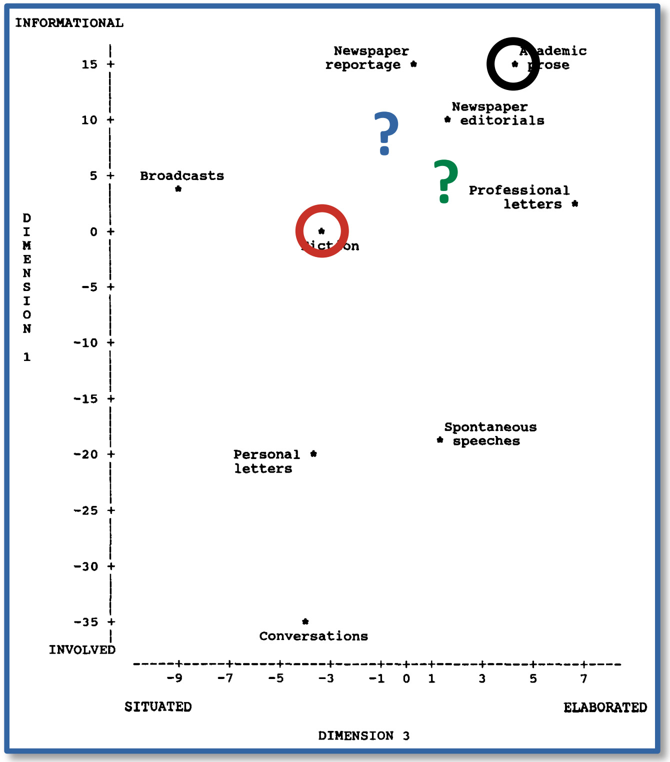
correlation matrix



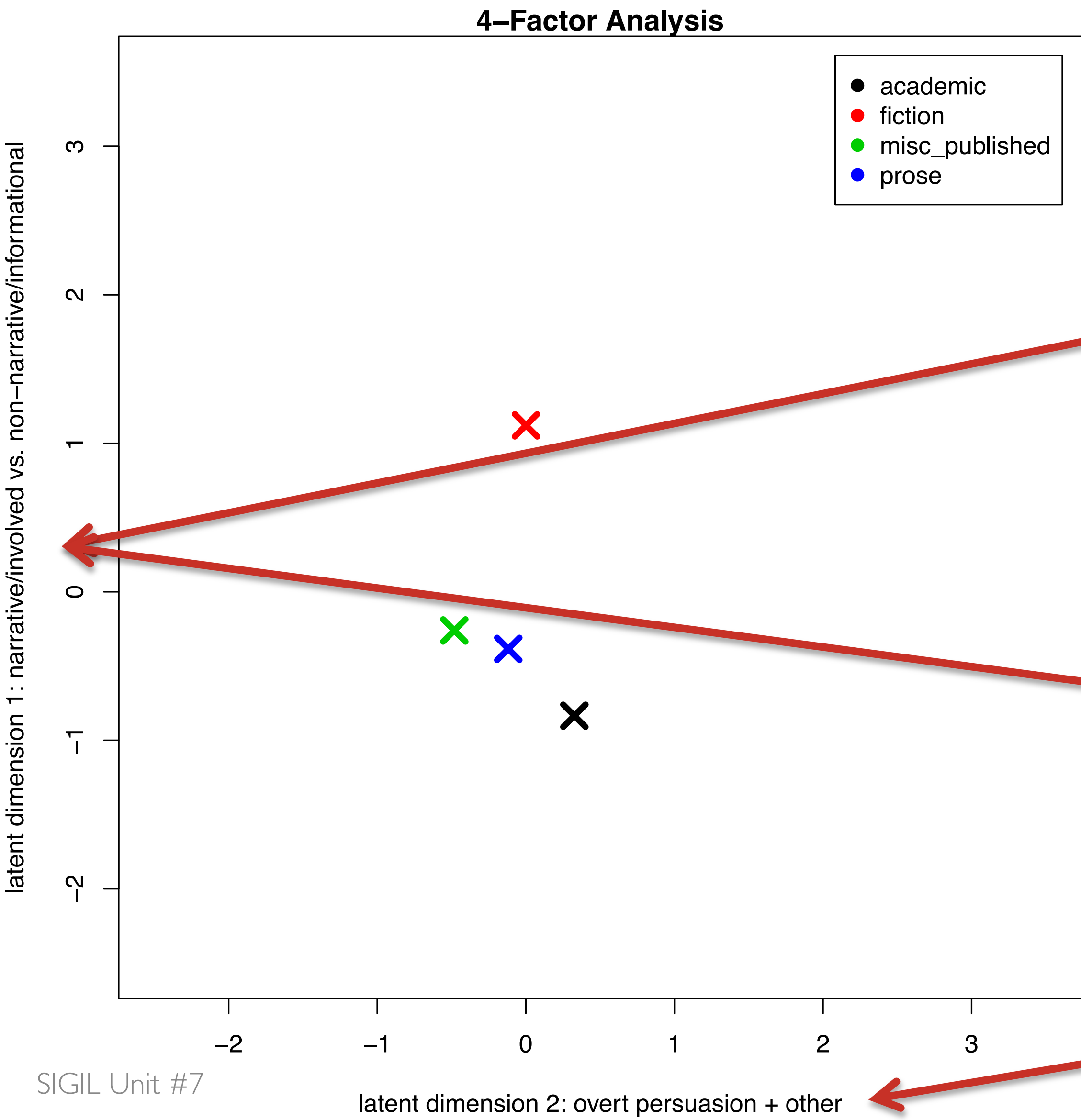
correlated with verb frequency (all feat's measured per 1000 words)

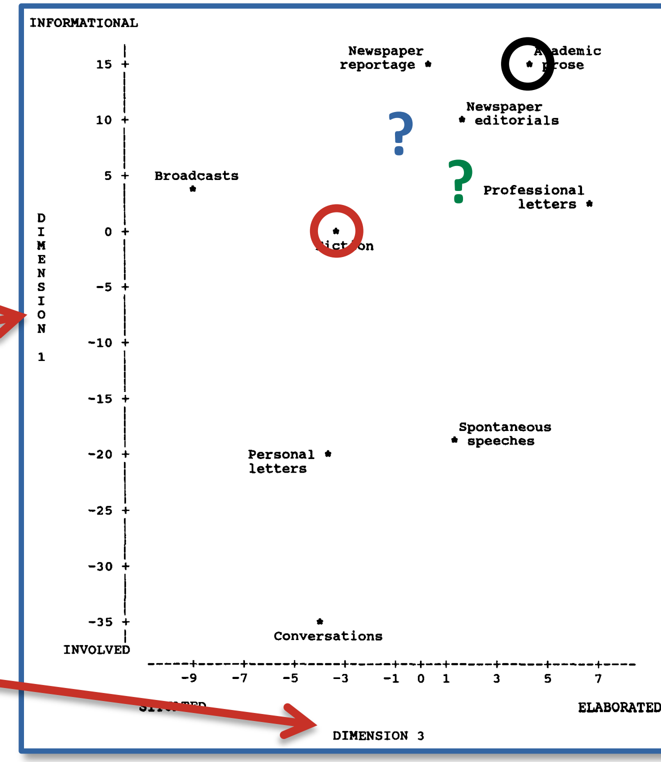
Design bias: choice of text samples





Interpretation bias





?

Interpretation bias

TABLE 2
Summary of the co-occurrence patterns underlying five major dimensions of English.

DIMENSION 1 (Informational vs. Involved)	
nouns	0.80
word length	0.58
prepositional phrases	0.54
type / token ratio	0.54
attributive adjs.	0.47
private verbs	−0.96
that deletions	−0.91
contractions	−0.90
present tense verbs	−0.86
2nd person pronouns	−0.86
do as pro-verb	−0.82
analytic negation	−0.78
demonstrative pronouns	−0.76
general emphatics	−0.74
first person pronouns	−0.74
pronoun *it*	−0.71
be as main verb	−0.71
causative subordination	−0.66
discourse particles	−0.66
indefinite pronouns	−0.62
general hedges	−0.58

DIMENSION 2 (Narrative versus Non-Narrative)	
past tense verbs	0.90
third person pronouns	0.73
perfect aspect verbs	0.48
public verbs	0.43
synthetic negation	0.40
present participial clauses	0.39
present tense verbs	−0.47
attributive adjs.	−0.41

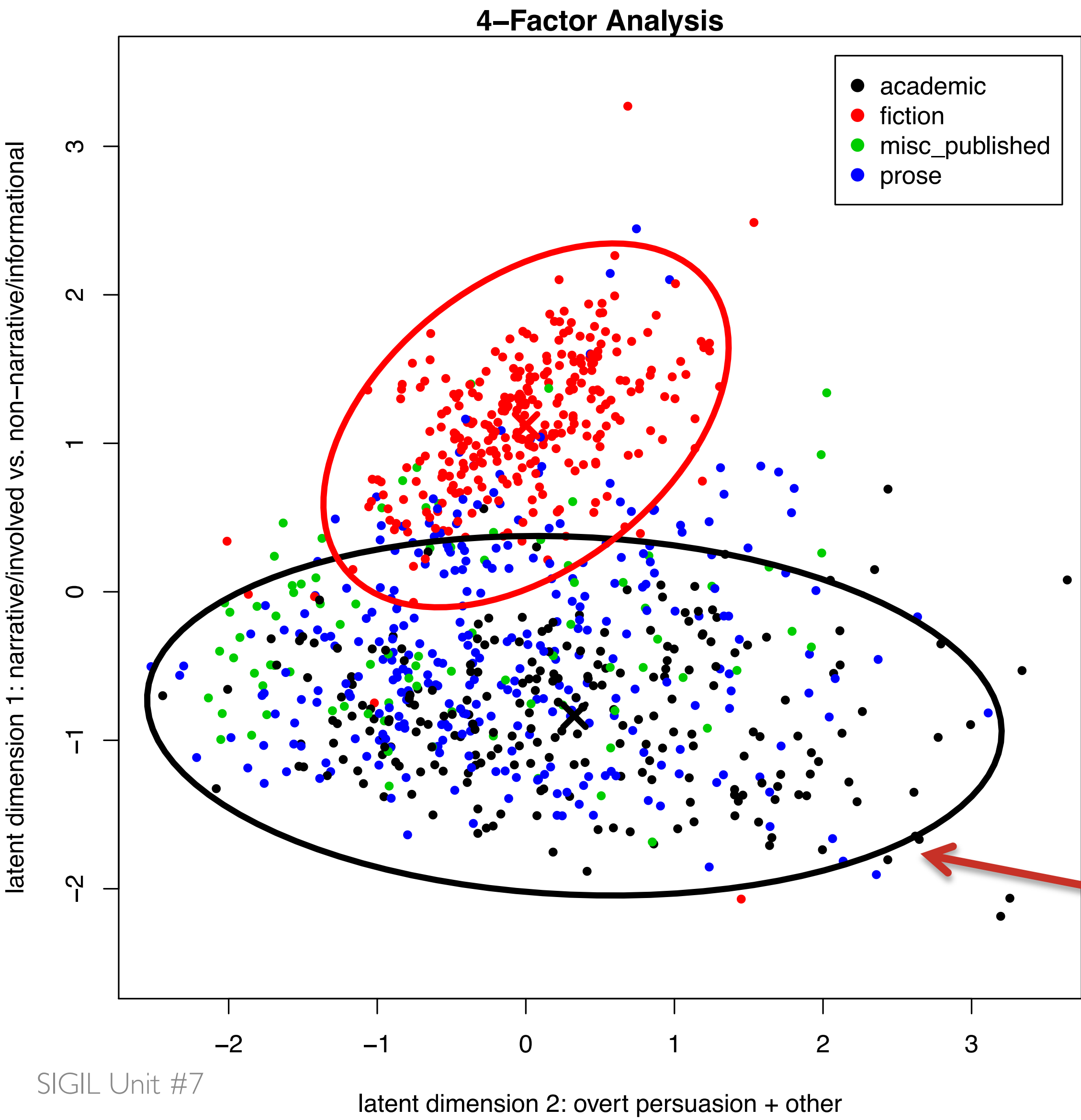
DIMENSION 3 (Elaborated vs. Situated Reference)	
WH relative clauses on object positions	0.63
pied piping constructions	0.61
WH relative clauses on subject position	0.45
phrasal coordination	0.36
nominalizations	0.36
time adverbials	−0.60
place adverbials	−0.49
other adverbs	−0.46

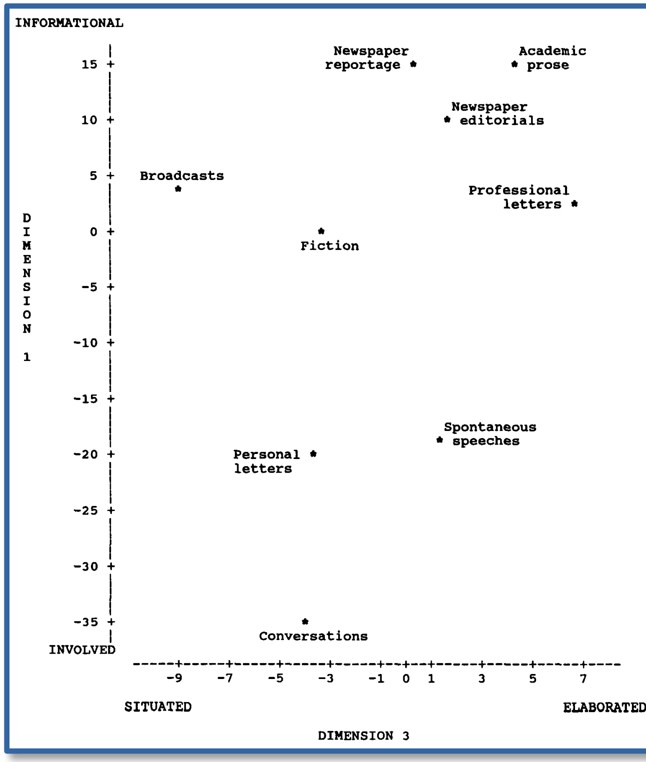
DIMENSION 4 (Overt Expression of Persuasion)	
infinitives	0.76
prediction modals	0.54
suasive verbs	0.49
conditional subordination	0.47
necessity modals	0.46
split auxiliaries	0.44
possibility modals	0.37
[No complementary features]	

DIM (Abs Non-A
conjun
agentle
past pa claus
BY-pas
past pa WHI
other a subo
[No co

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Variation between texts is ignored

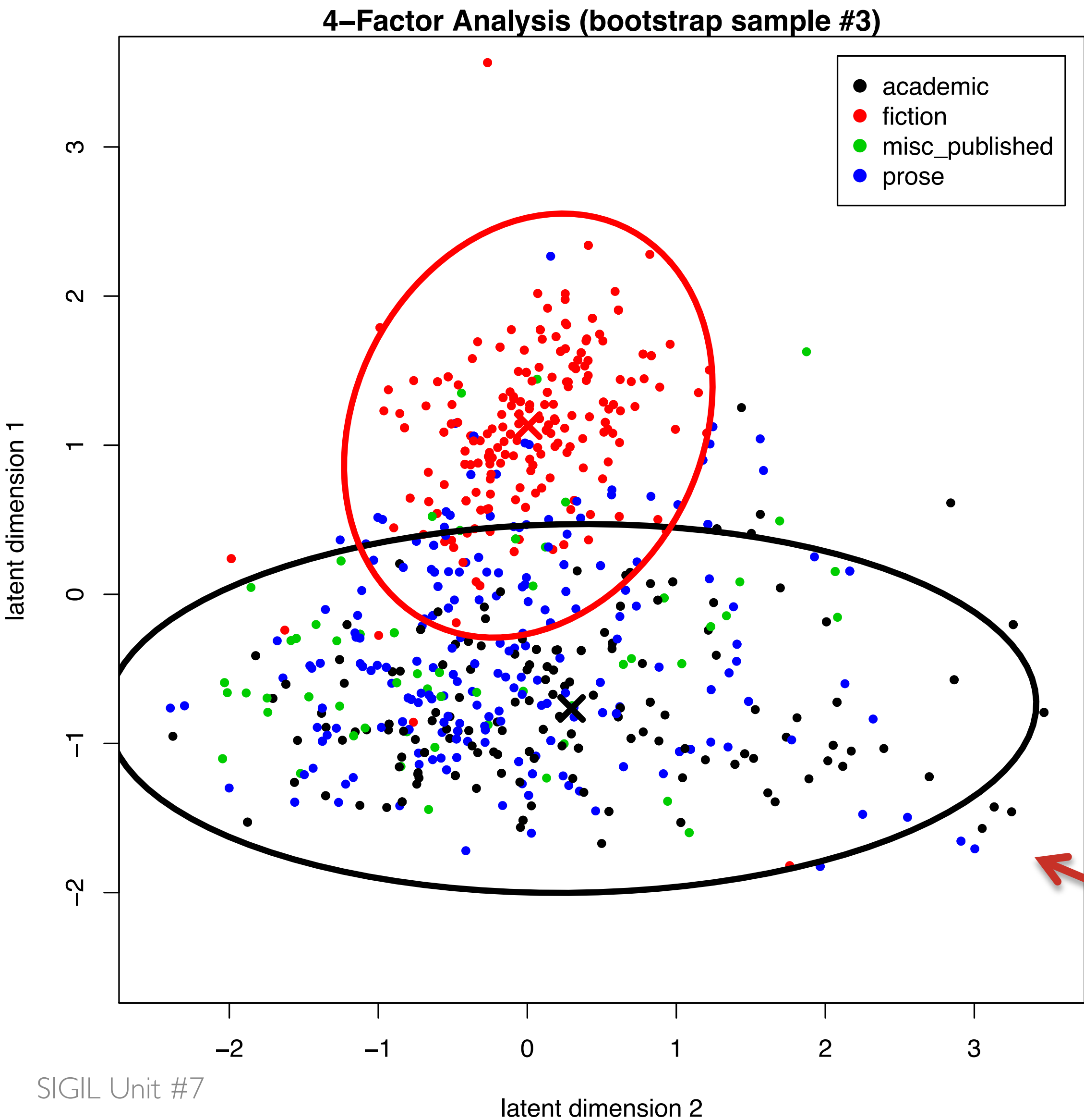


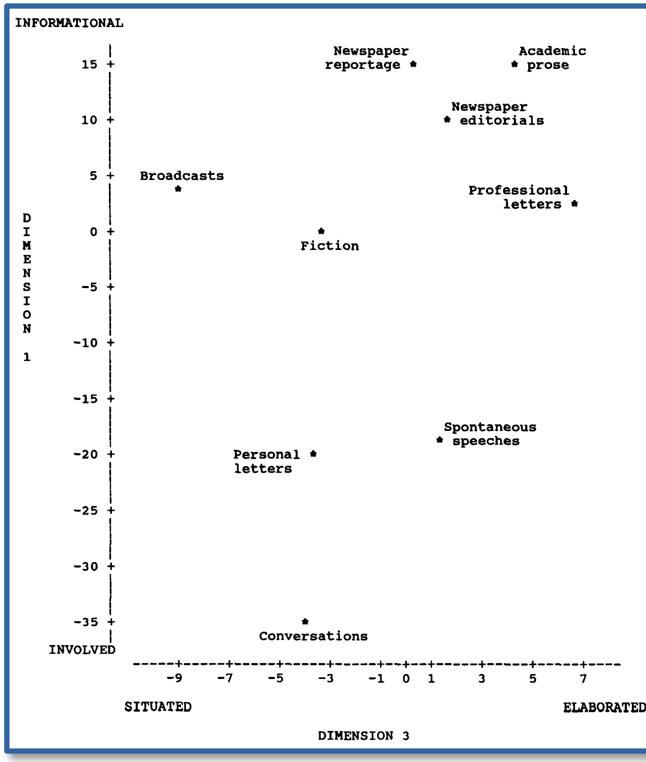


“confidence” ellipse

(→ significance)

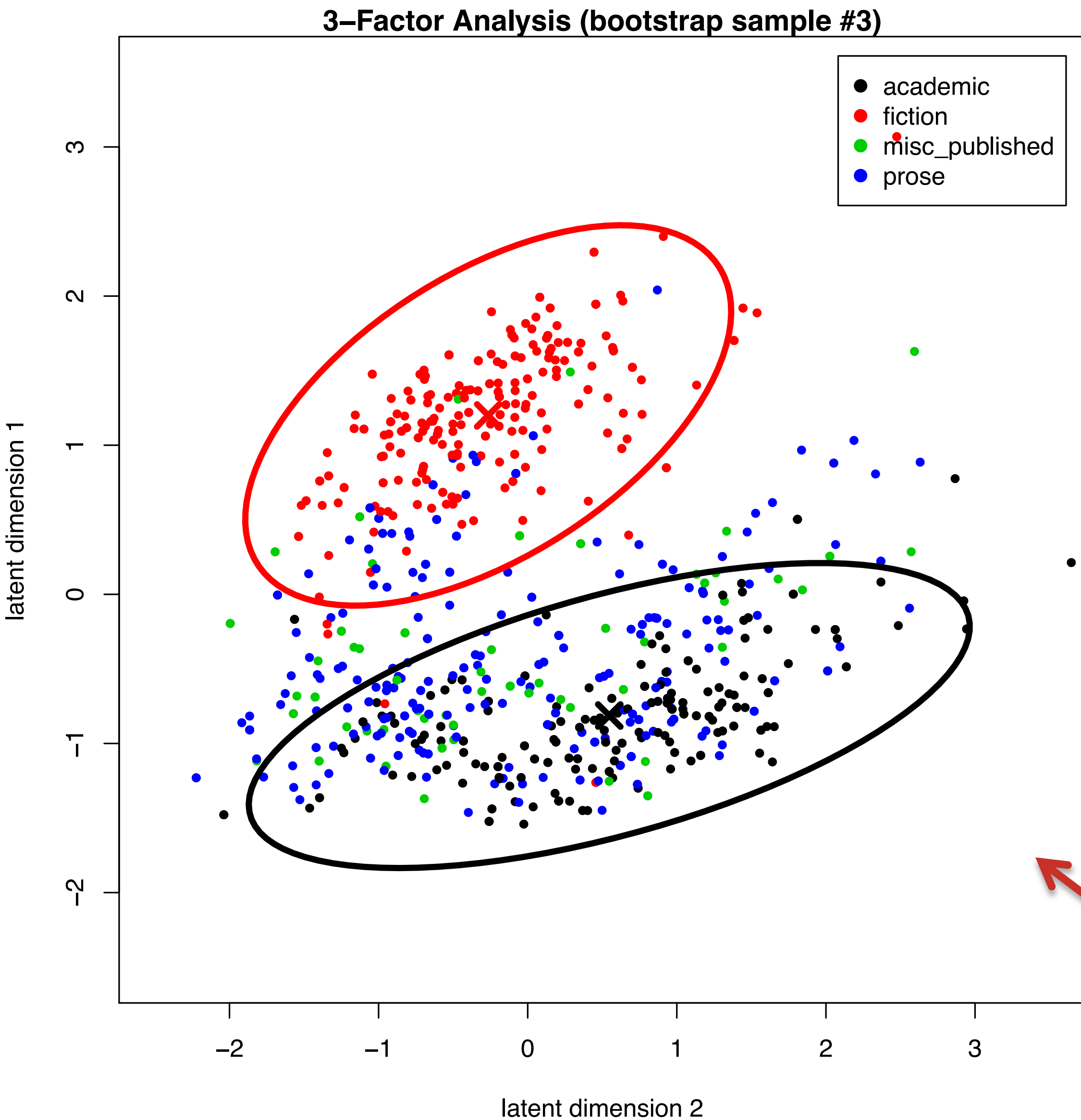
Design bias: choice of texts (redux)

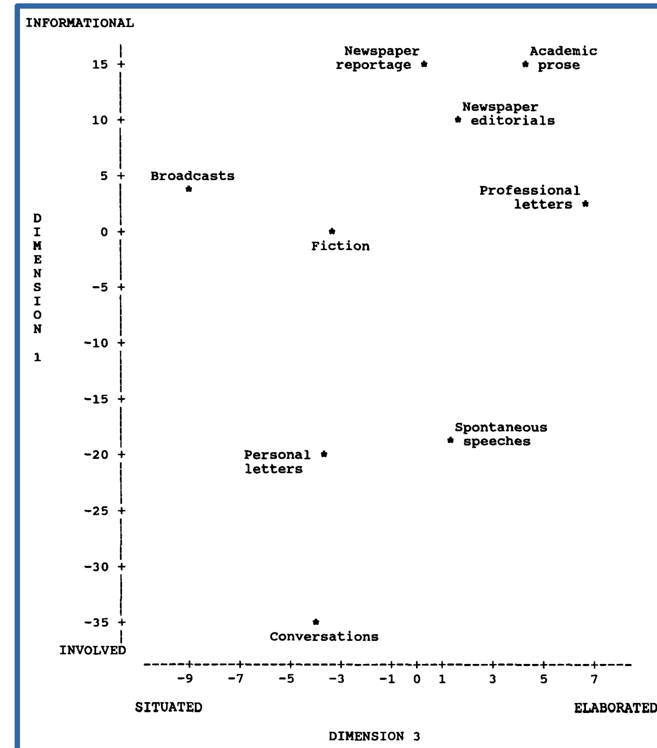




Bootstrapping

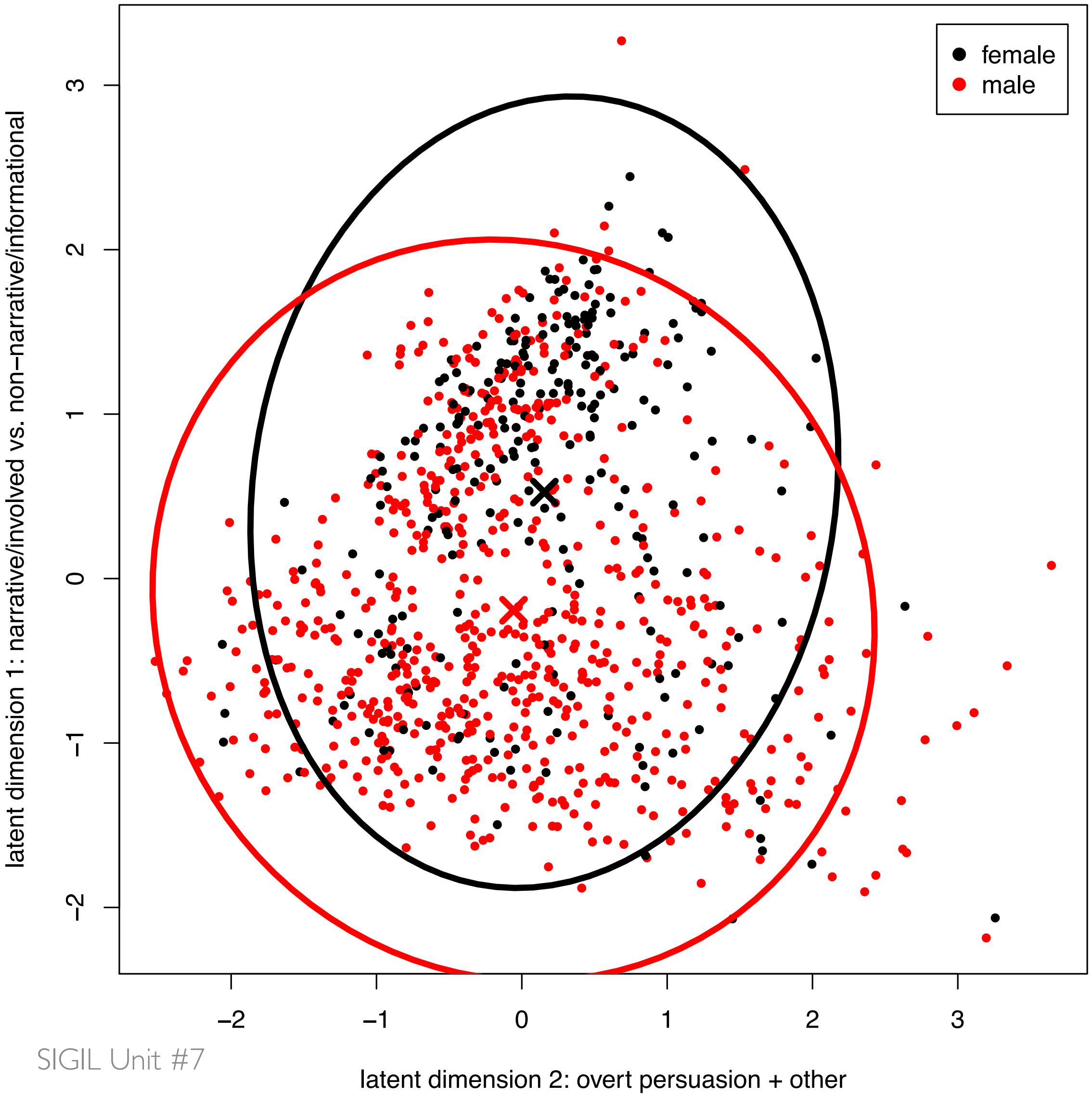
And there's the magic number …





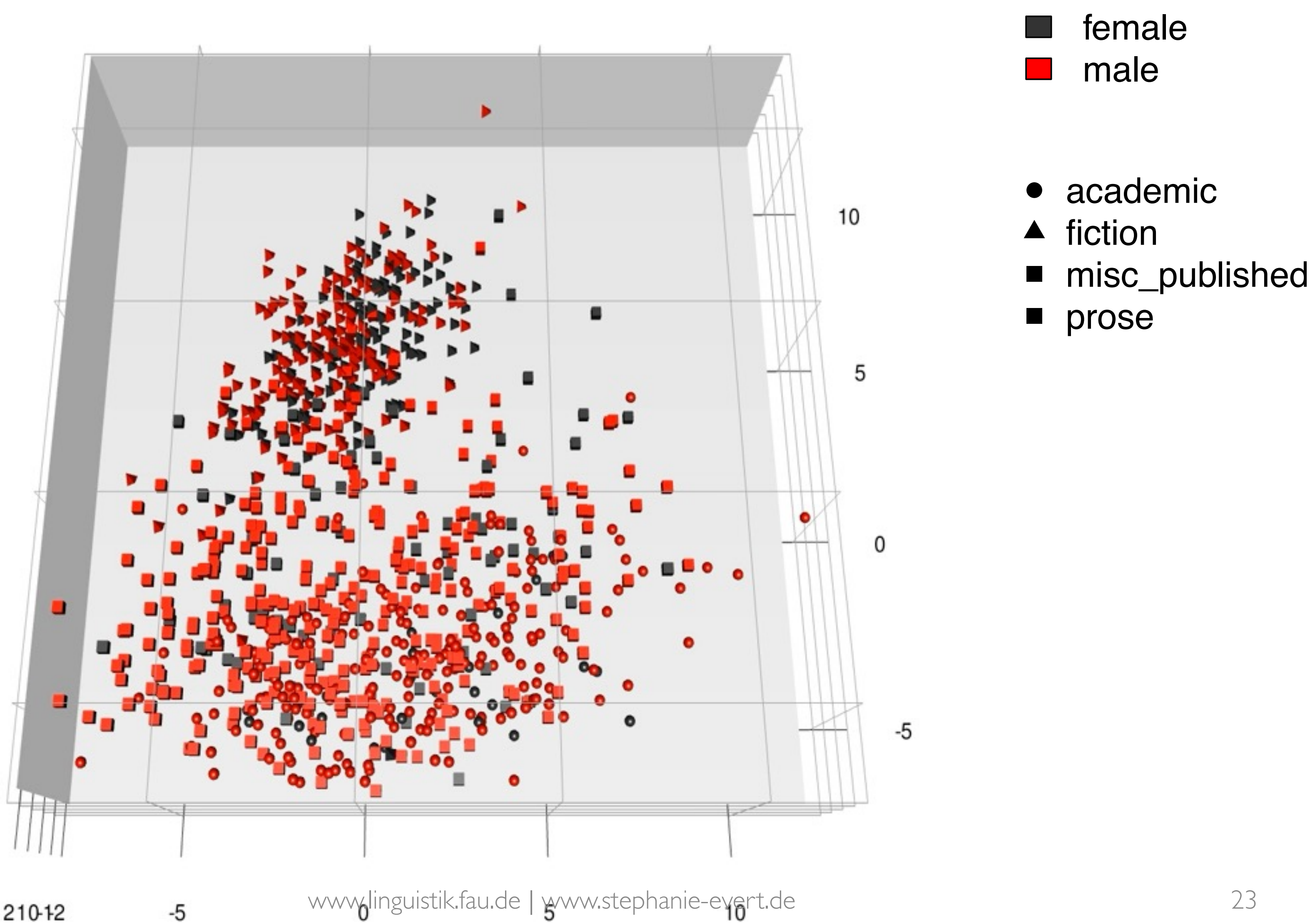
3-factor analysis
(instead of 4 factors)

Blindness to subtle patterns



- But research shows that author gender can be identified with high accuracy
  - Koppel et al. (2003): 77.3% with function words + POS n-grams
  - Gasthaus (2007): 82.9% with SVM on Biber features
- This dataset: 82.3% accuracy
  - baseline: 73.1%

Blindness to subtle patterns



Geometric Multivariate Analysis

(Diwersy, Evert & Neumann 2014; Evert & Neumann 2017; Neumann & Evert 2021)

Online supplements:

https://www.stephanie-evert.de/PUB/EvertNeumann2017/

https://www.stephanie-evert.de/PUB/NeumannEvert2021/

Geometric Multivariate Analysis

(Diwersy, Evert & Neumann 2014; Evert & Neumann 2017; Neumann & Evert 2021)

- Axiom: (Euclidean) distance = similarity of texts
  - depends crucially on theoretically motivated features
- Visualization ➞ interpret geometric configuration
  - latent dimensions as geometric projections
  - orthogonal projection = perspective on data
  - method: principal component analysis (PCA)
- Minimally supervised intervention
  - based on externally observable, theory-neutral information
  - method: linear discriminant analysis (LDA)
- Bootstrapping / cross-validation to assess significance
- Cautious interpretation of feature weights

Case study: CroCo

(Neumann 2013; Evert & Neumann 2017)

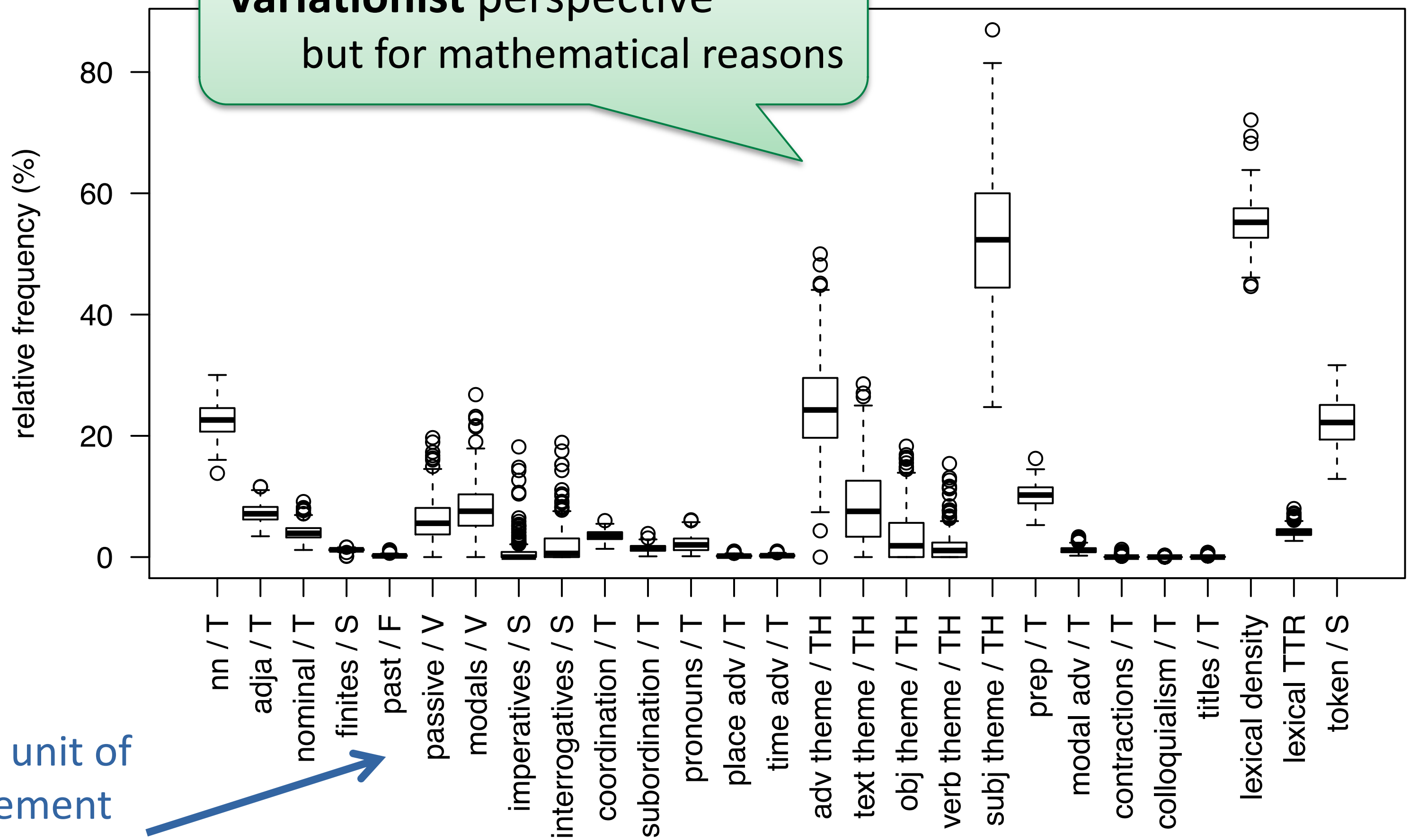
genre: language-external situation + purpose

register: language-internal co-occurrence patterns of linguistic features

- CroCo: parallel corpus English/German
  - English-German and German-English translation pairs
  - we use 298 texts from 5 different genres (excluded: instruction manuals, tourism, fiction)
- 28 lexico-grammatical features (Neumann 2013)
  - comparable between languages
  - inspired by SFL and translation studies
- Text = point in 28-dimensional feature space
- Research hypotheses: shining through (Teich 2003), prestige effect (Toury 2012)

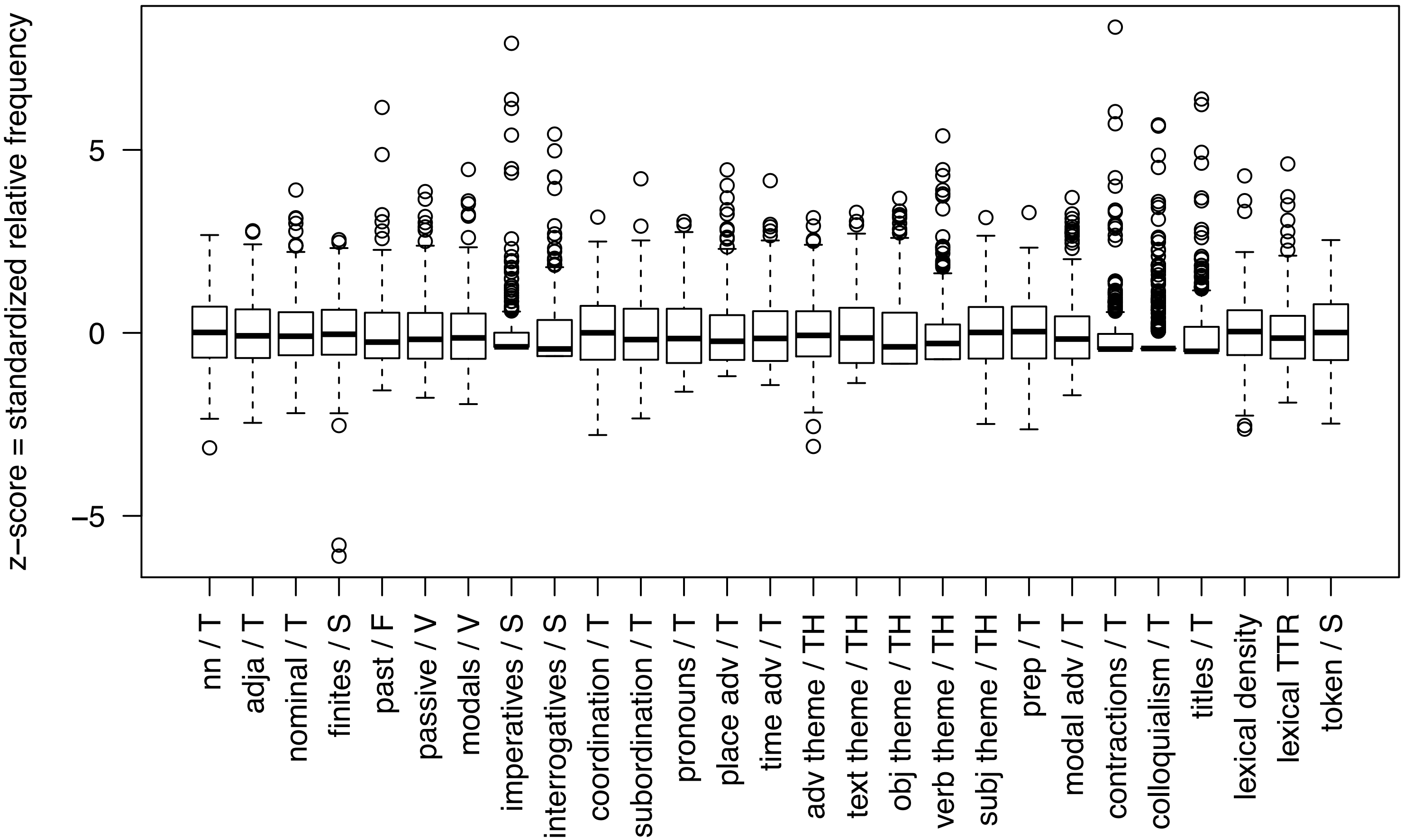
Feature design: avoid "obvious" correlations

variationist perspective
but for mathematical reasons

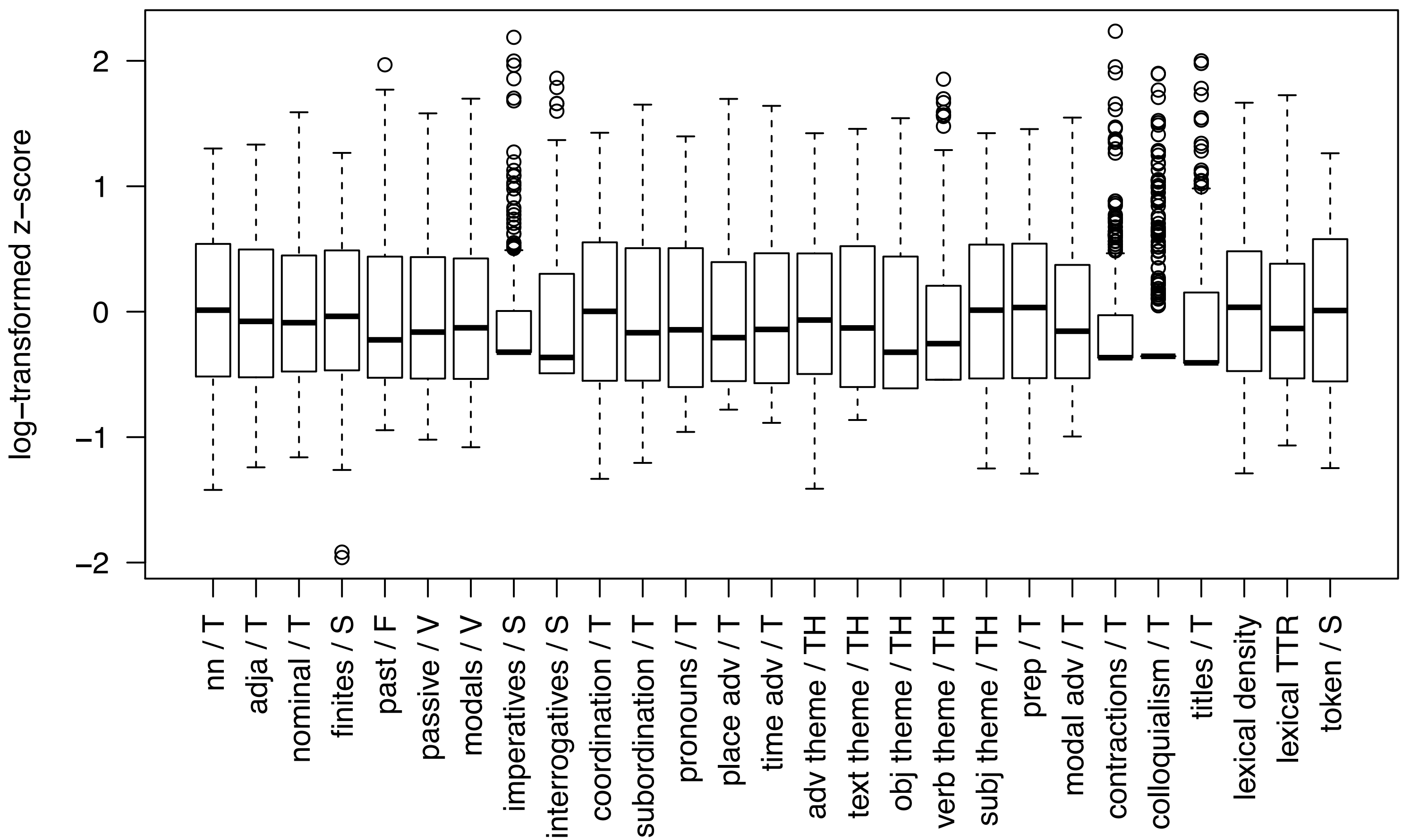


suitable unit of measurement (not always per 1000 words!)

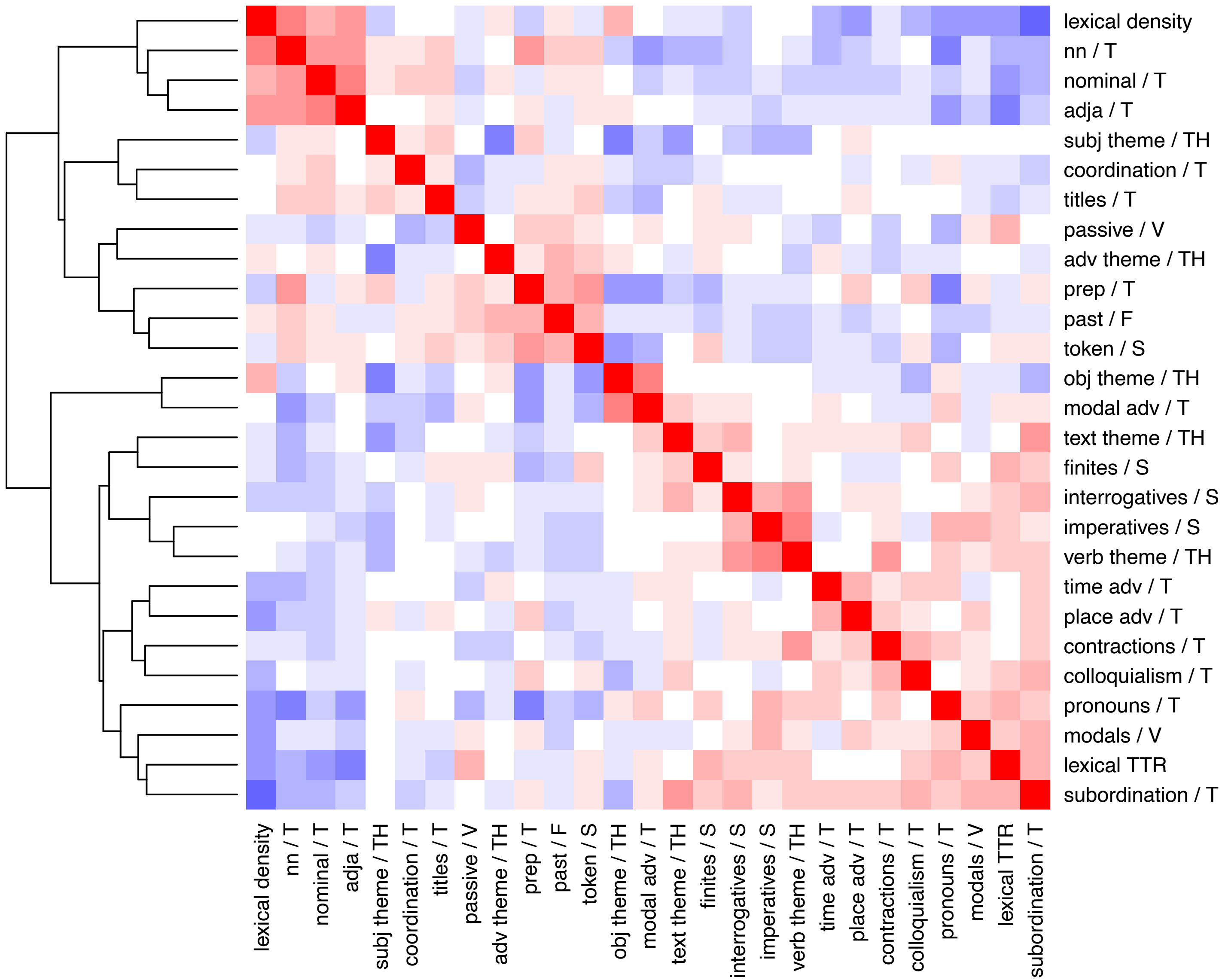
Feature scaling: same contribution to Euclidean distances



Feature scaling: optional signed log transformation



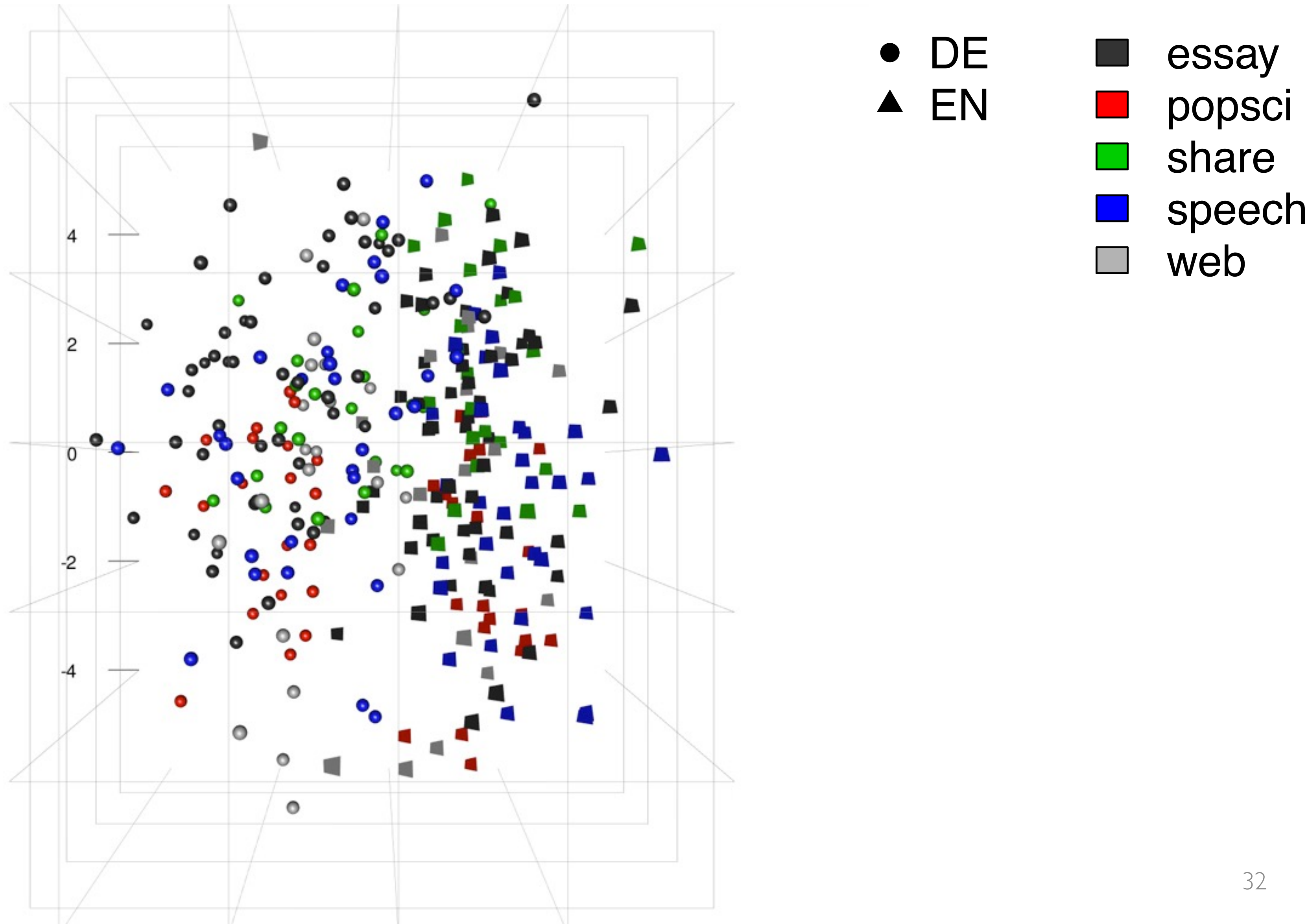
CroCo: correlation matrix



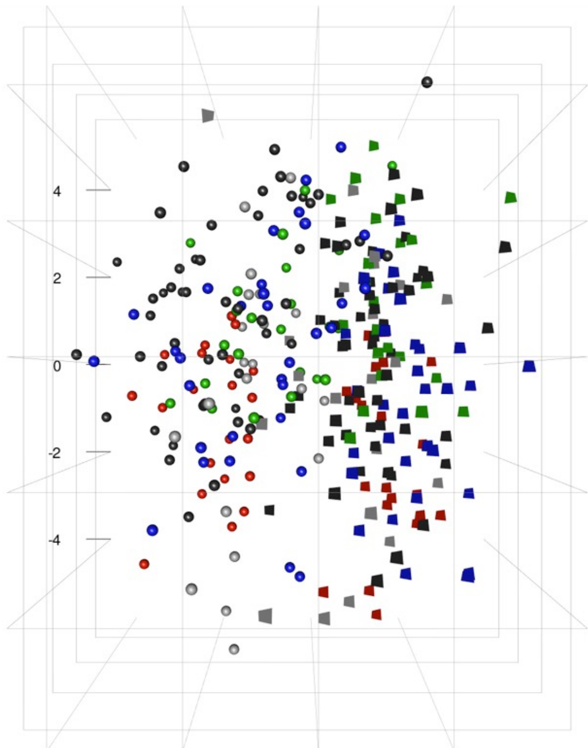
Latent dimensions as perspective on data configuration

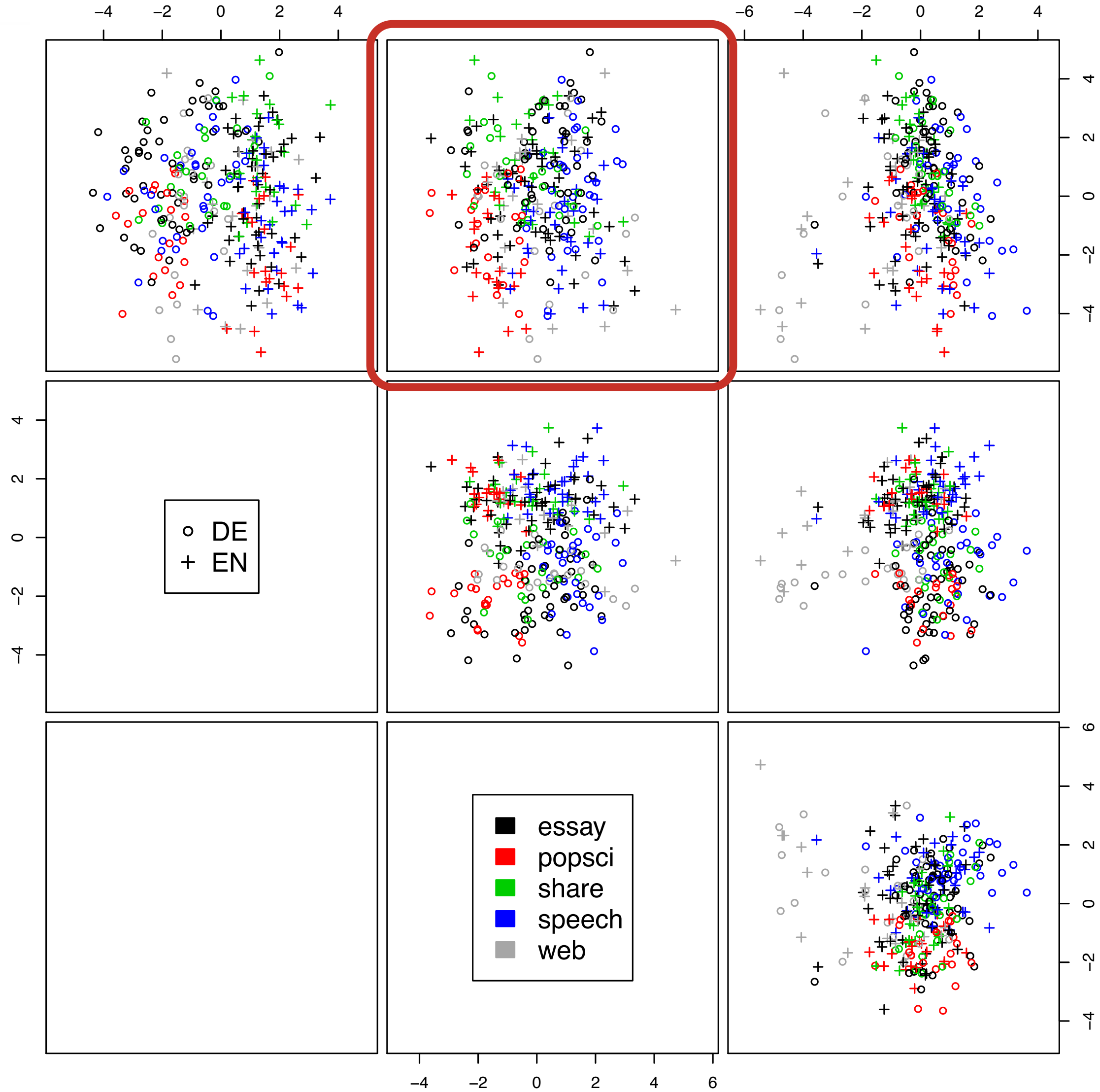
- Instead of "magical" latent dimensions we focus on orthogonal projections as perspectives on the data
  - cf. photograph as 2D perspective on 3D object
- Different perspectives highlight different aspects
- Multivariate analysis ➞ choice of perspective
  - principal component analysis (PCA) = perspective that reflects distances between texts as accurately as possible
  - should reveal major dimensions of variation
  - advantage over factor analysis (FA): dimensionality does not have to be fixed *a priori*

CroCo: 3-dimensional projection



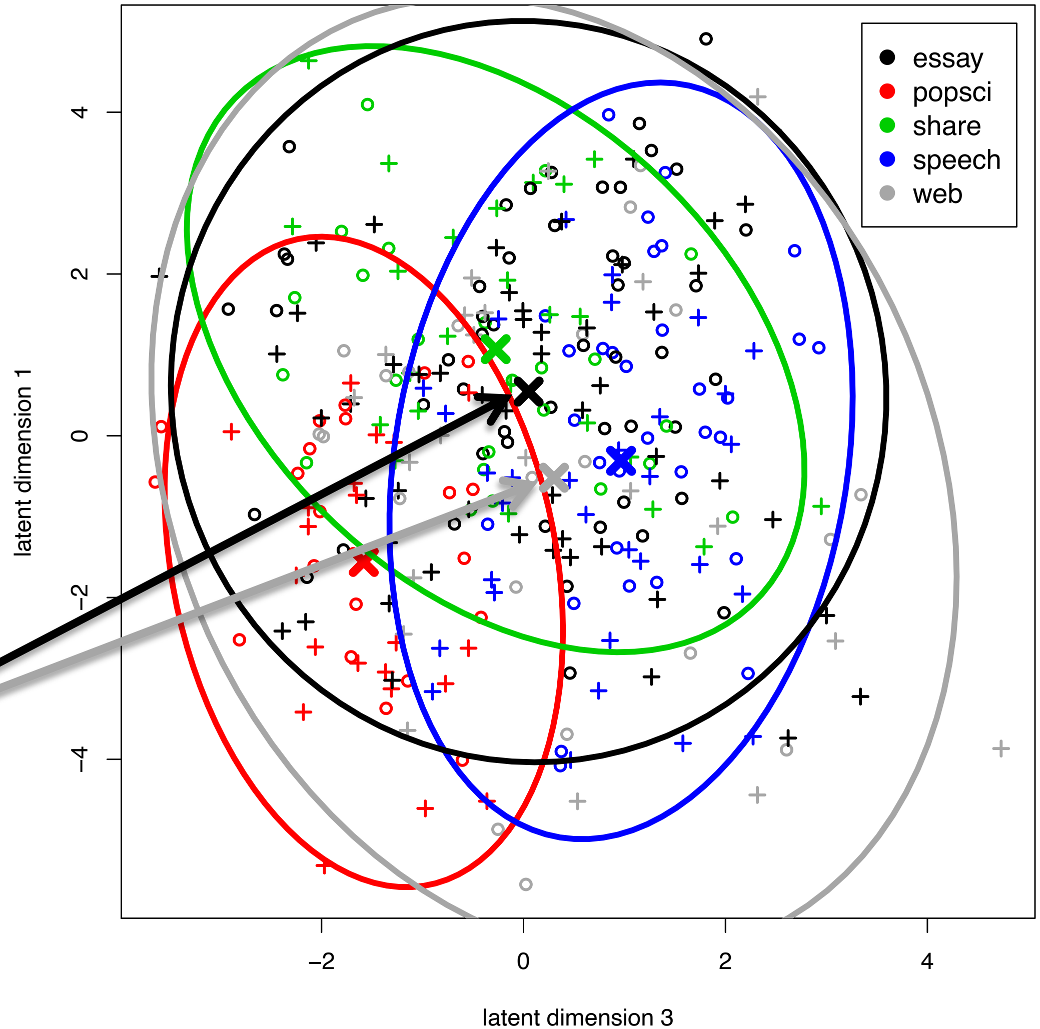
CroCo: 4-dimensional projection





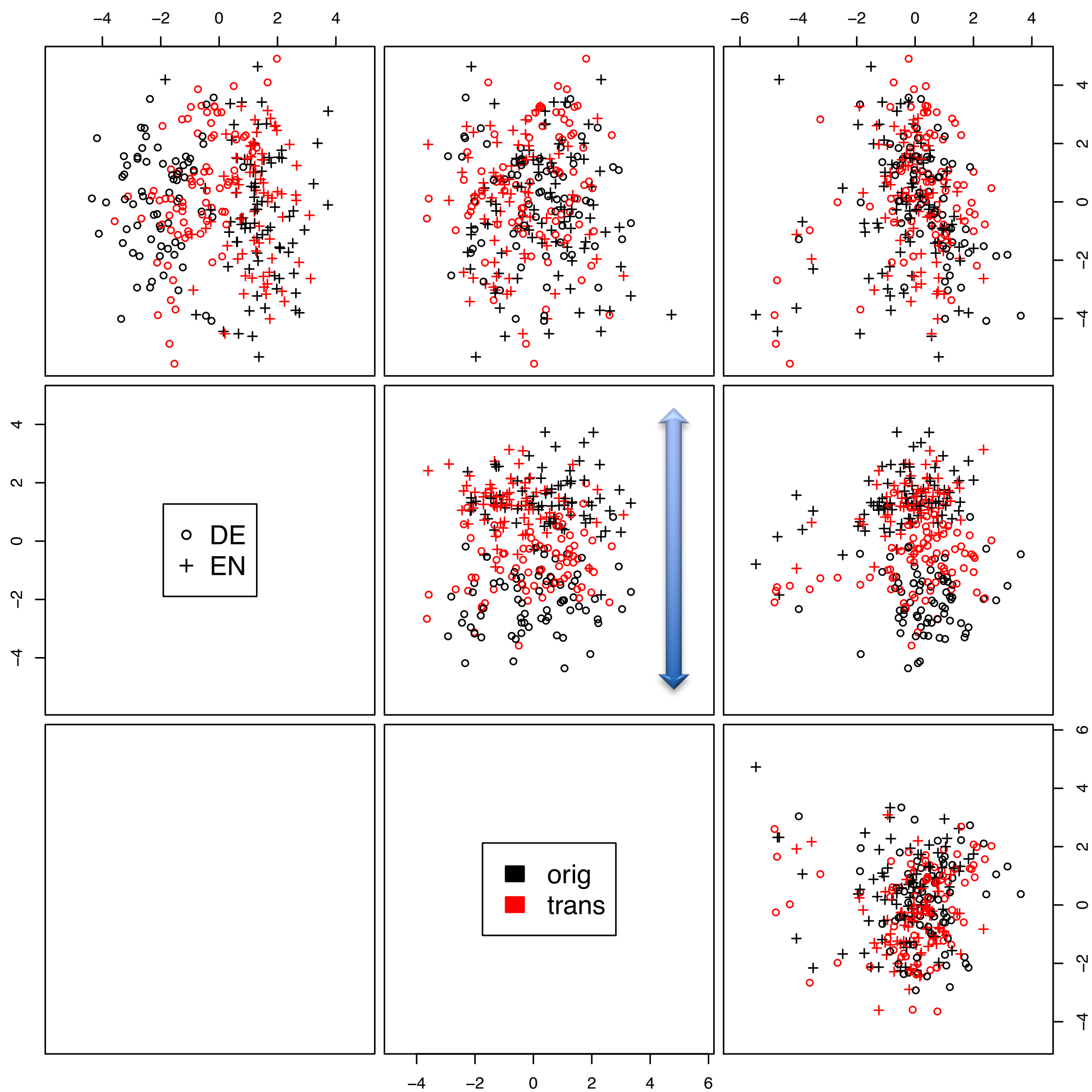
CroCo: genre distribution

- Focus on latent dim's 1 and 3 (register variation)
- Describe genre by centroid + ellipse
- Comparison with Hotelling's t^2 test
  - essays vs. Web
  - t^2=4.21, df=2/141, p=.0167 *



How about translationese?

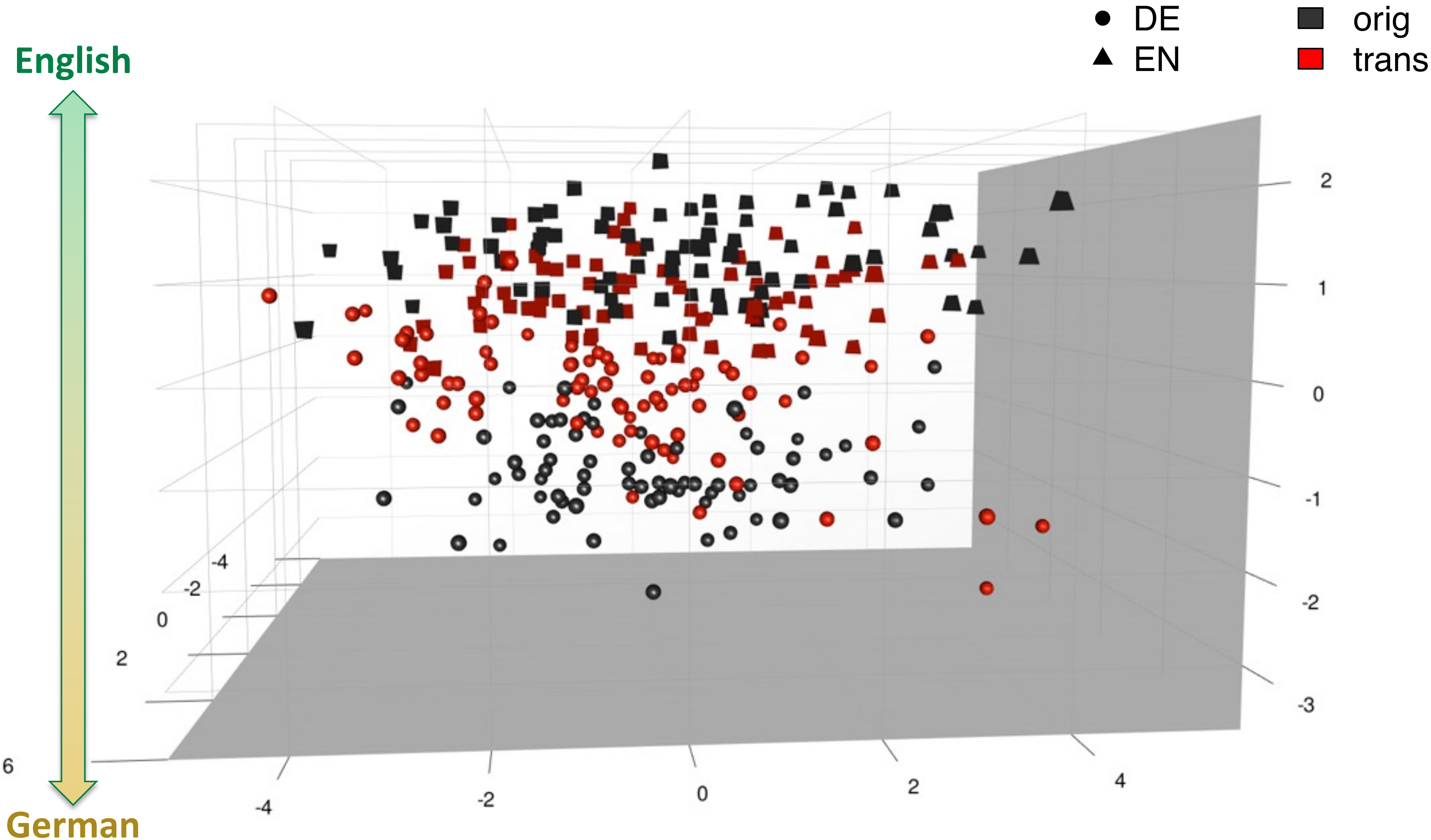
- PCA dim's can't separate translations from original texts
  - 62.1% accuracy on first 3 PCA dim's
- But SVM machine learner can do this with >80% accuracy
  - RBF kernel
  - 10-fold c.v.
- Hints at shining through, but no clear-cut evidence



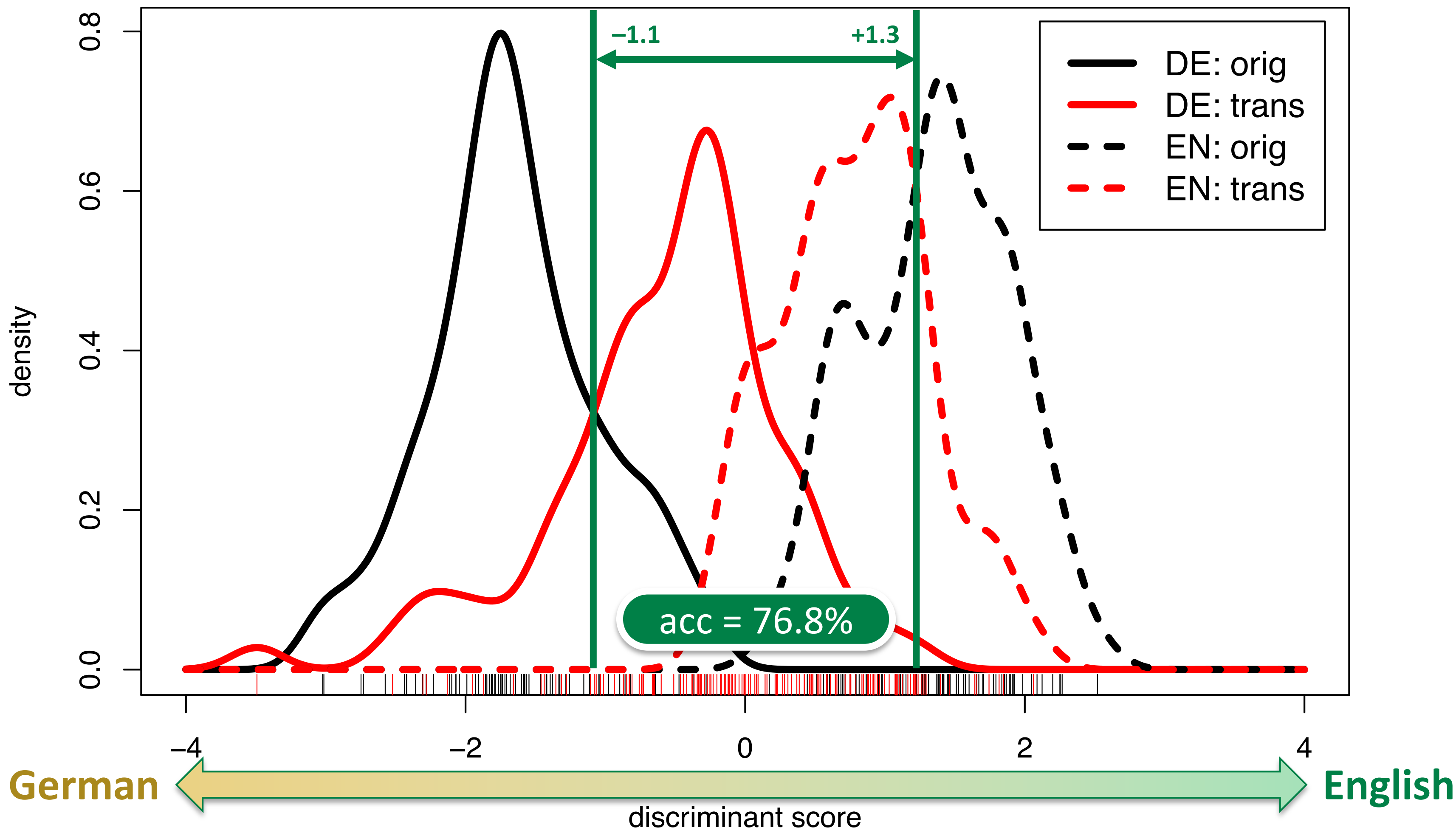
Minimally supervised LDA

- Add minimal amount of supervised knowledge to find a more informative perspective
  - evidence for shining through hypothesis from dimension that corresponds to contrast German vs. English
  - supervised knowledge: language of original texts only
- Linear discriminant analysis (LDA)
  - maximize separation between German / English originals
  - minimize variability within each group
  - classical technique related to PCA and ANOVA
- Project *all* texts onto LDA discriminant
  - complemented by additional PCA dim's for visualization

CroCo: LDA perspective

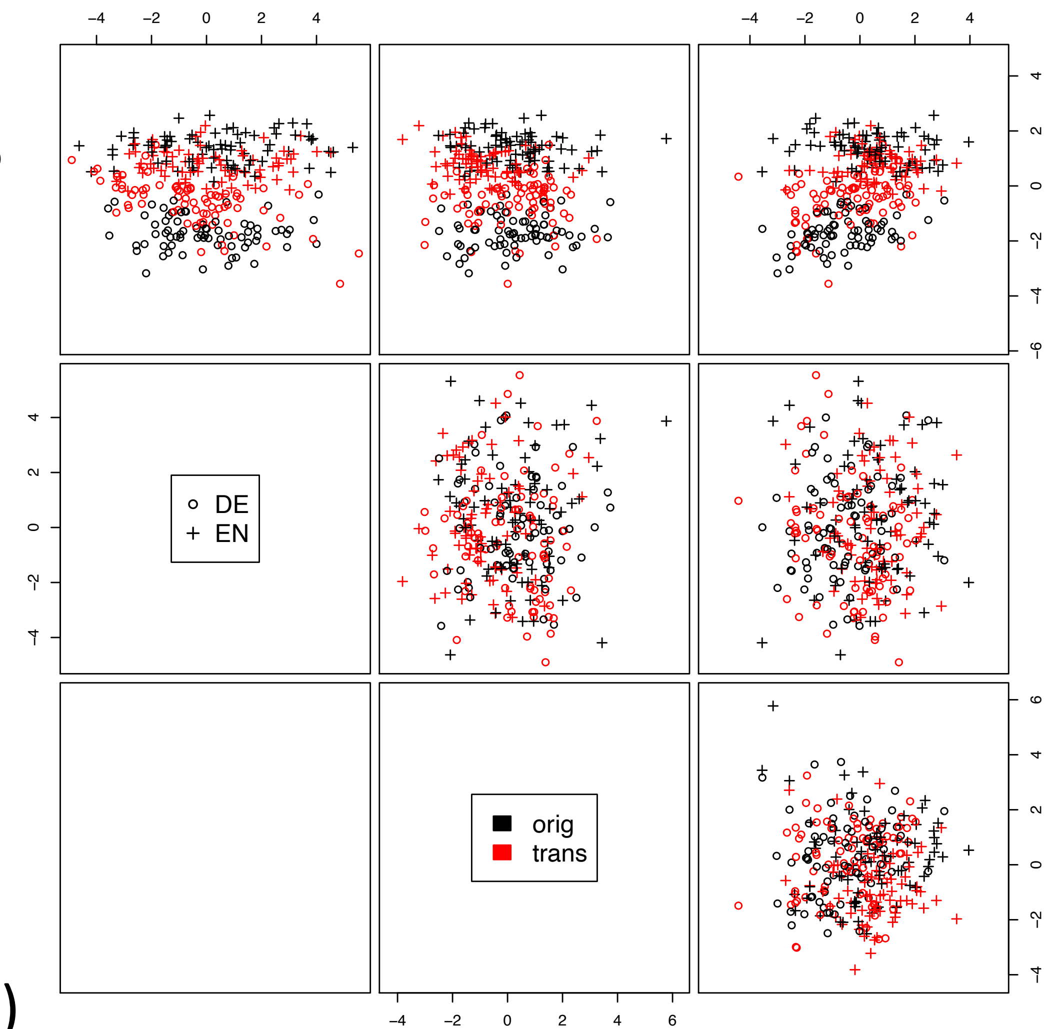


Discriminant for DE vs. EN confirms shining through & prestige effect



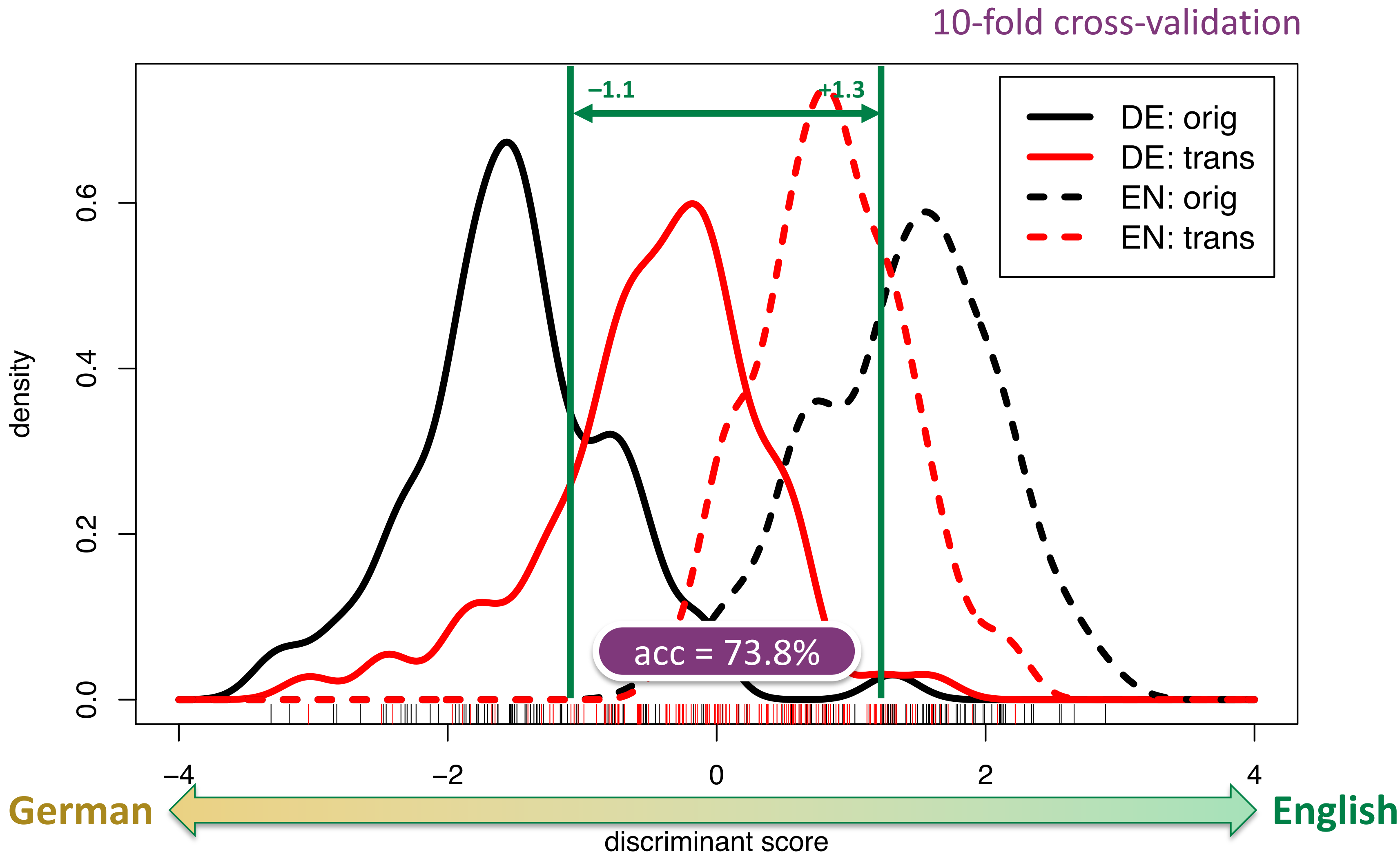
LDA significance: bootstrapping / cross-validation

- LDA is a supervised ML technique ➞ overtrained?
  - Would we find the same discriminant if we trained on a different set of texts?
- Verification with bootstrap resampling or 10-fold cross-validation
  - LDA trained on 90% of data
  - discriminant axis shows "wobble" of approx. 10°
- Discriminant scores from c.v. (10% test data per fold)

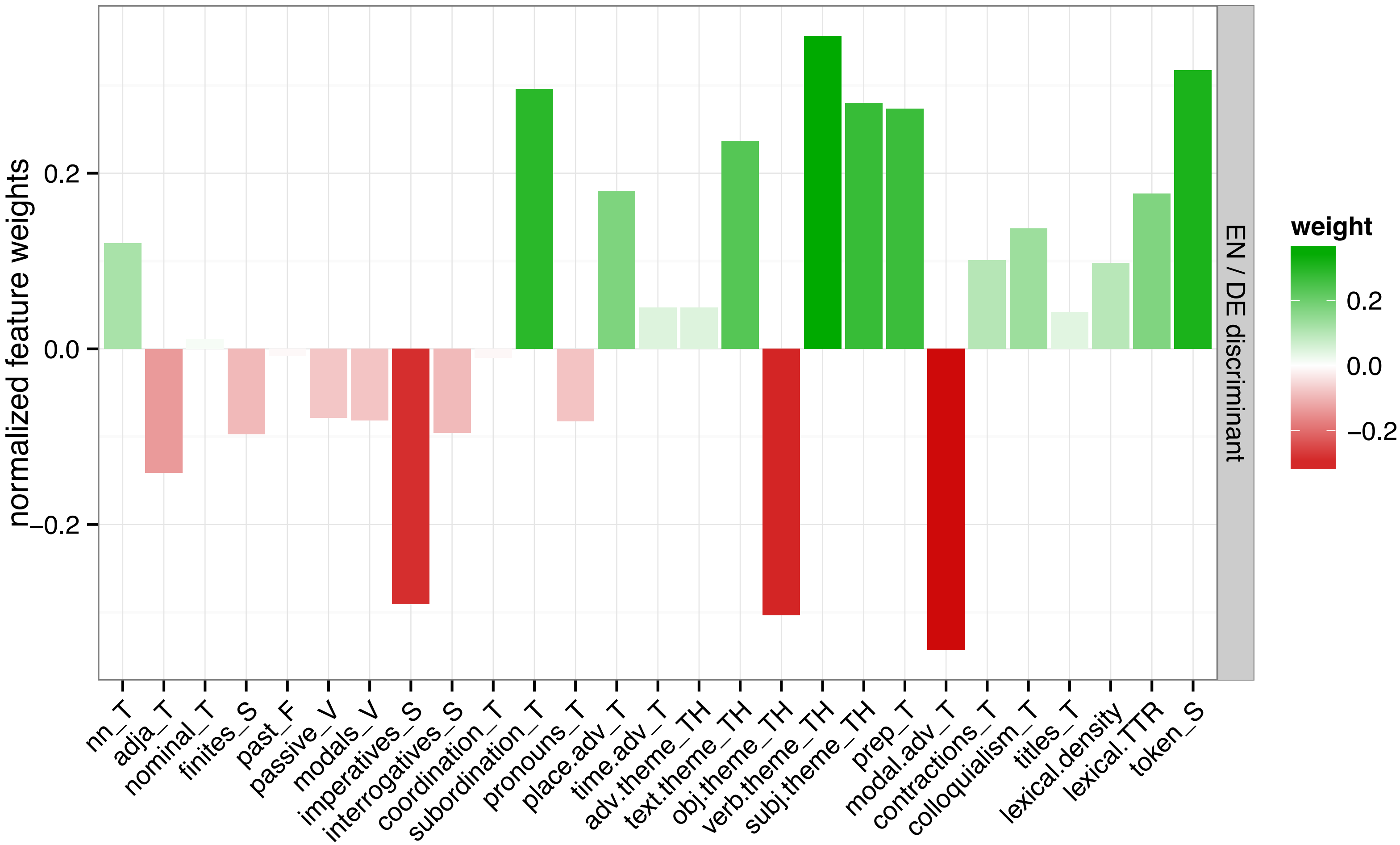


CV fold #5

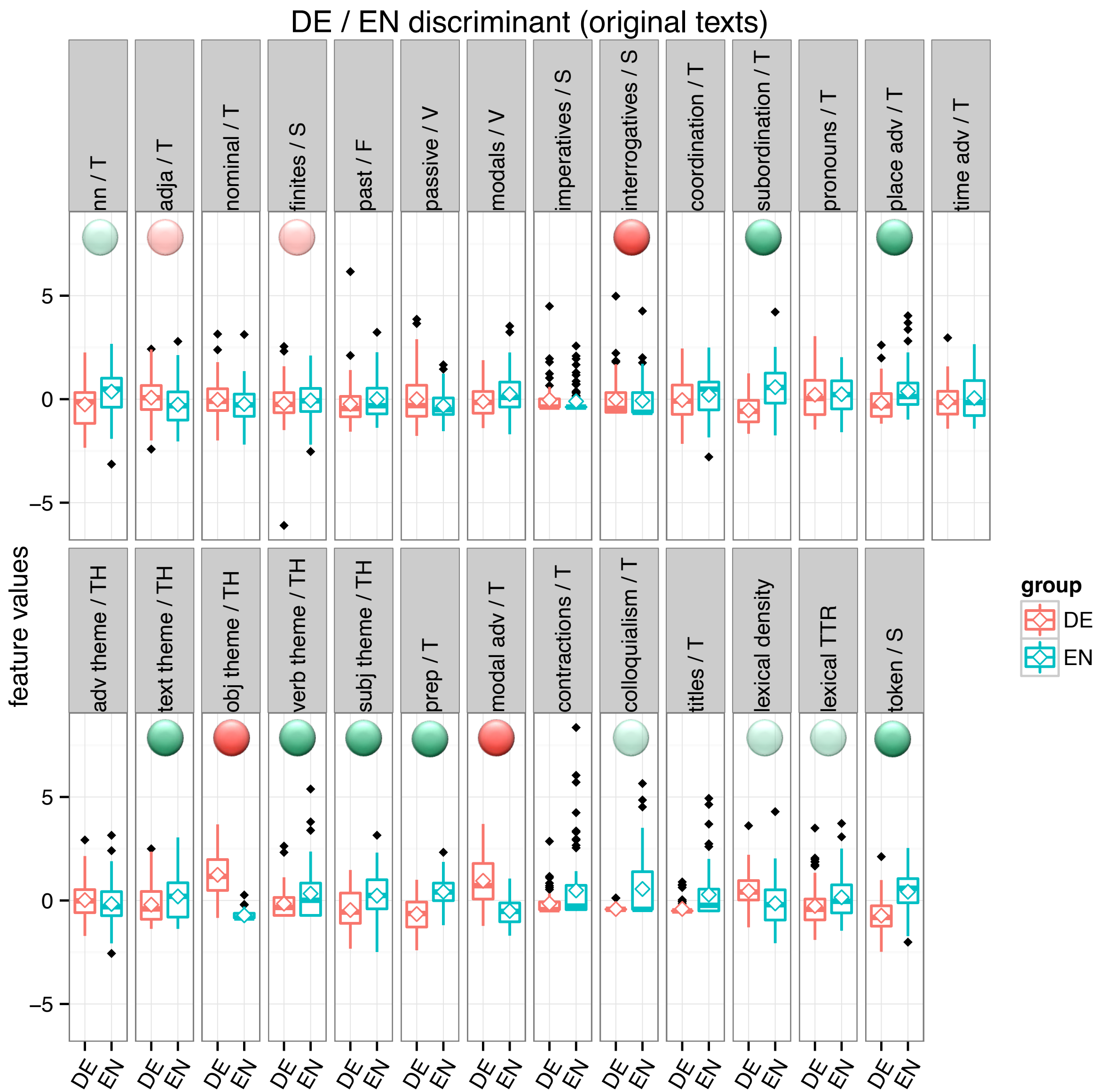
Cross-validated discriminant



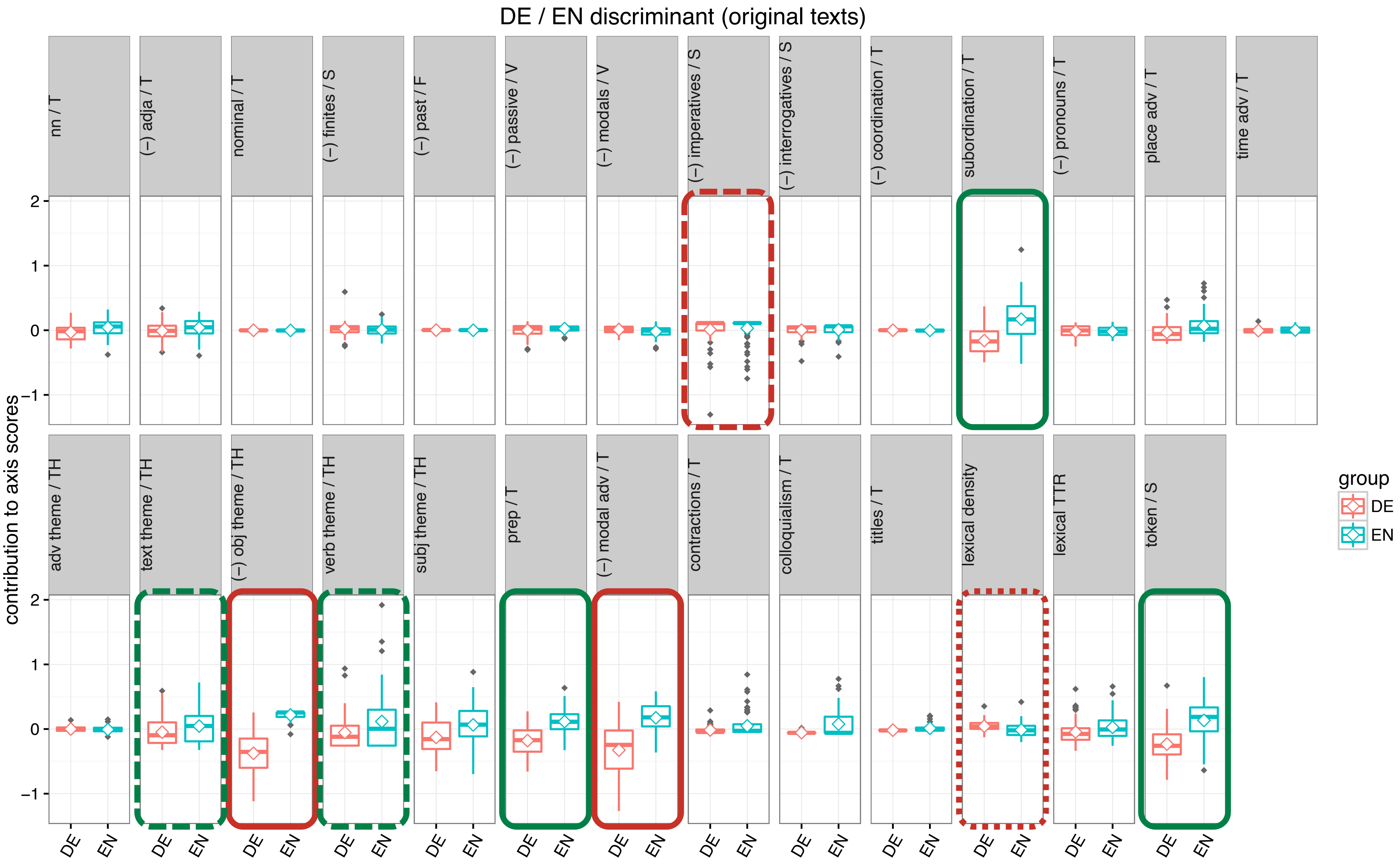
Interpreting discriminant features



Interpreting discriminant features



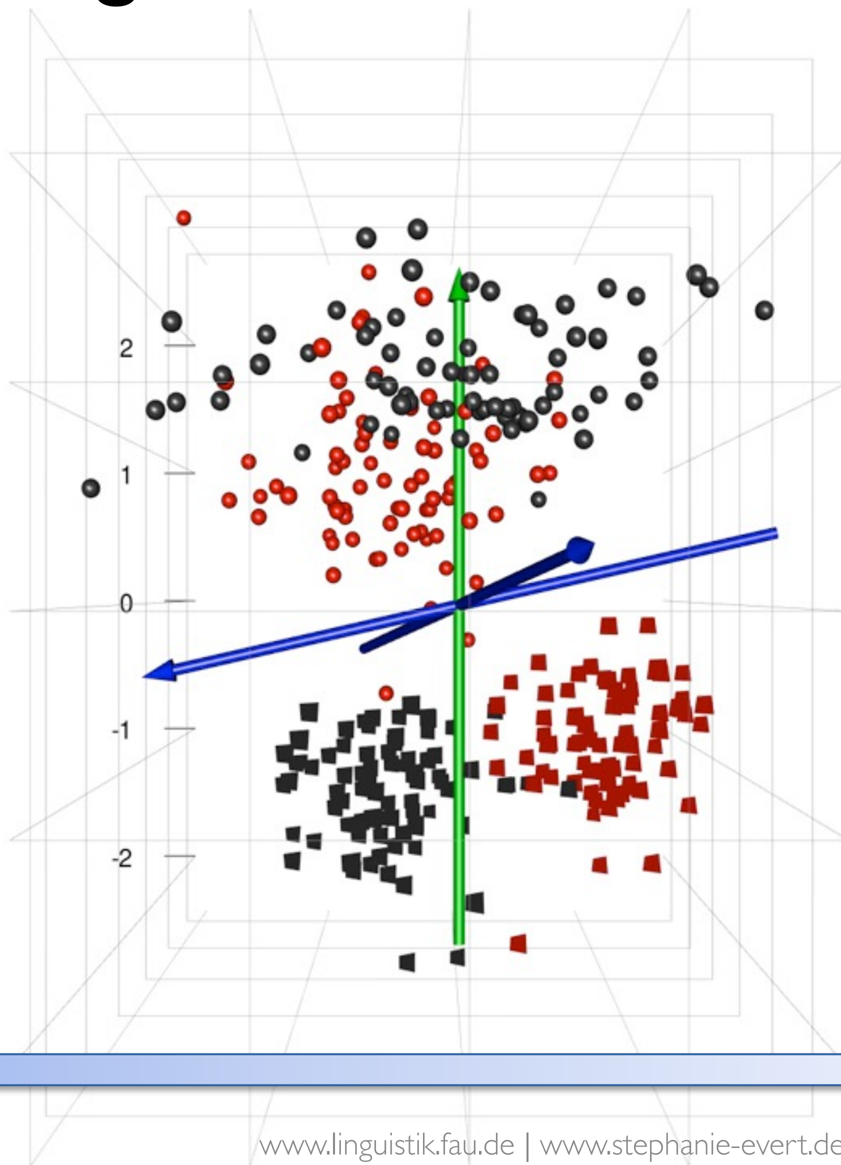
Interpreting discriminant features



Unravelling translationese

German

English



● DE
▲ EN
orig
trans

LDA for trans vs. orig in each language

Case study 2: French regional varieties

(Diwersy, Evert & Neumann 2014)

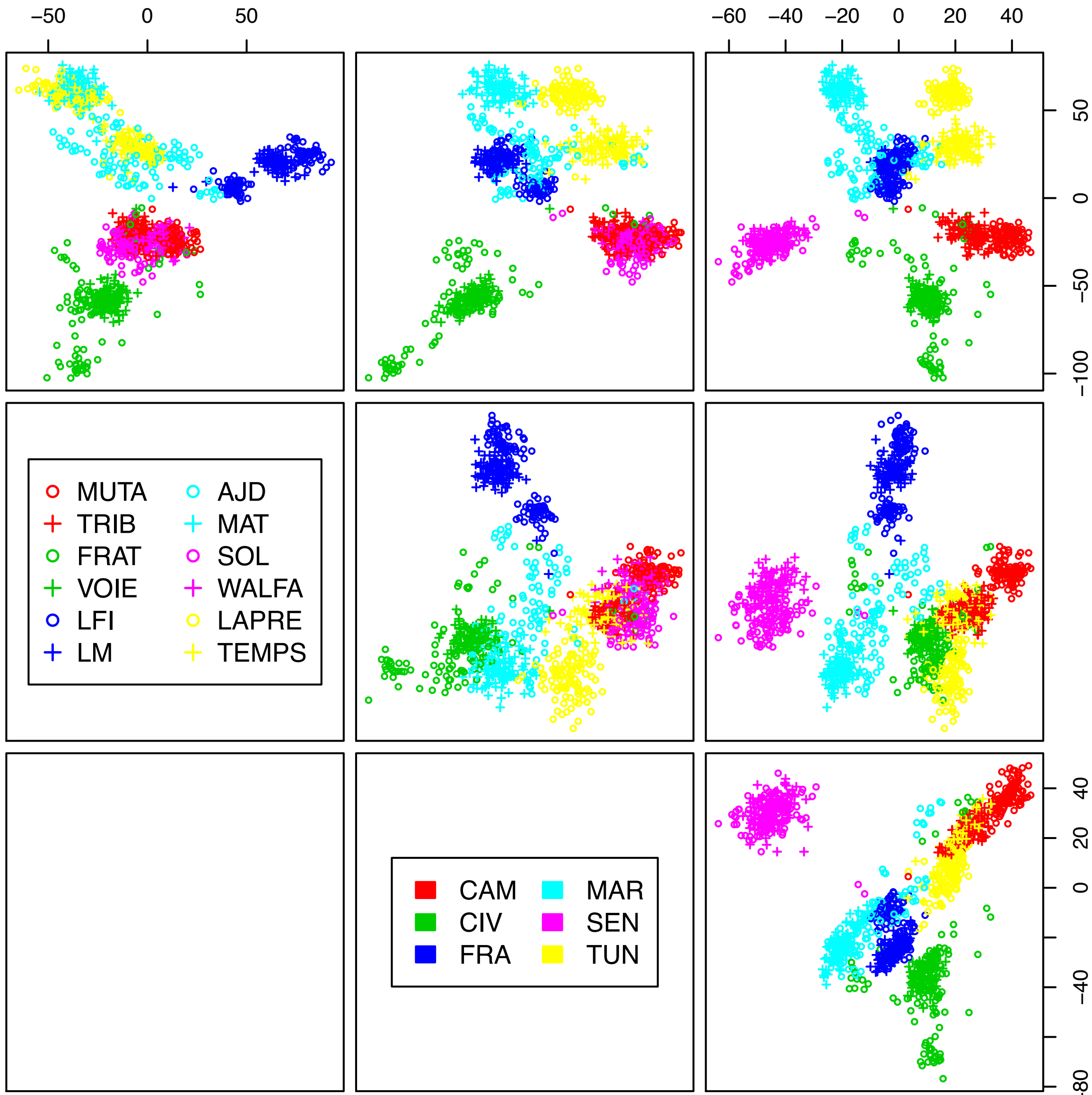
- Lexical differences in regional varieties of French
- Two nation-wide newspapers each from 6 countries
  – Cameroon, France, Ivory Coast, Morocco, Senegal, Tunisia
  – two consecutive volumes from each newspaper
  – total size approx. 14.5 million tokens
- Text samples = one week each
- Features: frequencies of shared colligations
  – colligation = lemma-function pairs
  – must occur in all subcorpora with $f \geq 100$

FRV: poor choice of features

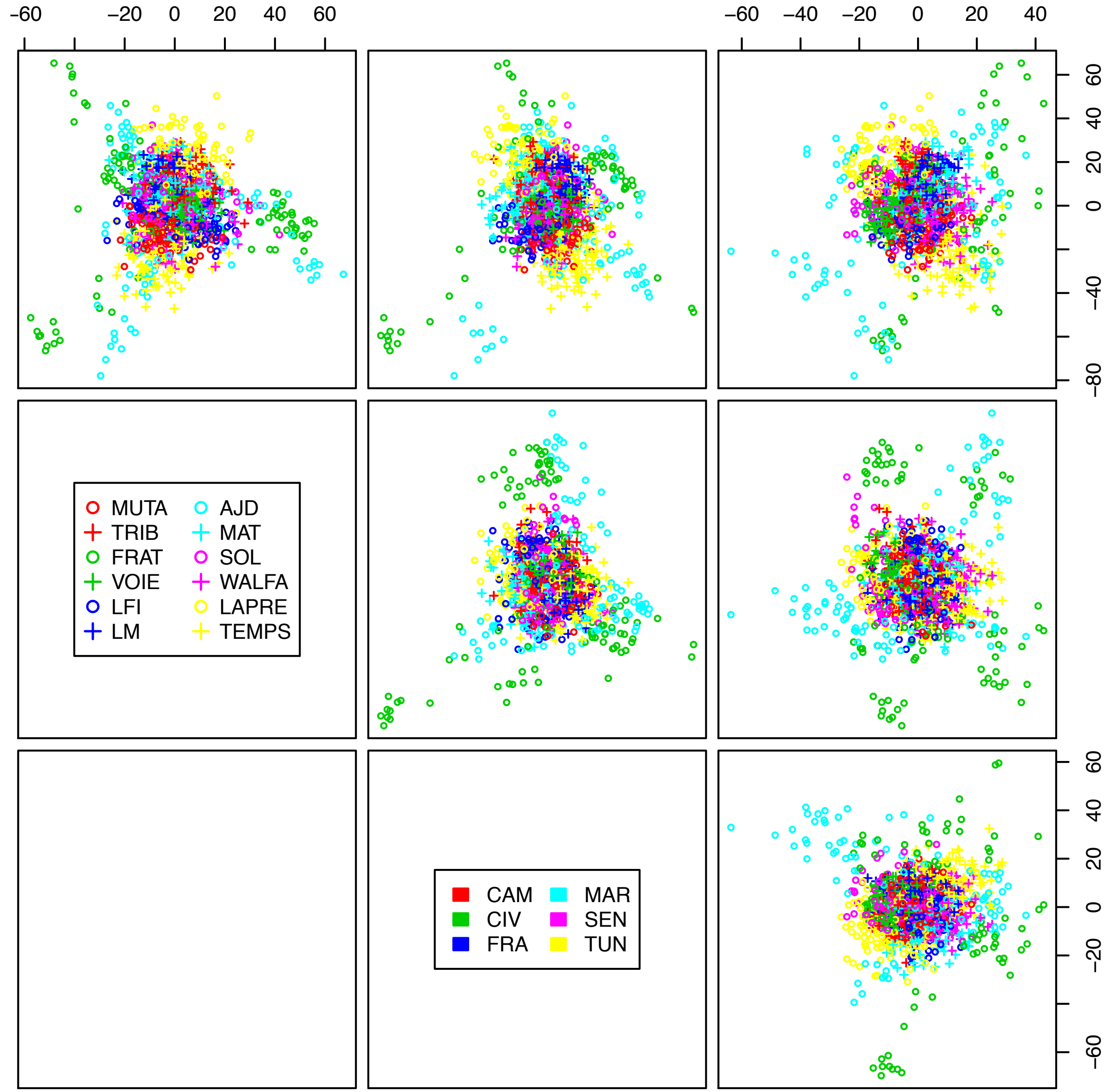
PCA not excluding country-specific words as features: perfect separation

Design bias results in a completely uninteresting model

FA not applicable: features >> texts



FRV: PCA dimensions

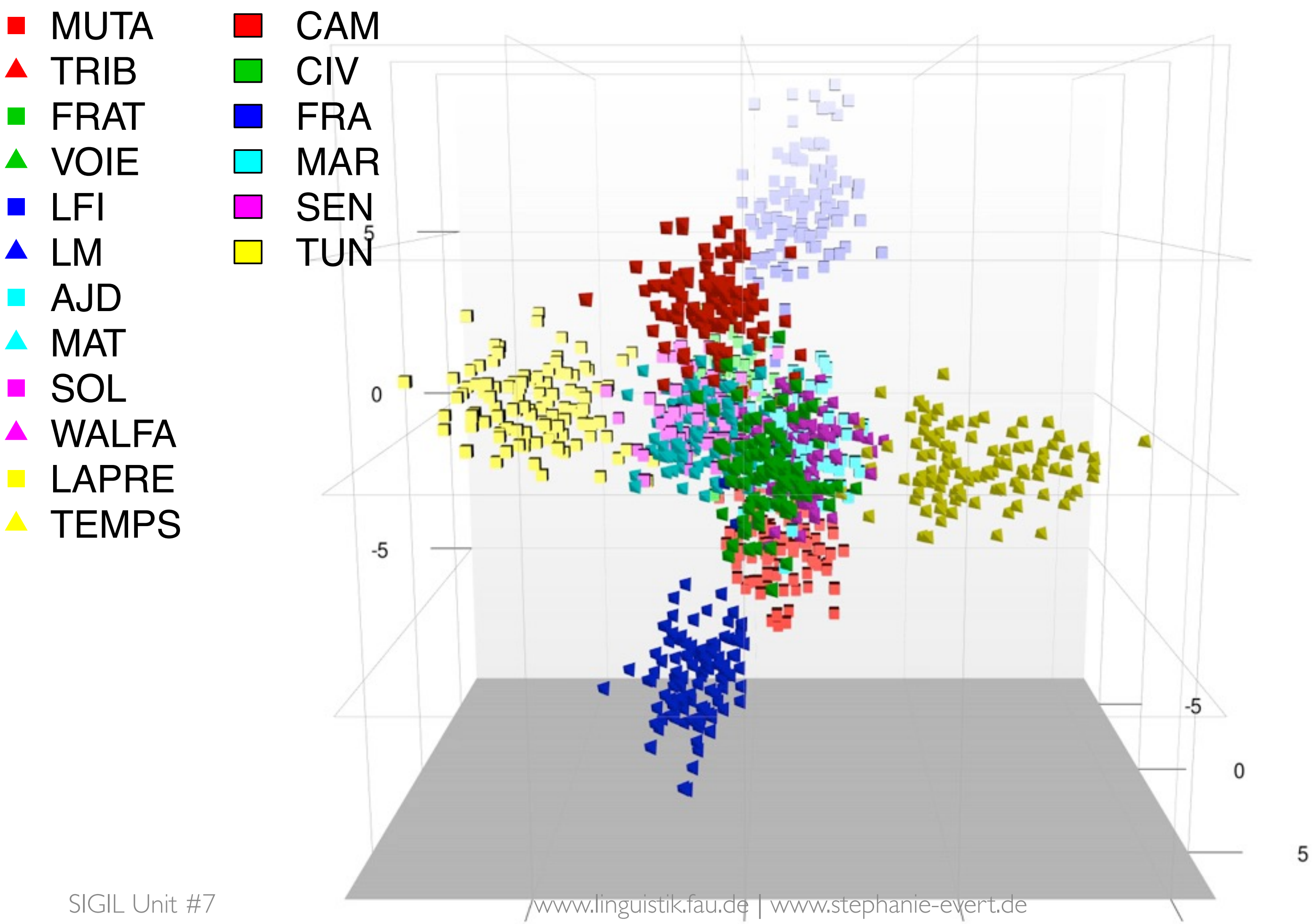


Using only shared words as features, PCA no longer reveals any patterns (just a few outliers)

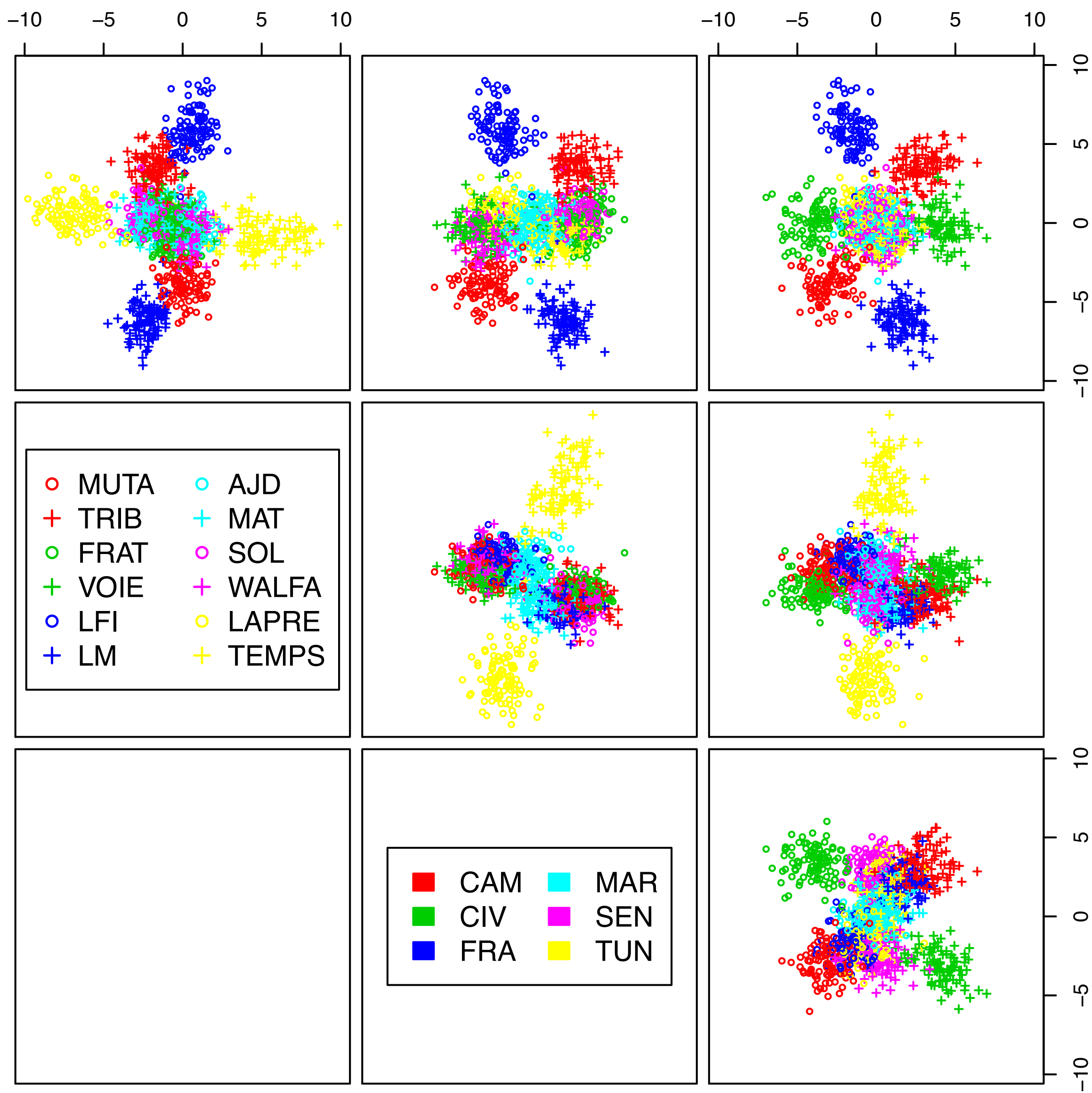
Use LDA to find a meaningful perspective, based on newspaper source

Country would presume regional varieties exist!

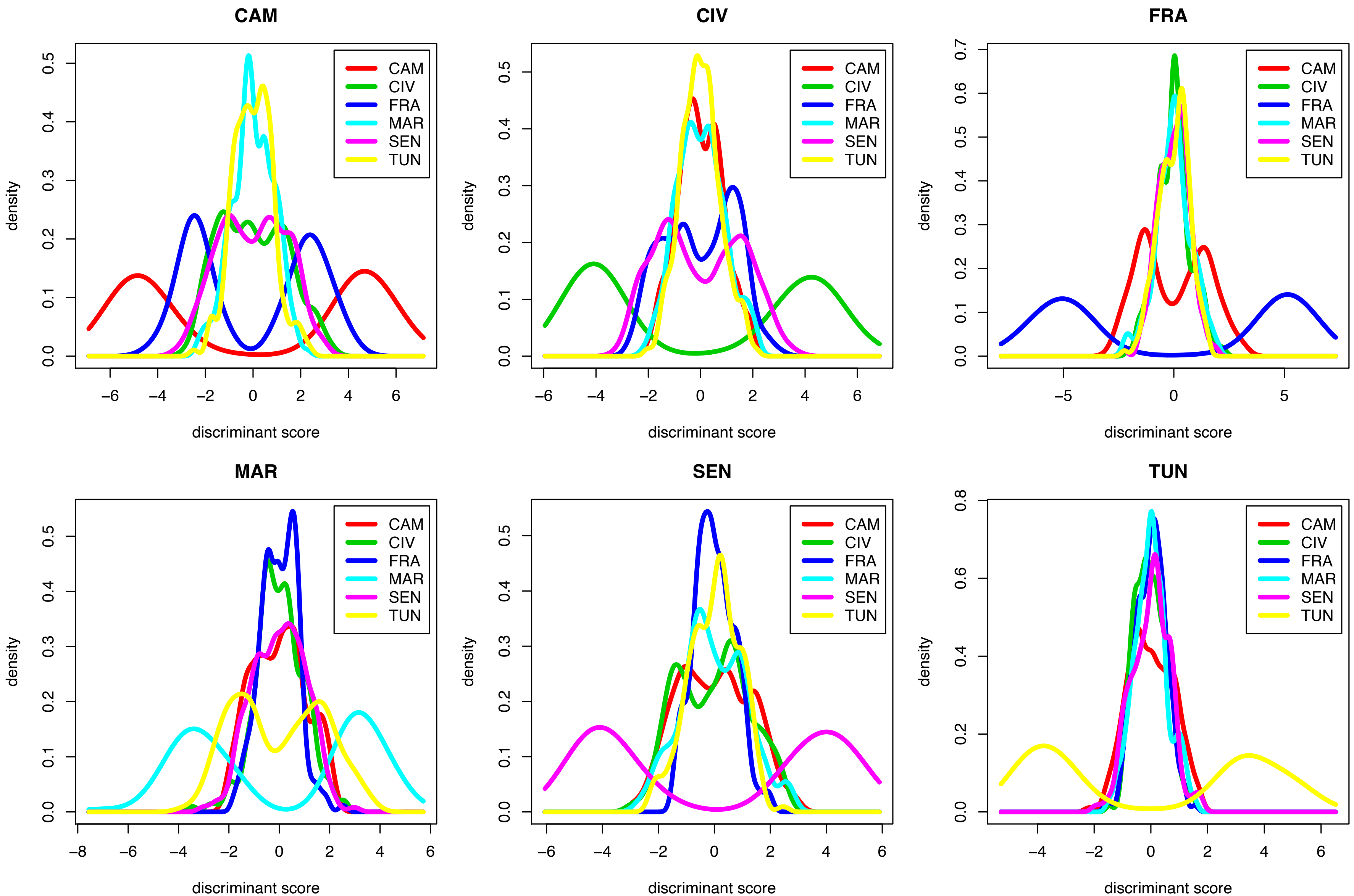
FRV: LDA dimensions (newspapers)



FRV: LDA dimensions (newspapers)



FRV: discriminant axes



References

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

Diwersy, S.; Evert, S.; Neumann, s. (2014). A weakly supervised multivariate approach to the study of language variation. In B. Szmrecsanyi & B. Wälchli (eds.), *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*. De Gruyter, Berlin.

Evert, S. & Neumann, S. (2017). The impact of translation direction on the characteristics of translated texts: a multivariate analysis for English and German. In G. De Sutter, M.-A. Lefer & I. Delaere (eds.), *Empirical Translation Studies. New Theoretical and Methodological Traditions* (TiLSM 300), pages 47–80. Mouton de Gruyter, Berlin.

Gasthaus, J. (2007). Prototype-Based Relevance Learning for Genre Classification. B.Sc. thesis, Universität Osnabrück, Institute of Cognitive Science.

Koppel, M.; Argamon, S.; Shimoni, A. R. (2003). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, **17**(4), 401–412.

Neumann, S. (2013). *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. de Gruyter Mouton, Berlin.

Neumann, S. & Evert, S. (2021). A register variation perspective on varieties of English. In E. Seoane & D. Biber (eds.), *Corpus based approaches to register variation*. Benjamins, Amsterdam.

Teich, E. (2003). *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts*. Berlin: Mouton de Gruyter.

Toury, G. (2012). *Descriptive Translation Studies – and beyond: Revised edition.* 2nd ed. Amsterdam: John Benjamins.