Distributional Methods in Corpus Linguistics
Towards a Hermeneutic Cyborg

Stefan Evert
Computational Corpus Linguistics Group
FAU Erlangen-Nürnberg

www.stefan-evert.de | www.linguistik.fau.de
Distributional semantics

“Die Bedeutung eines Wortes liegt in seinem Gebrauch.”
— Ludwig Wittgenstein (1953)

“You shall know a word by the company it keeps!”
— J. R. Firth (1957)

“What people know when they say that they know a word is not how to recite its dictionary definition – they know how to use it [...] in everyday discourse.”
— George A. Miller (1986)
Distributional semantics

What is a “bardiwac”?  
- He handed her her glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- Nigel staggered to his feet, face flushed from too much bardiwac.
- Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.
- I dined off bread and cheese and this excellent bardiwac.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.
## Distributional semantics

### A thought experiment: deciphering hieroglyphs

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(knife)</td>
<td>51</td>
<td>20</td>
<td>84</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>(cat)</td>
<td>52</td>
<td>58</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td>???</td>
<td>115</td>
<td>83</td>
<td>10</td>
<td>42</td>
<td>33</td>
<td>17</td>
</tr>
<tr>
<td>(boat)</td>
<td>59</td>
<td>39</td>
<td>23</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(cup)</td>
<td>98</td>
<td>14</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(pig)</td>
<td>12</td>
<td>17</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>27</td>
</tr>
<tr>
<td>(banana)</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>18</td>
<td>0</td>
</tr>
</tbody>
</table>
Act I

DISTRIBUTIONAL METHODS
Distributional models

- A **distributional model** is a scaled / transformed co-occurrence matrix which represents the distribution of linguistic items across contexts.

<table>
<thead>
<tr>
<th></th>
<th>get</th>
<th>see</th>
<th>use</th>
<th>hear</th>
<th>eat</th>
<th>kill</th>
</tr>
</thead>
<tbody>
<tr>
<td>knife</td>
<td>0.027</td>
<td>-0.024</td>
<td>0.206</td>
<td>-0.022</td>
<td>-0.044</td>
<td>-0.042</td>
</tr>
<tr>
<td>cat</td>
<td>0.031</td>
<td>0.143</td>
<td>-0.243</td>
<td>-0.015</td>
<td>-0.009</td>
<td>0.131</td>
</tr>
<tr>
<td>dog</td>
<td>-0.026</td>
<td>0.021</td>
<td>-0.212</td>
<td>0.064</td>
<td>0.013</td>
<td>0.014</td>
</tr>
<tr>
<td>boat</td>
<td>-0.022</td>
<td>0.009</td>
<td>-0.044</td>
<td>-0.040</td>
<td>-0.074</td>
<td>-0.042</td>
</tr>
<tr>
<td>cup</td>
<td>-0.014</td>
<td>-0.173</td>
<td>-0.249</td>
<td>-0.099</td>
<td>-0.119</td>
<td>-0.042</td>
</tr>
<tr>
<td>pig</td>
<td>-0.069</td>
<td>0.094</td>
<td>-0.158</td>
<td>0.000</td>
<td>0.094</td>
<td>0.265</td>
</tr>
<tr>
<td>banana</td>
<td>0.047</td>
<td>-0.139</td>
<td>-0.104</td>
<td>-0.022</td>
<td>0.267</td>
<td>-0.042</td>
</tr>
</tbody>
</table>

- **rows** = target items
- **columns** = features
Distributional models

- **Wordspace** (Schütze 1998) / **HAL** (Lund/Burgess 1996)
  
  targets = words / features = words

\[
M = \begin{bmatrix}
\cdots & m_1 & \cdots \\
\cdots & m_2 & \cdots \\
\vdots \\
\cdots & m_k & \cdots 
\end{bmatrix}
\]

<table>
<thead>
<tr>
<th></th>
<th>breed</th>
<th>tail</th>
<th>feed</th>
<th>kill</th>
<th>important</th>
<th>explain</th>
<th>likely</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>83</td>
<td>17</td>
<td>7</td>
<td>37</td>
<td>–</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>dog</td>
<td>561</td>
<td>13</td>
<td>30</td>
<td>60</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>animal</td>
<td>42</td>
<td>10</td>
<td>109</td>
<td>134</td>
<td>13</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>time</td>
<td>19</td>
<td>9</td>
<td>29</td>
<td>117</td>
<td>81</td>
<td>34</td>
<td>109</td>
</tr>
<tr>
<td>reason</td>
<td>1</td>
<td>–</td>
<td>2</td>
<td>14</td>
<td>68</td>
<td>140</td>
<td>47</td>
</tr>
<tr>
<td>cause</td>
<td>–</td>
<td>1</td>
<td>–</td>
<td>4</td>
<td>55</td>
<td>34</td>
<td>55</td>
</tr>
<tr>
<td>effect</td>
<td>–</td>
<td>–</td>
<td>1</td>
<td>6</td>
<td>60</td>
<td>35</td>
<td>17</td>
</tr>
</tbody>
</table>
**Distributional models**

- **Latent semantic analysis** *(Landauer/Dumais 1996)*
  targets = *words* / features = *documents*

\[
\mathbf{F} = \begin{bmatrix}
\ldots & \mathbf{f}_1 & \ldots \\
\ldots & \mathbf{f}_2 & \ldots \\
\vdots & \vdots & \vdots \\
\ldots & \mathbf{f}_k & \ldots 
\end{bmatrix}
\]

<table>
<thead>
<tr>
<th></th>
<th>Felidae</th>
<th>Pet</th>
<th>Feral</th>
<th>Bloat</th>
<th>Philosophy</th>
<th>Kant</th>
<th>Back pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>dog</td>
<td>-</td>
<td>10</td>
<td>4</td>
<td>11</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>animal</td>
<td>2</td>
<td>15</td>
<td>10</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>time</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>reason</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>cause</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>effect</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>
Distributional models

- Latent semantic indexing (Deerwester et al. 1990)
  targets = documents / features = words

\[
D = \begin{bmatrix}
  \vdots & \vdots \\
  \vdots & \vdots \\
  f_1 & \cdots & f_k \\
  \vdots & \vdots \\
\end{bmatrix}
\begin{array}{c|c|c|c|c|c|c}
\text{cat} & \text{dog} & \text{animal} & \text{time} & \text{reason} & \text{cause} & \text{effect} \\
\hline
\text{Felidae} & 10 & - & 2 & 1 & - & - & - \\
\text{Pet} & 10 & 10 & 15 & - & 1 & - & - \\
\text{Feral} & 7 & 4 & 10 & - & - & - & - \\
\text{Bloat} & - & 11 & 2 & - & - & 2 & 1 \\
\text{Philosophy} & - & - & - & 2 & 1 & 1 & - \\
\text{Kant} & - & - & - & 1 & 4 & 2 & 1 \\
\text{Back pain} & - & - & - & - & 1 & 6 & - \\
\end{array}
\]
Distributional models

- **Multidimensional analysis** (Biber 1988, 1993, ...)
  
  targets = texts / features = stylometric features

\[
S = \begin{bmatrix}
\vdots & s_1 & \vdots \\
\vdots & s_2 & \vdots \\
\vdots & \vdots & \vdots \\
\vdots & s_k & \vdots \\
\end{bmatrix}
\]

<table>
<thead>
<tr>
<th></th>
<th>nominal</th>
<th>pass</th>
<th>prep</th>
<th>subord</th>
<th>ttr</th>
</tr>
</thead>
<tbody>
<tr>
<td>orig1</td>
<td>1.205</td>
<td>5.013</td>
<td>6.883</td>
<td>4.483</td>
<td>1.285</td>
</tr>
<tr>
<td>orig2</td>
<td>0.738</td>
<td>2.537</td>
<td>6.486</td>
<td>6.157</td>
<td>1.714</td>
</tr>
<tr>
<td>orig3</td>
<td>1.252</td>
<td>4.462</td>
<td>8.463</td>
<td>4.785</td>
<td>2.476</td>
</tr>
<tr>
<td>orig4</td>
<td>1.105</td>
<td>2.899</td>
<td>8.119</td>
<td>3.966</td>
<td>1.519</td>
</tr>
<tr>
<td>orig5</td>
<td>1.764</td>
<td>4.268</td>
<td>7.167</td>
<td>3.947</td>
<td>1.792</td>
</tr>
<tr>
<td>orig8</td>
<td>1.545</td>
<td>7.268</td>
<td>7.461</td>
<td>5.455</td>
<td>1.572</td>
</tr>
<tr>
<td>trans1</td>
<td>0.463</td>
<td>2.208</td>
<td>6.297</td>
<td>6.089</td>
<td>2.339</td>
</tr>
<tr>
<td>trans2</td>
<td>1.131</td>
<td>2.597</td>
<td>6.307</td>
<td>4.844</td>
<td>1.810</td>
</tr>
<tr>
<td>trans4</td>
<td>0.935</td>
<td>1.744</td>
<td>7.098</td>
<td>4.012</td>
<td>1.403</td>
</tr>
<tr>
<td>trans5</td>
<td>0.867</td>
<td>3.604</td>
<td>7.511</td>
<td>5.154</td>
<td>1.902</td>
</tr>
<tr>
<td>trans7</td>
<td>1.387</td>
<td>4.290</td>
<td>8.211</td>
<td>3.998</td>
<td>1.822</td>
</tr>
</tbody>
</table>
Distributional models

- **Authorship attribution with Δ** (Burrows 2002)
  - targets = texts / features = most frequent words
- **Relational similarity for analogies** (Turney 2006)
  - targets = word pairs / features = connectors
- **Unsupervised POS induction** (Schütze 1995)
  - targets = word forms / features = grammatical contexts
- ...
Parameters of distributional models

- pre-processed corpus with linguistic annotation
  - term-context matrix
  - define target terms
  - context tokens or types
  - geometric analysis
  - feature scaling
  - similarity/distance measure + normalization
  - dimensionality reduction

- term-term matrix
  - define target & feature terms
  - type & size of co-occurrence
  - probabilistic analysis
  - embedding learned by neural network
Dimensionality reduction

\[ n \]

\[ M \]

\[ k \]
Dimensionality reduction

Matrix factorization

\[ \mathbf{M} \approx \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T \]
Types of matrix factorization

- SVD: $\mathbf{U}, \mathbf{V}$ orthogonal + Euclidean approximation cost
  → wordspace, most of distributional semantics

- NMF: $\mathbf{U}, \mathbf{V} \geq 0$, Euclidean or cross-entropy cost
  ($\mathbf{U}/\mathbf{V}$ probabilistic if $\Sigma$ absorbs weight of components)

- Probabilistic topic model: $\mathbf{U}, \mathbf{V}$ probabilistic
  → PLSA, equivalent to cross-entropy NMF
  → LDA with sparseness constraint on $\mathbf{U}/\mathbf{V}$ (Dirichlet)

- Neural embeddings: unconstrained, softmax cost
  → word2vec SGNS (Levy/Goldberg 2014)

- Or invent your own bizarre heuristics ...
  → GloVe (Pennington et al. 2014)
Some interesting connections

Special case: co-occurrence of words within short text segments (e.g. tweets, translation beads, ...)

- Term-document matrix $F$ is binary with SVD $U\Sigma V^T$
- Cosine similarity in $F \rightarrow$ first-order association ($\text{MI}^2$)
- Term-term matrix $M = FF^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T$
  $\rightarrow$ second-order association = scaling of components
- Higher-order co-occurrence based on inner products
  $M M^T = U\Sigma^4 U^T$ $\rightarrow$ further scaling of components

“... cuius rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.” (Fermat 1637)
Act II

MODELLING CORPUS

FREQUENCIES
Leech et al. (2009, 164) state that “the data in the Brown family of corpora support the hypothesis of an ongoing decline in passive constructions in written English ... AmE ... is leading this change that, most likely, has to be attributed to the sustained attack on the passive in usage guides.”

**Frequency comparison:** 
**passive voice**

data from joint research with Gerold Schneider (U Zürich)

AmE:
- **Brown**: 12.6% (12,623 / 100,102 VP)
- **1960s**
- **Frown**: 10.4% (10,779 / 104,078 VP)

BrE:
- **LOB**: 13.3% (13,713 / 103,385 VP)
- **FLOB**: 12.3% (12,597 / 102,473 VP)

**1990s**
Frequency comparison: **passive voice**

Confirm significance with chi-squared test

- **Brown vs. LOB:**
  - $X^2 = 19.24$, df = 1
  - $p = .000012 < .001$ ***
  - $\delta = 0.36\% \ldots 0.95\%$

- **Brown vs. Frown:**
  - $X^2 = 255.12$, df = 1
  - $p < 10^{-16}$ ***
  - $\delta = 1.98\% \ldots 2.53\%$

- **Brown:** 12.6%  
  12,623 / 100,102 VP

- **LOB:** 13.3%  
  13,713 / 103,385 VP

- **Frown:** 10.4%  
  10,779 / 104,078 VP

- **FLOB:** 12.3%  
  12,597 / 102,473 VP

- $\delta > 0.36\%$ ***  
- $\delta > 1.98\%$ ***  
- $\delta > 0.68\%$ ***  
- $\delta > 1.66\%$ ***
Frequency comparison: **passive voice**
data from joint research with Gerold Schneider (U Zürich)

- **Confirm significance with Student's t-test**
  - $H_0$: macro-avg. (texts) instead of micro-avg.

- **Brown vs. LOB:**
  - $t = 1.22$, df = 997
  - $p = .223$ n.s.

- **Brown vs. Frown:**
  - $t = -4.03$, df = 991.5
  - $p = .000061 < .001^{***}$
  - $\delta = 1.03\% \ldots 3.00\%$

- MW: $p = .049^{*}$

- MW: $p = .090$ n.s.

- MW: $p < .001^{***}$

Brown: 14.2% micro-avg. 12.6%
LOB: 14.9% micro-avg. 13.3%

Frown: 11.6% micro-avg. 10.4%
FLOB: 13.7% micro-avg. 12.3%
Variation as a nuisance parameter

- Many aspects of linguistic variation are nuisance parameters in corpus linguistics
  - Leech et al. (2009) are interested in the comparison of AmE vs. BrE and 1960s vs. 1990s
  - ignore factors such as register variation
  - pooled data: inflated significance
  - t-test: huge within-group variance
Modelling linguistic variation

- Quantitative analysis must account for the variability of relative frequencies between texts (Gries 2006)

- The contrast of interest is one possible factor behind this variation → can be tested for significance

- Gerold and I propose that (linear) statistical models are the appropriate technique for such an approach
Modelling linguistic variation

- Linear model predicts relative frequency in each text based on variety, genre, and other factors
  - goodness-of-fit = $R^2 = \text{proportion of explained variance}
    \begin{equation}
    1 - R^2 = \frac{\sum_i (\epsilon_i)^2}{\sum_i (p_i - \beta_0)^2}
    \end{equation}

\[
p_i = \beta_0 + \beta_1 (\text{genre}) + \beta_2 (\text{AmE/BrE}) + \cdots + \epsilon_i
\]

- Grand mean
- Effect of genre differences
- Relative frequency in text $i$
- Unexplained residuals (binomial sampling + other factors)
LM analysis: genre

English passives in the 1960s: genre

R² = 46.0% (var = 51253.4)
LM analysis: genre + AmE/BrE

English passives in the 1960s: genre + AmE/BrE

R² = 46.1% (var = 51113.1)
LM analysis: **genre × AmE/BrE**

English passives in the 1960s: genre + AmE/BrE + interaction

- **R² = 45.9% (var = 50548.2)**
## Analysis of variance

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Expl. Var</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>genre</td>
<td>14</td>
<td>44951</td>
<td>46.72%</td>
<td>61.6131</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>region</td>
<td>1</td>
<td>140</td>
<td>0.14%</td>
<td>2.6923</td>
<td>0.1012</td>
</tr>
<tr>
<td>genre:region</td>
<td>14</td>
<td>565</td>
<td>0.58%</td>
<td>0.7743</td>
<td>0.6980</td>
</tr>
<tr>
<td>Residuals</td>
<td>970</td>
<td>50548</td>
<td>52.54%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- AmE / BrE has no significant effect
  - also no evidence for genre-specific differences (interaction)

- But does this model really account for variation?
  - still **52.5% residual variance**!
  - need more explanatory factors
Distributional methods to the rescue

- Biber (1993) identifies latent register dimensions from correlations btw. linguistic features
- A distributional model: texts × stylistic features
- Use register dimensions as predictive factors in linear model analysis
Latent register dimensions

- Simplified: latent registers based on POS frequency profiles (C7 tags)
- W/o any verb tags to avoid circularity
- PCA on standardised frequency counts
Latent register dimensions

- Simplified: latent registers based on POS frequency profiles (C7 tags)
- W/o any verb tags to avoid circularity
- PCA on standardised frequency counts
- First 10 latent dim's as explanatory factors in LM model
Latent register dimensions

- Simplified: latent registers based on POS frequency profiles (C7 tags)
- W/o any verb tags to avoid circularity
- PCA on standardised frequency counts
- First 10 latent dim's as explanatory factors in LM model
Latent topic dimensions

- Domain & theme may also have effect on passives
- LSI = distributional topic model: texts × words
- First 10 topic dim's as explanatory factors in LM model
Latent topic dimensions

- Domain & theme may also have effect on passives
- **LSI** = distributional topic model: texts × words
- First 10 topic dim's as explanatory factors in LM model
- Must exclude last topic dim to avoid explaining away!
LM analysis: genre + var

English passives in the 1960s: genre + AmE/BrE

R2 = 46.1% (var = 51113.1)
LM analysis: genre + var + register

English passives in the 1960s: genre + 4 register dim's + AmE/BrE

R² = 69.3% (var = 29004.1)
LM analysis: genre + var + register + topic

English passives in the 1960s: genre + 4 register dim's + 6 topic dim's + AmE/BrE

R² = 73.4% (var = 24973.4)
## The final result

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Expl. Var</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>genre</td>
<td>14</td>
<td>44951</td>
<td>46.72%</td>
<td>125.2242</td>
<td>&lt;2.2e-16 ***</td>
</tr>
<tr>
<td>reg1</td>
<td>1</td>
<td>15425</td>
<td>16.03%</td>
<td>601.6077</td>
<td>&lt;2.2e-16 ***</td>
</tr>
<tr>
<td>reg3</td>
<td>1</td>
<td>4262</td>
<td>4.43%</td>
<td>166.2120</td>
<td>&lt;2.2e-16 ***</td>
</tr>
<tr>
<td>reg5</td>
<td>1</td>
<td>171</td>
<td>0.17%</td>
<td>6.6538</td>
<td>0.01004 *</td>
</tr>
<tr>
<td>reg6</td>
<td>1</td>
<td>1076</td>
<td>1.11%</td>
<td>41.9841</td>
<td>1.459e-10 ***</td>
</tr>
<tr>
<td>top2</td>
<td>1</td>
<td>118</td>
<td>0.12%</td>
<td>4.5965</td>
<td>0.03228 *</td>
</tr>
<tr>
<td>top4</td>
<td>1</td>
<td>1129</td>
<td>1.17%</td>
<td>44.0381</td>
<td>5.337e-11 ***</td>
</tr>
<tr>
<td>top5</td>
<td>1</td>
<td>742</td>
<td>0.77%</td>
<td>28.9237</td>
<td>9.428e-08 ***</td>
</tr>
<tr>
<td>top8</td>
<td>1</td>
<td>1191</td>
<td>1.24%</td>
<td>46.4364</td>
<td>1.656e-11 ***</td>
</tr>
<tr>
<td>top9</td>
<td>1</td>
<td>754</td>
<td>0.78%</td>
<td>29.4211</td>
<td>7.348e-08 ***</td>
</tr>
<tr>
<td>top10</td>
<td>1</td>
<td>3</td>
<td>0.00%</td>
<td>0.1349</td>
<td>0.71352</td>
</tr>
<tr>
<td>region</td>
<td>1</td>
<td>1409</td>
<td>1.47%</td>
<td>54.9514</td>
<td>2.680e-13 ***</td>
</tr>
</tbody>
</table>

Residuals | 974 | 24973 | 25.96%

👉 AmE / BrE is highly significant now!
The final result

gener effect plot

region effect plot

difference: 2.07% ... 3.56%
Act III

THE HERMENEUTIC CYBORG
Corpus linguistics in 2018

- Concordance
- Collocation
- Keywords
- Frequency analysis

But seriously:

- A lot of highly successful work is still based on methods from the 1970s
- Just bigger, faster and more convenient
Corpus linguistics in 2018

But seriously: A lot of highly successful work is still based on methods from the 1970s. Just bigger, faster and more convenient.
Digital humanities

- Humanities research transformed by possibilities of computational analysis & information visualization
- Origins in 1950s, but explosive growth in recent years
- Eager adoption of new analysis techniques

Ignore for today:

- large part of DH is compilation of digital editions, creation of software platforms & online interfaces
Digital humanities by example

Digital humanities by example

https://voyant-tools.org/
Digital humanities by example

“distant reading”

Deep learning & AI

- Artificial neural networks → general ML algorithm
- Origins in 1950s, but recent revival due to improvements in processing power (esp. GPGPU)
- Substantial improvements in language modelling, text categorization, analogies, machine translation, visual object recognition, OCR, playing Go, ...
- Simulate humans (AI phone calls, Obama lip-sync)
- Super-human performance (e.g. relational reasoning)
- End-to-end learning w/o any human insight
- Zero-shot learning w/o any training data
Deep learning


[Diagram of a neuron with inputs and output]

CC BY-SA 4.0 | https://commons.wikimedia.org/wiki/File:Kernel_Machine.svg
Deep learning


CC BY-SA 4.0 | https://commons.wikimedia.org/wiki/File:Kernel_Machine.svg
Deep learning

https://einstein.ai/research/the-natural-language-decathlon
The future of corpus linguistics
The future of corpus linguistics

1) Interoperability
The future of corpus linguistics

2) Interactivity
The future of corpus linguistics

3) Integration
The future of corpus linguistics

Hermeneutic
Cyborg

Corpus

Insight

Applications
Step 1: Interoperability

- Standard toolbox of CL: CQPweb, SketchEngine, AntConc, WordSmith, MonoConc, LancsBox, ...
  - provide almost the same basic functionality
- Slow adoption of innovative techniques (if at all)
- Goal: two-way interoperability
  - query tool $\rightarrow$ quantitative data $\rightarrow$ analysis/visualization
  - visualization $\rightarrow$ concordance in query tool $\rightarrow$ sort/context
- The Coquery way: tabular data
  - enables Coquery users to carry out flexible analysis within the software itself

https://www.coquery.org/
# Tabular data

- **Text-feature matrix for multivariate analysis**

**CQPweb: Analyse corpus function**

<table>
<thead>
<tr>
<th>id</th>
<th># sent</th>
<th># word</th>
<th>ld</th>
<th>nn</th>
<th>poss</th>
<th>pron</th>
<th>adj</th>
<th>fin</th>
<th>past</th>
<th>pass</th>
<th>will</th>
<th>coord</th>
<th>sub</th>
</tr>
</thead>
<tbody>
<tr>
<td>dpc-aby-002270-en</td>
<td>32</td>
<td>795</td>
<td>457</td>
<td>228</td>
<td>11</td>
<td>16</td>
<td>77</td>
<td>51</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>50</td>
<td>8</td>
</tr>
<tr>
<td>dpc-aby-002271-en</td>
<td>9</td>
<td>201</td>
<td>118</td>
<td>52</td>
<td>3</td>
<td>6</td>
<td>19</td>
<td>13</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>dpc-aby-002272-en</td>
<td>9</td>
<td>208</td>
<td>131</td>
<td>57</td>
<td>5</td>
<td>9</td>
<td>18</td>
<td>19</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>dpc-aby-002273-en</td>
<td>26</td>
<td>494</td>
<td>296</td>
<td>145</td>
<td>10</td>
<td>16</td>
<td>67</td>
<td>35</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>dpc-aby-002274-en</td>
<td>8</td>
<td>138</td>
<td>75</td>
<td>27</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>dpc-aby-002275-en</td>
<td>13</td>
<td>319</td>
<td>185</td>
<td>83</td>
<td>6</td>
<td>19</td>
<td>27</td>
<td>24</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>dpc-aby-002276-en</td>
<td>16</td>
<td>378</td>
<td>229</td>
<td>92</td>
<td>4</td>
<td>9</td>
<td>50</td>
<td>27</td>
<td>11</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>dpc-aby-002277-en</td>
<td>14</td>
<td>297</td>
<td>180</td>
<td>85</td>
<td>3</td>
<td>9</td>
<td>27</td>
<td>26</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>dpc-aby-002278-en</td>
<td>41</td>
<td>832</td>
<td>496</td>
<td>215</td>
<td>5</td>
<td>15</td>
<td>110</td>
<td>72</td>
<td>5</td>
<td>11</td>
<td>7</td>
<td>33</td>
<td>12</td>
</tr>
</tbody>
</table>
Tabular data

- Token-level data for distribution, word sketch, ...

**CQPweb: Download query as plain-text tabulation**

<table>
<thead>
<tr>
<th>head</th>
<th>head cpos</th>
<th>rel</th>
<th>dep</th>
<th>dep cpos</th>
<th>year</th>
<th>genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>book_V</td>
<td>42221</td>
<td>subj</td>
<td>i_P</td>
<td>42219</td>
<td>1929</td>
<td>thriller</td>
</tr>
<tr>
<td>book_V</td>
<td>42221</td>
<td>ncmod</td>
<td>will_V</td>
<td>42220</td>
<td>1929</td>
<td>thriller</td>
</tr>
<tr>
<td>book_V</td>
<td>42221</td>
<td>dobj</td>
<td>room_N</td>
<td>42223</td>
<td>1929</td>
<td>thriller</td>
</tr>
<tr>
<td>give_V</td>
<td>42231</td>
<td>subj</td>
<td>i_P</td>
<td>42219</td>
<td>1929</td>
<td>thriller</td>
</tr>
<tr>
<td>give_V</td>
<td>42231</td>
<td>dobj</td>
<td>card_N</td>
<td>42235</td>
<td>1929</td>
<td>thriller</td>
</tr>
<tr>
<td>give_V</td>
<td>42231</td>
<td>iobj</td>
<td>you_P</td>
<td>42232</td>
<td>1929</td>
<td>thriller</td>
</tr>
<tr>
<td>give_V</td>
<td>42231</td>
<td>ncmod</td>
<td>then_R</td>
<td>42218</td>
<td>1929</td>
<td>thriller</td>
</tr>
<tr>
<td>card_N</td>
<td>42235</td>
<td>ncmod</td>
<td>credit_N</td>
<td>42234</td>
<td>1929</td>
<td>thriller</td>
</tr>
<tr>
<td>use_V</td>
<td>42237</td>
<td>dobj</td>
<td>card_N</td>
<td>42235</td>
<td>1929</td>
<td>thriller</td>
</tr>
</tbody>
</table>
From step 2 to step 3

grouping & structuring

feedback

Insight

Applications

Corpus

CC BY 4.0 | http://fontawesome.com/
Step 2: Interactivity

- Interpretation procedure should be **interactive**
  - larger part of workflow integrated into corpus software
  - maintains connection to concordance
  - requires dedicated visualization component in analysis tool

- For keywords & collocations: grouping/sorting
  - usually first stage of interpreting keywords/collocations
  - central to **corpus-based discourse analysis** (Baker 2006)
  - could easily be supported by interactive visualization

- **Word embeddings** provide semantic map
  - for the interactive visualization
Work in progress: MMDA
Step 3: Integration

- Key challenge: how to feed back information from manual grouping into quantitative procedure
- Distributional word embeddings provide suitable framework → derived from co-occurrence patterns
  - collocations / keywords = first-order statistics
  - embeddings = second-order statistics
  - more than just a convenient semantic map
- MMDA operationalization
  - collocate groups (and topic nodes) form “discoursemes”
  - discourse positions = co-occurrences of discoursemes
MMDA: discourse similarity
MMDA: discourse similarity

tough stance

rules
vow
moves
approach
stance
way
stand
against

stresses
insists
understands
said
constructive
careful
calm
smart
liar
nasty

understands
stresses

wants

insists

said

smart
careful
calm

understands

rules

vow

moves

way

approach

stance

stand

against
THANK YOU!
ありがとうございます