FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

A multivariate approach to linguistic variation and distribution

Stefan Evert

Corpus Linguistics Group
FAU Erlangen-Nürnberg

Linguistic variation

Variation of a quantitative linguistic feature

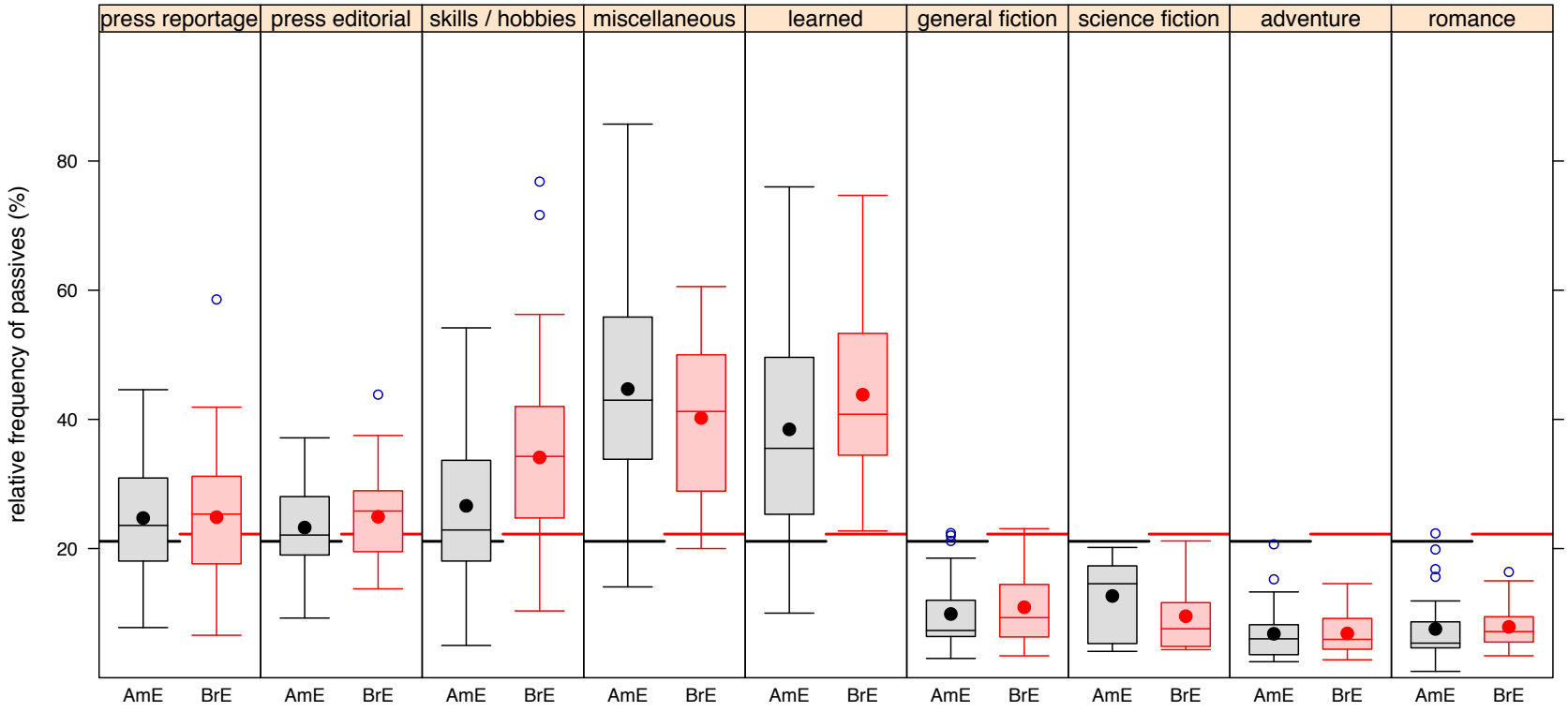
- frequency of passive, past perfect, split infinitive, …
- frequency of expression, semantic field, topic, …
- etc.

across

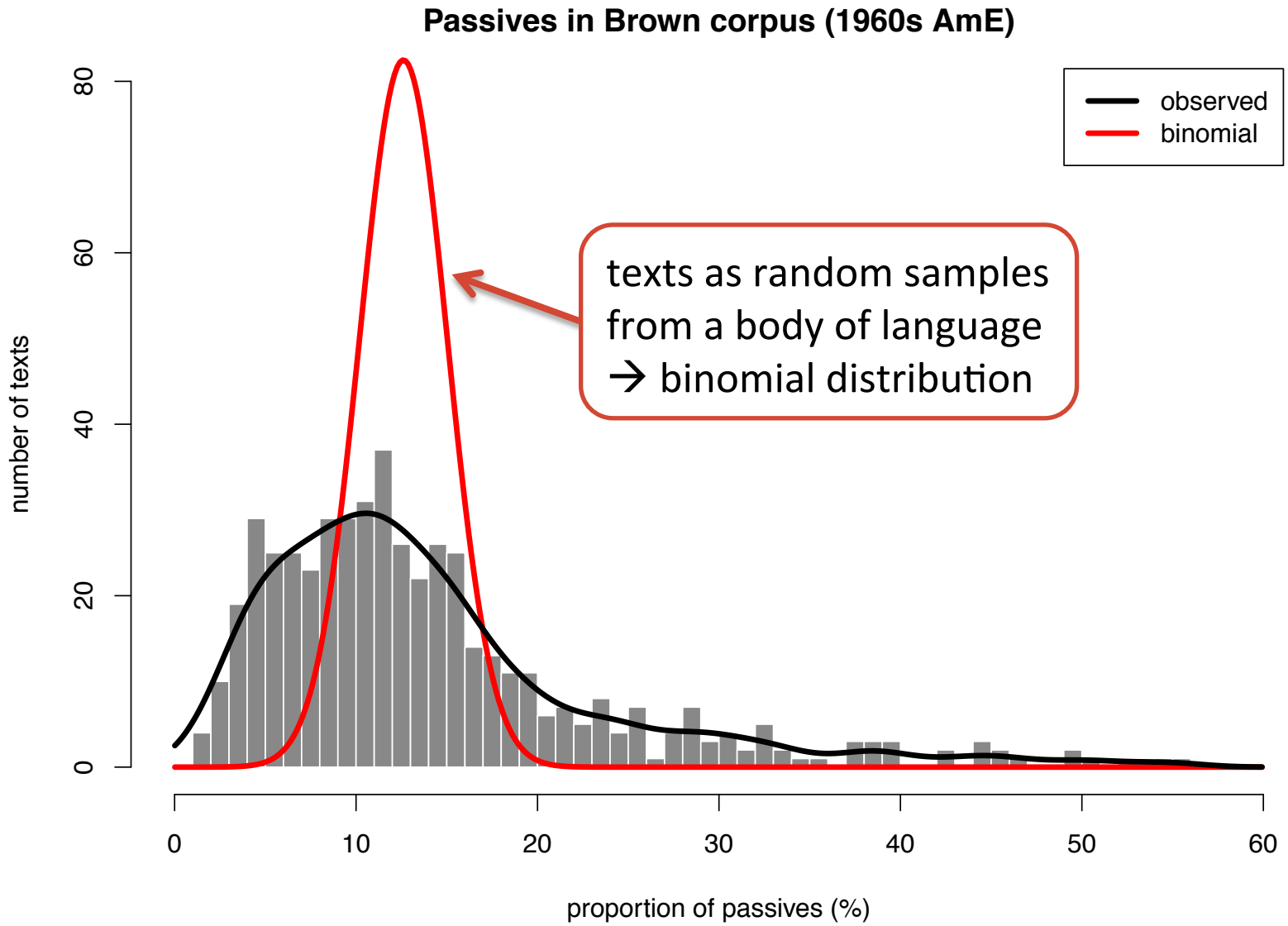
- languages and language varieties
- regions
- social strata
- time
- individual speakers
- etc.

The traditional approach

- Select a linguistic feature (e.g. passive voice)
- Compare its frequency across different categories (genres, language varieties, speakers, …)

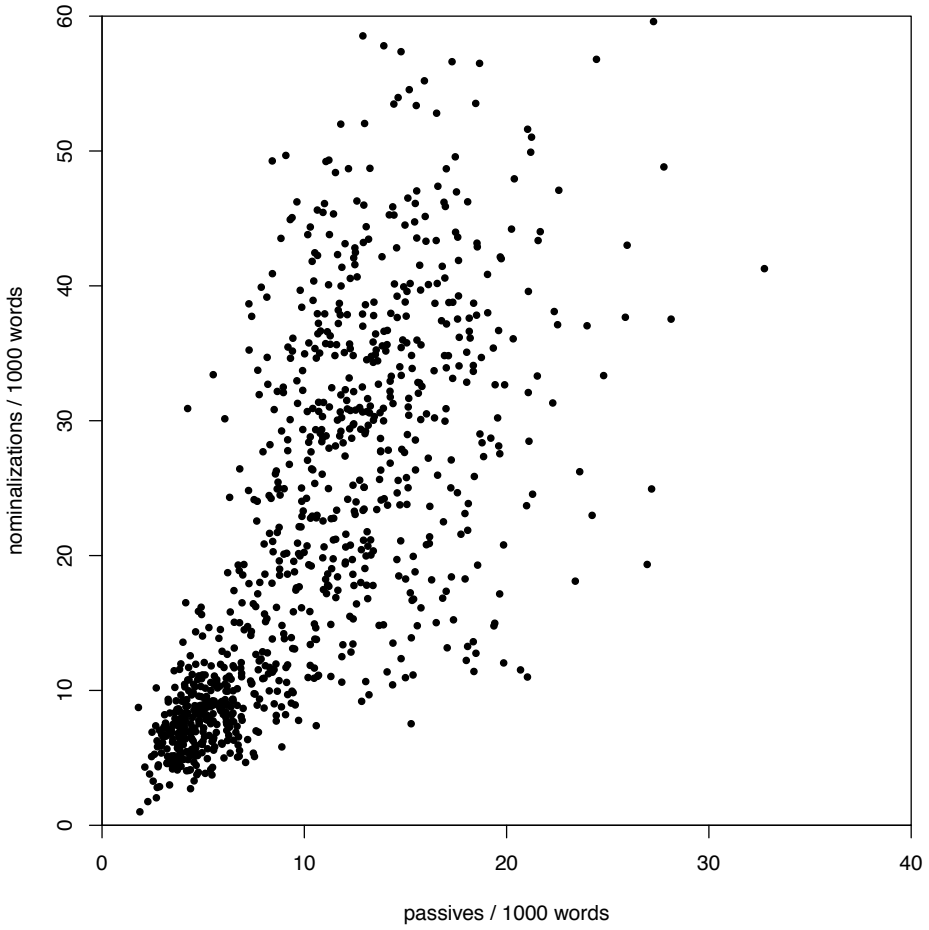


Language variation as a nuisance parameter in corpus linguistics



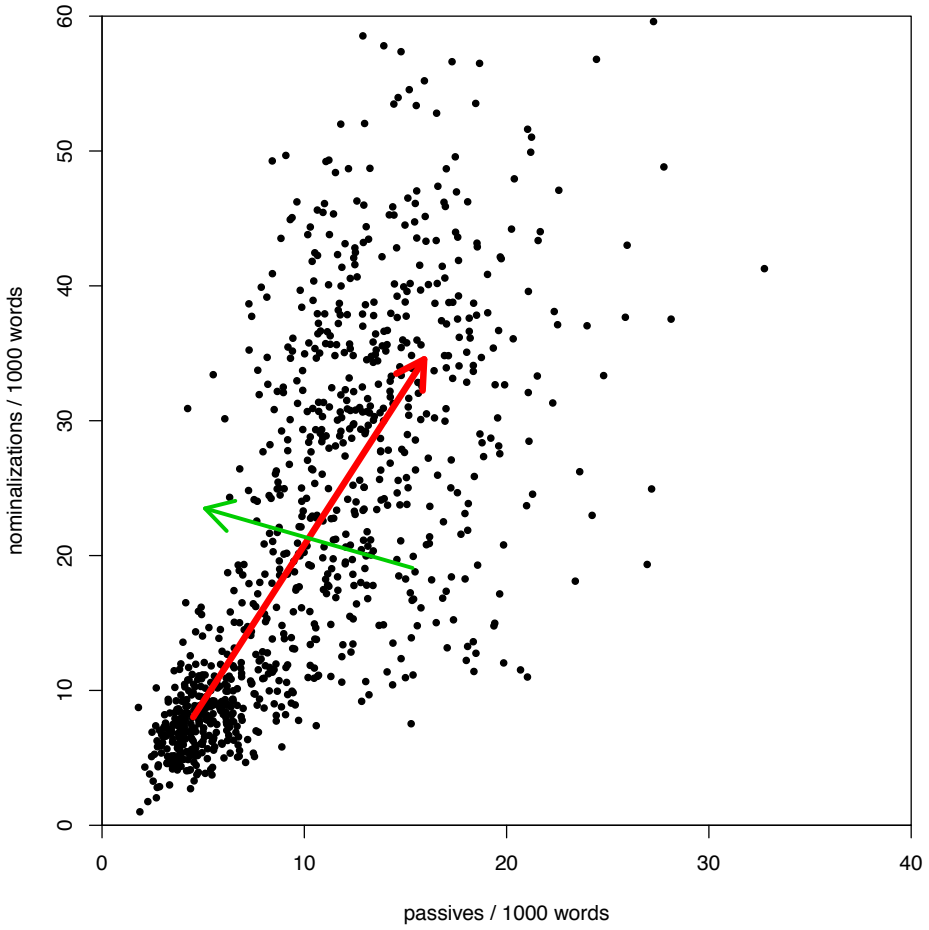
The multivariate approach

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

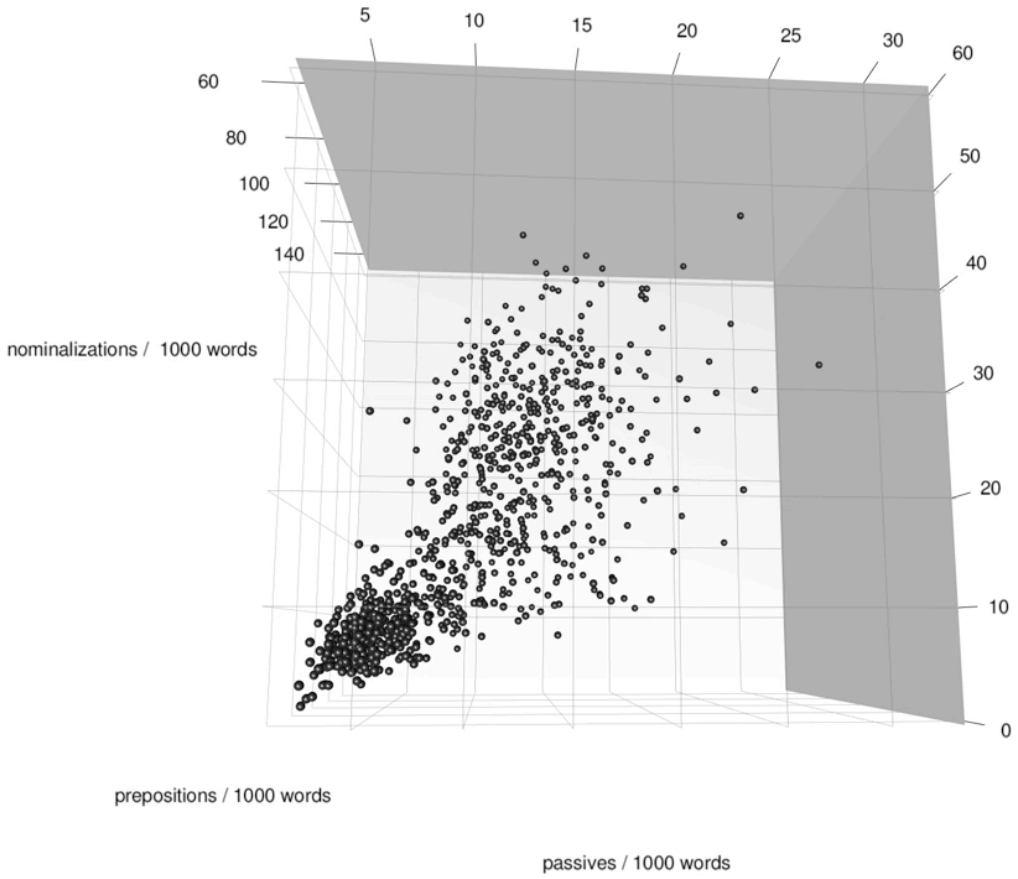


The multivariate approach

- 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

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*)

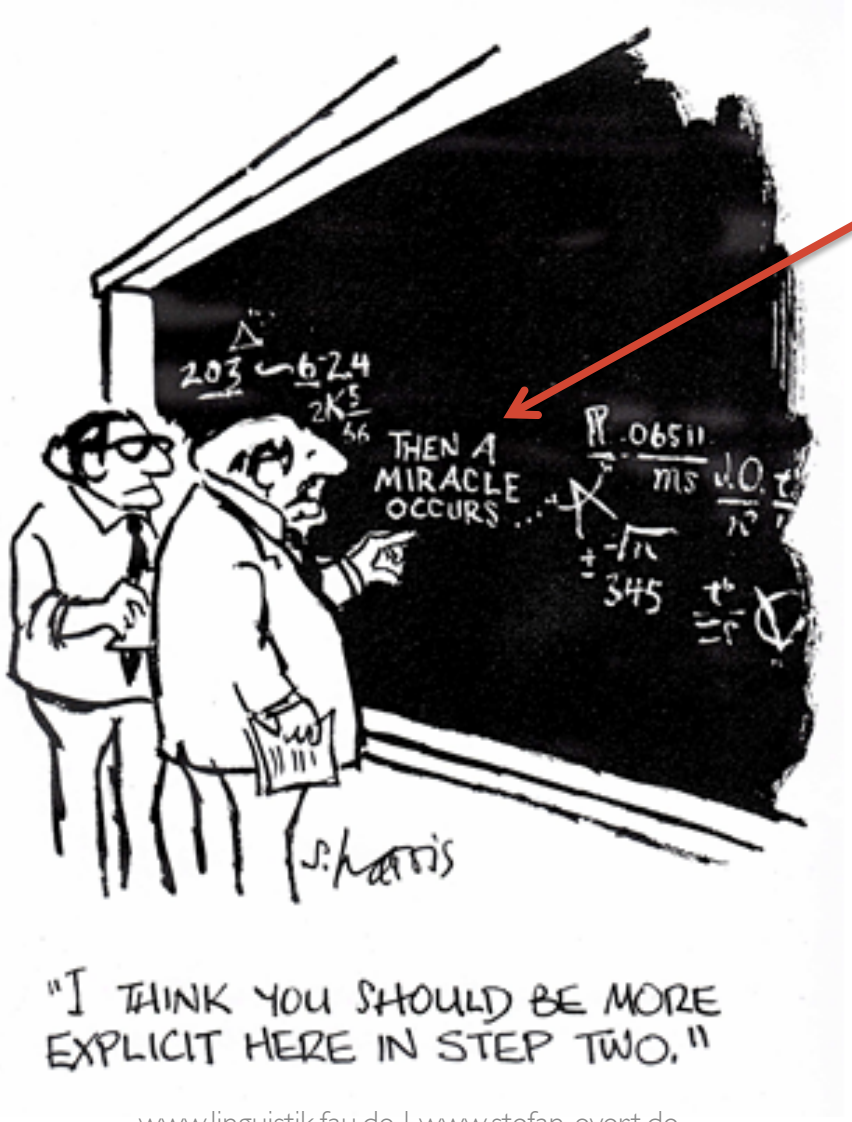
M. Specialized verb classes
- 55 Public verbs (e.g., *assert, declare, mention*)
- 56 Private verbs (e.g., *assume, believe, doubt, know*)
- 57 Suasive verbs (e.g., *command, insist, propose*)
- 58 *seem* and *appear*

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 multidimensional analysis



factor analysis (FA)

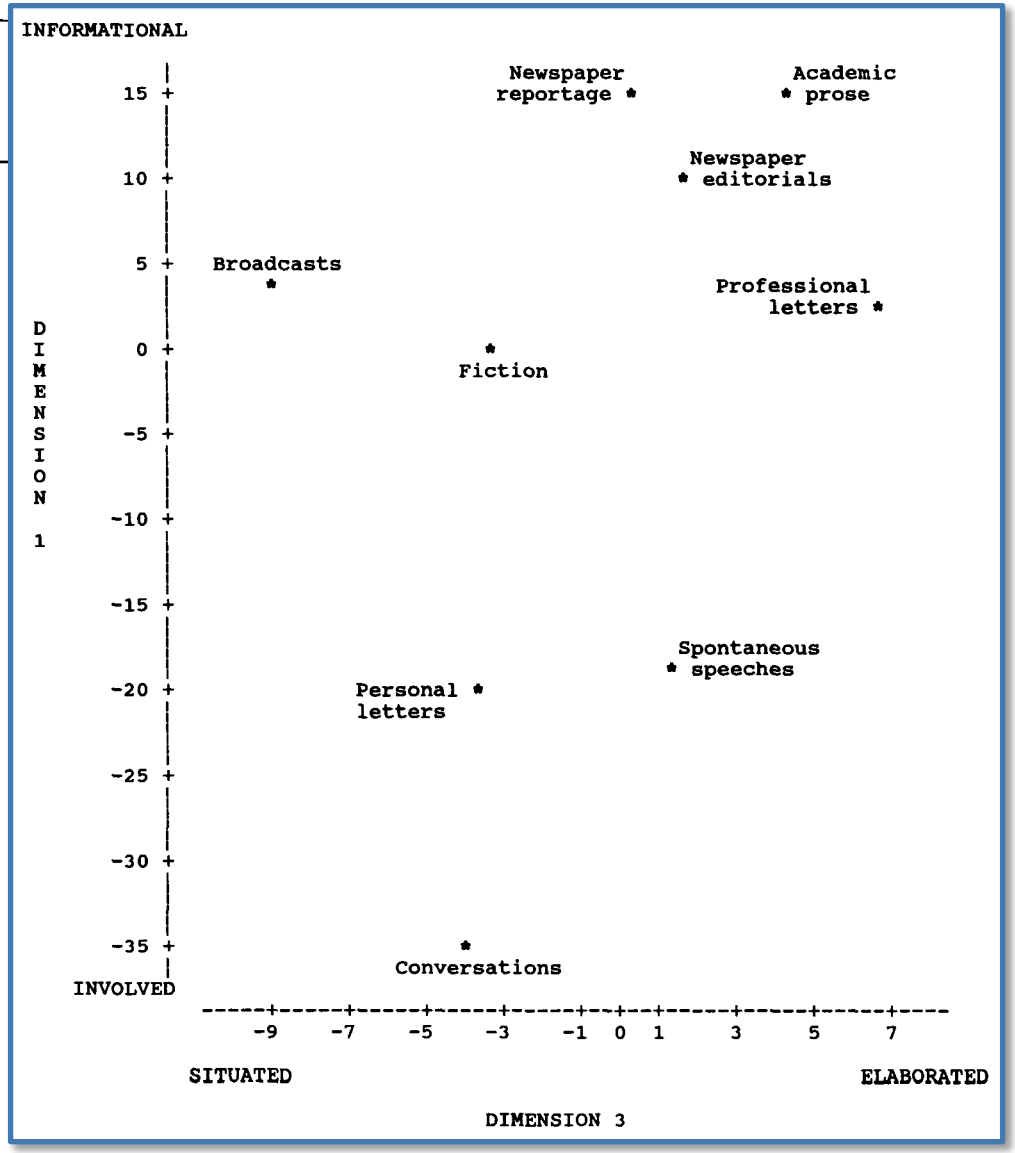
Biber's multidimensional analysis

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

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		



Problems

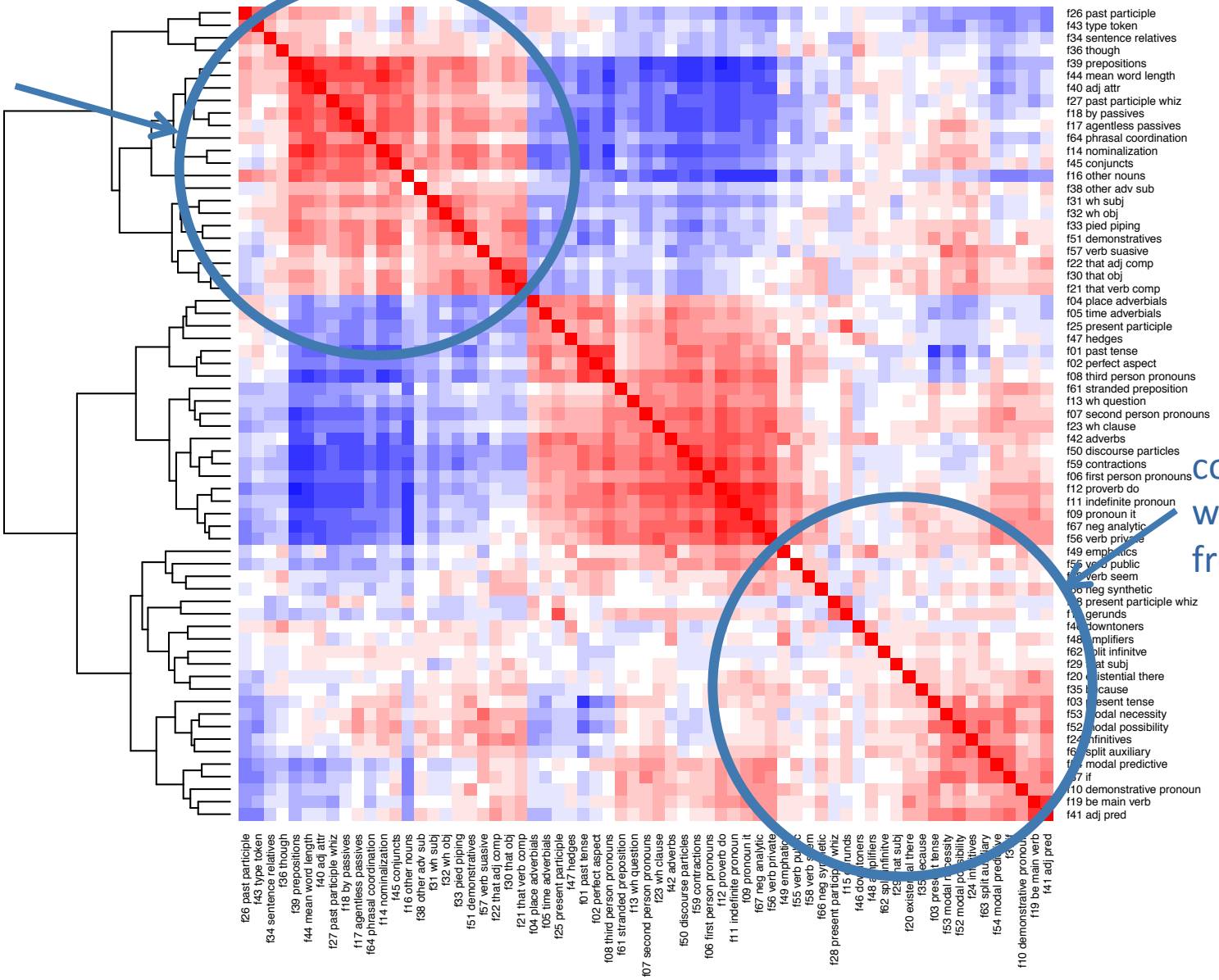
- Design bias
  - choice of features
  - selection of text samples
- Involves a miracle
  - and it isn't even a very robust one
- Interpretation bias
  - arbitrary cutoff for feature weights (“loadings”)
  - risk of reading one's own expectations into features
- More subtle patterns of variation invisible

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)
- Factor analysis with 4 latent dimensions + varimax
  - seems to yield the most clearly structured analysis

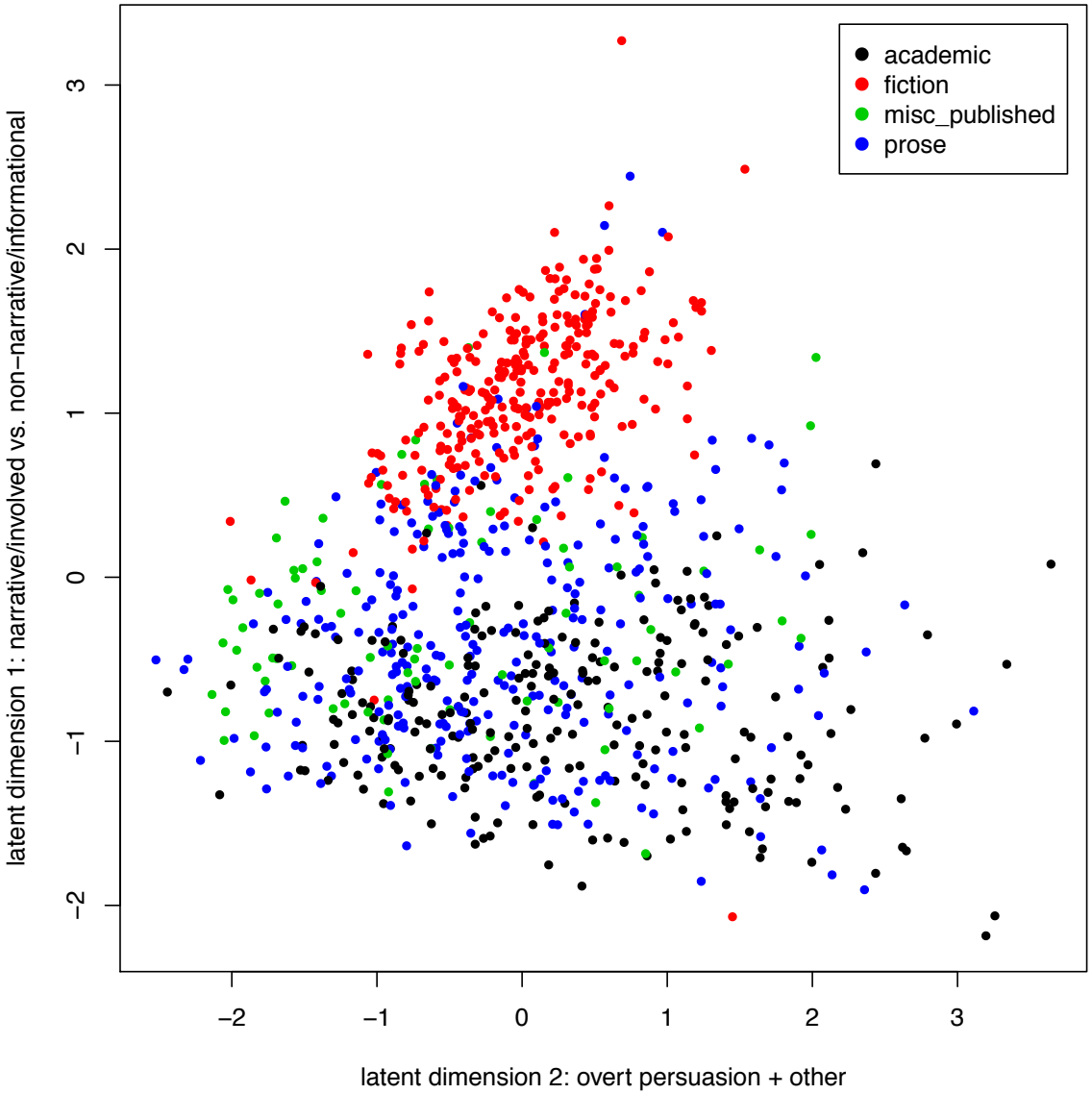
Design bias: choice of features

correlated with noun frequency

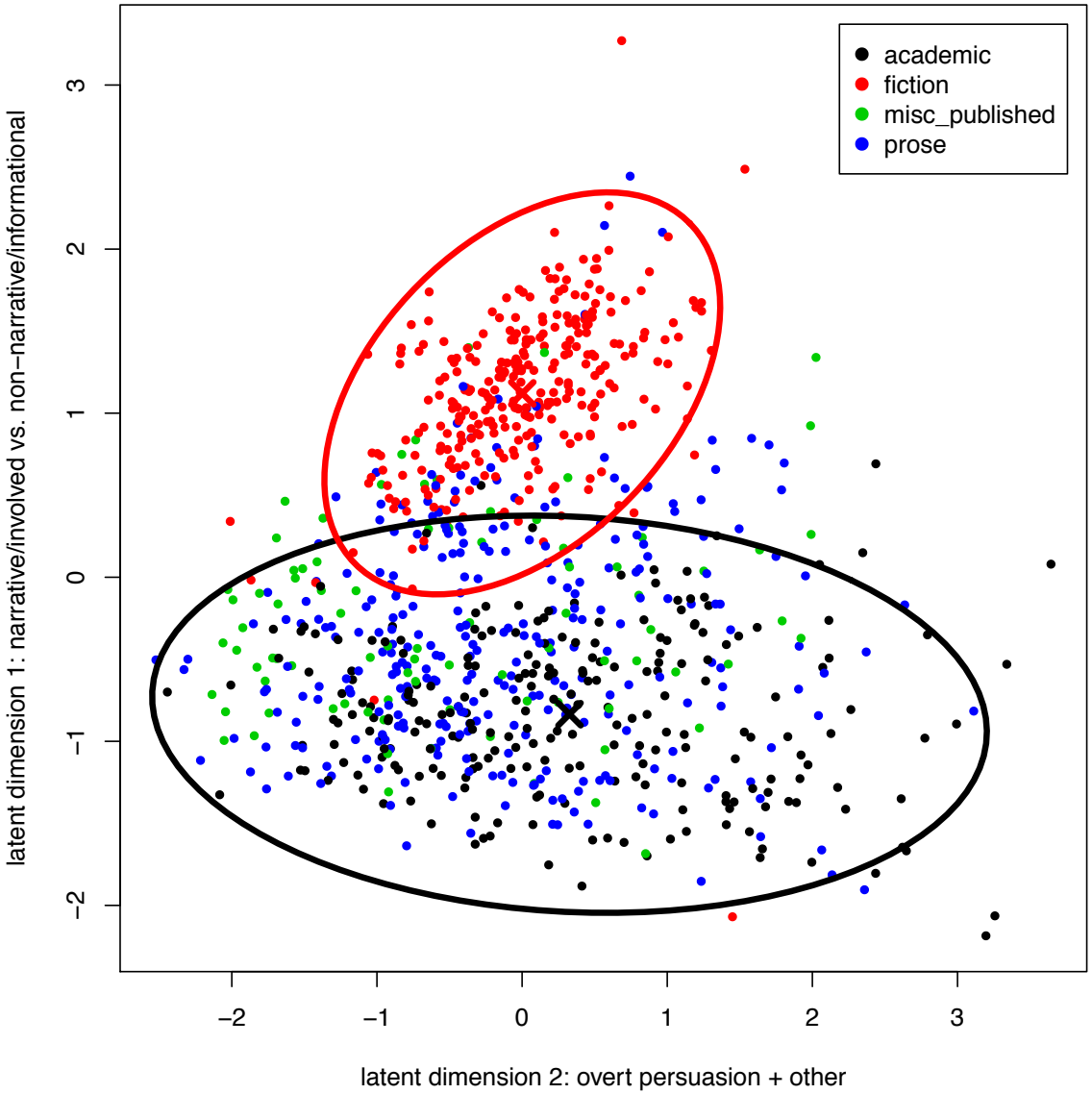


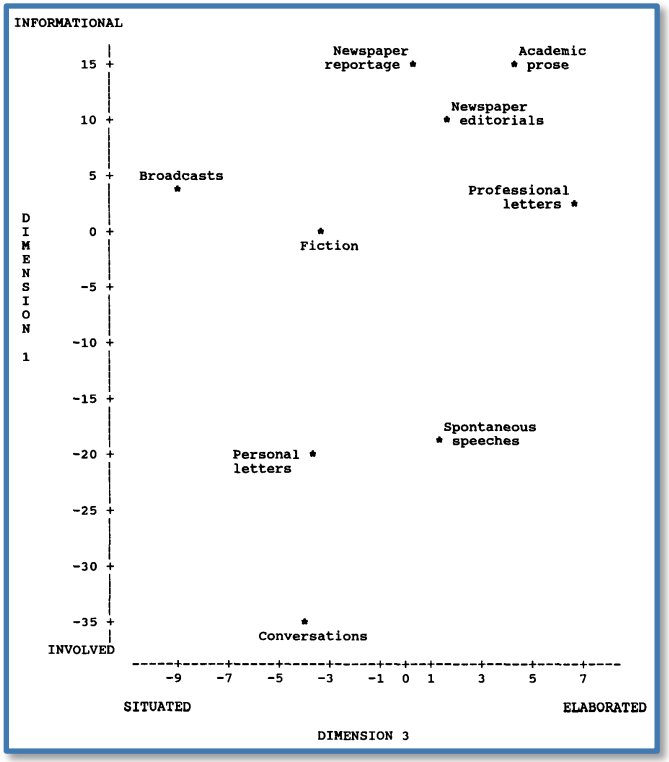
correlated with verb frequency

Design bias: choice of texts

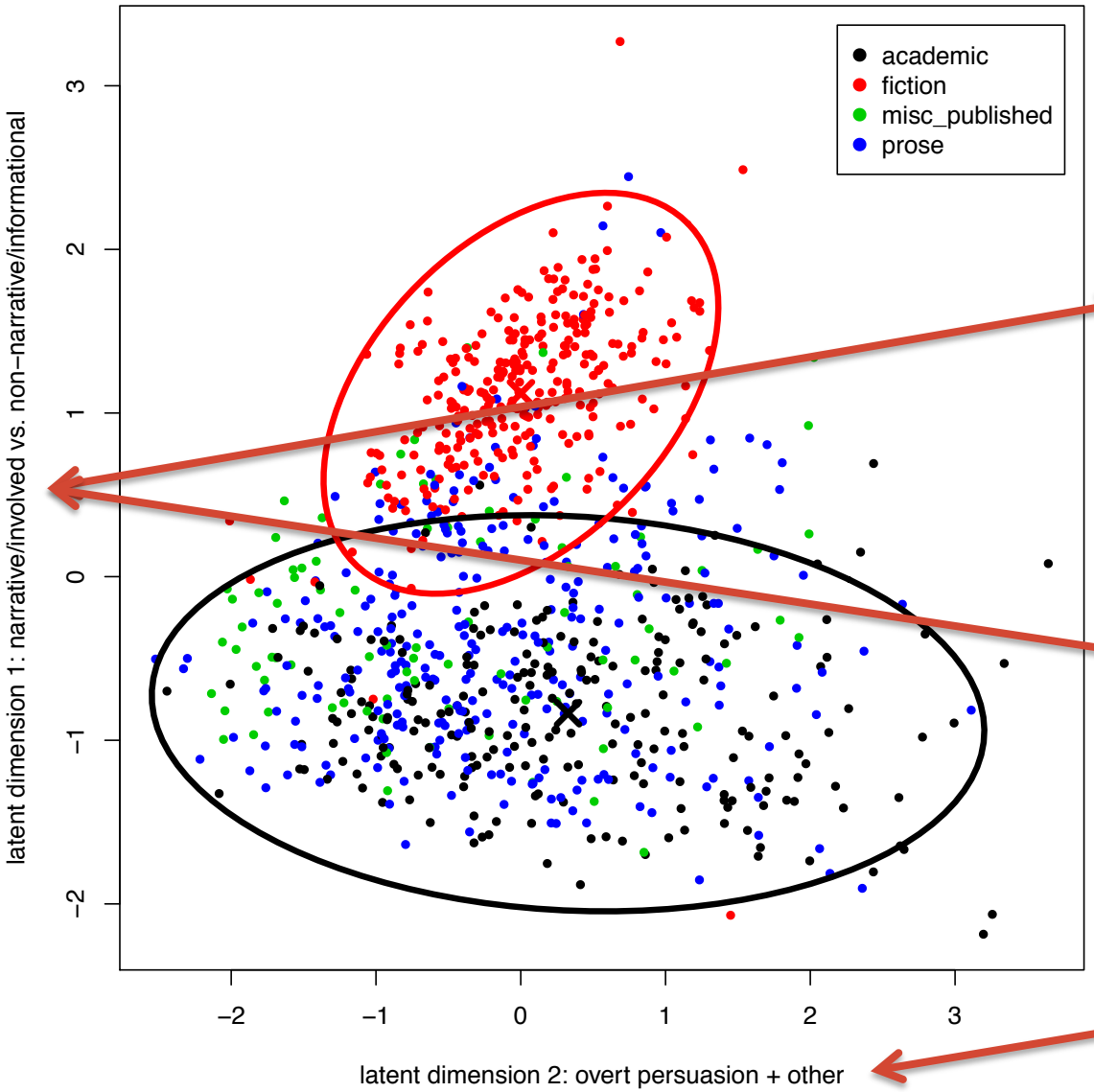


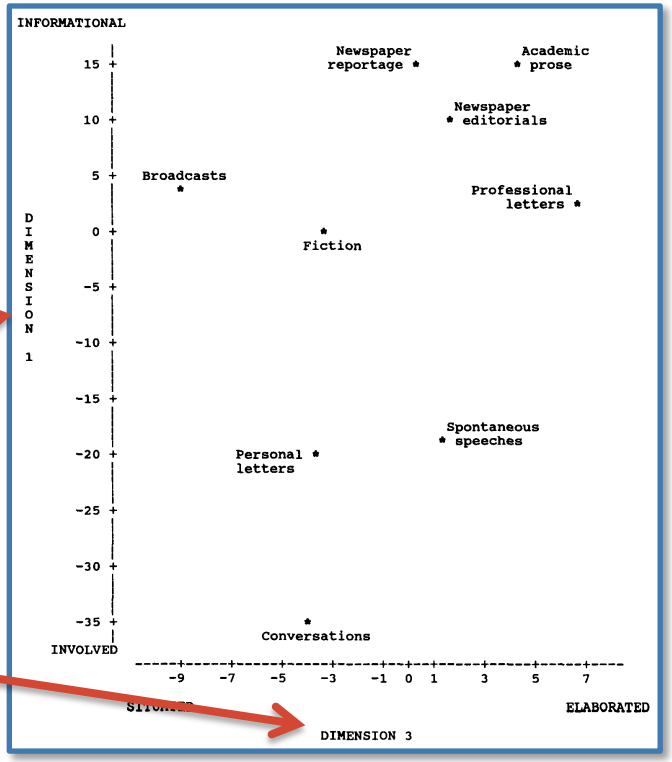
Design bias: choice of texts





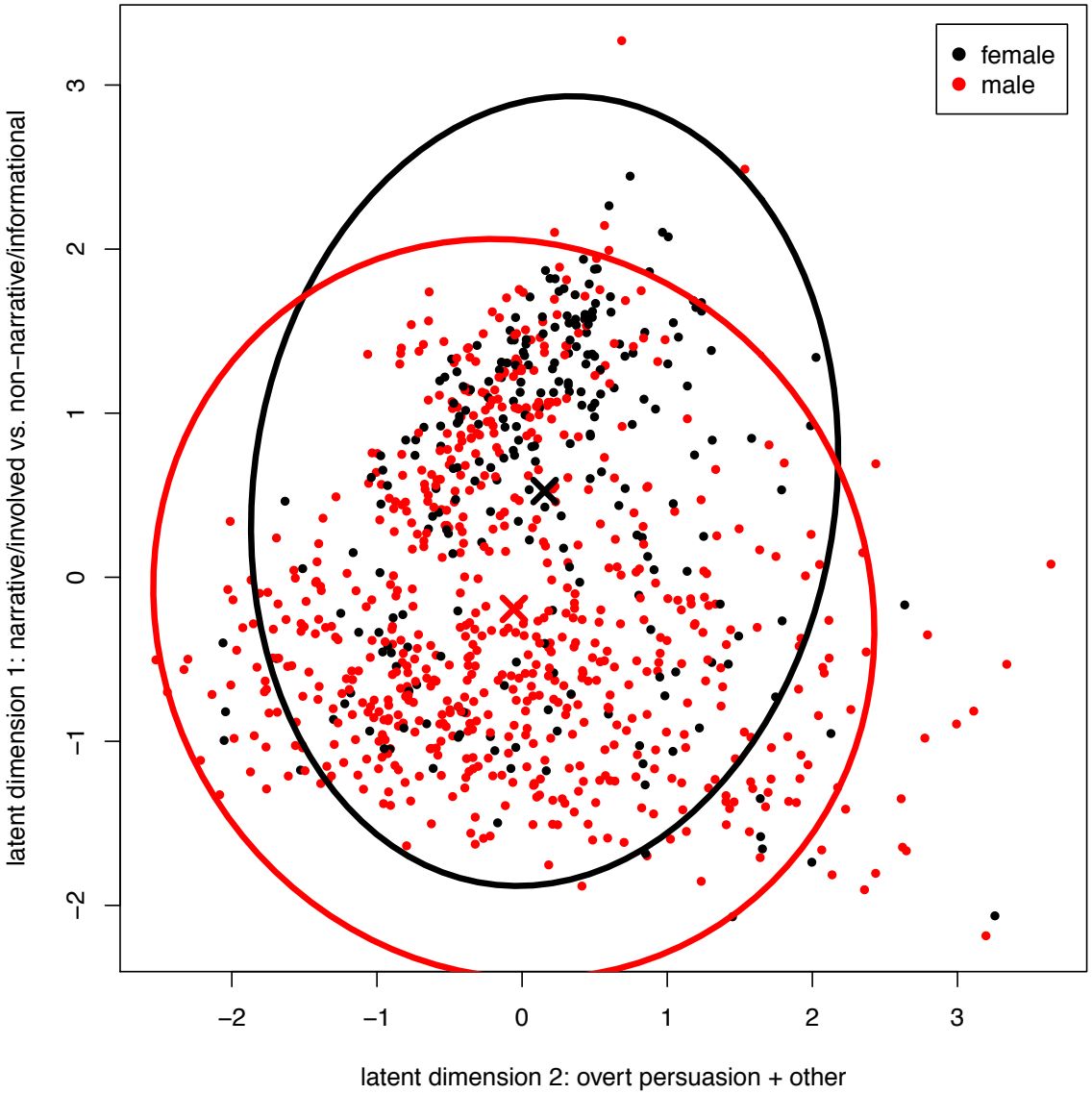
Interpretation bias





?

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%

Our approach

(Diwersy, Evert & Neumann 2014)

Our approach

(Diwersy, Evert & Neumann 2014)

- Assumption: (Euclidean) distances meaningful
  - as a measure of linguistic similarity of texts
  - depends crucially on choice of features
- Visualization to interpret geometric configuration
- Orthogonal projection = perspective on data
  - (squared) distances decompose into preserved structure + orthogonal (hidden) component
  - optimal projection: principal component analysis (PCA)
- Minimally supervised intervention
  - based on externally observable, theory-neutral information
  - method: linear discriminant analysis (LDA)

Case studies

- Translation effects and register variation in German and English (Evert & Neumann in prep.)
- Regional varieties of French, based on colligational frequencies in newspaper texts (Diwersy et al. 2014)
- Work in progress: Authorship attribution with Burrows Delta (Evert et al. 2015)

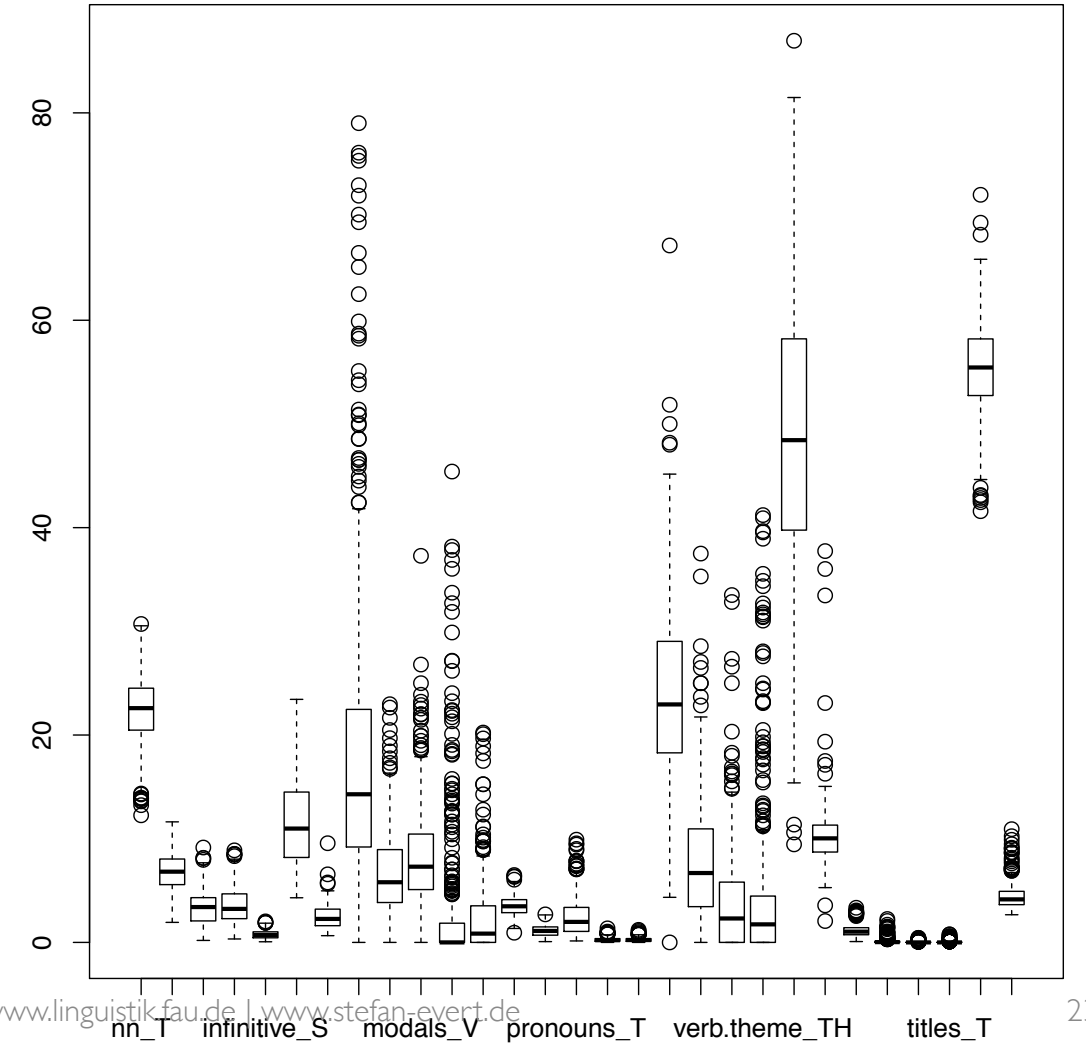
Case study 1: CroCo

Diwersy, Evert & Neumann (2014); Evert & Neumann (in prep.)

- CroCo: parallel corpus English/German
  - English-German and German-English translation pairs
  - 454 texts from 8 different genres
- 28 lexico-grammatical features (Neumann 2013)
  - comparable btw. languages, try to reduce correlations
  - inspired by SFL and translation studies
- Text = point in 28-dimensional feature space
- PCA identifies latent dimensions of variation
  - FA results are very similar → comparable to Biber approach
- Focus on English texts here (originals and translations)

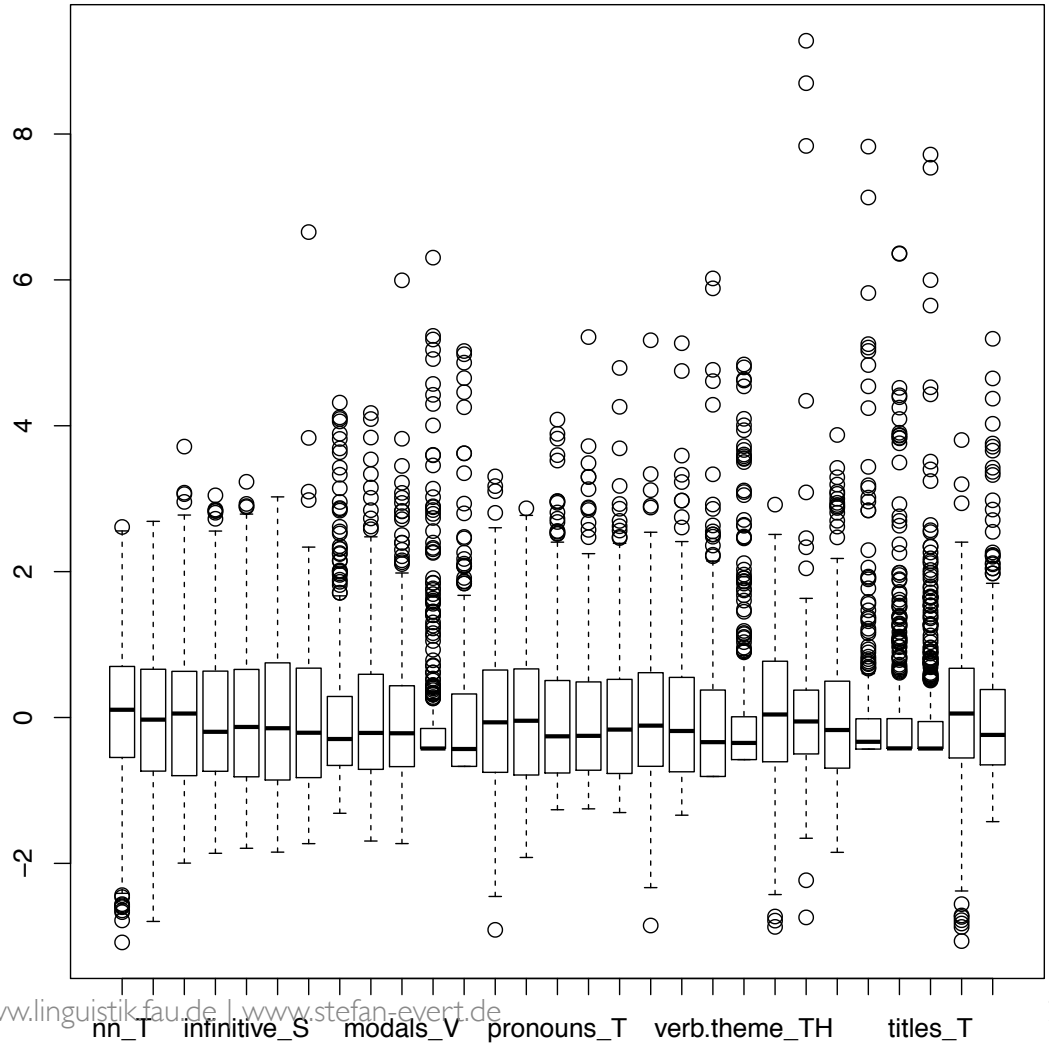
Methodological issues

- Feature scaling
- Choice of features
- Choice of texts
- Delicate effects are obscured

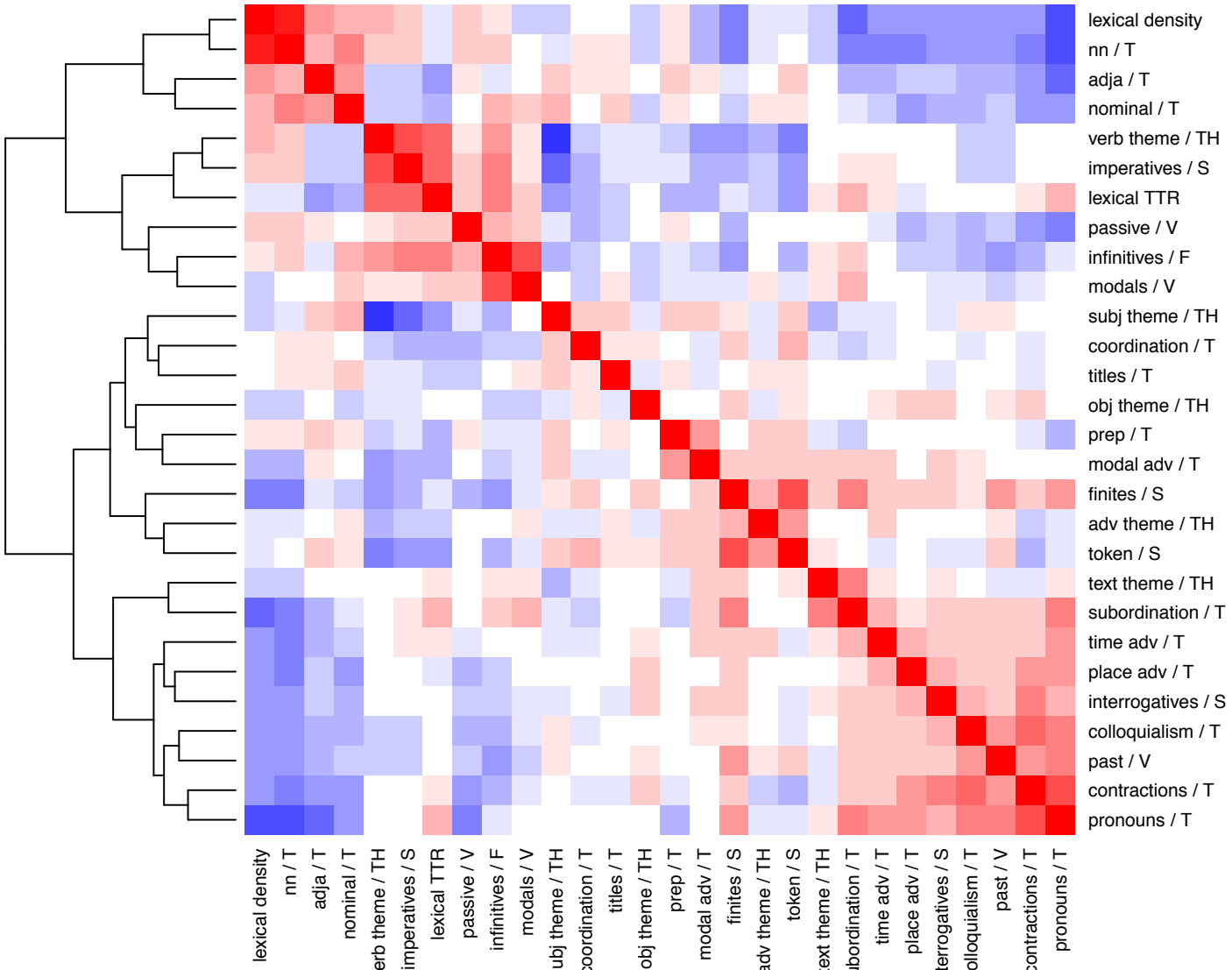


Methodological issues

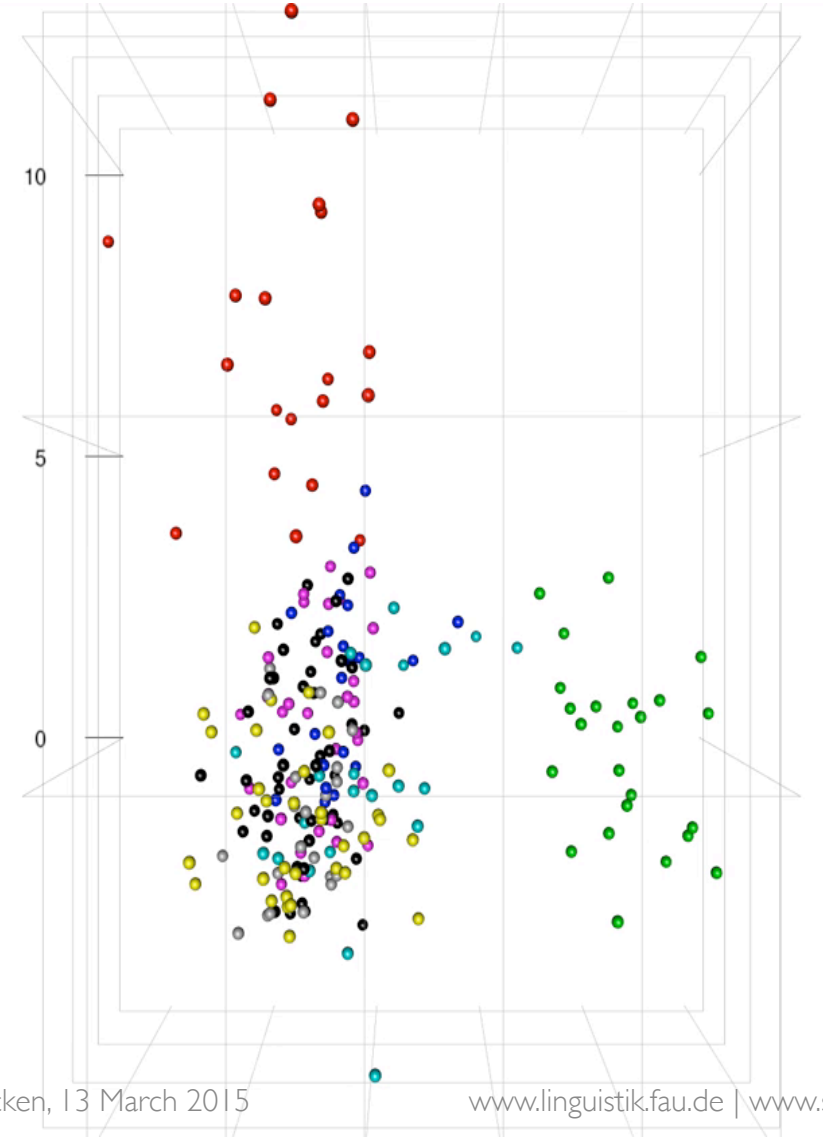
- Feature scaling
- Choice of features
- Choice of texts
- Delicate effects are obscured



Case study 1: CroCo

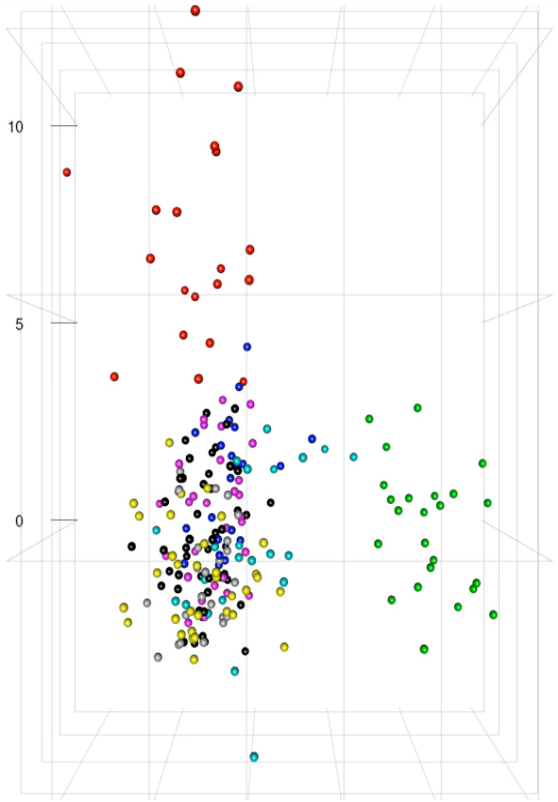


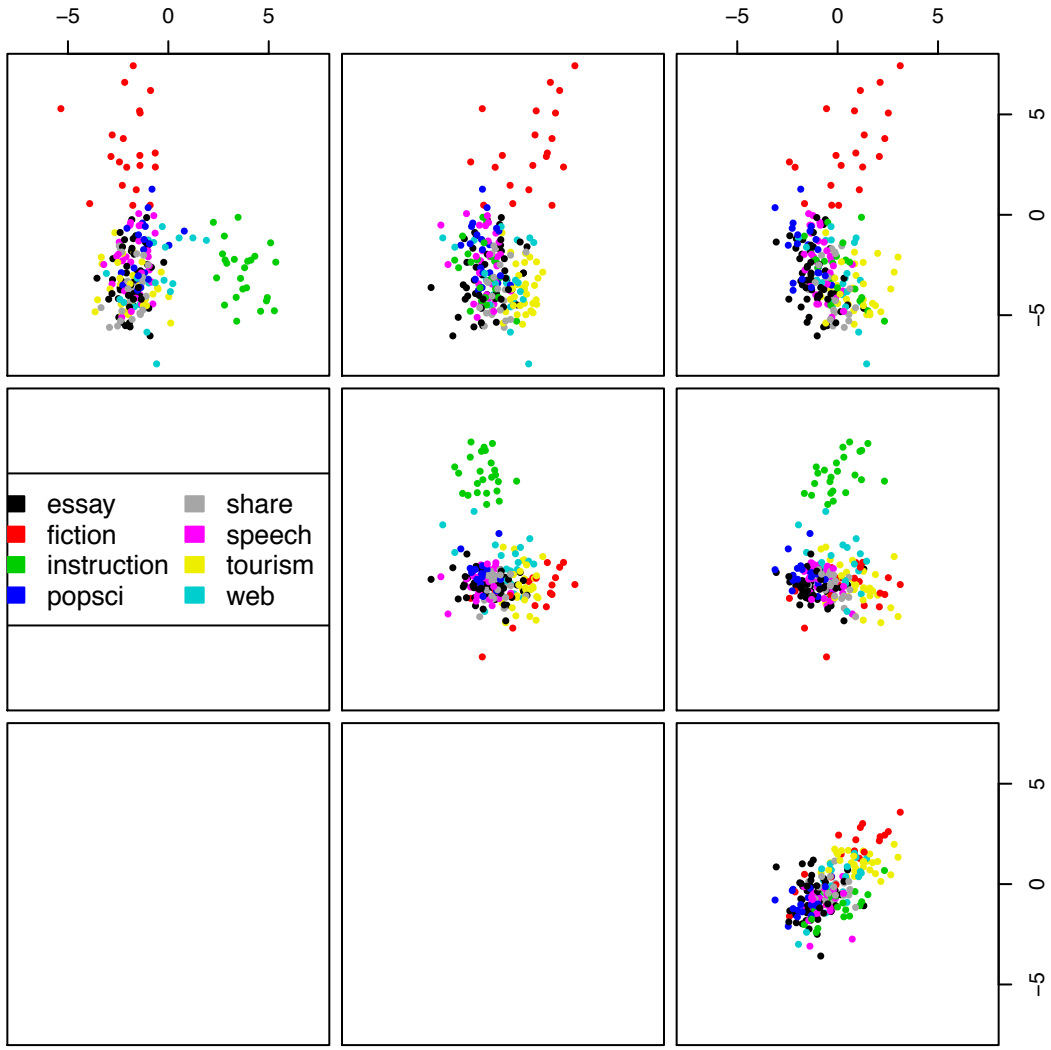
Case study 1: CroCo



- essay
- fiction
- instruction
- popsci
- share
- speech
- tourism
- web

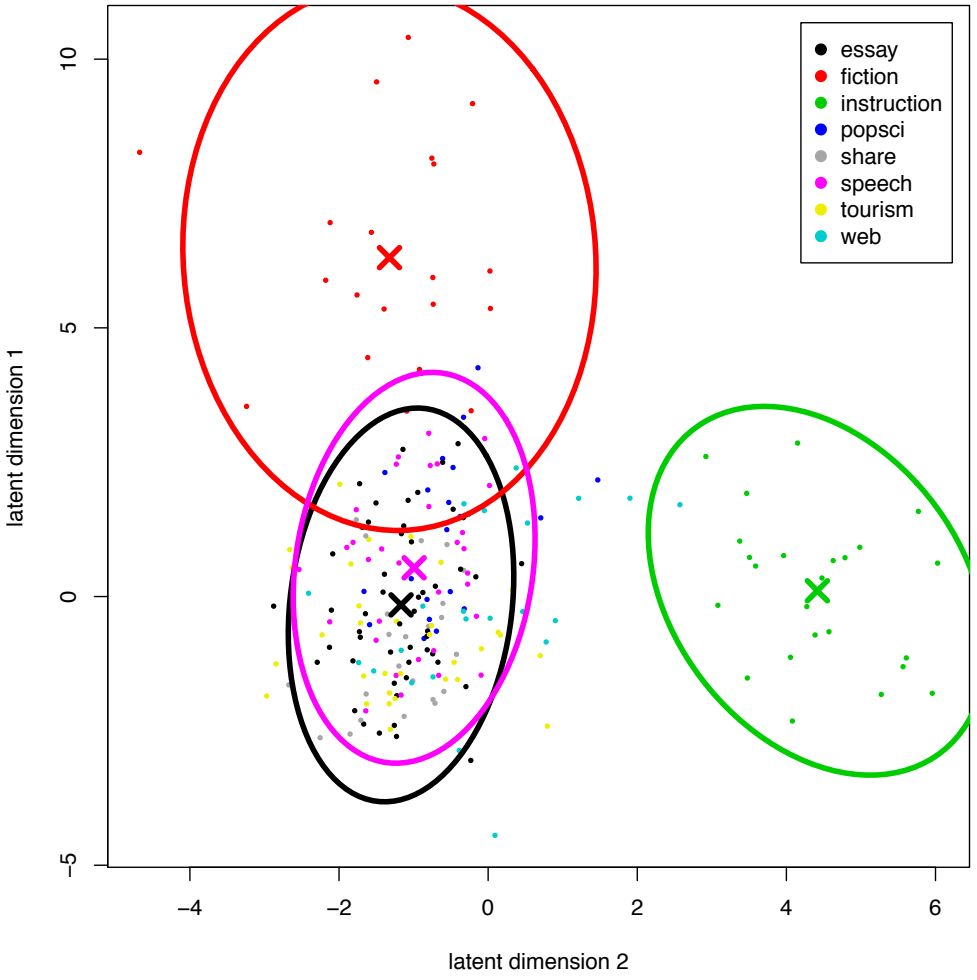
Case study 1: CroCo





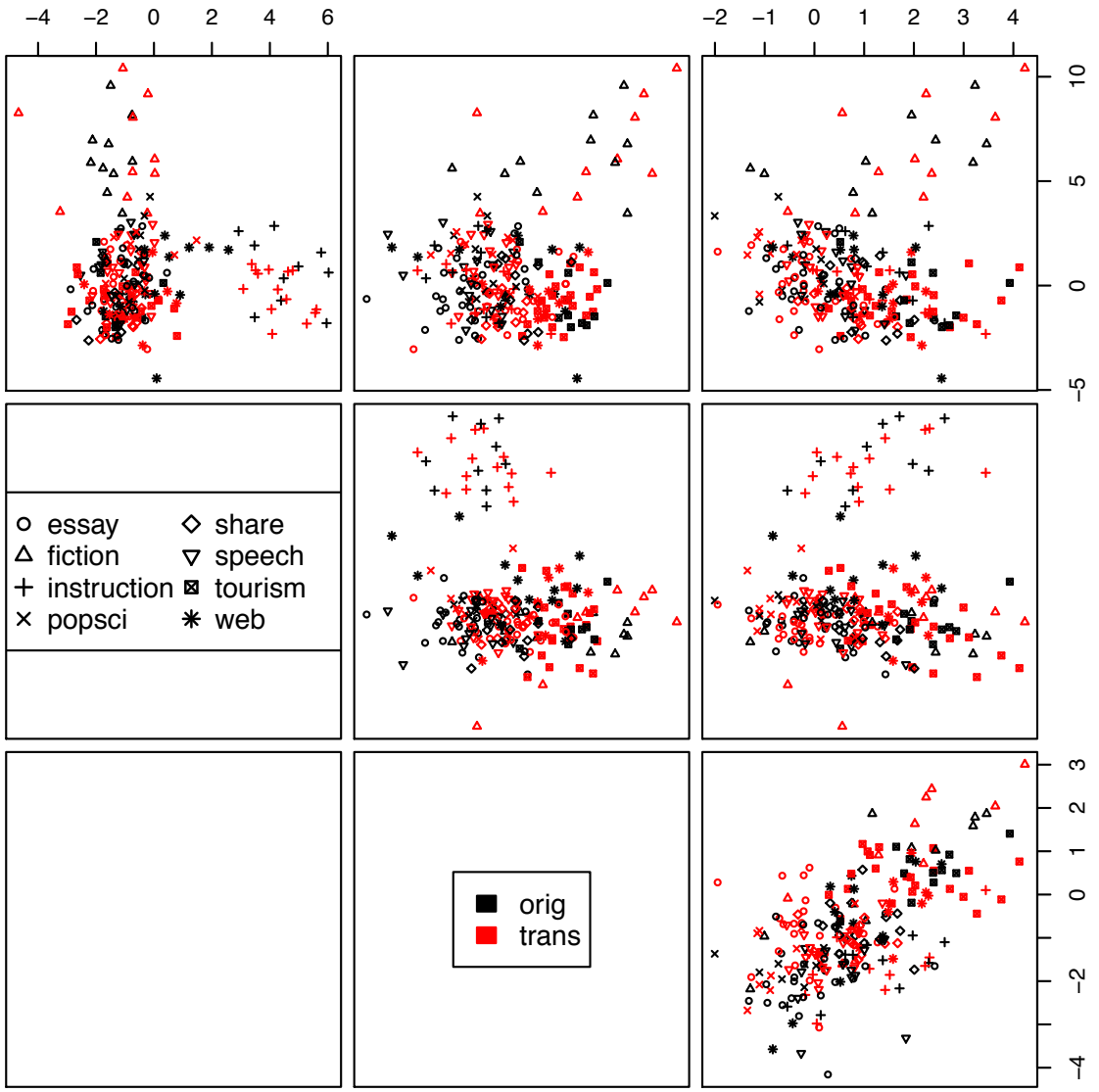
Case study 1: CroCo

- Focus on first two latent dimensions (→ Biber's map)
- Describe genre by centroid and confidence ellipse
- Comparison with Hotelling's t^2 test
  - essay vs. speech
  - t^2=2.512, df=2/80, p=.0875 n.s.

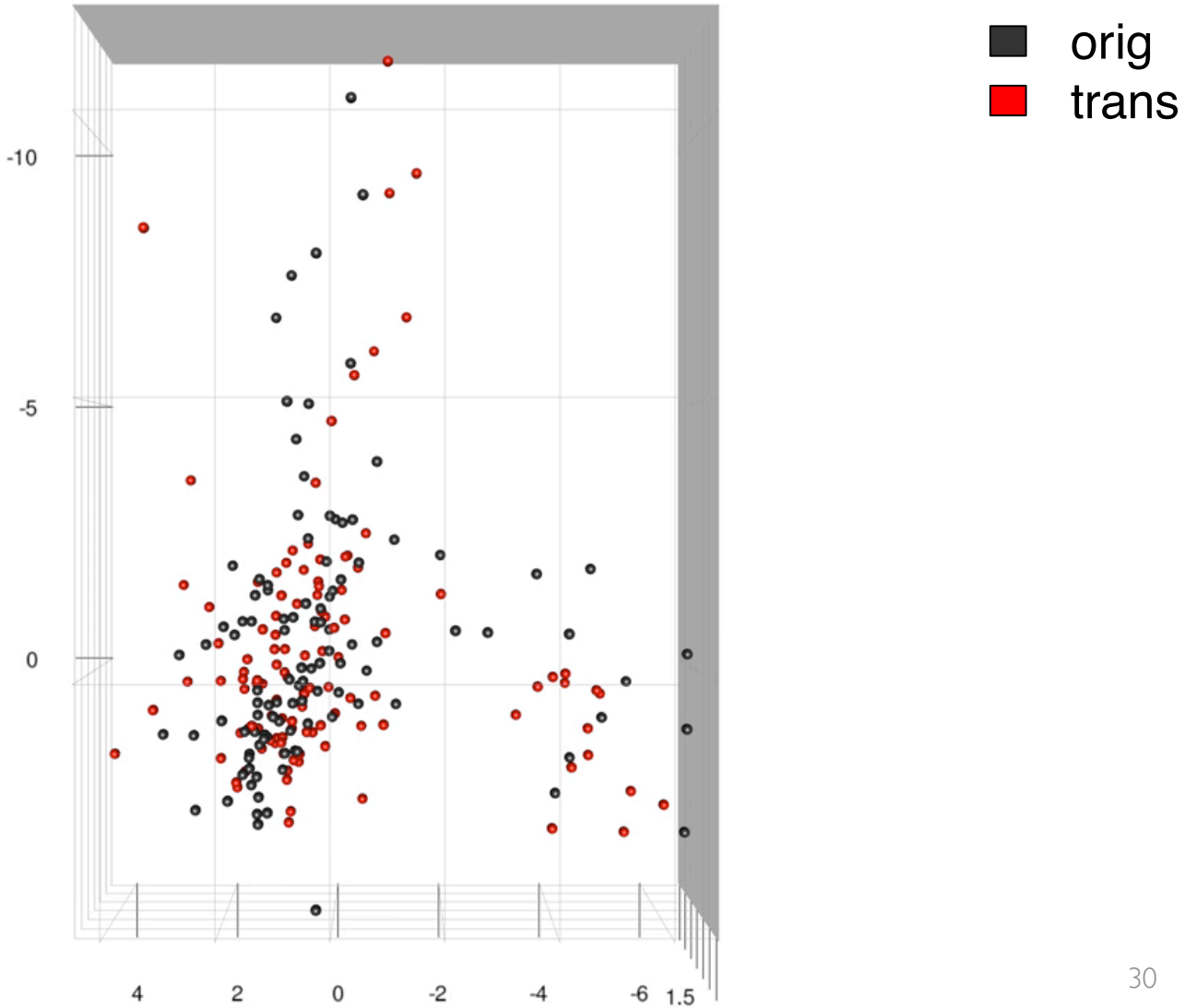


How about subtle patterns?

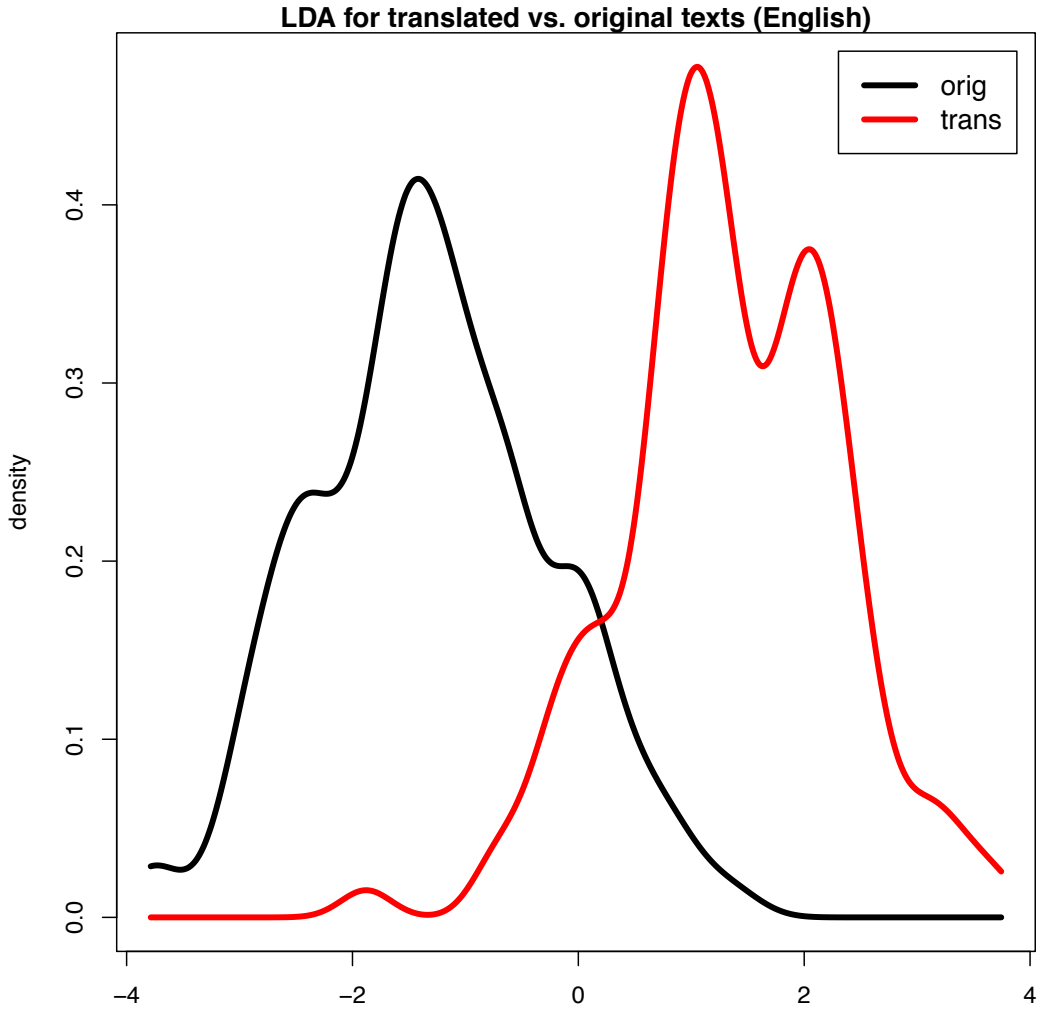
- PCA dimensions fail to distinguish translations from original texts
- But a SVM machine learner can do this with 85% accuracy
- Replace one PCA dimension with LDA discriminant for **orig** vs. **trans**
  - external & theory-neutral information



Finding the right perspective



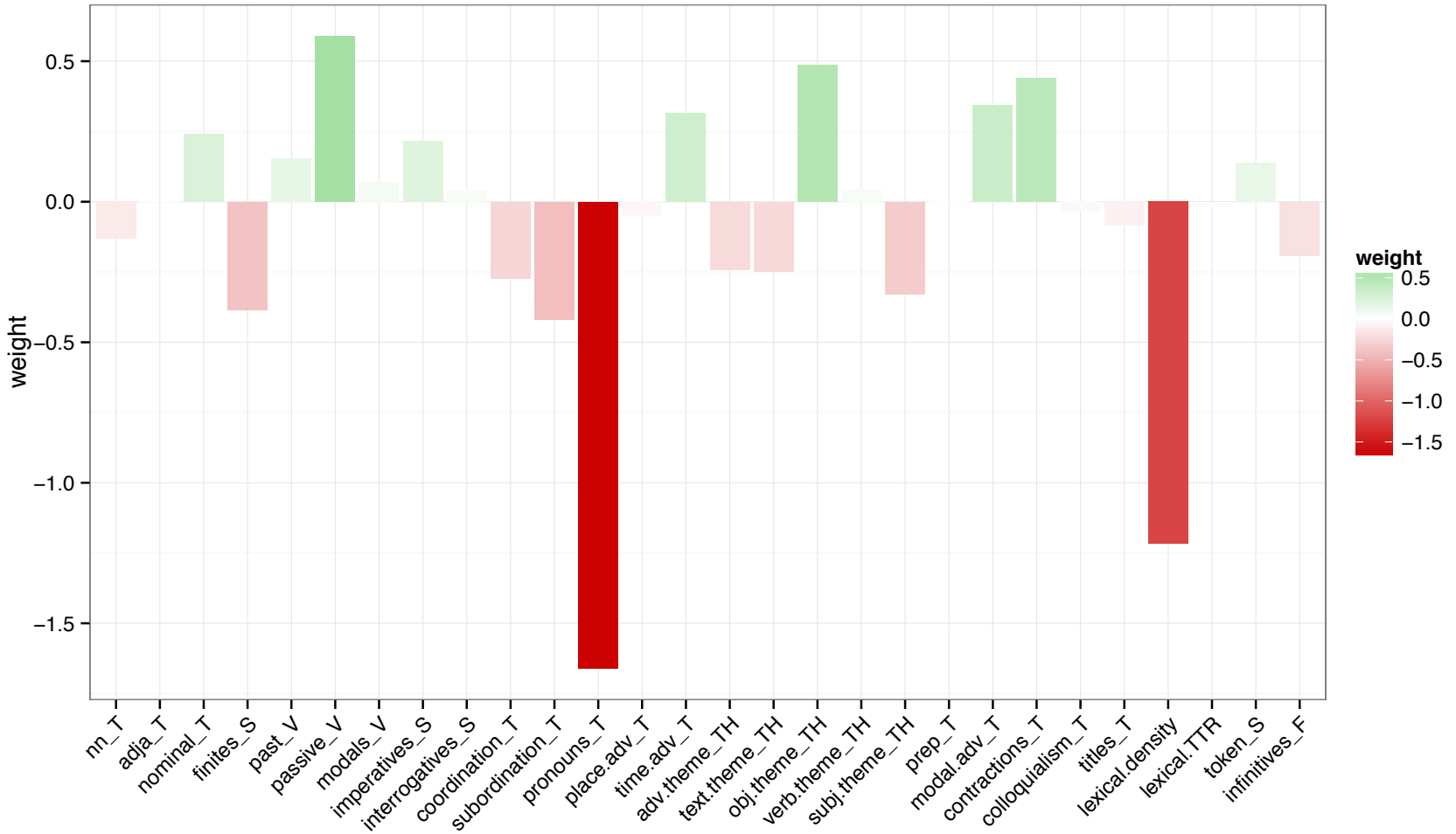
Interpreting discriminant features



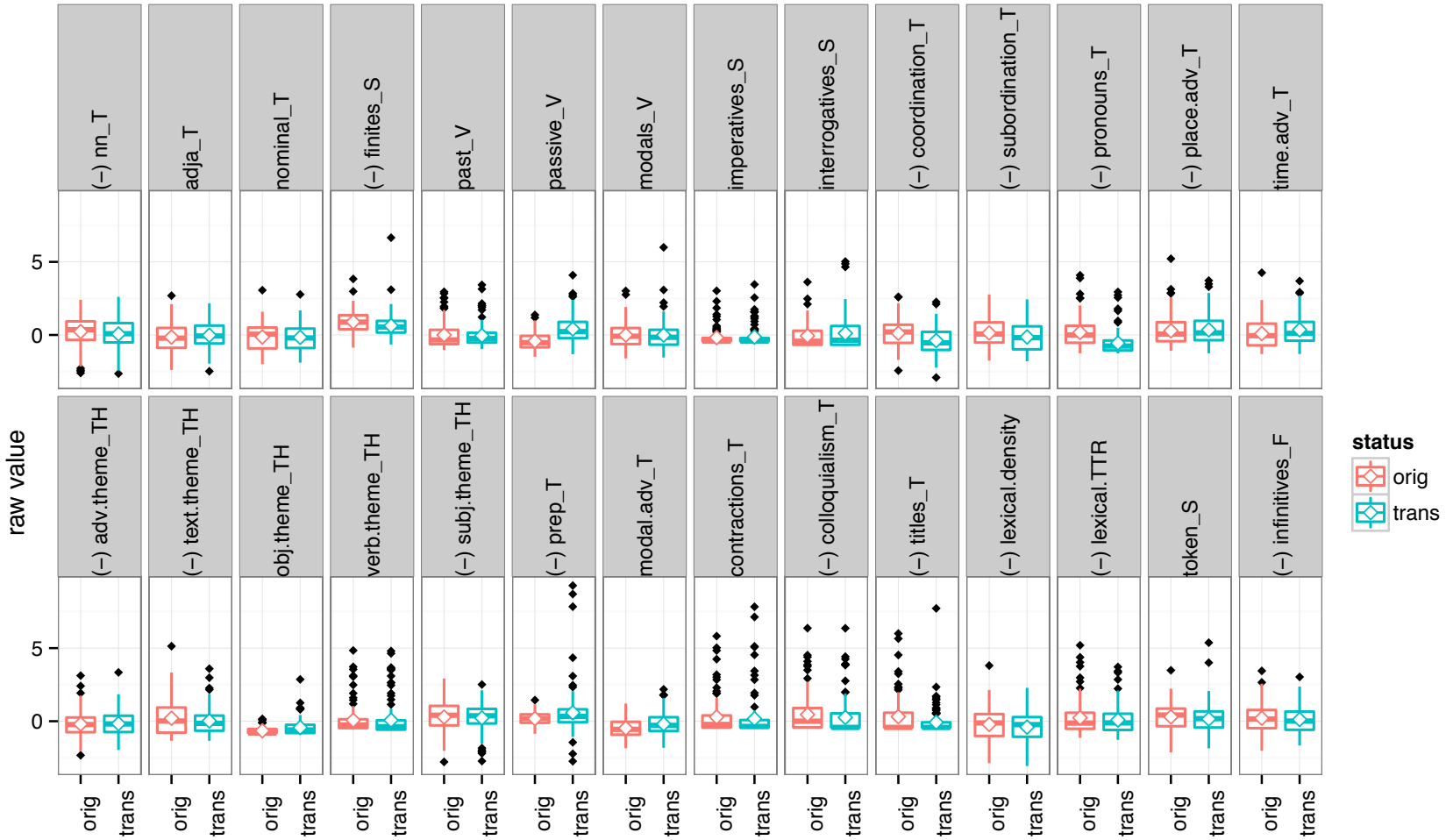
originals

translations

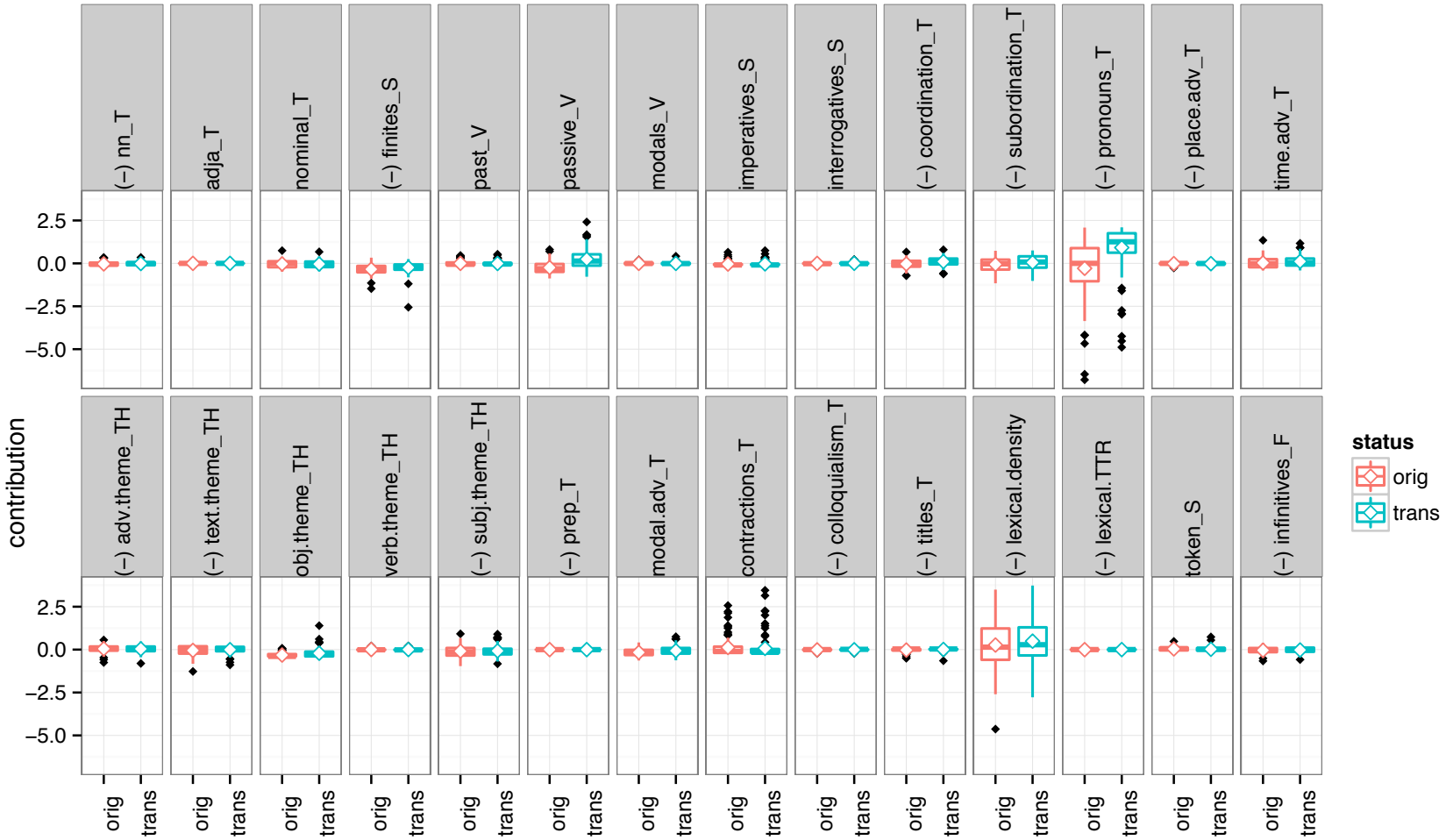
Interpreting discriminant features



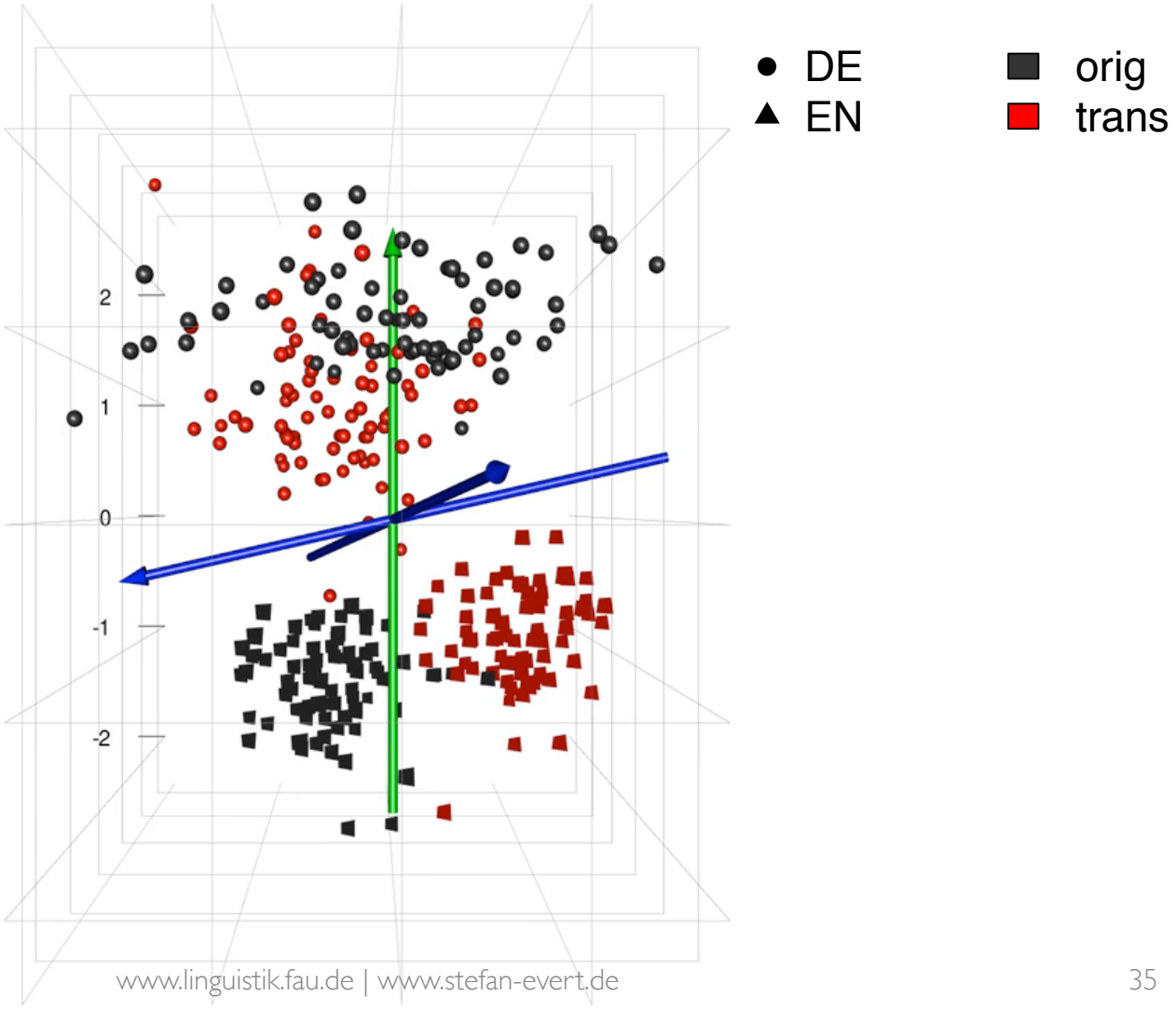
Interpreting discriminant features



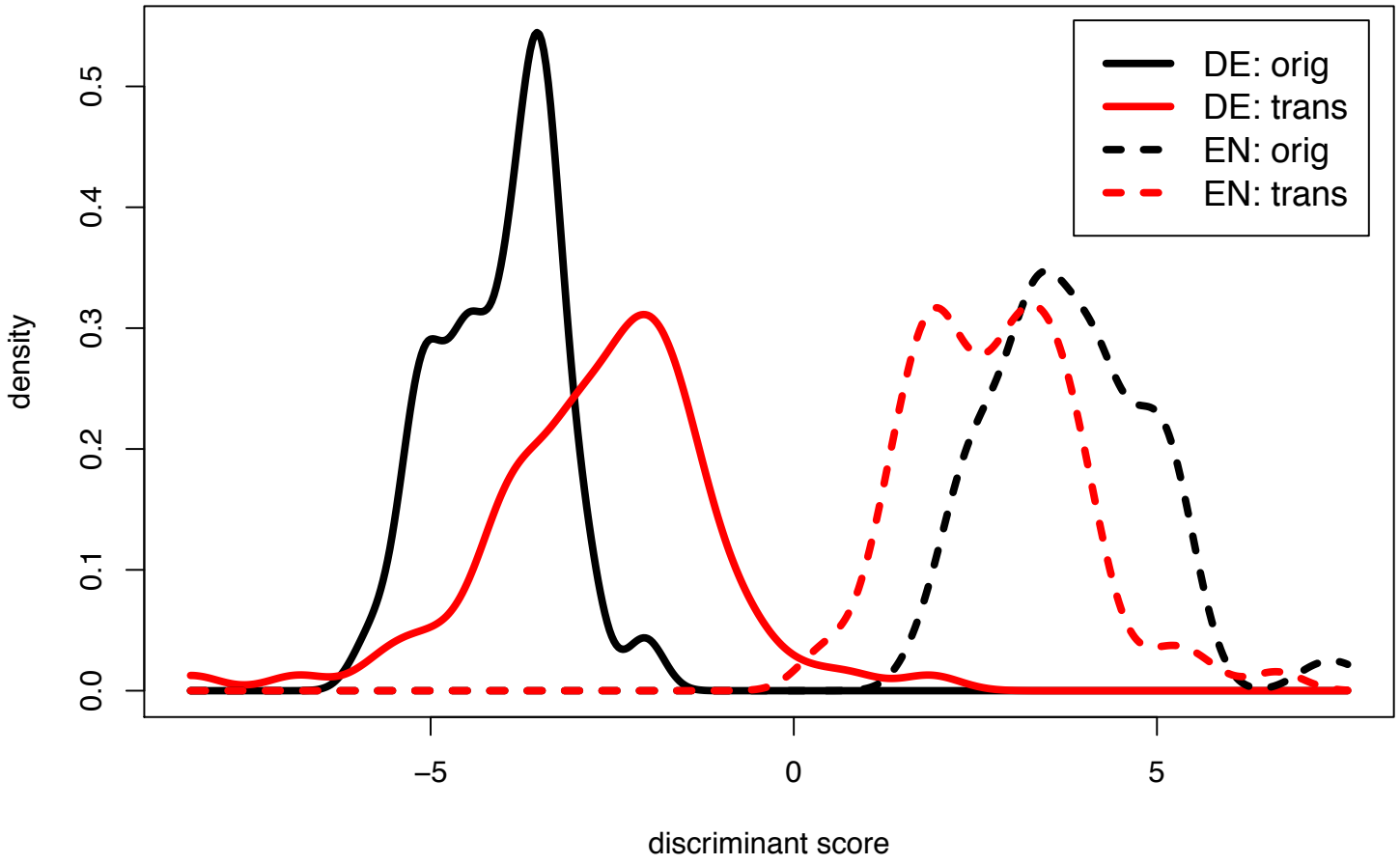
Interpreting discriminant features



Interpreting geometric configurations: German vs. English



Discriminant for DE/EN: Evidence for shining through & prestige?



Case study 2: French regional varieties

Diwersy, Evert & Neumann (2014)

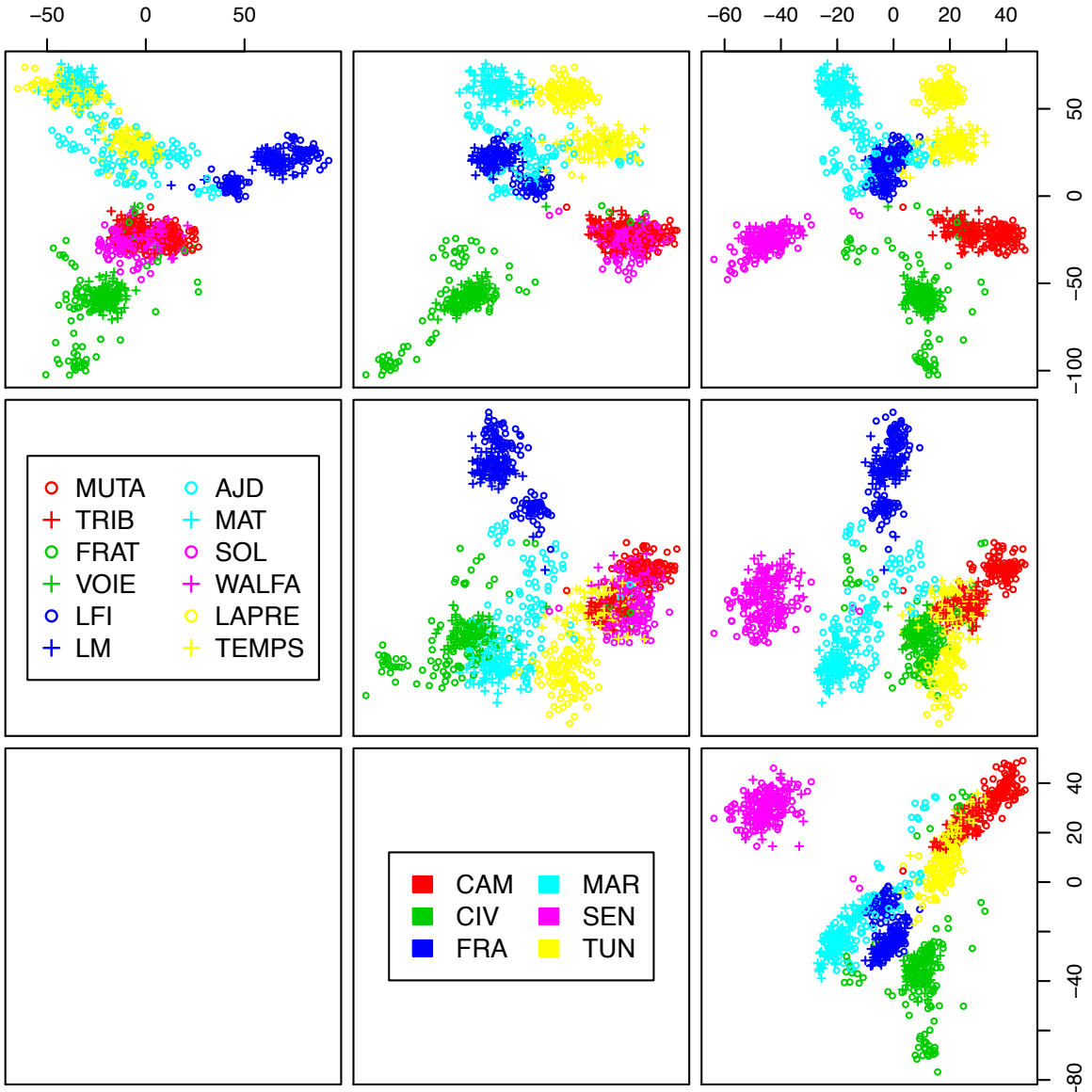
- Lexical differences in regional varieties of French
- Two nation-wide newspapers each from 6 countries
  - Cameroon, France, Ivory Coast, Morocco, Senegal, Tunisia
  - two consecutive volumes from each newspaper
  - total size approx. 14.5 million tokens
- Text samples = one week each
- Features: frequencies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

Case study 2: French regional varieties

PCA including country-specific words as features: perfect separation

Design bias results in a completely uninteresting model

FA not applicable: features >> texts

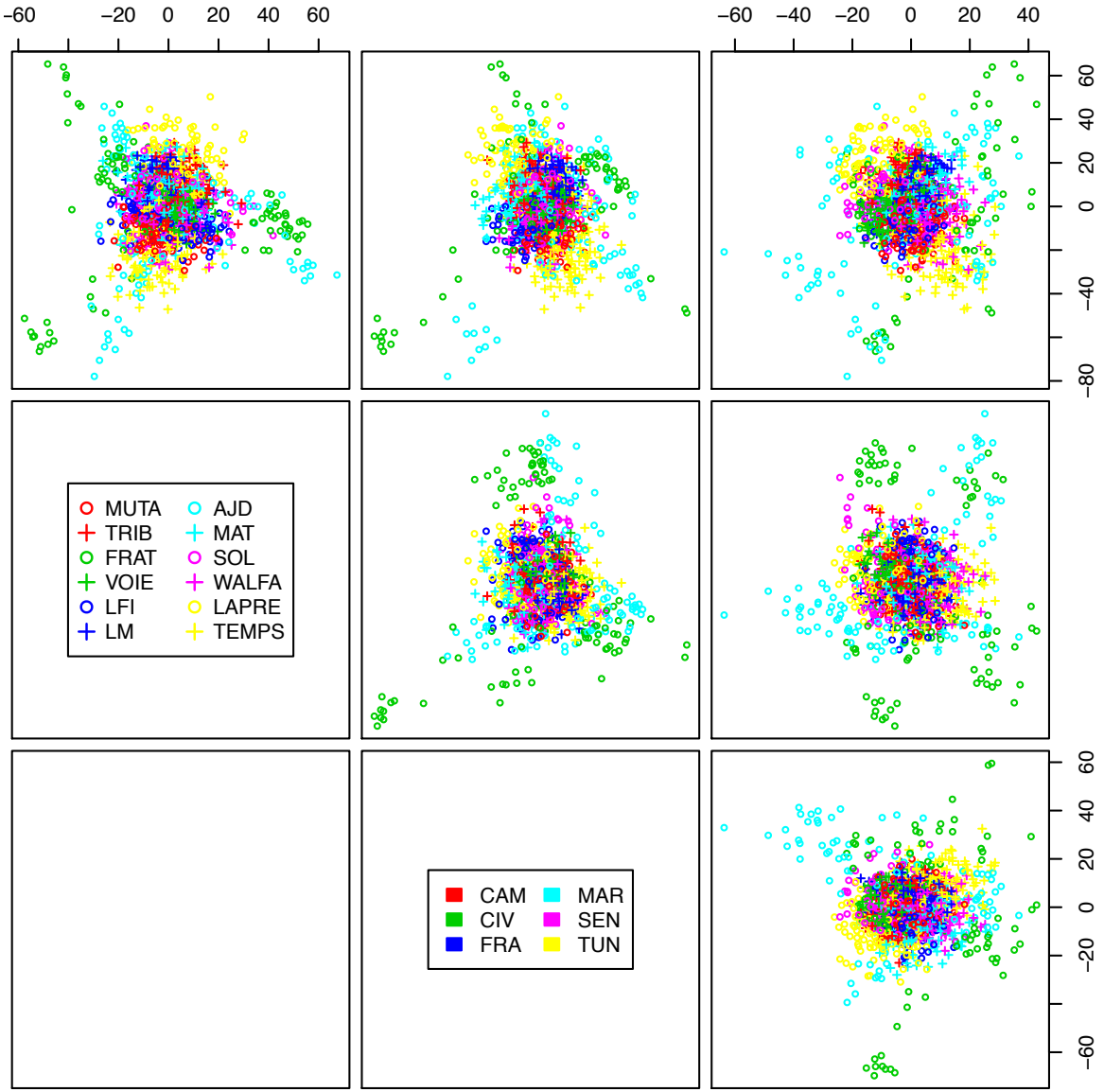


Case study 2: French regional varieties

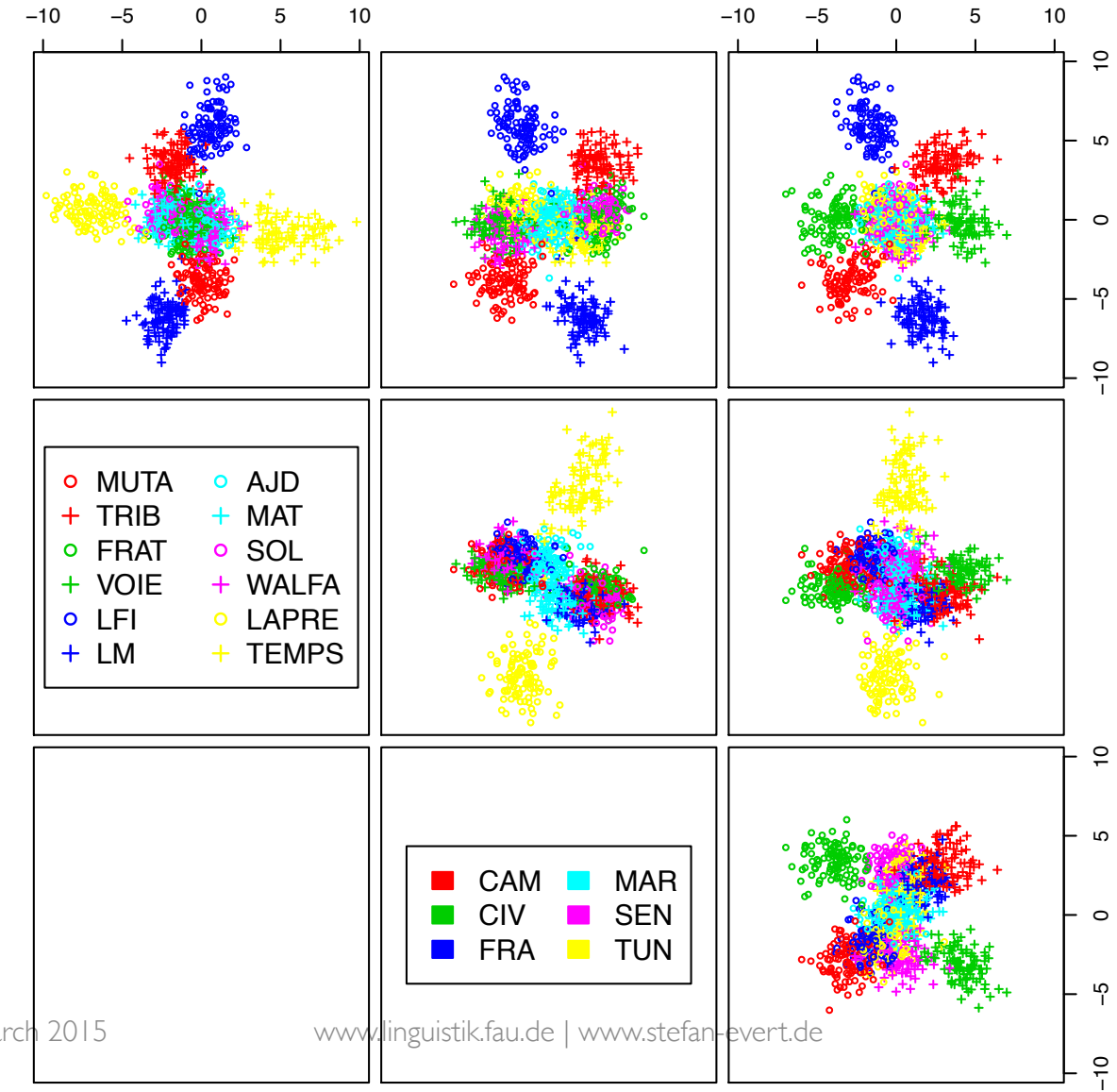
Using only shared words as features, PCA no longer reveals any patterns (just a few outliers)

Use LDA to find a meaningful per-spective, based on newspaper source

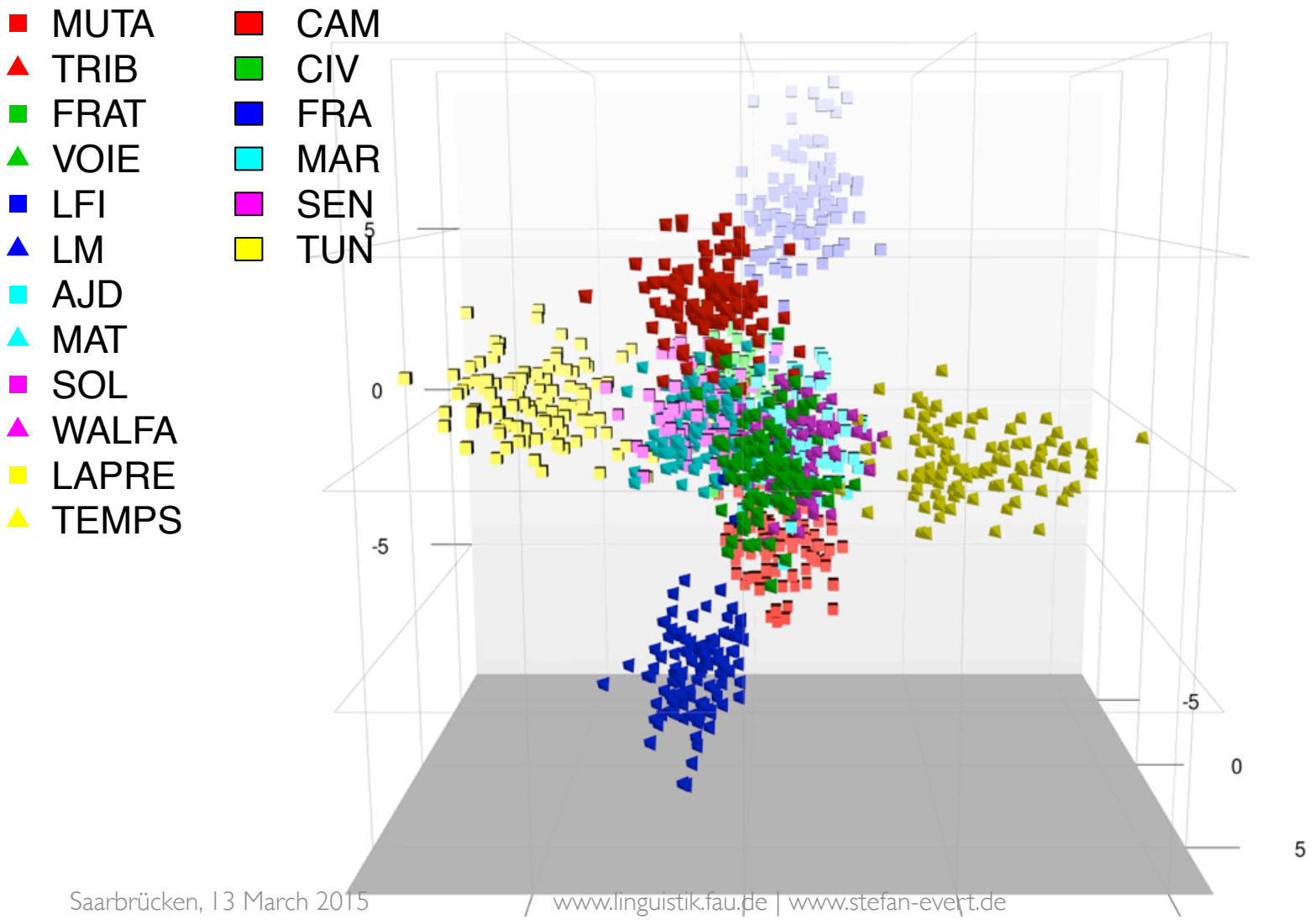
Country would presume regional varieties exist!



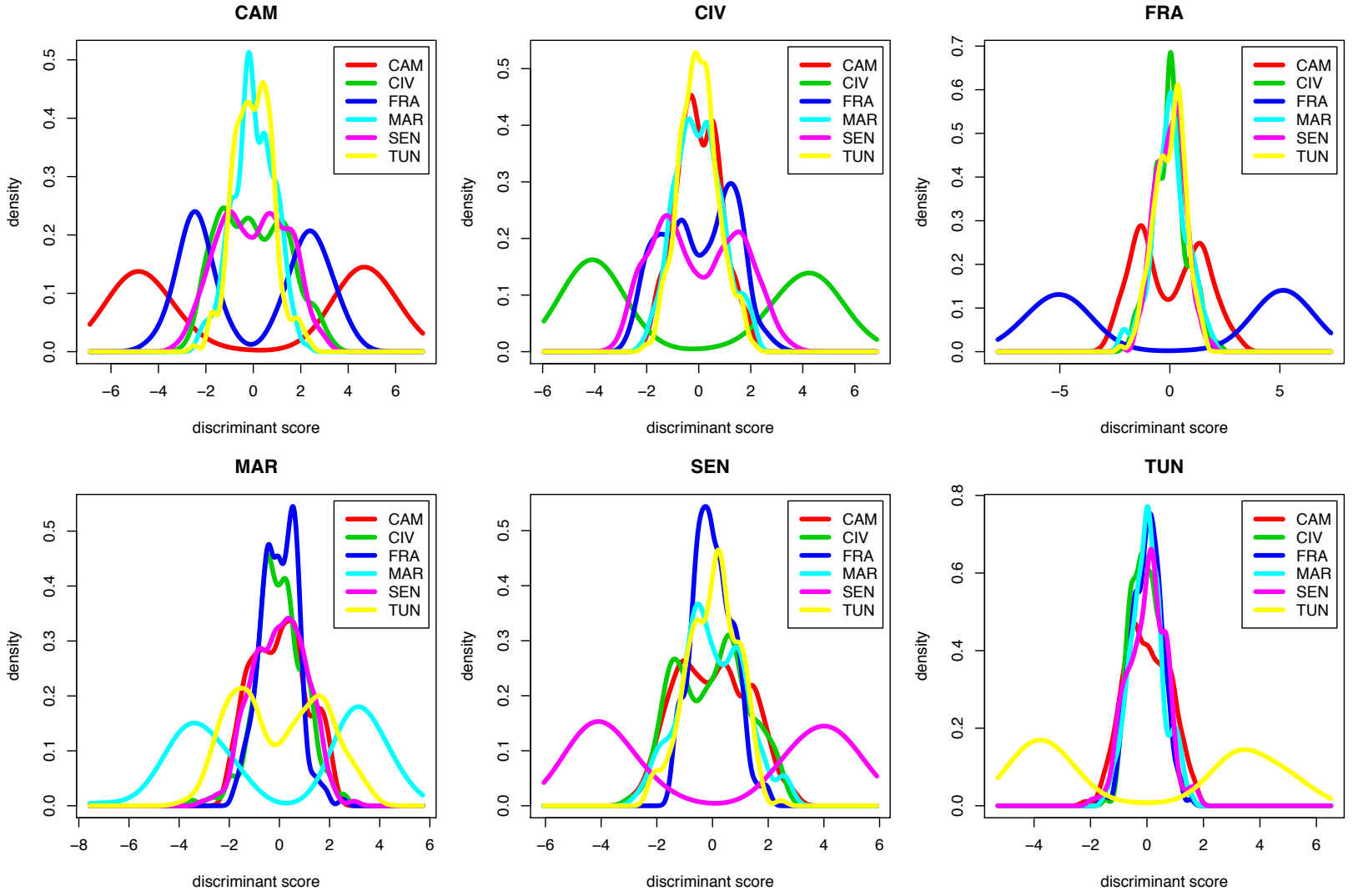
Case study 2: French regional varieties



LDA dimensions (newspapers)



Discriminant axes (newspapers)



THANK YOU!

References

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

Diwersy, Sascha; Evert, Stefan; Neumann, Stella (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, Stefan & Neumann, Stella (in prep.). The impact of translation direction on the characteristics of translated texts: a multivariate analysis for English and German.

Evert, Stefan; Proisl, Thomas; Schöch, Christof; Jannidis, Fotis; Pielström, Steffen; Vitt, Thorsten (2015). Explaining Delta, or: How do distance measures for authorship attribution work? *Presentation at Corpus Linguistics 2015*, Lancaster, UK.

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

Koppel, Moshe; Argamon, Shlomo; Shimoni, Anat Rachel (2003). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, **17**(4), 401–412.

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