Corpus statistics: key issues and controversies
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?
1 EXPERIMENTAL DESIGN

Sean Wallis, University College London
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

3A model of Corpus Linguistics
(Wallis and Nelson 2001)
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
  \[ p(X) \in [0, 1] \]
Baselines alter statistical models

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

\[ p(X) \in [0, ?] \]
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?

\[ p(X) \in [0, ?] \]
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

![Graph showing correct error interval width and increasing non-alternating forms with different initial f values and N values ranging from 100 to 10,000. Wider intervals indicate more conservative methods.](image)
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?
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

• $\chi^2$, 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

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

- Observed
- Predicted
- $p < 0.01$
Statistics vs. the truth

• ‘Language is never, ever, ever, random’ (Kilgarriff, 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. $T$

- $\chi^2$, log-likelihood ratio / $G$, Fisher Exact test assume independence of every instance

<table>
<thead>
<tr>
<th></th>
<th>$Y = \text{true}$</th>
<th>$Y = \text{false}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = \text{true}$</td>
<td>$R$</td>
<td>$S$</td>
</tr>
<tr>
<td>$X = \text{false}$</td>
<td>$T$</td>
<td>$U$</td>
</tr>
</tbody>
</table>
Why the problem matters

• If the assumptions are false, \( p \)-values can be too high or too low, \emph{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
  \[-/\rightarrow\text{ statistical testing is an ill choice}\]

• Often, there are alternatives
  1. Manipulate the representation (adjusted counts)
  2. Select only appropriate data (use dispersion)
  3. Use another test
     \((t\text{-test, anova are almost always fine})\)
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?
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 (≠ 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”

  ![Graph showing log10 occurrence count against frequency.](image)

  - $P(f = 1) = \alpha(1 - \gamma)$ (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 $\alpha$, $\gamma$, $p$ 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?
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.*
1. Paste tab delimited data including header row and id column. For help on importing data, see \\
**log likelihood**

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?
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 ***  
CONJ bevor/before -0.819629   0.098658 -8.308  < 2e-16 ***  
CONJ nachdem/after -0.064951   0.085287  -0.762 0.446     
CONJ weil/because  2.968349   0.088017  33.725  < 2e-16 ***  
LENGTH_DIFF      0.111711   0.009568  11.675  < 2e-16 ***  
CONJ bevor/before:LENGTH_DIFF -0.147896   0.015960  -9.267  < 2e-16 ***  
CONJ nachdem/after:LENGTH_DIFF -0.089803   0.013054  -6.934 8.3ef-08 ***  
CONJ weil/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

<table>
<thead>
<tr>
<th></th>
<th>animate</th>
<th>inanimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>s</td>
<td>50</td>
<td>15</td>
</tr>
</tbody>
</table>
On perspectives and uncertainty

Figure 12. The interaction POSSEDSPEC : LANG
On granularity
On axis limits
On axis limits and on uncertainty
On curvature
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 ≠ (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, ...)

<table>
<thead>
<tr>
<th>No</th>
<th>Filename</th>
<th>Solution 1 to 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AmE95_A07</td>
<td>future tides could be shallow indeed. HUNTSVILLE, Ala. — Somewhere between curriculum committees, writing conference...</td>
</tr>
<tr>
<td>2</td>
<td>AmE95_A42</td>
<td>about time. I know I have a wife and two kids. somewhere, &quot; said Steines of &quot; Entertainment Tonight. &quot; Steine...</td>
</tr>
<tr>
<td>3</td>
<td>AmE95_B20</td>
<td>that their DNA will be as much as 99.96 percent identical. Somewhere in that 0.04 percent difference presumably lies the...</td>
</tr>
<tr>
<td>4</td>
<td>AmE95_D04</td>
<td>For the author of the Testament of Moses, Israel now stands somewhere in the wilderness between the period of Moses’s passage...</td>
</tr>
<tr>
<td>5</td>
<td>AmE95_D14</td>
<td>camping with McKenzie engaged the Blackfeet in a &quot;severe battle. &quot; somewhere on the middle Snake Path, and when the victorious...</td>
</tr>
<tr>
<td>6</td>
<td>AmE95_E17</td>
<td>&quot; there ’ s not a book in there you can ’ t get somewhere else. &quot; In any case, the bricks-and-mortar debate...</td>
</tr>
<tr>
<td>7</td>
<td>AmE95_E31</td>
<td>add fifteen to twenty minutes lobby time whenever we have to be somewhere. &quot; Michael says. He describes how his daughter...</td>
</tr>
<tr>
<td>8</td>
<td>AmE95_F06</td>
<td>As I sit talking to her, I realize Donna lands somewhere in the middle. She is okay. Not extraordinary but...</td>
</tr>
<tr>
<td>9</td>
<td>AmE95_F07</td>
<td>play, meeting new people ( everyone who ’ s logged in is somewhere in the SL landscape ), building a rocket, even having...</td>
</tr>
<tr>
<td>10</td>
<td>AmE95_F48</td>
<td>fired, any death that results, is documented by someone somewhere and crochets instantly across the world. Add to this...</td>
</tr>
<tr>
<td>11</td>
<td>AmE95_G01</td>
<td>our Dutch baronet. Ehrase as a pearl he was , posing somewhere unfailingly in high boots and a long sword and a sword...</td>
</tr>
<tr>
<td>12</td>
<td>AmE95_G02</td>
<td>sing it together often, over martins, in bed, driving somewhere I used it in the most performed of my plays.</td>
</tr>
<tr>
<td>13</td>
<td>AmE95_G02</td>
<td>with me so often, might get a response. I read somewhere that musicians with Alzheimer’s disease remember...</td>
</tr>
<tr>
<td>14</td>
<td>AmE95_G11</td>
<td>virtually all ethnography is parenchymatically located, in some literal sense somewhere — in some kind of place or nexus of places, or...</td>
</tr>
<tr>
<td>15</td>
<td>AmE95_G16</td>
<td>The answer is simple: &quot; A bliss began to happen somewhere. There was America. &quot; Retell’s love of America...</td>
</tr>
</tbody>
</table>
Pattern discovery (Sketch Engine)

<table>
<thead>
<tr>
<th>object_of</th>
<th>score</th>
<th>achieve</th>
<th>concede</th>
<th>accomplish</th>
<th>reach</th>
<th>net</th>
<th>pursue</th>
<th>attain</th>
<th>grab</th>
<th>set</th>
<th>pull</th>
<th>disallow</th>
</tr>
</thead>
<tbody>
<tr>
<td>ukWaC freq</td>
<td>58924</td>
<td>9422</td>
<td>1421</td>
<td>585</td>
<td>1924</td>
<td>337</td>
<td>648</td>
<td>400</td>
<td>406</td>
<td>2413</td>
<td>501</td>
<td>190</td>
</tr>
<tr>
<td>3.2</td>
<td>11.28</td>
<td>9.9</td>
<td>9.39</td>
<td>7.97</td>
<td>7.66</td>
<td>7.42</td>
<td>7.41</td>
<td>7.35</td>
<td>7.34</td>
<td>7.01</td>
<td>6.88</td>
<td>6.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>subject_of</th>
<th>score</th>
<th>disallow</th>
<th>concede</th>
<th>gape</th>
<th>come</th>
<th>kick</th>
<th>rule</th>
<th>orientate</th>
<th>arrive</th>
<th>cap</th>
<th>beat</th>
<th>direct</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ukWaC freq</td>
<td>25451</td>
<td>223</td>
<td>204</td>
<td>76</td>
<td>1316</td>
<td>76</td>
<td>61</td>
<td>34</td>
<td>90</td>
<td>20</td>
<td>53</td>
<td>53</td>
<td>611</td>
</tr>
<tr>
<td>2.4</td>
<td>8.04</td>
<td>7.53</td>
<td>6.5</td>
<td>5.44</td>
<td>5.44</td>
<td>5.24</td>
<td>5.06</td>
<td>4.43</td>
<td>4.43</td>
<td>4.38</td>
<td>4.31</td>
<td>4.22</td>
<td>6.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>modifier</th>
<th>ultimate</th>
<th>long-term</th>
<th>league</th>
<th>winning</th>
<th>primary</th>
<th>second</th>
<th>common</th>
<th>strategic</th>
<th>realistic</th>
<th>achievable</th>
<th>stated</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ukWaC freq</td>
<td>67879</td>
<td>875</td>
<td>638</td>
<td>401</td>
<td>993</td>
<td>2000</td>
<td>1529</td>
<td>645</td>
<td>422</td>
<td>290</td>
<td>259</td>
<td>611</td>
</tr>
<tr>
<td>1.6</td>
<td>9.27</td>
<td>7.66</td>
<td>7.38</td>
<td>7.33</td>
<td>7.24</td>
<td>7.19</td>
<td>7.17</td>
<td>7.1</td>
<td>7.05</td>
<td>6.97</td>
<td>6.8</td>
<td>6.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>modifies</th>
<th>scorer</th>
<th>kick</th>
<th>tally</th>
<th>keeper</th>
<th>scramble</th>
<th>drought</th>
<th>difference</th>
<th>cushion</th>
<th>lead</th>
<th>setting</th>
<th>kicker</th>
<th>post</th>
</tr>
</thead>
<tbody>
<tr>
<td>ukWaC freq</td>
<td>11026</td>
<td>634</td>
<td>129</td>
<td>204</td>
<td>50</td>
<td>78</td>
<td>676</td>
<td>53</td>
<td>267</td>
<td>405</td>
<td>25</td>
<td>482</td>
</tr>
<tr>
<td>0.3</td>
<td>9.39</td>
<td>8.86</td>
<td>7.9</td>
<td>7.31</td>
<td>6.75</td>
<td>6.65</td>
<td>6.28</td>
<td>6.26</td>
<td>6.24</td>
<td>6.14</td>
<td>6.04</td>
<td>5.91</td>
</tr>
</tbody>
</table>
Pattern discovery
(Phrase Net, GraphColl, ...)

Phrase Net of Beatles Lyrics
Finding typical raw data (ProtAnt)
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?
5 WHICH MODELS?

Gerold Schneider, Universität Zürich & Universität Konstanz
Which models can we use?

- Violated assumption
  - Random distribution
  - Independence
  - Idiosyncratic Data
- Improvement
  - Models of choice ↔ frequency
  - 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
  Labov 1969, Church 2000,
- e.g. passives: per article
- Models of choice ↔ frequency
  Evert 2006, Sean Wallis’ Baseline
- 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)
Which models can we use?

- Independence
- Genre and subgenre

<table>
<thead>
<tr>
<th>Multifactorial models</th>
</tr>
</thead>
<tbody>
<tr>
<td>-&gt; regression</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>subgenre</td>
<td>3</td>
<td>1.7816</td>
<td>0.5939</td>
<td>29.126</td>
</tr>
<tr>
<td>region</td>
<td>1</td>
<td>0.1218</td>
<td>0.1218</td>
<td>5.975</td>
</tr>
<tr>
<td>period</td>
<td>1</td>
<td>0.4852</td>
<td>0.4852</td>
<td>23.795</td>
</tr>
<tr>
<td>subgenre:region</td>
<td>3</td>
<td>0.0876</td>
<td>0.0292</td>
<td>1.433</td>
</tr>
<tr>
<td>subgenre:period</td>
<td>3</td>
<td>0.0821</td>
<td>0.0274</td>
<td>1.343</td>
</tr>
<tr>
<td>region:period</td>
<td>1</td>
<td>0.0350</td>
<td>0.0350</td>
<td>1.714</td>
</tr>
<tr>
<td>subgenre:region:period</td>
<td>3</td>
<td>0.0624</td>
<td>0.0208</td>
<td>1.021</td>
</tr>
<tr>
<td>Residuals</td>
<td>148</td>
<td>3.0176</td>
<td>0.0204</td>
<td></td>
</tr>
</tbody>
</table>
```
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.
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:

<table>
<thead>
<tr>
<th>Merkmal</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>'ve</td>
<td>0.6455</td>
</tr>
<tr>
<td>'re</td>
<td>0.6443</td>
</tr>
<tr>
<td>nation</td>
<td>0.6443</td>
</tr>
<tr>
<td>it_'s</td>
<td>0.6336</td>
</tr>
<tr>
<td>men</td>
<td>0.6333</td>
</tr>
<tr>
<td>–</td>
<td>0.6312</td>
</tr>
<tr>
<td>i_'m</td>
<td>0.6286</td>
</tr>
<tr>
<td>'m</td>
<td>0.6286</td>
</tr>
<tr>
<td>you_all</td>
<td>0.6273</td>
</tr>
<tr>
<td>freedom</td>
<td>0.6261</td>
</tr>
<tr>
<td>we_'re</td>
<td>0.6254</td>
</tr>
<tr>
<td>well</td>
<td>0.6224</td>
</tr>
<tr>
<td>&lt;PERIOD&gt; .he</td>
<td>0.6219</td>
</tr>
<tr>
<td>&lt;PERIOD&gt; .and</td>
<td>0.6203</td>
</tr>
<tr>
<td>great</td>
<td>0.6192</td>
</tr>
<tr>
<td>'s</td>
<td>0.6177</td>
</tr>
<tr>
<td>one</td>
<td>0.6159</td>
</tr>
<tr>
<td>government</td>
<td>0.6158</td>
</tr>
<tr>
<td>america</td>
<td>0.6153</td>
</tr>
<tr>
<td>military</td>
<td>0.6147</td>
</tr>
</tbody>
</table>

Typisch unrepulikanischste Merkmale (Auswahl):

<table>
<thead>
<tr>
<th>Merkmal</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>nra</td>
<td>0.0014</td>
</tr>
<tr>
<td>equal_pay</td>
<td>0.0014</td>
</tr>
<tr>
<td>of_climate</td>
<td>0.0014</td>
</tr>
<tr>
<td>racial_&lt;COMMA&gt;</td>
<td>0.0014</td>
</tr>
<tr>
<td>insurance_program</td>
<td>0.0014</td>
</tr>
<tr>
<td>high-wage</td>
<td>0.0014</td>
</tr>
<tr>
<td>our_steel</td>
<td>0.0014</td>
</tr>
<tr>
<td>without_health</td>
<td>0.0014</td>
</tr>
<tr>
<td>in_clean</td>
<td>0.0013</td>
</tr>
<tr>
<td>together_across</td>
<td>0.0013</td>
</tr>
<tr>
<td>campaign_finance</td>
<td>0.0013</td>
</tr>
<tr>
<td>to_hillary</td>
<td>0.0013</td>
</tr>
<tr>
<td>service_program</td>
<td>0.0013</td>
</tr>
<tr>
<td>fugitives</td>
<td>0.0013</td>
</tr>
<tr>
<td>stalkers</td>
<td>0.0013</td>
</tr>
<tr>
<td>our_planet</td>
<td>0.0013</td>
</tr>
<tr>
<td>financial_system</td>
<td>0.0013</td>
</tr>
<tr>
<td>after_high</td>
<td>0.0013</td>
</tr>
<tr>
<td>student_loans</td>
<td>0.0013</td>
</tr>
<tr>
<td>toxic_waste</td>
<td>0.0013</td>
</tr>
<tr>
<td>&lt;PERIOD&gt; .hillary</td>
<td>0.0013</td>
</tr>
<tr>
<td>from_welfare</td>
<td>0.0013</td>
</tr>
<tr>
<td>national_service</td>
<td>0.0013</td>
</tr>
<tr>
<td>more_police</td>
<td>0.0013</td>
</tr>
</tbody>
</table>
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;

http://www.nature.com/srep/publish/guidelines