The Hermeneutic Cyborg

Sinclair Lecture 2018

Stefan Evert

FAU Erlangen-Nürnberg

www.stefan-evert.de
Foreword

- Disclaimer: programmatic, no quantitative results → my vision for the future of corpus linguistics

- Promise: no dinosaurs, orangutans, cats, bunnies, dragons or chobits in this presentation

- Promise: no authorship attribution, collocation evaluation, non-randomness or Zipf's law
Corpus linguistics in 2018

- Concordance
- Collocation
- Frequency analysis

Ignore for today:
- corpus-based research in theoretical and cognitive linguistics
- method-based research (e.g. Biber dimensions)
- a few innovative studies with statistics/CS experts
Corpus linguistics in 2018

- Today: bigger, faster and more convenient
Corpus linguistics in 2018

- Today: bigger, faster and more convenient

But very successful:
- Birmingham, Lancaster, CASS, Erlangen ;-), ...
- Language description, sociolinguistics & politics, literary stylometrics, language variation, ...
- Transformative for lexicography & language teaching
Research: Age of Austerity

Workshop
Texts and Images of Austerity in Britain. A Multimodal Multimedia Analysis

25.09.-29.09.2017
Research: Multidrug resistance

Research: EFE

EFE – Exploring the Fukushima Effect

Stefan Evert, Fabian Schäfer, Christina Holtz-Bacha, Marc Stamminger

https://www.flickr.com/photos/25058419@N05/5172953200
© 2010 Bjoern Schwarz (CC-by 2.0)
Research: EFE

Transnational Algorithmic Public Sphere

- Germany
  - Mass media: "Public sphere"
  - Social media: "Semi-public sphere"
- Japan
  - Mass media: "Public sphere"
  - Social media: "Semi-public sphere"

Corpus Linguistics
- Automatic text analysis
- Corpus-based CDA

Computer Science
- Network analysis
- Information visualization

Cultural Studies
- Intercultural comparison
- Theory transfer

Communication Science
- Manual comparative analysis
- Quantitative content analysis

Twitter mining

Monthly salience of topics

<table>
<thead>
<tr>
<th>Year</th>
<th>elections</th>
<th>#elections</th>
<th>#nuclear phase-out</th>
<th>nuclear phase-out</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

“distant reading”
Digital humanities by example

Digital humanities software
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)
- Zero-shot learning w/o any training data
- CommAI: raise general AI agent from scratch
Deep learning

Deep learning

Deep learning

https://einstein.ai/research/the-natural-language-decathlon
An illustration
An illustration

Insight

Applications

Corpus

CC BY 4.0 | http://fontawesome.com/
Why corpus linguistics?

- DH draws conclusions from (more or less well-understood) quantitative techniques
- Far removed from object of study (text)
- Often aesthetic appreciation of visualizations
- No established methodology for interpretation

Corpus linguists know how to ...

... combine distant and close reading

... draw conclusions from quantitative data
FET Time Machine proposal

http://timemachine.eu/
Why corpus linguistics?

- AI & deep learning have transformed information technology (Google, Facebook, Microsoft, Apple, ...)
- Focus on end-to-end learning w/o human insight
- Basically: (i) supervised classification or (ii) prediction of observable external quantity

Corpus linguists needed if there is ...

... no clear-cut *a priori* classification scheme
... too little training data for end-to-end learning
Financial narrative processing


- Goals: predict future performance of company & identify red flags by analysis of financial reports etc.
  - needs to be automated for large-scale application
- 1st FNP workshop at LREC 2018
- End-to-end machine learning can predict short-term effects of ad-hoc disclosures – but for wrong reasons
  - exceed+, fall−, improvement+, rise+, weak−, lower−, turbon+, AG+, reason−, expect−, cancer+, ...
  - 48% accuracy (vs. 33%), but would still make money
- Need to read between the lines
  - e.g vague reporting
Financial narrative processing

- Need to read between the lines in FNP
  - vague reporting, focus on performance of industry sector, contradictory sentiment, conflicts with prior reports, ...

- Data: U.S. financial reports (EDGAR 10-k) 2006–2015

![Diagram showing t-SNE projection of linguistic observables into a 2-dimensional space.](image)
Argumentation mining

- Traditional approach: supervised classification
  CLAIM, QUALIFIER, PREMISE, BACKING, REFUTATION, REBUTTAL

- Traditional approach: knowledge mining
  e.g. *we need A because B* (PREMISE $\rightarrow$ CLAIM)

- Difficult to apply to social media such as Twitter
  - short texts don't allow for complex arguments
  - few coherent discussions (reply threads)

tweethead: PLEASE THINK: You have a choice between Independence or being ruled by foreigners for the rest of yours & your childrens lives! Vote BREXIT!
Charnoski: @tweethead A #Brexit vote will finally finish off the UK. I'm generally in favour of an end to the UK so I'm sorely tempted.
73UMP2016: @Charnoski @anncaswell2 @tweethead Absolute crap If we leave the EU will collapse! Its the only way to stop Merkel's genocide plan! #Brexit
Argumentation mining

- **RANT** = Reconstructing Arguments from Noisy Text
- Corpus-linguistic analysis of argumentation in Twitter
  - esp. implicit arguments, opinion instead of reasoning
- Interactive development of corpus-based patterns for extraction & classification of argument expressions
  - *[[Entity]]* (AdvP)? {say/V} {PP}* that
    
    George Osborne says that a Brexit vote will reduce cost ...

  - *[[Human]]* (NEG)? be bothered [about,by] *[[Topic]]*
    
    Cameron ... he won't be bothered about party or country ...
    
    I doubt many professionals will be bothered about the European chair throwing championship.
The future of corpus linguistics
The future of corpus linguistics

1) Interoperability
The future of corpus linguistics

2) Interactivity
2) Interactivity
The future of corpus linguistics

3) Integration

feedback

Insight

Applications

Corpus

DL

Lr

CC BY 4.0 | http://fontawesome.com/
The future of corpus linguistics

Hermeneutic
Cyborg

Insight

Applications
Step 1: Interoperability

- Standard toolbox of CL: CQPweb, SketchEngine, AntConc, WordSmith, MonoConc, LancsBox, ...
  - provide almost the same basic functionality (frequency, collocations, keywords linked to concordance)
  - specific extensions (e.g. collocation networks)

- Slow adoption of innovative techniques (if at all)
- Need better **interoperability** of software tools
  1. data format (TEI, ISO LAF)
  2. corpus query (ISO CQLF)
  3. analysis & visualization (“plug-ins”)
Step 1: Interoperability

- Goal: two-way interoperability
  - query tool ➔ quantitative data ➔ analysis/visualization
  - visualization ➔ concordance in query tool ➔ sort/context

- The Coquery way: tabular data
  - enables Coquery users to carry out flexible analysis within the software itself

- A minimalistic approach to interoperability
  - query tool exports user-defined tabular data
  - user has to understand & specify analysis requirements
  - query tool provides API to link back to concordance lines

https://www.coquery.org/
### Text-feature matrix for multivariate analysis

**CQPweb: Analyse corpus function**

<table>
<thead>
<tr>
<th>id</th>
<th># sent</th>
<th># word</th>
<th>id</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>ncmode</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>ncmode</td>
<td>then_R</td>
<td>42218</td>
<td>1929</td>
<td>thriller</td>
</tr>
<tr>
<td>card_N</td>
<td>42235</td>
<td>ncmode</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>
## Tabular data: Ziggurat corpus

<table>
<thead>
<tr>
<th>#</th>
<th>word</th>
<th>pos</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>A</td>
<td>DET</td>
<td>a</td>
</tr>
<tr>
<td>1</td>
<td>fine</td>
<td>ADJ</td>
<td>fine</td>
</tr>
<tr>
<td>2</td>
<td>example</td>
<td>NN</td>
<td>example</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>PUN</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>Very</td>
<td>ADV</td>
<td>very</td>
</tr>
<tr>
<td>5</td>
<td>fine</td>
<td>ADJ</td>
<td>fine</td>
</tr>
<tr>
<td>6</td>
<td>examples</td>
<td>NN</td>
<td>example</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>PUN</td>
<td>.</td>
</tr>
</tbody>
</table>

*primary layer*
Tabular data: Ziggurat corpus

<table>
<thead>
<tr>
<th>#</th>
<th>word</th>
<th>pos</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>A</td>
<td>DET</td>
<td>a</td>
</tr>
<tr>
<td>1</td>
<td>fine</td>
<td>ADJ</td>
<td>fine</td>
</tr>
<tr>
<td>2</td>
<td>example</td>
<td>NN</td>
<td>example</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>PUN</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>Very</td>
<td>ADV</td>
<td>very</td>
</tr>
<tr>
<td>5</td>
<td>fine</td>
<td>ADJ</td>
<td>fine</td>
</tr>
<tr>
<td>6</td>
<td>examples</td>
<td>NN</td>
<td>example</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>PUN</td>
<td>.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#</th>
<th>type</th>
<th>id</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>text</td>
<td>t0042</td>
</tr>
<tr>
<td>1</td>
<td>s</td>
<td>s42_001</td>
</tr>
<tr>
<td>2</td>
<td>pause</td>
<td>p42_001</td>
</tr>
<tr>
<td>3</td>
<td>s</td>
<td>s42_002</td>
</tr>
</tbody>
</table>

Tree layer ordered by start tag

Primary layer
# Tabular data: Ziggurat corpus

<table>
<thead>
<tr>
<th>#</th>
<th>word</th>
<th>pos</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Stefan</td>
<td>NE</td>
<td>Stefan</td>
</tr>
<tr>
<td>1</td>
<td>likes</td>
<td>VV</td>
<td>like</td>
</tr>
<tr>
<td>2</td>
<td>fluffy</td>
<td>ADJ</td>
<td>fluffy</td>
</tr>
<tr>
<td>3</td>
<td>cats</td>
<td>NN</td>
<td>cat</td>
</tr>
<tr>
<td>4</td>
<td>and</td>
<td>CONJ</td>
<td>and</td>
</tr>
<tr>
<td>5</td>
<td>friendly</td>
<td>ADJ</td>
<td>friendly</td>
</tr>
<tr>
<td>6</td>
<td>dogs</td>
<td>NN</td>
<td>dog</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>PUN</td>
<td>.</td>
</tr>
</tbody>
</table>

# label

<table>
<thead>
<tr>
<th>#</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>subj</td>
</tr>
<tr>
<td>1</td>
<td>obj</td>
</tr>
<tr>
<td>2</td>
<td>obj</td>
</tr>
<tr>
<td>3</td>
<td>mod</td>
</tr>
<tr>
<td>4</td>
<td>mod</td>
</tr>
</tbody>
</table>

graph layer

ordered by tail, then head

primary layer
Step 2: Interactivity

- Step 1 connects analysis result back to concordance
  - corpus linguist interacts with concordance, but for analysis/visualization only control over parameters

- 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
  - could easily be supported by interactive visualization
Interactivity in progress
Step 3: Integration
Step 3: Integration
Integration: MMDA

- Discourse operationalized as topic + attitude
- Topic = node (search term)
- Attitude = set of collocates
- Grouping of collocates in MMDA interfaces provides feedback to visualization
  - modify semantic similarity
  - modify collocation weights
  - second-order collocations

A first draft of MMDA-THC
23 August 2017

Stefan Evert & Philipp Heinrich
Computational Corpus Linguistics Group
FAU Erlangen-Nürnberg

Overview
MMDA-THC stands for Mixed-Methods Discourse Analysis by Thematic Hierarchies of Collocations.¹

¹ The acronym MMDA was chosen because the new method is mood-lifting, empathogenic and has not been approved yet for any human applications.
MMDA: discourse similarity

Merkel
Angela
Vows
Chancellor
oust
says
Hollande
German
no
keep
survive
Jugcker
Bundestag
Frau
German
her
host
center
speedy
ost
fumes
Geller
Pamela
juggles
tells
warns
Blames
Vows
plays
Senior
Chance
Stage
Informal
juncker
Blames
speedy
understands
discuss
urges
oes
wants
By no means
...
MMDA: discourse similarity

- tough stance
- stands
- way
- constructive
- smart
- calm
- careful
- liar
- nasty
- said
- insists
- understands
- stresses
- rules
- vow
- moves
- approach
- stance
- constructive
THANK YOU!