Exploring and analyzing linguistic variation

Elke Teich
Saarbrücken
Germany

IDeaL
SFB 1102

M2CI
CLUSTER OF EXCELLENCE

CLARIN-D

UNIVERSITÄT
DES
SAARLANDES

Linguistic variation

Methods

text-based vs. variationist

data-based vs. data-driven

macro vs. micro-analysis

Description

variables: time, age, gender, context of situation/register

linguistic levels: phonetic, lexical, syntactic, semantic

Scientific English
(1650-1850)

Models

entrenchment

innovation

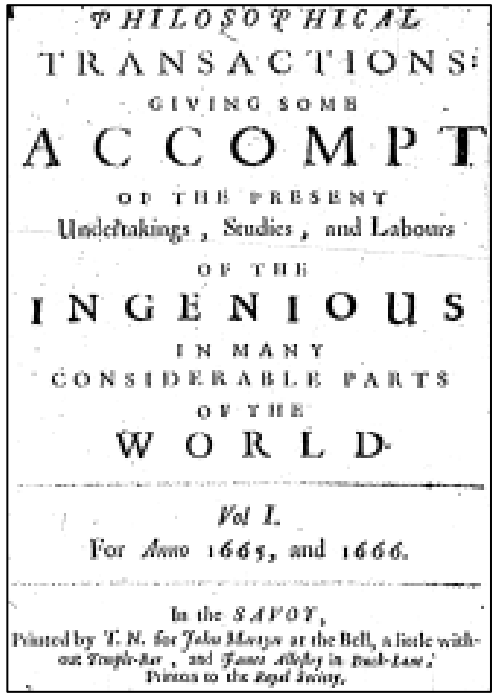
diffusion

typicality

productivity

...

Historical example: Royal Society (article)



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

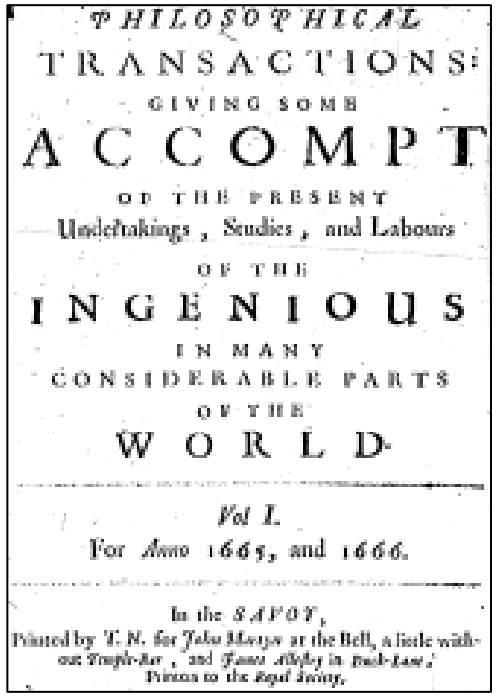
Vol. I.
For Anno 1665, and 1666.

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 (article)



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

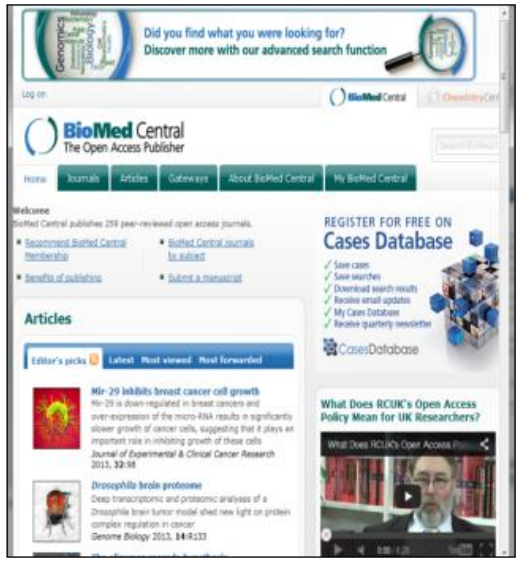
Vol. I.
For Anno 1665, and 1666.

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.

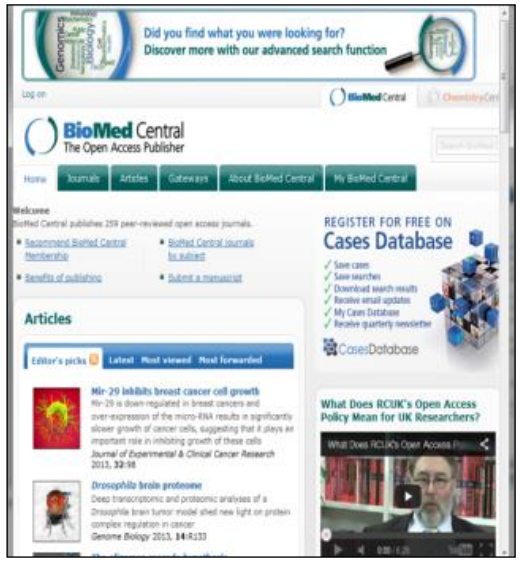
Contemporary example: Biomed Central (abstract)



We report the discovery of a novel downstream target of BCR-ABL signalling, PRL-3 (PTP4A3), an oncogenic tyrosine phosphatase. Analysis of CML cancer cell lines and CML patient samples reveals the upregulation of PRL-3.

http://www.biomedcentral.com/

Contemporary example: Biomed Central (abstract)



We report the discovery of a novel downstream target of BCR-ABL signalling, PRL-3 (PTP4A3), an oncogenic tyrosine phosphatase. Analysis of CML cancer cell lines and CML patient samples reveals the upregulation of PRL-3.

http://www.biomedcentral.com/

Linguistic development of scientific discourse

Assumptions

diversification, specialization

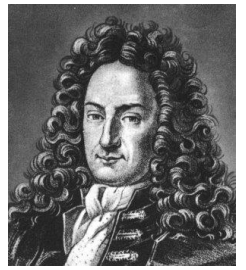
→ *denser encoding of information*

use of compact/reduced linguistic forms
(e.g. compounds, reduced relative clauses)

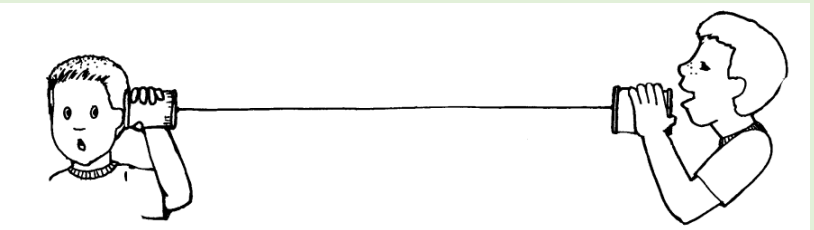
professionalization, institutionalization

→ *conventionalization*

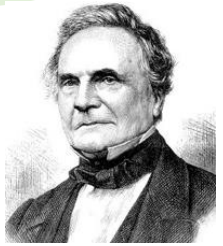
use of fairly fixed vocabularies
(terminology, formulaic expressions)



optimal code



for communication

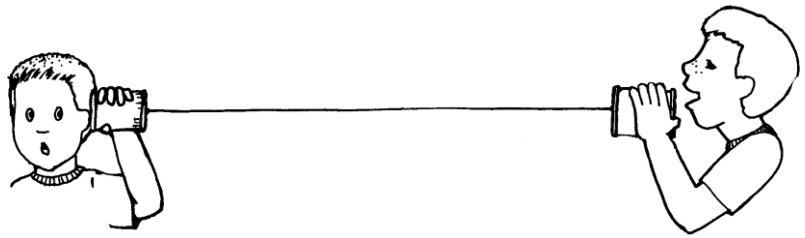


(cf. Aylett & Turk 2004; Jaeger 2010; Piantidosi, Tily & Gibson 2011)

Information Density and Linguistic Encoding

- choice of a particular linguistic encoding depends on (predictability in) context

Information Density (ID)



Surprisal

$$-\log_2 P(unit \mid \text{Context})$$

- contextually determined predictability is appropriately indexed by Shannon's notion of information

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

http://www.sfb1102.uni-saarland.de/

Which features are involved in diachronic change in scientific writing?
How can ID/surprisal help capture this change?

Feature detection: typicality

- Discover representative and distinctive features

relative entropy

Feature inspection: productivity

- Observe ambient context

average surprisal

Analyses

- grammatical typicality
- lexical productivity

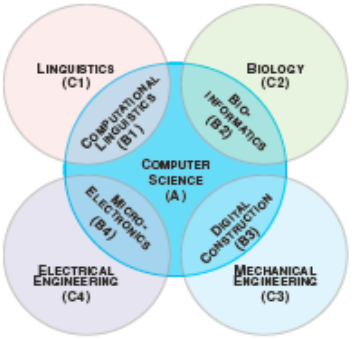
Data

SCIENTIFIC

1506–1700

Royal Society Corpus

PHILOSOPHICAL TRANSACTIONS



1970s & 2000s

EME (1500–1700) LME (1700–1900) ME (1900–present)

MIXED

HELSINKI CORPUS
1500–1700

EEBO

1641–1700

CLMET
1700-1920

BROWN FAMILY
1931, 1961,1991

ARCHER
1600–1999

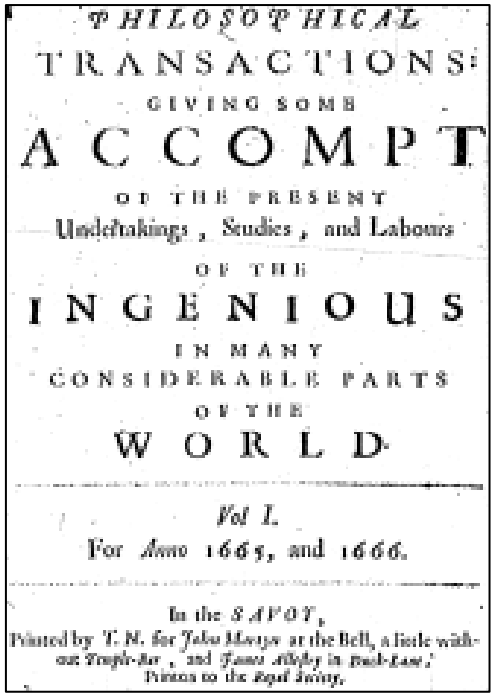
1640–1740

Google books

Royal Society Corpus (RSC)

Journal\Type	brv	fla	mis	nws	total
Philosophical Transactions (1665-1678)	124	641	154	–	919
Philosophical Transactions (1683-1775)	154	3903	338	–	4395
Philosophical Transactions of the Royal Society of London	–	2531	283	–	2814
Abstracts of Papers Printed ...	–	1316	15	–	1331
Abstracts of Papers Communicated ...	–	429	5	–	434
Proceedings of the Royal Society of London	–	1476	38	14	1528
total	278	10296	833	14	11421

book reviews
articles
miscellaneous
obituaries


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 Pauls-Lane,
Printers to the Royal Society.

size: ca. 35 million tokens, source: XML (JSTOR)

1-, 10-, 50-year time periods

(Kermes et al. 2016)

available from: CLARIN-D

Feature detection

Typical linguistic features

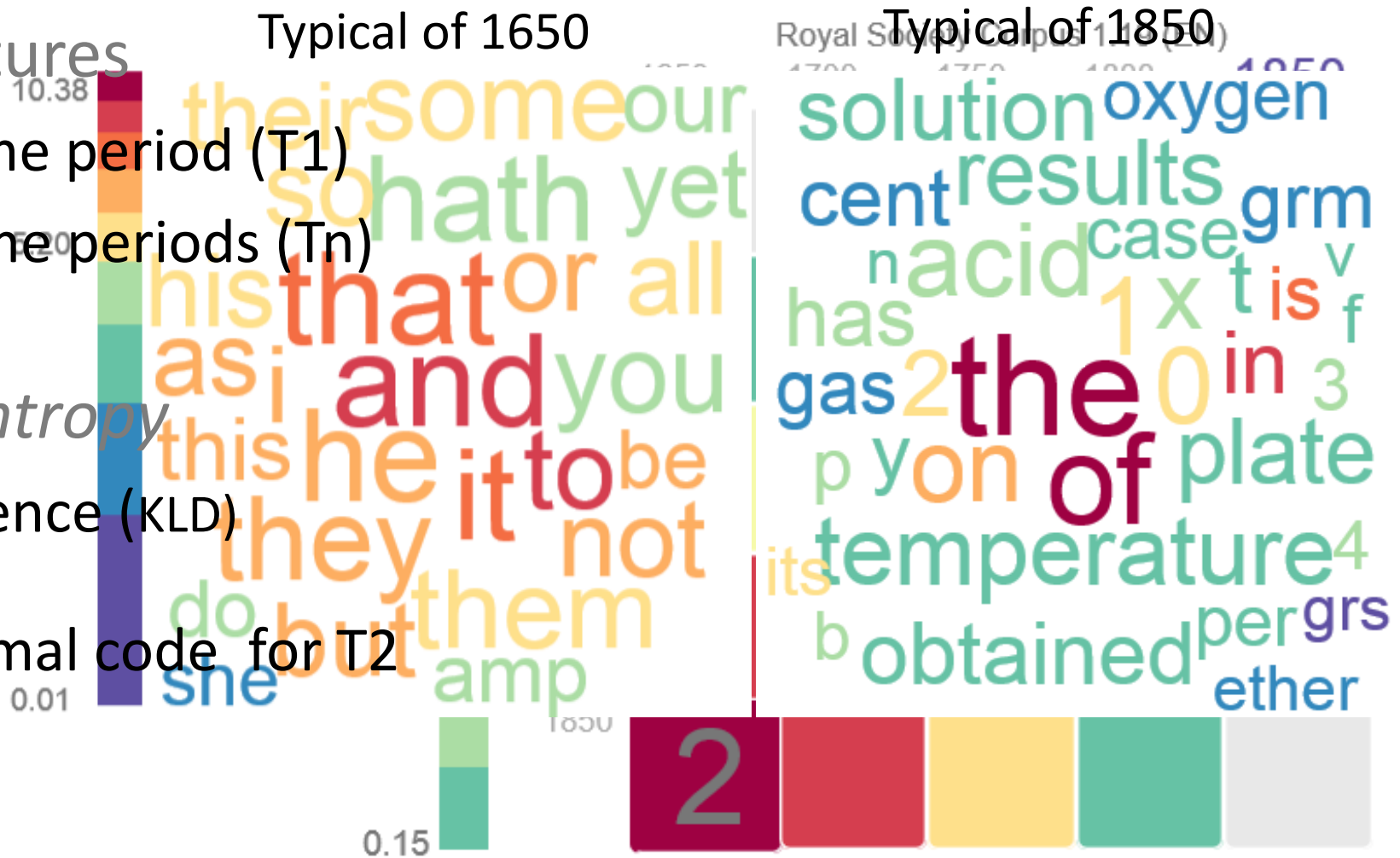
- *representative* of a time period (T1)
- *distinctive* to other time periods (Tn)

Measure: *Relative Entropy*

Kullback-Leibler Divergence (KLD)

→ encode T1 with optimal code for T2

(Fankhauser et al. 2014)



Feature inspection

Typical features in context

In a given context, a unit has
low predictability → high surprise
high predictability → low surprise

Measure: *Average Surprisal* (AvS)

$$AvS(unit) = \frac{1}{|unit|} \sum_{i} -\log_2 p(unit|context_i)$$

(Genzel & Charniak 2002; Degaetano-Ortlieb et al. to appear)

I Cannot ***enough wonder*** *at the* ***strange agreement*** *of the* ***thoughts*** *of that* ***acute French*** *Gentleman, Monsieur Auzont, in the Hypothesis of the* ***Comets motion***, *with* ***mine***;

Analyses

Typical lexico-grammatical patterns over time

A1: Grammatical typicality
A2: Lexical productivity

A1: Approach

Approximate grammatical patterns with POS-trigrams

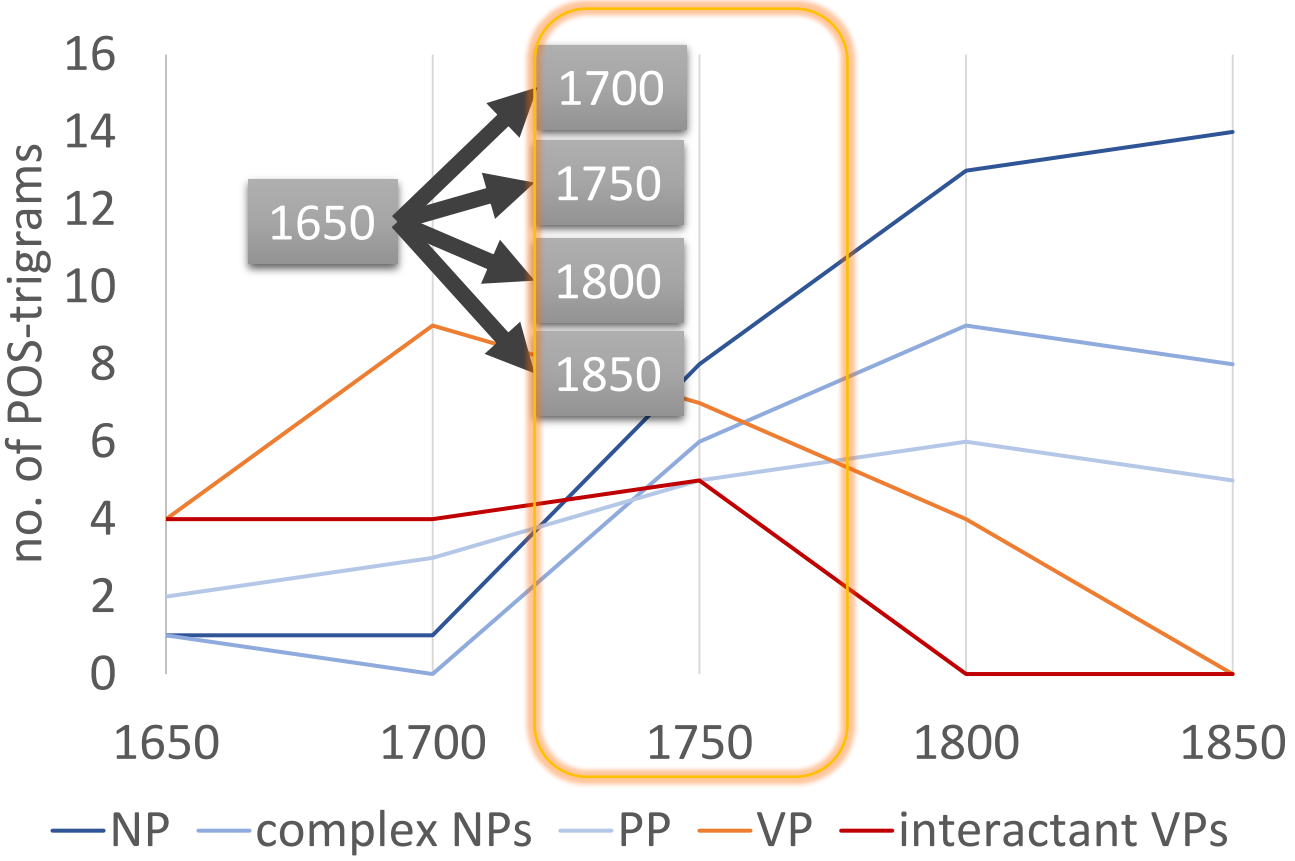
$$D_{KL}(T1||T2) = -\sum_i p(trigram_i|T1) log_2 \frac{p(trigram_i|T1)}{p(trigram_i|T2)}$$

- Extract all POS-trigrams from corpus
  (e.g. Det-Adj-N such as *a few minutes*)
- Exclude sentence markers, symbols and foreign words
- Group into phrase types (NP, VP etc)

A1: Grammatical typicality

Comparison

Typical phrase types over time



verbal (interactional)

→

nominal (informational)

A1: Grammatical typicality

1750s – a period of transition

1750 vs.	**phrase type**	**example**
1650/1700	NP	*the freezing point, the same time*
	NP (complex)	*the effects of, degree of heat*
	PP	*of nitrous air, an account of*
	AdjP (complex)	*small quantity of, great number of*
1800/1850	VP	*repeated the experiment, found the sum*
	VP (interact.)	*I then took, I did not*
	VP (modal)	*could not perceive/find*
	NP (complex)	*part of it, end of it*
	Cl (interact.)	*as I found*

new: informational

- nominal patterns
- *specialized* terminology

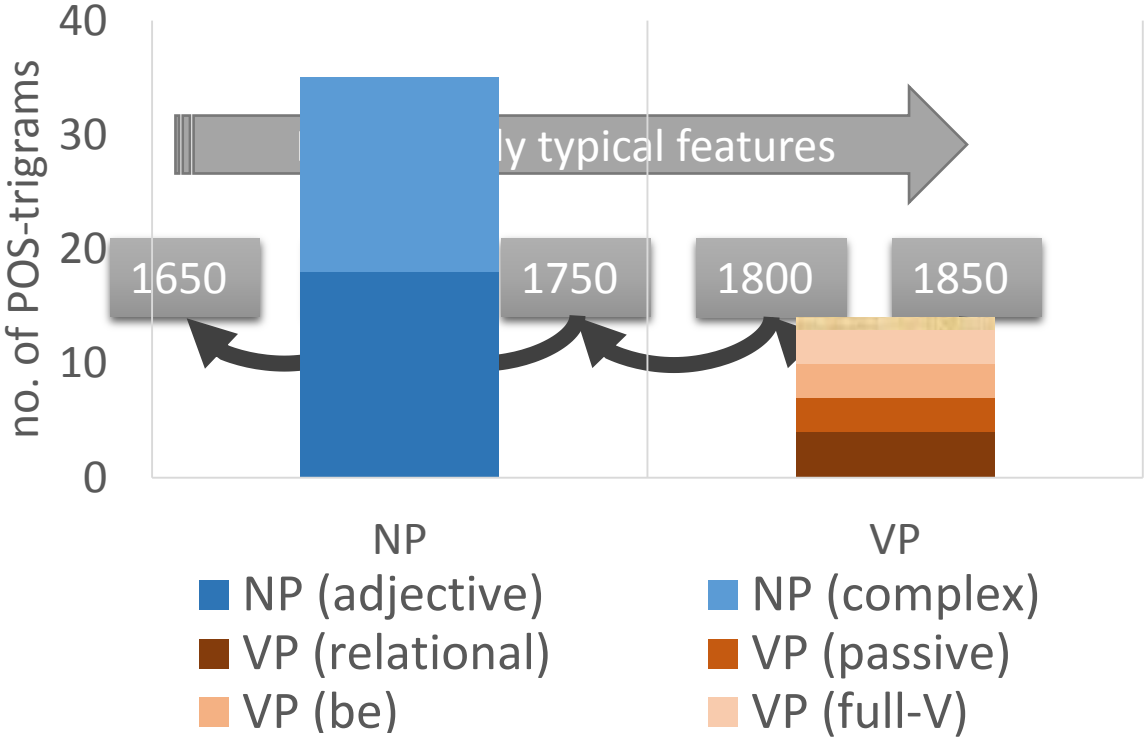
old: interactional

- verbal patterns (interactant, past tense, modals), *author-as-agent*
- nominal patterns referring to *general concepts*

(1751: Royal Society starts review process → *professionalization*)

A1: Grammatical typicality

Increasingly typical phrase types



NPs related to terminology
- NP adjective (*carbonic/muriatic acid gas*)
- NP complex (*center of gravity*)

VPs related to "scientific" style
- VP relational (*star is blue, r is odd*)
- VP passive (*effect is produced, account is given*)
- VP full-V (*the author concludes/states/considers*)
- VP red-rel (*effect produced by, line drawn from*)

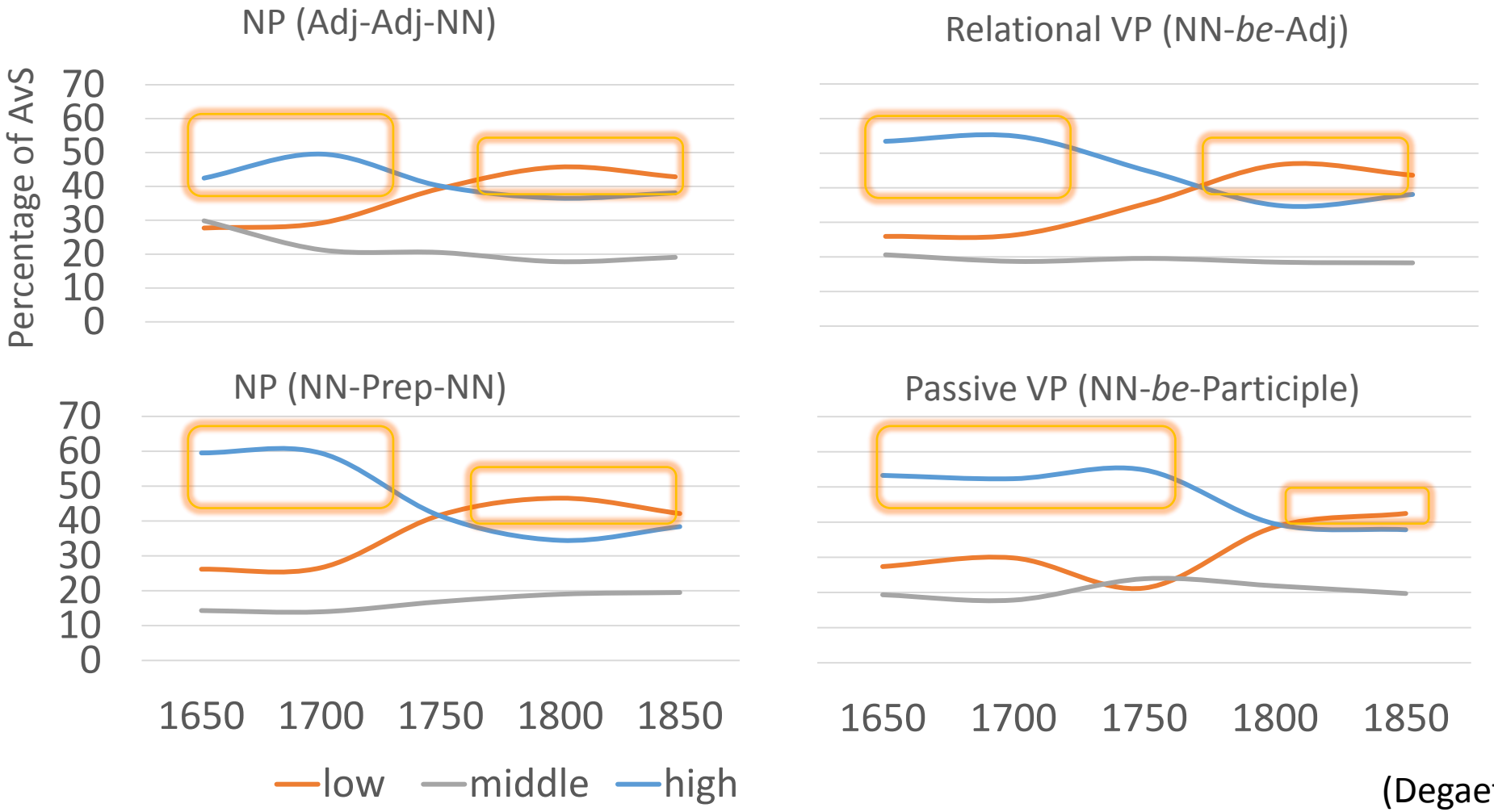
A2: Lexical productivity

- Inspect increasingly typical POS-trigrams in ambient context
- Do they become more or less productive?

High AvS values indicate higher productivity → type variation
i.e. grammatical patterns attract new lexical items (and spread to new contexts)

Low AvS values indicate lower productivity → conventionalized use
i.e. grammatical patterns are used in same contexts

A2: Lexical productivity



From 1650 to 1750:
quite unpredictable
→ high variation

From 1750 onwards:
more predictable
→ conventionalized use

(Degaetano-Ortlieb & Teich 2016)

A2: Lexical productivity

Lexical realizations of Adj-Adj-NN

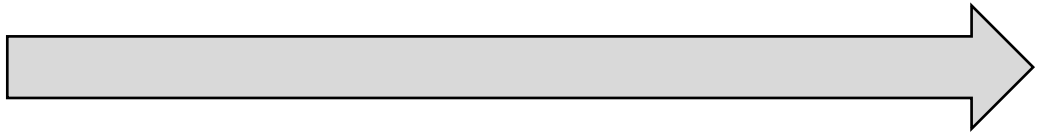
period	examples	freq. (pM)
1650	*dark brown colour*	7 (2.70)
	next foregoing tract	6 (2.32)
	cold fair weather	4 (1.55)
1750	*obedient/obliged humble servant*	135 (23.25)
	heavy/light inflammable air	110 (17.47)
	diluted vitriolic acid	29 (6.96)
1850	*concentrated sulphuric acid*	104 (8.93)
	carbonic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

- from general to specific concepts
- towards terminological patterns that increase in frequency

Results (so far)

1750

interactant verbal patterns

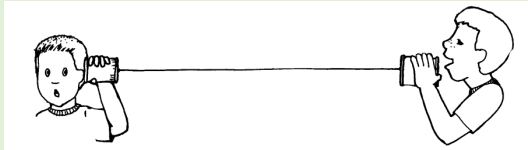


complex nominal patterns
relational verbal patterns

→increasing use of denser forms of encoding

→conventionalization

optimal code



for communication

Approach (so far)

- Relative entropy
  - good for comparing whole corpora
  - reveals typical features (beyond simple frequency)
  - indicator of degree of difference

Macro-Analysis: Contrast (Linguistics)

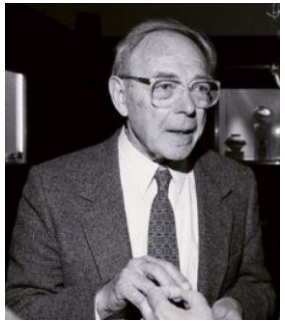
- Average Surprisal
  - good for inspecting ambient context
  - reveals differences in ID of alternative encodings
  - indicator of productivity

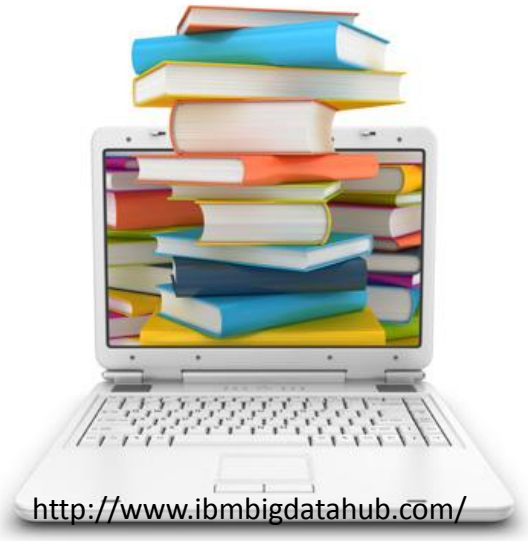
Micro-Analysis: Context (Philology)

A variationist‘s (*my*) dream come true

Linguistics (Science)

- valid generalizations
- modeling linguistic processes
- (psychology)



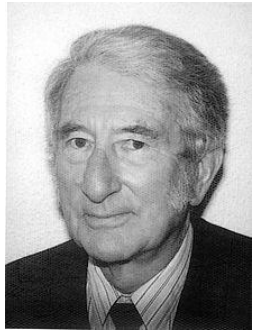


Informatics

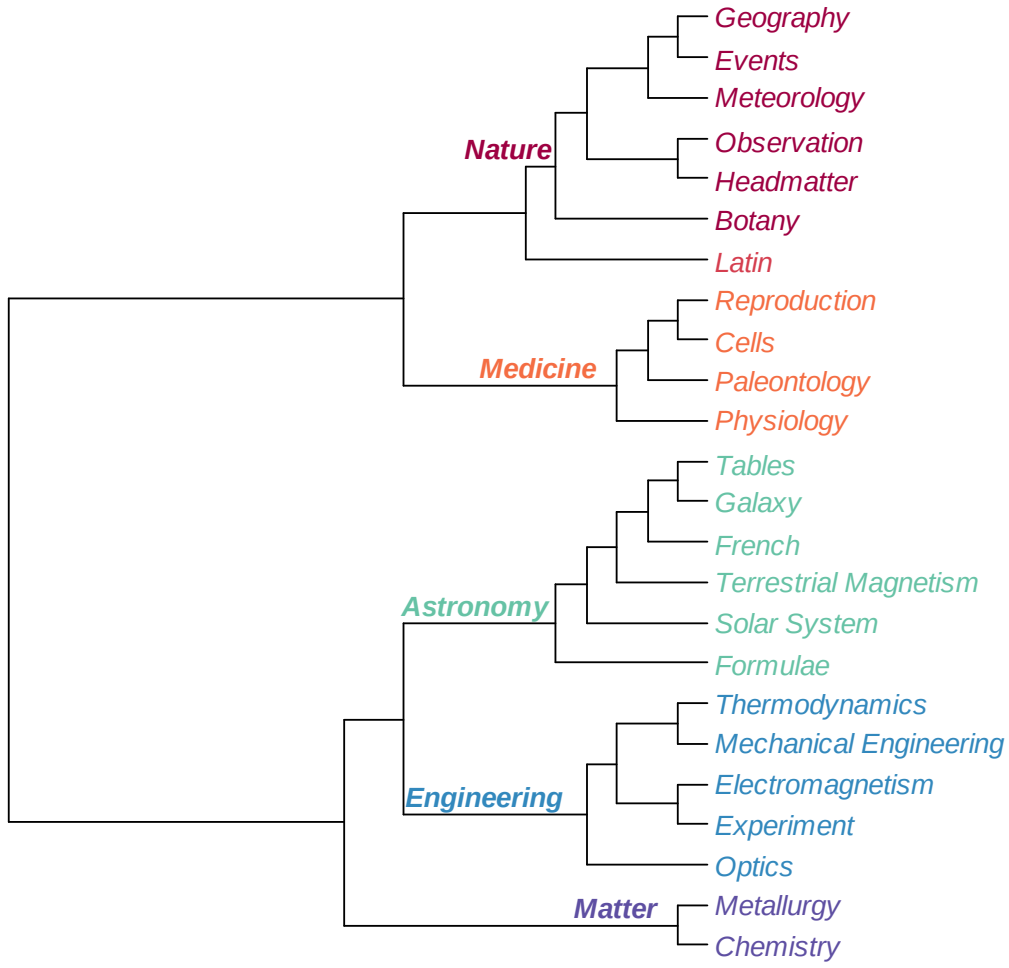
- modeling information processes
- robust, reliable systems
- (engineering)

Linguistics (Philology)

- textual scrutiny
- modeling linguistic experience in context (culture, time)
- (sociology)



Current & future work



water sea tide high found river coast north land tides miles height
great time account stone ground house fire letter place miles found
day ditto rain wind cloudy weather fair clear april year days night
great made make parts found body time small part water nature long
years year author society age number time royal life great letter
leaves plant plants tree tab bark folio foliis trees seeds seed flowers

quae quam sed ab sit vero hoc ac sunt esse qui etiam autem pro erit

cells animal blood fluid eggs membrane found egg part animals ova
fibres structure form surface portion cells anterior part section side
part bone bones teeth surface upper side lower anterior length
blood heart muscles part animal nerves vessels left parts stomach

weight water oo oz parts gr grain io grains fat increase weights grs
distance position stars star obs small hill double equatorial vf diff st
la le les des en du par dans qui il une qu pour ou ce sur ne
observations needle ship magnetic direct force made variation
sun time observations moon made observed difference observation
cos equation sin equal series point equations number line terms form

air water heat temperature experiments tube experiment glass made
made length weight end diameter iron instrument experiments brass
force electricity current wire action body power direction fluid motion
present general subject case results similar nature author state result
light rays glass eye red colours spectrum colour surface lines angle

water acid salt grains quantity iron found solution colour substance
acid water solution gas oxygen hydrogen carbonic cent action obtained

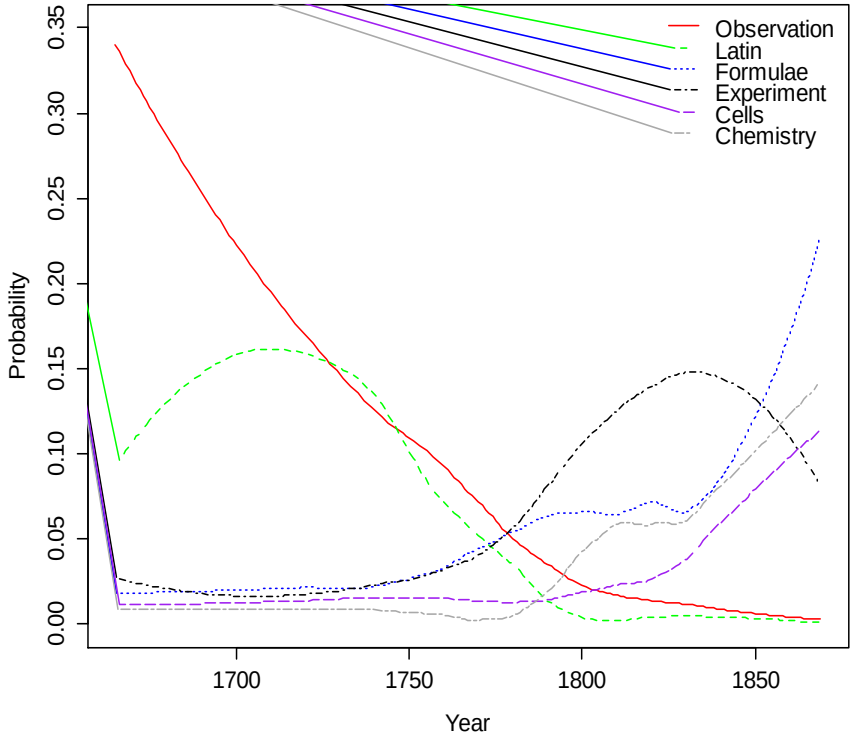
(Fankhauser et al. 2016)

Topic Models

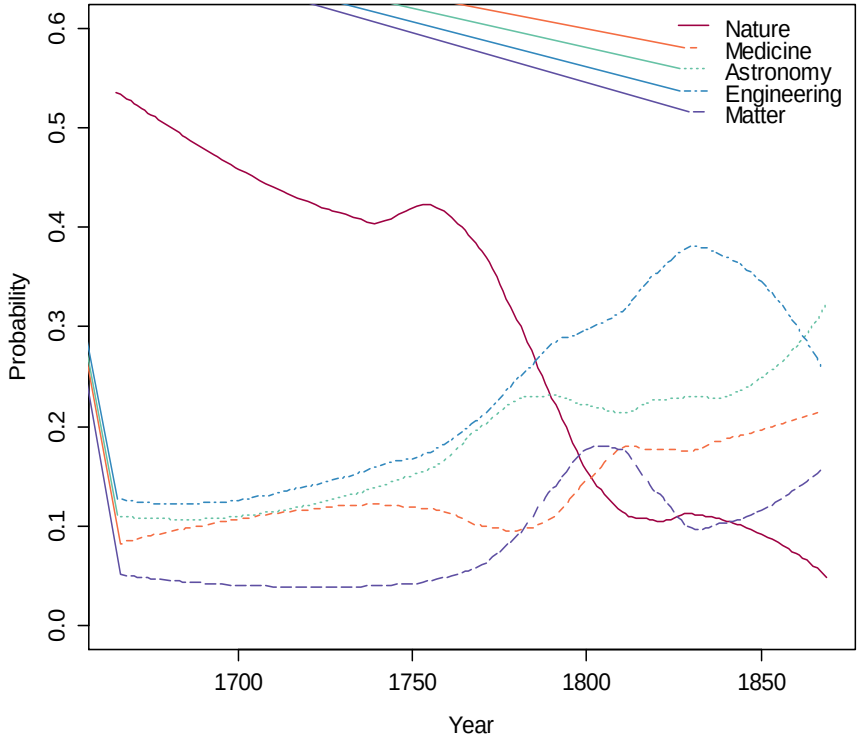
- Factorize Document-Word distribution $P(w_i|d)$ into Topic-Word distributions $P(w_i|z_k)$ and Document-Topic distributions $P(z_k|d)$

$$P(w_i|d) = \sum_k P(w_i|z_k)P(z_k|d).$$

- Dimensionality Reduction: Represent documents by means of 20-100 topics instead of > 100.000 types (different words)
- Here: 24 Topics on documents with stopword exclusion



Selected Topics



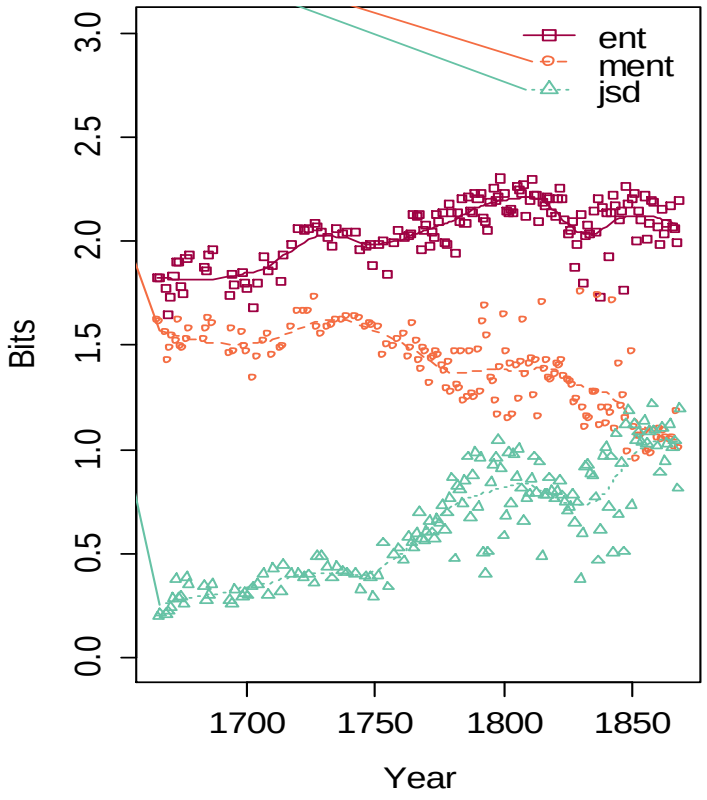
Topic Groups

- Observation
  - First half: Few, dominant topic (groups) – skewed distribution
  - Second half: many topic (groups) – even distribution
- Measure for topical diversification
  - Shannon-Entropy on topic distributions per year (*ent*)

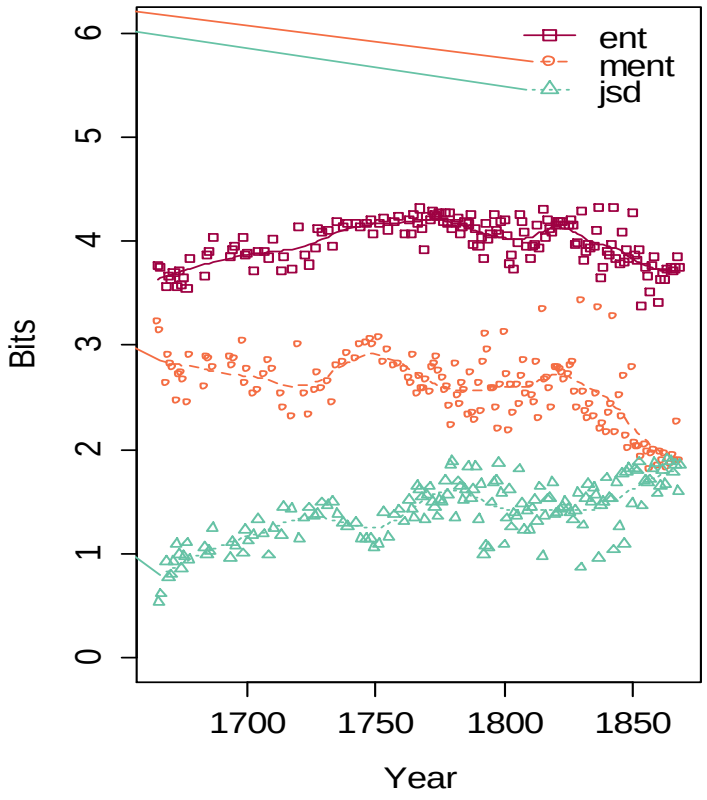
$$H(P_y) = -\sum_k P(z_k|y)\ log_2\ P(z_k|y)$$

  - Skewed -> low entropy, Even -> high entropy
- Measure for topical specialization
  - Mean Shannon-Entropy of *individual* document-topic distributions per year (*ment*)

$$H_{mean}(P_y) := 1/ny \sum_{d_j \in y} H(P_{d_i})$$



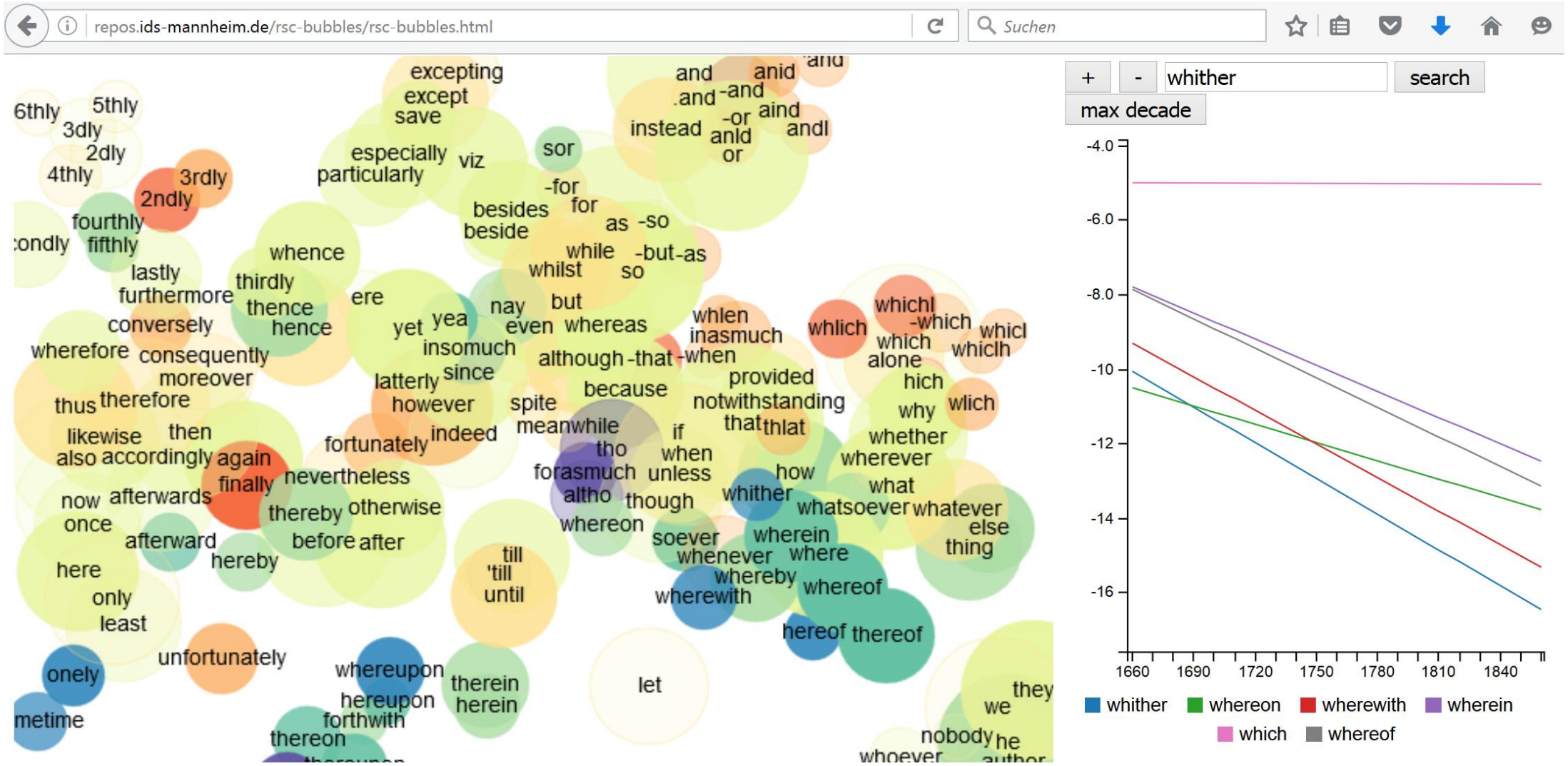
Topic Groups



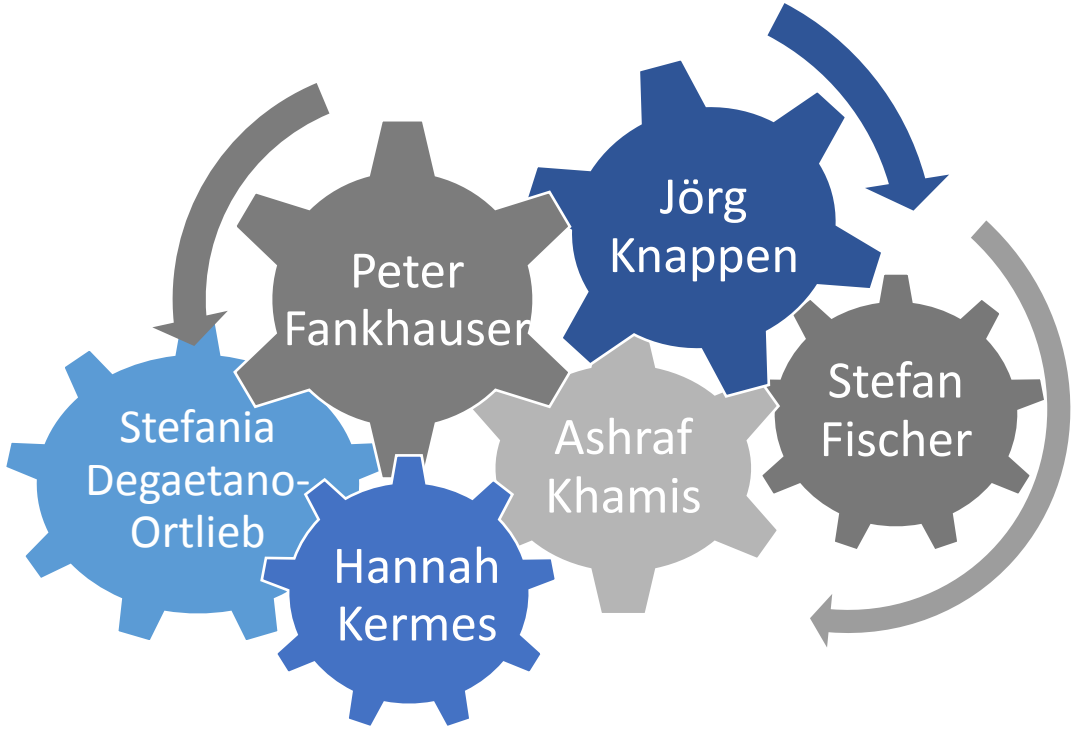
Topics

- Topical evolution of science
  - Diversification: more evenly distributed topics overall: *ent* increases
  - Specialization: topically more specific documents: *ment* decreases
  - Balance: Overall Complexity (*ent*) vs. Individual Complexity (*ment*): *ent* – *ment* = (generalised) Jensen Shannon Divergence *jsd* increases
- Complexity balance as a general principle?
  - Other distributions/models (e.g. word/pos ngrams)
  - Other contextual dimensions (e.g. register)
  - Example (Brown family): Registers become more diverse, documents more specific

Word embeddings (word2vec)



Thanks to team and sponsors



References

Matthew Aylett and Alice Turk. The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech *Language and Speech* 47:31-56, March 2004 .

Matthew W. Crocker, Vera Demberg, and Elke Teich. Information Density and Linguistic Encoding (IDeaL). *KI - Künstliche Intelligenz*, 30(1):77-81, 2016.

Stefania Degaetano-Ortlieb and Elke Teich. Information-based modeling of diachronic linguistic change: from typicality to productivity. In *Proceedings of Language Technologies for the Socio-Economic Sciences and Humanities* (LATECH'16), ACL, Berlin, Germany, 2016.

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis and Elke Teich. An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English. In Carla Suhr, Terttu Nevalainen and Irma Taavitsainen (eds), *Selected Papers from Varieng - From Data to Evidence* (d2e) 2015, Helsinki, to appear.

Peter Fankhauser, Jörg Knappen, and Elke Teich. Exploring and Visualizing Variation in Language Resources. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC), Reykjavik, 2014.

Peter Fankhauser, Jörg Knappen and Elke Teich. Topical Diversification over Time in the Royal Society Corpus, *Proceedings of DH 2016*, Krakow, Poland, 2016.

Dmitriy Genzel and Eugene Charniak. Entropy rate constancy in text. In *Proceedings of ACL*, Philadelphia, 2002.

References

John Hale. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics,* volume 2, pages 159–166, Pittsburgh, PA, 2001.

Florian T. Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology* 61(1):23-62, 2010.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen and Elke Teich. The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC), Portorož, Slovenia, 2016.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3):1126–1177.

Steven T. Piantadosi, Harry Tily and Edward Gibson. Word lengths are optimized for efficient communication, *Proceedings of the National Academy of Sciences* 108(9):3526-3530, 2011,.