A resource for the diachronic study of scientific English: Introducing the Royal Society Corpus

Ashraf Khamis, Stefania Degaetano-Ortlieb, Hannah Kermes, Jörg Knappen, Noam Ordan, Elke Teich

Main goal

- Diachronic comparison of different historical stages of scientific text from the mid 17th century to present
- Focus on the relationship between linguistic encoding and information density

Hypothesis

- As scientific activity becomes more specialized over time
  - Particular meanings become more predictable
  - Denser encodings are used to optimize efficiency in communication

Background

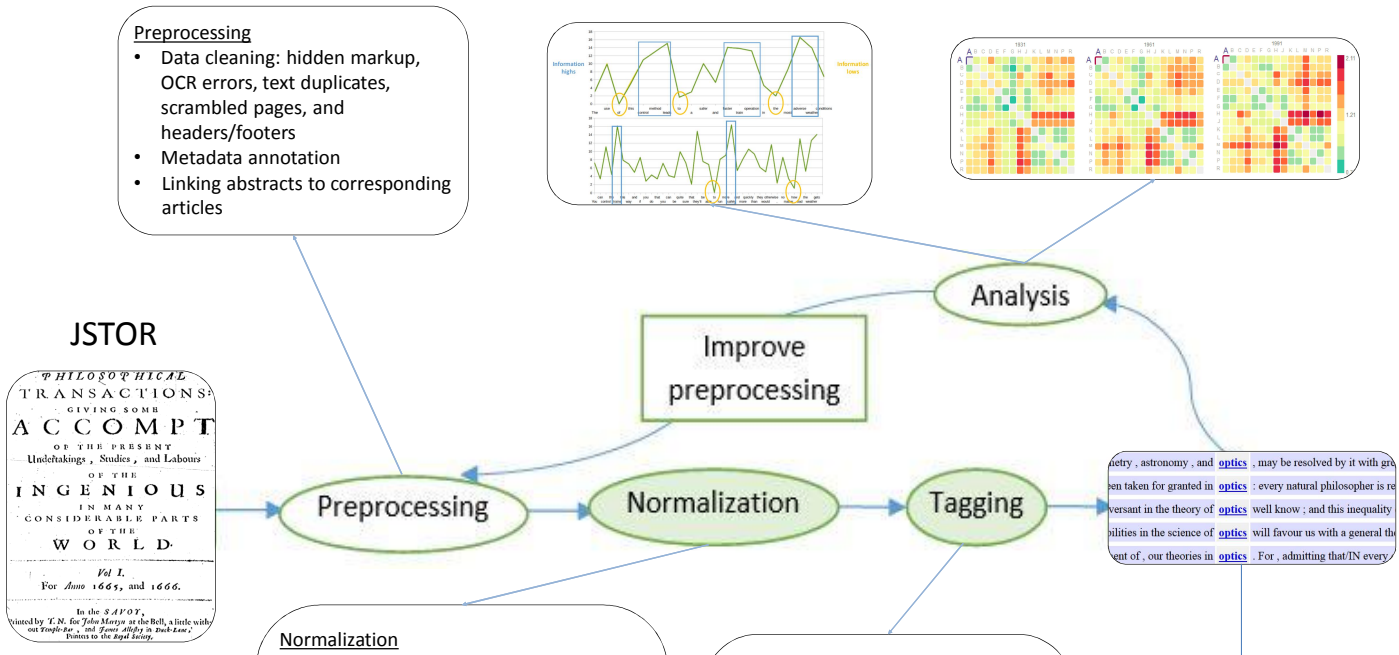
- Available corpora of scientific texts are limited in
  - Size (e.g. the Coruña Corpus (Moskowich and Crespo 2007))
  - Scope (e.g. ARCHER (Biber et al. 1994) and the corpus of Early Modern English Medical Texts (Taavitsainen et al. 2011))
- The Philosophical Transactions as the first scientific journal offers an insight into the formation of scientific language (Atkinson 1998)

Corpus design

Journal	Period	Text type				
		Book reviews	Articles	Miscellaneous	Obituaries	Total
Philosophical Transactions	1665–1678	124	641	154	–	919
Philosophical Transactions	1683–1775	154	3,903	338	–	4,395
Philosophical Transactions of the Royal Society of London	1776–1869	–	2,531	283	–	2,814
Abstracts of Papers Printed in the Philosophical Transactions of the Royal Society of London	1800–1842	–	1,316	15	–	1,331
Abstracts of Papers Communicated to the Royal Society of London	1843–1861	–	429	5	–	434
Proceedings of the Royal Society of London	1862–1869	–	1,476	38	14	1,528
Total		278	10,296	833	14	11,421

Material Transactions and Proceedings of the Royal Society of London
Period 1665–1869
Register Multi-disciplinary scientific writing (e.g. biology, chemistry, physics, geography, medicine)
Metadata Author, title, journal, year of publication, JSTOR link
Size 34.9 million tokens

Agile corpus building



Preprocessing
- Data cleaning: hidden markup, OCR errors, text duplicates, scrambled pages, and headers/footers
- Metadata annotation
- Linking abstracts to corresponding articles

Normalization
- Normalize spelling variation with VARD (Baron and Rayson 2008)
- Train VARD on a gold-standard set (47,642 tokens)

Evaluation of untrained and trained VARD models
- Precision = 61.8% → 72.8%
- Recall = 31.4% → 57.7%

Error types
- Remaining OCR errors (e.g. *asifit* → *as if it*, *irlstruments* → *instruments*)
- Mathematical equations (e.g. $y = a + bx$)

Tagging
- Use TreeTagger (Schmid 1994) for tokenization, lemmatization, and part-of-speech tagging

Evaluation on original and normalized word forms
- Precision = 94.4% → 95.1%

Main errors
- NN ↔ NP (29.5%)
- WP → WDT (18.0%)
- JJ ↔ N.*/V.* (14.2%)

Corpus encoding
- Encode final version in Corpus Query Processor (CQP) format (Evert 2011)

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG) under grant *SFB 1102: Information Density and Linguistic Encoding,* www.sfb1102.uni-saarland.de