Exploring Diachronic Lexical Semantics with JESEME

Johannes Hellrich\textsuperscript{1,2} & Udo Hahn\textsuperscript{2}

\textsuperscript{1}: Graduate School ‘The Romantic Model’, Friedrich-Schiller–Universität Jena, Jena, Germany
\textsuperscript{2}: Jena University Language & Information Engineering (JULIE) Lab

Motivation

- You shall know a word by the company it keeps! (Firth, 1957)
- State-of-the-art distributional semantics methods for diachronic linguistics
- Accessible for non-experts

http://jeseme.org

Distributional Semantics

- SVD\textsubscript{PPMI} word embeddings for similarity (Levy et al., 2015)
- PPMI and \(\chi^2\) for specific context words
- Relative word frequency

Example Insights

- Meaning of “heart” in COHA shifts from metaphor to medical:
  - Lowered similarity to “soul” and increased similarity to “lungs” and “stroke”
  - Increased \(\chi^2\) for context words “attack” and “disease”
- Better understanding of “electricity” in RSC:
  - Increased similarity with “conductor” and “spark” (controlled experiments) as well as “magnetism” (electromagnetism)
  - Decreased similarity with natural phenomenon “lightning”

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Description</th>
<th>Opaque</th>
<th>Coverage</th>
<th>Tokens</th>
<th>5-grams</th>
<th>Temporal Slices</th>
<th>Words in JeSemE</th>
</tr>
</thead>
<tbody>
<tr>
<td>COHA</td>
<td>genre balanced</td>
<td>no</td>
<td>1830–2009</td>
<td>376-10^6</td>
<td>–</td>
<td>18</td>
<td>5,101</td>
</tr>
<tr>
<td>Deutsches TextArchiv ‘German Text Archive’</td>
<td>representative texts</td>
<td>no</td>
<td>1751–1900</td>
<td>79.6-10^6</td>
<td>–</td>
<td>5</td>
<td>5,338</td>
</tr>
<tr>
<td>Google Book Fiction</td>
<td>fiction</td>
<td>yes</td>
<td>1820–2009</td>
<td>–</td>
<td>14.7-10^6</td>
<td>19</td>
<td>6,492</td>
</tr>
<tr>
<td>Google Book German</td>
<td>mixed</td>
<td>yes</td>
<td>1830–2009</td>
<td>–</td>
<td>5.25-10^6</td>
<td>18</td>
<td>4,449</td>
</tr>
<tr>
<td>RSC</td>
<td>natural science</td>
<td>no</td>
<td>1750–1869</td>
<td>24.7-10^6</td>
<td>–</td>
<td>3</td>
<td>3,080</td>
</tr>
</tbody>
</table>

Processing

- Hyperwords for computing embeddings
- PostgreSQL
- Spark micro web framework with Thymeleaf template engine
- C3.js chart library

This research was conducted within the Graduate School “The Romantic Model. Variation – Scope – Relevance” supported by grant GRK 2041/1 from the Deutsche Forschungsgemeinschaft (DFG). http://www.modellromantik.uni-jena.de