The taming of the data:
Using text mining in building a corpus for diachronic analysis

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

IDeaL
SFB 1102

UNIVERSITÄT
DES
SAARLANDES

Background

- Typically poorly contextualized
  some meta-data (e.g. time, book title)
  no structural data
  no linguistic data
- Considered of limited value for linguistic analysis

- Annotated with
  meta-data (e.g. variety, register and time)
  structural data (e.g. page, section)
  linguistic data (e.g. pos, lemma)
- Readily usable for analysis

Google n-grams example

Taming of the shrew. Winter's tale. Comedy of errors
https://books.google.de/books?id... - Diese Seite übersetzen
William Shakespeare, Joseph Dennie, Isaac Reed - 1805 - Lesen - Mehr Ausgaben

Taming of the shrew. All's well that ends well. Twelfth ...
https://books.google.de/books?id... - Diese Seite übersetzen
William Shakespeare - 1807 - Lesen

Taming of the shrew. All's well that ends well
https://books.google.de/books?id... - Diese Seite übersetzen
William Shakespeare, Samuel Johnson, George Steevens - 1788 - Lesen

The Taming of the Press: Cohen V. Cowles Media Company
https://books.google.de/books?isbn... - Diese Seite übersetzen
Elliot C. Rothenberg - 1999 - Vorschau - Mehr Ausgaben
Cohen v.

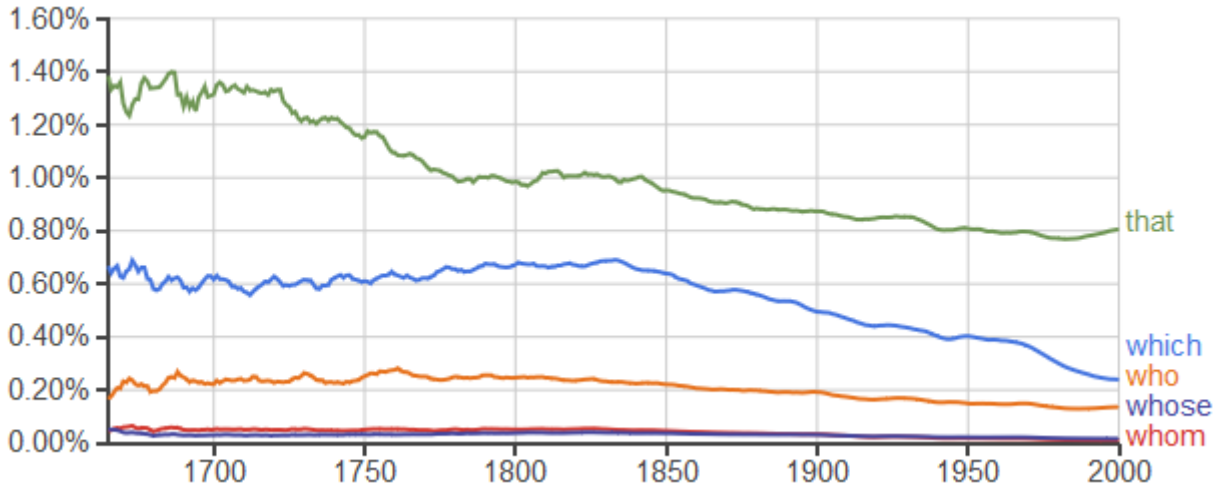
The Taming of the Shrew
https://books.google.de/books?isbn... - Diese Seite übersetzen
William Shakespeare, Harold James Oliver - 1999 - Vorschau - Mehr Ausgaben
A comedy of Petruchio's determination to subdue the irascible Katherine and to make her his wife

The Taming of the Samurai: Honorific Individualism and the ...
https://books.google.de/books?isbn... - Diese Seite übersetzen
Eiko Ikegami - 1995 - Vorschau - Mehr Ausgaben
This book demonstrates how Japan's so-called harmonious collective culture is paradoxically connected with a history of conflict.

Insights:

- Development over time
- More variation of the phrase *taming of the NOUN* over time

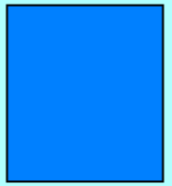
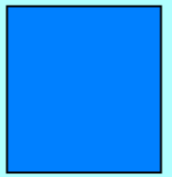
Google n-grams vs. Brown corpora



A resolution of some cases of conscience which respect ...
https://books.google.de/books?id... - Diese Seite übersetzen
William Sherlock - 1683 - Lesen - Mehr Ausgaben

An abstract of the grievances of trade which oppress our poor...
https://books.google.de/books?id... - Diese Seite übersetzen
Abstract - 1694 - Lesen

- Diachronic development
- No linguistic distinction possible (*that*)
- No context for inspection

1961	1991
	
11835	11190
2.34	2.37
5048.17	4713.65

by the huge amount of preliminary work	**that**	had to be got through and was actually b
am a cordial supporter of the decision to	**which**	he has come " and so forth , while Minis
round table conference in India . And by	**whom**	had the words In " India " been inserted
th , as Mr Baldwin had wished , the truth	**that**	comes out a collision with hard realities
has received a letter from Mr Baldwin in	**which**	the Tory leader states : . The " principle

- Limited diachronic perspective
- Linguistic distinction possible (*that* as relativizer)
- Contextualized search (kwic, register, etc.)

Rationale

Assumption

- Scientific language becomes more informationally dense over time
- Due to specialization greater encoding density over time

→ shorter ling. forms used to maximize efficiency in communication

Approach

- Detection of linguistic features of densification
- Comparison across historical stages

Example

The use of this control method leads to a safer and faster train operation in the most adverse weather conditions.

→ more dense linguistic encoding

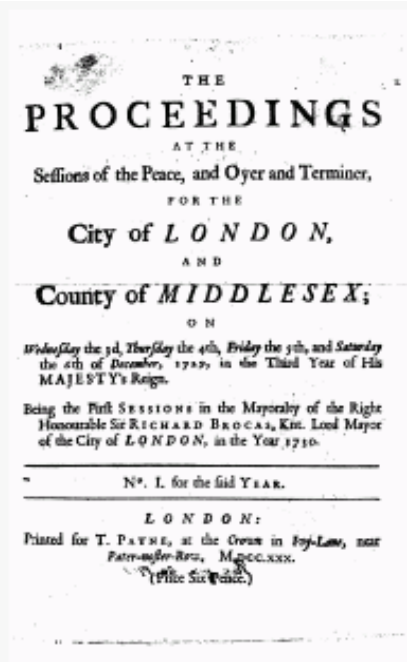
You can control the trains this way and if you do that you can be quite sure that they'll be able to run more safely and more quickly than they would otherwise, no matter how bad the weather gets.

→ less dense linguistic encoding

Building new corpora

Sources for new corpora → all relevant meta-, structural and ling. data?

Old Bailey sources vs. richly annotated corpus (Huber, 2007)



THE
PROCEEDINGS
AT THE
Sessions of the Peace, and Oyer and Terminer,
FOR THE
City of *LONDON*,
AND
County of *MIDDLESEX*;
ON
Wednesday the 3d, *Thursday* the 4th, *Friday* the 5th, and *Saturday* the 6th of *December*, 1729, in the Third Year of His MAJESTY's Reign.
Being the First SESSIONS in the Mayoralty of the Right Honourable Sir RICHARD BROCAS, Knt. Lord Mayor of the City of *LONDON*, in the Year 1730.
N°. I. for the said YEAR.
LONDON:
Printed for T. PAYNE, at the *Crown* in *Fox-Lane*, near *Paternoster-Row*, M.DCC.XXX.
(Price Sixpence.)

Keyword(s) which

And ○ Or ○ Phrase ◉ Advanced ○ what's this?

Surname

Given Name

Alias

Offence <All Offences>

Verdict <All Verdicts>

Punishment <All Punishments>

Search In <All Text>

Time Period *From (month/year)* Any Any *To (month/year)* Any Any

Reference Number

SEARCH

1. **Front matter from Proceedings, 29th April 1674.**

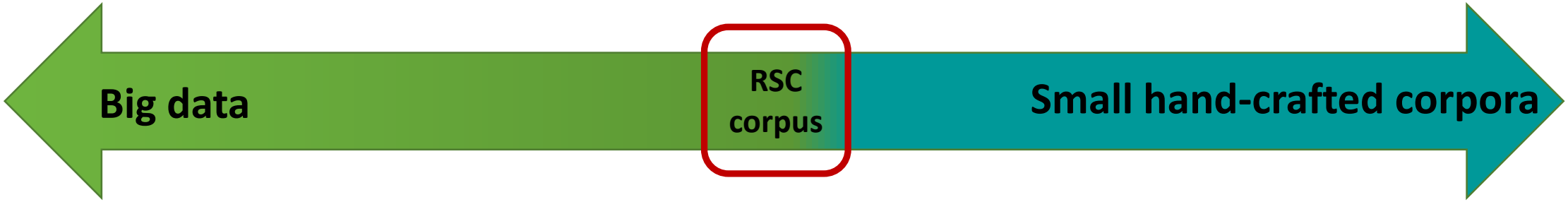
... ving are forced to seek out those ways and means, which either are destructive in themselves, or purchase shame and destruction in their end?

2. **Violent Theft > highway robbery, 29th April 1674.**

... in his other Pocket, pulled out about four pound, which they took from him, and going their way, a Neighbour of the said Mr. Stutely coming by followed them crying Thieves and with other help one of them w ...

Motivation

- Create a corpus from uncharted material
  - of the *Philosophical Transactions and Proceedings of the Royal Society of London* (RSC Corpus)
  - JSTOR material in XML
  - Containing some meta-data (e.g. time, title), but no structural data
- Enrich corpus with relevant meta-, structural, and linguistic data for diachronic linguistic analysis

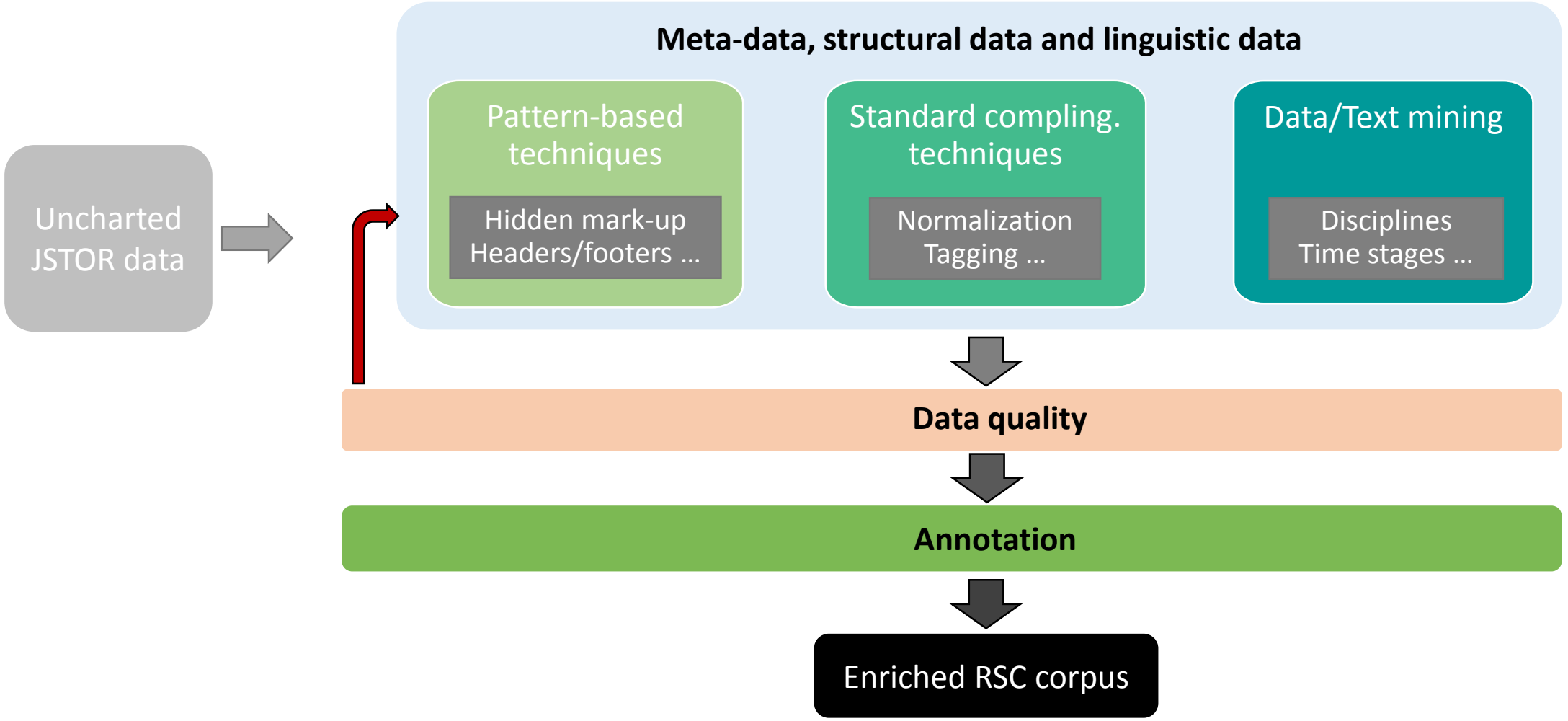


Royal Society Corpus (RSC)

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 (PTRSL)	1776–1869	–	2,531	283	–	2,814
Abstracts of Papers Printed in PTRSL	1843-1861	–	1,316	15	–	1,331
Abstracts of Papers Communicated to RSL	1862-1869	–	429	5	–	434
Proceedings of RSL	1862–1869	–	1,476	38	14	1,528
Total		278	10,296	833	14	11,421

Size: approx. **35 million tokens**
Source: **XML** (JSTOR)

Methods
From uncharted to enriched data



Pattern-based techniques

Structural data

- Uncover and clean hidden markup
- Identify article beginnings and endings and order scrambled pages
- Detect headers/footers, toc, errata

Data quality

- Detect and remove duplicates
- Eliminate OCR errors by adaptation of patterns from Underwood and Auvil (n.d.)
  (1,282 correction patterns)

<page>III. Experiments and Obserznations on the Production of Ligbt from dXerent Bodies, Zoy Heat and by Attrition. By Mr. Thomas \Vedgwood; cornmunicated; by Sir Joseph 13ankst Bart. P. R. S. Read December 22, 1791C BEFORE I begin to state the experimellts ^rhich are the subject of this Paper, it may

Standard comp. linguistic techniques

Normalization

- Spelling variation with VARD (Baron and Rayson 2008)
- Manual normalization of an extract of the RSC used to train VARD

Tokenization, segmentation, PoS tagging and lemmatization

- TreeTagger (Schmid 1994) + Perl scripts

Data quality

- Additions to abbreviation list of TreeTagger to improve segmentation

Standard comp. linguistic techniques

Feature extraction

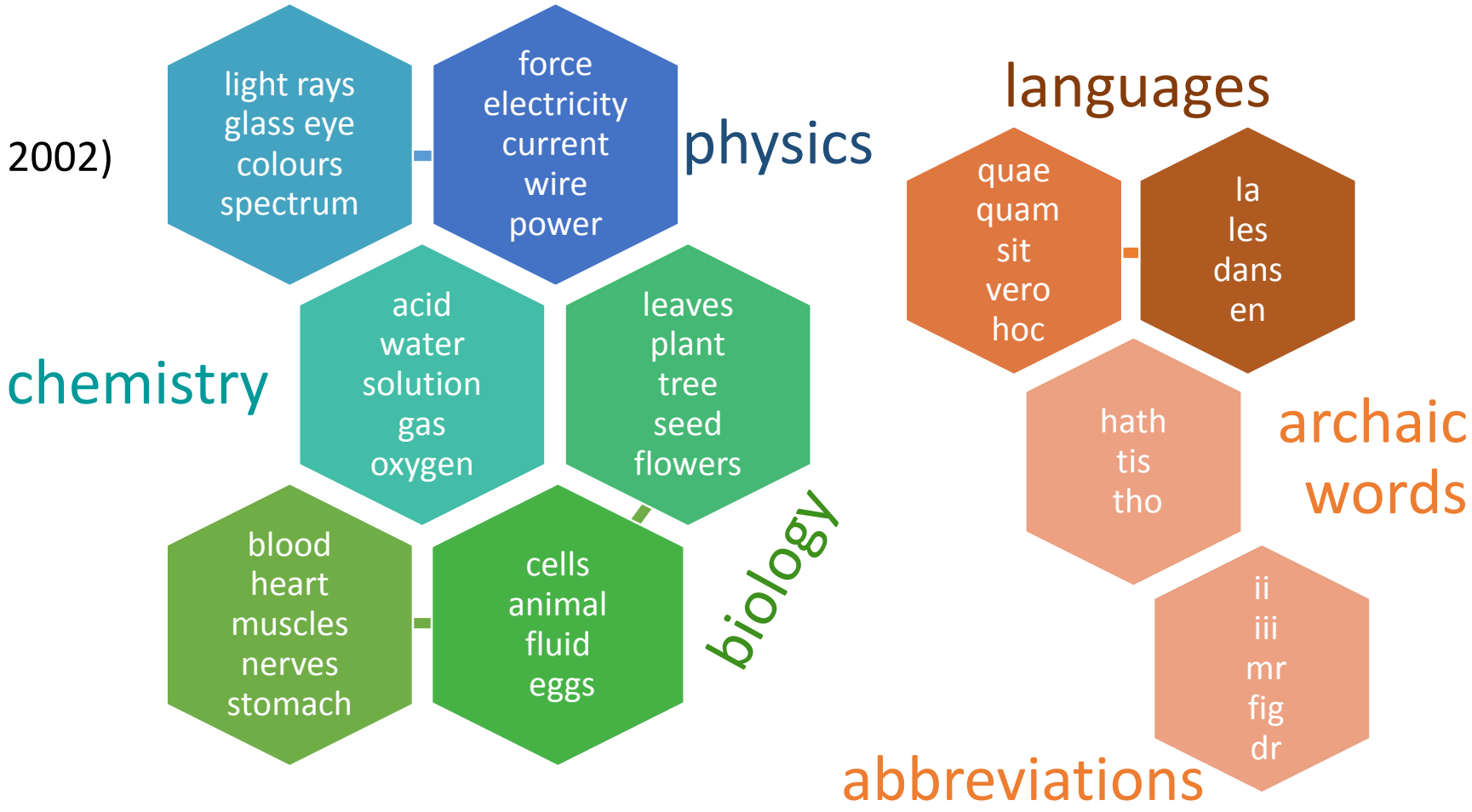
- Semi-automatic extraction of features relevant for diachronic analysis (Harris 1991) with CQP (CWB2010)
- Use of word/PoS sequences in manually designed macros

Feature	Extraction pattern	Example
Reduction by prefix by suffix	[lemma="anti-.*"] [pos="VV.*" & lemma="\w{1,}ify"]	*anti-rheumatic remedies* *surfaces solidify simultaneously*
Omission of relativizer	[pos="DT"][pos="N.*"][pos="P.*"][pos="V.*"]	*the Bodies ^ we are acquainted*
Nominalization	[pos="NN.*" & lemma="\w{1,}ness"]	*there is a Lake of that bigness*

Data mining – Discipline detection

Topic modeling

- MALLET (McCallum 2002)
- Limit of 24 topics



Data mining – Detection of time periods

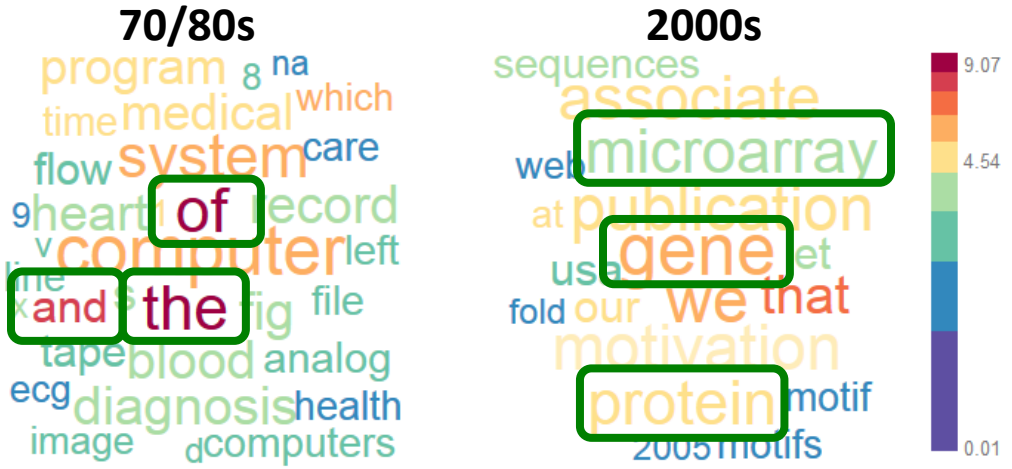
Distance measures

- Identification of ling. changes in corpora (Fankhauser et al. 2014a & 2014b)
- Based on Information Theory
  Kullback-Leibler Divergence (relative entropy)
  Unigram model + smoothing
- Assessing how typical an n-gram is to a corpus/subcorpus

Example: Bioinformatics Abstracts across time

→Function words typical for 70/80s

→Nominal (denser) style in 2000s



Data mining – Detection of time periods

Clustering

Variability-based neighbor clustering algorithm (Gries & Hilpert 2008)

- Detection of stages in diachronic data
- Tailored to specific linguistic phenomena

Piotrowski law

Language changes as a result of interaction between old forms and new forms

- Complete change
- Partial change
- Reversible change

Data mining – Feature detection

Classification/Ranking

- Classify time periods
- by linguistic features relevant for dense/less dense encodings
- Use feature weights to detect relevant features

Pattern mining

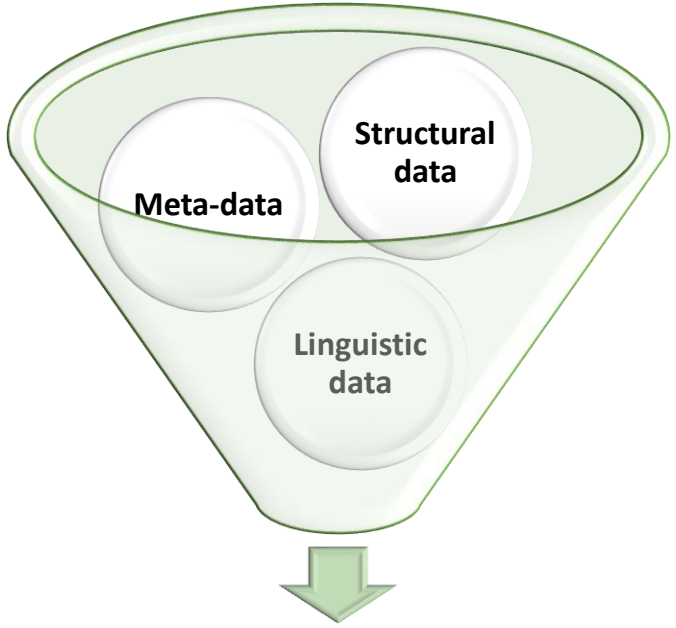
- *Squeeze* looks for interesting patterns (Vreeken 2010)
- *Desq* (Gemulla forthcoming)
  - Looks for patterns of a desired “form”
  - Makes use of a hierarchy (e.g. WordNet)

anti-SMTH → PersPron *opposes* SMTH

is against

…

Conclusions

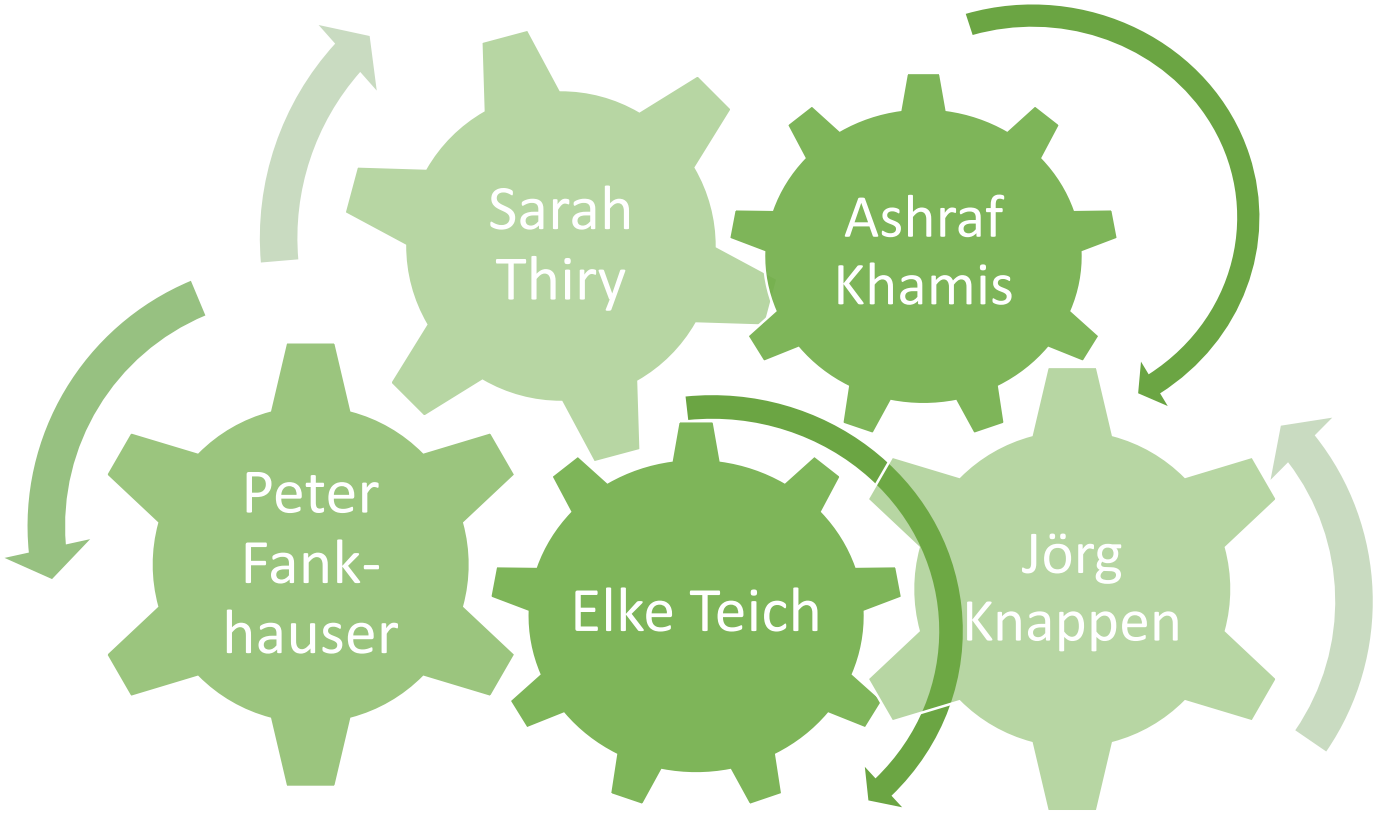


Royal Society Corpus

- High quality corpus from rel. big data with affordable automatic and manual effort
- Continuously improve data quality
- Tailored to linguistic research
  - → Comparison of historical stages / disciplines over time
  - → Inspection of linguistic features of densification

Thank you for your attention!

Thanks to the team!



IDeaL
SFB 1102