Helsinki Computational History Group

“Computational history” refers to an integrated mixed methods approach to study large digitized historical sources. “Integrated” means that data science is combined to specialized subject knowledge; in the case of COMHIS, intellectual history and book history.

http://helsinki.fi/computational-history

Helsinki Computational History Group: Public Communication in Early-Modern Europe

Movement of ideas

- Metadata work based on several different library catalogues
- genres (poetry, pamphleteering); intellectual traditions (natural law tradition, ancient texts
- text reuse: genres (historical works, quoting practices)

Research data releases

- ESTC; Fennica; Kunglica; CERL; ECCO text reuse (+ EEBO text reuse); Finnish Newspapers

Conceptual change

- concepts are crucial, but not directly jumping into this for various reasons
- Theoretical underpinning (historians + linguists)
- Concepts as linguistic objects (linguists + historians + CS)

Tools for others

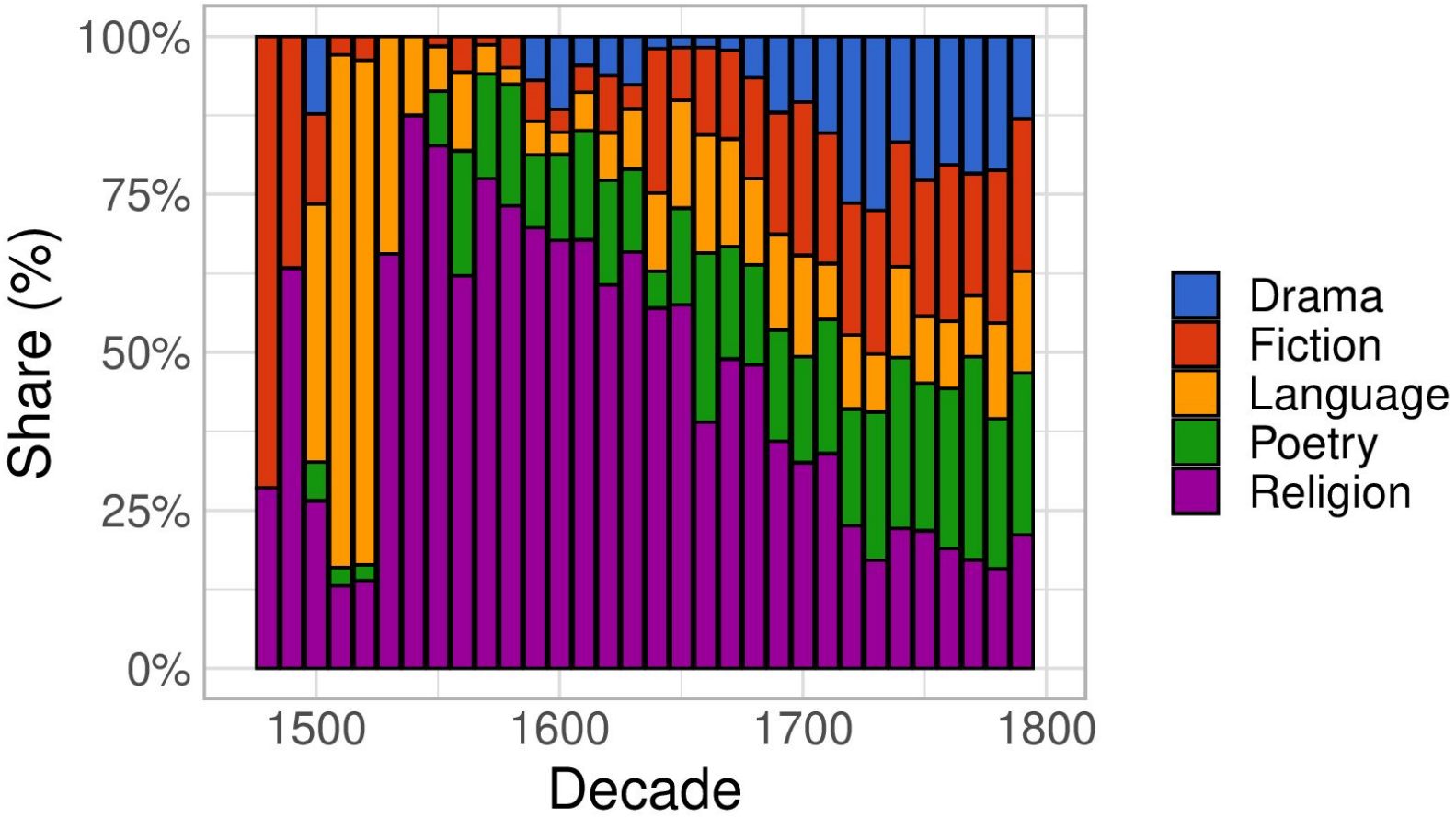
- UIs, APIs, shiny apps etc.

What have we done with respect to ESTC harmonization?

- Created an ecosystem for harmonizing ESTC -> changes traceable and reproducible
- Harmonized most relevant MARC fields
- Extracted printer, publisher and bookseller information from imprint information
- Unique IDs for all actors
- Linked ESTC to VIAF and BBTI for actors
- Algorithmically linked edition/workfield information across ESTC data
- Enriched ESTC, for example, with gender names drawn from birth record sources
- Ready to use ESTC and ECCO to enrich both (same applies for EEBO)
- Ongoing evaluations of ESTC as a record (f.e. comparing to STCN)

Data-driven approach to constructing and examining the English canon (ca. 1500-1800)

- Quantitatively constructed canon of works that were **a) published most often**, **b) most frequently** and **c) for the longest period** of time in Britain and North-America
- Making use of a processed version of the ESTC
- Keys to the analysis: **1) edition field information** and **2) information extracted from imprints about publishers and printers**
- analyzing the canon in terms of time, people, places, and materiality.
    - Main interest: epistemological shifts during early modern era.



The most popular subject-topics for the ten most printed works in each decade from 1500 to 1800.

Virtuous cycle of better data

- Combining **harmonized metadata to full-text sources (ESTC & ECCO)** -> Enables text mining in a new way, upcoming this academic year.
- Using **full-texts to enrich metadata (ESTC & ECCO)** -> Feeding back to the loop, better quality data, detecting subject/topics for example.
- Combining **text reuse information to metadata (ESTC & ECCO)** -> feeds back to edition information.
- **Re-OCRing (ECCO)** -> Feeds back to all processes that combine ECCO and ESTC.

Comhis
Helsinki Computational History Group

Humanities collaboration for better data

- Crowdsourcing experts
- Collaboration with different field of science, national libraries, infrastructures and projects
- Collaboration with companies that do digitization
- Interoperability & dealing with noise and bias

→ We need right kind of infrastructures for specific purposes that enable collaboration between researchers, companies and libraries.

NEWS EYE

A Digital Investigator for Historical Newspapers

NewsEye is a research project advancing the state of the art and introducing new concepts, methods and tools for digital humanities by providing enhanced access to historical newspapers for a wide range of users. With the tools and methods created by NewsEye, crucial user groups will be able to investigate views and perspectives on historical events and development and, as a consequence, the project will change the way European digital heritage data is (re)searched, accessed, used and analysed.

Workflow

The core concept of NewsEye is a seamlessly integrated armoury of tools and methods that will improve the users' capability to access, analyse and use the content in the digital libraries of historical newspapers.

Four Case Studies: the inner test material used by the project's humanities research groups will be from European newspaper datasets from the three partner libraries focusssing on the period 1850-1950.

4 NewsEye case study topics

1. Media & Journalism
2. Gender
3. Migration

Computer Science

Digital Humanities

La Rochelle Université

Universität Rostock

University of Helsinki

universität innsbruck

universität wien

Université Paul Valéry Montpellier 3

National Libraries

Österreichische Nationalbibliothek 600K

16 40 The National Library of Finland 500K

BnF Bibliothèque nationale de France 250K

Total number of pages used for the project (1.35M)

NewsEye consortium

Text Recognition & Article Separation

> Article Separation
> Automatic Text Recognition
> Layout Analysis

Semantic Text Enrichment

> Named entity recognition
> Stance detection
> Novelty detection
> Event detection

Science and hermeneutics

Tangible objects ≠ Subjective experience

Need for mixed methods!

photo: Time Machine project