Data Mining Tools for Analysis of Linguistic Variation

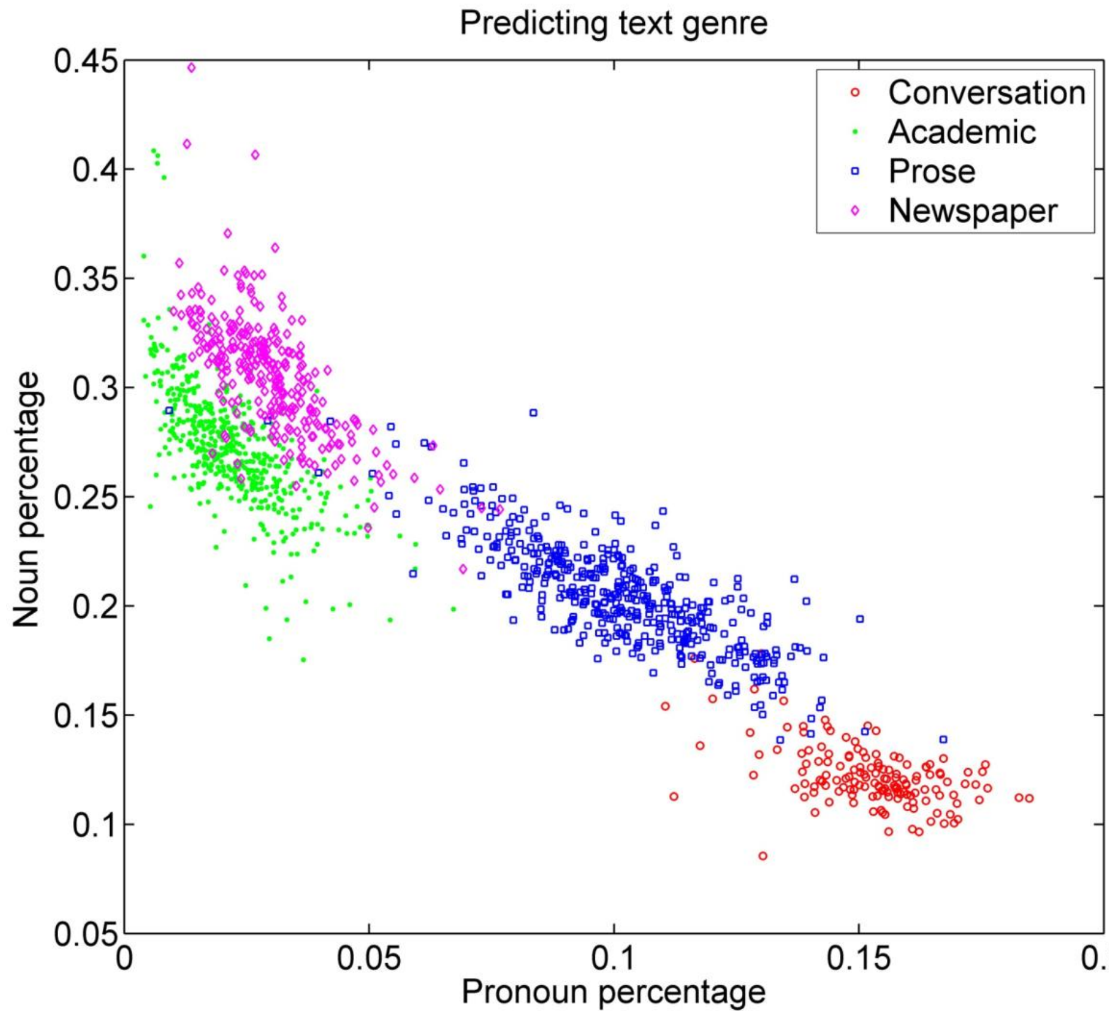
Jefrey Lijffijt

The Dammoc Project

Over the past decades, linguists have compiled large electronic **text corpora** of various kinds, enabling the study of diverse aspects of language. The development of **tools** for analysis of corpora has received far less attention. In a combined effort with researchers in **data mining**, **linguistics** and **information visualization**, we develop advanced and interactive tools, specifically for analysis of natural language corpora. We use these tools to study differences in **writing style** throughout genres in modern texts, and development of genres and **language change** starting at Early Modern English (ca. 1400).

Finding Differences In Writing Style

Similarity measures and genre



Why

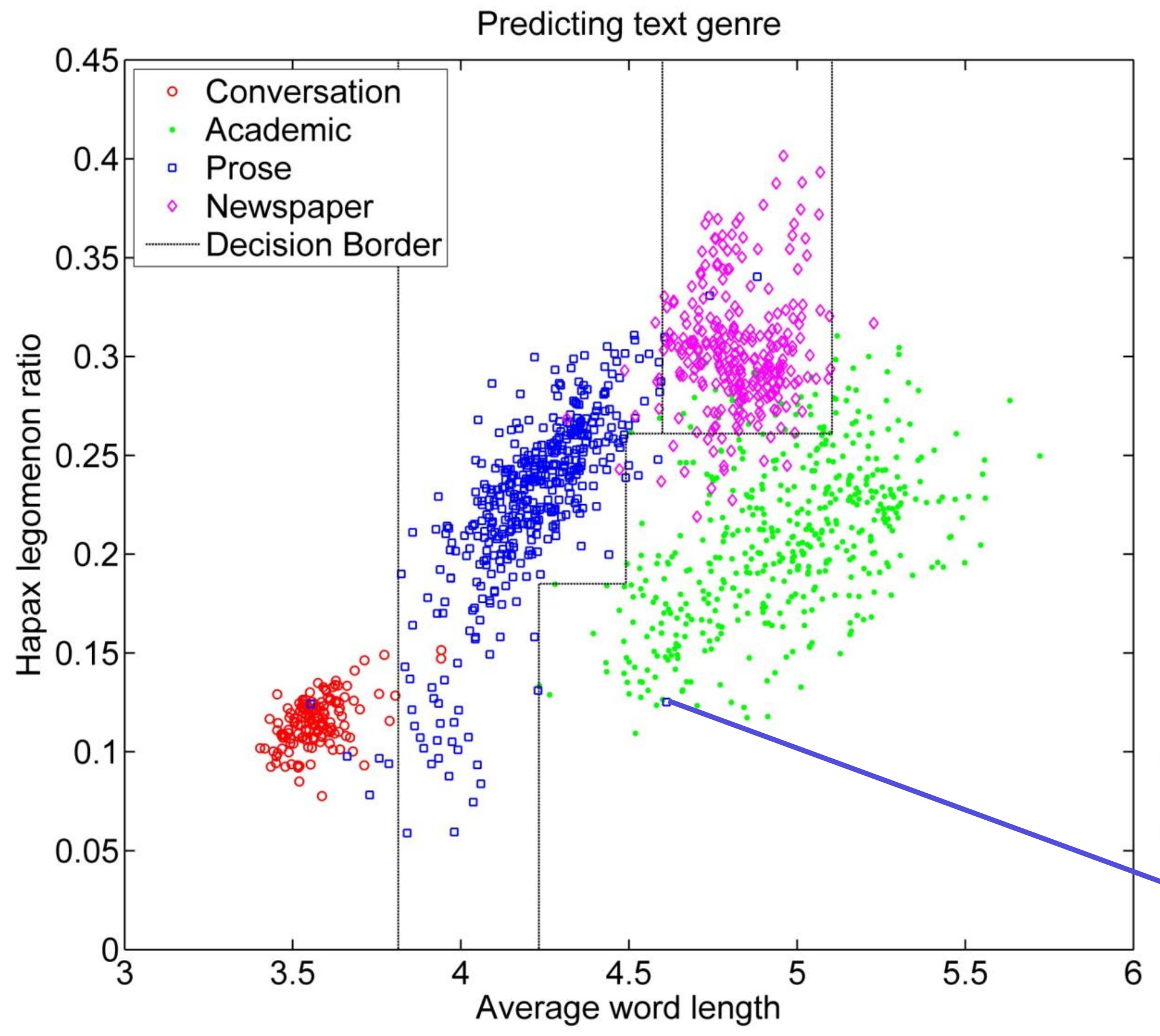
- Sampling homogeneity
- Improve insight
- Search for similar texts

Result

- > 90% Accurate
  - Novel result
- Strong cluster structure
  - Simpler measures

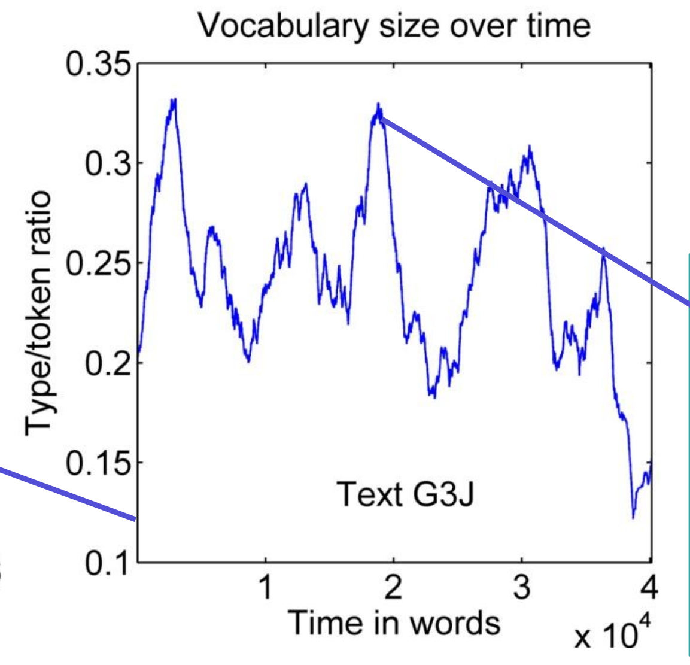
How

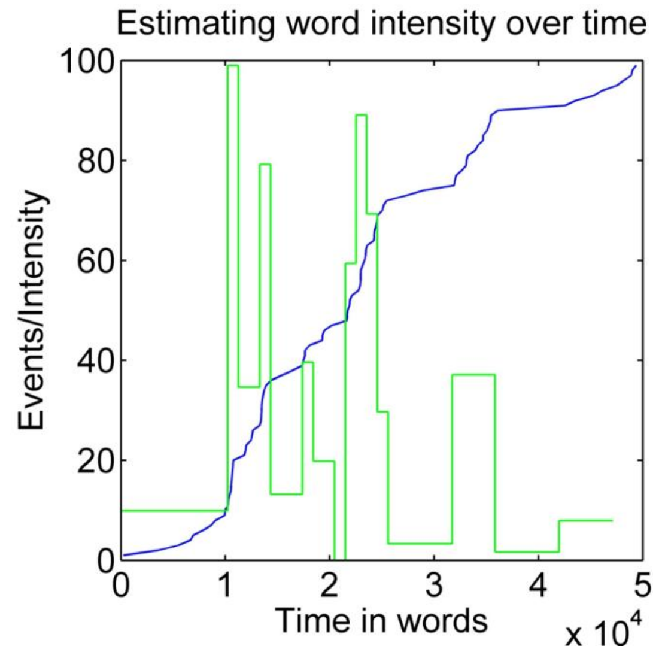
- Simple measures
  Noun % / Pronoun %
- Simple ML-algorithms



Interactive tools

- Get more information
- Identify interesting texts (parts)
- Multi-dimensional





The_AT0 procedure_NN1 provided_VVN by_PRP subs_NN2 ._PUN
Words: 19003-19007

(_PUL 2_CRD)_PUR is_VBZ not_XX0 open_AJ0 to_PRP him_PNP.
Words: 19008-19013

If_CJS his_DPS bargain_NN1 is_VBZ concluded_VVN between_PRP quarterly_AJ0-NN1 meetings_NN2 ,_PUN he_PNP must_VM0 wait_VVI until_CJS he_PNP obtains_VVZ a_AT0 transfer_NN1 at_PRP the_AT0 next_ORD quarterly_AJ0-NN1 meeting_NN1 of_PRF the_AT0 licensing_NN1-AJ0 board_NN1 before_CJS he_PNP may_VM0 legally_AV0 carry_VVI on_AVP business_NN1 on_PRP his_DPS own_DT0 behalf_NN1 under_PRP the_AT0 licence_NN1 ._PUN
Words: 19014-19052

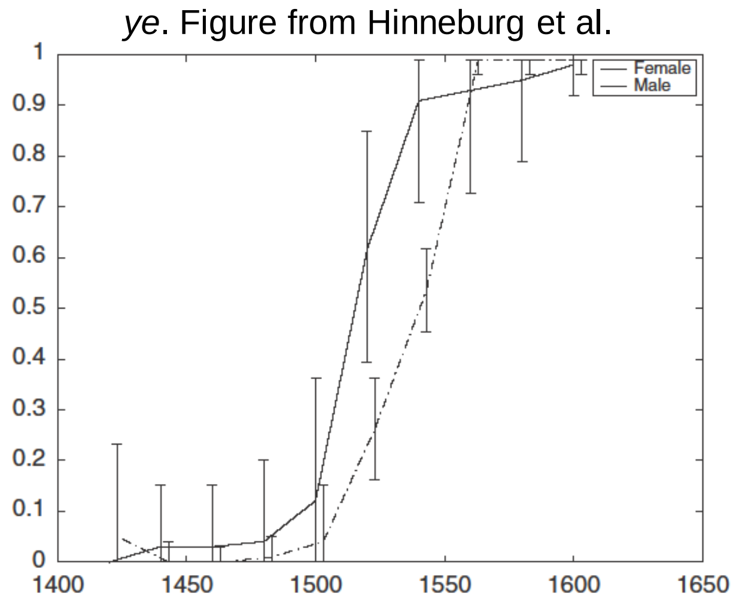
Assessing How The English Language Has Changed

Early Modern English

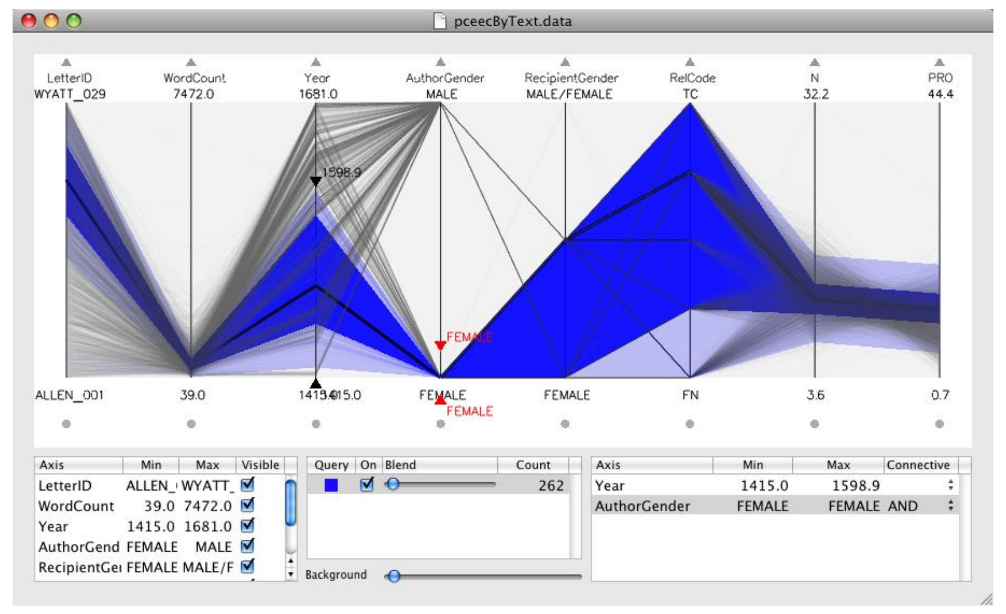
- Understanding change
- Who leads change
- What influences writing style
  - Gender
  - Social class
  - Age

- Based on letter collections
  - Parsed Corpus of Early English Correspondence
  - Penn-Helsinki Parsed Corpus of Early Modern English
- Based on book collections
  - Early English Books Online

Bayesian estimates for frequency of *you* vs. *ye*. Figure from Hinneburg et al.



Parallel coordinates plot of PCEEC.

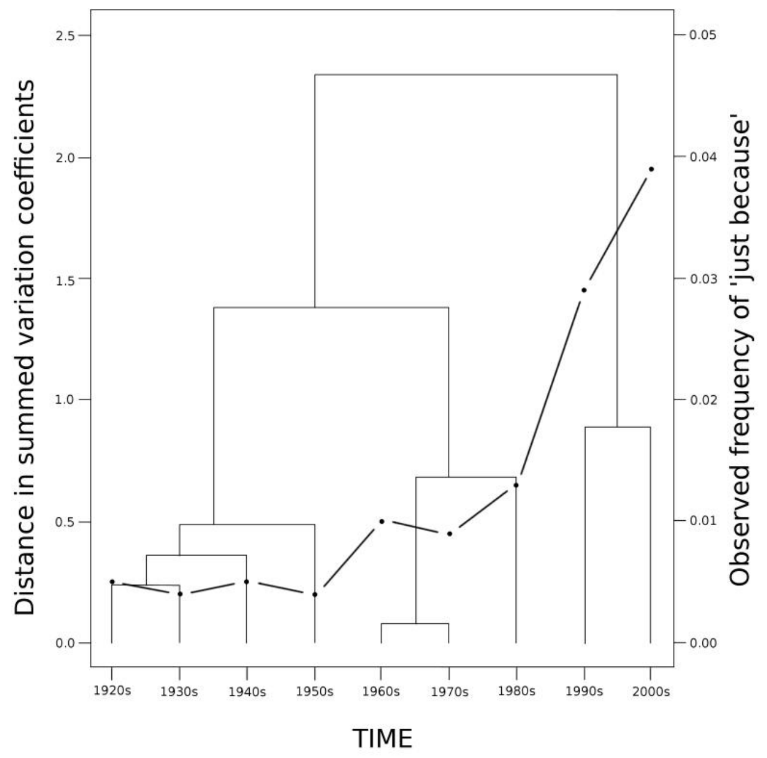


Frequency estimation

- Sparse data
- High dimensionality
- Solution: Bootstrapping or Bayesian estimation

Clustering

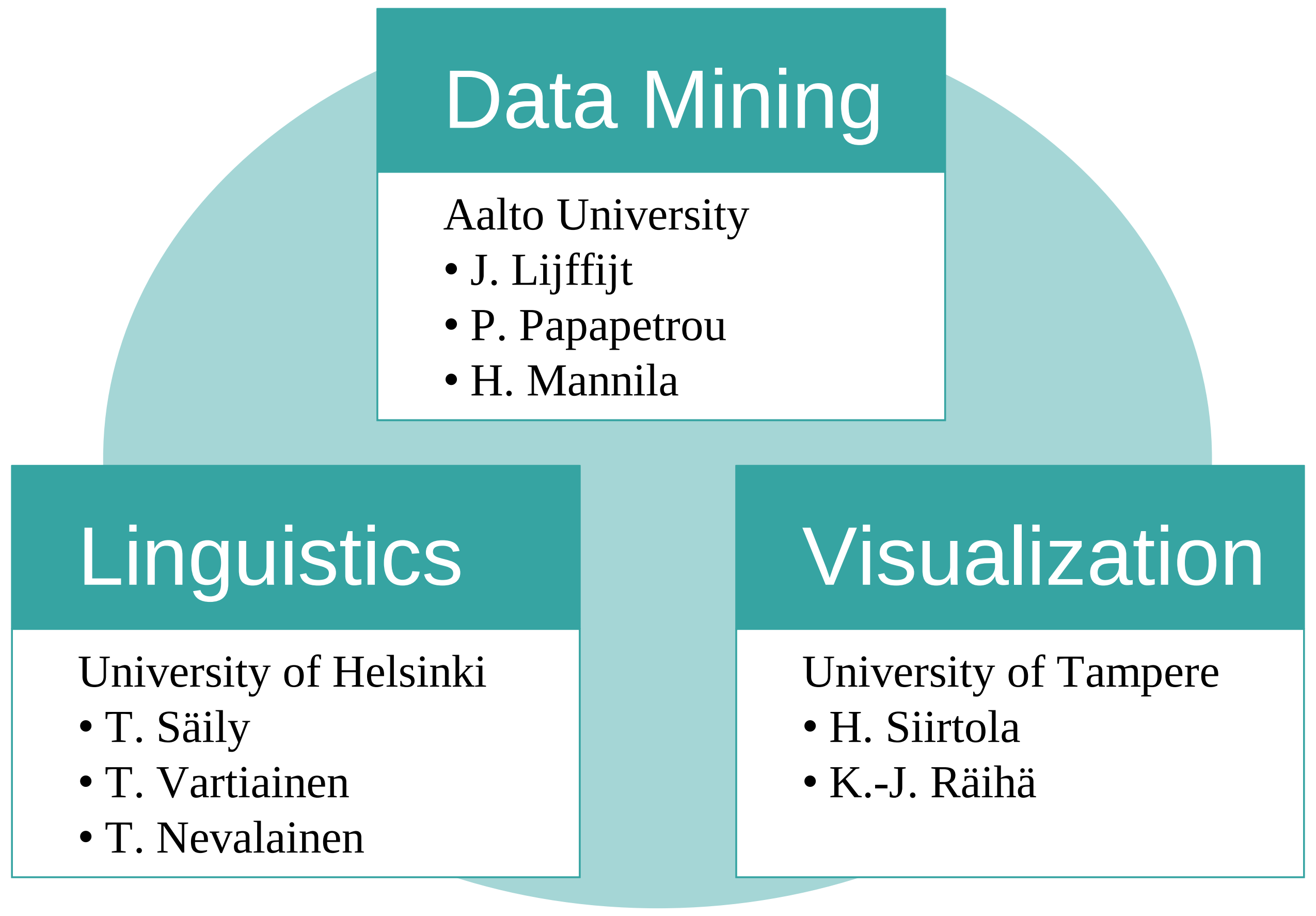
- Data-driven discovery of genres
- Evolution of genres
- Assess stages of change



Time-constrained clustering of a frequency change. Figure from Hilpert and Gries.

Complexity and productivity

- Measure productivity of
  - Suffixes
  - Prefixes
  - Grammatical constructions
- Part-of-speech tagging and grammatical parsing is often imperfect
- Finding all relevant instances
- Precision / recall estimation



Data Mining

Aalto University
- J. Lijffijt
- P. Papapetrou
- H. Mannila

Linguistics

University of Helsinki
- T. Säily
- T. Vartiainen
- T. Nevalainen

Visualization

University of Tampere
- H. Siirtola
- K.-J. Räihä

Contact Information

Jefrey Lijffijt
jefrey.lijffijt@tkk.fi
users.ics.tkk.fi/lijffijt

Department of Information
and Computer Science
Aalto University

Aalto University

ALGODAN Algorithmic Data Analysis

DAMMOC
DATA MINING TOOLS
for changing modalities of communication