Corpus-Driven Investigation of Language Use, Variation and Change

*Resources, Models, Tools*

Elke Teich
Universität des Saarlandes
Saarbrücken

CLARIN Annual Conference 2019
Leipzig

Das also bedeutet es, wenn behauptet wird: die Lautgesetze an sich sind ausnahmslos.


Auch wer der Lehre von der Ausnahmslosigkeit der Lautgesetze huldigt, ist gezwungen, eine Menge von Einzelfällen anzuerkennen, die er mit dem Gesetz nicht in Uebereinstimmung bringen kann, und findet kein Arg darin, diese Fälle als Ausnahmen zu bezeichnen (d. h. als solche Erscheinungen, welche bis jetzt noch nicht unter ein Gesetz zu bringen sind), und auf der anderen Seite gibt es für denjenigen, der die Möglichkeit beliebiger, d. h. von dem absolut freien Willen abhängiger Ausnahmen behauptet, keine grössere Freude, als wenn es ihm gelingt, solche Ausnahmen zu beseitigen.


Humanist-as-scientist

How to adapt the “scientific method”?

How to integrate macro- and micro-analysis?

Discipline

Linguistics

Literary Studies

Cultural Studies

History

Method

Resources

Processes

Metadata

Information

Statistics

Machine learning

Information theory

https://laits.utexas.edu/crit/home
Language Use, Variation and Change

Research questions

• What are the mechanisms of language variation and change?
• Which linguistic features are involved in change?
• How does change proceed?
• What are the effects of change?

(Weinreich/Labov/Herzog 1968)
Resources, Models, Tools

Methodological questions

• Which kinds of resources (corpora) are needed?
• Which computational models are suited?
• Which tools are needed to support the analytic tasks?
Today’s Talk

• Diachronic Development of Scientific English
• Resources; Research Questions, Analytic Tasks, Models
• Selected Analyses and Results
• Summary and Envoi
Diachronic Development of Scientific English
Diachronic Development of Scientific English

- diversification
- specialization
- standardization

What are the linguistic reflexes?

This I did with much solicitude further inquire into; whereupon I found not only one hollowness, but as often as I cut the Nerve asunder, the hollowness still continued therein, and I found in some places not only one cavity, but two or three cavities at once;

In contrast to the complete and temporary visual motion blindness which occurs during stimulation of V5, a less-prominent interference with the perception of visual motion occurs at 70–80 ms after the onset of the visual stimulus when TMS is applied to V1.
In contrast to the complete and temporary visual motion blindness which occurs during stimulation of V5, a less-prominent interference with the perception of visual motion occurs at 70–80 ms after the onset of the visual stimulus when TMS is applied to V1.
Observations and hypothesis

- diversification
- specialization
- standardization

condensed linguistic forms
linguistic patterns
terminology

Optimal code for scientific communication
Back to research questions

Optimal code for scientific communication

(1) What are the linguistic features involved in change?
   → discover typical, distinctive features

(2) How does change proceed?
   → capture course of development

(3) What are the effects of change?
   → observe communicative and register forming effects
Resources

Research Questions, Analytic Tasks, Models
Resources: Royal Society Corpus (RSC)

- Philosophical Transactions and Proceedings of the Royal Society of London
- sources: JSTOR/Royal Society (XML)
- versions up to 4.0: 1665–1869
- size: 35 million tokens, ca. 11,000 texts
- upcoming: extensions up to 1920s (free) and 1990s (restricted); ca. 300 mill. tokens
- meta-data: time, author, no discipline!
- 1-, 10-, 50-year time periods
- CQP-encoded
- annotations: token, lemma, PoS

(Kermes et al. 2016@LREC)
**Research questions, analytic tasks and models**

<table>
<thead>
<tr>
<th>RQ</th>
<th>Task</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) What are the features involved?</td>
<td>Detect features</td>
<td>Relative Entropy (KLD)</td>
</tr>
<tr>
<td>(2) How does change proceed?</td>
<td>Capture phases</td>
<td>Average Surprisal</td>
</tr>
<tr>
<td>(3) What are the effects of change?</td>
<td>Analyze effects</td>
<td>Word embeddings (Wang2Vec)</td>
</tr>
</tbody>
</table>

\[ D(A||B) = \sum_i p(item_i|A) \log \frac{p(item_i|A)}{p(item_i|B)} \]

\[ AvS(item) = \frac{1}{|item|} \sum_{i=1}^{n} -\log p(item_i|item_{i-n}) \]
Detect features

Typical linguistic features
- representative of a time period
- distinct from other time periods

Model
- uni-grams (words), 50-year periods
- Relative Entropy (Kullback-Leibler Divergence; KLD)

$$D(A||B) = \sum_i p(item_i|A)\log_2\frac{p(item_i|A)}{p(item_i|B)}$$

- e.g. encode 1850 with optimal code for 1650 (and vice versa)
Capture phases

Method: Slide KLD over time line

Overall trend by KLD

Single features by pointwise KLD

(Degaetano-Ortlieb & Teich 2018@LaTeCH)
Analyze effects: communication

Model: *Surprisal* (4gram, averaged across 10-year time periods)

\[
AvS(\text{unit}) = \frac{1}{|\text{unit}|} \sum_{i=1}^{n} -\log_2 p(\text{unit}_i|\text{unit}_{i-1} \text{ unit}_{i-2} \text{ unit}_{i-3})
\]

Examples

**ADJ ADJ NOUN**: diluted vitriolic acid

**NOUN of NOUN**: oxide of iron
Analyze effects: register / subsystem

1670-79

1860-69

Model: *Word Embeddings* (Wang2Vec, 10-year time periods)

Selected Analyses and Results
Typical features: Scientific language over time

1650s (RSC)

- pronouns, conjunctions
- reporting genre

1850s (RSC)

- nominal markers, lexical words
- expository genre

Color: RelFreq, Size: KLD score
Phasing of change: Micro-analysis

1791: technical terms

- **oxygen**
- **hydrogen**

*hydrogen (inflammable air)*
Henry Cavendish (1766)

*oxygen (dephlogisticated air)*
Joseph Priestley (1774)
Phasing of change: Micro-analysis

PoS trigram contribution to KLD peak in 1791
(20-year windows, 2-year slider)

PoS trigrams:
prepositional (15)
nominal (4)
passive/relational (3)
relative clause (2)
Phasing of change: Micro-analysis

**ADJ ADJ NOUN: lexical realizations**

<table>
<thead>
<tr>
<th>period</th>
<th>examples</th>
<th>freq. (pM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1650s</td>
<td>dark brown colour</td>
<td>7 (2.70)</td>
</tr>
<tr>
<td></td>
<td>next foregoing tract</td>
<td>6 (2.32)</td>
</tr>
<tr>
<td></td>
<td>cold fair weather</td>
<td>4 (1.55)</td>
</tr>
<tr>
<td>1750s</td>
<td>obedient/obliged humble servant</td>
<td>135 (23.25)</td>
</tr>
<tr>
<td></td>
<td>light inflammable air</td>
<td>110 (17.47)</td>
</tr>
<tr>
<td></td>
<td>diluted vitriolic acid</td>
<td>29 (6.96)</td>
</tr>
<tr>
<td>1850s</td>
<td>concentrated sulphuric acid</td>
<td>104 (8.93)</td>
</tr>
<tr>
<td></td>
<td>carbonic acidic gas</td>
<td>64 (5.50)</td>
</tr>
<tr>
<td></td>
<td>complete differential coefficient</td>
<td>49 (4.21)</td>
</tr>
</tbody>
</table>

1751: Royal Society starts reviewing process ➔ *standardization*

- from general to specific
- pattern for terminology
Phasing of change: Macro-analysis

**lexical level (lemmas)**

**grammatical level (PoS-trigrams)**

- lexical usage goes in waves (innovation vs. conservation)
- grammatical usage is consolidated over time
Effects of change: communication

Surprisal of NOUN of NOUN

- Surprisal going down over time, then leveling out

Average Surprisal of oxide of iron

- Stable average surprisal over time
Examples

(ex1)  
A_5.306 little_6.066 green_11.011 or _6.038 blackish_11.741 
oxide_12.587 of_1.894 copper_10.546 adhered_13.363 to_3.004 
their_6.937 surfaces_10.713.
(1796, G. Pearson, Observations on Some Ancient Metallic Arms and Utensils)

(ex2)  
The_2.895 oxide_7.169 of_1.536 iron_5.131, _2.224 precipitated_8.235 
(1802, E. Howard, Experiments and Observations on Certain Stony and Metalline Substances)
Effects of change: register/system

- paradigmatic productivity of *ing*-forms increases

**1700s**
- examining, observing, discovering, showing, effecting, deciding, describing, dissecting, comparing, making, knowing

**1800s**
- showing, assuming, believing, conceiving, investigating, calculating, admitting, assuming, proving, examining, exhibiting, indicating, showing, establishing, demonstrating, arguing, investigating, observing, measuring, considering, deciding, watching, confirming, detecting, estimating, calculating, accounting, embracing

bubble size: $\sqrt{\text{relative frequency}}$
Effects of change: register/system

paradigmatic productivity of *wh-relativizers* decreases
Summary and Envoi
Summary

Language Use, Variation and Change
• What are the mechanisms of language variation and change?
• What are the linguistic features involved?
• How does change proceed?
• What are the effects of change?

Resources, Models, Tools
• Which kinds of resources (corpora) are needed?
• Which computational models are suited?
• Which tools are needed to support the analytic tasks?
Summary

Diachronic Development of Scientific English

• Based on relative entropy and surprisal
  – lexis shows peaks and troughs over time → INNOVATION
  – typical syntactic patterns (e.g. N of N, ADJ ADJ N) develop over time → CONVENTIONALIZATION

• Based on diachronic word embeddings
  – sets of words become more/less productive paradigmatically
  – specific vocabularies develop → SPECIALIZATION/DIVERSIFICATION
Summary

How to adapt the “scientific method”?

**Discipline**
- Linguistics
- Literary Studies
- Cultural Studies
- History

**Methods**
- Statistics
- Machine learning
- Information theory

**Infrastructure**
- Research
- Processes
- Models

How to integrate macro- and micro-analysis?
Application in historical / cultural analysis

Chemical revolution
Lavoisier's theory of oxygen replacing the former phlogiston theory

Envoi

Methods
• Model quality: up-/down-sampling, randomized control tests, simulation
• Model interpretability: (interactive) visualization

Community
• Culture & Technology European Summer University (ESU) / CLARIN-D
• Computational Approaches to Historical Language Change (Workshop @ ACL 2019)
• Computational Socio-Linguistics (special issue *Frontiers*)
• SPP Computational Literary Studies (DFG *Schwerpunktprogramm*)
Envoi

• Research questions, interpretation: humanistic tasks
• Tools, models: computational tasks
• Don’t leave the data to the data scientist
• Team up ...

Become a “data humanist”!
Thanks to team, collaborators and sponsors

Yuri Bizzoni  Stefania Degaetano-Ortlieb
Stefan Fischer  Tom S. Juzek
Jörg Knappen  Pauline Krielke
Katrin Menzel  &
Peter Fankhauser
(Leibniz-IDS Mannheim)
References


• Stefania Degaetano-Ortlieb and Elke Teich, 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, COLING 2018, pages 22–33, Santa Fe, NM, USA.


