

Corpus statistics: key issues and controversies

Panel Discussion

The panellists

- Vaclav Brezina (Lancaster)
 - Stefan Evert (Erlangen-Nürnberg)
 - Stefan Th. Gries (Santa Barbara)
 - Andrew Hardie (Lancaster)
 - Jefrey Lijffijt (Bristol)
 - Gerold Schneider (Zürich/Konstanz)
 - Sean Wallis (London)
- Session chair:** MC Paul Rayson (Lancaster)

Topics

1. Experimental design:
Which factors should we measure?
2. Non-randomness, dispersion and
the assumptions of hypothesis tests
3. Teaching and curricula
4. Visualisation
5. Which models can we use?

Sean Wallis, University College London

1 EXPERIMENTAL DESIGN

Are we getting statistics right?

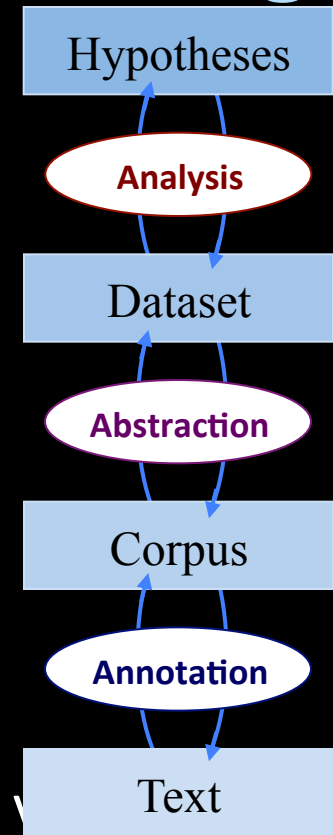
- In a single recent top CL journal volume:
 - 1 article employed a method using a per-word baseline
 - No argument as to why this was optimal
 - 1 article cited a statistical test without specifying the baseline (expected distribution)
 - Implication: constant proportion of words
 - 2 articles quoted naïve frequencies or probabilities without any inferential statistical evaluation
 - Arguably one of these was justified in doing so

The centrality of experimental design

- Experimental design is **central**:
 - Clarification of testable hypotheses
 - Abstraction / operationalisation
 - Map corpus events to regular dataset
 - Frequently necessary to reformulate
 - The experimental model determines:
 - instances of phenomena to capture
 - how to express aspects of phenomena as variables
 - the appropriate statistical model

The centrality of experimental design

- Experimental design is **central**:
 - Clarification of testable hypotheses
 - Abstraction / operationalisation
 - Map corpus events to regular dataset
 - Frequently necessary to reformulate
 - The experimental model determines:
 - instances of phenomena to capture
 - how to express aspects of phenomena as variables
 - the appropriate statistical model



CLAIM: experimental design > statistical method

- Researchers usually focus on selecting their **research object** (e.g. noun phrase)
 - Often unsure about the **baseline** to use
 - Often just use ‘words’ (per 1,000 or 1,000,000)
 - Mistaken concept of ‘normalisation’
- Baselines determine **meaning** and **comparability** of results
 - Most statistical methods also depend on assumptions that the item is free to vary

Research questions and baselines

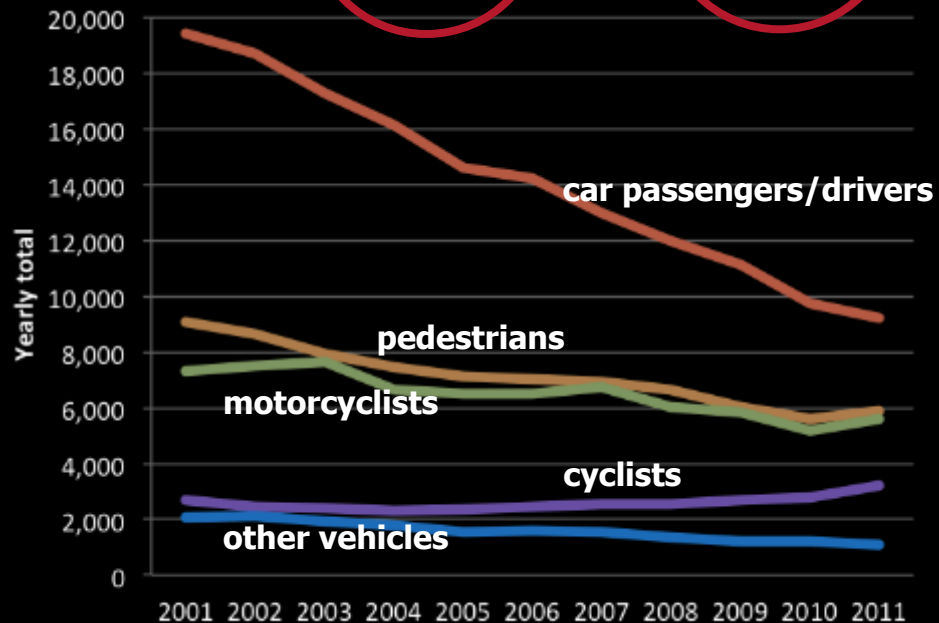
- Suppose you are told that
 - cycling is getting safer
- Do you believe them?
 - would **you** start cycling?
- Facts
 - fatalities have **increased**
- What is the most **meaningful** statistic?
 - $p(\text{accident} \mid \text{population})$ – $p(\text{accident} \mid \text{cyclist})$
 - $p(\text{accident} \mid \text{journey})$ – $p(\text{accident} \mid \text{km})$



See e.g. <http://cyclinginfo.co.uk/blog/323/cycling/how-dangerous-is-cycling>

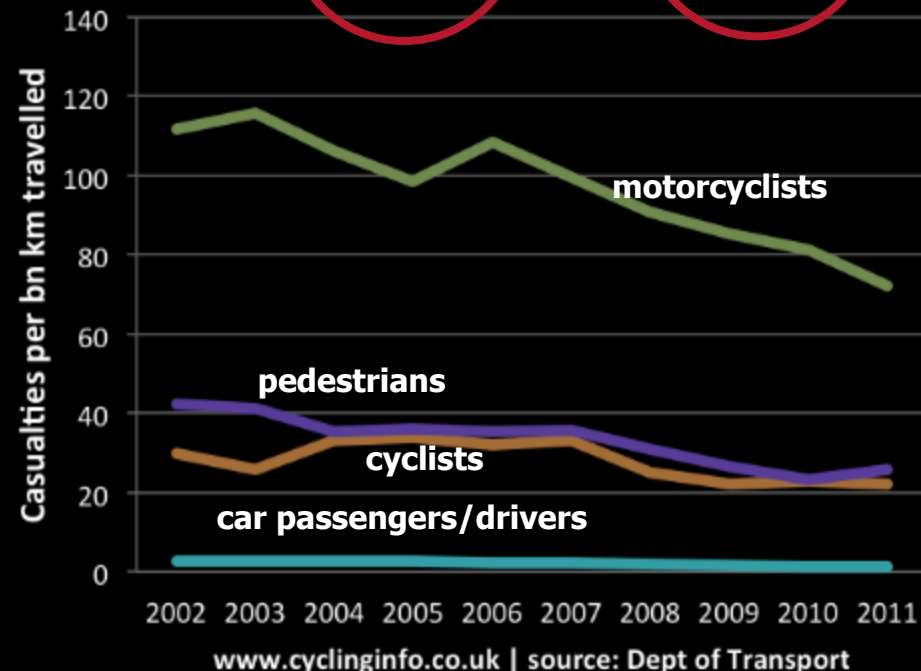
Research questions and baselines

- Suppose you are told that
 - cycling is getting safer
- Do you believe them?
 - would **you** start cycling?
- Facts
 - fatalities **increased**
 - are there **more cyclists** now?
 - **BUT...**



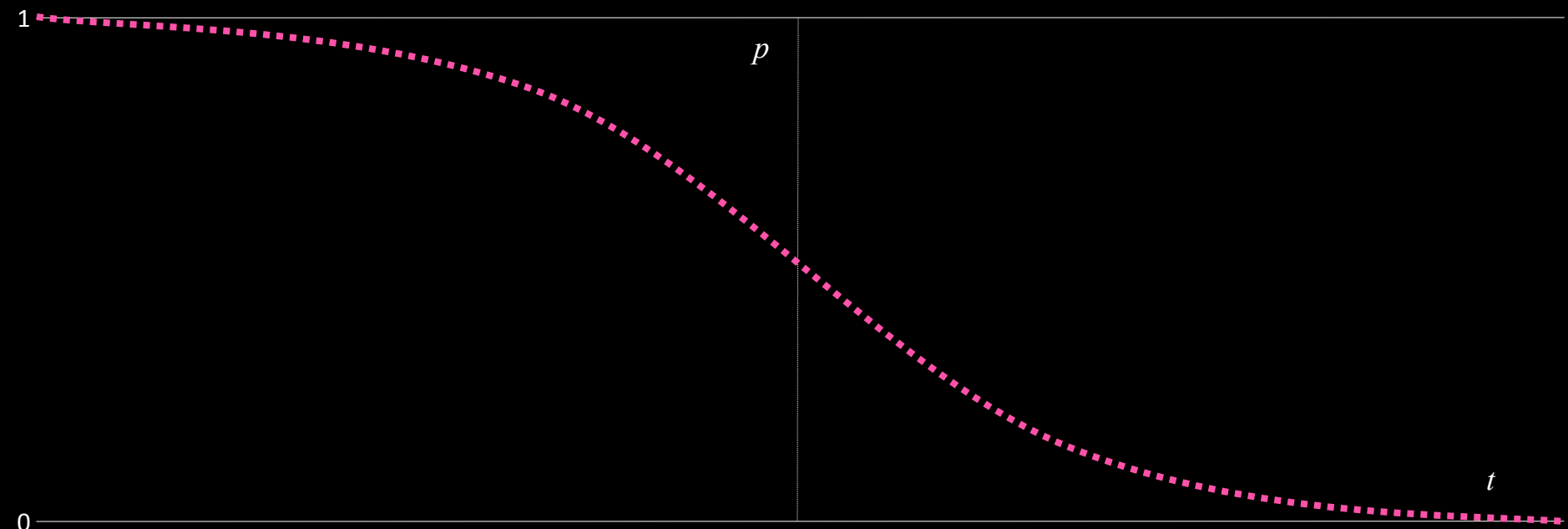
Research questions and baselines

- Suppose you are told that
 - cycling is getting safer
- Do you believe them?
 - would **you** start cycling?
- Facts
 - fatalities **increased**
 - there *are* **more cyclists** now
 - **BUT...** death rates per km have **fallen**



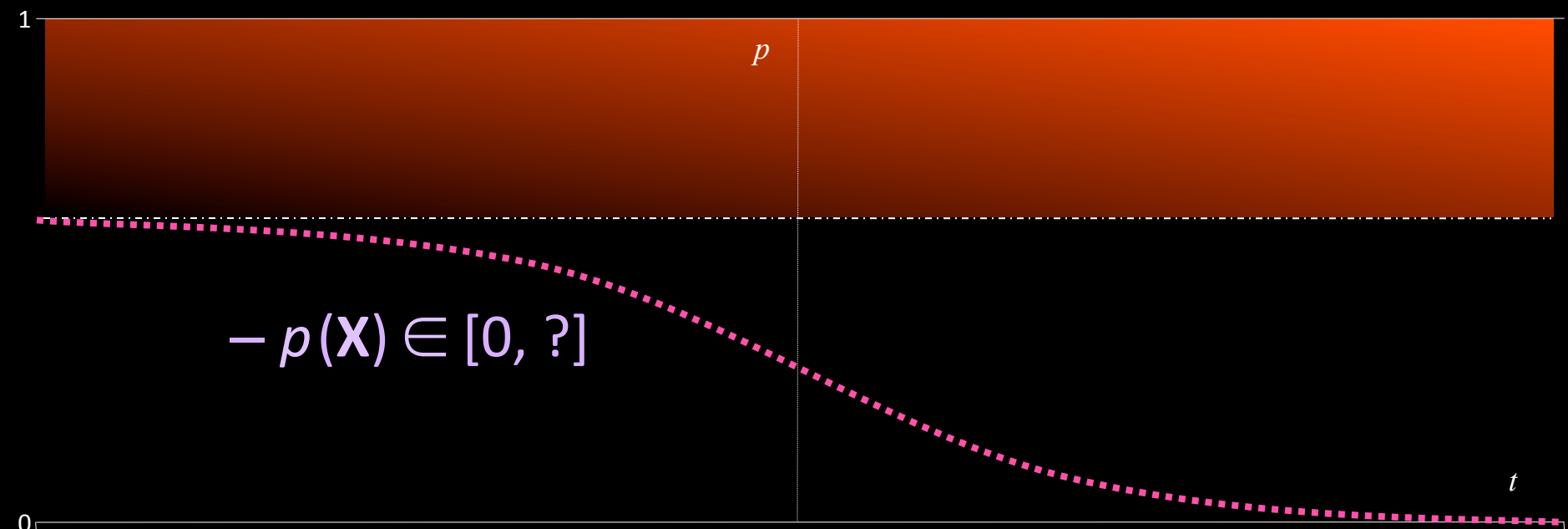
Baselines alter statistical models

- Logistic ‘S’ curve assumes **freedom to vary**
 - $p(\mathbf{X}) \in [0, 1]$



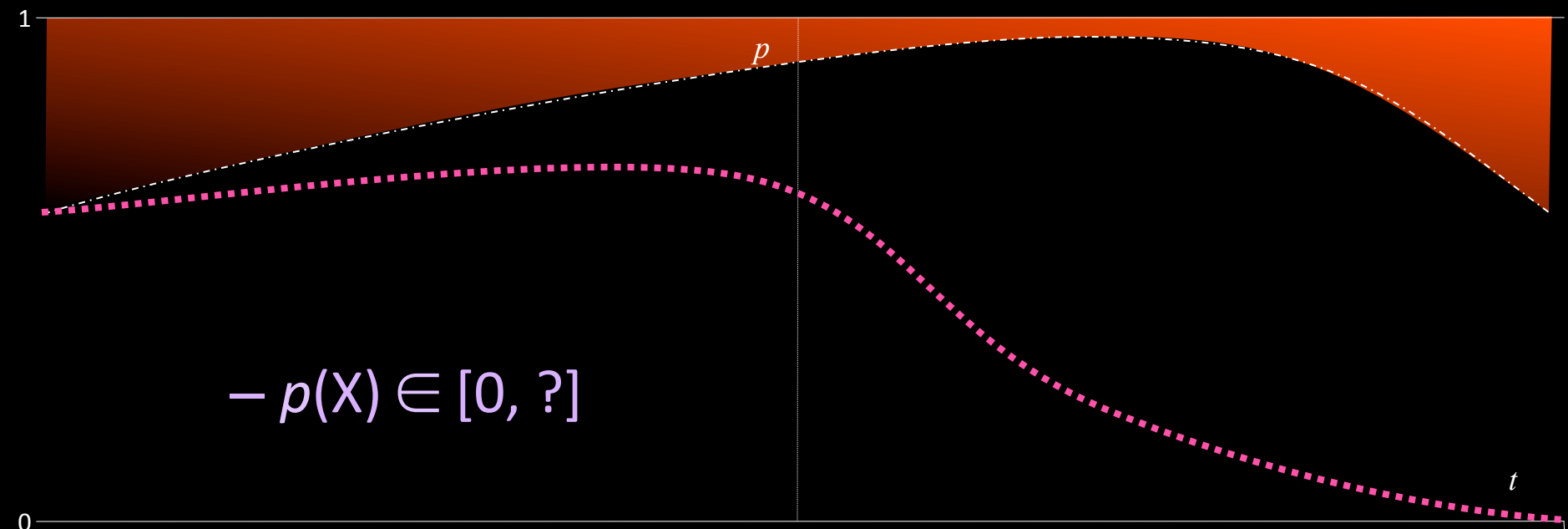
Baselines alter statistical models

- Logistic ‘S’ curve assumes **freedom to vary**
 - what happens if that freedom is **limited**?



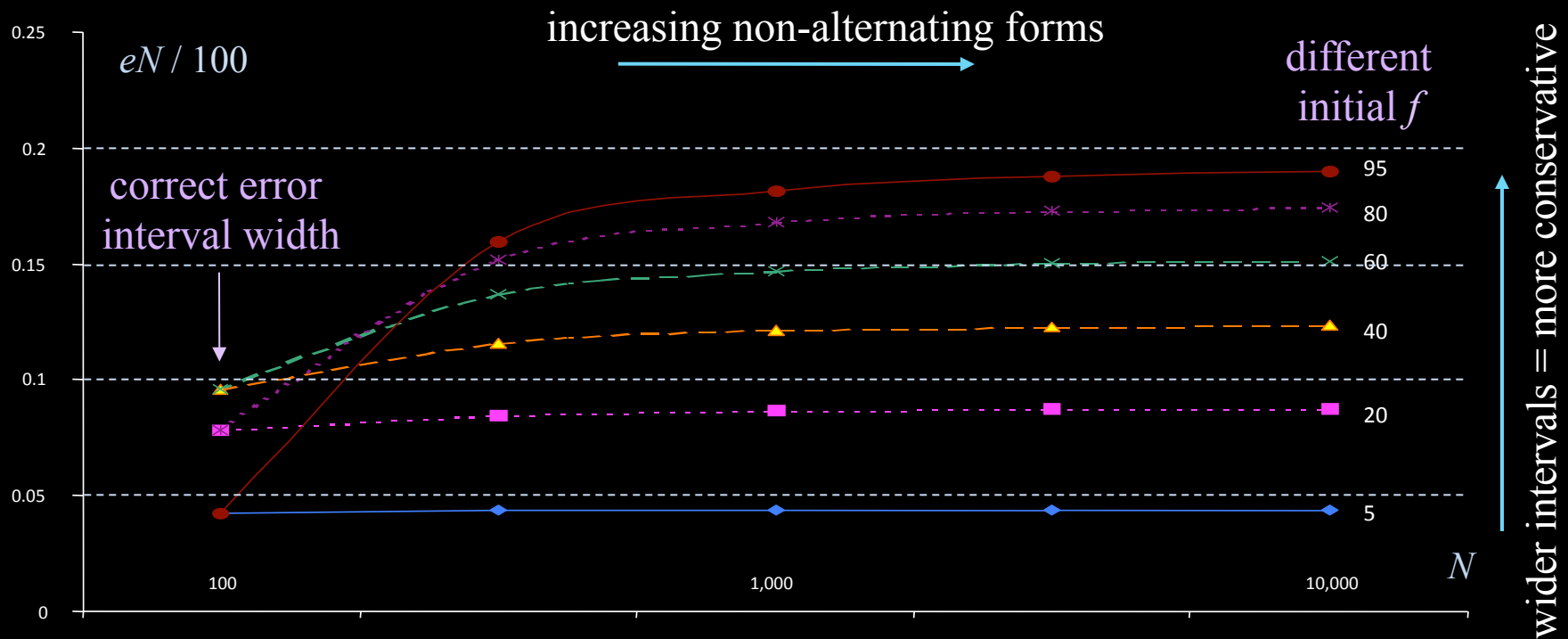
Baselines alter statistical models

- Logistic ‘S’ curve assumes **freedom to vary**
 - what happens if that freedom is limited?
 - or the opportunity to use a construction also varies?



Baselines alter statistical models

- Statistical models assume data is free to vary
 - Add large numbers of invariant terms to the dataset and methods become more **conservative**



Discussion

1. Experimental design:
Which factors should we measure?
2. Non-randomness, dispersion and the assumptions of hypothesis tests
3. Teaching and curricula
4. Visualisation
5. Which models can we use?

Jefrey Lijffijt, University of Bristol

2 NON-RANDOMNESS

The problem

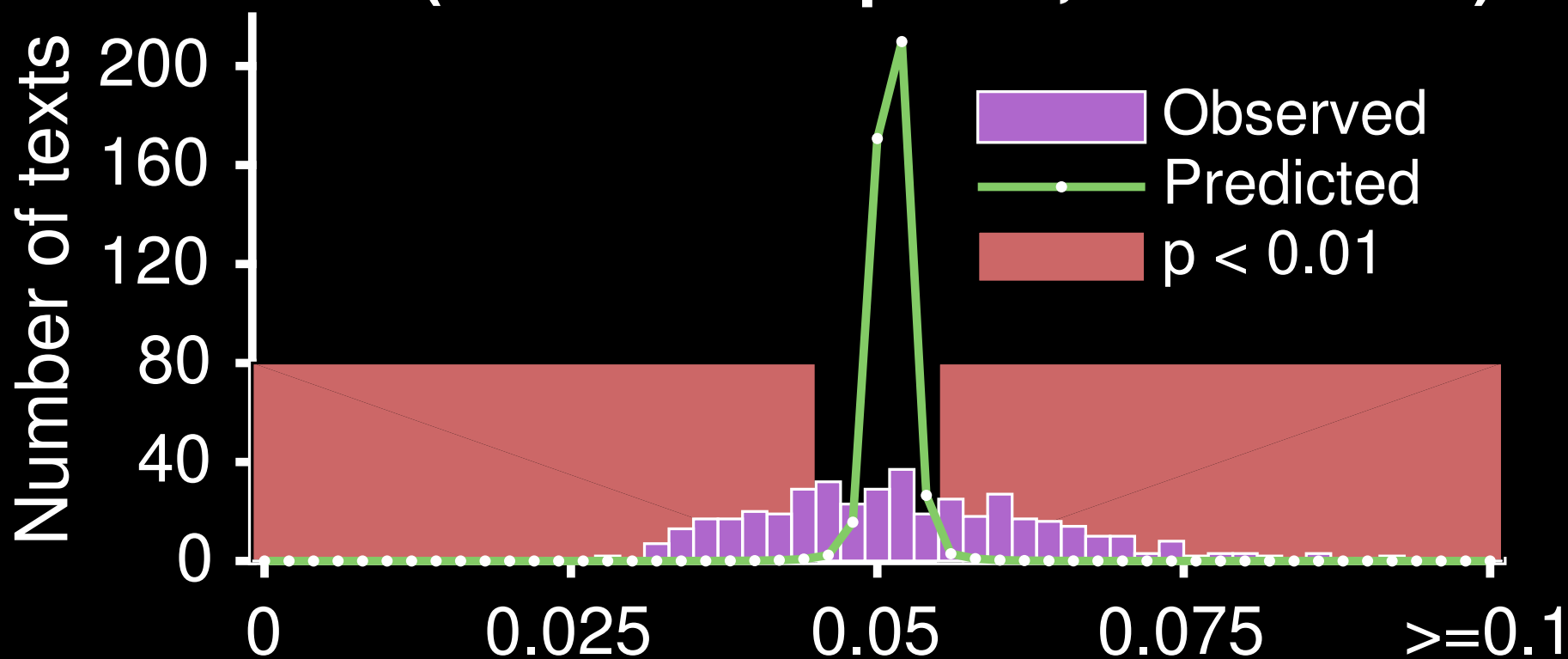
- Statistical tests/models are *always* based on assumptions
- What if the assumptions are false?
- How would you know?
- What to do?

An example problem

- χ^2 , log-likelihood ratio / G, Fisher Exact test, etc., assume **independence of all counts**
 - **Expectation of variance** over texts
(binomial distribution)
- Unless samples contain at most one instance, such as extremely short texts (tweets),
this expectation is *always* wrong
(Church, COLING 2000, Evert, ZAA 2006, Lijffijt *et al.*, DSH 2014)

Overly simple example

'the' (BNC fiction prose, $\sigma = 820573$)



Statistics vs. the truth

- ‘Language is never, ever, ever, random’ (Kilgarrieff, CLLT 2005)
 - [These models are very far from the truth
→ you failed to model the ‘true’ variation]
- **Why to model text as random process**
 - Corpus is sample of texts (= true randomness)
 - Complex structure (= remaining variation)

Tests and assumptions

- $p = \Pr(T \geq x)$
 - Probability that the test statistic is the same or higher in random data
 - This assumes a **stochastic model** for the r.v. T
- χ^2 , log-likelihood ratio / G, Fisher Exact test assume independence of *every instance*

	Y = true	Y = false
X = true	R	S
X = false	T	U

Why the problem matters

- If the assumptions are false, *p-values can be too high or too low, to any degree*
- Conjecture: *p-values* derived under invalid assumptions do not add any value
- The assumptions underlying a statistical test have to be correct

However

- Some tests require invalid assumptions
–/–> statistical testing is an ill choice
- Often, there are alternatives
 1. Manipulate the representation (adjusted counts)
 2. Select only appropriate data (use dispersion)
 3. Use another test
(*t*-test, anova are almost always fine)

The open questions

- Often, there are alternatives
 1. Manipulate the representation (adjusted counts)
 2. Select only appropriate data (use dispersion)
 3. Use another test
- What approach to prefer?
- What if it is not clear how to do any of the above?

Stefan Evert, Friedrich-Alexander-Universität Erlangen-Nürnberg

A SECOND OPINION

Three views of corpus studies

- Topic 1: controlled experiment
 - is there a significant difference btw conditions?
- Topic 2: **observational study**
 - inference about property of population
 - problem of non-randomness (\neq random text!)
- Topic 5: predictive model
 - which factors affect linguistic behaviour?

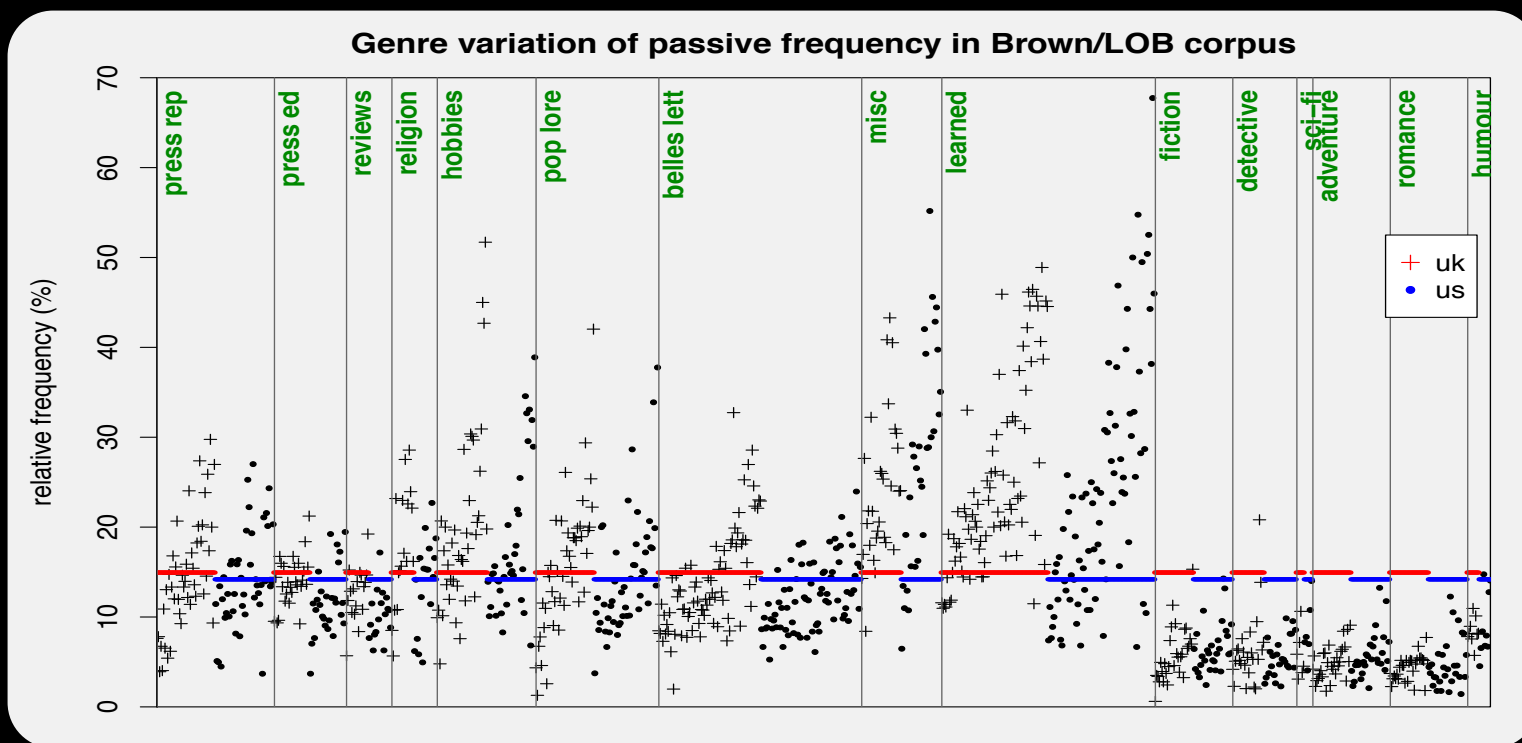
Methodological questions

- Do corpus + statistical analysis accurately reflect the underlying population?
 - statistics: yes, if corpus = truly random sample
- What property do we want to measure?
 - and is it the one we're actually measuring?

☞ population parameter vs. sample statistic

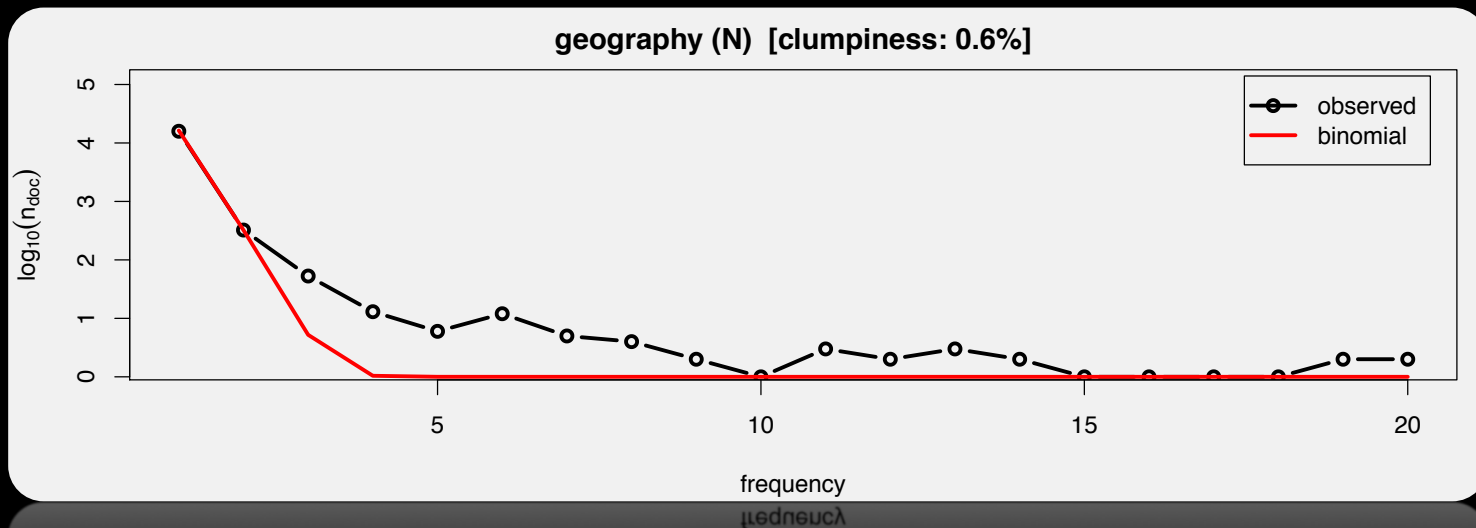
Example 1: frequency comparison

- Passive VPs more frequent in BrE than AmE
 - 13.3% vs. 12.6% → significant?
 - chi-squared: **yes!**; t-test: **no!**; GLM: **yes!!**



Example 2: burstiness

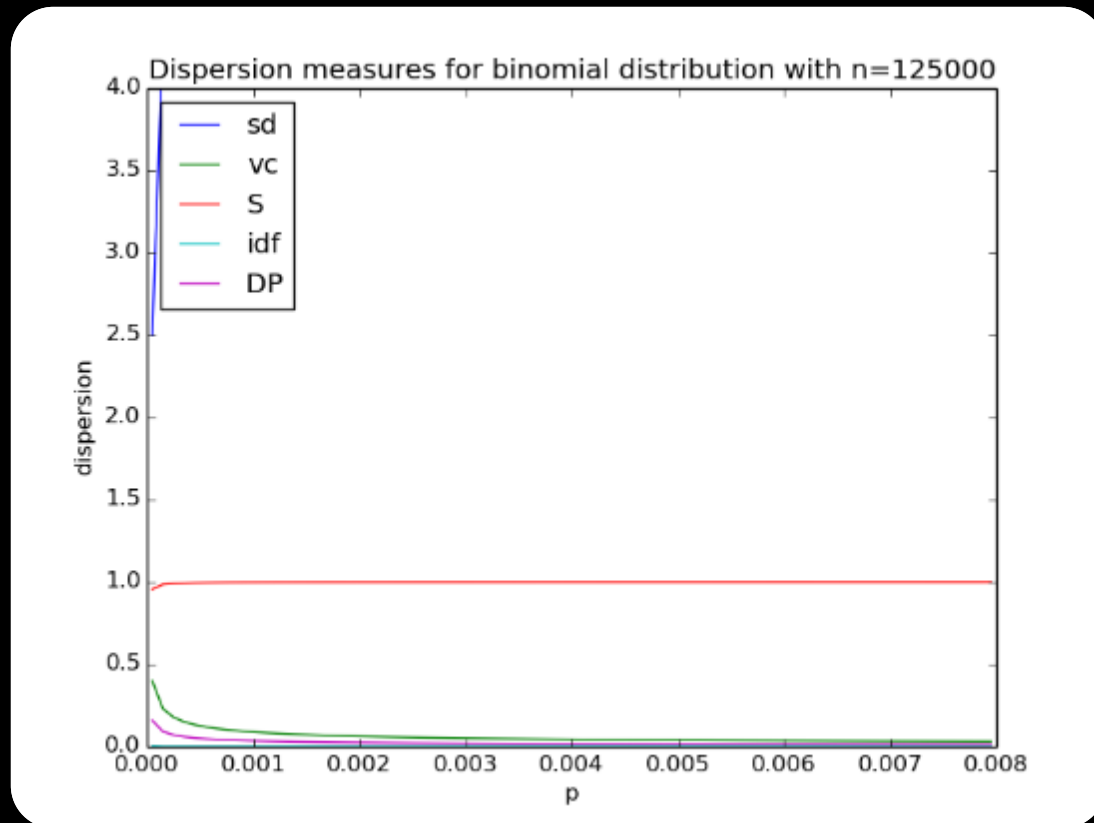
- Content words tend to occur in “bursts”



- $P(f = 1) = \alpha(1 - \gamma)$ (Katz 1996)
- $P(f = 2) = \alpha\gamma / (1 - p)$
- $P(f = k) = \alpha\gamma \times p^{k-2} / (1 - p)$ for $k \geq 3$
- Which of α , γ , p is “frequency”?

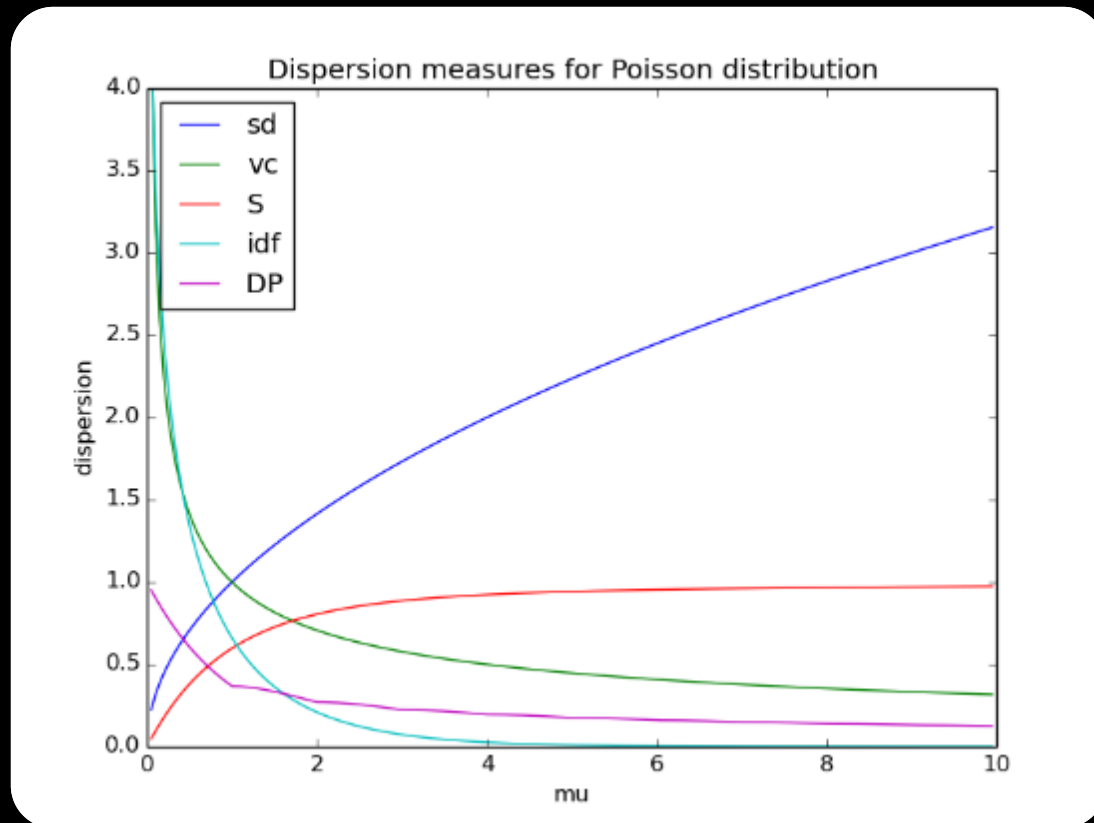
Example 3: dispersion

- Many dispersion measures (e.g. Gries 2008)
- Clear: **binomial sample = perfect dispersion**



Example 3: dispersion

- Many dispersion measures (e.g. Gries 2008)
- Clear: **binomial sample = perfect dispersion**



Discussion

1. Experimental design:
Which factors should we measure?
2. Non-randomness, dispersion and
the assumptions of hypothesis tests
3. Teaching and curricula
4. Visualisation
5. Which models can we use?

Vaclav Brezina, Lancaster University

3 TEACHING & CURRICULA

Teaching and curricula

View 1:

1. CL is a quantitative discipline.
2. Efficient quantification *requires* detailed knowledge of statistics.


Hence: **CL requires detailed knowledge of statistics.**


Teaching and curricula (cont.)


View 2 (loose syllogism):

1. CL combines linguistics and quantitative (statistical) methods.
2. Corpus *linguists* primarily specialise in understanding linguistic processes.

Hence: **It's good to have an expert statistician on the team.**

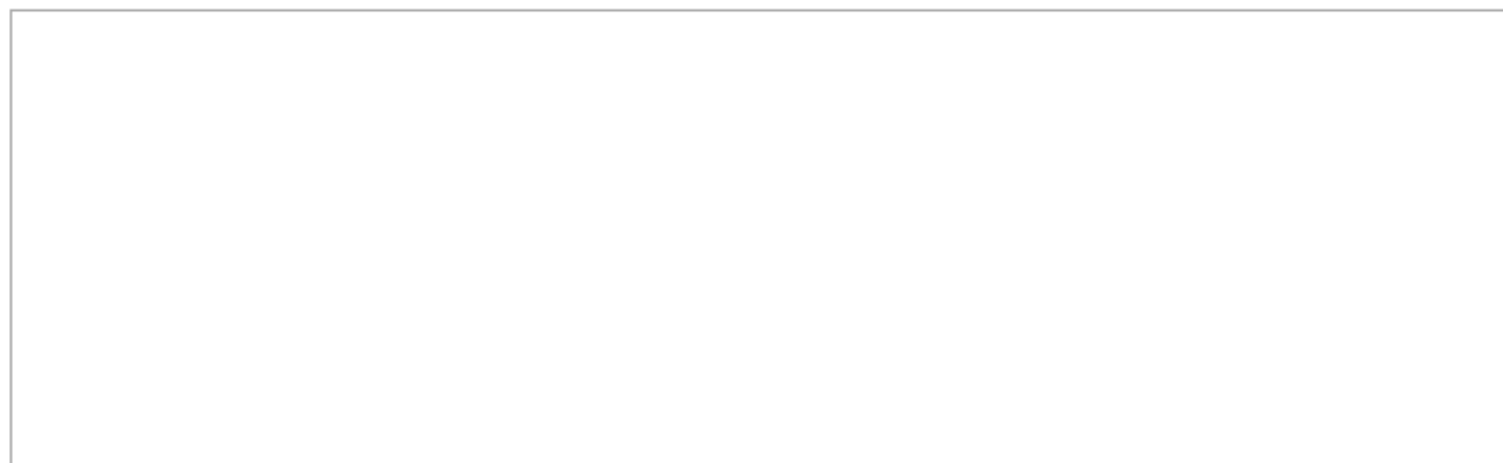
Stats calculator 

Graph tool 

Randomizer 

UCREL tools

1. Paste tab delimited data including header row and id column. For help



2. Select parameters.

- ☐ One linguistic variable ☐ Multiple linguistic variables (relationship)
- ☐ Description ☐ Prediction

Create graph

Clear

Discussion

1. Experimental design:
Which factors should we measure?
2. Non-randomness, dispersion and
the assumptions of hypothesis tests
3. Teaching and curricula
4. Visualisation
5. Which models can we use?

Stefan Th. Gries, University of California, Santa Barbara

4 VISUALISATION

On why we need to visualize

```
> summary(model.01)
```

```
Call:
glm(formula = ORDER ~ CONJ * LENGTH_DIFF, family = binomial,
    data = CLAUSE.ORDERS)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8803	-0.4005	-0.3954	0.8620	2.2867

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.463366	0.056357	8.222	< 2e-16	***
CONJbevor/before	-0.819629	0.098658	-8.308	< 2e-16	***
CONJnachdem/after	-0.064951	0.085287	-0.762	0.446	
CONJweil/because	-2.968349	0.088017	-33.725	< 2e-16	***
LENGTH_DIFF	0.111711	0.009568	11.675	< 2e-16	***
CONJbevor/before:LENGTH_DIFF	-0.147896	0.015960	-9.267	< 2e-16	***
CONJnachdem/after:LENGTH_DIFF	-0.069803	0.013054	-5.347	8.93e-08	***
CONJweil/because:LENGTH_DIFF	-0.109476	0.013802	-7.932	2.15e-15	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

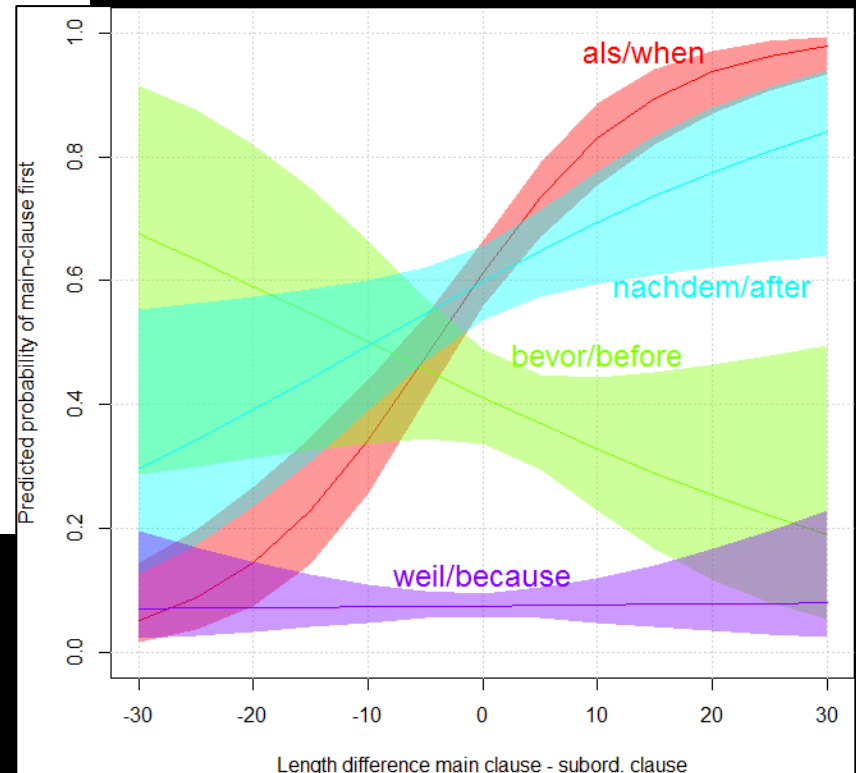
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 8060.8 on 6447 degrees of freedom
Residual deviance: 5897.4 on 6440 degrees of freedom
AIC: 5913.4

Number of Fisher scoring iterations: 5

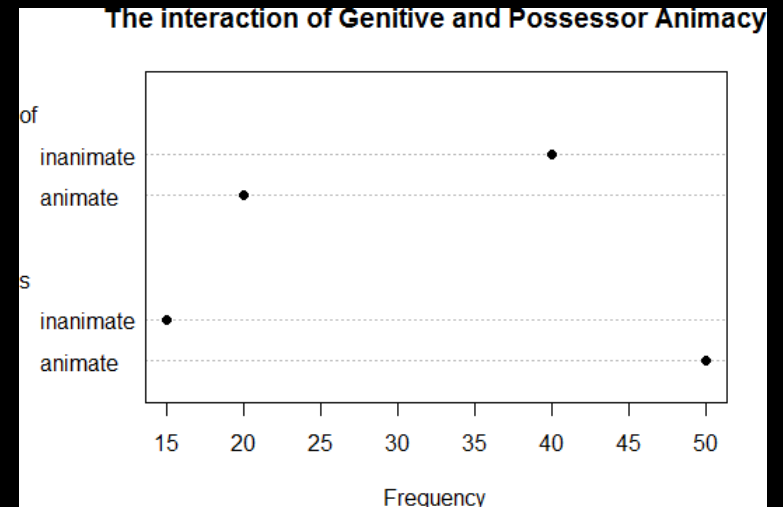
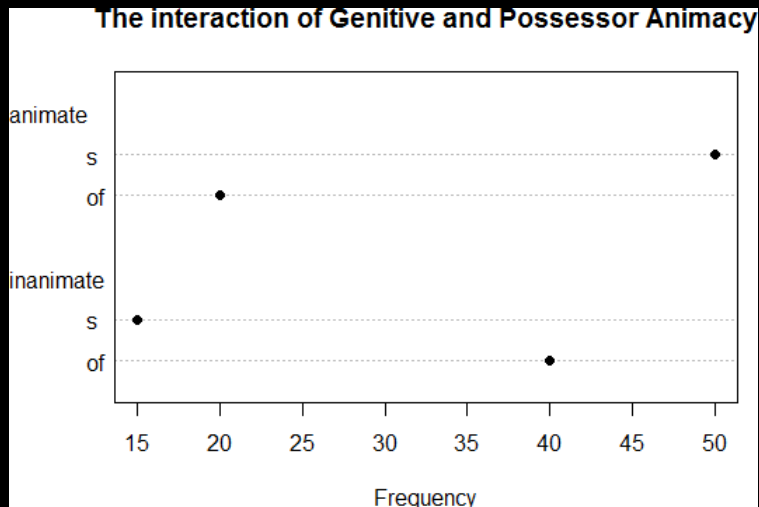
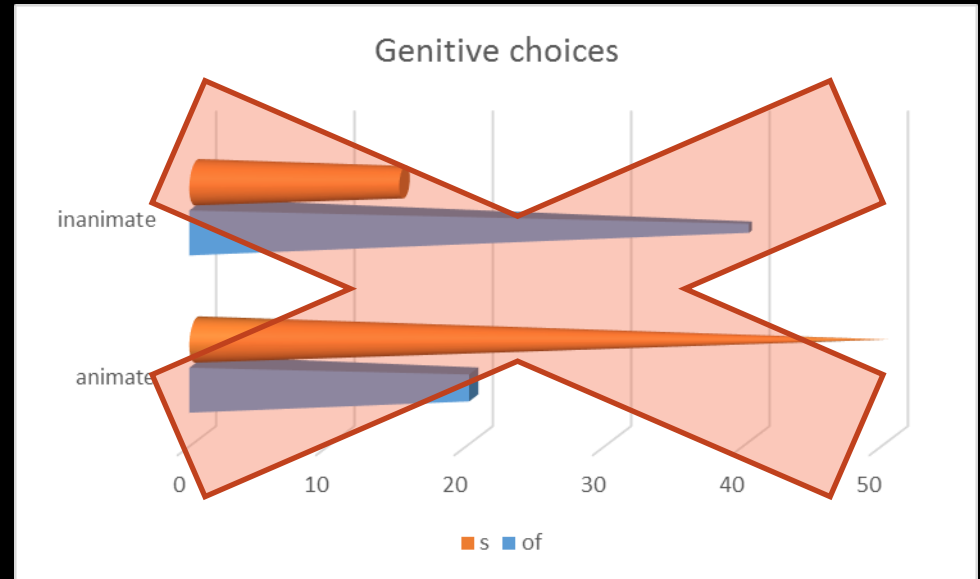
>

>



On ink-to-information ratio

	animate	inanimate
of	20	40
s	50	15



On perspectives and uncertainty

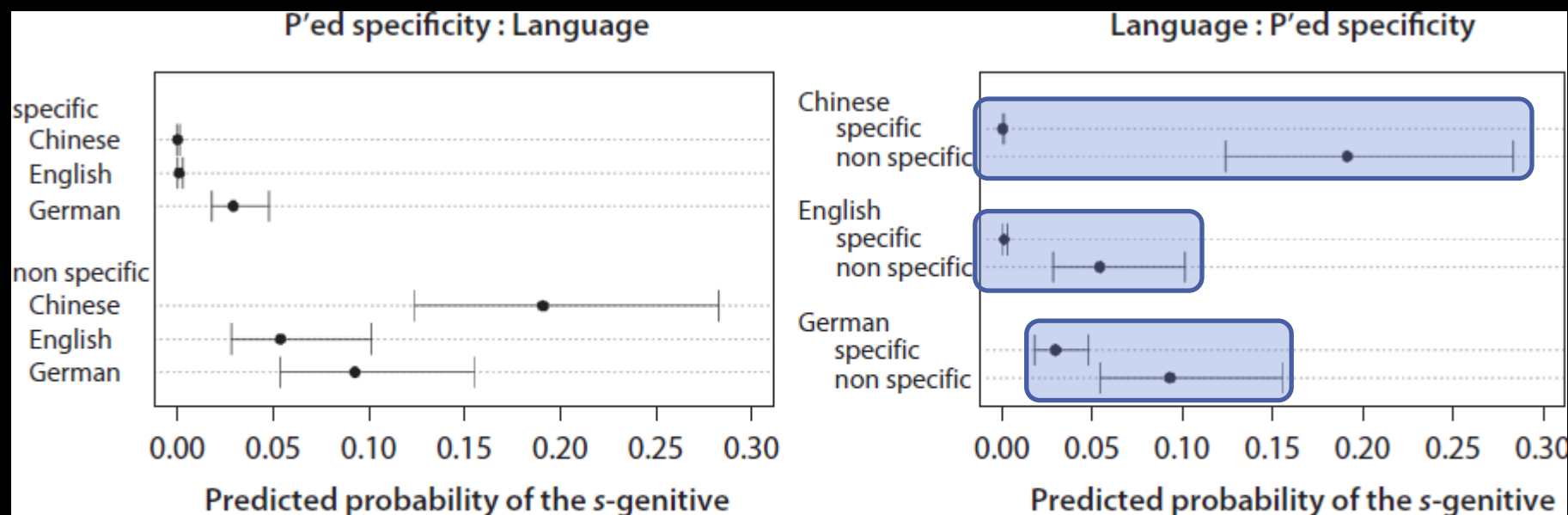
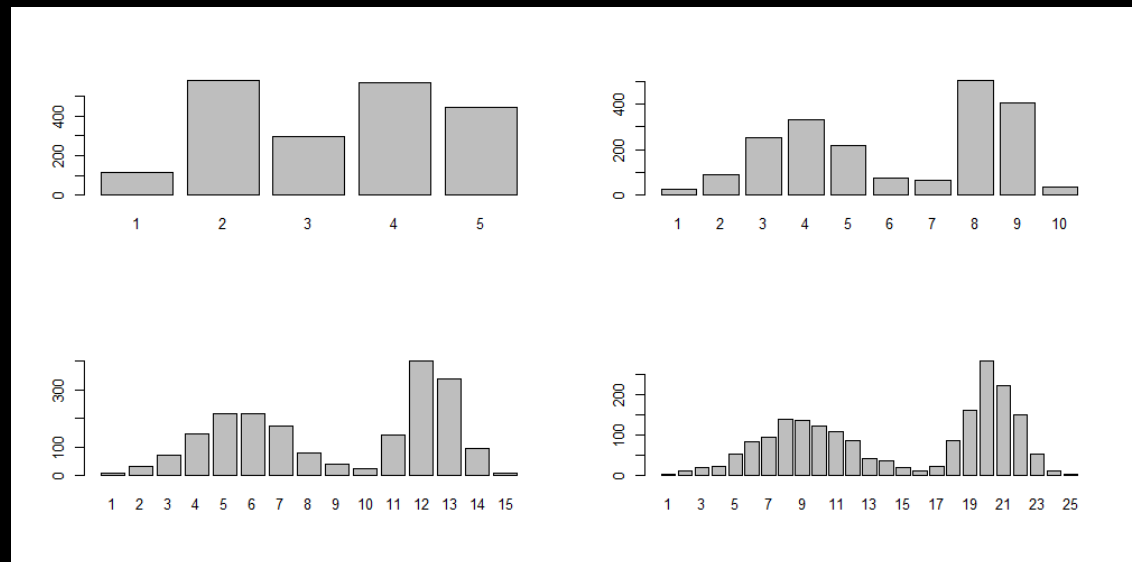
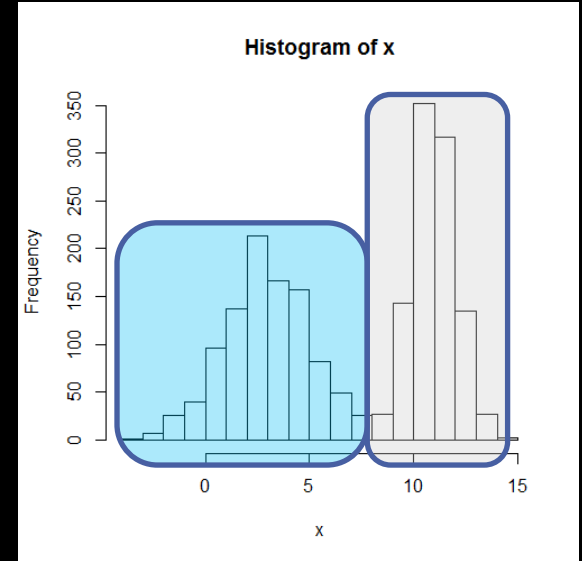
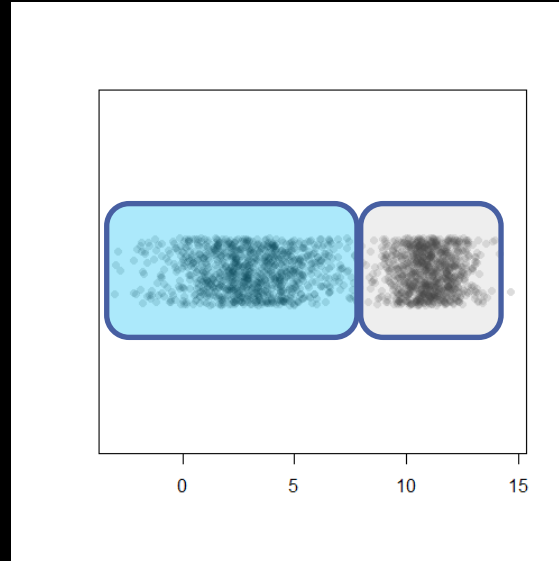
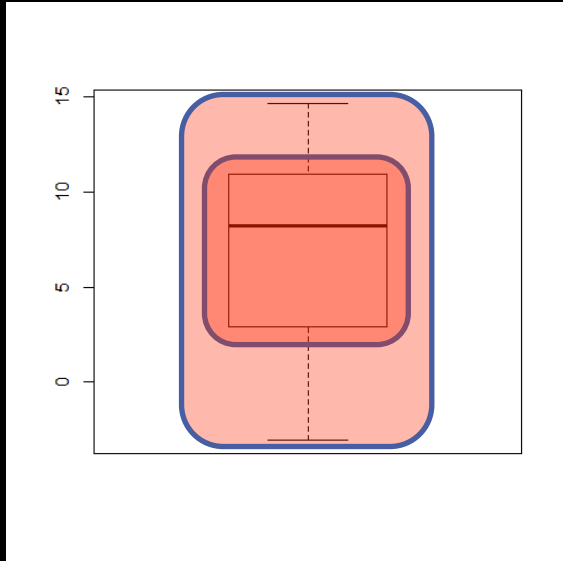
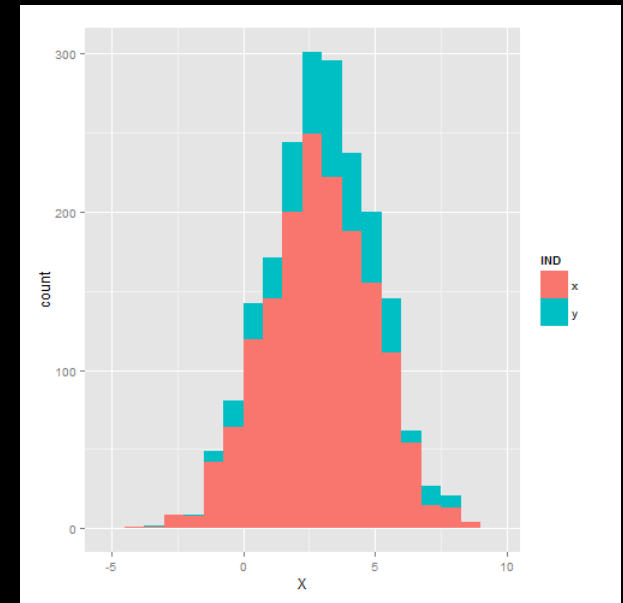
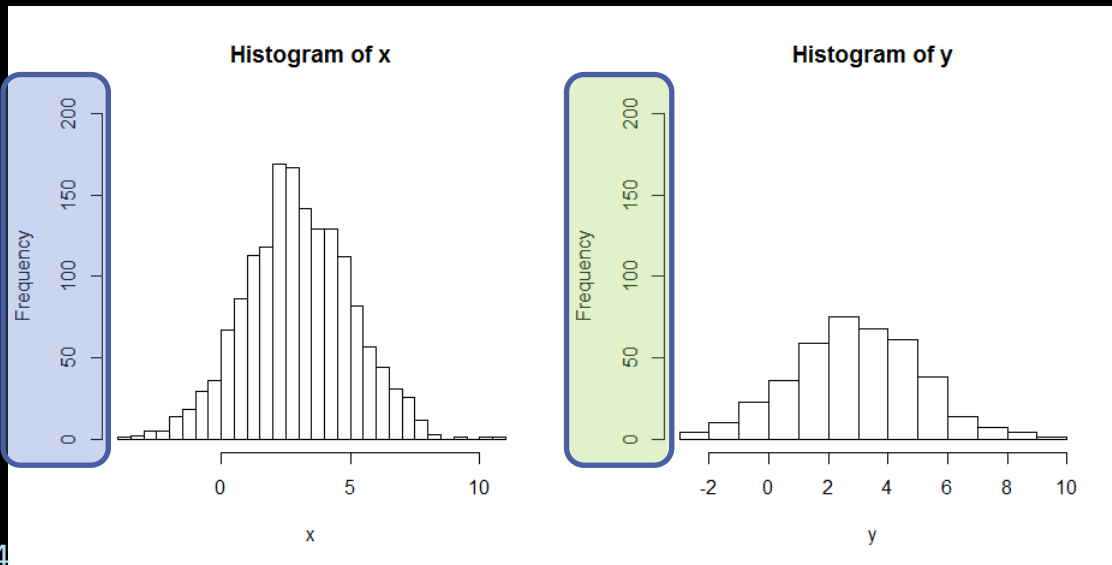
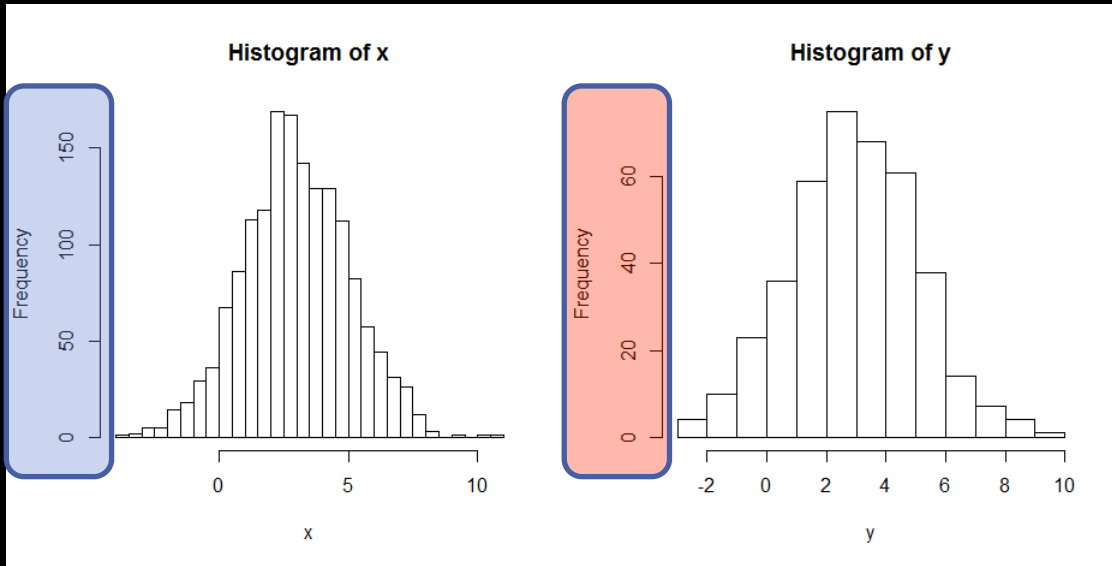


Figure 12. The interaction POSSEDSPEC : LANG

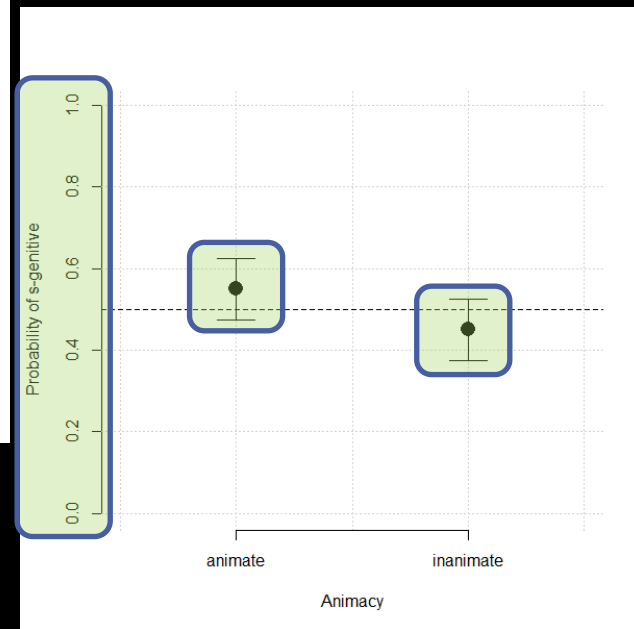
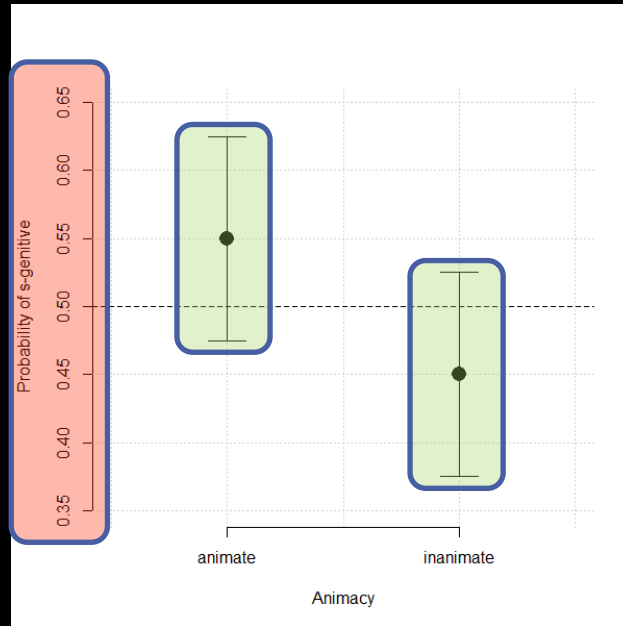
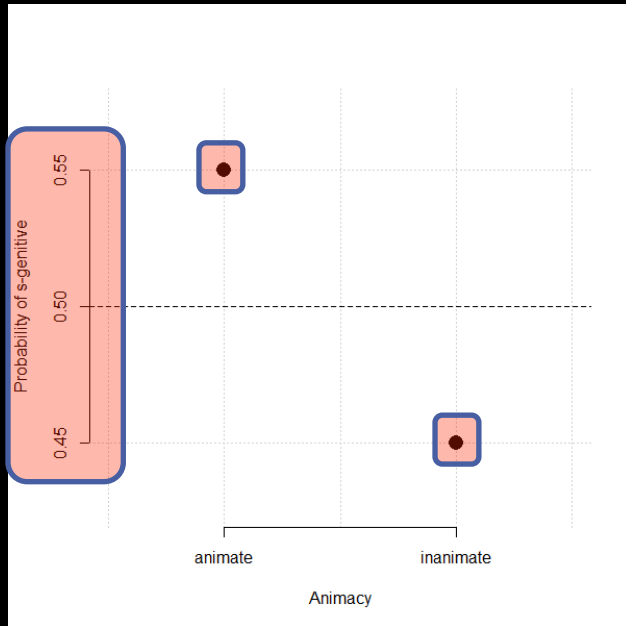
On granularity



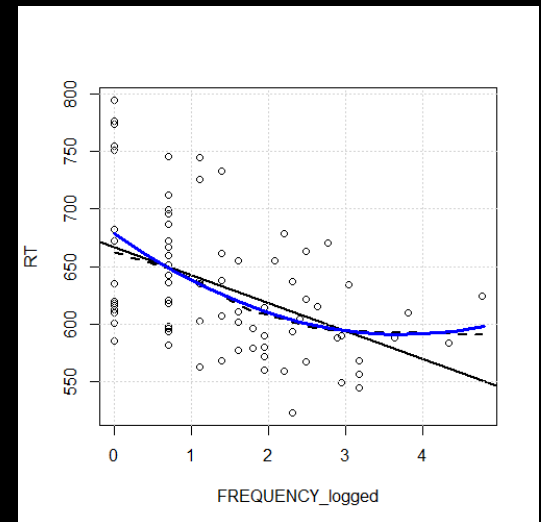
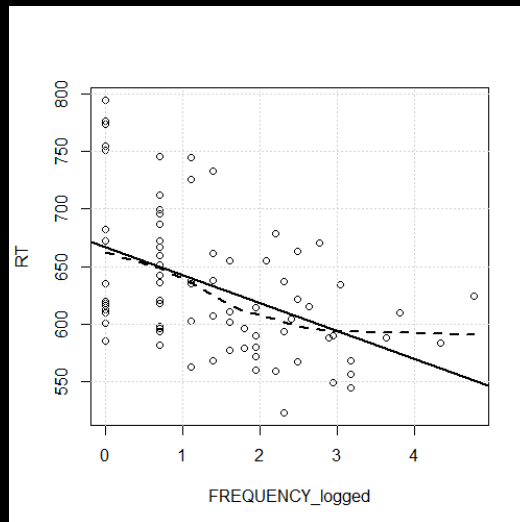
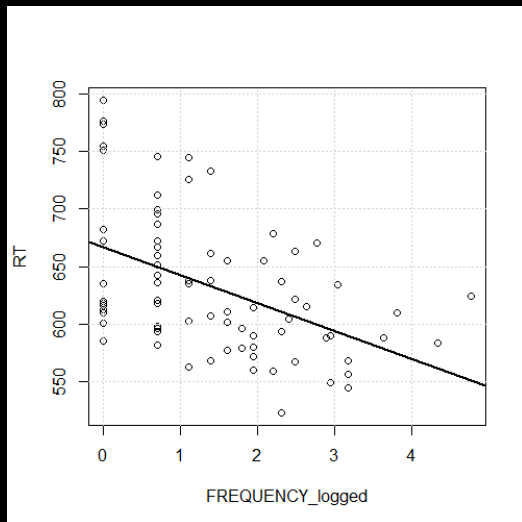
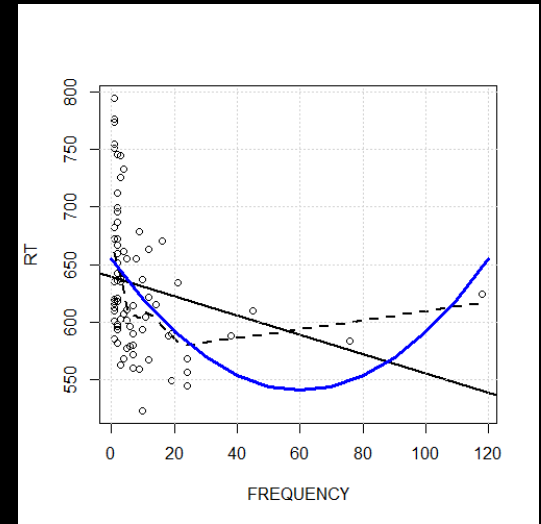
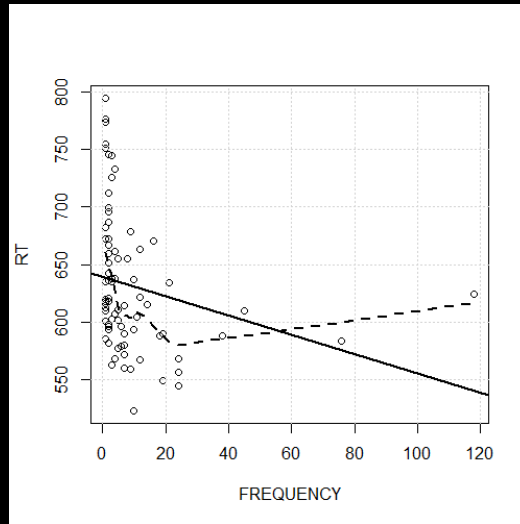
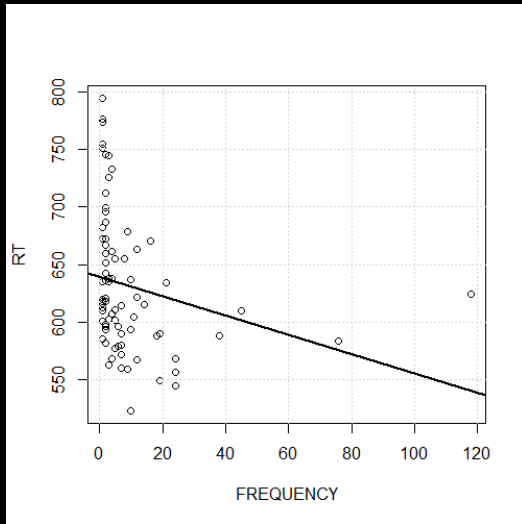
On axis limits



On axis limits and on uncertainty



On curvature



Jefrey Lijffijt, University of Bristol

A SECOND OPINION – VISUALISATION AND GOOD SCIENTIFIC PRACTICE

The aim of info-vis

1. Enable efficient exploration of data
2. Discover patterns
 - Exploratory data analysis \neq (just) graphs
 - Corpus linguists are *great* data explorers

Inspecting raw data

[Figure removed in order to respect copyrights]

Querying/playing with data (CQPweb, WordSmith Tools, ...)

The screenshot shows a web browser window with the URL `cqpweb.lancs.ac.uk/ame06/concordance.php?theData=somewhere&qmode=sq_nocase&pp=50&del=begin&t=8&del=end&uT=y`. The page displays the results of a query for the word "somewhere".

Query Results Summary:
 Your query "somewhere" returned 73 matches in 63 different texts (in 1,175,965 words [500 texts]; frequency: 62.08 instances per million words) [0.23 seconds - retrieved from cache]

Navigation and Controls:
 - Navigation buttons: < << >> >
 - Show Page: 1
 - Line View
 - Show in random order
 - New query
 - Go!

No	Filename	Solution 1 to 50	Page 1 / 2
1	AmE06_A27	future tides could be shallow indeed . HUNTSVILLE , Ala. --	Somewhere between curriculum committees , writing conference
2	AmE06_A42	about time . I know I have a wife and two kids	somewhere , " said Steines of " Entertainment Tonight . " Steine
3	AmE06_B20	that their DNA will be as much as 99.96 percent identical .	Somewhere in that 0.04 percent difference presumably lies the c
4	AmE06_D04	For the author of the Testament of Moses , Israel now stands	somewhere in the wilderness between the period of Moses 's pa
5	AmE06_D14	camping with McKenzie engaged the Blackfeet in a " severe battle "	somewhere on the middle Snake Plain , and when the victorious
6	AmE06_E17	" there 's not a book in there you ca n't get	somewhere else . " In any case , the bricks-and-mortar debate ,
7	AmE06_E21	add fifteen to twenty minutes lobby time whenever we have to be	somewhere , " Michael says . He describes how his daughter w
8	AmE06_F06	As I sit talking to her , I realize Donna lands	somewhere in the middle . She is okay . Not extraordinary but r
9	AmE06_F07	play : meeting new people (everyone who 's logged in is	somewhere in the SL landscape) , building a rocket , even havi
10	AmE06_F48	fired , any death that results , is documented by someone ,	somewhere and ricochets instantly across the world . Add to thi
11	AmE06_G01	our Dutch baronet . Elusive as a pearl he was , posing	somewhere unfindable in high boots and a long sword and a ser
12	AmE06_G02	sing it together often , over martinis , in bed , driving	somewhere . I used it in the most performed of my plays .
13	AmE06_G02	with me so often , might get a response . I read	somewhere that musicians with Alzheimer 's disease remember :
14	AmE06_G11	virtually all ethnography is parochially located , in some literal sense ,	somewhere - in some kind of place or nexus of places , or
15	AmE06_G16	The answer is simple : " A bliss began to happen	somewhere . There was America . " Revell 's love of America is
16	AmE06_G34	Is there a Zool ...	somewhere ...

Pattern discovery (Sketch Engine)

goal (*noun*) ukWaC freq = **168345** (107.5 per million)

<u>object of</u>	<u>58924</u>	<u>3.2</u>	<u>subject of</u>	<u>25451</u>	<u>2.4</u>	<u>modifier</u>	<u>67879</u>	<u>1.6</u>	<u>modifies</u>	<u>11026</u>	<u>0.3</u>
score	8390	11.28	score	903	8.59	ultimate	1911	9.27	scorer	389	9.39
achieve	9422	9.9	disallow	223	8.04	long-term	875	7.66	kick	634	8.86
concede	1421	9.39	concede	204	7.53	league	638	7.38	tally	129	7.9
accomplish	585	7.97	gape	76	6.5	winning	401	7.33	keeper	204	7.31
reach	1924	7.66	come	1316	5.44	primary	993	7.24	scramble	50	6.75
net	337	7.42	kick	76	5.44	second	2000	7.19	drought	78	6.65
pursue	648	7.41	rule	61	5.24	common	1529	7.17	difference	676	6.28
attain	400	7.35	orientate	34	5.06	strategic	645	7.1	cushion	53	6.26
grab	406	7.34	arrive	90	4.43	realistic	422	7.05	lead	267	6.24
set	2413	7.01	cap	20	4.38	achievable	290	6.97	setting	405	6.14
pull	501	6.88	beat	53	4.31	stated	259	6.8	kicker	25	6.04
disallow	190	6.67	direct	53	4.22	score	611	6.75	post	482	5.91

52



Finding typical raw data (ProtAnt)

ProtAnt 1.0.0

File Help

Target Corpus

Add Corpus File Add Corpus Directory Clear

1.txt
2.txt
3.txt
4.txt
5.txt
6.txt
7.txt
8.txt
9.txt
10.txt
11.txt

Reference Corpus

Add Corpus File Add Corpus Directory Clear

A01.txt
A02.txt
A03.txt
A04.txt
A05.txt
A06.txt
A07.txt
A08.txt
A09.txt
A10.txt
A11.txt

Keyness Statistic Log Likelihood

Threshold (p-value) 0.001 Keyword Min. Freq. 1

Norm Freq. per 1000 words

Case ☒ Ignore case (Change to lowercase)

Token Def (Regex) abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMN
OPQRSTUVWXYZ

Progress Start Stop

Results

Clear

	File	KeyTypes	KeyTokens	NormKeyTypes	NormKeyTokens	AllTypes	AllTokens
1	7.txt	67	237	150.901	227.011	444	1044
2	5.txt	80	267	149.813	221.026	534	1208
3	4.txt	70	267	144.628	228.01	484	1171
4	6.txt	49	242	133.88	232.692	366	1040
5	8.txt	55	203	129.108	203	426	1000
6	3.txt	44	170	107.843	172.239	408	987

	File	Key 1	Key 2	Key 3	Key 4	Key 5	Key 6	Key 7	Key 8	Key 9	Key 10
1	7.txt	islam	blair	muslim	brotherhood	assad	syria	arabia	saudi	extremism	speech
2	5.txt	islam	blair	muslim	brotherhood	islamic	assad	syria	arabia	saudi	speech
3	4.txt	islam	blair	muslim	brotherhood	islamic	assad	syria	arabia	saudi	extremism
4	6.txt	islam	muslims	muslim	islamic	pastor	speech	mccconnell	sermon	religion	congregation
5	8.txt	islam	blair	muslim	brotherhood	assad	syria	extremism	speech	intervention	ideology
6	3.txt	islam	muslims	batten	muslim	islamic	arabia	saudi	ukin	men	religion

	Keyword	Freq	Keyness
1	islam	41	239.602
2	muslims	37	210.93
3	batten	17	147.137
4	football	32	135.558
5	blair	35	132.355
6	joffrey	12	103.862

Summary

- Graphics can be very helpful
- For big data, graphs are often necessary
- But, please do not forget to carefully inspect your raw data

Discussion

1. Experimental design:
Which factors should we measure?
2. Non-randomness, dispersion and
the assumptions of hypothesis tests
3. Teaching and curricula
4. Visualisation
5. Which models can we use?

Gerold Schneider, Universität Zürich & Universität Konstanz

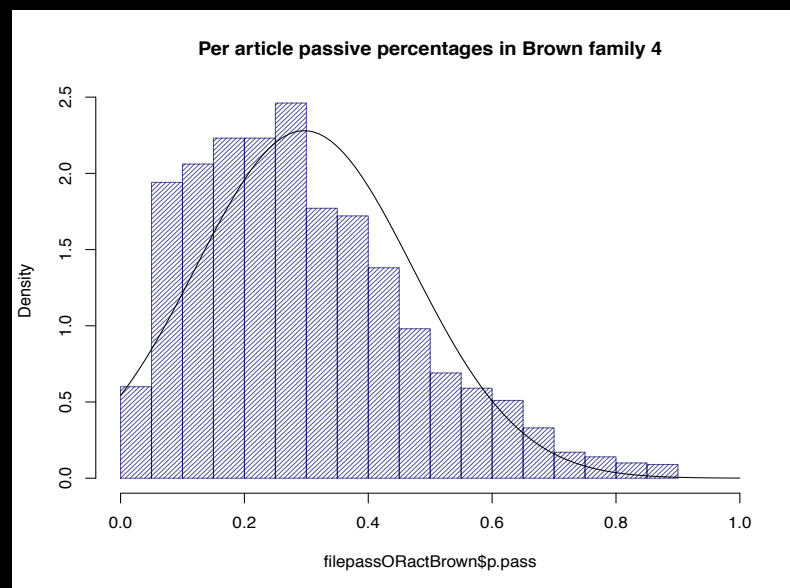
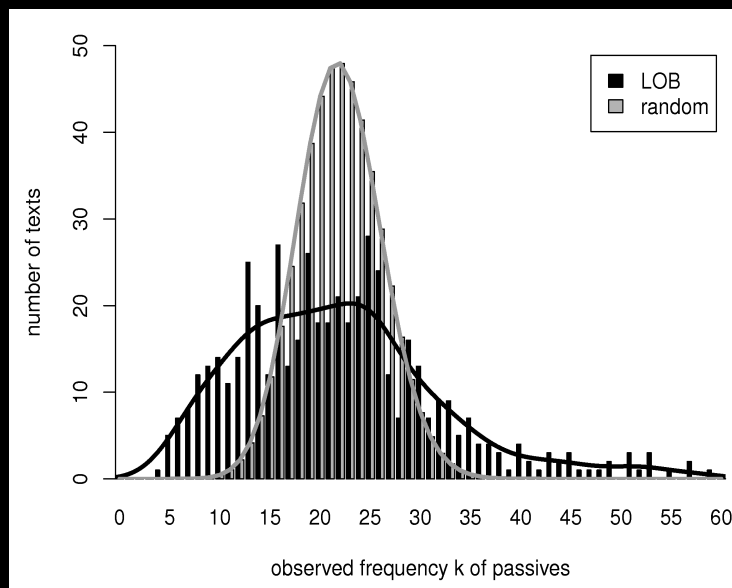
5 WHICH MODELS?

Which models can we use?

- Violated assumption
 - Random distribution | Improvement
 - Independence | Models of choice \Leftrightarrow frequency
 - Idiosyncratic Data | Multifactorial models
 - | Predictive models
 - Regression
 - Machine learning
- Characteristics of models
 - Model fit
 - Evaluation
 - Get to know your data!

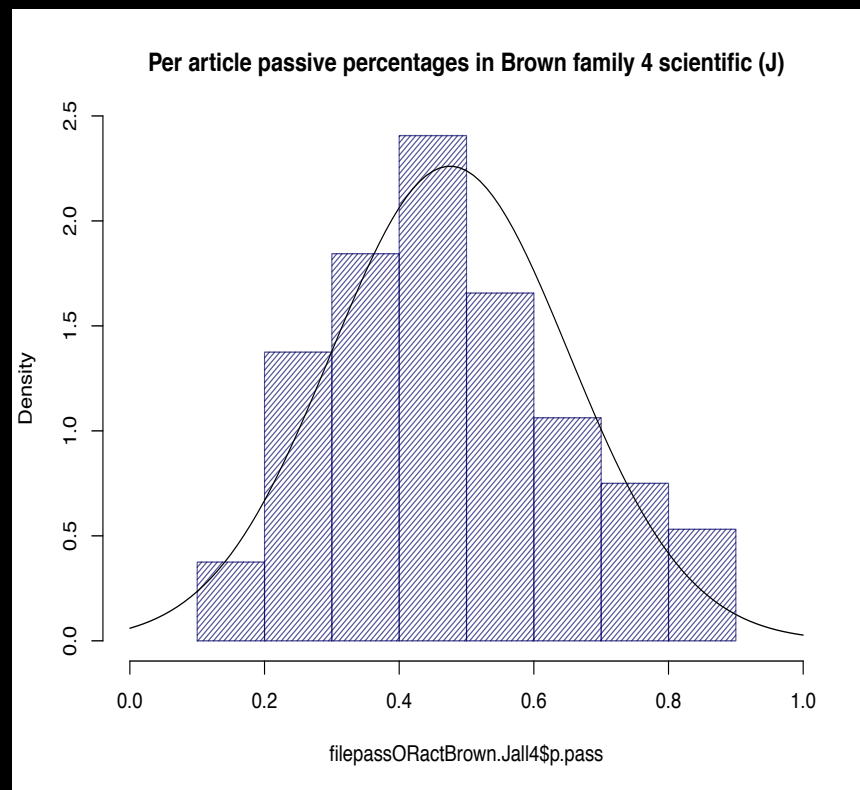
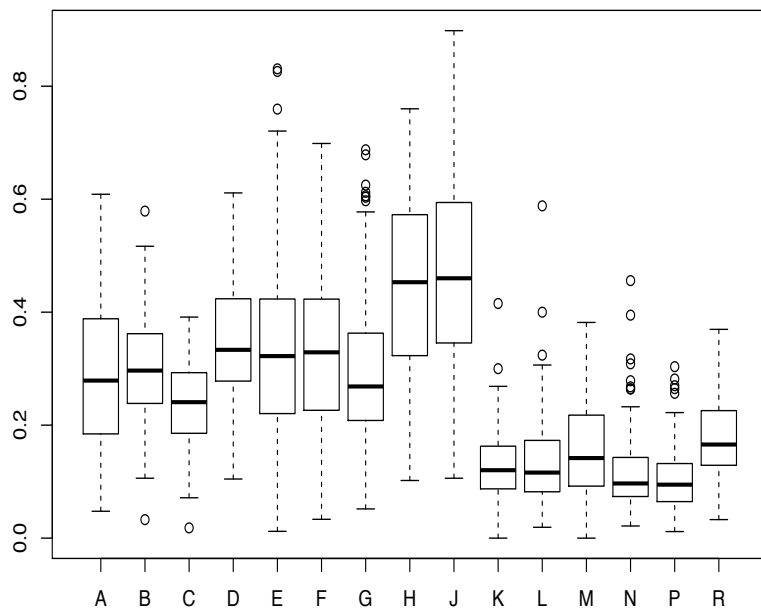
Which models can we use?

- Random distribution | Models of choice \Leftrightarrow frequency
Labov 1969, Church 2000, Evert 2006, Sean Wallis' Baseline
- e.g. passives: per article | restricted to transitive verbs



Which models can we use?

- Independence / discourse | Multifactorial models
Gries 2006, Gries 2010, Gries 2015
- e.g. genre (here Brown passives)

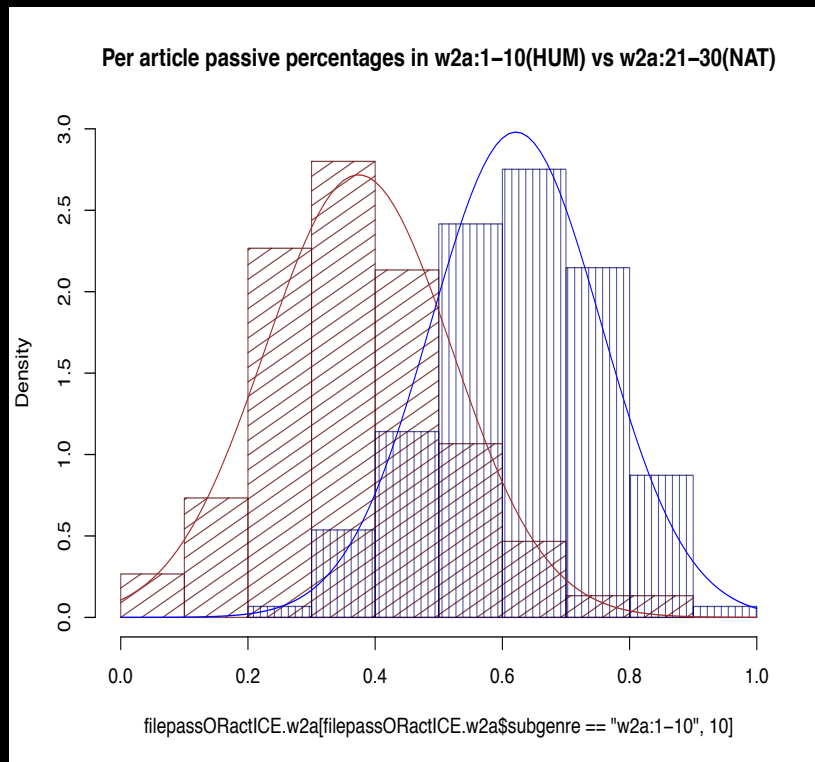


filepassORactBrown.Jall4\$p.pass

Which models can we use?

- Independence
- Genre and subgenre

| Multifactorial models
-> regression



```
filepassORactBrown.J4.subgenreregionperiod=
  aov(p.pass ~ subgenre * region * period,
      data=filepassORactBrown.J4);
summary(filepassORactBrown.J4.subgenreregionperiod)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
subgenre	3	1.7816	0.5939	29.126	7.37e-15	***
region	1	0.1218	0.1218	5.975	0.0157	*
period	1	0.4852	0.4852	23.795	2.74e-06	***
subgenre:region	3	0.0876	0.0292	1.433	0.2356	
subgenre:period	3	0.0821	0.0274	1.343	0.2628	
region:period	1	0.0350	0.0350	1.714	0.1924	
subgenre:region:period	3	0.0624	0.0208	1.021	0.3853	
Residuals	148	3.0176	0.0204			

Which models can we use?

- Predictive models: Machine learning
“Depending on definitional boundaries, predictive modelling is synonymous with, or largely overlapping with, the field of machine learning, as it is more commonly referred to in academic or research and development contexts.” (Wikipedia)
- Regression, naïve Bayes, SVM, ...
- There is a vast selection of tools out there.

A screenshot of the 'Train' window in the LightSide Researcher's Workbench. The window title is 'Train' and the name is 'bayes__12grams_length_1'. It shows a list of trained models, a table of model evaluation metrics, and a model confusion matrix.

Trained Models:

bayes__12gra...

Model Evaluation Metrics:

Metric	Value
Accuracy	0.9024
Kappa	0.7977

Model Confusion Matrix:

Act \ Pred	dem	rep
dem	1795	27
rep	290	1135

Trained Model: bayes__12gra...

LightSide

Report a Bug

Multithreaded 5.0 GB used, 16.0 GB max

Which models can we use?

- *Predictive* models
 - Data loss and compression / smoothing and generalisation
 - Effect sizes
 - Generalising power
 - Permits evaluation on different / held-out dataset
→ Evaluate! Get to know your data!
 - Massive feature set / feature selection
 - Interpretability vs. complexity of algorithm
- Computational linguistics tools:
 - Taggers
 - Parsers
 - Machine Translation
 - Distributional Semantics
 - ...
 - Different methods give complementary views (ML). Triangulate!



Which models can we use?

- Characteristics of models: Feature engineering
e.g. US party speech features from CORPS corpus:

Die typisch republikanischsten Merkmale sind:

Merkmal	F-score
've	0.6455
're	0.6443
nation	0.6443
it_'s	0.6336
men	0.6333
—	0.6312
i_'m	0.6286
'm	0.6286
you_all	0.6273
freedom	0.6261
we_'re	0.6254
well	0.6224
<PERIOD>_he	0.6219
<PERIOD>_and	0.6203
great	0.6192
's	0.6177
one	0.6159
government	0.6158
america	0.6153
military	0.6147

Typisch **un**republikanischste Merkmale (Auswahl):

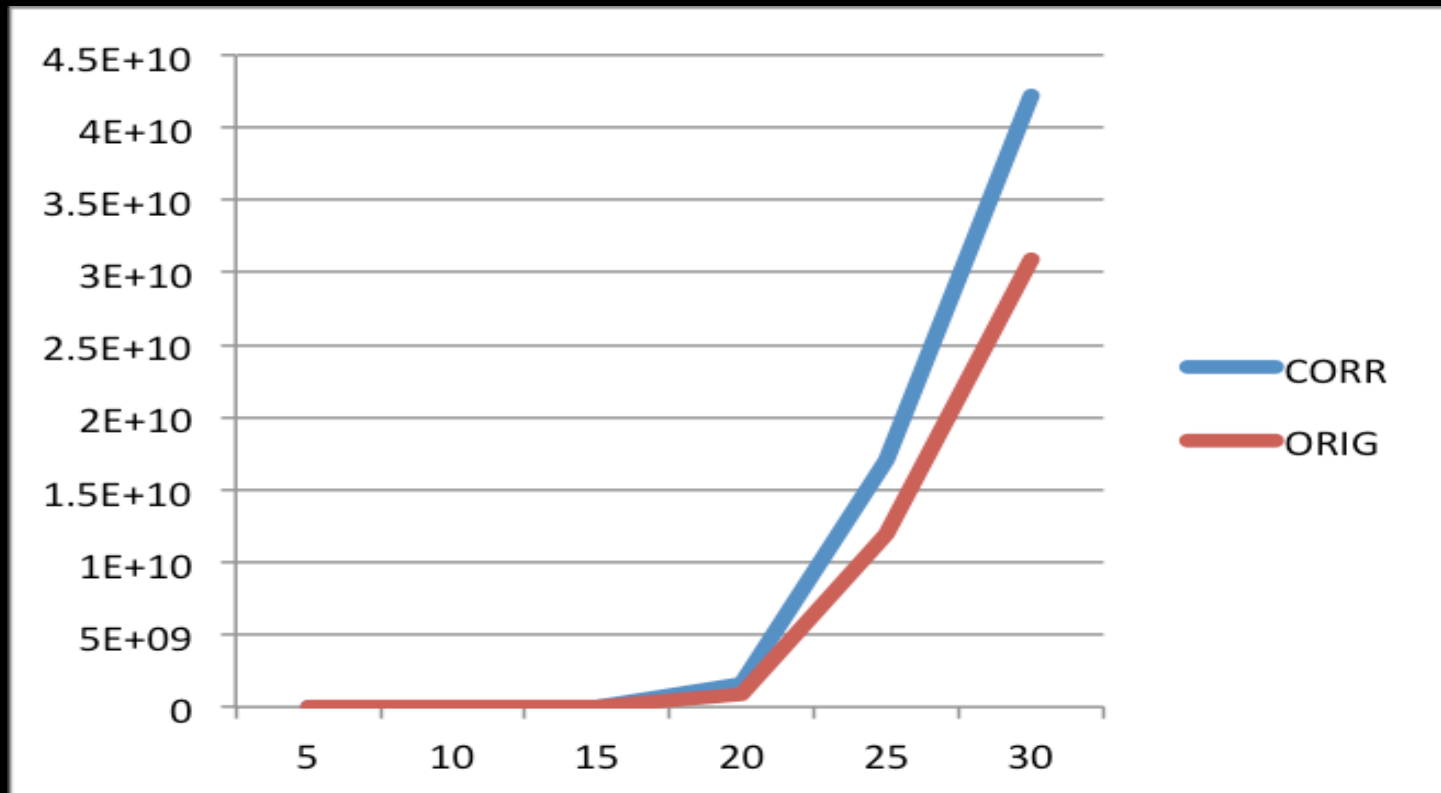
Merkmal	F-score
nra	0.0014
equal_pay	0.0014
of_climate	0.0014
racial_<COMMA>	0.0014
insurance_program	0.0014
high-wage	0.0014
our_steel	0.0014
without_health	0.0014
in_clean	0.0013
together_across	0.0013
campaign_finance	0.0013
to_hillary	0.0013
service_program	0.0013
fugitives	0.0013
stalkers	0.0013
our_planet	0.0013
financial_system	0.0013
after_high	0.0013
student_loans	0.0013
toxic_waste	0.0013
<PERIOD>_hillary	0.0013
from_welfare	0.0013
national_service	0.0013
more_police	0.0013

Which models can we use?

- Characteristics of model: Model fit / prediction accuracy
- A bad accuracy can mean:
 - You have a bad model. Get more features and semantic classifications (manual or automated) and take interactions into account
 - There is no pattern here. People have truly free choice, there is no story to be found. Your models aims to fit random distribution
 - The problem that you are dealing with is really challenging and deserved further, detailed research
 - You have some serious outliers in your data
- A good accuracy can mean:
 - I have a good model which respects all important factors
 - I have overfitted the data
 - My problem is trivial
 - The decisions are already taken in my features → independence?
- Use model fit as parameter, e.g. model fit of my syntactic parser is higher on corrected learner corpus than on original learner corpus

Which models can we use?

- Characteristics of models: Model fit *as parameter*: Learner English: parser scores = model fit is higher on corrected data



Which models can we use?

- Models! Multifactorial! Many! ML!
- Get to know your data. Evaluation and model refinement / feature selection / outlier analysis is a cyclical process.
- Corpus as a bicycle of the mind



John Sinclair 2014: “I am advocating that we should trust the text. We should be open to what it may tell us ... We should search for models that are especially appropriate to the study of text and discourse. The study of language is moving into a new era in which the exploitation of modern computers will be at the centre of progress”

George Box 1987: “all models are wrong, but some are useful”

Discussion

1. Experimental design:
Which factors should we measure?
2. Non-randomness, dispersion and
the assumptions of hypothesis tests
3. Teaching and curricula
4. Visualisation
5. Which models can we use?

GENERAL DISCUSSION

Thank you!

Statistical guidelines from *Nature*

Every article that contains statistical testing should state

- the name of the statistical test,
- the n value for each statistical analysis,
- the comparisons of interest,
- a justification for the use of that test (including, for example, a discussion of the normality of the data when the test is appropriate only for normal data),
- the alpha level for all tests, whether the tests were one-tailed or two-tailed, and
- the actual P value for each test (not merely "significant" or " $P < 0.05$ "). It should be clear what statistical test was used to generate every P value. Use of the word "significant" should always be accompanied by a P value;