Computational Approaches to Textual Data

Dr. Martin Schweinberger
slides available at
www.martinschweinberger.de
Aims of this talk

- Give an overview of computational approaches to analyzing textual data
- Exemplify computational approaches to analyzing textual data
- Provide information about The Language Technology and Data Analysis Laboratory (LADAL)
Introduction

Corpus Linguistics

Amplification in SLA

LADAL
About me

- Postdoctoral Research Fellow in Language Technologies at UQ’s School of Languages and Cultures
- PhD in English linguistics (U Hamburg, Germany)
- Studied Philosophy, English Philology, and Psychology at U Kassel (Germany) and the National University of Ireland, Galway
- Language Technology Group at the Computer Science department of Universität Hamburg
About me

- Focus on computational approaches to language data with a specialization in statistical modeling.
- Building *The Language Technology and Data Analysis Laboratory* (LADAL).
- Consultant for issues relating to statistics, text analysis, and research design (methodology)
- Concerned with Best Practices and Quality Control in Data Analysis and Research Data Management (as a result of the Replication Crisis)
What I study

Phenomenon: Adjective Amplification

(1) And you just have to hint well then it’s a very good hint (ICE-AUS:S1A-012$A)

(2) They’re all really cheap <#> They’re all really nice, the t-shirts in there (ICE-AUS:S1A-009$B)

(3) It was so bad (ICE-AUS:S1A-044$B)
What I study

Variationist Sociolinguistics

- Language is not homogeneous: variation is ubiquitous
  - Social factors: language use
  - Linguistic variation not random
  - Systematic correlation between certain social factors (age, gender, class, ethnicity, etc.) and language use
- Linguistic differentiation ↔ social stratification
Diffusion of Innovations

young ↔ old

upper

working

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations

young $\leftrightarrow$ old

upper
working
Diffusion of Innovations

young $\leftrightarrow$ old

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations

\[ \text{young} \leftrightarrow \text{old} \]
Diffusion of Innovations

young ←→ old

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations
Diffusion of Innovations

young $\leftrightarrow$ old

upper

working

%

low

high

Construct - i - con | Lexicon
emotional attributive $\leftrightarrow$ non-emotional predicative

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations

young $\leftrightarrow$ old

upper $\leftrightarrow$ working

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations

young $\leftrightarrow$ old

upper

working

high

% ✓

low

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations

young $\leftrightarrow$ old

upper

working

young $\leftrightarrow$ old

high

low

%

innovative

traditional

Construct - i-con | Lexicon

emotional attributive $\leftrightarrow$ non-emotional predicative

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations

young ↔ old

upper

working

young ↔ old

high

% 

low

innovative

traditional

Construct - i - con | Lexicon

emotional

attributive

non-emotional

predicative
Diffusion of Innovations
Computational Approaches to Textual Data

Introduction

Diffusion of Innovations

young ↔ old

upper
class

working

young ↔ old

high

% emotional

Construct - i - con | Lexicon

attributive

non-emotional

predicative

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations
Diffusion of Innovations

young $\leftrightarrow$ old

upper

working

high

%

low

innovative

traditional

Construct - i - con | Lexicon
emotional attributive $\leftrightarrow$ non-emotional predicative

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations

young $\leftrightarrow$ old

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations

- young $\leftrightarrow$ old

- upper $\leftrightarrow$ working

Graph showing:
- young $\leftrightarrow$ old
- innovative vs. traditional
- emotional attributive vs. non-emotional predicative
Diffusion of Innovations

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations

young $\leftrightarrow$ old

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Diffusion of Innovations

young $\leftrightarrow$ old

upper

working

\begin{itemize}
  \item young
  \item old
\end{itemize}

\begin{itemize}
  \item high
  \item low
\end{itemize}

\begin{itemize}
  \item emotional attributive
  \item non-emotional predicative
\end{itemize}

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
Adjective Amplification in Australian English

Figure 1: Adjective Amplification In AusE by age of speaker.
Corpus Linguistics
Corpus Linguistics

What do I use to investigate language change?
→ Corpora
  ▶ Corpora are digitized collections of texts
  ▶ I use transcriptions of spoken conversations that are accompanied by socio-demographic information about the speakers (age, gender, education level, socio-economic status, etc.)

Advantages
  ▶ Cheap and relatively easy to analyze
  ▶ Allow other researchers to check what I have done (full transparency & allows replication)
Corpus Linguistics

What can you do with corpora?

→ e.g. data-driven classification

<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>31</td>
<td>16</td>
<td>69</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>cat</td>
<td>36</td>
<td>38</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>???</td>
<td>66</td>
<td>58</td>
<td>9</td>
<td>34</td>
<td>28</td>
<td>12</td>
</tr>
<tr>
<td>boat</td>
<td>46</td>
<td>21</td>
<td>17</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>cup</td>
<td>59</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>pig</td>
<td>4</td>
<td>15</td>
<td>3</td>
<td>1</td>
<td>7</td>
<td>21</td>
</tr>
<tr>
<td>banana</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>12</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Co-occurrences of selected nouns and verbs in the British National Corpus.
How data-driven classification works

<table>
<thead>
<tr>
<th></th>
<th>knife</th>
<th>cat</th>
<th>???</th>
<th>boat</th>
<th>cup</th>
<th>pig</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>???</td>
<td>.60</td>
<td>.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>boat</td>
<td>.48</td>
<td>.33</td>
<td>.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cup</td>
<td>.76</td>
<td>.59</td>
<td>.58</td>
<td>.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pig</td>
<td>.72</td>
<td>.36</td>
<td>.45</td>
<td>.64</td>
<td>.88</td>
<td></td>
</tr>
<tr>
<td>banana</td>
<td>.71</td>
<td>.57</td>
<td>.47</td>
<td>.60</td>
<td>.72</td>
<td>.64</td>
</tr>
</tbody>
</table>

Table 2: Distance matrix based on co-occurrences of selected nouns and verbs in the British National Corpus.
Cluster Dendrogram

- Height: 0.2, 0.5, 0.8
- Items: banana, pig, cat, ??, knife, boat, cup

mxclust(*, "ward.D")
Q₁

How can this help in the analysis of adjective amplification?

→ Are there meaningful clusters of amplifiers?
Are all adjective amplifiers the same?

(4) That’s very good!
(5) That’s really good!
(6) That’s completely good!
(7) That’s absolutely good!
(8) That’s absolutely amazing!
(9) That’s very amazing!
Data Processing

- Part-of-speech tagged every word
- Extracted all adjectives
- Identified adj. preceded by an amplifier
- Determined the type of amplifier
- Tabulated co-occurrences of amplifiers and adjectives
- Visualized results
Figure 2: Rooted dendrogram showing the clustering of amplifiers in Australian English based on the semantic vector space model.
Figure 3: Unrooted dendrogram showing the clustering of amplifiers in Australian English based on the semantic vector space model.
Figure 4: Results of a correspondence analysis based on amplifiers and their co-occurrences with adjectives in Australian English.
Q₂

What is this good for?

How and where do learners of English (NNS) differ from native speakers (NS) with respect to adjective amplification?
Amplification in Second Language Acquisition (Language Learning and Teaching)
Data

- **International Corpus of Learners of English (ICLE)**
  - 2.5 mil. words representing argumentative writing by intermediate to advanced Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, and Swedish learners of English

- **Louvain Corpus of Native English Essays (LOCNESS)**
  - 290,000 words of argumentative essays by American and British university students and British A-level students
  - LOCNESS was specifically designed to allow meaningful comparisons between the learner data represented in the ICLE.

▶ Processing as described above.
Figure 5: Percentages of amplifiers by L1-background.
Figure 6: Bar graphs showing the difference to expected frequencies of amplifier types based on NS use.
Computational Approaches to Textual Data

The Language Technology and Data Analysis Laboratory (LADAL)

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
LADAL

What is LADAL?

- HASS eResearch support infrastructure for digital HASS at the UQ School of Languages and Cultures
- Targeted at humanities researchers
- Offers pathways into new research possibilities
  - Specialist computing lab for language-based computational and experimental work (the Computational and Experimental Workshop)
  - Online virtual lab (the LADAL website https://slcladala.github.io/index.html)

Dr. Martin Schweinberger, slides available at, www.martinschweinberger.de
LADAL

Services

- Specialized training/support (workshops) on digital research methods and technologies
- Information and self-guided study materials
- Hands-on practical tutorials on topics relating to digital tools, computational methods for data extraction and processing, data visualization, and statistical analyses (learning to “code”)
- Face-to-face consultations
LADAL

Aims of LADAL?

- Development of skills in
  - Digital tools and data management
  - Computational methods and (basic) programming skills
  - Data extraction / transformation / processing / analysis
  - Data visualization (including geo-spatial mapping and interactive web apps)
  - NLP applications (text analysis) and various statistical procedures (including classification and machine learning)
Thank you so, really, very much!
Computational Approaches to Textual Data

Dr. Martin Schweinberger
slides available at
www.martinschweinberger.de