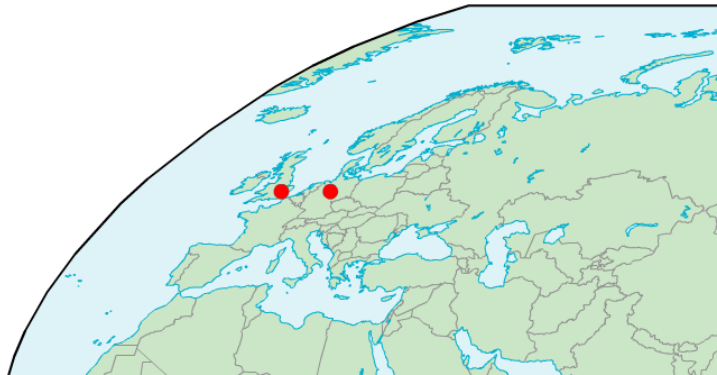
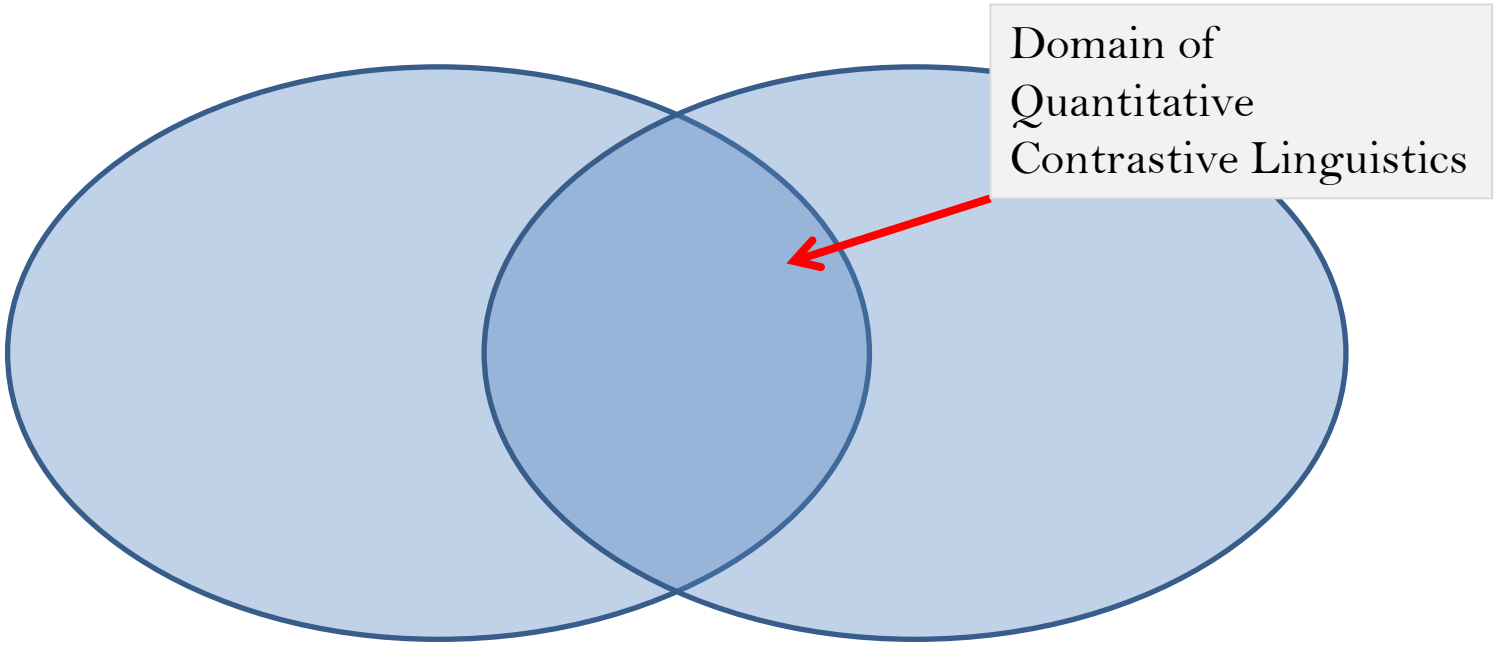
5th FLaRN Conference
Formulaic Language Research Network
28-30 March 2012
Tilburg University

Towards a Constructionist Quantitative Contrastive Linguistics

—A methodological Proposal—



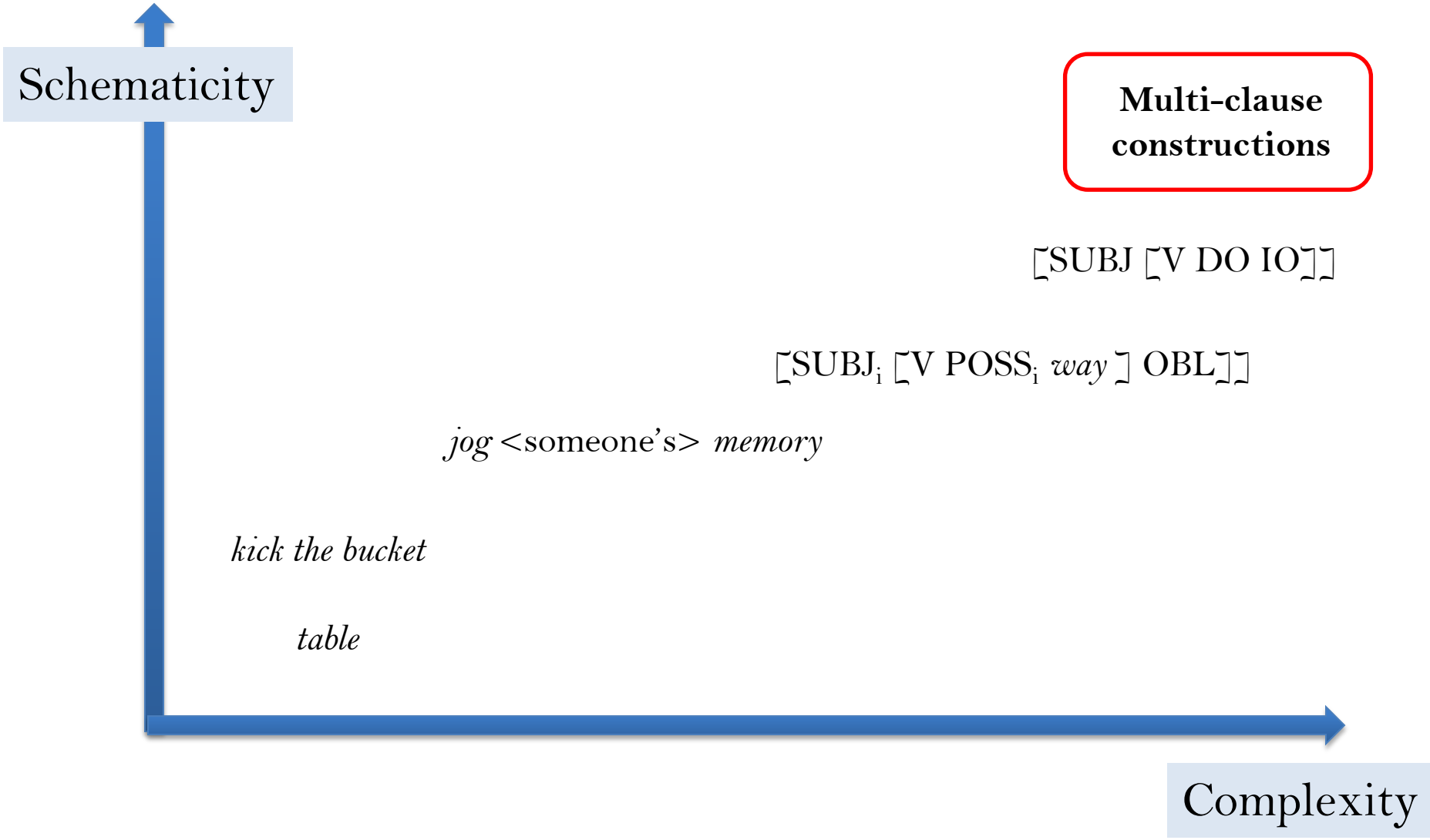


Grammatical system A:
English

Grammatical system B:
German

Circles represent the **set of all constructions** of a given **system:**

Constructions



Goal of this talk

- sketch a corpus-based exploratory methodology for a quantitative contrastive analysis at the level of complex & schematic constructions

- **detect patterns** (feature bundles) and
- **compare** their **usage frequency** across languages
  - Why? Because usage frequency influences…
    - strength of mental representation (entrenchment)
    - online processing demand
    - L2 learning & L2 errors (→ over- & underuse)

Example:
Contrasting English and German Relative Clause Constructions (RCC)

Two tasks

1

What is the space in which we look for patterns?
For any given phenomenon of interest, there are usually many potentially relevant contrasts.
How can we tell which ones are most interesting?
► **Identify variables that strongly distinguish language A and B**

2

► How can we **measure degrees of entrenchment of complex, schematic constructions** so that we can compare pattern-usage in different languages?

An example:

Contrasting English and German
Relative Clause Constructions

Procedure

I. Sample from comparable linguistic resources (corpora)
   - GERMAN: 250 tensed RC-constructions IDS corpus compilation via COSMAS II_{web}
   - ENGLISH: 250 tensed RC-constructions from written part of ICE-GB R2

II. Describe patterns: Code for a wide range of properties

III. Task 1: Identify VARIABLES that strongly distinguish language A from language B

IV. Task 2: Identify typical PATTERNS and compare their usage frequencies

Description of Relative Clause Constructions

Examples:

i. *Peter hates* ***everything*** $_{\text{REL}}[$ *that* ***John*** *likes* __ $]$.

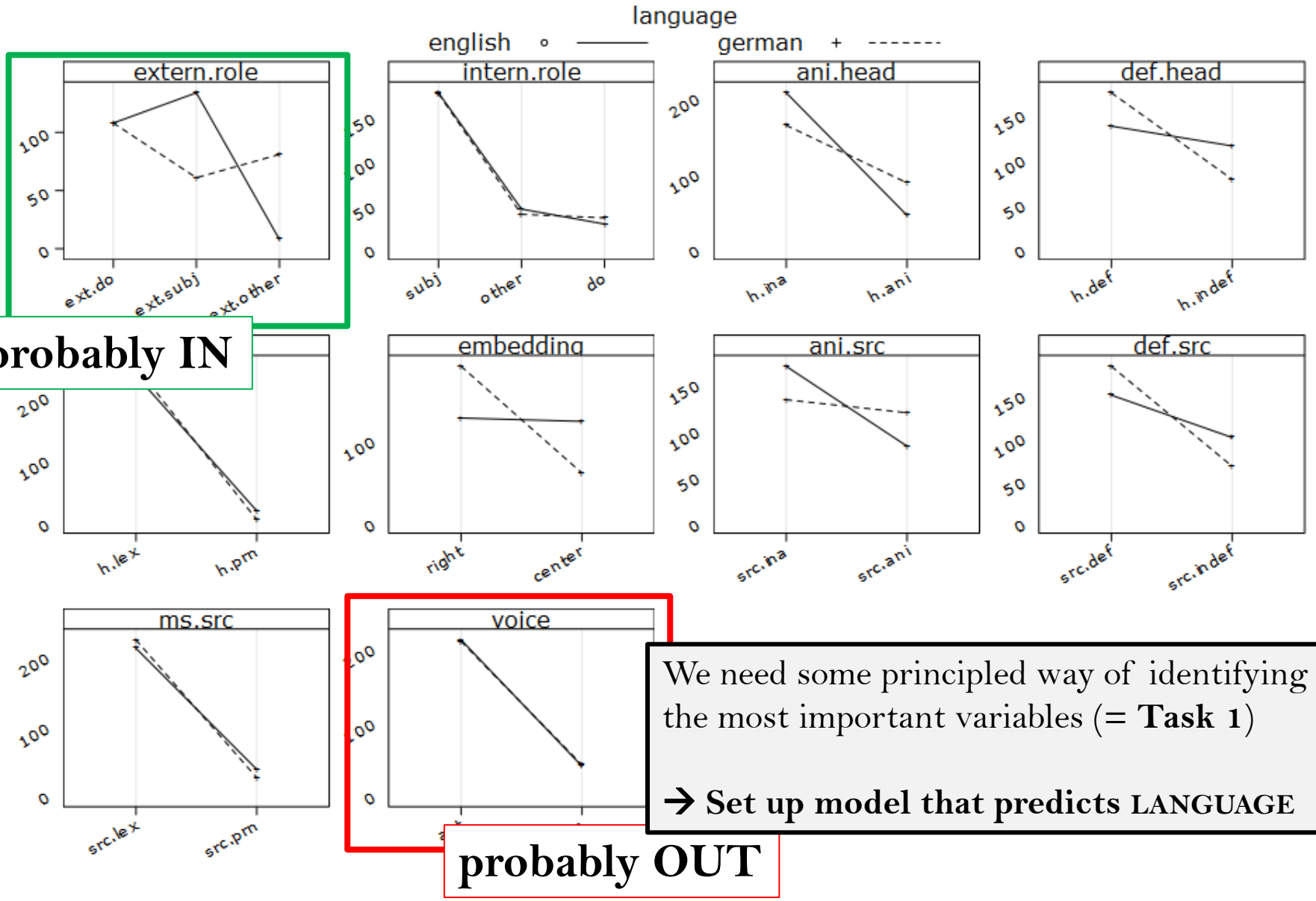
ii. *Peter hasst* ***alles*** , $_{\text{REL}}[$ *was* ***Johann*** *mag* __ $]$.

Potentially interesting **dimensions of contrast**

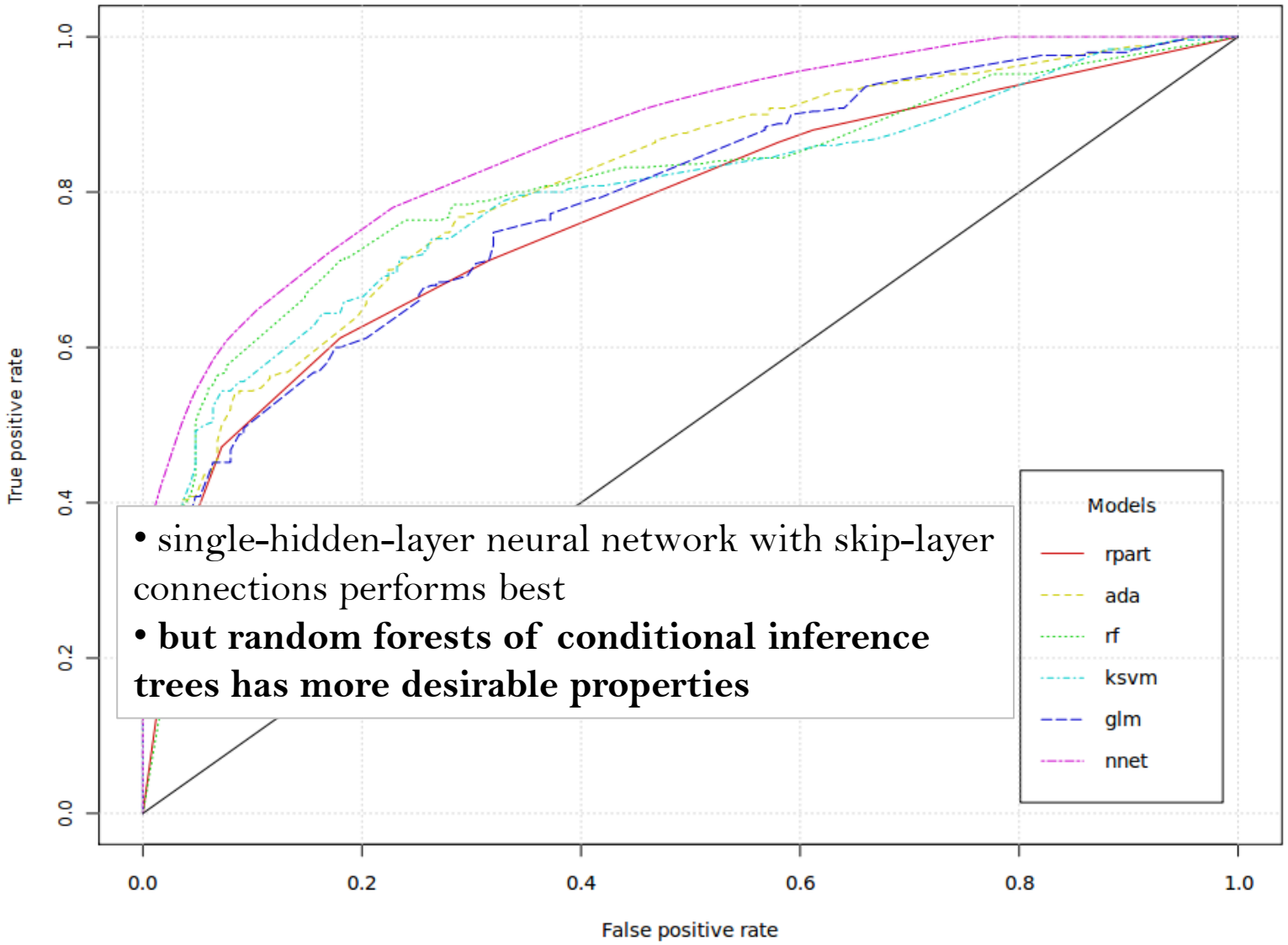
- **PROPERTIES OF THE RELATIVE CLAUSE**
  1. the internal grammatical role of head
  2. the external grammatical role of head
  3. the type of embedding of the RC
  4. the voice of the RC
- **PROPERTIES ENCODED ON THE HEAD NOUN**
  5. the animacy (of the referent) of the head
  6. the definiteness of the head
  7. the type of NP_{head}
- **PROPERTIES ENCODED ON THE SUBJECT OF RC**
  8. the animacy (of the referent) of the RC subject
  9. the definiteness of the RC subject
  10. the type of $NP_{RCsubject}$

Task 1

Identify **VARIABLES** that strongly distinguish language A from language B


language
english
german
extern.role
intern.role
ani.head
def.head
ext.do
ext.subj
ext.other
subj
other
do
h.ina
h.ani
h.def
h.indef
h.lex
h.prn
embedding
ani.src
def.src
right
center
src.ina
src.ani
src.def
src.indef
ms.src
voice
src.lex
src.prn
probably IN
probably OUT
We need some principled way of identifying the most important variables (= **Task 1**)
→ **Set up model that predicts LANGUAGE**

Predictive models:
6 model builders



- single-hidden-layer neural network with skip-layer connections performs best
- **but random forests of conditional inference trees has more desirable properties**

Predictive model: Method

- **Random forests of conditional inference trees***

Advantages:

- can **handle correlating predictors**
- conditional inference tree-based models do **not favor variables with many factor levels** (or large ranges)
- output of random forests is **more robust** than that of single-tree models

---

*

- Hothorn, Hornik and Zeileis (2006) on conditional inference framework
- Breiman (2001) on random forests
- Strobl, Boulesteix, Kneib, Augustin and Zeileis (2008) on permutation variable importance for random forests

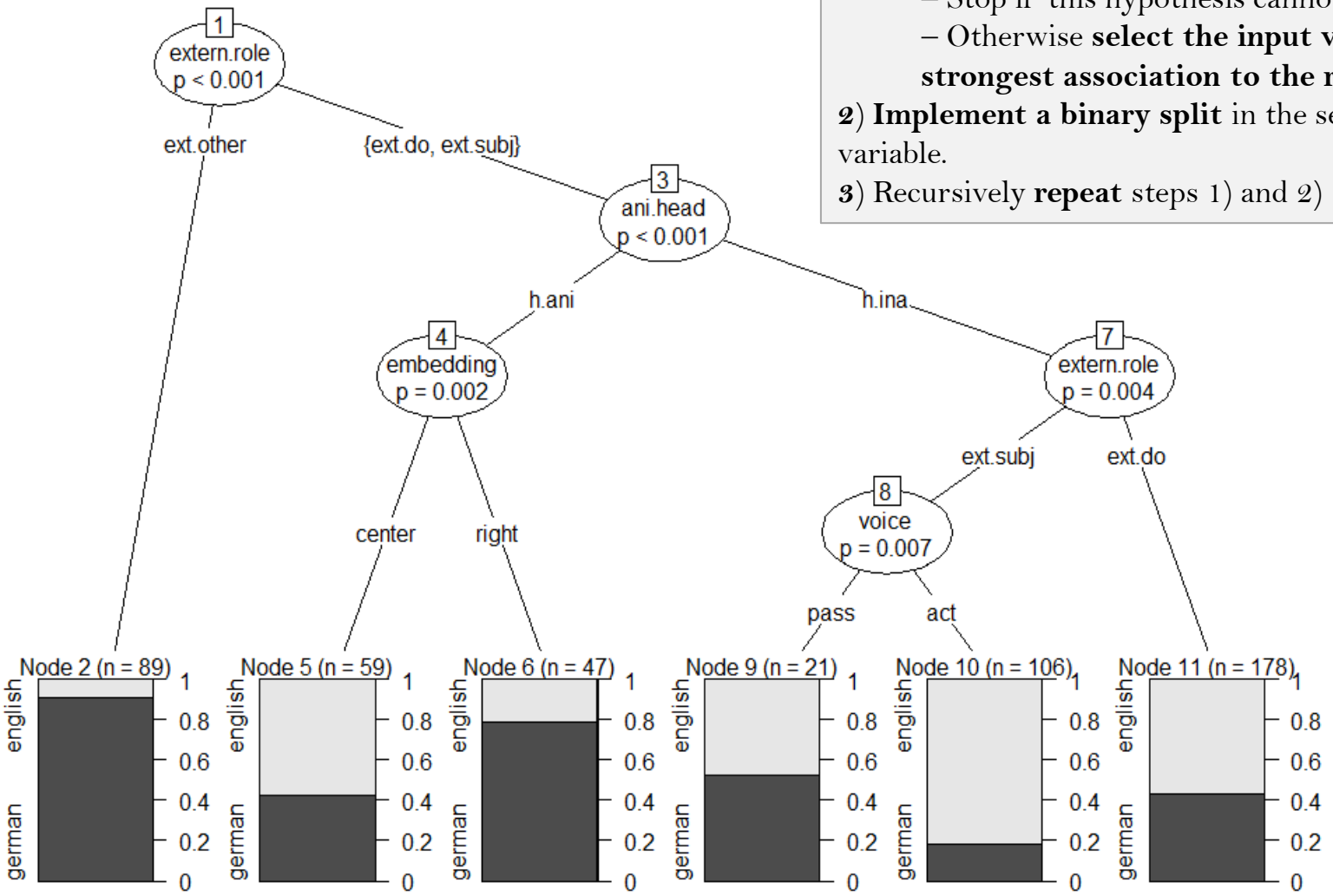
A conditional inference tree-based model (single tree)

Algorithm

1) **Test** the global null hypothesis of **independence** between any of the **input variables** and the **response**
   - Stop if this hypothesis cannot be rejected.
   - Otherwise **select the input variable with strongest association to the response**

2) **Implement a binary split** in the selected input variable.

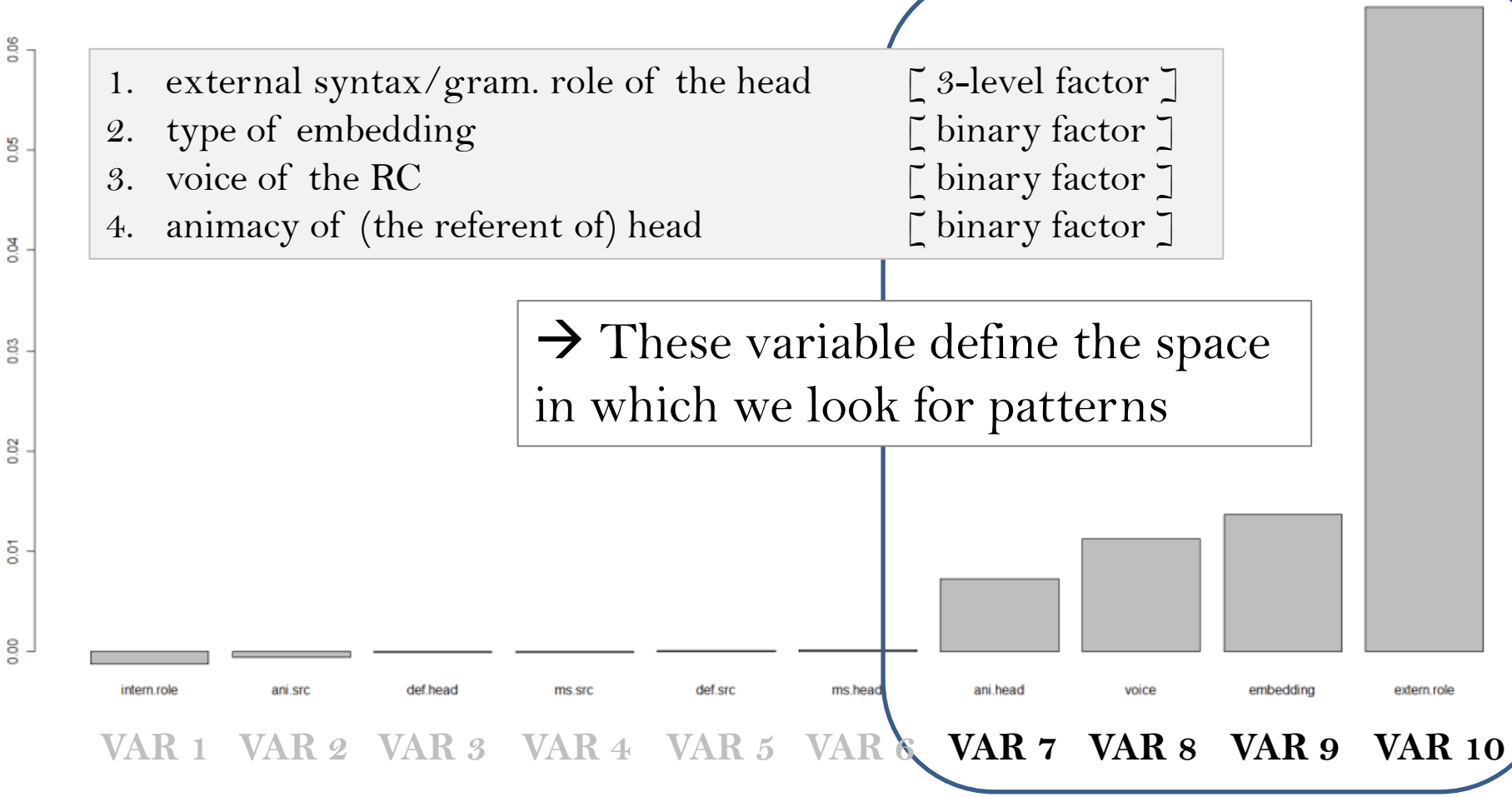
3) Recursively **repeat** steps 1) and 2)



Conditional permutation
Variable importance

1. external syntax/gram. role of the head [3-level factor]
2. type of embedding [binary factor]
3. voice of the RC [binary factor]
4. animacy of (the referent of) head [binary factor]

→ These variable define the space in which we look for patterns



Task 2

Identifying preferred patterns
and comparing their usage frequencies

Pattern identification: Method: Configural Frequency Analysis (CFA; von Eye, 1990)

- CFA evaluates cell-frequencies in an n-dimensional contingency table

- It compares the observed frequency of a configuration with the expected frequency of that configuration

  - configurations with observed frequency $>_{sig}$ expected frequency are called TYPES
  - configurations with observed frequency $<_{sig}$ expected frequency are called ANTI-TYPES

English relative clause constructions

Configuration (feature bundle)					Stats for configuration						
pattern	extern.role	ani.head	embedding	voice	Freq	Exp	Cont.chisq	Obs-exp	P.adj.bin	Dec	Q
c1	EXT.SUBJ	H.ANI	CENTER	ACT	24	7.8162	33.5093	>	9.40E-05	***	0.033
c2	EXT.SUBJ	H.ANI	CENTER	PASS	1	1.5333	0.1855	<	26.2304	ns	0.001
c3	EXT.SUBJ	H.ANI	RIGHT	ACT	6	13.5395	4.1984	<	0.84787	ns	0.015
c4	EXT.SUBJ	H.ANI	RIGHT	PASS	0	2.6561	2.6561	<	3.34697	ns	0.005
c5	EXT.SUBJ	H.INA	CENTER	ACT	19	22.0165	0.4133	<	14.391	ns	0.006
c6	EXT.SUBJ	H.INA	CENTER	PASS	11	4.319	10.3347	>	0.22939	ns	0.013
c7	EXT.SUBJ	H.INA	RIGHT	ACT	0	38.1379	38.1379	<	2.83E-16	***	0.083
c8	EXT.SUBJ	H.INA	RIGHT	PASS	0	7.4816	7.4816	<	0.02555	*	0.015
c9	EXT.OTHER	H.ANI	CENTER	ACT	1	3.5674	1.8477	<	6.14461	ns	0.005
c10	EXT.OTHER	H.ANI	CENTER	PASS	1	0.6998	0.1288	>	24.1713	ns	0.001
c11	EXT.OTHER	H.ANI	RIGHT	ACT	20	6.1795	30.9097	>	0.00031	***	0.028
c12	EXT.OTHER	H.ANI	RIGHT	PASS	3	1.2123	2.6362	>	5.90267	ns	0.004
c13	EXT.OTHER	H.INA	CENTER	ACT	3	10.0486	4.9442	<	0.4542	ns	0.014
c14	EXT.OTHER	H.INA	CENTER	PASS	0	1.9712	1.9712	<	6.65958	ns	0.004
c15	EXT.OTHER	H.INA	RIGHT	ACT	39	17.4065	26.7877	>	0.00017	***	0.045
c16	EXT.OTHER	H.INA	RIGHT	PASS	14	3.4147	32.8136	>	0.00062	***	0.021
c17	EXT.DO	H.ANI	CENTER	ACT	0	8.6579	8.6579	<	0.00773	**	0.018
c18	EXT.DO	H.ANI	CENTER	PASS	0	1.6984	1.6984	<	8.75715	ns	0.003
c19	EXT.DO	H.ANI	RIGHT	ACT	30	14.9976	15.0072	>	0.0159	*	0.031
c20	EXT.DO	H.ANI	RIGHT	PASS	1	2.9421	1.282	<	9.93962	ns	0.004
c21	EXT.DO	H.INA	CENTER	ACT	0	24.3875	24.3875	<	6.65E-10	***	0.051
c22	EXT.DO	H.INA	CENTER	PASS	0	4.7842	4.7842	<	0.3922	ns	0.01
c23	EXT.DO	H.INA	RIGHT	ACT	66	42.245	13.3578	>	0.01096	*	0.052
c24	EXT.DO	H.INA	RIGHT	PASS	11	8.2873	0.888	>	10.1805	ns	0.006

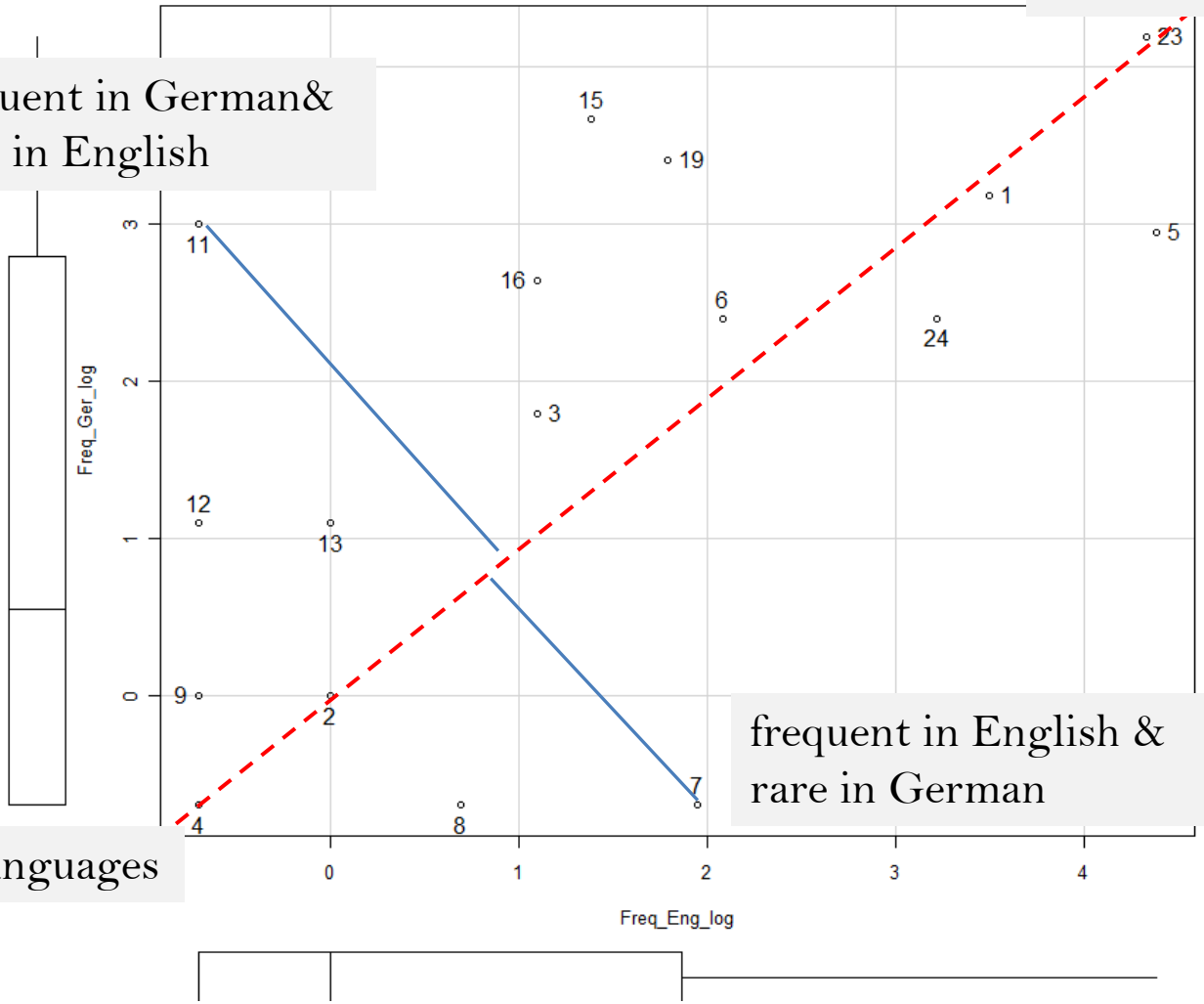
German relative clause constructions

Configuration (feature bundle)					Stats for configuration						
pattern	**extern.role**	**ani.head**	**embedding**	**voice**	**Freq**	**Exp**	**Cont.chisq**	**Obs-exp**	**P.adj.bin**	**Dec**	**Q**
c1	EXT.SUBJ	H.ANI	CENTER	ACT	33	7.8162	81.1422	>	4.50E-10	***	0.051
c2	EXT.SUBJ	H.ANI	CENTER	PASS	1	1.5333	0.1855	<	26.2304	ns	0.001
c3	EXT.SUBJ	H.ANI	RIGHT	ACT	3	13.5395	8.2042	<	0.02929	*	0.022
c4	EXT.SUBJ	H.ANI	RIGHT	PASS	0	2.6561	2.6561	<	3.34697	ns	0.005
c5	EXT.SUBJ	H.INA	CENTER	ACT	80	22.0165	152.7076	>	1.77E-21	***	0.121
c6	EXT.SUBJ	H.INA	CENTER	PASS	8	4.319	3.1372	>	3.43542	ns	0.007
c7	EXT.SUBJ	H.INA	RIGHT	ACT	7	38.1379	25.4227	<	1.32E-08	***	0.067
c8	EXT.SUBJ	H.INA	RIGHT	PASS	2	7.4816	4.0162	<	0.95527	ns	0.011
c9	EXT.OTHER	H.ANI	CENTER	ACT	0	3.5674	3.5674	<	1.33782	ns	0.007
c10	EXT.OTHER	H.ANI	CENTER	PASS	0	0.6998	0.6998	<	23.8287	ns	0.001
c11	EXT.OTHER	H.ANI	RIGHT	ACT	0	6.1795	6.1795	<	0.09567	ms	0.013
c12	EXT.OTHER	H.ANI	RIGHT	PASS	0	1.2123	1.2123	<	14.2602	ns	0.002
c13	EXT.OTHER	H.INA	CENTER	ACT	1	10.0486	8.1481	<	0.02109	*	0.018
c14	EXT.OTHER	H.INA	CENTER	PASS	0	1.9712	1.9712	<	6.65958	ns	0.004
c15	EXT.OTHER	H.INA	RIGHT	ACT	4	17.4065	10.3257	<	0.00534	**	0.028
c16	EXT.OTHER	H.INA	RIGHT	PASS	3	3.4147	0.0504	<	26.6324	ns	0.001
c17	EXT.DO	H.ANI	CENTER	ACT	0	8.6579	8.6579	<	0.00773	**	0.018
c18	EXT.DO	H.ANI	CENTER	PASS	0	1.6984	1.6984	<	8.75715	ns	0.003
c19	EXT.DO	H.ANI	RIGHT	ACT	6	14.9976	5.398	<	0.33613	ns	0.019
c20	EXT.DO	H.ANI	RIGHT	PASS	1	2.9421	1.282	<	9.93962	ns	0.004
c21	EXT.DO	H.INA	CENTER	ACT	0	24.3875	24.3875	<	6.65E-10	***	0.051
c22	EXT.DO	H.INA	CENTER	PASS	0	4.7842	4.7842	<	0.3922	ns	0.01
c23	EXT.DO	H.INA	RIGHT	ACT	76	42.245	26.9712	>	2.52E-05	***	0.074
c24	EXT.DO	H.INA	RIGHT	PASS	25	8.2873	33.7039	>	7.85E-05	***	0.034

Pattern usage across languages

highly frequent in both languages

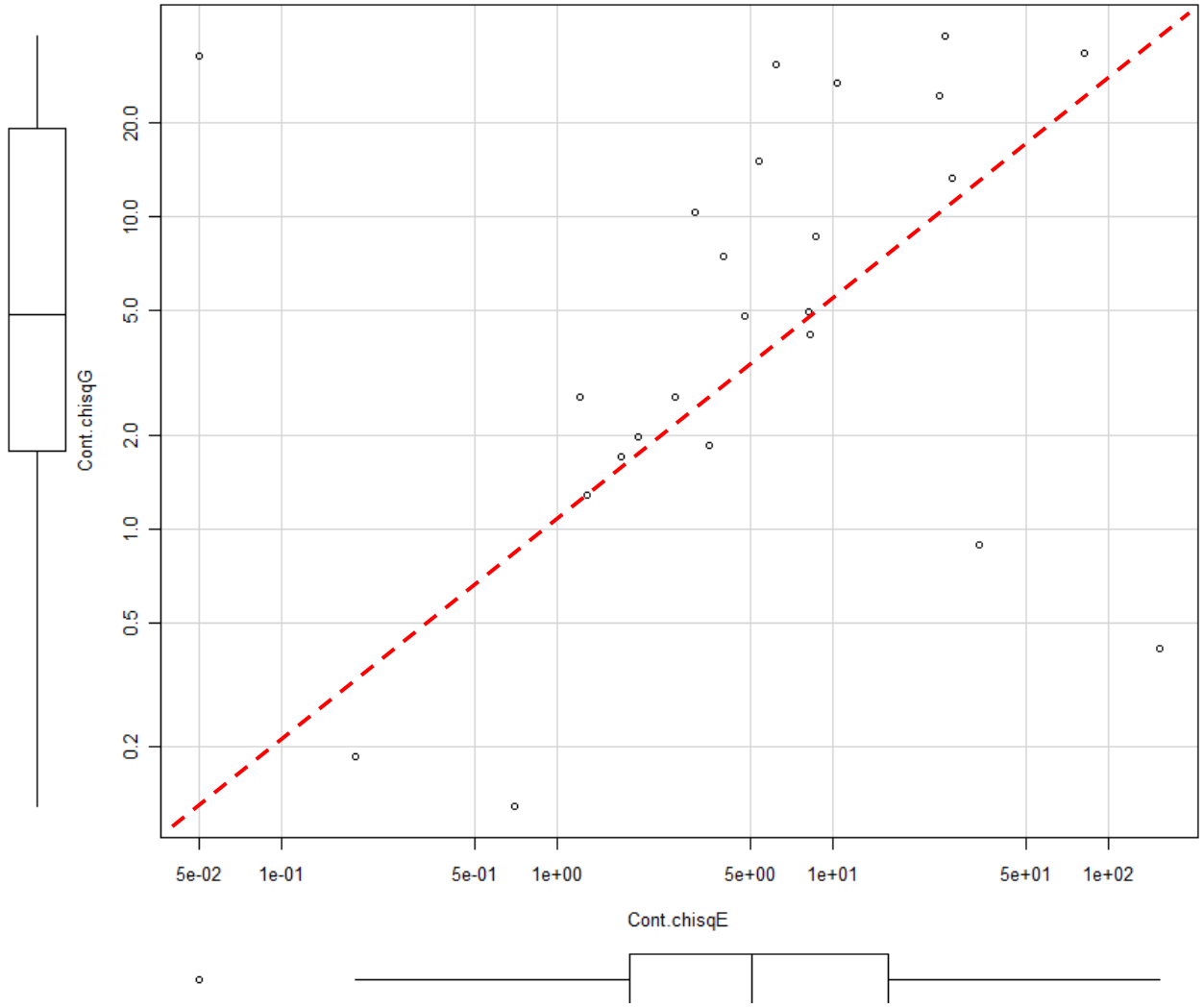
frequent in German& rare in English



frequent in English & rare in German

rare in both languages

Contribution to χ2 instead of **frequency**



Summing up

What?

The **goal** of the talk was to sketch a corpus-based **exploratory methodology** for the **quantitative contrastive analysis** of (genetically related) languages

- Step 1: Identify **variables** that strongly **distinguish language A from language B**
- Step 2: Identify **preferred patterns in each language** and **comparison of usage frequencies**

Why?

Comparing usage frequencies of patterns allows us to...

- make predictions about ***relative processing demand*** of a pattern in given L
- make predictions ***of L2 learner errors*** (over- & underuse)
- Investigate difference in ***form-function mapping*** across languages
- ....

References

von Eye, A. 1990. *Introduction to Configural Frequency Analysis: The search for types and antitypes in cross-classifications*. Cambridge, UK: Cambridge University Press.

Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1), 5–32.

Hothorn, Torsten, Kurt Hornik and Achim Zeileis (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, **15** (3), 651-674.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin and Achim Zeileis (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, **9**, 307.

Thank you very much