Corpus-Driven Investigation of Language Use, Variation and Change

Resources, Models, Tools

Elke Teich

Universität des Saarlandes

Saarbrücken

CLARIN Annual Conference 2019

Leipzig

Suche im Deutschen Textarchiv

Treffer 101 - 110 von 177

Neue Suche · KWIC Lautgesetze suchen Hilfe

10 Treffer pro Seite Sortierung: Datum aufsteigend/absteigend · zufällig

gehe zu: Anfang · -10 · -5 · vorherige · nächste · +5 · +10 · Ende

101: Delbrück, Berthold: Die neueste Sprachforschung. Betrachtungen über Georg Curtius Schrift zur Kritik der neuesten Sprachforschung. Leipzig, 1885. #24 [mehr]

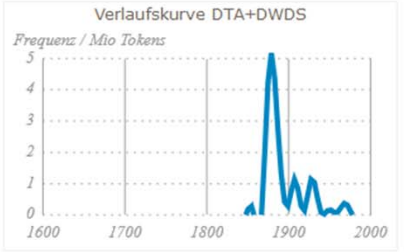
Das also bedeutet es, wenn behauptet wird: die **Lautgesetze** an sich sind ausnahmslos.

102: Delbrück, Berthold: Die neueste Sprachforschung. Betrachtungen über Georg Curtius Schrift zur Kritik der neuesten Sprachforschung. Leipzig, 1885. #25 [mehr]

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 giebt 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.

103: Delbrück, Berthold: Die neueste Sprachforschung. Betrachtungen über Georg Curtius Schrift zur Kritik der neuesten Sprachforschung. Leipzig, 1885. #30 [mehr]

Curtius hatte in seinem Aufsatz über die Tragweite der **Lautgesetze** gemeint, dass das ι in δοίην erhalten, in ποέω dagegen verschwunden sei, weil es in δοίην als bedeutungstragend empfunden wurde, in ποιέω aber nicht.



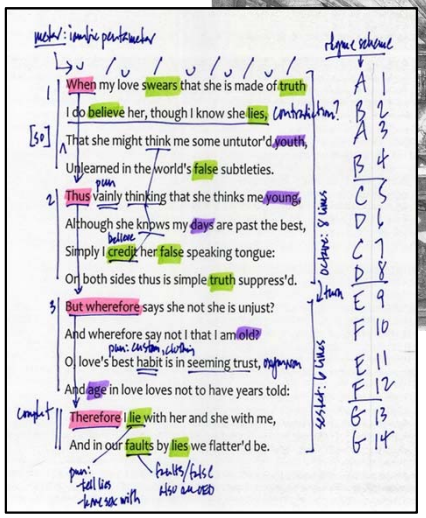
Suchergebnisse herunterladen: Text, Text/KWIC, JSON, YAML, ATOM 1.0, RSS 2.0, TCF 0.4.

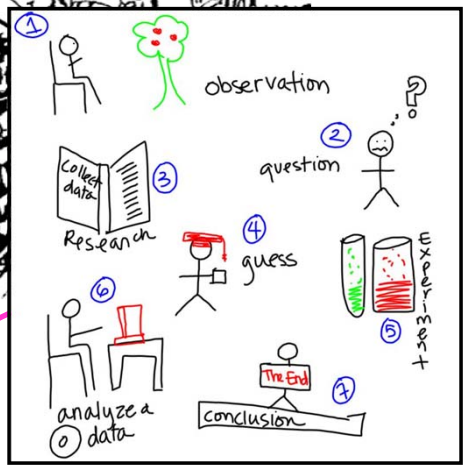
Fahren Sie über die einzelnen Tokens mit der Maus, um folgende Informationen zu sehen:

- `u:` Originaltext, UTF-8-kodiert
- `w:` approximierter Latin-1-Text
- `v:` CAB-normalisierte Wortform
- `l:` Lemma (unflektierte Form)
- `p:` Part-of-Speech-Analyse

Humanist-as-scientist

How to adapt the “scientific method”?





Linguistics
Literary Studies
Cultural Studies
History

Methods

Resources
Processes
Metadata

Statistics
Machine learning
Information theory

How to integrate macro- and micro-analysis?

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)

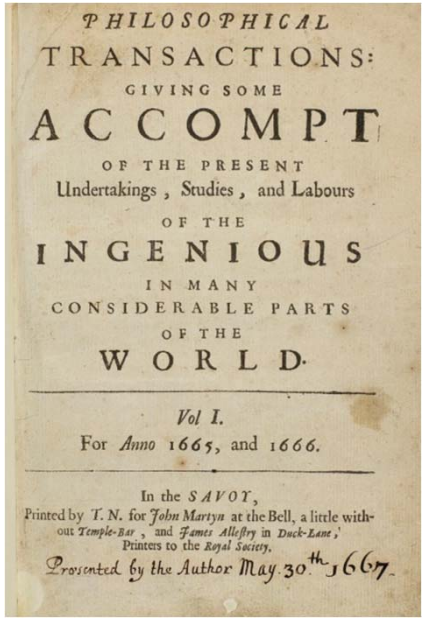
Resources, Models, Tools

Methododological 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


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.
Presented by the Author May 30th. 1667.

Diachronic Development of Scientific English

Diachronic Development of Scientific English

Late Modern Period (1700-1900)

- diversification
- specialization
- standardization

⟶ What are the linguistic reflexes?

(Halliday 1988, Atkinson 1999, Halliday & Martin 1993, Banks 2008, Biber & Gray 2016)

Historical example: Royal Society (1674)

PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.

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;

Coxe, Daniel. 1674. "A continuation of Dr. Daniel Coxe's Discourse, Touching the Identity of All Volatil Salts, and Vinous Spirits; Together with Two Surprizing Experiments Concerning Vegetable Salts, Perfectly Resembling the Shape of the Plants, Whence They Had Been Obtained". *Philosophical Transactions (1665-1678)* 9. The Royal Society: 169–82.

Historical example: Royal Society (1674)

PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

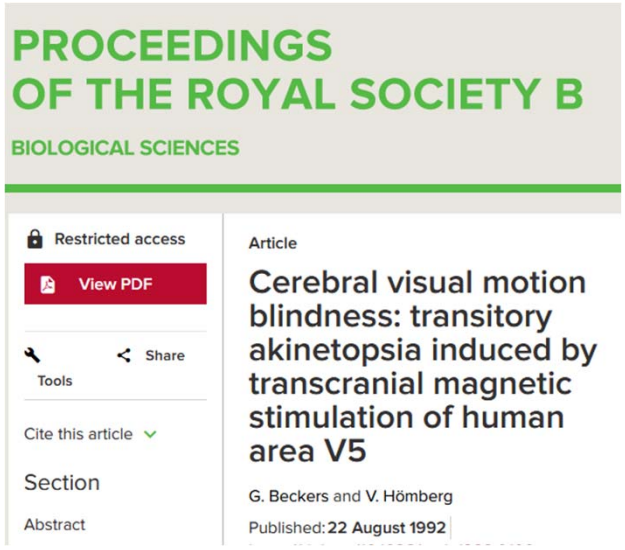
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane;
Printers to the Royal Society.

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;

Coxe, Daniel. 1674. "A continuation of Dr. Daniel Coxe's Discourse, Touching the Identity of All Volatil Salts, and Vinous Spirits; Together with Two Surprizing Experiments Concerning Vegetable Salts, Perfectly Resembling the Shape of the Plants, Whence They Had Been Obtained". *Philosophical Transactions (1665-1678)* 9. The Royal Society: 169–82.

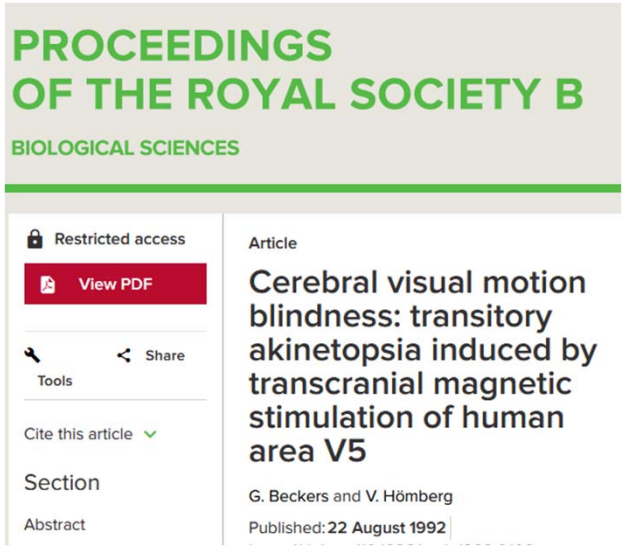
- reporting genre; oral mode
- pronouns, coordinating conjunctions

Contemporary example: Royal Society (1992)



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.

Contemporary example: Royal Society (1992)



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.

- expository genre; written mode
- complex nominal structure
- terminology

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)

https://fedora.clarin-d.uni-saarland.de/rsc/
V4.0: http://hdl.handle.net/21.11119/0000-0001-7E8B-6

Research questions, analytic tasks and models

RQ	Task	Model
(1) What are the features involved?	Detect features	Relative Entropy (KLD) $D(A\|\|B) = \sum_i p(item_i\|A) log_2 \frac{p(item_i\|A)}{p(item_i\|B)}$
(2) How does change proceed?	Capture phases	
(3) What are the effects of change?	Analyze effects	Average Surprisal $AvS(item) = \frac{1}{\|item\|} \sum_{i=1}^{n} -log_2 p(item_i\|item_{i-n})$
		Word embeddings (Wang2Vec)

(Fankhauser et al. 2014@LREC)

Detect features

Typical linguistic features

- *representative* of a time period
- *distinct* from other time periods

Model

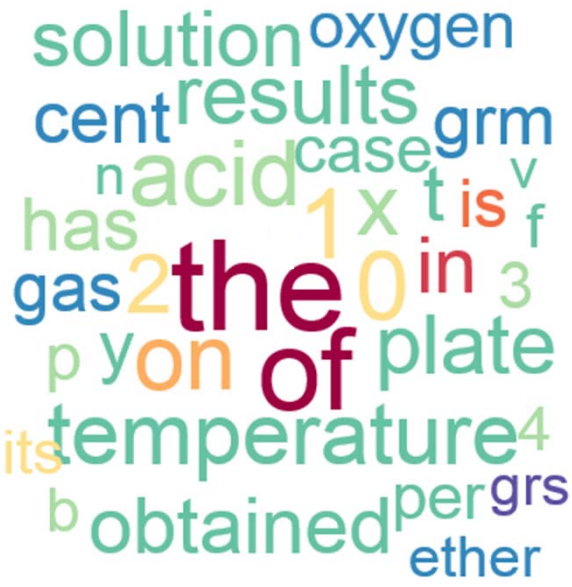
- uni-grams (words), 50-year periods
- *Relative Entropy* (Kullback-Leibler Divergence; KLD)

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

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

Typical of 1650s

Typical of 1850s

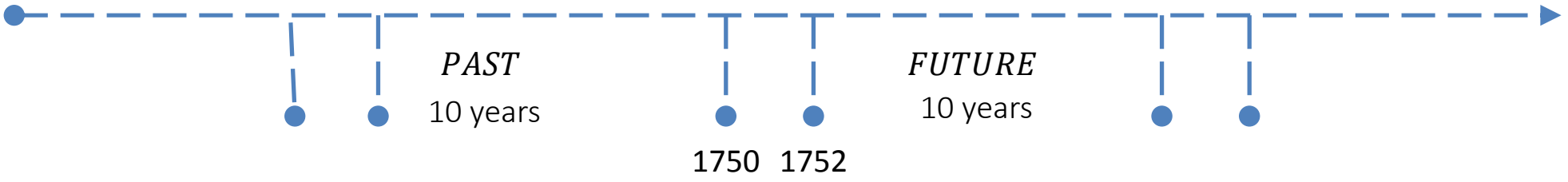


Color: RelFreq, Size: KLD score

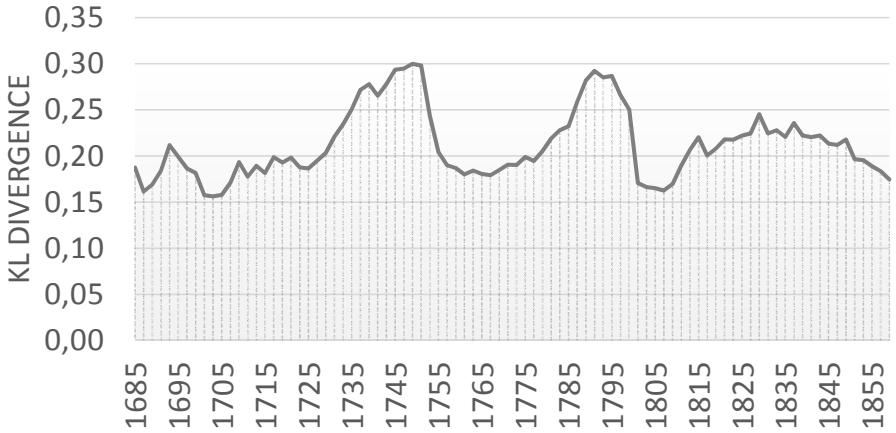
(Degaetano-Ortlieb & Teich 2018@LaTeCH)

Capture phases

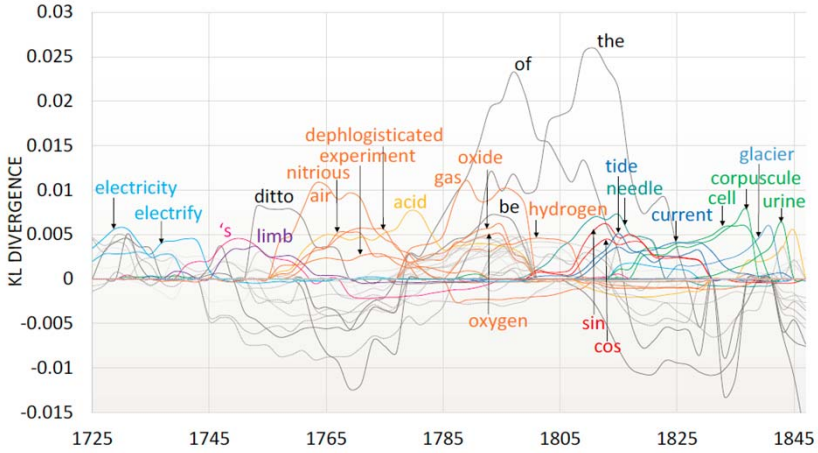
Method: Slide KLD over time line



Overall trend by KLD



Single features by pointwise KLD



(Degaetano-Ortlieb & Teich 2019@CLLT)

Analyze effects: communication

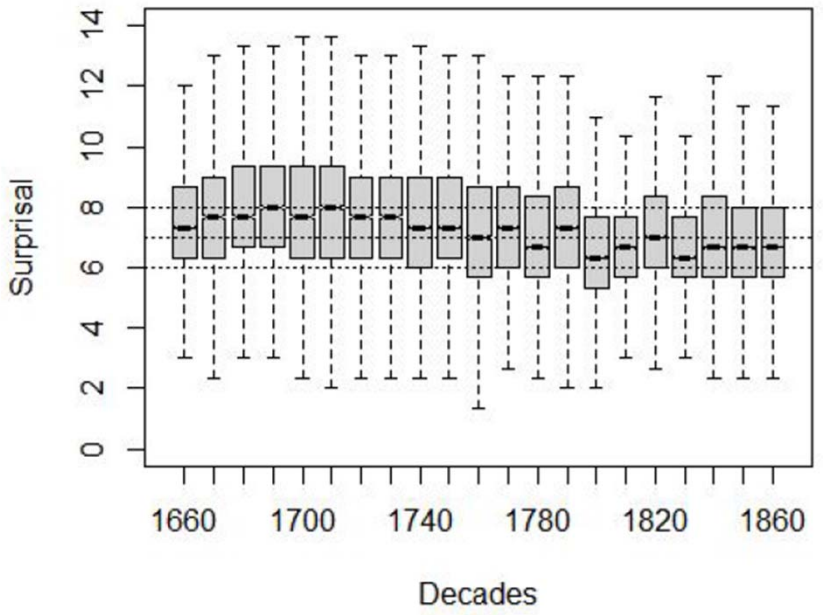
Model: *Surprisal* (4gram, averaged across 10-year time periods)

$$AvS(unit) = \frac{1}{|unit|}\sum_{i=1}^{n} -log_2 p(unit_i | unit_{i-1}\, unit_{i-2}\, unit_{i-3})$$

Examples

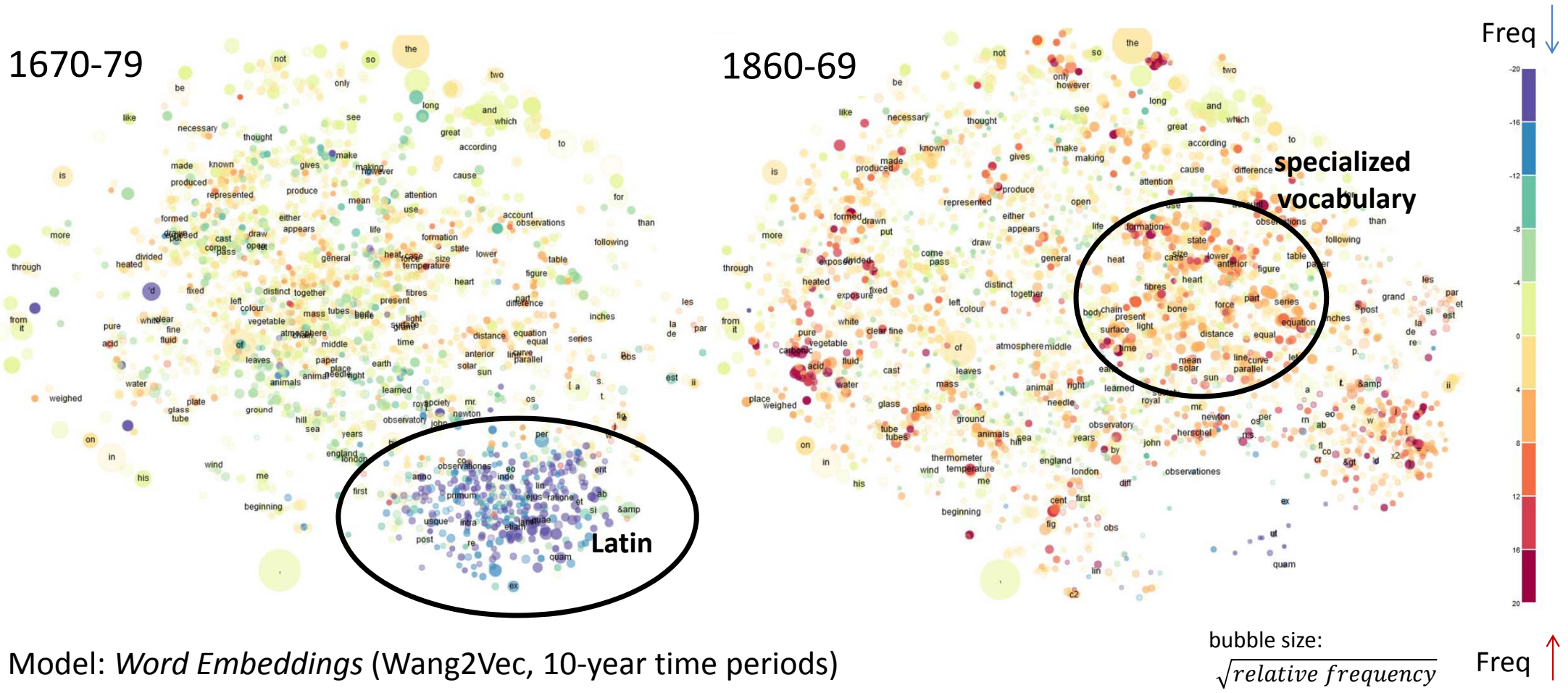
ADJ ADJ NOUN: *diluted vitriolic acid*

NOUN *of* NOUN: *oxide of iron*



(Bizzoni et al. 2019@HistComp-ACL)

Analyze effects: register / subsystem



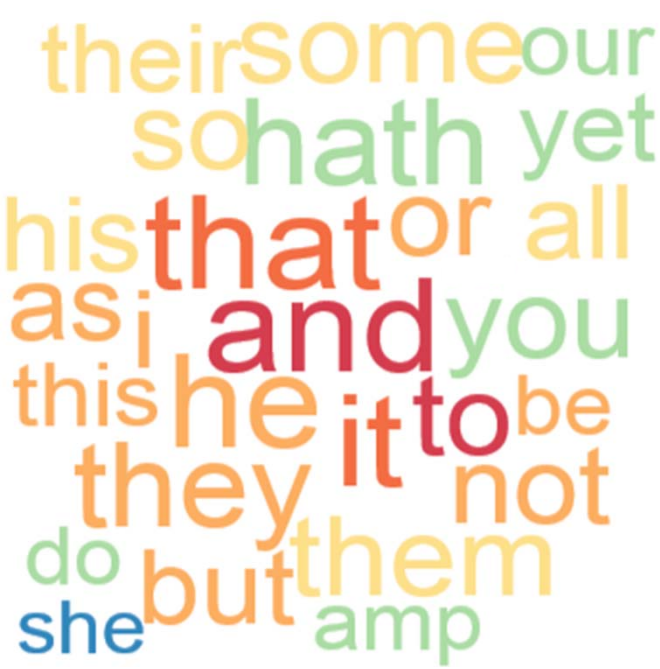
Model: *Word Embeddings* (Wang2Vec, 10-year time periods)

(Fankhauser and Kupietz 2017, Hamilton et al. 2016, Dubossarsky et al. 2017)

Selected Analyses and Results

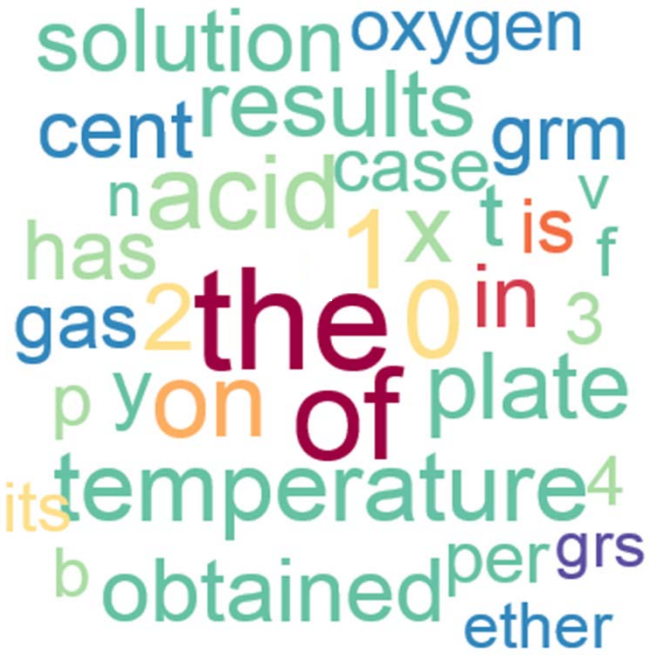
Typical features: Scientific language over time

1650s (RSC)



- pronouns, conjunctions
- reporting genre

1850s (RSC)



- nominal markers, lexical words
- expository genre

Color: RelFreq, Size: KLD score

Phasing of change: Micro-analysis

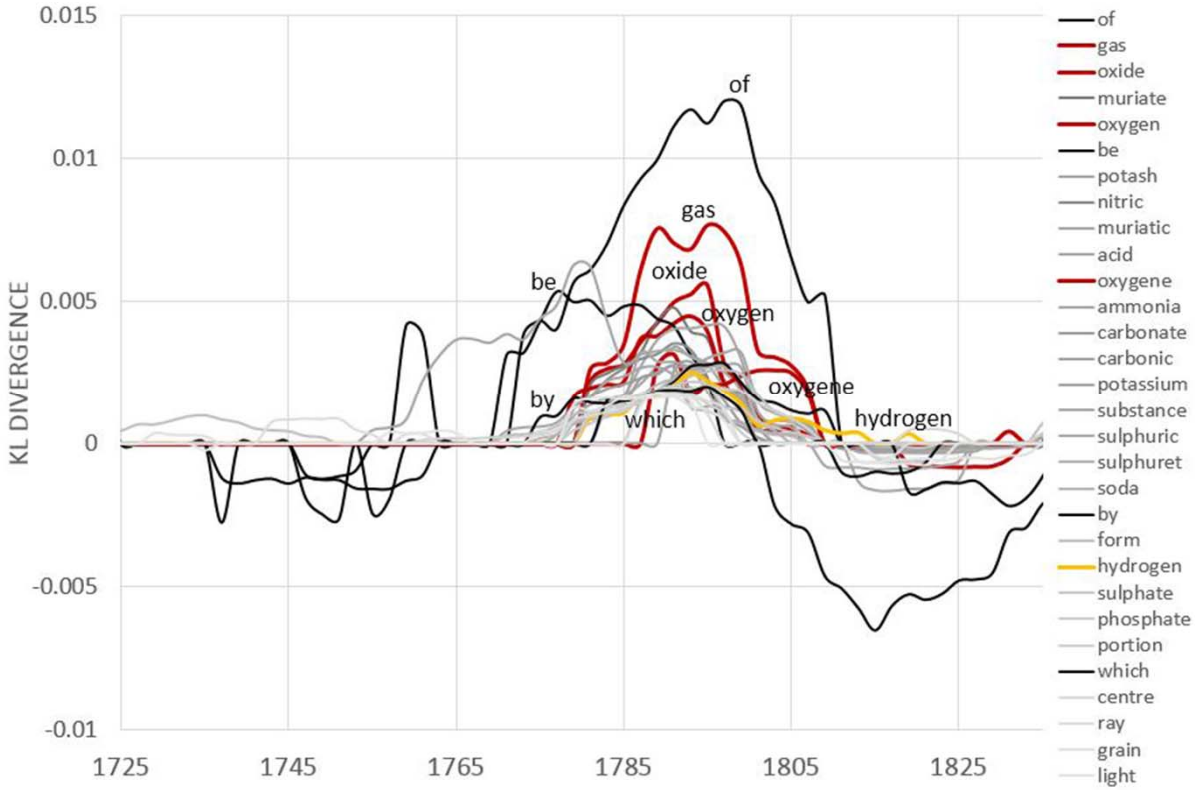
1791: technical terms

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)

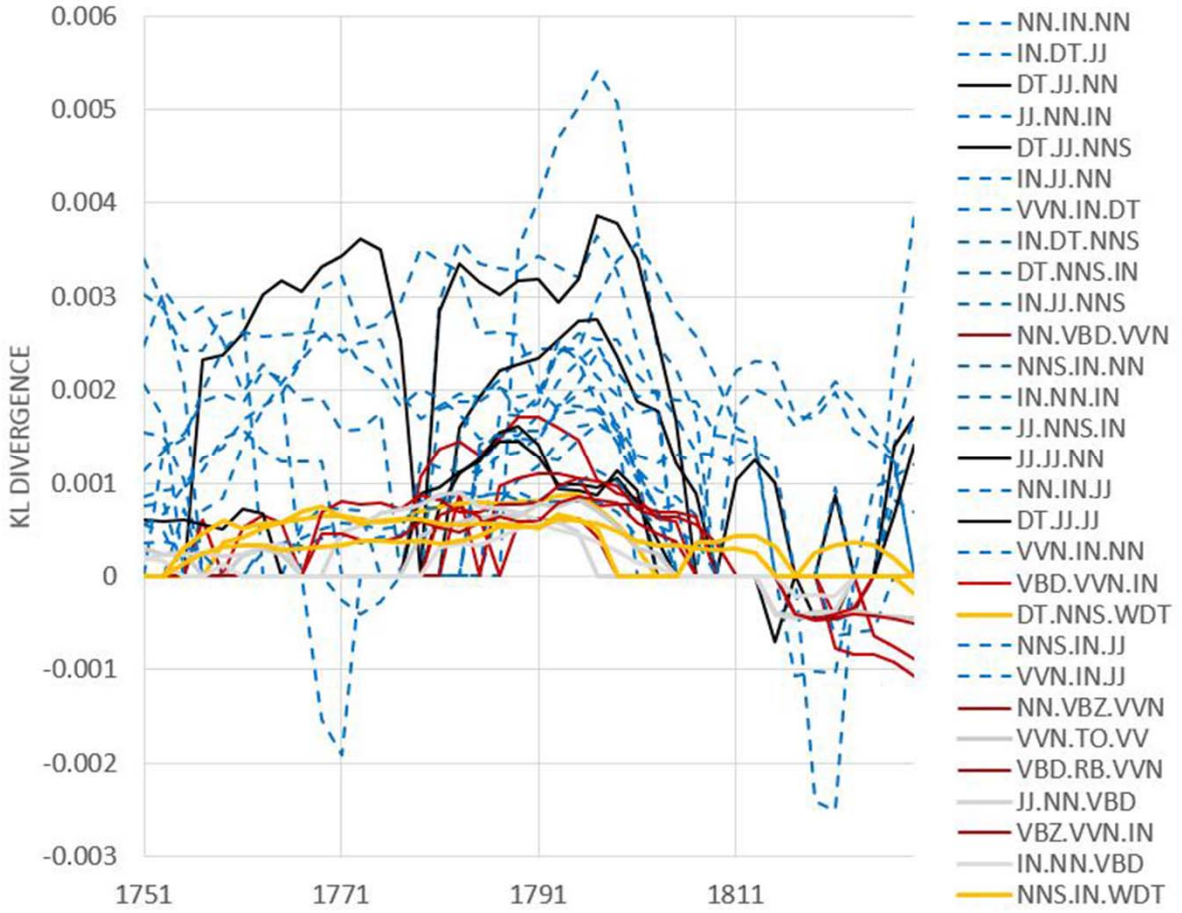
PoS trigrams:

prepositional (15)

nominal (4)

passive/relational (3)

relative clause (2)



Phasing of change: Micro-analysis

ADJ ADJ NOUN: lexical realizations

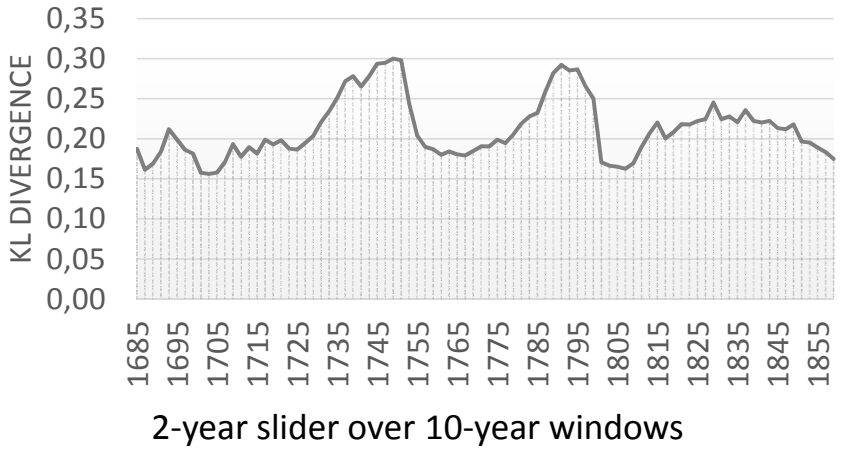
period	examples	freq. (pM)
1650s	*dark brown colour*	7 (2.70)
	next foregoing tract	6 (2.32)
	cold fair weather	4 (1.55)
1750s	*obedient/obliged humble servant*	135 (23.25)
	light inflammable air	110 (17.47)
	diluted vitriolic acid	29 (6.96)
1850s	*concentrated sulphuric acid*	104 (8.93)
	carbonic acidic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

- from general to specific
- pattern for terminology

1751: Royal Society starts reviewing process → *standardization*

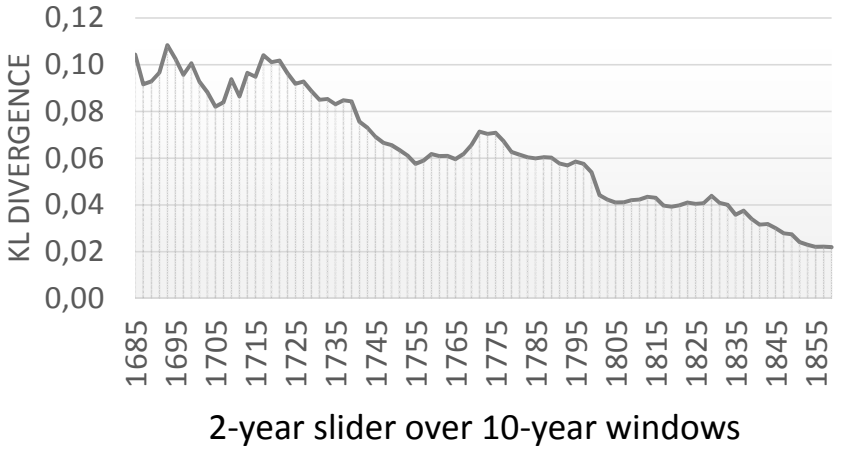
Phasing of change: Macro-analysis

lexical level (lemmas)



2-year slider over 10-year windows

grammatical level (PoS-trigrams)



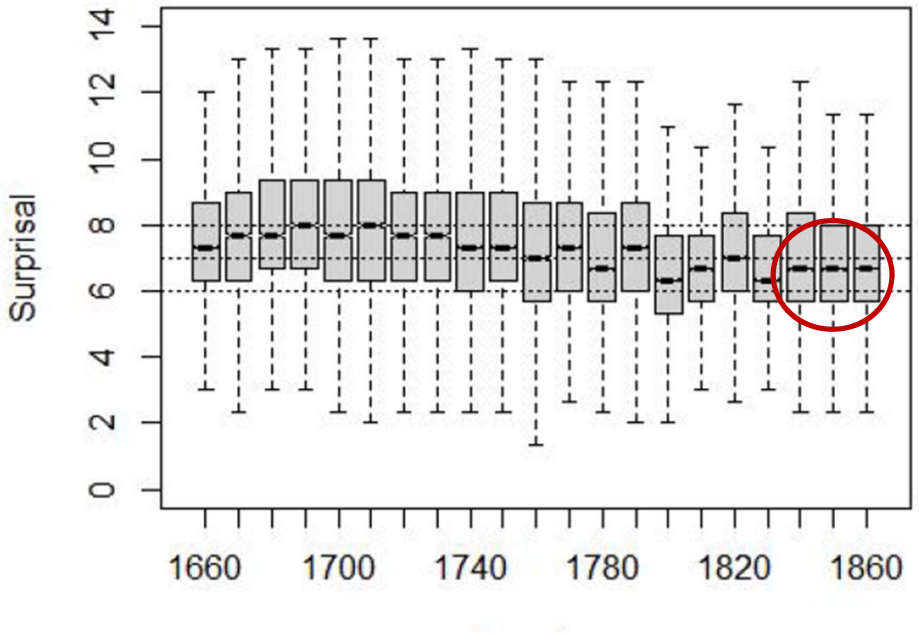
2-year slider over 10-year windows

- lexical usage goes in waves (innovation vs. conservation)

- grammatical usage is consolidated over time

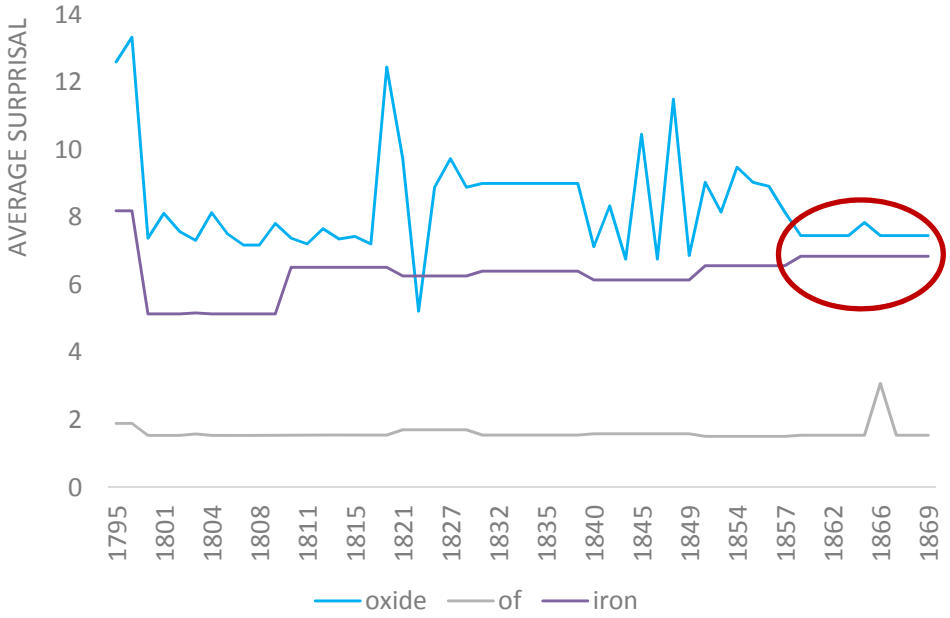
Effects of change: communication

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)

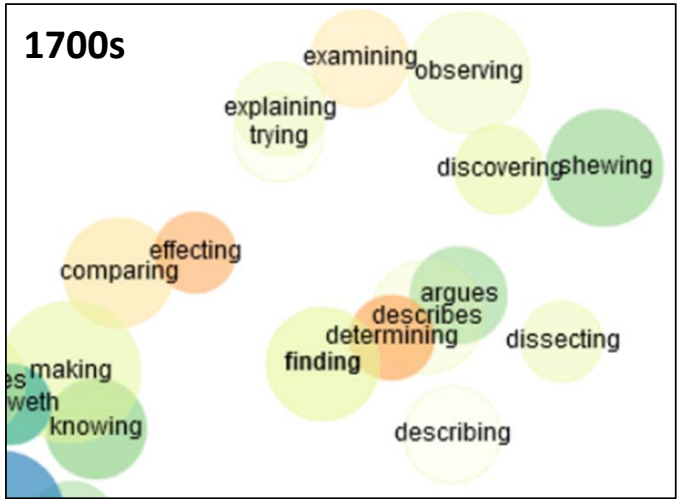
(ex2)

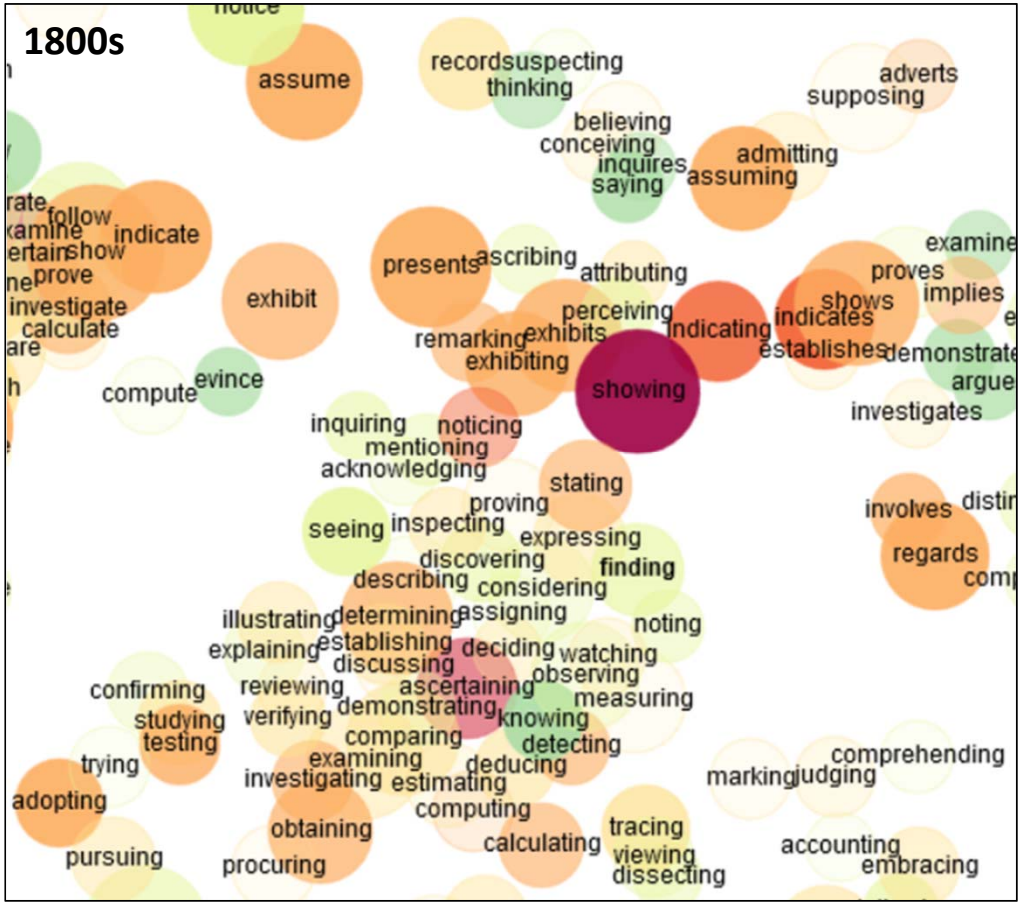
The_2.895 ***oxide_7.169 of_1.536 iron_5.131****,_2.224 precipitated_8.235 by_3.985 ammonia_6.799 ,_2.224 weighed_9.726 8_9.408 grains_6.263.*

(1802, E. Howard, Experiments and Observations on Certain Stony and Metalline Substances)

Effects of change: register/system

- paradigmatic productivity of ***ing*-forms** increases



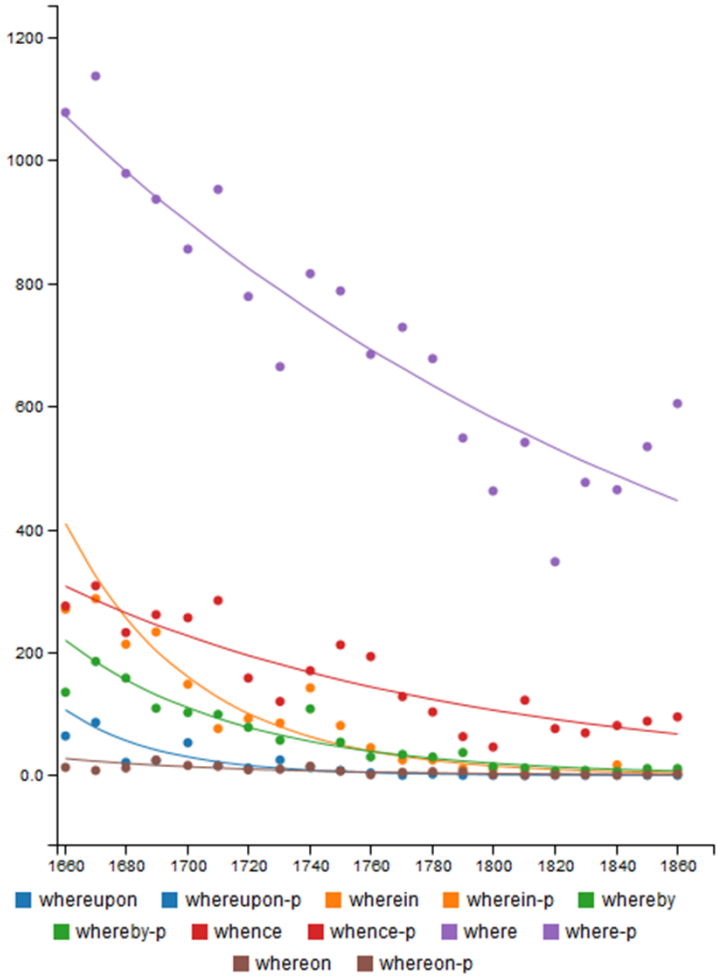


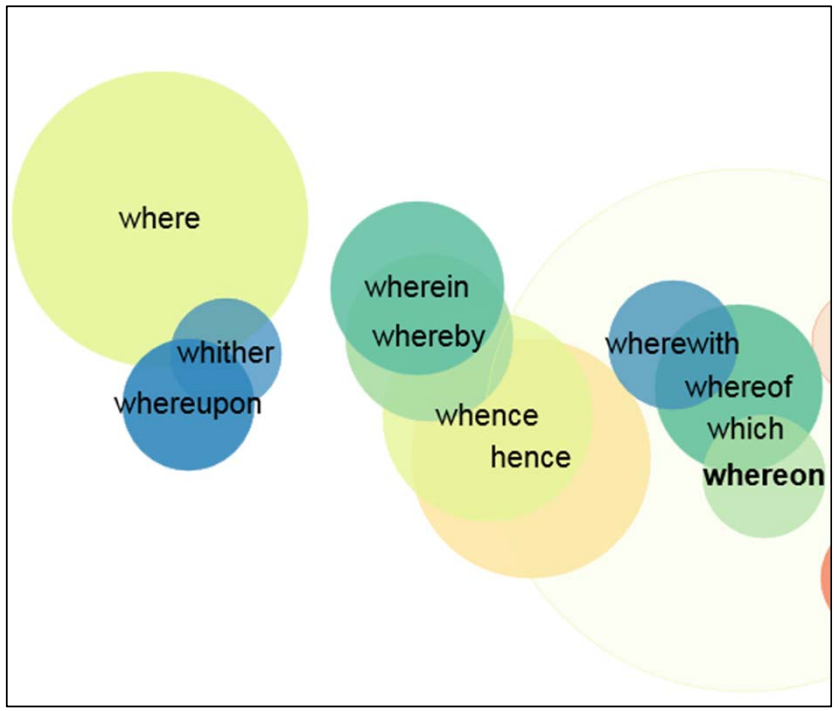
Freq ↓

Freq ↑

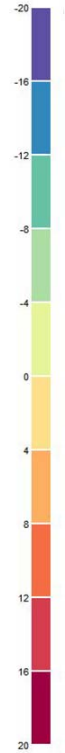
bubble size: $\sqrt{relative\ frequency}$

Effects of change: register/system





Freq ↓



Freq ↑

- paradigmatic productivity of ***wh-relativizers*** decreases

bubble size: $\sqrt{relative\ frequency}$

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?

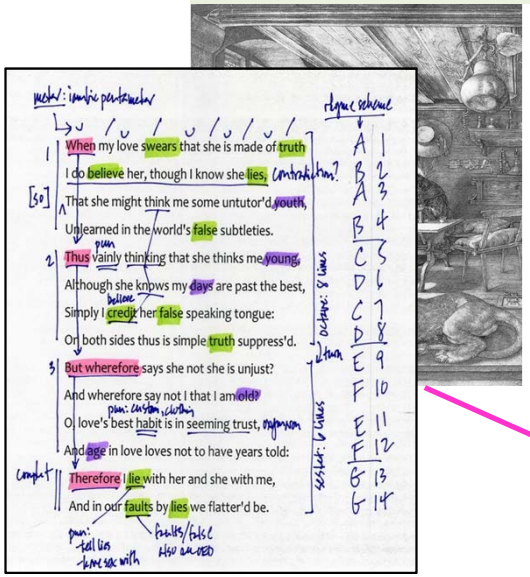
Summary

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

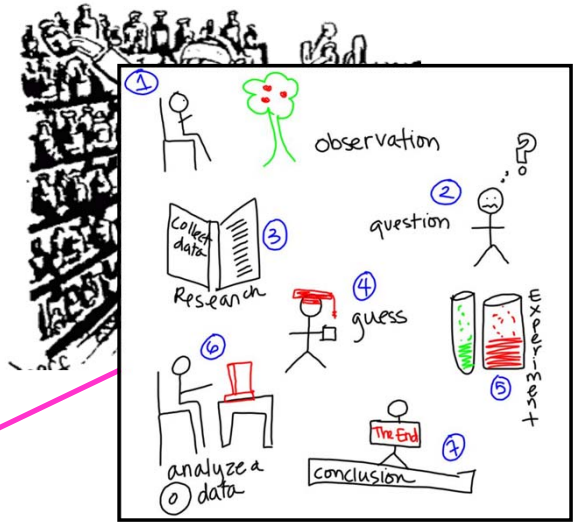
Summary

How to adapt the “scientific method”?



Discipline

Linguistics
Literary Studies
Cultural Studies
History

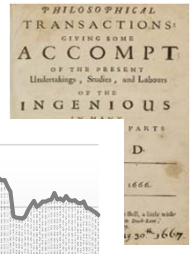


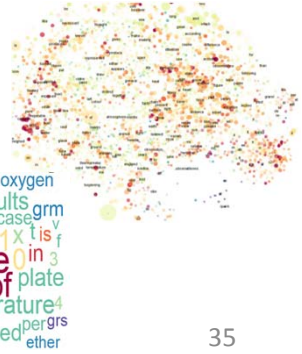
Methods

Res
Pro
Met

Statistics
e learning
ation theory

How to integrate macro- and micro-analysis?





Infrastructure

Models

Application in historical / cultural analysis

Chemical revolution

Lavoisier's theory of oxygen

replacing the former phlogiston theory

https://www.chemistryworld.com/news/the-oxygen-revolution/3003970.article

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“!

CLARIN
Common Language Resources and Technology Infrastructure

Thanks to team, collaborators and sponsors

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

References

- Dwight Atkinson, 1999. *Scientific discourse in sociohistorical context: The Philosophical Transactions of the Royal Society of London, 1675-1975*. Erlbaum, New York.
- David Banks, 2008. *The Development of Scientific Writing: Linguistic Features and Historical Context*. Equinox, London/Oakville.
- Douglas Biber and Bethany Gray, 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing. Studies in English Language*. Cambridge University Press, Cambridge, UK.
- Yuri Bizzoni, Stefania Degaetano-Ortlieb, Katrin Menzel, Pauline Krielke, and Elke Teich, 2019. Grammar and meaning: Analysing the topology of diachronic word embeddings. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, ACL 2019, pages 175–185, Florence, Italy.
- Stefania Degaetano-Ortlieb and Elke Teich, 2016. Information-based modeling of diachronic linguistic change: From typicality to productivity. In Proceedings of the 10th LaTeCH Workshop, ACL 2016, pages 165–173, Berlin, Germany.
- 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.
- Stefania Degaetano-Ortlieb and Teich, Elke, 2019. Towards an optimal code for communication: the case of scientific English. *Corpus Linguistics and Linguistic Theory* (open access), pages 1-33, 2019.

- Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman, 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark.
- Peter Fankhauser, Jörg Knappen, and Elke Teich, 2014. Exploring and visualizing variation in language resources. In Proceedings of the 9th Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland.
- Peter Fankhauser and Marc Kupietz., 2017. Visual correlation for detecting patterns in language change. In Visualisierungsprozesse in den Humanities. Linguistische Perspektiven auf Prägungen, Praktiken, Positionen (VisuHu2017), Zürich, Switzerland.
- M.A.K. Halliday, 1988. On the language of physical science. In Mohsen Ghadessy, editor, *Registers of Written English: Situational Factors and Linguistic Features*, pages 162–177. Pinter, London.
- M.A.K. Halliday and J.R. Martin, 1993. *Writing Science: Literacy and Discursive Power.* Falmer Press, London.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky, 2016. Cultural Shift or Linguistic Drift? Comparing Two Computational Models of Semantic Change. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, Texas.
- Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich, 2016. The Royal Society Corpus: From Uncharted Data to Corpus. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), Portoroz, Slovenia.
- Uriel Weinreich, William Labov, and Marvin I. Herzog, 1968. Empirical foundations for a theory of language change. In W.P. Lehmann and Y. Malkiel (eds.), *Directions for Historical Linguistics*. University of Texas Press, Austin, Texas, pages 95-195.