KU LEUVEN

A!

Aalto University

**HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI**

The first steps of PreCEM
Tracking the Presence of Classics in Early Modern England

Margherita Fantoli, Mark Depauw, Jukka Suomela, Toon Van Hal, Mikko Tolonen
KU Leuven, Helsinki University, Aalto University

How did classical texts become cultural constituents?

Introduction
The impact of Classics on Early Modern cultural production is well known, but little has been done to provide quantitative information related to the topic. With PreCEM (Presence of Classics in Early Modern Book History) we aim at exploiting the largest collections of printed (meta)data for British books (1470–1800) available in machine-readable form: the English Short Title Catalogue (ESTC), Eighteenth Century Collections Online (ECCO) and Early English Books Online (EEBO-tcp phase II).

Workflow

Identification of classical authors in the ESTC

Metadata analysis

Text reuse analysis
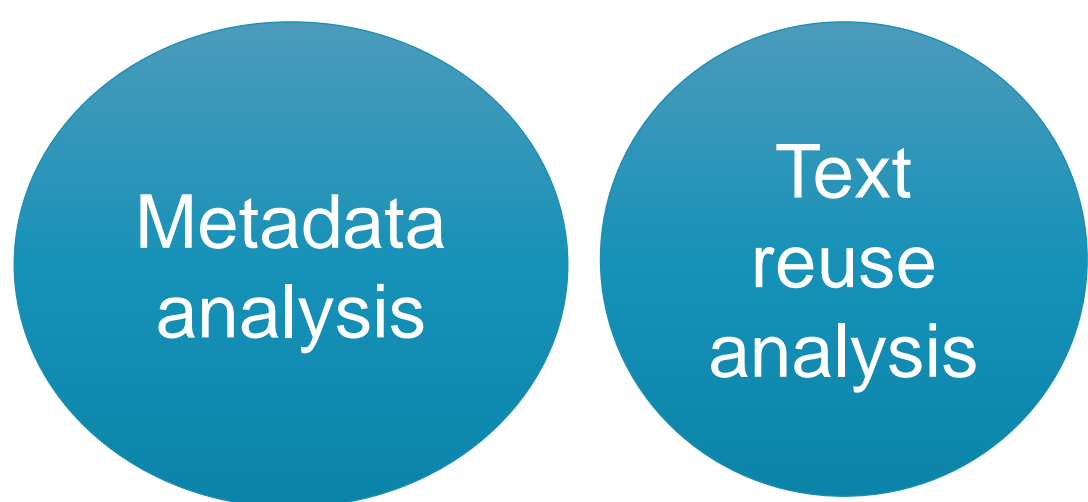
Authors
- **Criterium:** authors writing in Latin and Ancient Greek who lived up to the end of 5th century CE.
- **Method 1:** filtering on dates in the DB ESTC authors
- **Method 2:** matching VIAF identifiers from Trismegistos database (800 BCE-800 CE)
- Total: **312** ESTC classical authors for **6307 classical works** (6260 when removing the 1800 decade)

Metadata Analysis: trends, authors and language

ESTC number of classics publications
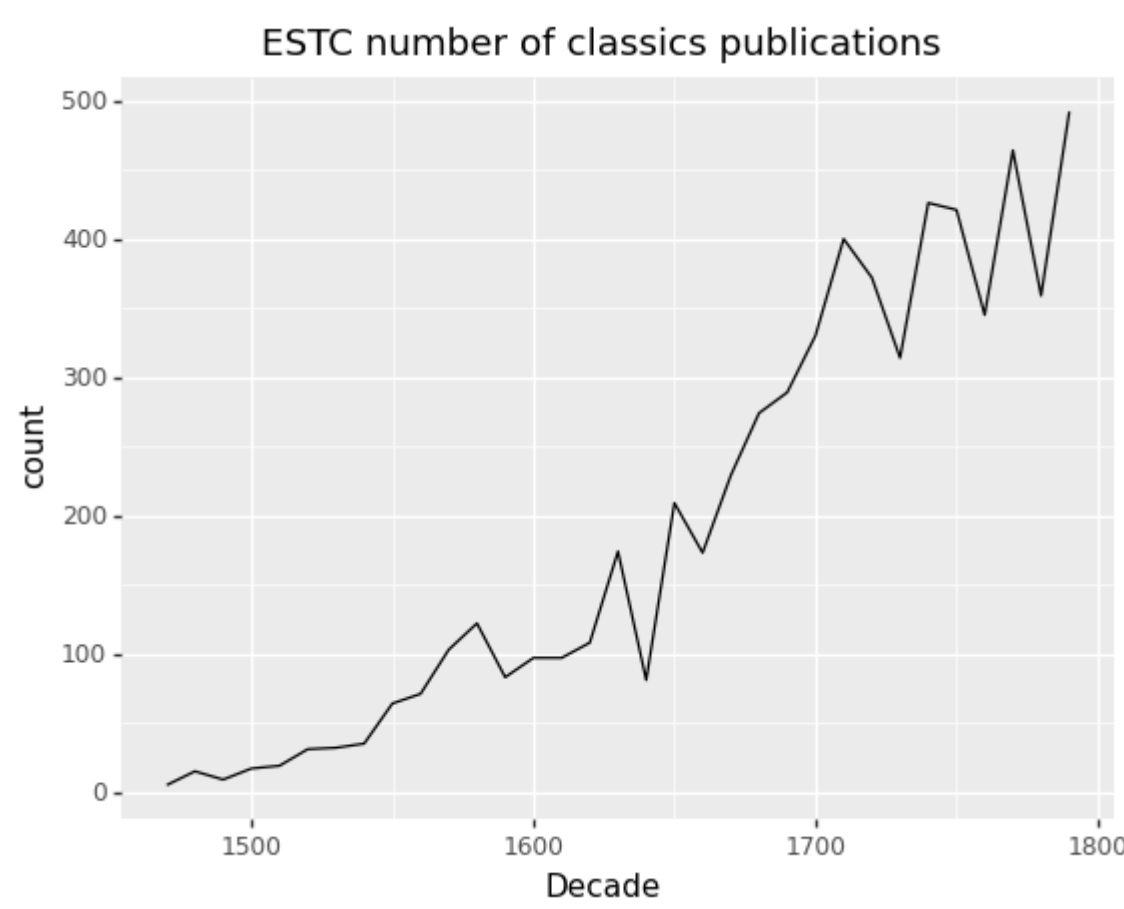
Despite a general growth of the number of classical texts printed, their impact on the total of publications decreases.

ESTC numer of publications

ESTC rate of classics
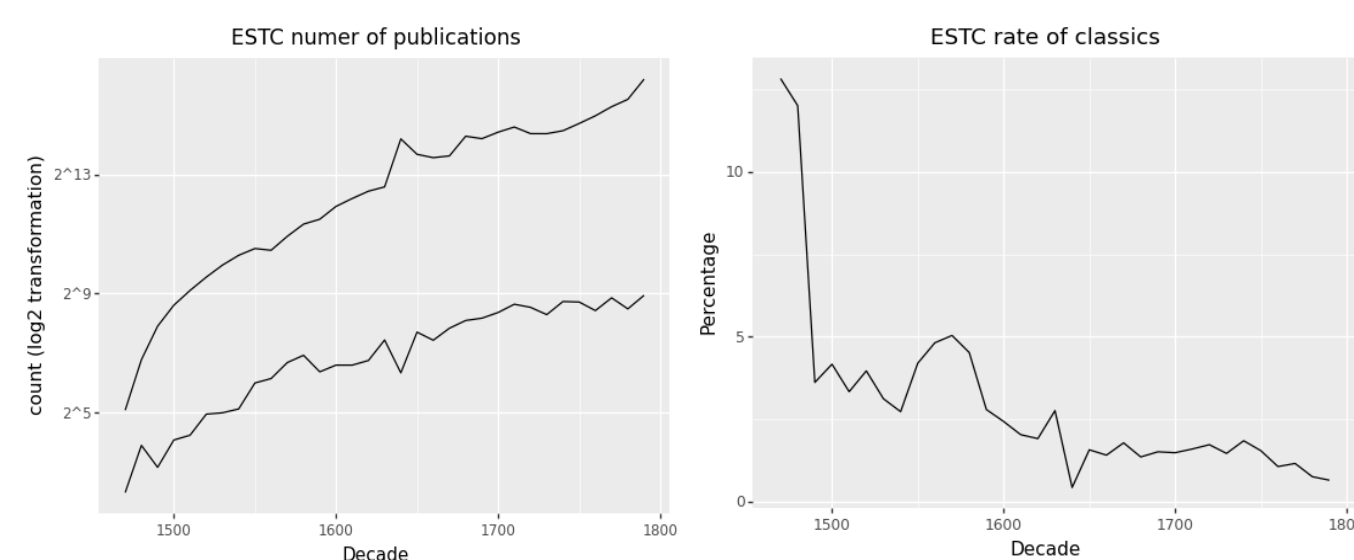
The number of different authors printed drops from the 17th (258) to the 18th century: (197), despite an increase in the number of books (from 2337 to 3923). In general, the number of different authors published per year stabilizes and decreases.
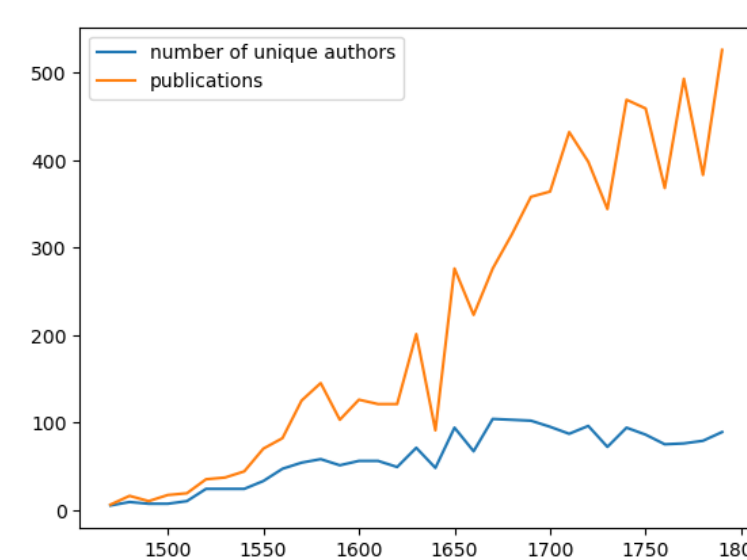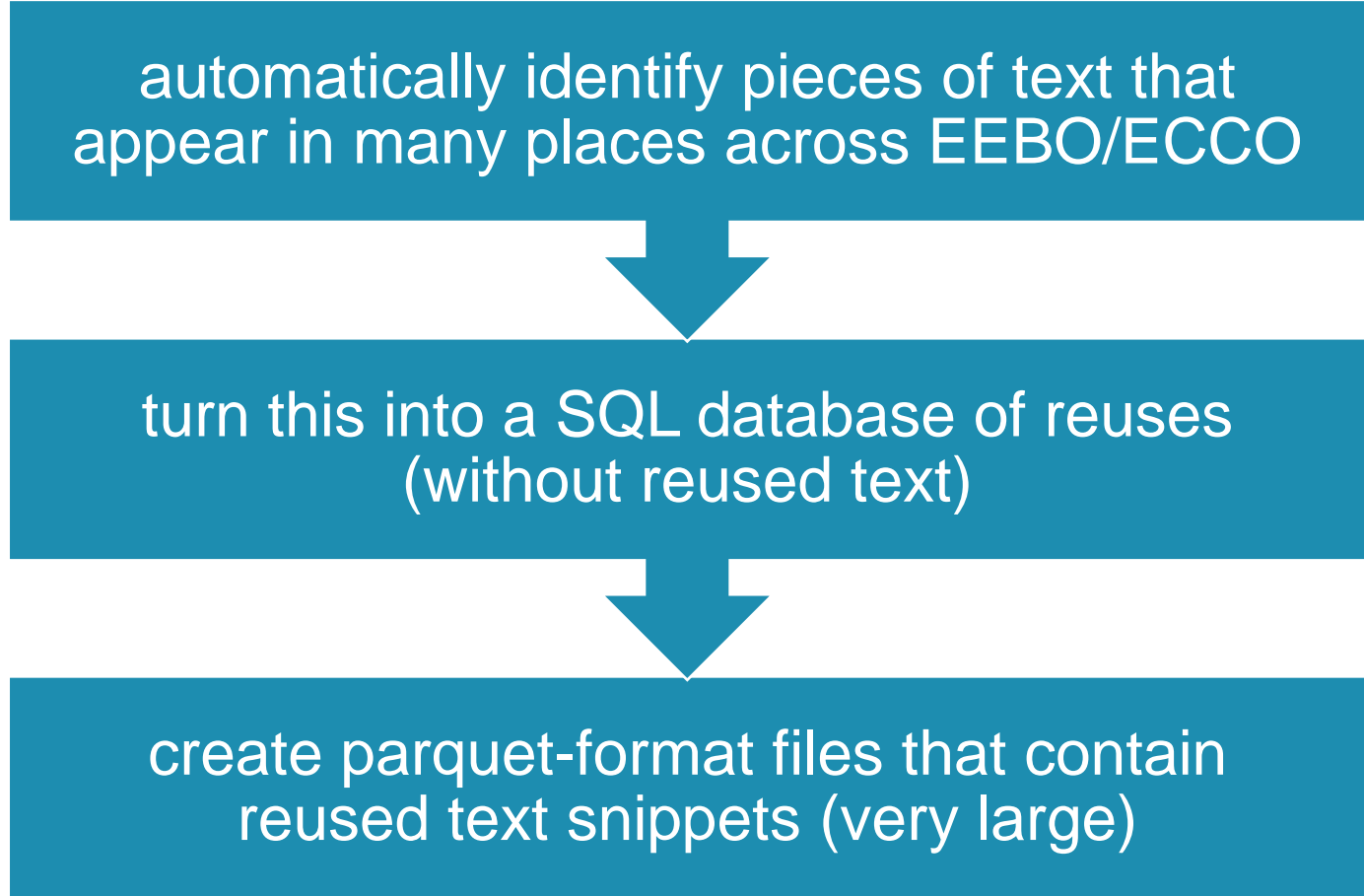

Most published authors	
Ovid, 447	Homer (241)
Cic., 433	Arist. (227)
Hor., 365	Ter. (180)
Virgil, 332	Plut. (159)
Aes., 319	Xen. (131)

The languages of publication are mainly English and Latin, whereas Ancient Greek is clearly underrepresented.

languages_Publication_of_classics_estc
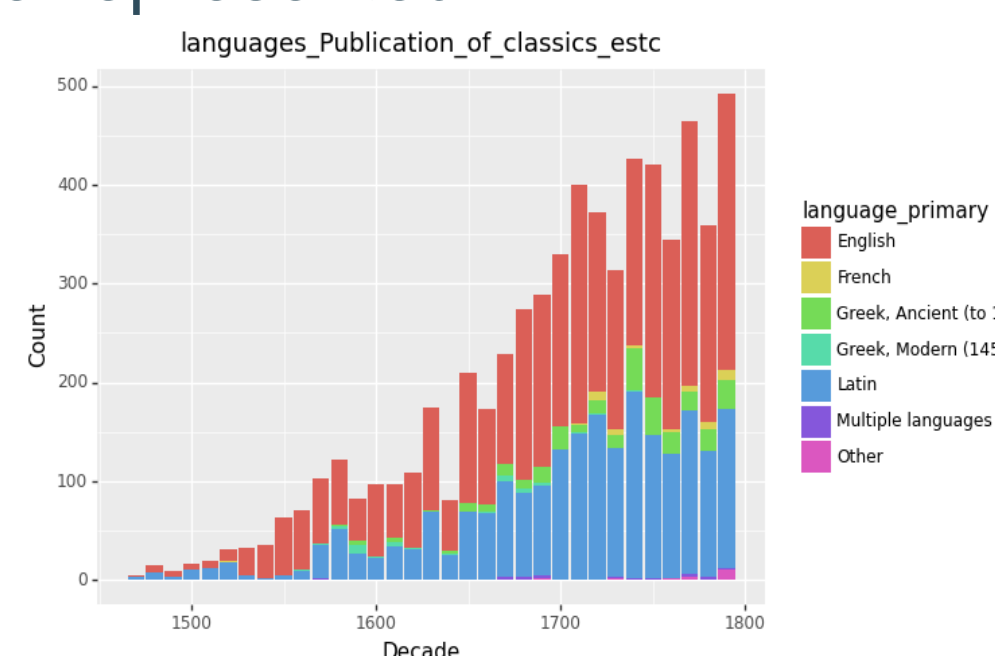
Capturing text-reuse

EEBO and ECCO collections provide machine-readable versions of (part of) the texts belonging to the catalogue. The Helsinki and Aalto team has developed a workflow to:

automatically identify pieces of text that appear in many places across EEBO/ECCO

turn this into a SQL database of reuses (without reused text)

create parquet-format files that contain reused text snippets (very large)

And in particular:
- extract all reuses where one of the works is in our list of classical texts
- perform some automatic cleanup + summarization
- present in a form suitable for human consumption & annotation
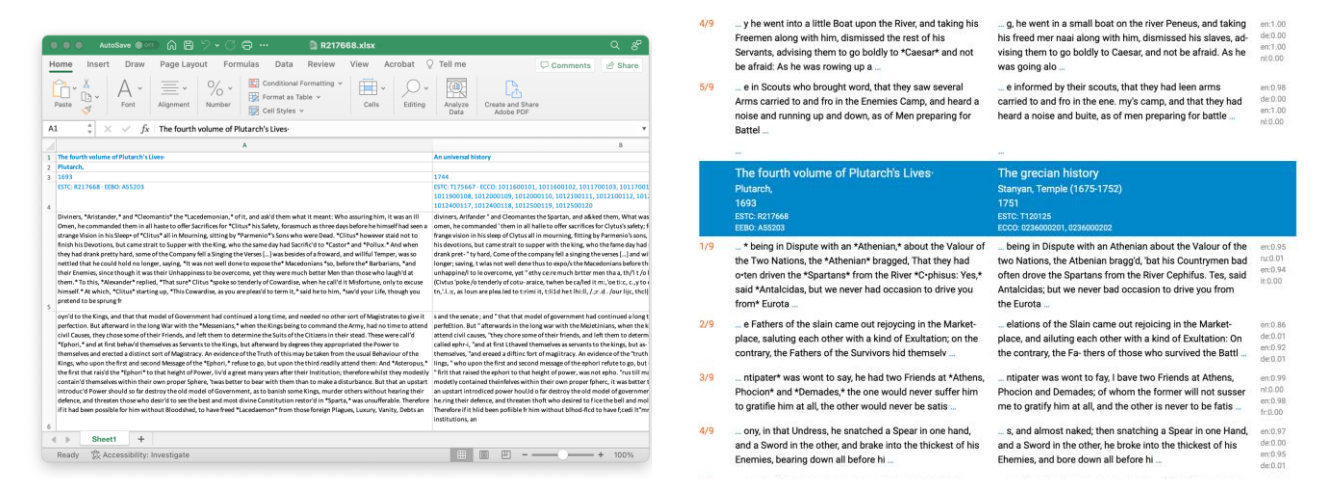
Main challenges
- EEBO: XML tags create noise & passages in Greek are not transcribed
- ECCO: low-quality OCR transcription
- Sorting out reuse of paratext (mention of publisher etc.) and actual text-reuse (e.g. 17000 snippets already removed, out of 280000)

WIP and future work
- Language identification of the snippets: identifying Latin quotes & their translations in English context
- Fine-grained analysis of specific authors (Horace)