Exploring Diachronic Lexical Semantics with JESEME

Johannes Hellrich1,2 & Udo Hahn2
1: Graduate School 'The Romantic Model',
2: Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena,
Jena, Germany
http://www.julielab.de

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

• SVDPPM word embeddings for similarity (Levy et al., 2015)
• PPMI and χ² 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 χ² 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”

| Corpus                          | Description             | Opaque | Coverage        | Tokens          | 5-grams | Temporal Slices | Words in JeSemE |
|---------------------------------|-------------------------|--------|-----------------|-----------------|---------|-----------------|-----------------|
| Corpus of Historical American English (COHA) | genre balanced          | no     | 1830–2009       | 376-10⁶         | –       | 18              | 5,101           |
| Deutsches TextArchiv ‘German Text Archive’ | representative texts   | no     | 1751–1900       | 79.6-10⁶        | –       | 5               | 5,338           |
| Google Book Fiction             | fiction                 | yes    | 1820–2009       | –               | 14.7-10⁵ | 19              | 6,492           |
| Google Book German              | mixed                   | yes    | 1830–2009       | –               | 5.25-10⁵ | 18              | 4,449           |
| Royal Society Corpus (RSC)      | natural science         | no     | 1750–1869       | 24.7-10⁵        | –       | 3               | 3,080           |

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.modelromantik.uni-jena.de