IDeaL
SFB 1102

UNIVERSITÄT
DES
SAARLANDES

Modeling Registerial Developments with Information Theory: Variation and Change in 300 Years of Scientific Written English

PD Dr. Stefania Degaetano-Ortlieb

Saarland University

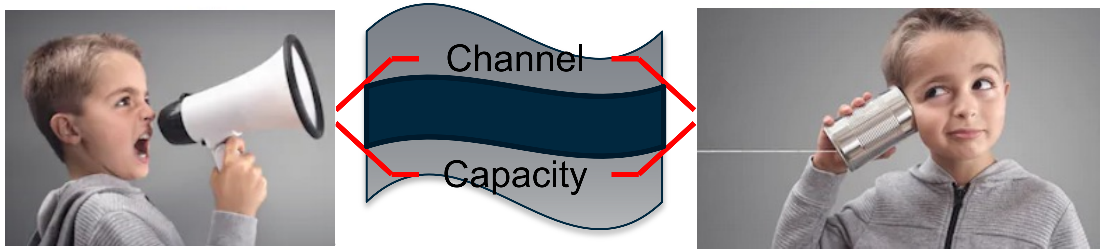
Department of Language Science and Technology

@ICLaVE panel on embracing variability in NLP 10.07.24

COMMUNICATION through language

Assumptions

The *language system* approximates an optimal code for human communication (close to channel capacity)



Language use is rational:
Interlocutors

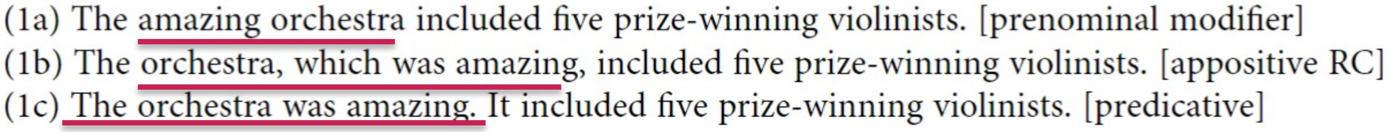
- strive for successful communication
- want to keep effort reasonable

Human language processing is tied to expectancy: predictability in context

Theoretical setting

Language variation and register theory (Halliday 1985, Biber 1988)

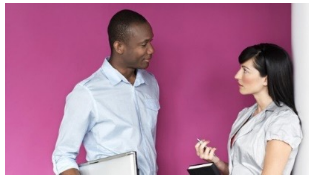
» variation given the local linguistic context

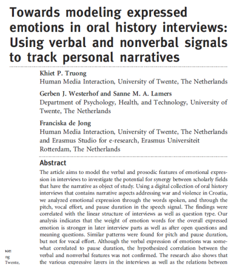


(1a) The amazing orchestra included five prize-winning violinists. [prenominal modifier]
(1b) The orchestra, which was amazing, included five prize-winning violinists. [appositive RC]
(1c) The orchestra was amazing. It included five prize-winning violinists. [predicative]

(Kaiser & Wang 2021)

» language use is determined by the situational context





Towards modeling expressed emotions in oral history interviews: Using verbal and nonverbal signals to track personal narratives

Khiet P. Truong
Human Media Interaction, University of Twente, The Netherlands

Gerben J. Westerhof and Sanne M. A. Lamers
Department of Psychology, Health, and Technology, University of Twente, The Netherlands

Franciska de Jong
Human Media Interaction, University of Twente, The Netherlands and Erasmus Studio for e-research, Erasmus Universiteit Rotterdam, The Netherlands

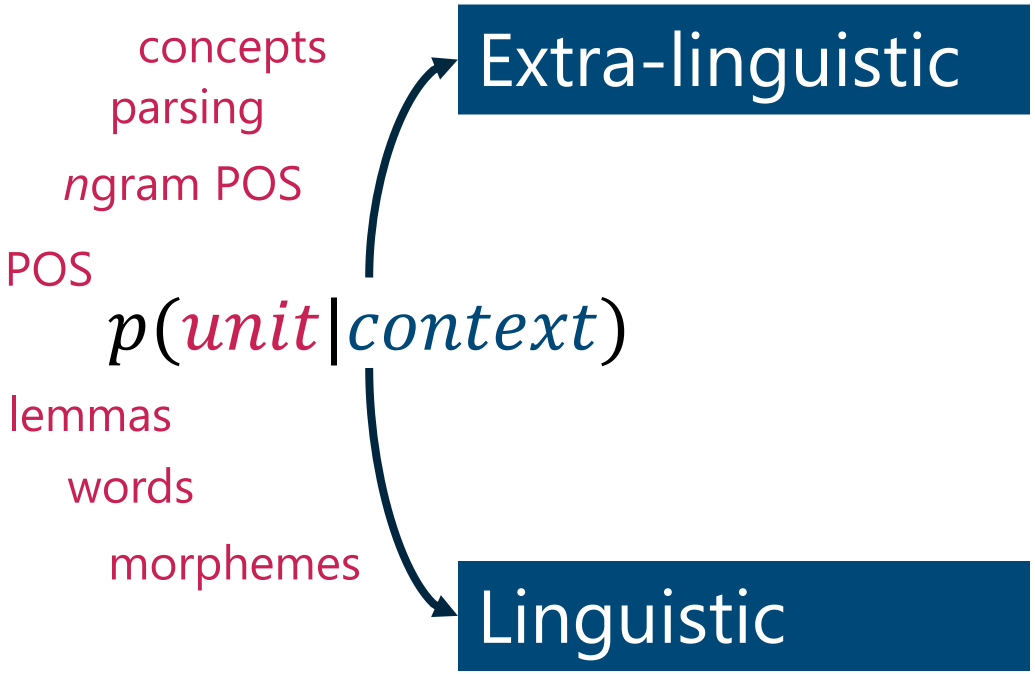
Abstract

The article aims to model the verbal and prosodic features of emotional expression in interviews to investigate the potential for synergy between scholarly fields that have the narrative as object of study. Using a digital collection of oral history interviews that contains narrative aspects addressing war and violence in Croatia, we analyzed emotional expression through the words spoken, and through the pitch, vocal effort, and pause duration in the speech signal. The findings were correlated with the linear structure of interviews as well as question type. Our analysis indicates that the weight of emotion words for the overall expressed emotion is stronger in later interview parts as well as after open questions and meaning questions. Similar patterns were found for pitch and pause duration, but not for vocal effort. Although the verbal expression of emotions was somewhat correlated to pause duration, the hypothesized correlation between the verbal and nonverbal features was not confirmed. The research also shows that the various expressive layers in the interviews as well as the relations between

Rational communication and information theory

- usage-based and communicative perspective (Bybee 2007, Aitchinson 2008, Kirby et al. 2015, Crocker et al. 2016)
- variation helps modulate the information content leading to optimization effects for efficient communication (Jaeger and Levy 2007, Piantadosi et al. 2011)

Probability and context



concepts

parsing

*n*gram POS

POS

$p(unit|context)$

lemmas

words

morphemes

Extra-linguistic

Linguistic

→ detect variation across situational contexts (e.g. time, registers, authors) with relative entropy

→ analyze variation in linguistic context

$syntagmatic\ context$

$paradigmatic\ context$

Scenario: Scientific writing

In collaboration with people from SFB1102, Project B1

Elke Teich

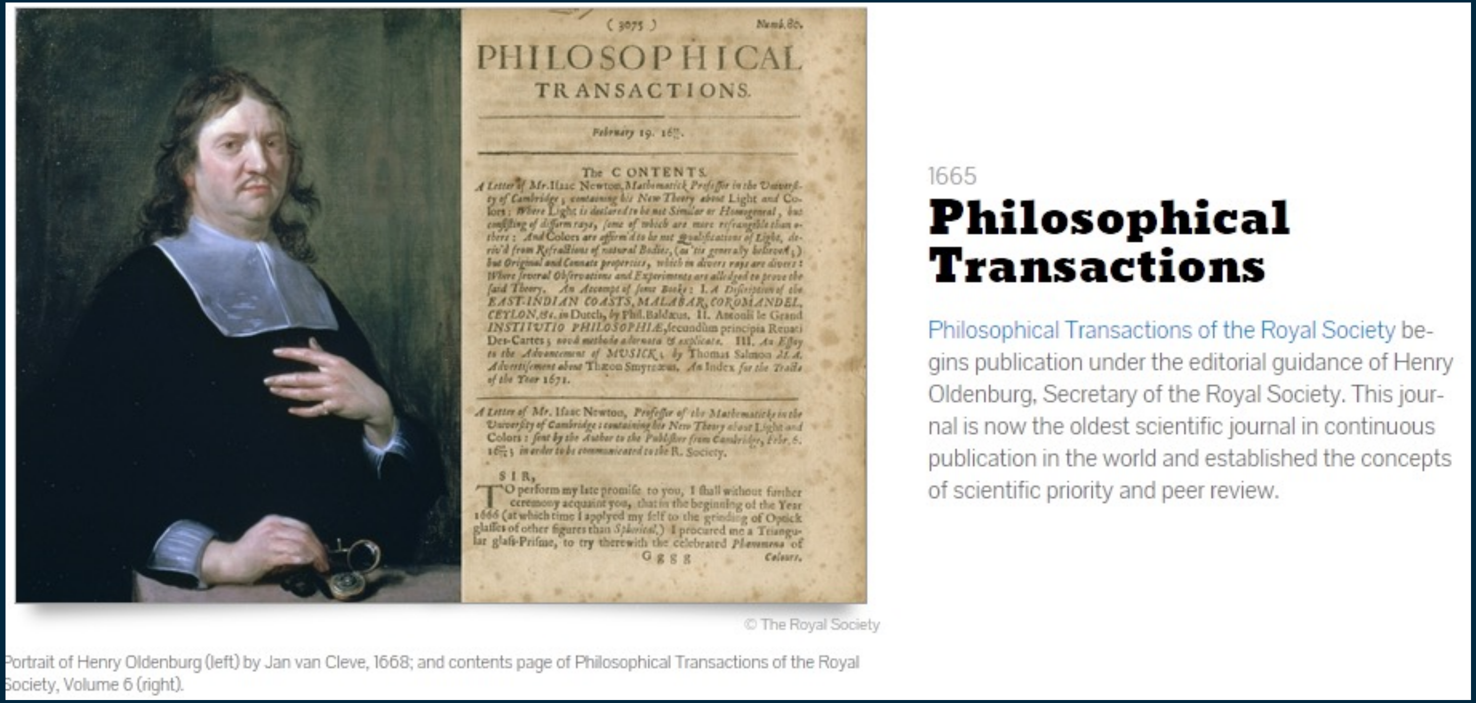
Diego Alves

Pauline Krielke

Isabell Landwehr

Sergei Bagdasarov

1665



(3075) Numb. 80.

PHILOSOPHICAL TRANSACTIONS.

February 19. 16⁷¹⁄₇₂.

The CONTENTS.

A Letter of Mr. Iſaac Newton, Mathematick Profeſſor in the Univerſity of Cambridge; containing his New Theory about Light and Colors: Where Light is declared to be not Similar or Homogeneal, but conſiſting of difform rays, ſome of which are more refrangible than others: And Colors are affirm'd to be not Qualifications of Light, deriv'd from Refractions of natural Bodies, (as 'tis generally believed;) but Original and Connate properties, which in divers rays are divers: Where ſeveral Obſervations and Experiments are alledged to prove the ſaid Theory. An Accompt of ſome Books: I. A Deſcription of the EAST-INDIAN COASTS, MALABAR, COROMANDEL, CEYLON, &c. in Dutch, by Phil. Baldæus. II. Antonii le Grand INSTITUTIO PHILOSOPHIÆ, ſecundùm principia Renati Des-Cartes; novâ methodo adornata & explicata. III. An Eſſay to the Advancement of MUSICK; by Thomas Salmon M. A. Advertiſement about Thæon Smyrnæus. An Index for the Tracts of the Year 1671.

A Letter of Mr. Iſaac Newton, Profeſſor of the Mathematicks in the Univerſity of Cambridge; containing his New Theory about Light and Colors: ſent by the Author to the Publiſher from Cambridge, Febr. 6. 16⁷¹⁄₇₂; in order to be communicated to the R. Society.

SIR,

To perform my late promiſe to you, I ſhall without further ceremony acquaint you, that in the beginning of the Year 1666 (at which time I applyed my ſelf to the grinding of Optick glaſſes of other figures than Spherical,) I procured me a Triangular glaſs-Priſme, to try therewith the celebrated Phænomena of

Gggg Colours,

© The Royal Society

Portrait of Henry Oldenburg (left) by Jan van Cleve, 1668; and contents page of Philosophical Transactions of the Royal Society, Volume 6 (right).

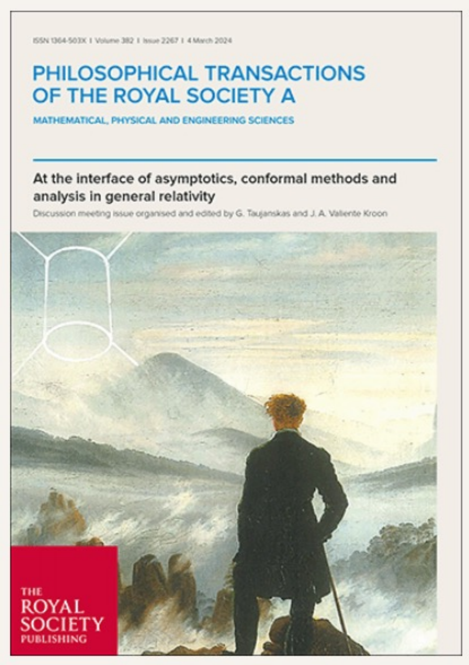
1665

Philosophical Transactions

Philosophical Transactions of the Royal Society begins publication under the editorial guidance of Henry Oldenburg, Secretary of the Royal Society. This journal is now the oldest scientific journal in continuous publication in the world and established the concepts of scientific priority and peer review.

https://royalsociety.org/about-us/history/

March, 4, 2024



ISSN 1364-503X | Volume 382 | Issue 2267 | 4 March 2024

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

MATHEMATICAL, PHYSICAL AND ENGINEERING SCIENCES

At the interface of asymptotics, conformal methods and analysis in general relativity

Discussion meeting issue organised and edited by G. Taujanskas and J. A. Valiente Kroon

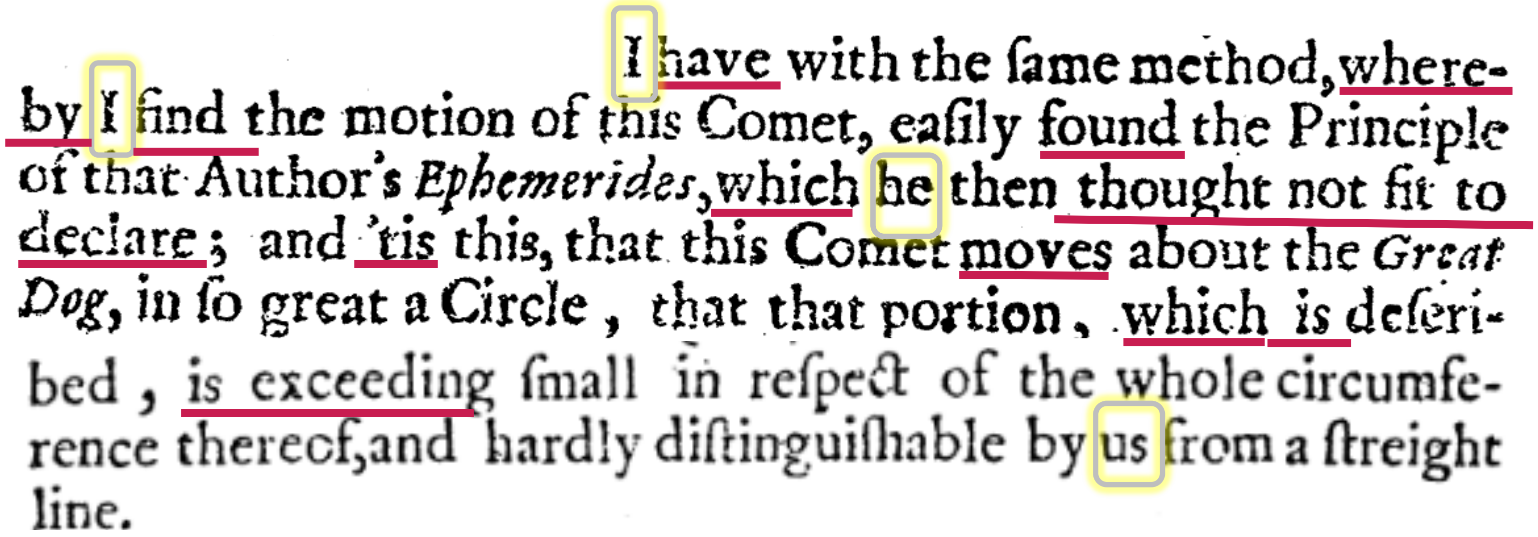
https://royalsocietypublishing.org/cms/attachment/047e986d-cb3c-4171-90ec-07418f1b0f4a/front.pdf

Diachronic variation

Reporting genre

- involved verbal style
- general vocabulary

1665

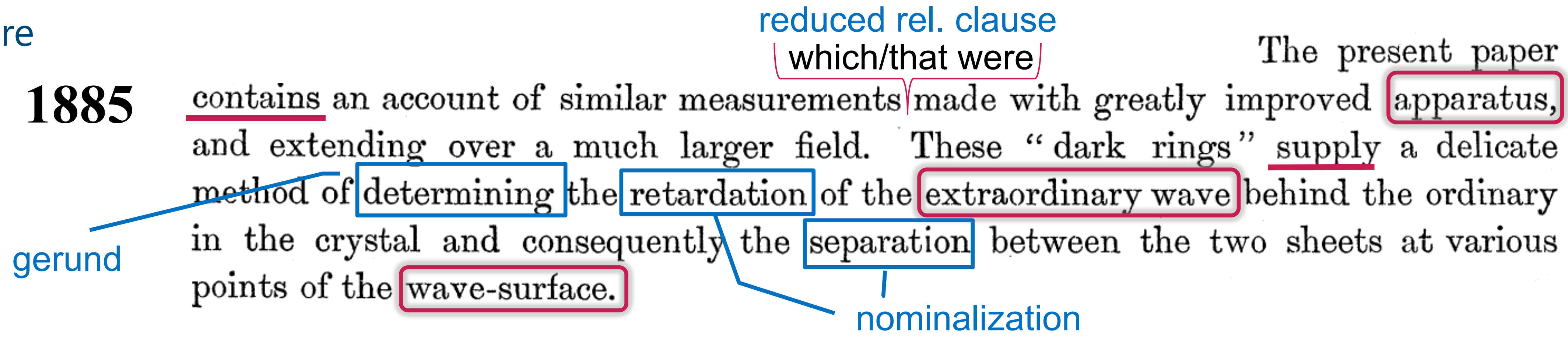


I have with the ſame method, whereby I find the motion of this Comet, eaſily found the Principle of that Author's *Ephemerides*, which he then thought not fit to declare; and 'tis this, that this Comet moves about the *Great Dog*, in ſo great a Circle, that that portion, which is deſcribed, is exceeding ſmall in reſpect of the whole circumference thereof, and hardly diſtinguiſhable by us from a ſtreight line.

Expository genre

- informational nominal style
- specialized vocabulary

1885



reduced rel. clause

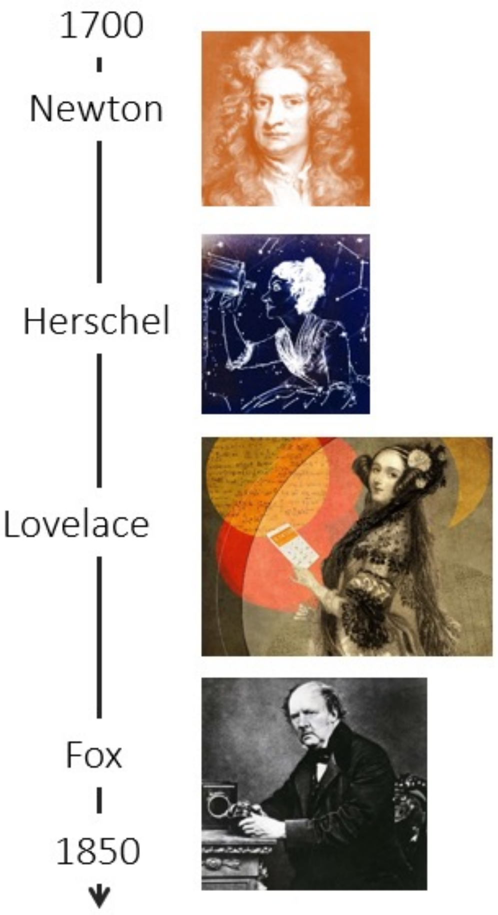
which/that were

The present paper contains an account of similar measurements made with greatly improved apparatus, and extending over a much larger field. These "dark rings" supply a delicate method of determining the retardation of the extraordinary wave behind the ordinary in the crystal and consequently the separation between the two sheets at various points of the wave-surface.

gerund

nominalization

Assumptions



Evolution of modern science → Development of scientific language

(cf. Ure 1982, Halliday 1988, Harris 1991)

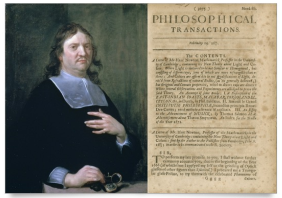
Diversification → distinctness from ‘general’ English

Specialization → expressivity (new expressions) (Säily et al. 2017)

Standardization → conventionalization (e.g. formulaic expressions, terminology) (DeSmet 2016)

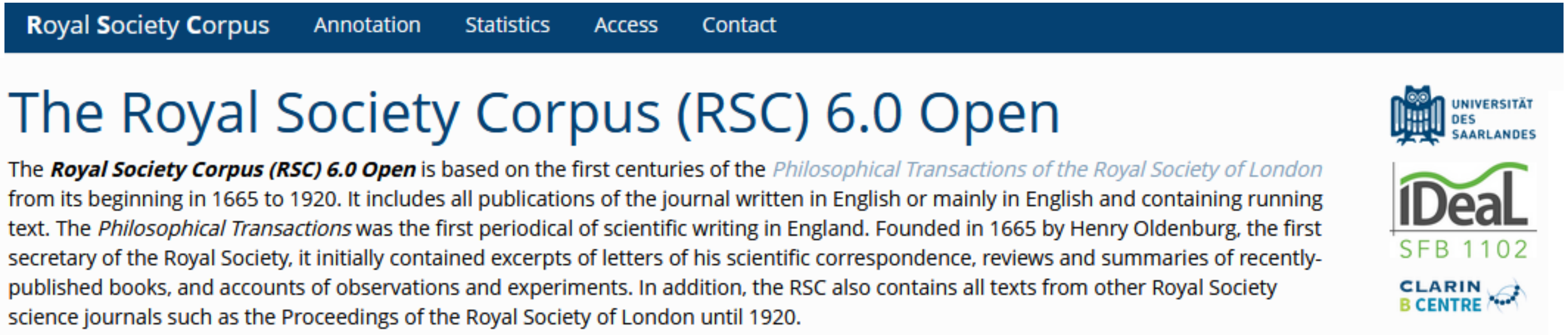
→ optimal code for expert-to-expert communication

Data



Portrait of Henry Oldenburg (left) by Jan van Cleve, 1668; and contents page of Philosophical Transactions of the Royal Society, Volume 6 (right).

© The Royal Society

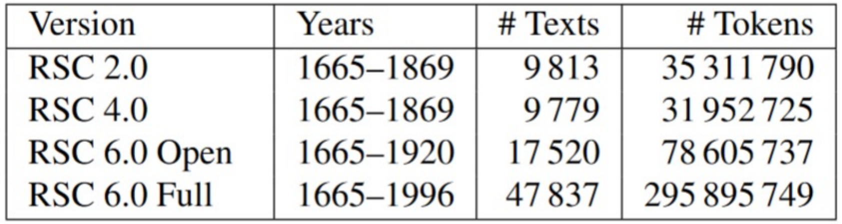


The Royal Society Corpus (RSC) 6.0 Open

The ***Royal Society Corpus (RSC) 6.0 Open*** is based on the first centuries of the *Philosophical Transactions of the Royal Society of London* from its beginning in 1665 to 1920. It includes all publications of the journal written in English or mainly in English and containing running text. The *Philosophical Transactions* was the first periodical of scientific writing in England. Founded in 1665 by Henry Oldenburg, the first secretary of the Royal Society, it initially contained excerpts of letters of his scientific correspondence, reviews and summaries of recently-published books, and accounts of observations and experiments. In addition, the RSC also contains all texts from other Royal Society science journals such as the Proceedings of the Royal Society of London until 1920.

(Kermes et al. 2016; Fischer et al. 2020, Menzel et al. 2021)

- Built in accordance to FAIR principles (Wilkinson et al. 2016)
- OCR-correction based on Noisy-Channel Spell Checker (Klaus et al. 2019)
- 295 mio tokens and 47k texts
- Comprehensive metadata (Menzel et al. 2021)

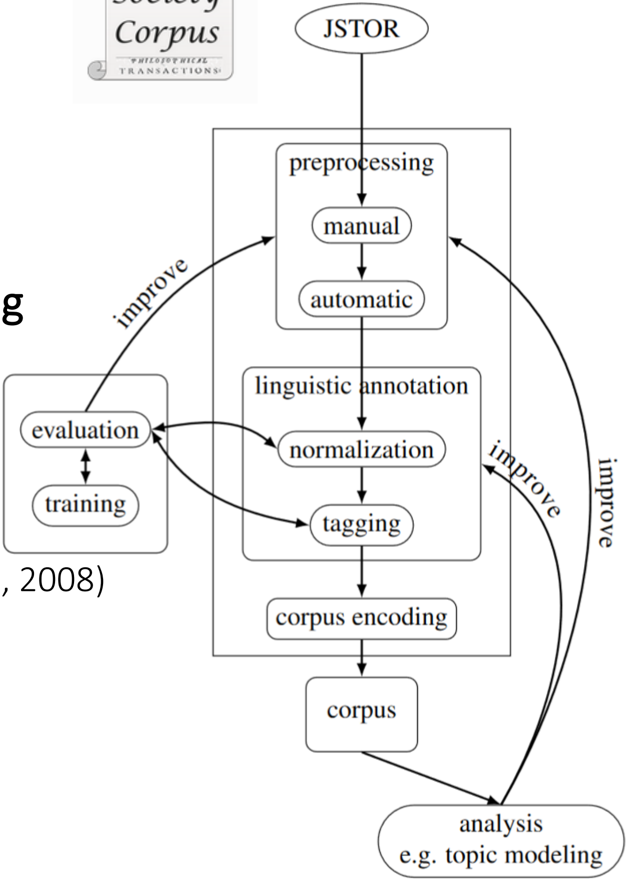


Version	Years	# Texts	# Tokens
RSC 2.0	1665–1869	9 813	35 311 790
RSC 4.0	1665–1869	9 779	31 952 725
RSC 6.0 Open	1665–1920	17 520	78 605 737
RSC 6.0 Full	1665–1996	47 837	295 895 749

Table 1: History of RSC releases. Compared to previous releases, the current *Open* version covers 51 additional years.

Corpus building inspired by Agile Software Development

(Cockburn, 2001; Voormann and Gut, 2008)



Overview of methods

1. Detect periods of Innovation vs. Conventionalization in a data-driven fashion
    - Kullback-Leibler Divergence

2. Inspect change:
    - Word embeddings: Inspect Specialization trends
    - Hyperbolic embeddings: Inspect emergence of specialized terminology

3. Model extra-linguistic factors:
    - Event cascades: modeling influencer and influencees on picking up new terms

4. Model linguistic context:
    - Surprisal: Context-aware analysis of evolving norms and expectations of the scientific community

Detect periods of change in language use

(rather than comparing predefined periods)

Divergence

parsing

*n*gram POS

POS

$$p(unit|context)$$

lemmas

words

morphemes

Extra-linguistic

Language models (LMs) → detect change across *situational contexts*

Relative entropy

(Kullback-Leibler Divergence)

Overall KLD

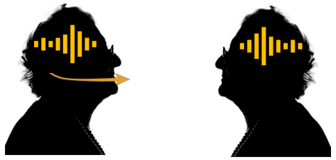
$$D(P||Q) = \sum_{i} p(unit_i|P)\ \log_2 \frac{p(unit_i|P)}{p(unit_i|Q)}$$

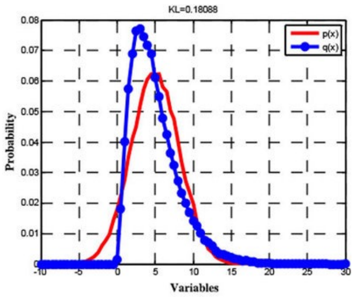
Pointwise KLD

$$D\_l(P||Q) = p(l)\ \log_2 \frac{p(l)}{q(l)}$$

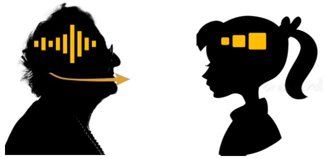
Unigram model with Jelinek-Mercer Smoothing p(w)=(1−λ)·p′(w)+λ·b(w), where p‘(w) subcorpus, b(w) entire corpus, λ=0.05

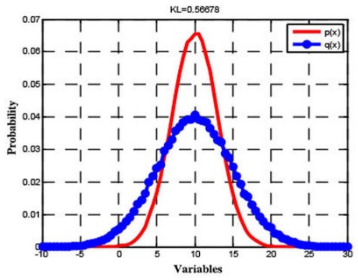
$D(1650s||1700s)$





$D(1650s||1990s)$





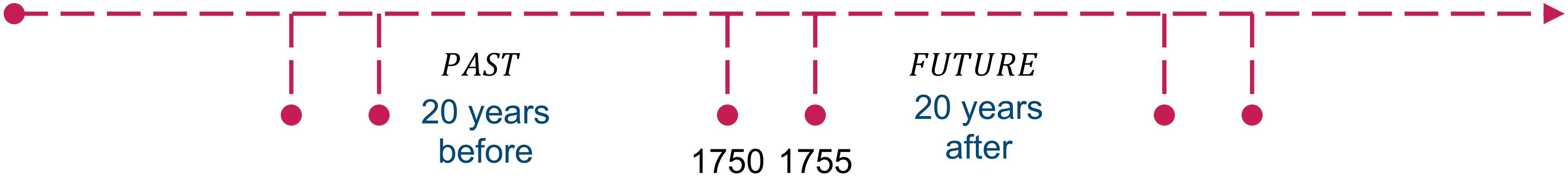
relatively similar → low divergence

differ → higher divergence

LMs to detect period of change (Degaetano-Ortlieb and Teich 2018, 2019)

Relative entropy (KLD)

$$D(FUTURE||PAST) = \sum_i p(unit_i|FUTURE)\ \log_2 \frac{p(unit_i|FUTURE)}{p(unit_i|PAST)}$$



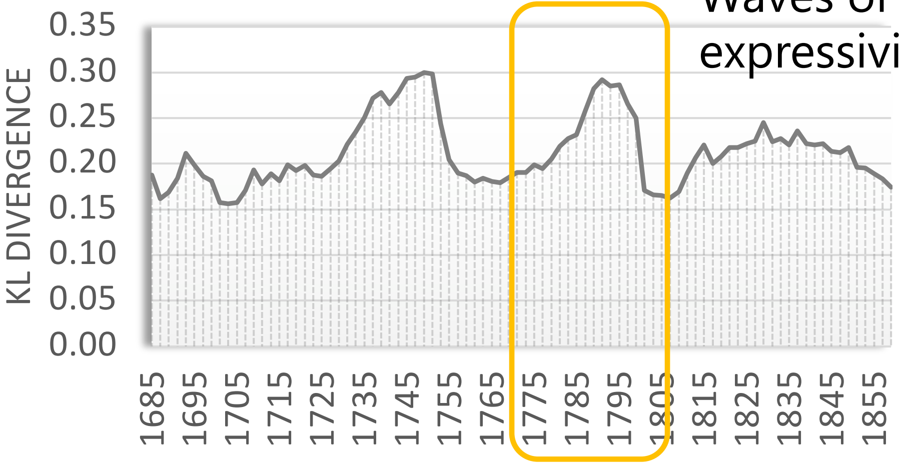
$PAST$

20 years before

1750 1755

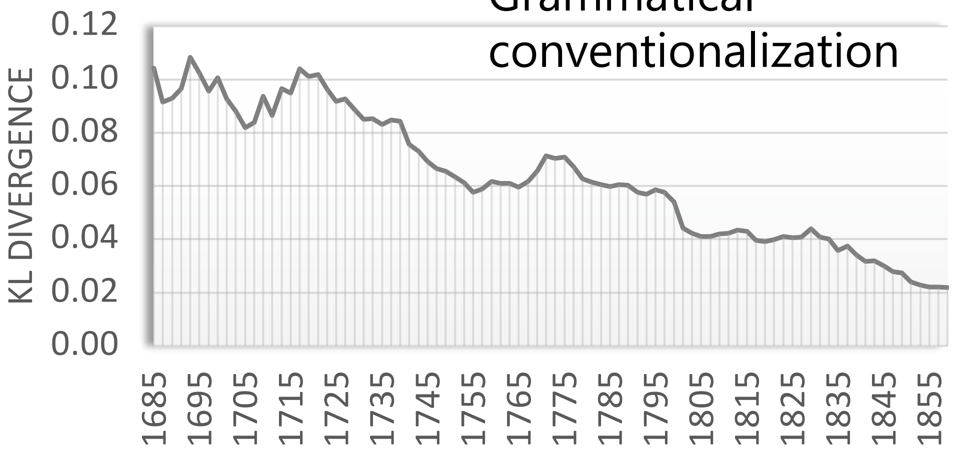
$FUTURE$

20 years after



Waves of increased expressivity

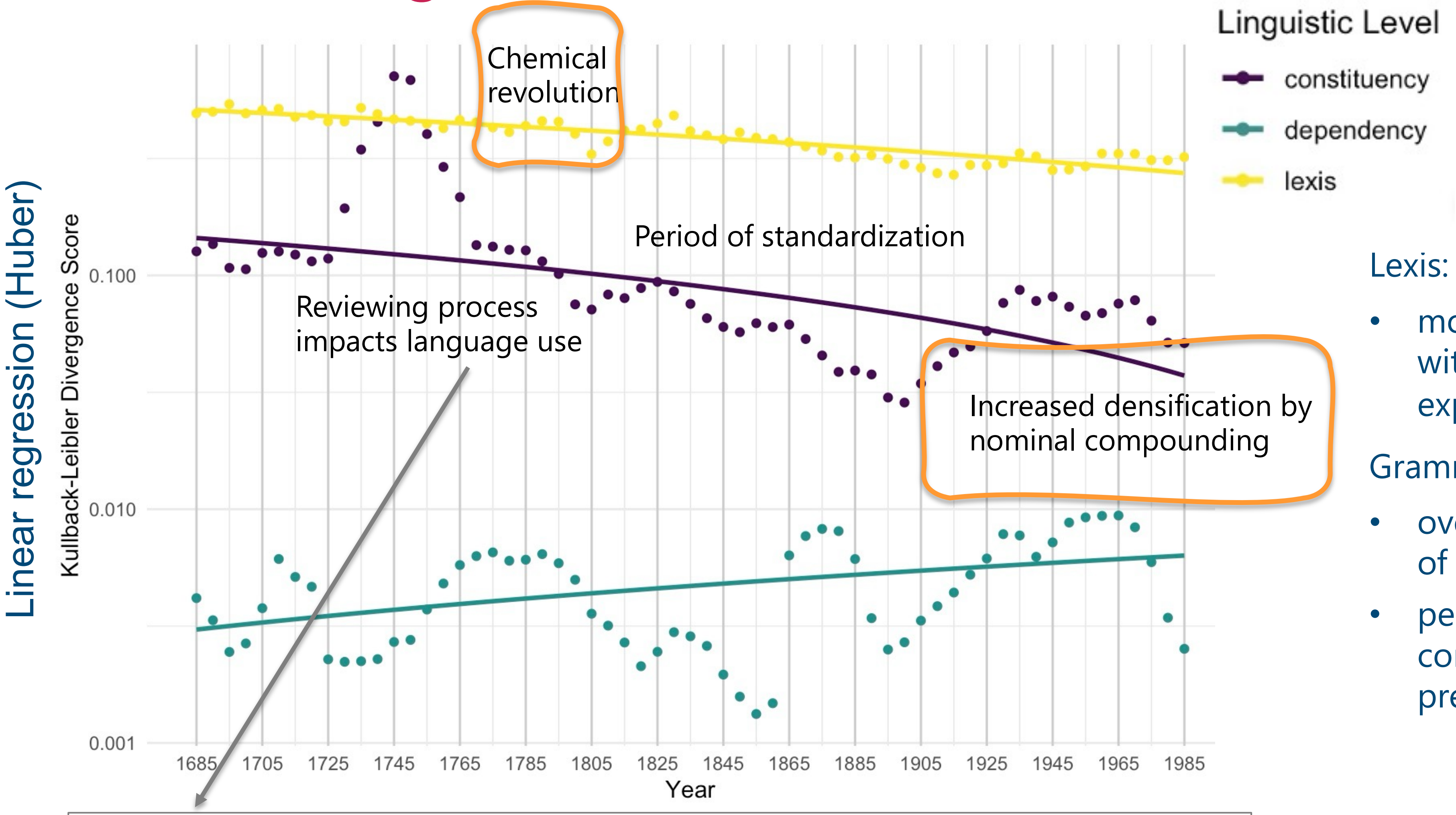
$unit = lemma$



Grammatical conventionalization

$unit =\ POStrigram$ (e.g Adj-Adj-Noun)

KLD across linguistic levels



Mr. Pound observed a similar eclipse at Wanstead … and has described it in ***the Philos. Trans.*** *N. 347 , p. 402.* (id 6094358)

Lexis:

- more stable vocabulary usage with periods of increased expressivity

Grammar

- overall conventionalization trend of grammatical structures
- peak in the 20th c. triggered by compounding and premodification

Analyze change in language use

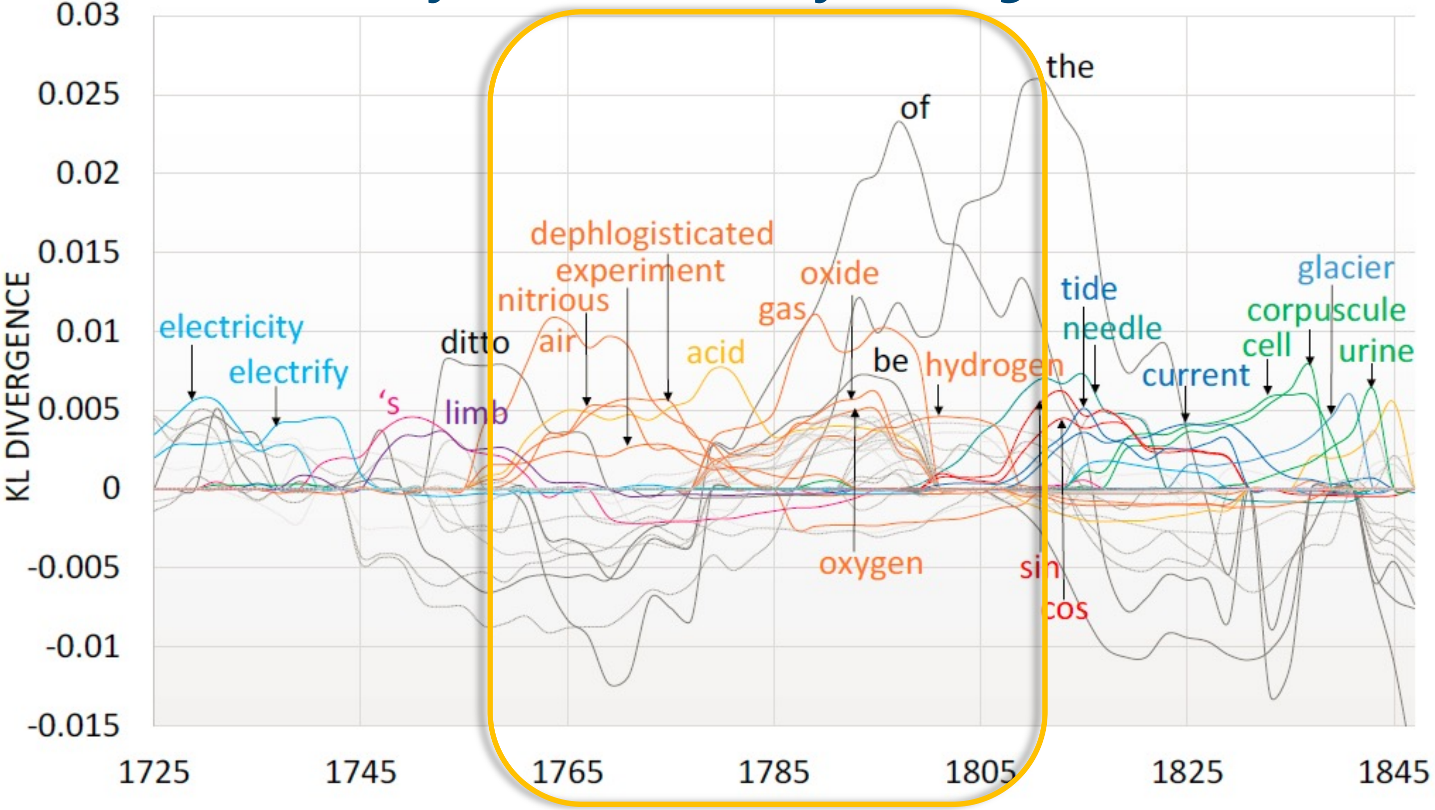
Which linguistic features contribute to a change?

Allows a qualitative perspective on the data!

Lexical contributions to change (Degaetano-Ortlieb and Teich 2018, 2019)

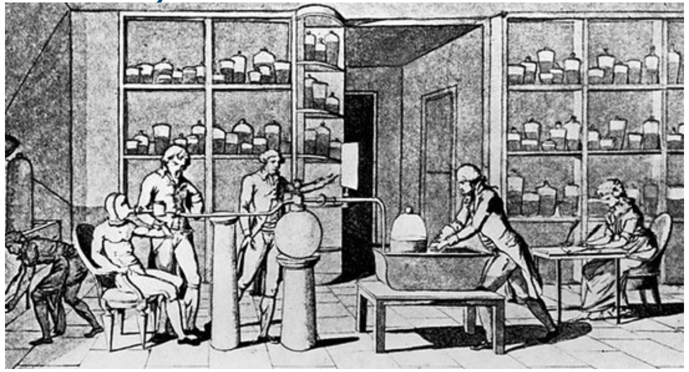
$$D\_lemma(FUTURE||PAST) = p(lemma|FUTURE)\ \log_2 \frac{p(lemma|FUTURE)}{p(lemma|PAST)}$$

2-year window, 20-year range



Discovery of hydrogen (*inflammable air*) by Henry Cavendish in 1766

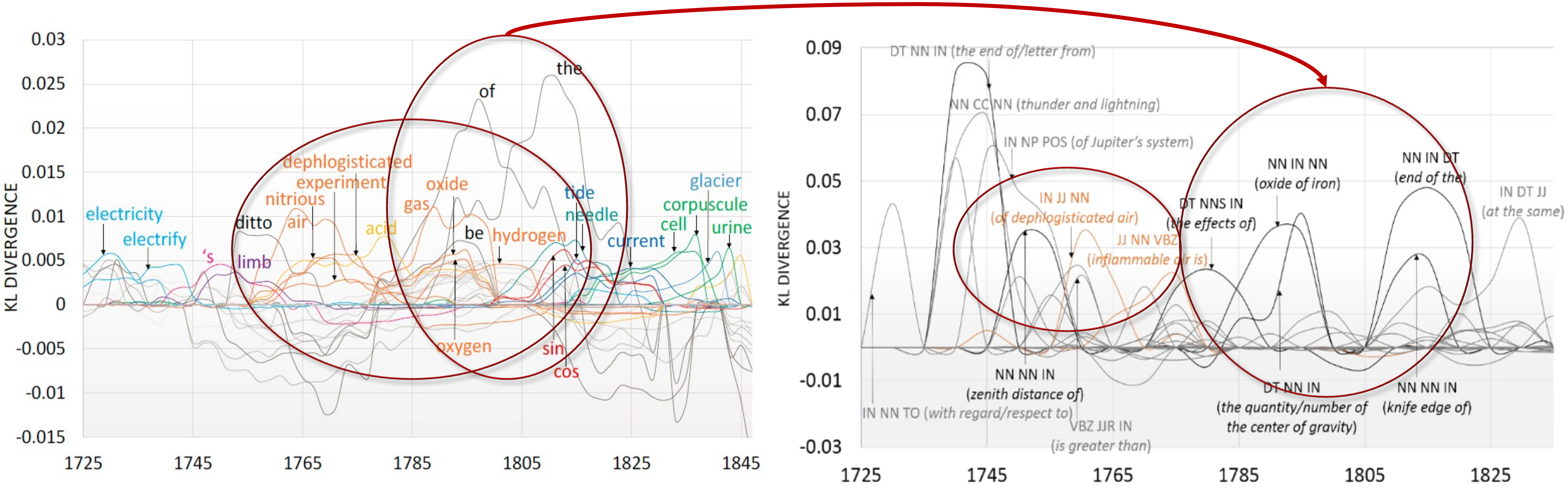
Discovery of oxygen (*dephlogisticated air*) by Joseph Priestley in 1774



Effects across linguistic levels

LEXIS

GRAMMAR



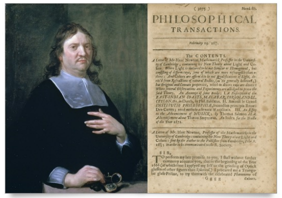
$unit = lemma$

$unit = POStrigram$

Methodology

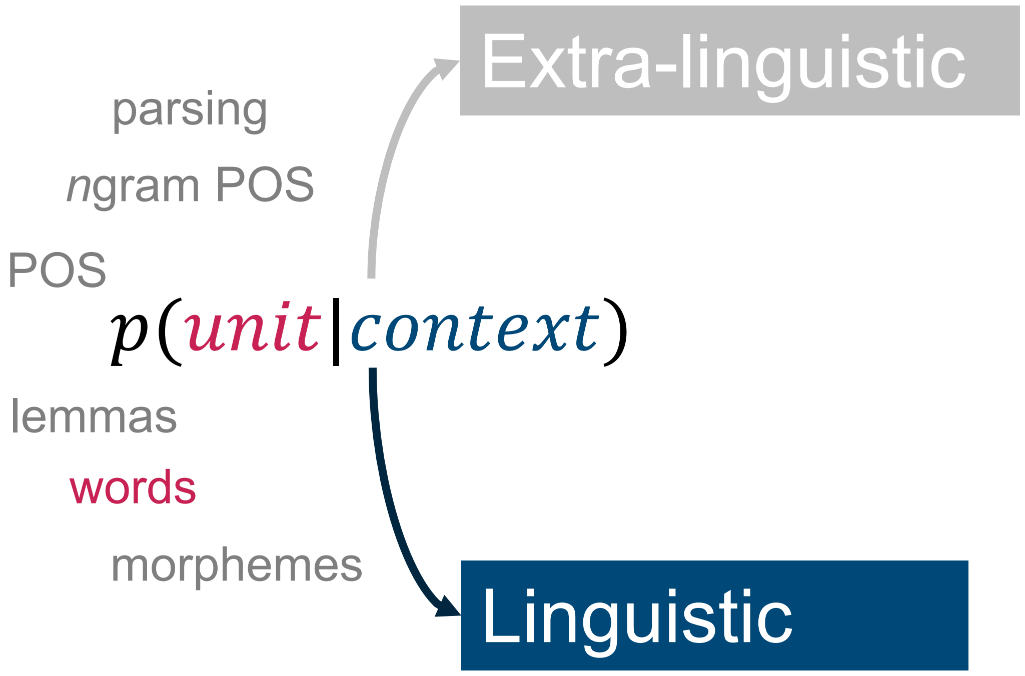
Royal Society Corpus

PHILOSOPHICAL TRANSACTIONS



© The Royal Society

Portrait of Henry Oldenburg (left) by Jan van Cleve, 1668; and contents page of Philosophical Transactions of the Royal Society, Volume 6 (right).



Extra-linguistic

parsing

*n*gram POS

POS

$p(unit|context)$

lemmas

words

morphemes

Linguistic

Language models (LMs)
Relative entropy
(Kullback-Leibler Divergence)

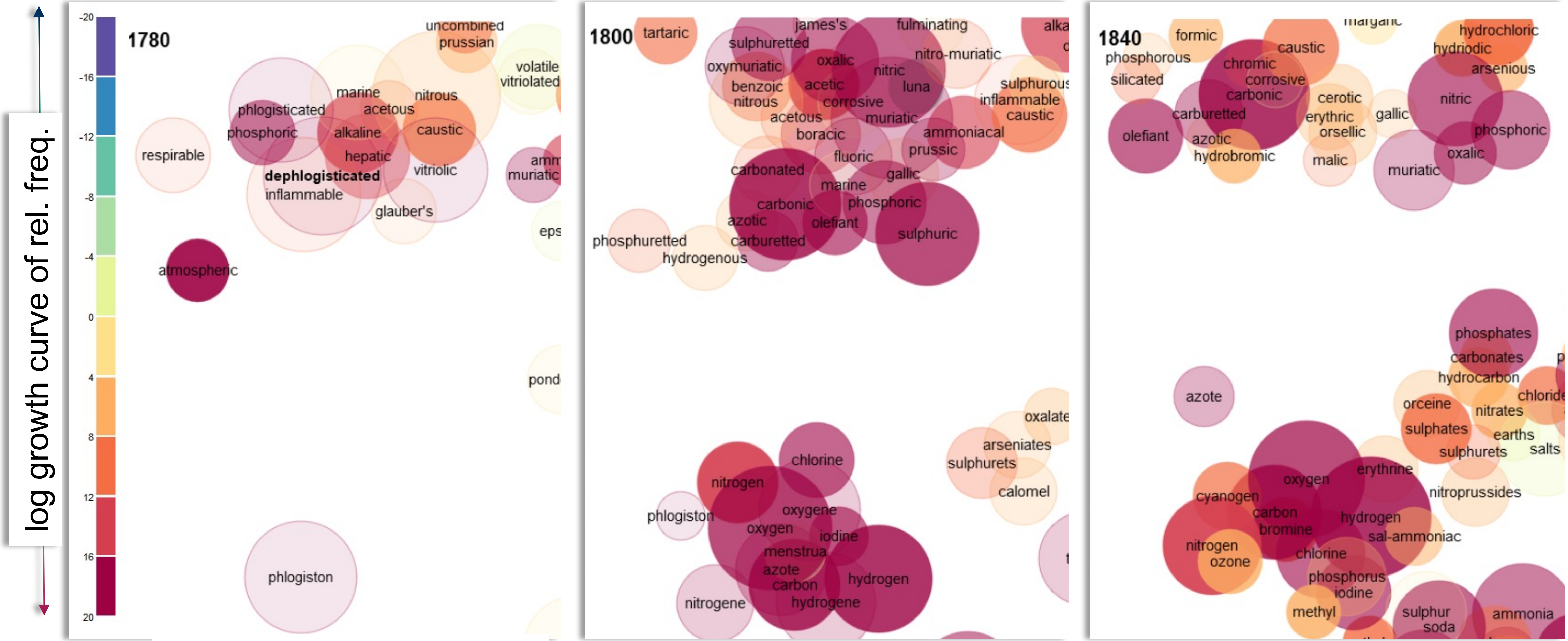
→ detect change across *TIME*

Word embeddings
word2vec:
surrounding words [-5,5]

→ inspect *paradigmatic context*

Paradigmatic context and change (Fankhauser et al. 2017, Bizzoni et al. 2019)

decreasing



increasing

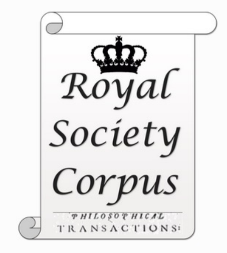
Bubble size: $\sqrt{relative\ frequency}$

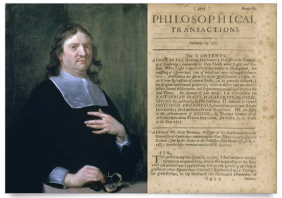
https://corpora.ids-mannheim.de/diaviz/royalsociety.html

Modeling linguistic context

Surprisal allows for context aware analysis of evolving norms and expectations within the scientific community

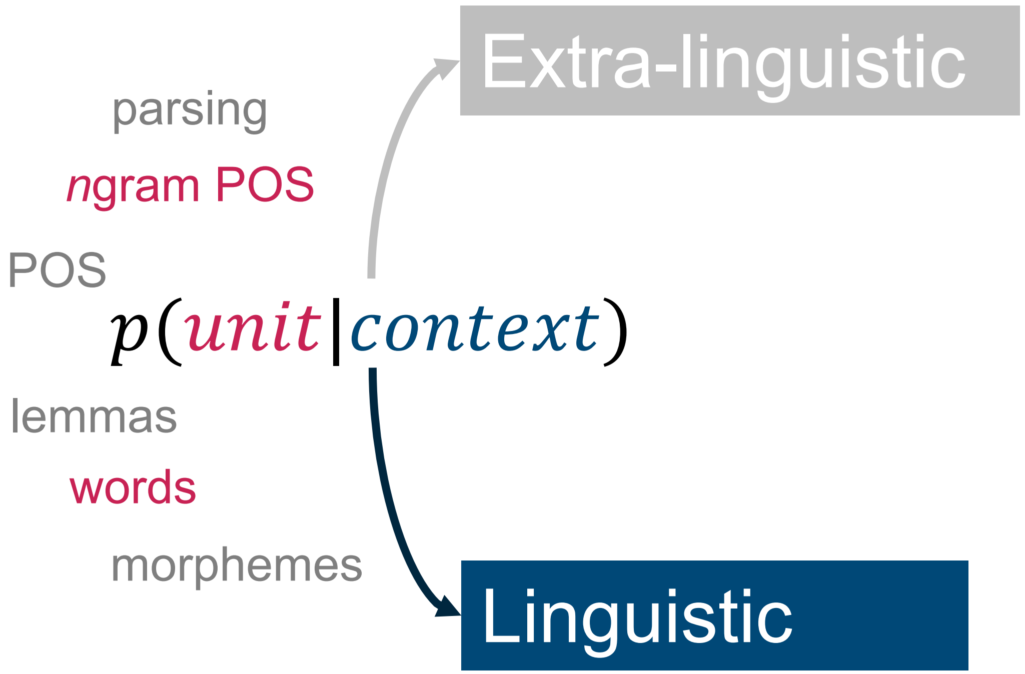
Methodology





© The Royal Society

Portrait of Henry Oldenburg (left) by Jan van Cleve, 1668; and contents page of Philosophical Transactions of the Royal Society, Volume 6 (right).



Extra-linguistic

parsing

ngram POS

POS

$p(unit|context)$

lemmas

words

morphemes

Linguistic

Language models (LMs)
Relative entropy
(Kullback-Leibler Divergence)

→ detect change across *TIME*

Word embeddings
Wang2vec:
surrounding words [-5,5]

→ inspect *paradigmatic context*

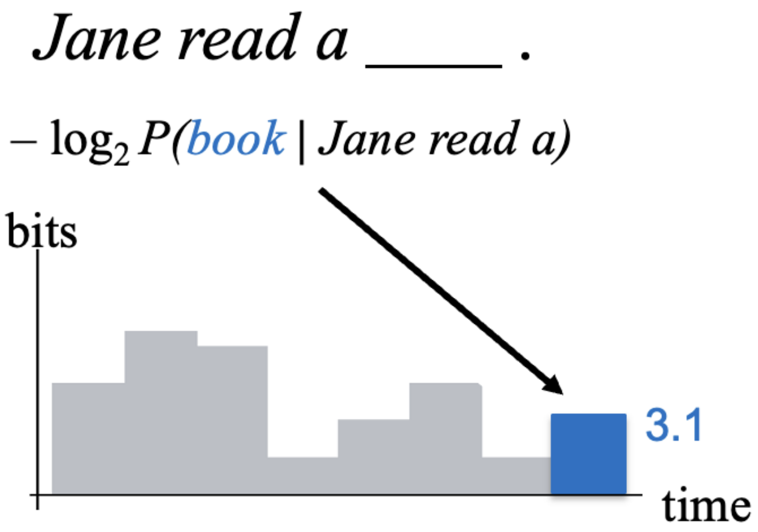
LMs
Average surprisal

→ analyze *syntagmatic context*

Surprisal

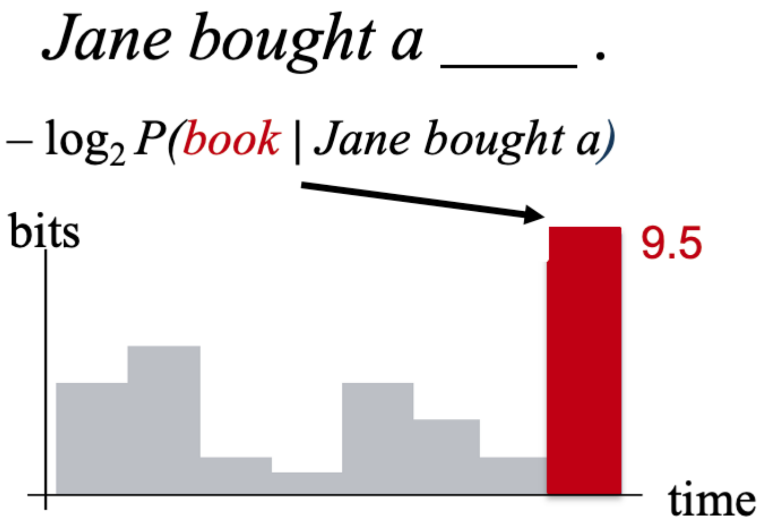
$$Surprisal(unit) = -\log_2 p(unit|context)$$

(cf. Shannon 1948, Hale 2001, Levy 2008)



Jane read a ____ .

$-\log_2 P(book \mid Jane\ read\ a)$



Jane bought a ____ .

$-\log_2 P(book \mid Jane\ bought\ a)$

$$Effort(unit) \propto Surprisal(unit)$$

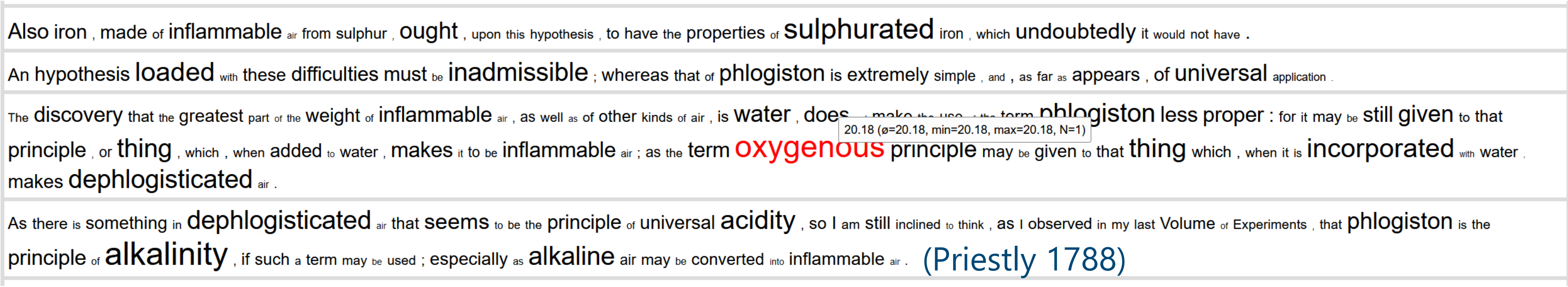
(cf. Hale 2001)

Shannon’s surprisal

(Shannon 1949)

$$Surprisal(unit) = -\log_2 P(unit|context)$$
$$Effort(unit) \propto Surprisal(unit)$$

(Hale 2001; Levy 2008; Crocker et al. 2016)



Also iron , made of inflammable air from sulphur , ought , upon this hypothesis , to have the properties of sulphurated iron , which undoubtedly it would not have .

An hypothesis loaded with these difficulties must be inadmissible ; whereas that of phlogiston is extremely simple , and , as far as appears , of universal application .

The discovery that the greatest part of the weight of inflammable air , as well as of other kinds of air , is water , does make the use the term phlogiston less proper : for it may be still given to that principle , or thing , which , when added to water , makes it to be inflammable air ; as the term oxygenous principle may be given to that thing which , when it is incorporated with water , makes dephlogisticated air .

20.18 (ø=20.18, min=20.18, max=20.18, N=1)

As there is something in dephlogisticated air that seems to be the principle of universal acidity , so I am still inclined to think , as I observed in my last Volume of Experiments , that phlogiston is the principle of alkalinity , if such a term may be used ; especially as alkaline air may be converted into inflammable air .

(Priestly 1788)

4-gram language model

$$Surprisal(w_i) = -\log_2 p(w_i|w_{i-1}w_{i-2}w_{i-3})$$

$$AvS_{1850}(w) = \frac{1}{|w|}\sum_i -\log_2 p(w_i|w_{i-1}w_{i-2}w_{i-3})$$

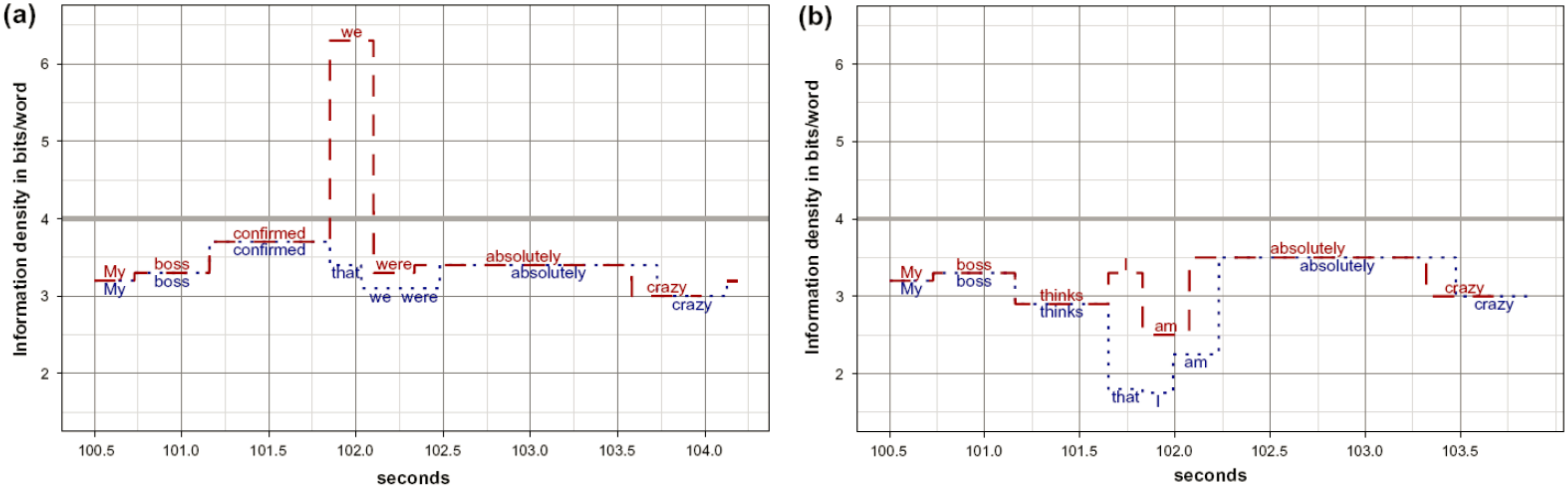
Analyze the syntagmatic context to trace

- *specialization* and *conventionalization*
- *densification* over time

Ongoing experiments with transformer-based surprisal (Steuer et. al 2024)

Uniform information density (UID hypothesis)

(Jäger 2010)

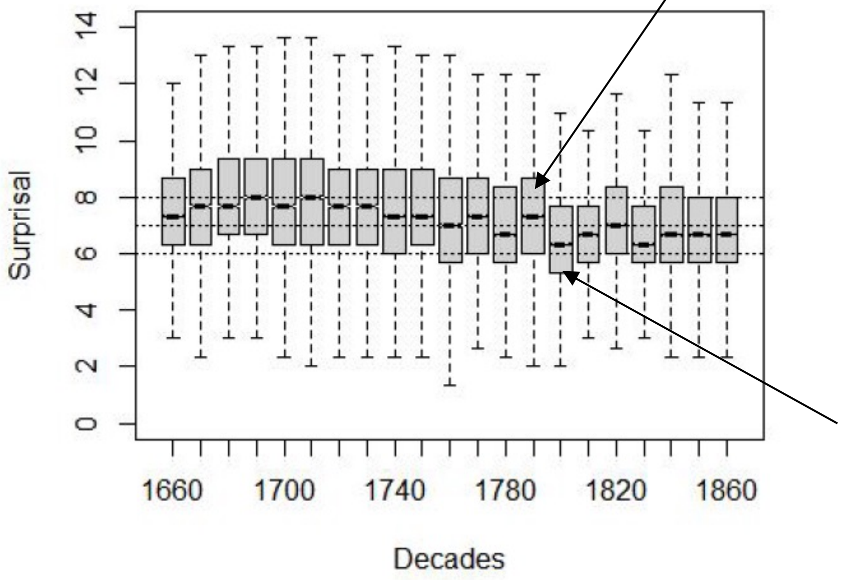


Syntagmatic context and change (Degaetano-Ortlieb and Teich 2018, 2019)

Surprisal averaged across time periods (four-gram model on decades)

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

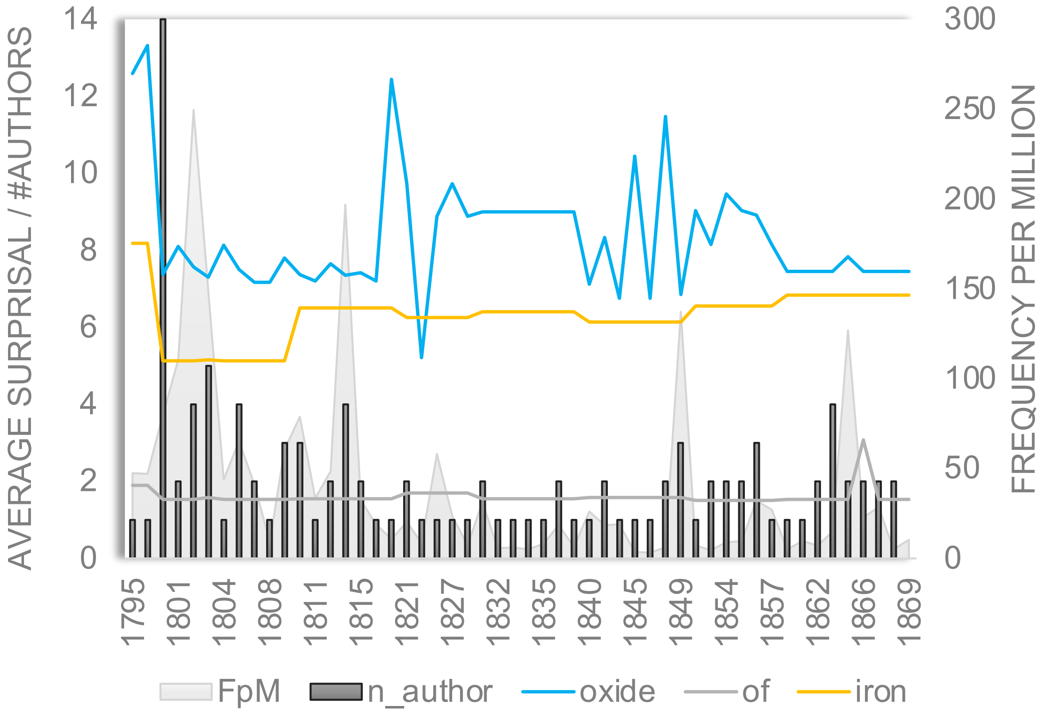
1790: new lexemes



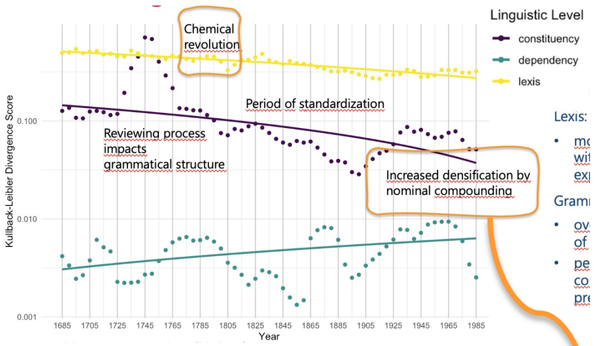
1800: established terms

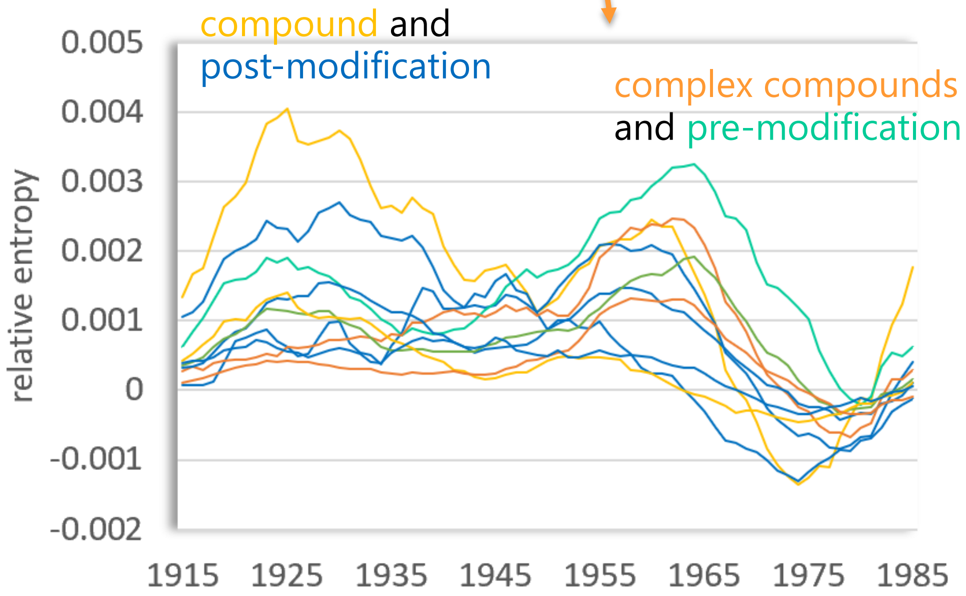
Surprisal of NN.IN.NN (lexical)

Average surprisal of *oxide of iron*



Types of patterns and changes

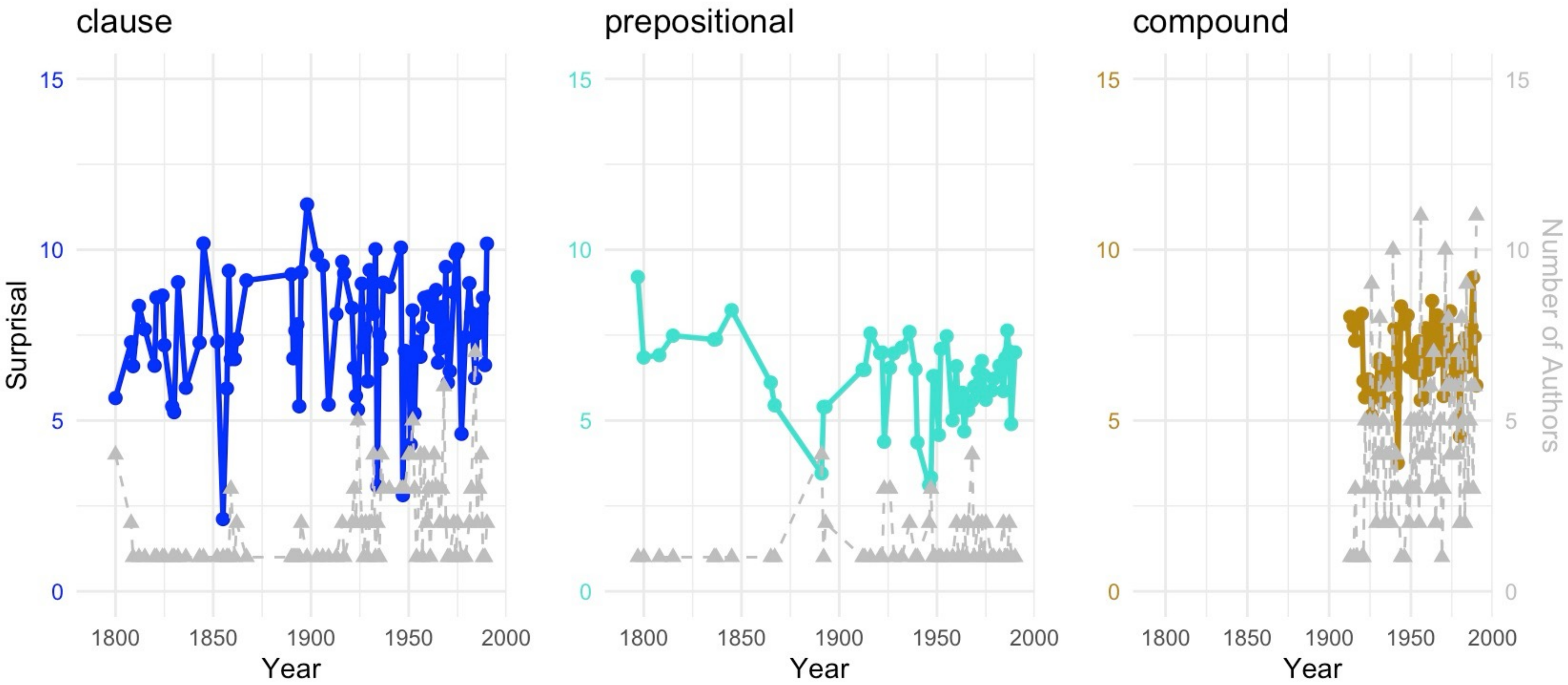




Structural compression strategies (cf. Biber and Gray 2016)

Single noun	The **oxygen** was consumed
Modification by clause	The quantity of **oxygen which was consumed**
Post-modification	The **consumption of oxygen**
Compound	The **oxygen consumption** plotted against
Pre-modification / complex compound	Animals have a **mean dermal oxygen consumption**

Surprisal to inspect cycles of change



The quantity of ***oxygen which was consumed***

The ***consumption of oxygen***

The ***oxygen consumption*** *plotted against*

Summary

Development of the scientific register

Balance between *specialization* and *conventionalization* procedures

→ Optimal code: sufficiently conventionalized while leaving room for innovation (interplay between lexis and grammar)

Relative entropy as a measure to detect changes

Surprisal as a notion of cognitive effort and predictability of changes

→ Optimization process in language as cycles of linguistic change

Combining different methods allows for validity and diverse insights to gain a more comprehensive picture

Thank you for your attention!

Yuri Bizzoni, Marius Mosbach, Dietrich Klakow

Hyperbolic embeddings

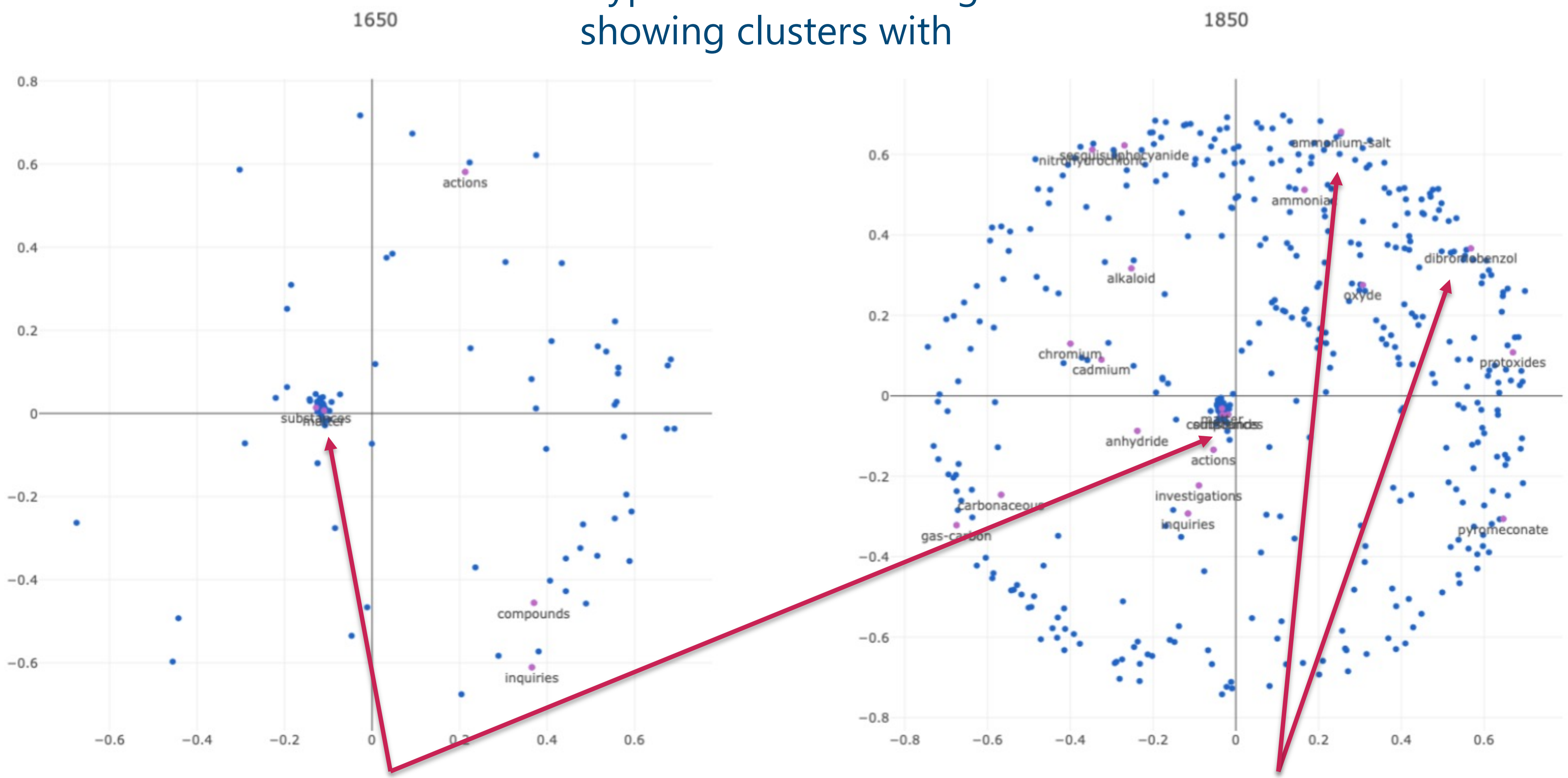
further trace the process of specialization

from the use of more abstract/general to more specific terms over time

(Bizzoni et al. 2019)

Trends of specialization

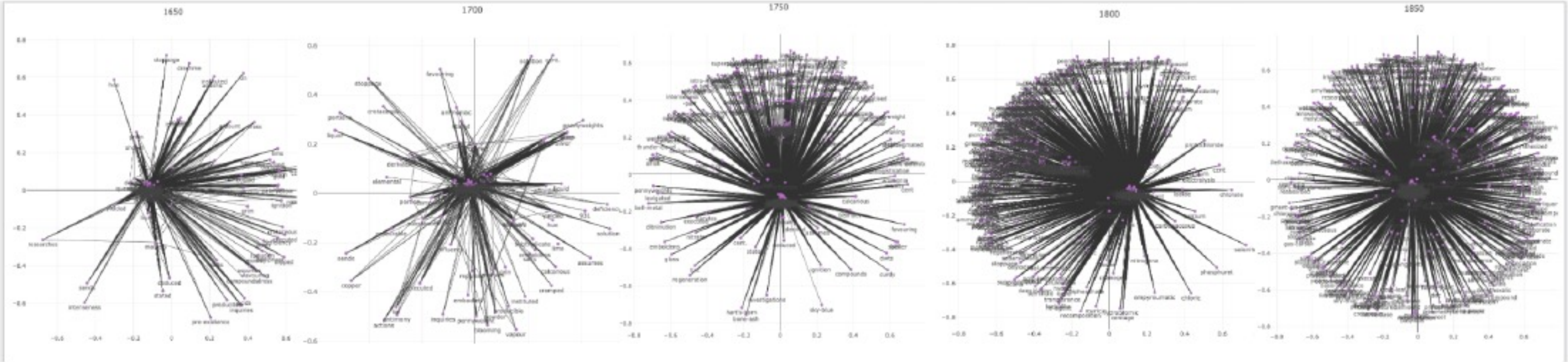
Hyperbolic embeddings showing clusters with



more abstract (potentially ambiguous) words towards the center,

specialized terms at the periphery of the cluster and more distant from the center

Trends of specialization



Population of the space towards the periphery indicating specialization

Yuri Bizzoni
Katrin Menzel
Elke Teich

Influencers and influencees in the RSC

the role of individual scientists
in the diffusion of new concepts
(Bizzoni et al. 2021)

Methods

KLD for term selection

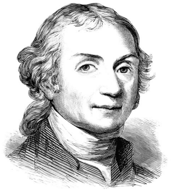
Discovery of oxygen and new nebulae, single-author papers (>7papers, min 1 occurrence)

Event cascades to model influencer and influencees

- Event intensity function $\lambda_j(t) = \lambda_{j,0} + \sum_{t'<t} \alpha_{s_e \to j} \kappa(t - t')$ with based intensity $\lambda_{j,0}$ and sum of influence effects (influence strength $\alpha_{se \to j}$ from source event se to target j from past events t′ $< t$ with decay function κ,
- Influence intensity between entities (se and j) over time interval Δt calculated as sum over B basis models (no. of authors)
- Each model $\phi_b(\Delta t)$ represents influence pattern, with dyad-specific weights $g_{se \to j}^{b}$determining contribution of each pattern

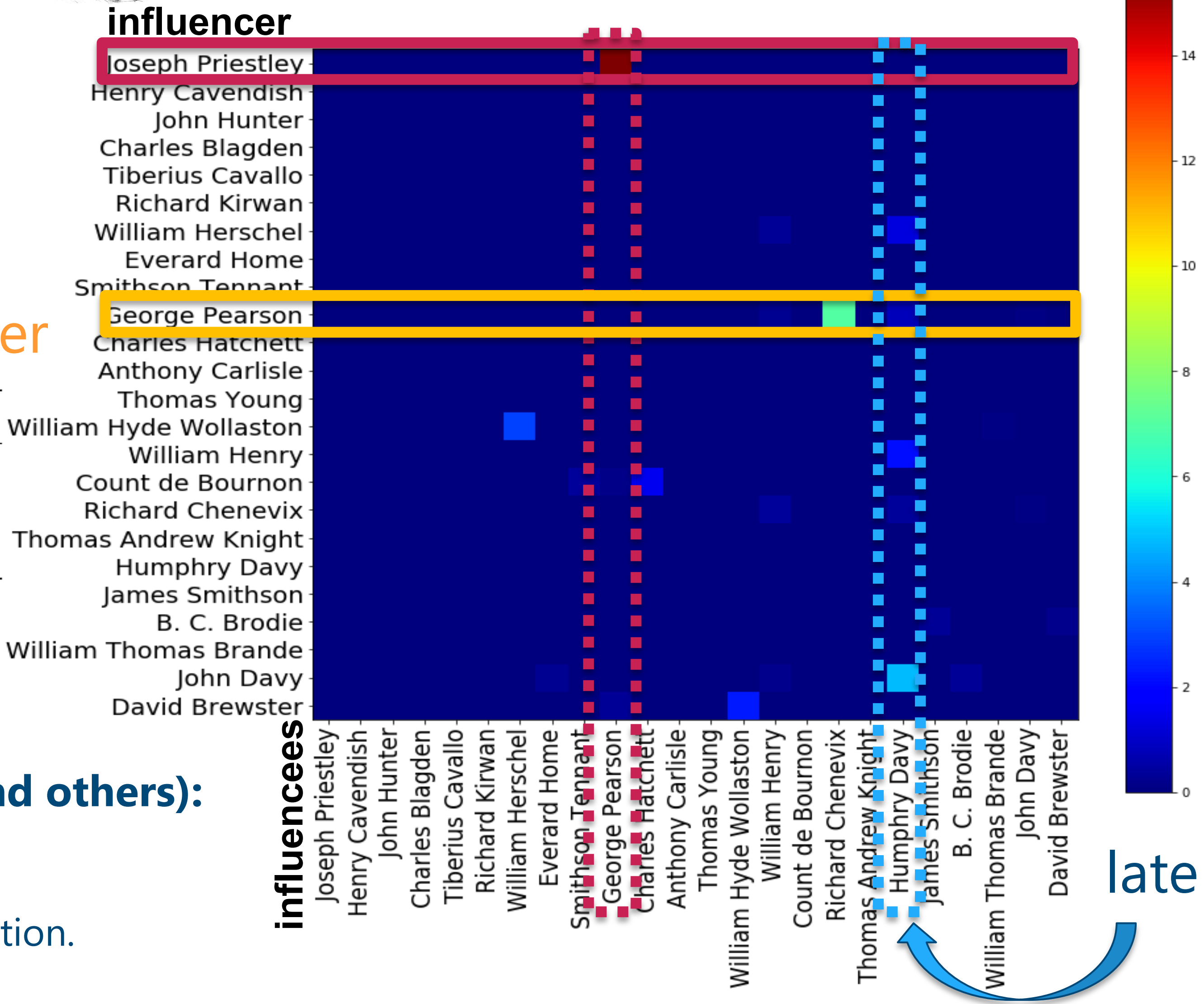
How much does each source event tend to excite each target event?

Innovator (Priestley):



Initiates use, exerting strong, focused influence;
catalyst for trend adoption.

innovator



Early Adopter (Pearson):

Reacts to innovator,
becomes influential among peers,
expanding trend reach.

spreader

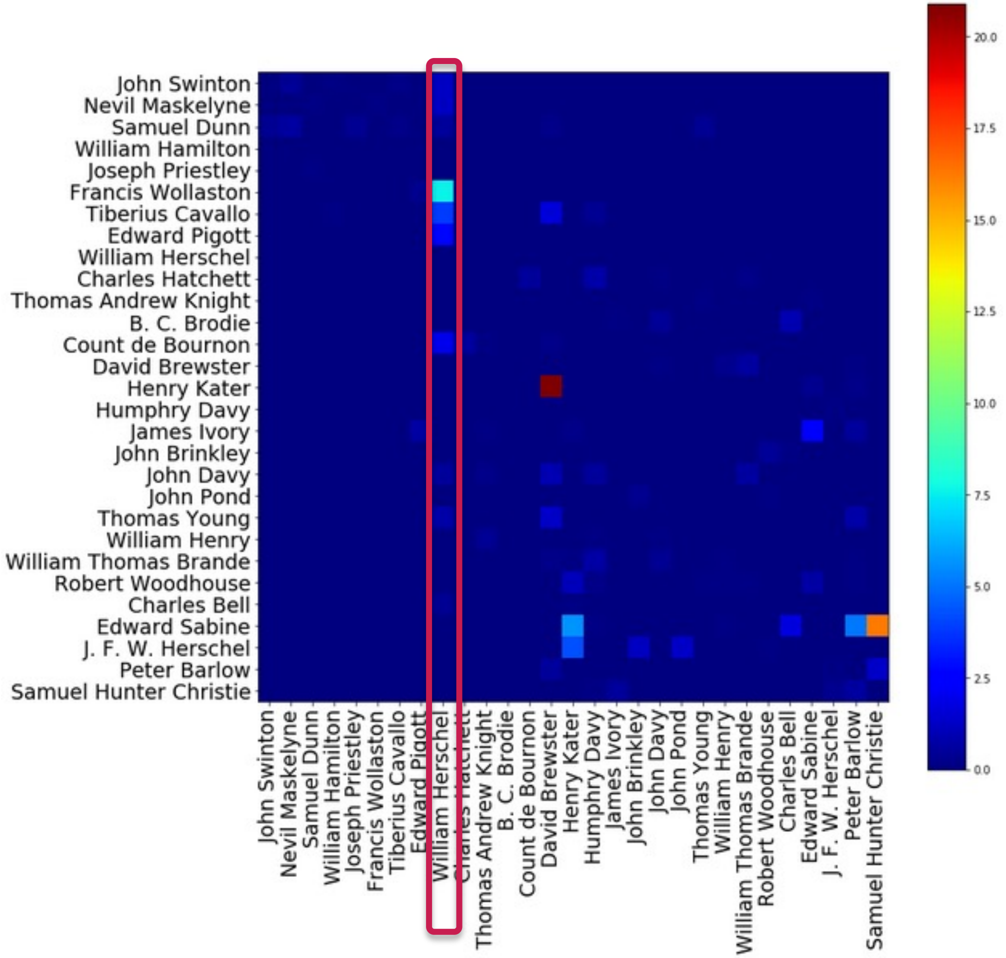
No. of Influenced Authors	Author
11	George Pearson
8	Richard Chenevix
6	William Herschel
4	John Davy
3	William Hyde Wollaston
3	Count de Bournon

early adopter

Early Majority (Davy and others):

Engages with trend widely
popularized by several
authors, solidifying its adoption.

late adopter



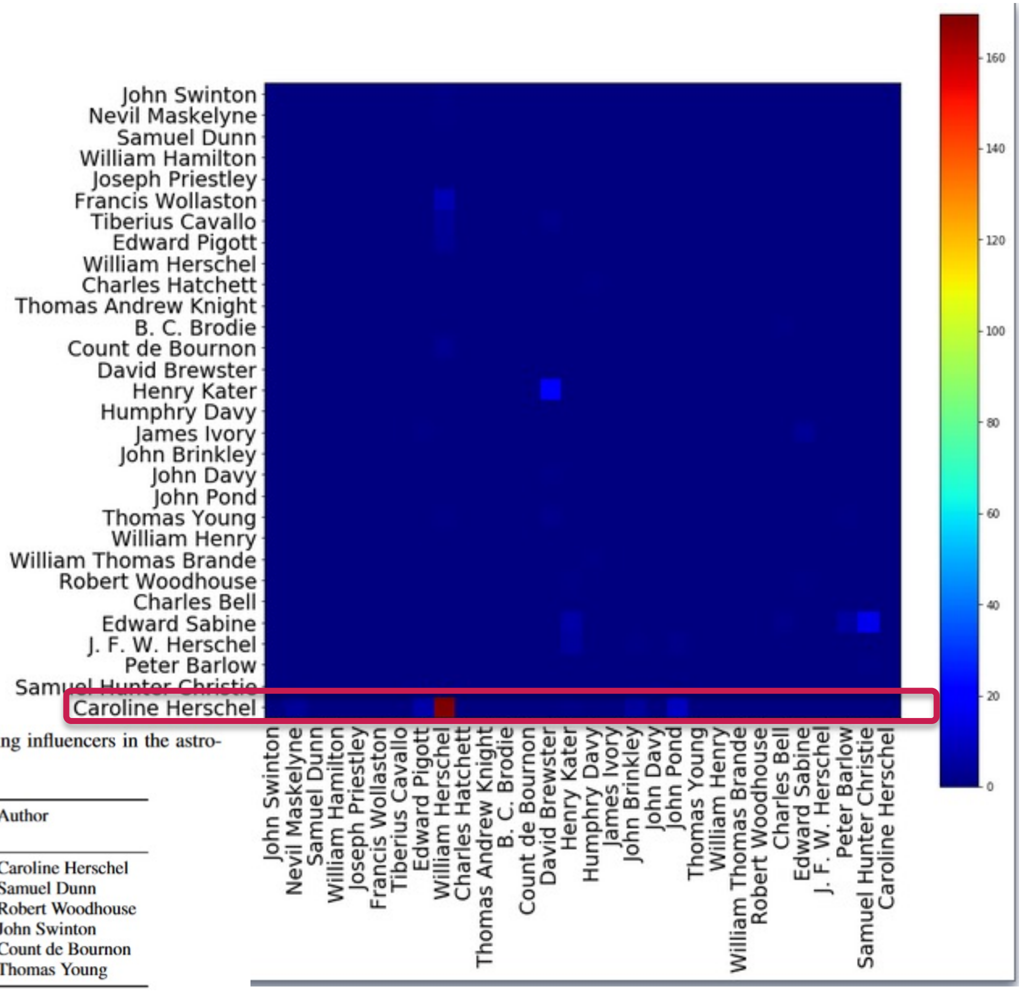


Table 4: Most wide-reaching influencers in the astronomical field.

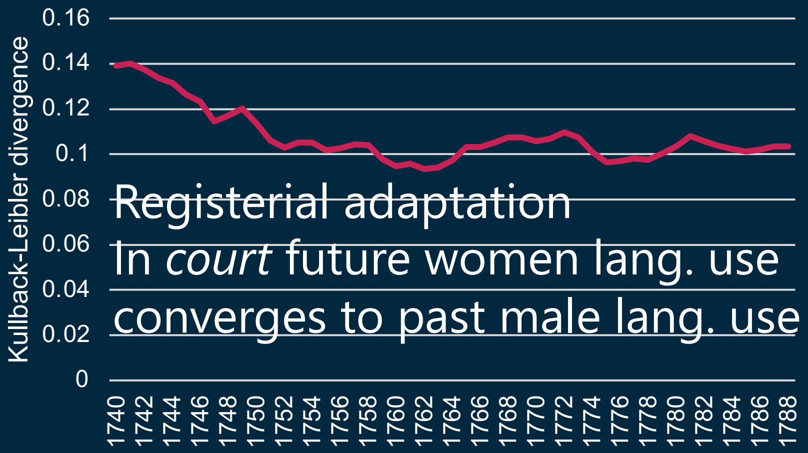
No. of Influenced Authors	Author
19	Caroline Herschel
17	Samuel Dunn
14	Robert Woodhouse
12	John Swinton
12	Count de Bournon
11	Thomas Young

Scenario: Formal vs. informal settings and gender

(Degaetano-Ortlieb, Tanja Säily & Yuri Bizzoni 2021)

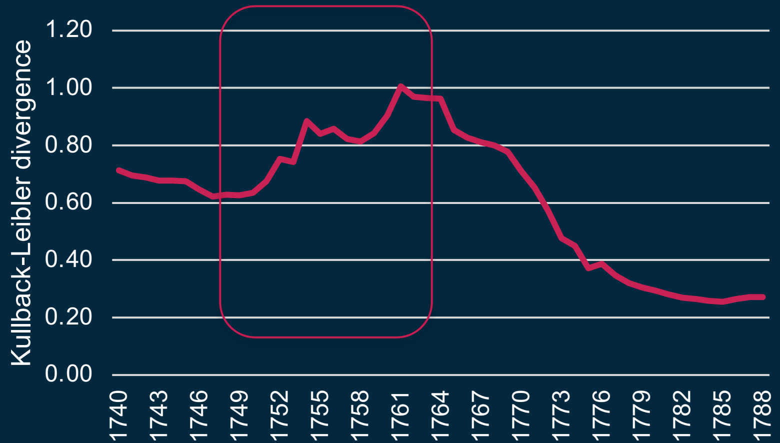
Registerial Adaptation vs. Innovation Across Situational Contexts: 18th Century Women in Transition

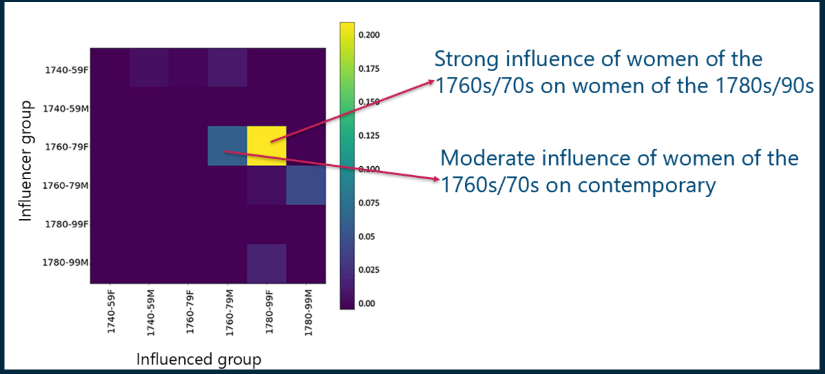
Old Bailey court proceedings
(OBC Corpus; Huber et al. 2016)



Registerial adaptation
In *court* future women lang. use converges to past male lang. use

Letter of correspondences
(TCEECE; Saario & Säily 2020)





Strong influence of women of the 1760s/70s on women of the 1780s/90s

Moderate influence of women of the 1760s/70s on contemporary

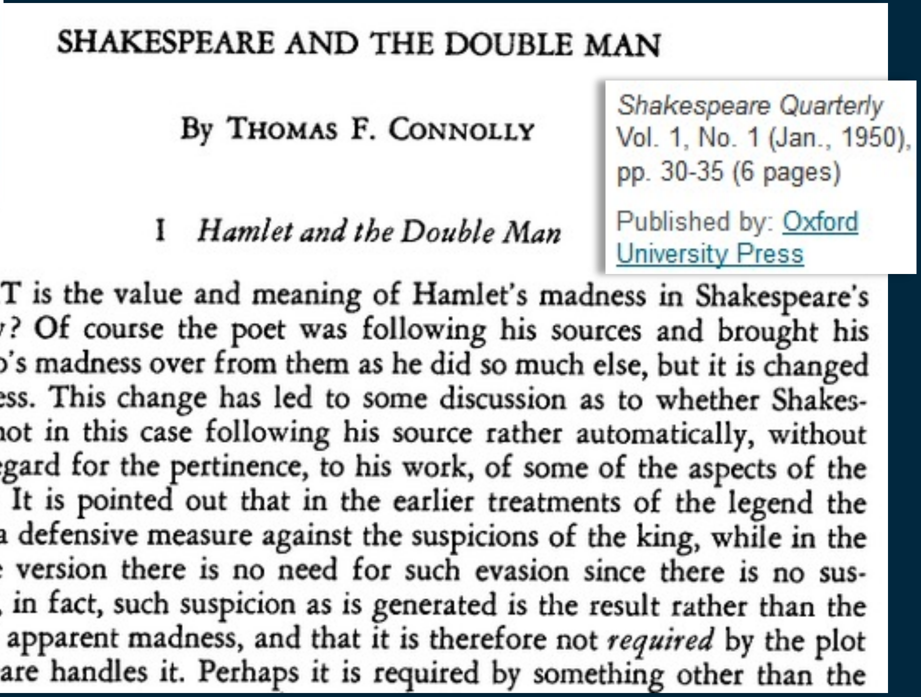
Registerial innovation in *letters:* future women language use diverges from the past (male and female)

Scenarios 1: Literary studies

(Degaetano-Ortlieb & Piper 2019)

(Nussbaum 1997, Lamont 2009, Biber and Gray 2016, Kramnick 2018)

1950



SHAKESPEARE AND THE DOUBLE MAN

By Thomas F. Connolly

Shakespeare Quarterly
Vol. 1, No. 1 (Jan., 1950),
pp. 30-35 (6 pages)

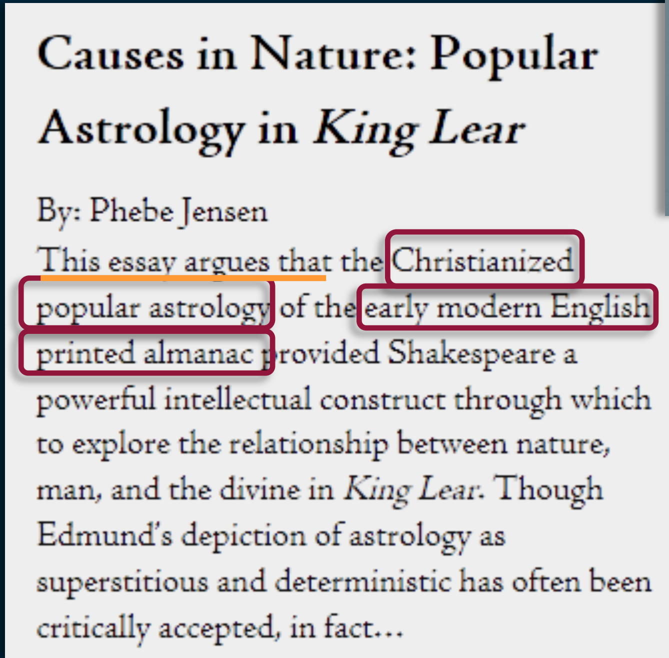
Published by: Oxford University Press

I *Hamlet and the Double Man*

WHAT is the value and meaning of Hamlet's madness in Shakespeare's play? Of course the poet was following his sources and brought his hero's madness over from them as he did so much else, but it is changed in the process. This change has led to some discussion as to whether Shakespeare was not in this case following his source rather automatically, without too much regard for the pertinence, to his work, of some of the aspects of the older plays. It is pointed out that in the earlier treatments of the legend the madness is a defensive measure against the suspicions of the king, while in the Shakespeare version there is no need for such evasion since there is no suspicion; that, in fact, such suspicion as is generated is the result rather than the cause of his apparent madness, and that it is therefore not *required* by the plot as Shakespeare handles it. Perhaps it is required by something other than the

https://www.jstor.org/stable/2866204?seq=1#metadata_info_tab_contents

2018



Causes in Nature: Popular Astrology in *King Lear*

By: Phebe Jensen

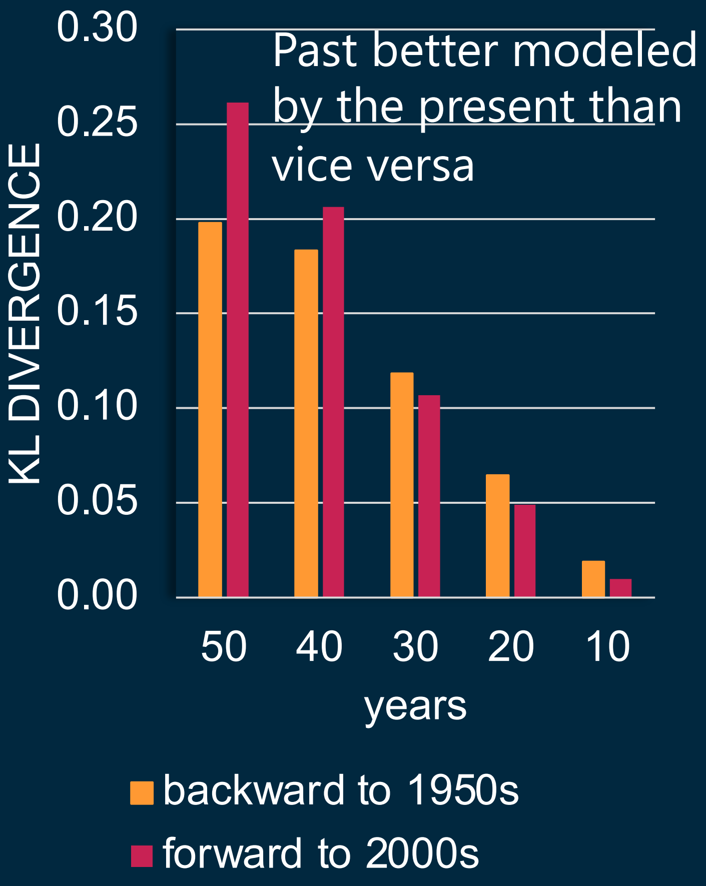
This essay argues that the Christianized popular astrology of the early modern English printed almanac provided Shakespeare a powerful intellectual construct through which to explore the relationship between nature, man, and the divine in *King Lear*. Though Edmund's depiction of astrology as superstitious and deterministic has often been critically accepted, in fact…

https://shakespearequarterly.folger.edu/essays/causes-nature-popular-astrology-king-lear/

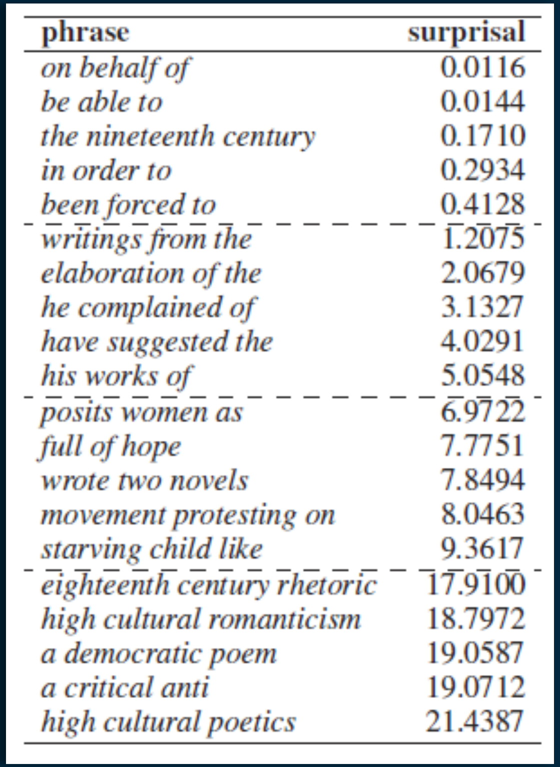
Scenarios 1: Literary studies

(Degaetano-Ortlieb & Piper 2019)

(Nussbaum 1997, Lamont 2009, Biber and Gray 2016, Kramnick 2018)

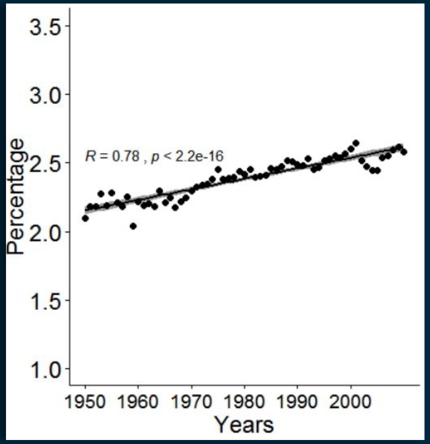


Past better modeled by the present than vice versa

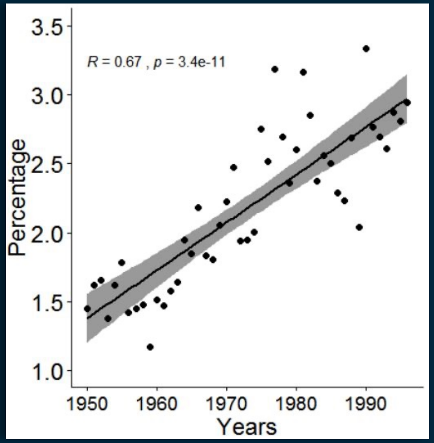


phrase	surprisal
on behalf of	0.0116
be able to	0.0144
the nineteenth century	0.1710
in order to	0.2934
been forced to	0.4128
writings from the	1.2075
elaboration of the	2.0679
he complained of	3.1327
have suggested the	4.0291
his works of	5.0548
posits women as	6.9722
full of hope	7.7751
wrote two novels	7.8494
movement protesting on	8.0463
starving child like	9.3617
eighteenth century rhetoric	17.9100
high cultural romanticism	18.7972
a democratic poem	19.0587
a critical anti	19.0712
high cultural poetics	21.4387

Literary studies



Royal Society



Information Density and Linguistic Encoding

IDeaL
SFB 1102

DFG Deutsche Forschungsgemeinschaft

UNIVERSITÄT
DES
SAARLANDES