Thoughts from the North: Humanities-Driven AI and LLMs in SSH

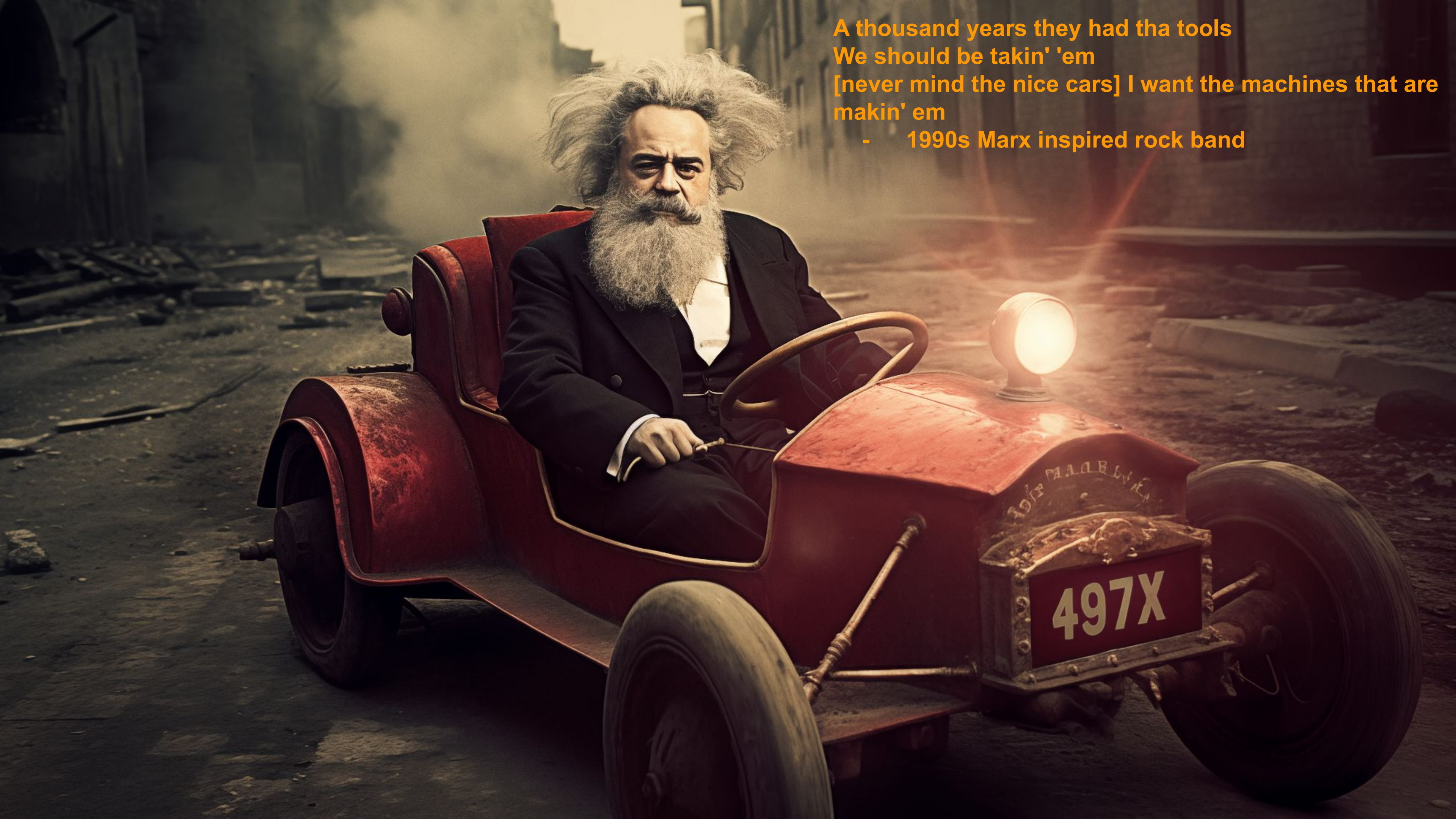
Mikko Tolonen, University of Helsinki

CLARIAH-EUS: A Vibrant Community, a Practical Infrastructure
21.11.2024

Comhis
Helsinki Computational History Group

What is Humanities Driven AI?

1) Debate about ***Digital* in *Humanities*** is rather futile. It won’t go away.
2) Ambition level should be high. We are lagging behind in humanities in the use of machine learning ~ AI in traditional research fields.
3) Interdisciplinary collaboration & focus on the subject field in SSH is the key.
4) Everyone does not have to turn *computational*. But we need to collaborate in new ways on new (and old) research questions.
5) More purpose built focus. If we need low resource digitization pipeline, let’s build it …



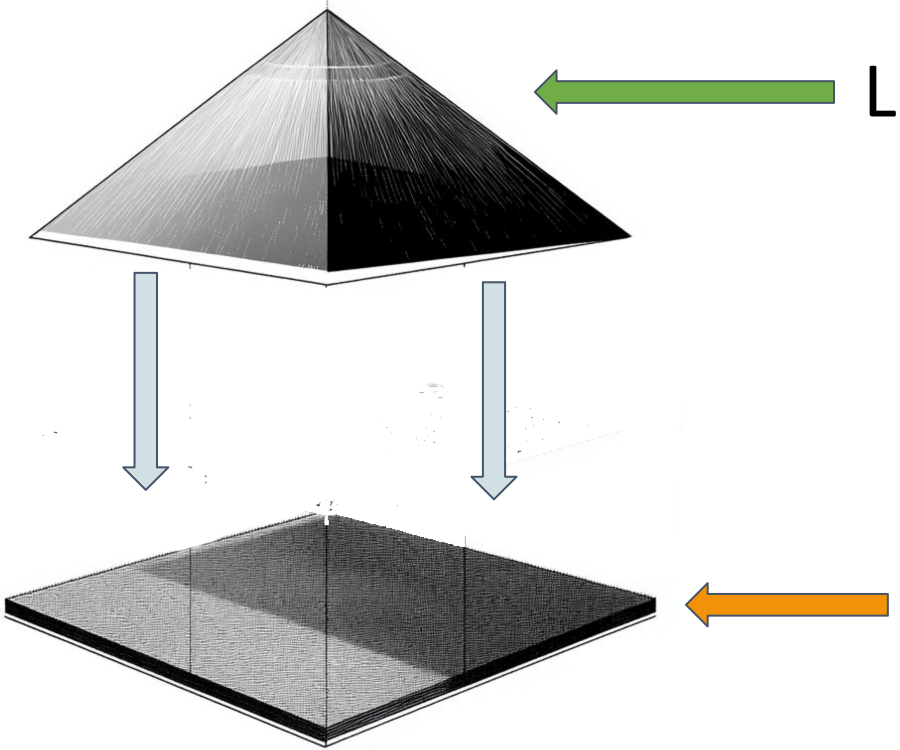
A thousand years they had tha tools
We should be takin' 'em
[never mind the nice cars] I want the machines that are makin' em

- 1990s Marx inspired rock band

Backbone of humanities driven AI

1) Leading (not only participating in) sustained collaboration with respect to use of machine learning (~ AI).
2) From borrowing methods to method development
3) Digitization of core research based on its own premises
4) With new possibilities of studying primary sources getting back to core theoretical questions -> i.e. what is an “idea” in history of philosophy?
5) We need to be clever so that SSH will be the winner of AI boom.

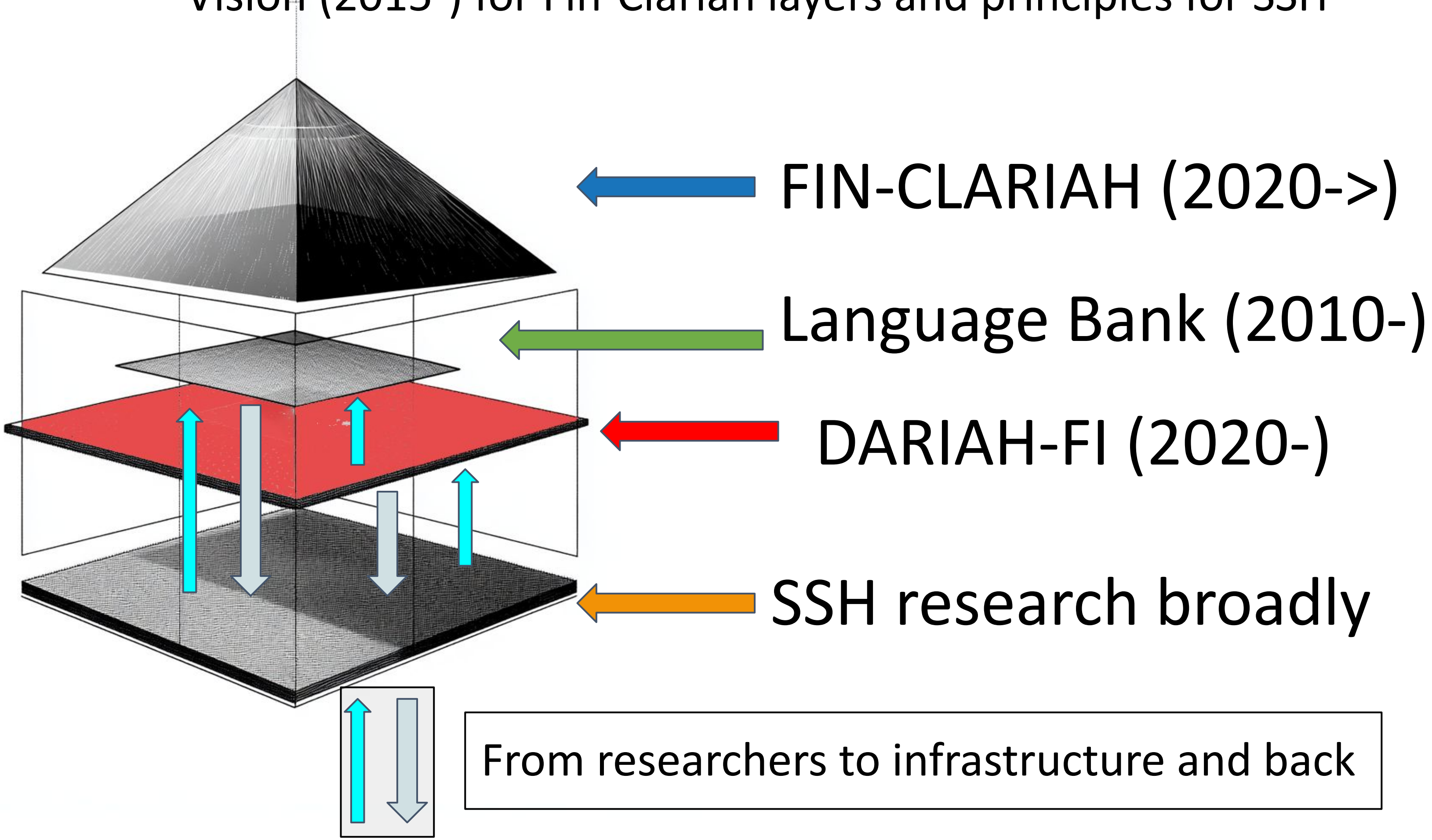
State of humanities infrastructure in 2010 in Finland



Language Bank (2010-)

Mainly linguists

Vision (2015-) for Fin-Clariah layers and principles for SSH



FIN-CLARIAH (2020->)

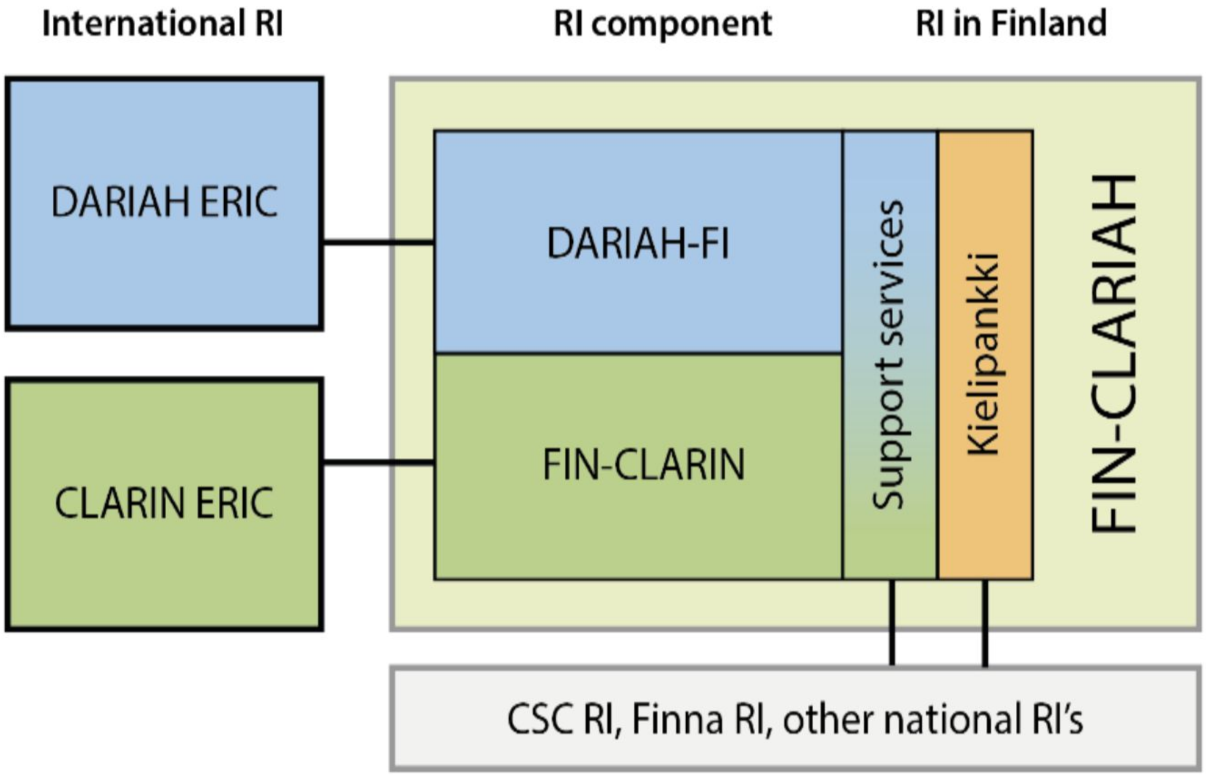
Language Bank (2010-)

DARIAH-FI (2020-)

SSH research broadly

From researchers to infrastructure and back

FIN-CLARIAH - DIGITIZING THE SSH DOMAIN



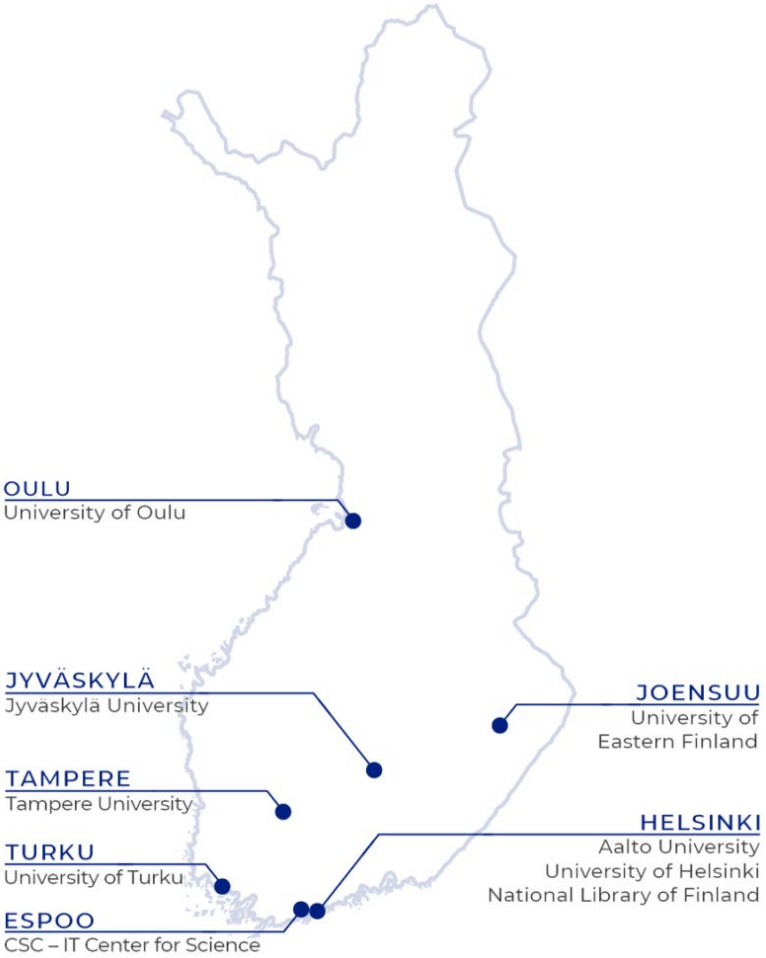
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

DARIAH-FI AS A NATIONAL NETWORK

Ambition: We want to represent all areas of SSH the best we can but understanding that we all have a past and there is a limit to funding.

What's next for the future? Large Language Models, visual cultures, ethics, new partnerships, and enhanced collaboration. In other words, heavy emphasis on AI.

What do we aim to enable? It's all about community; we are providing the tools that power AI, but it's up to different segments of the community to step up, deliver, and take charge of the overall development of humanities-driven AI in their respective fields.



Swedish choice for infrastructure is very different

- https://www.huminfra.se/
- No aim to build capacities at infrastructure level or deal centrally with data at all
- Whole Sweden divided into nodes that work autonomously but reporting to central organisation
- HUM-INFRA gathers common information from nodes and disseminates it and does some training on top. Main point for them is the website.
- Point being: no “one solution” will fix things!

If we aim at everything, we hit nothing.

Example of research group working together
with CLARIAH infrastructure

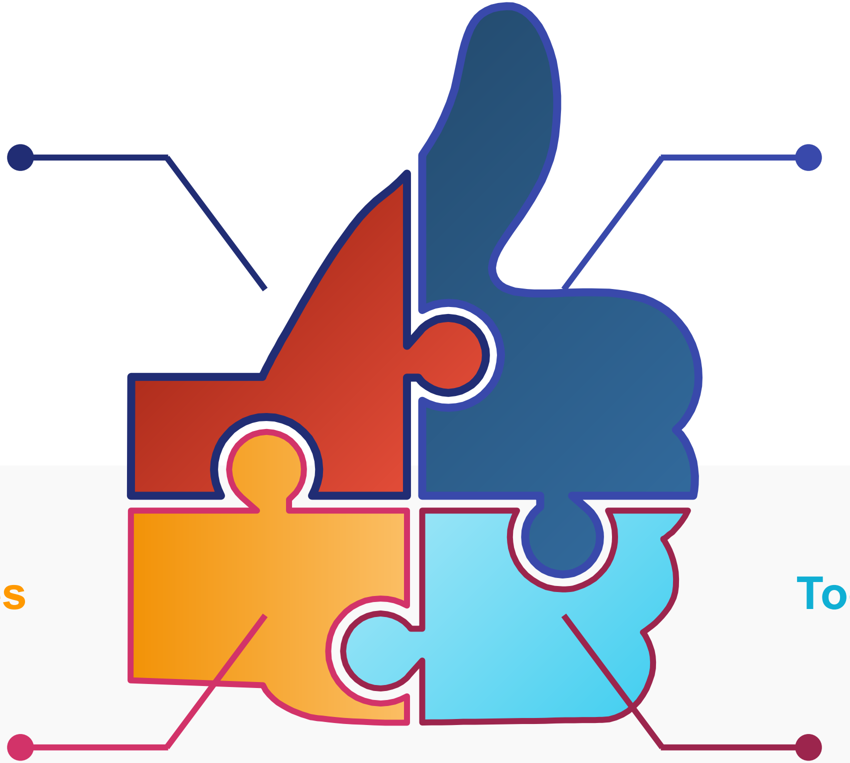
Helsinki Computational History Group

Computer scientists researching open workflows, algorithms and interfaces for humanities text and metadata

Linguists exploring the relationship between words and concepts

Historians interested in conceptual and actual historical processes

COMHIS aim: Understanding public communication covering the 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)

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)

Research data releases

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

Tools for others

- UIs, APIs, shiny apps etc.

Comhis
Helsinki Computational History Group

COMHIS: ongoing funded projects using HPC

Comhis
Helsinki Computational History Group

RiCEP: Rise of Commercial Society and Eighteenth-Century Publishing

- Academy of Finland funded project 2020-2024

HPC-HD: High Performance Computing for Historical Discourse Detection

- Academy of Finland funded consortium with 3 computer science groups 2022-2025

PreCEM: Presence of Classics in Early-Modern Book History

- Marie Curie Doctoral Training Network 2024-2028

From research to infrastructure and back

- We started work on text reuse in 2015 (examples will follow)
- We have been adding data and developed the ecosystem since then (e.g. newspaper process still ongoing)
- We were able to stabilise it in 2023 to the level that we thought public tool makes sense (part of DARIAH-FI) -- if it is cutting-edge LLMs does not matter because it serves a purpose!
- Because it does it’s job, there’s been spinoffs: Latin Reader; Edition Reader (and could be others too)
  - We consider these as part of other projects or infrastructure building i.e. funding from those who need it as part of their project or infrastructure
  - Maintenance part is part of infrastructure (ideally)
- Same goes for many others, e.g. Sampo-portals in Finland are a difficult question who owns them after the research group finishes. Infrastructure issue?

Meaning of meaning in intellectual history and LLMs

Jean-Jacques Rousseau wrote: “**Man is born free, and everywhere he is in chains.**" (1762).

This has been *reinterpreted, used and misused* in different contexts:

1) The cycle of vices into which the commercial society leads us.
2) In French revolution a suggestion to take arms against the monarchy.
3) A claim against slavery.
4) Eighteenth-century pro-slavery people managed to use this to claim that slavery and “living in chains” is natural for some people.
5) During the 250 years many have also criticised Rousseau that he did not apply this principle to women.

Here’s the point for SSH in general. It is obvious that Rousseau is relevant through different discussions even today, but *how has he influenced through* different times and contexts is the way to answer questions that humanities people are asking. So, direct answer to question “What did Rousseau mean when he said…” as such is not the way we ask questions in the humanities.

That we can hope that we can algorithmically start separating engagement in the way that presented above is only a possibility now across languages and across 500,000 historical books.

Text reuse, large language models and semantic similarity

Text reuse and large digital archives

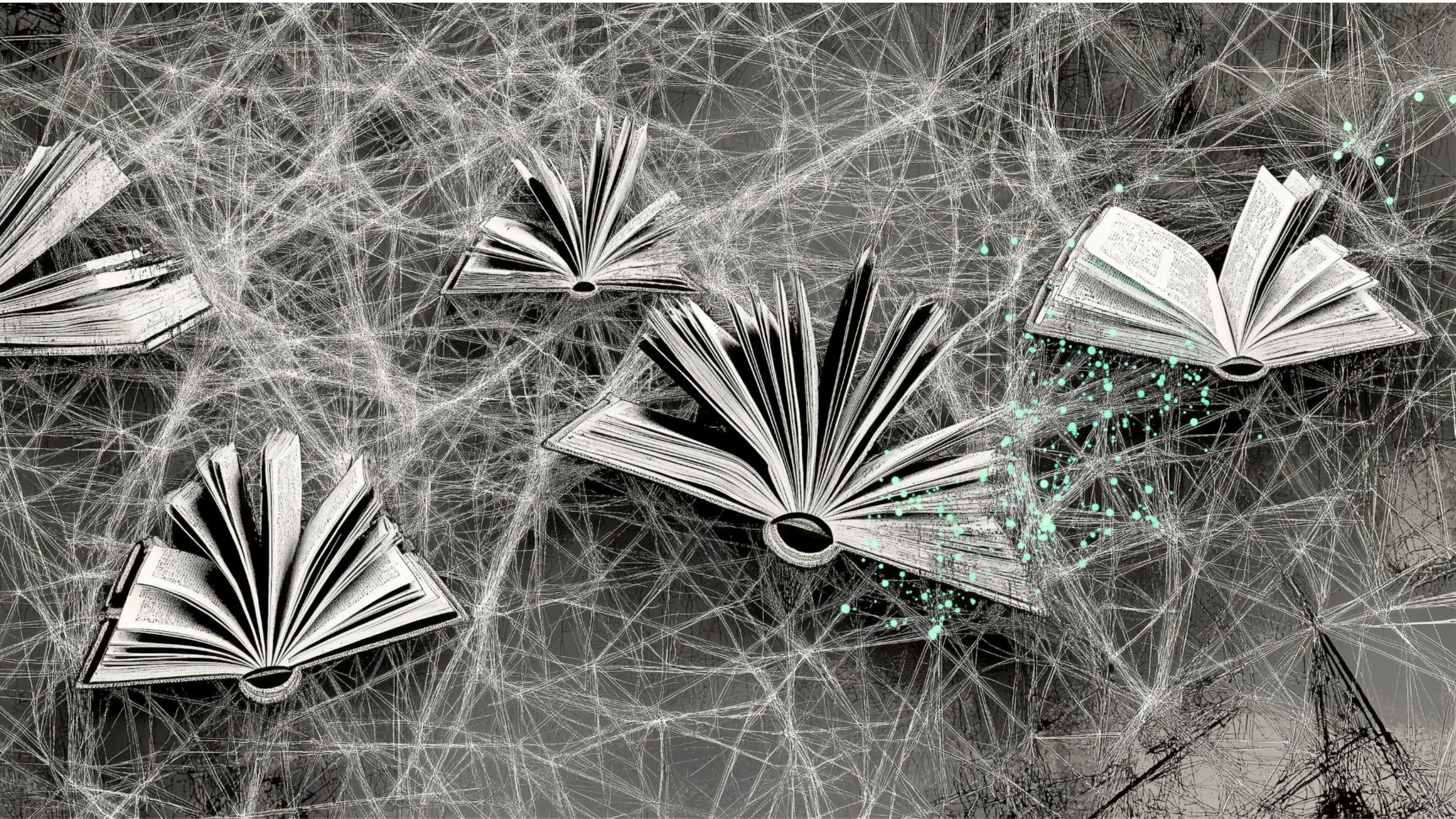
In 2018, we turned earlier Eighteenth-Century Collections Online (ECCO) and its 225,000 titles into 0.5B text reuse pairs… we consider this as the beginning.

With ECCO text reuse data we can build countless historical cases. Enough for lifetimes of historians (at least in the old school sense of it).

We can study 1) **reception** 2) **patterns of borrowing** 3) **impact** 4) **discourses** through dissemination of texts

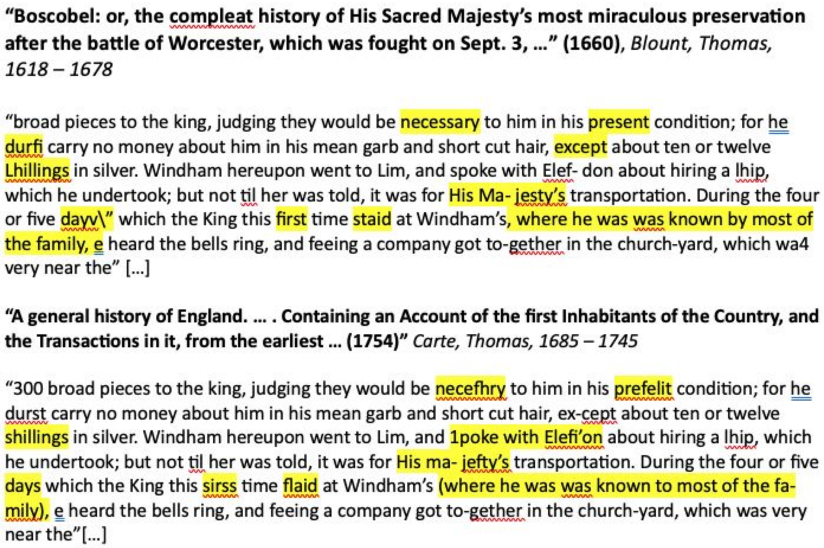
But we **ought to aim to go beyond single digital archives**

The **step to be taken is from simple textual overlap to mapping of ideas**. What has to be understood though is that these go hand in hand.



Textual overlaps and semantic similarity

Example of lexical text reuse



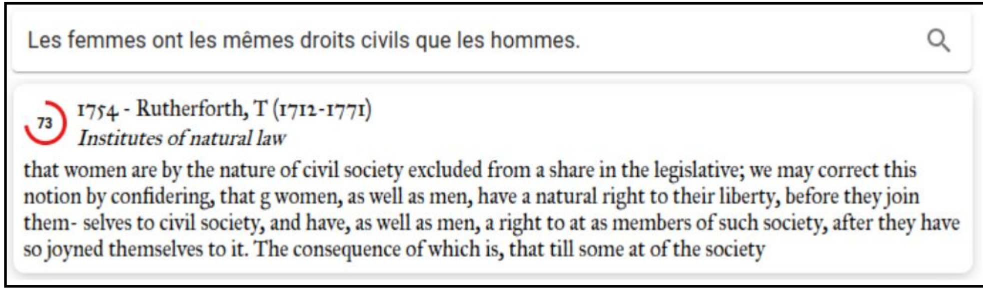
“Boscobel: or, the compleat history of His Sacred Majesty’s most miraculous preservation after the battle of Worcester, which was fought on Sept. 3, …” (1660), *Blount, Thomas, 1618 – 1678*

“broad pieces to the king, judging they would be necessary to him in his present condition; for he durfi carry no money about him in his mean garb and short cut hair, except about ten or twelve Lhillings in silver. Windham hereupon went to Lim, and spoke with Elef- don about hiring a lhip, which he undertook; but not til her was told, it was for His Ma- jesty’s transportation. During the four or five dayv\” which the King this first time staid at Windham’s, where he was was known by most of the family, e heard the bells ring, and feeing a company got to-gether in the church-yard, which wa4 very near the” […]

“A general history of England. … . Containing an Account of the first Inhabitants of the Country, and the Transactions in it, from the earliest … (1754)” *Carte, Thomas, 1685 – 1745*

“300 broad pieces to the king, judging they would be necefhry to him in his prefelit condition; for he durst carry no money about him in his mean garb and short cut hair, ex-cept about ten or twelve shillings in silver. Windham hereupon went to Lim, and 1poke with Elefi’on about hiring a lhip, which he undertook; but not til her was told, it was for His ma- jefty’s transportation. During the four or five days which the King this sirss time flaid at Windham’s (where he was was known to most of the fa-mily), e heard the bells ring, and feeing a company got to-gether in the church-yard, which was very near the”[…]

Example of semantic similarity



Les femmes ont les mêmes droits civils que les hommes.

73

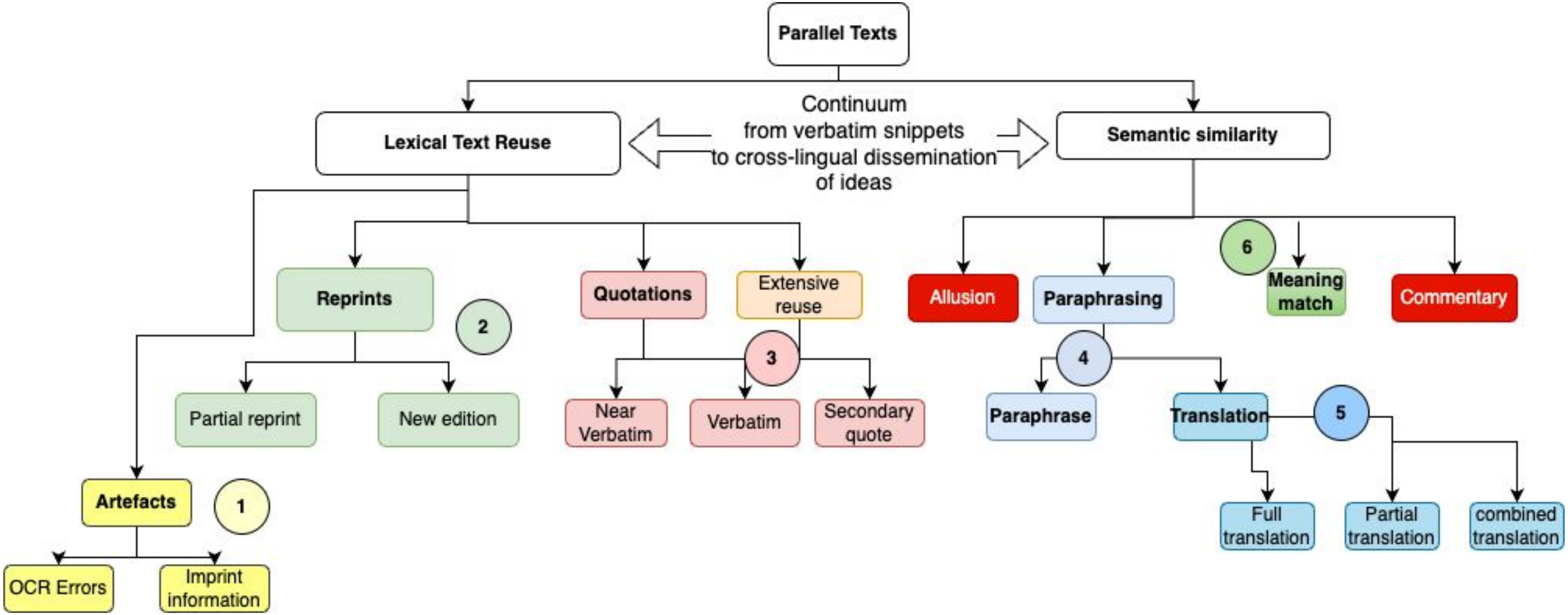
1754 - Rutherforth, T (1712-1771)

Institutes of natural law

that women are by the nature of civil society excluded from a share in the legislative; we may correct this notion by confidering, that g women, as well as men, have a natural right to their liberty, before they join them- selves to civil society, and have, as well as men, a right to at as members of such society, after they have so joyned themselves to it. The consequence of which is, that till some at of the society

Comhis
Helsinki Computational History Group

Conceptualising text reuse and semantic similarity



Towards a taxonomy of text reuse types

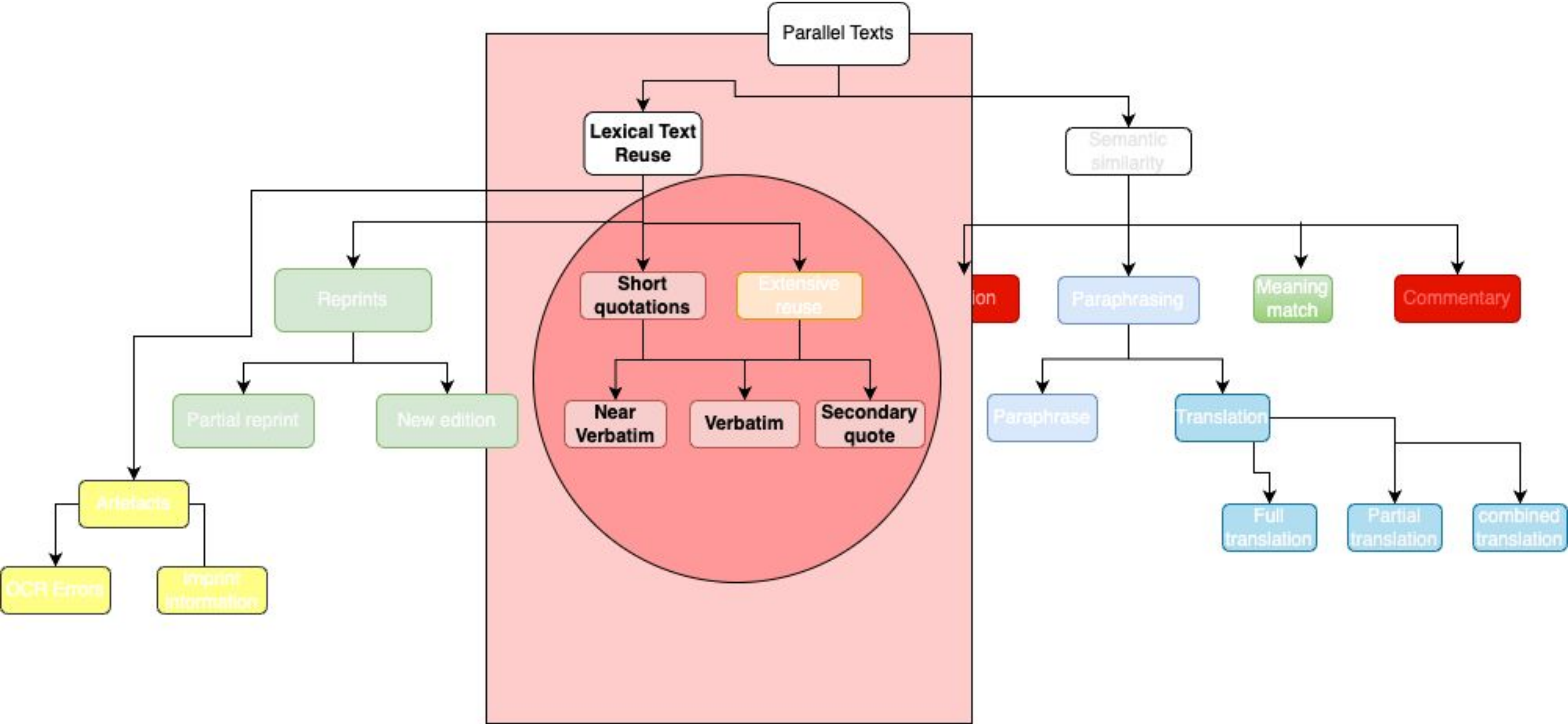
Lexical text reuse

1. Artefacts
   a. OCR errors
   b. Imprint information
2. Reprints
   a. Partial reprints
   b. New editions
3. Quotations
   a. Near verbatim
   b. Verbatim
   c. Secondary quote

Semantic similarity

4. Paraphrasing
5. Translations
   a. Full translation
   b. Partial translation
   c. Combined translation
6. Meaning matches
   a. Same language
   b. Different language

Quotations (lexical text reuse)



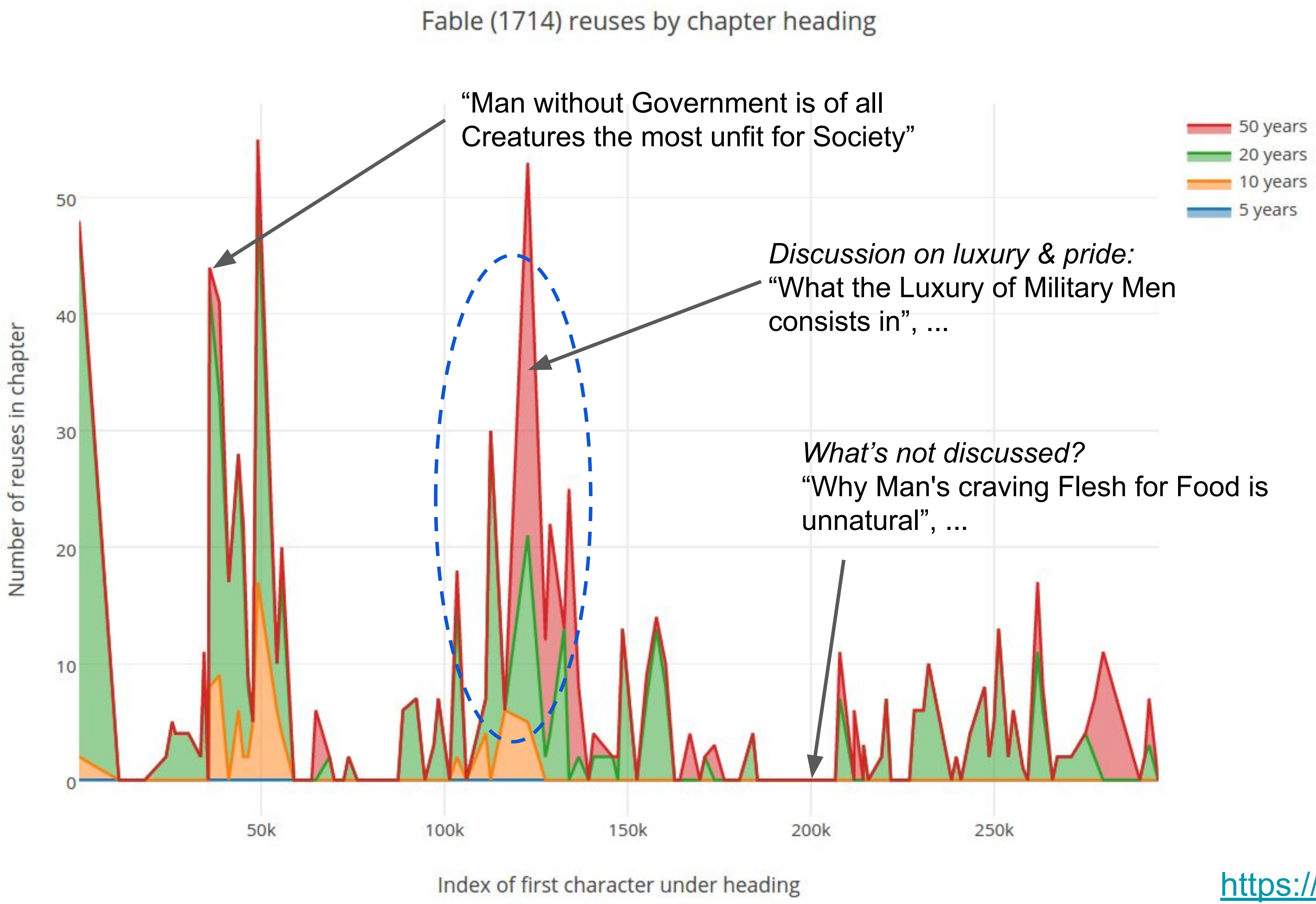
Quotes and intellectual history

What kind of patterns of borrowing Bernard Mandeville (1670-1733) had?

What kind of habits of reusing his own works did Mandeville have?

Who quoted Mandeville, when and why?

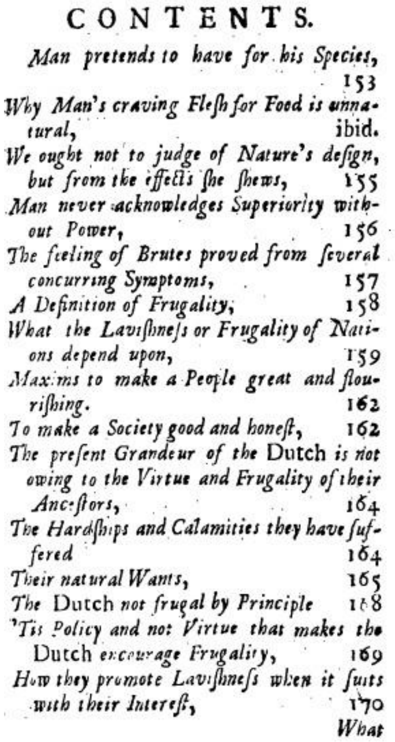
With computational methods it is possible to start thinking about the composition and the influence of any early modern work in a new way.



“Man without Government is of all Creatures the most unfit for Society”

Discussion on luxury & pride:
“What the Luxury of Military Men consists in”, ...

What’s not discussed?
“Why Man's craving Flesh for Food is unnatural”, ...

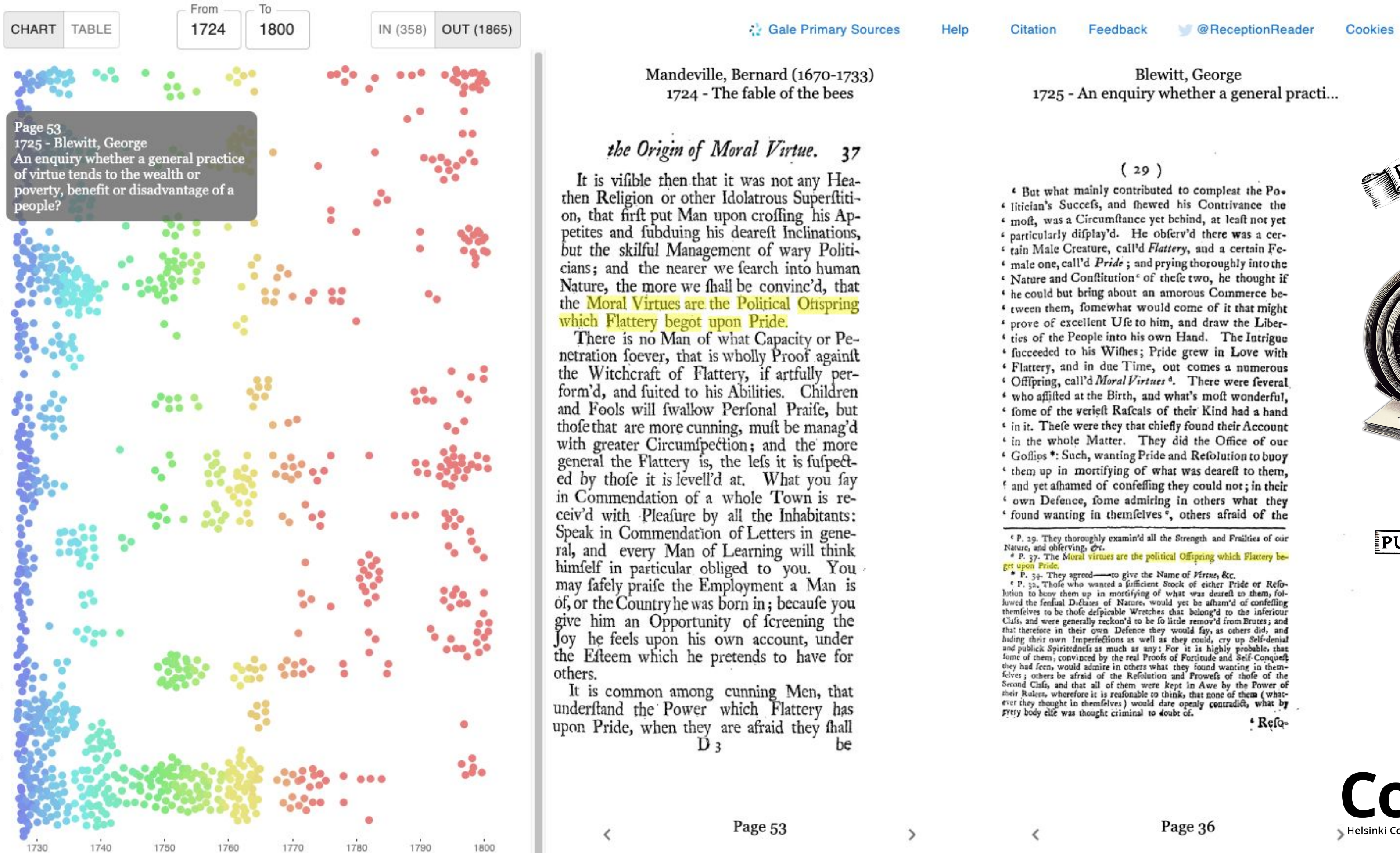


CONTENTS.

What

https://plot.ly/~villepvaara/7/

Mandeville's *Fable* in ReceptionReader.com interface that we made in 2023



Page 53
1725 - Blewitt, George
An enquiry whether a general practice of virtue tends to the wealth or poverty, benefit or disadvantage of a people?

Mandeville, Bernard (1670-1733)
1724 - The fable of the bees

the Origin of Moral Virtue. 37

It is viſible then that it was not any Heathen Religion or other Idolatrous Superſtition, that firſt put Man upon croſſing his Appetites and ſubduing his deareſt Inclinations, but the ſkilful Management of wary Politicians; and the nearer we ſearch into human Nature, the more we ſhall be convinc'd, that the Moral Virtues are the Political Offspring which Flattery begot upon Pride.

There is no Man of what Capacity or Penetration ſoever, that is wholly Proof againſt the Witchcraft of Flattery, if artfully perform'd, and ſuited to his Abilities. Children and Fools will ſwallow Perſonal Praiſe, but thoſe that are more cunning, muſt be manag'd with greater Circumſpection; and the more general the Flattery is, the leſs it is ſuſpected by thoſe it is levell'd at. What you ſay in Commendation of a whole Town is receiv'd with Pleaſure by all the Inhabitants: Speak in Commendation of Letters in general, and every Man of Learning will think himſelf in particular obliged to you. You may ſafely praiſe the Employment a Man is of, or the Country he was born in; becauſe you give him an Opportunity of ſcreening the Joy he feels upon his own account, under the Eſteem which he pretends to have for others.

It is common among cunning Men, that underſtand the Power which Flattery has upon Pride, when they are afraid they ſhall

D 3 be

Page 53

Blewitt, George
1725 - An enquiry whether a general practi...

(29)

' But what mainly contributed to compleat the Politician's Succeſs, and ſhewed his Contrivance the moſt, was a Circumſtance yet behind, at leaſt not yet particularly diſplay'd. He obſerv'd there was a certain Male Creature, call'd *Flattery*, and a certain Female one, call'd *Pride*; and prying thoroughly into the Nature and Conſtitution[c] of theſe two, he thought if he could but bring about an amorous Commerce between them, ſomewhat would come of it that might prove of excellent Uſe to him, and draw the Liberties of the People into his own Hand. The Intrigue ſucceeded to his Wiſhes; Pride grew in Love with Flattery, and in due Time, out comes a numerous Offſpring, call'd *Moral Virtues*[d]. There were ſeveral who aſſiſted at the Birth, and what's moſt wonderful, ſome of the verieſt Raſcals of their Kind had a hand in it. Theſe were they that chiefly found their Account in the whole Matter. They did the Office of our Goſſips *: Such, wanting Pride and Reſolution to buoy them up in mortifying of what was deareſt to them, and yet aſhamed of confeſſing they could not; in their own Defence, ſome admiring in others what they found wanting in themſelves[e], others afraid of the

[c] P. 29. They thoroughly examin'd all the Strength and Frailties of our Nature, and obſerving, &c.

[d] P. 37. The Moral virtues are the political Offſpring which Flattery beget upon Pride.

* P. 34. They agreed——to give the Name of *Virtue*, &c.

[e] P. 32. Thoſe who wanted a ſufficient Stock of either Pride or Reſolution to buoy them up in mortifying of what was deareſt to them, followed the ſenſual Dictates of Nature, would yet be aſham'd of confeſſing themſelves to be thoſe deſpicable Wretches that belong'd to the inferiour Claſs, and were generally reckon'd to be ſo little remov'd from Brutes; and that therefore in their own Defence they would ſay, as others did, and hiding their own Imperfections as well as they could, cry up *Self-denial* and publick *Spiritedneſs* as much as any: For it is highly probable, that ſome of them, convinced by the real Proofs of Fortitude and Self-Conqueſt they had ſeen, would admire in others what they found wanting in themſelves; others be afraid of the Reſolution and Proweſs of thoſe of the Second Claſs, and that all of them were kept in Awe by the Power of their Rulers, wherefore it is reaſonable to think, that none of them (whatever they thought in themſelves) would dare openly contradict, what by every body elſe was thought criminal to doubt of.

' Reſo-

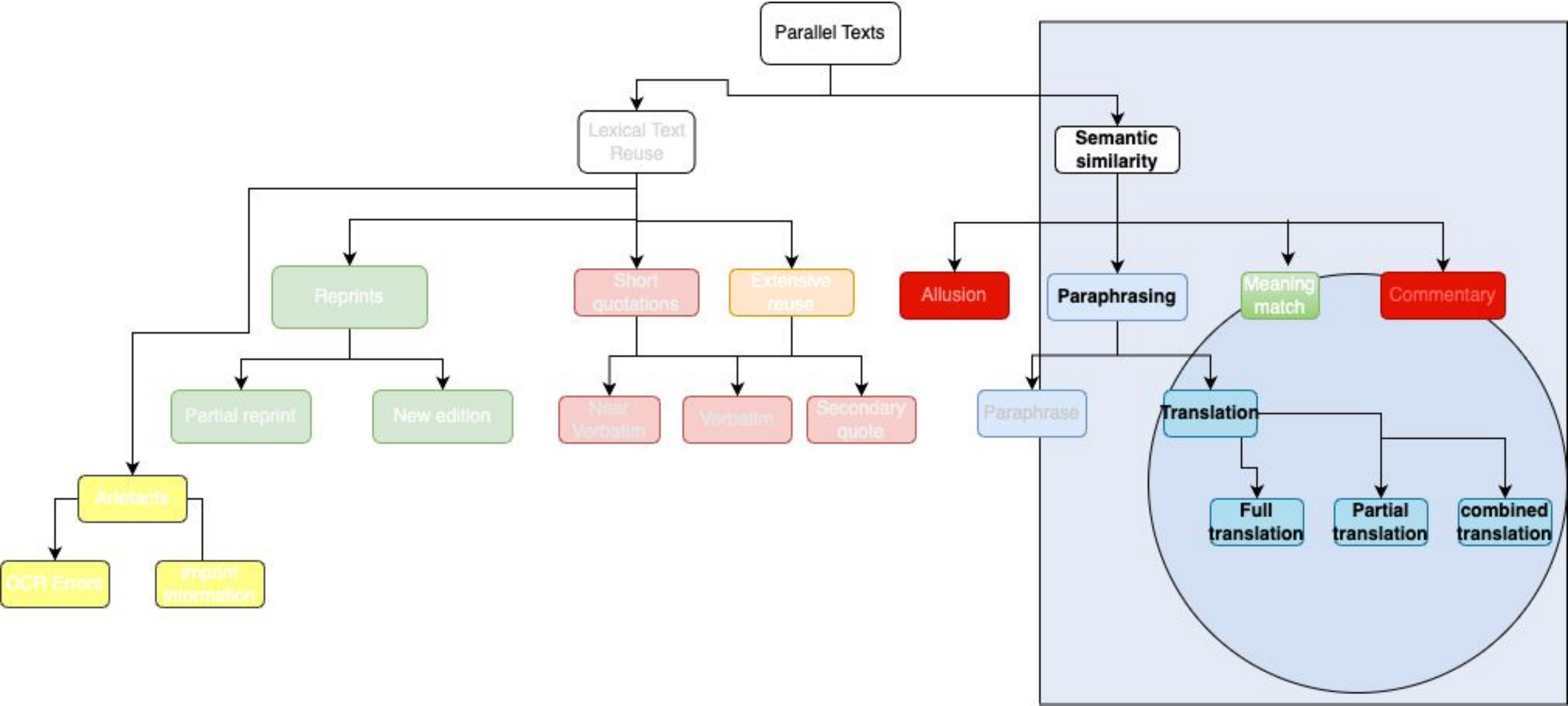
Page 36

PRIVATE VICES

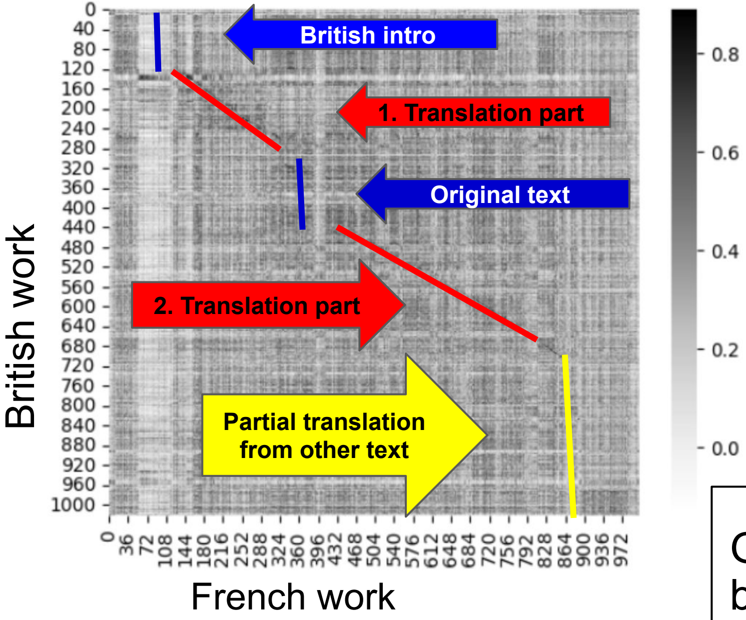
PUBLICK BENEFITS

Comhis
Helsinki Computational History Group

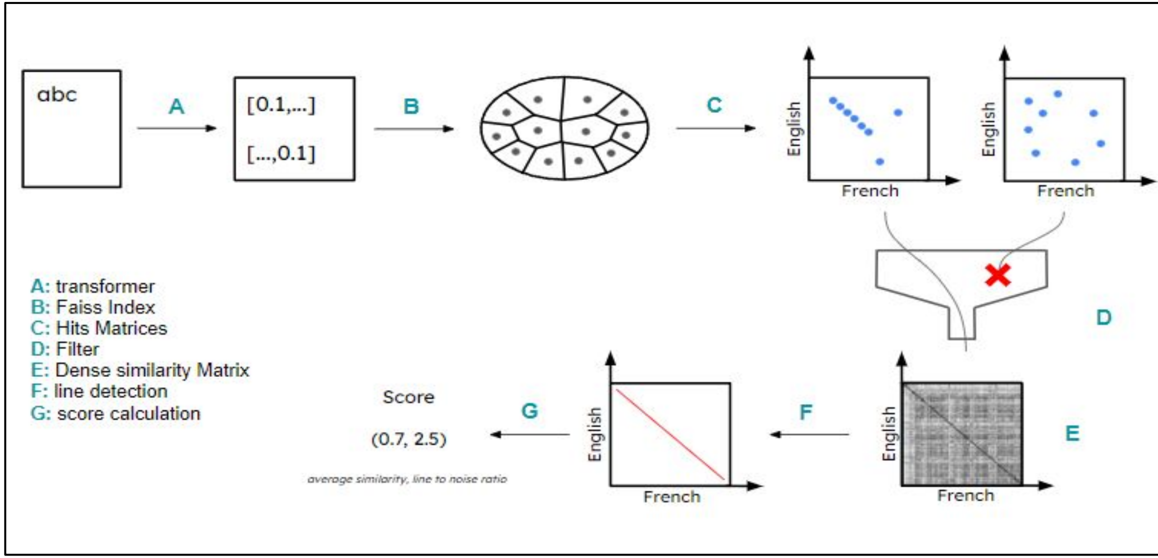
Translations (semantic similarity)



Translation mining



Previously unknown compiled translation, Stevenson, *Military instructions for officers*, 1770; **two red diagonals = translated from *La science des postes militaires*, 1759.**

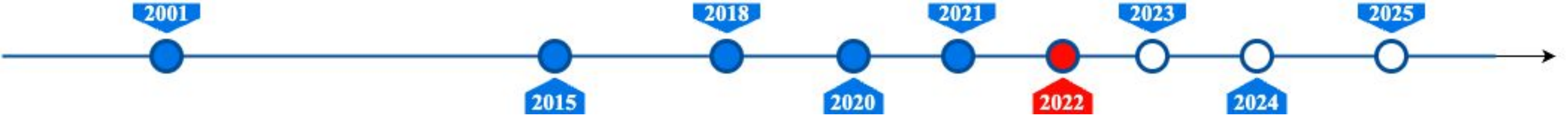


A: transformer
B: Faiss Index
C: Hits Matrices
D: Filter
E: Dense similarity Matrix
F: line detection
G: score calculation

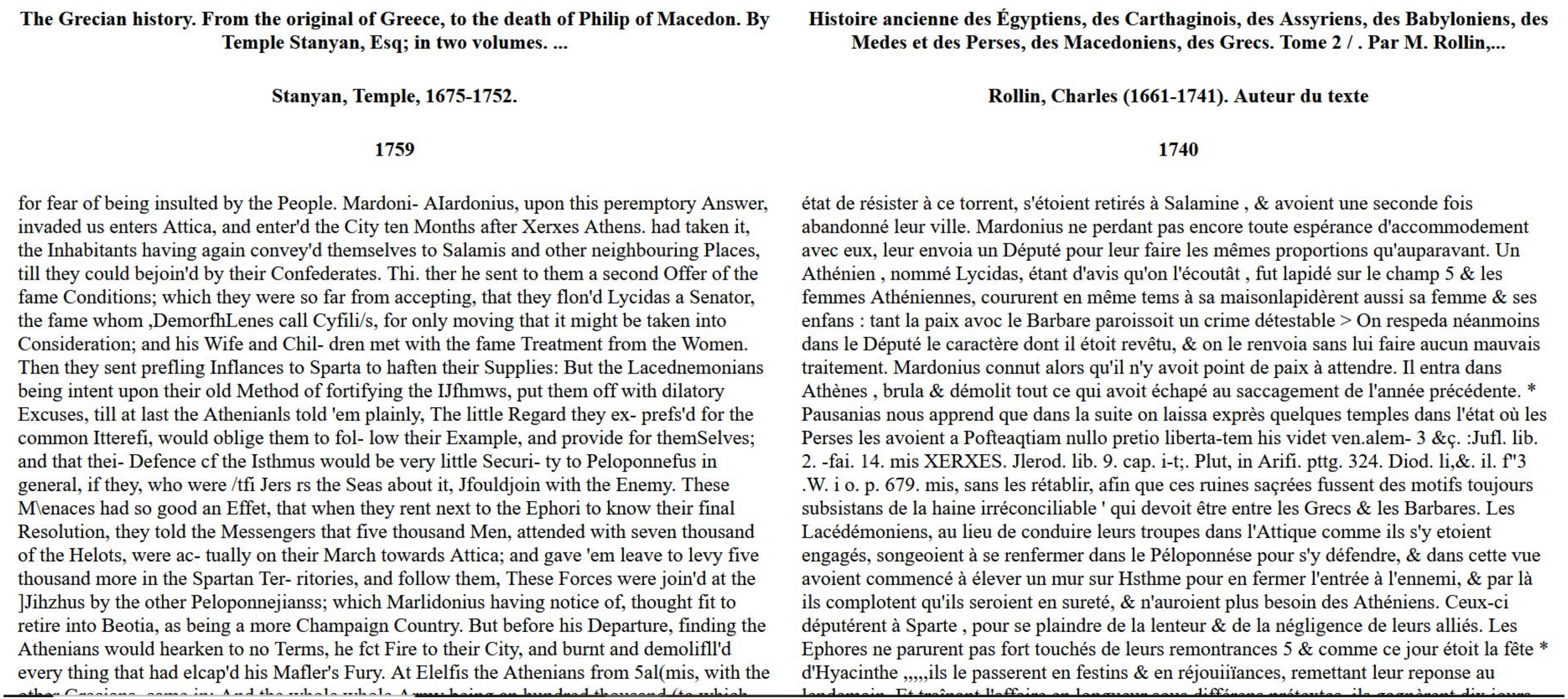
Gallica BNF (~300.000 French books) vs ECCO (~200.000 English books); find pairs of books translated from a language to the other.

TURKUNLP
.ORG

Comhis
Helsinki Computational History Group



Where to draw the line?: Translation vs. topic similarity



The Grecian history. From the original of Greece, to the death of Philip of Macedon. By Temple Stanyan, Esq; in two volumes. ...

Stanyan, Temple, 1675-1752.

1759

for fear of being insulted by the People. Mardoni- AIardonius, upon this peremptory Answer, invaded us enters Attica, and enter'd the City ten Months after Xerxes Athens. had taken it, the Inhabitants having again convey'd themselves to Salamis and other neighbouring Places, till they could bejoin'd by their Confederates. Thi. ther he sent to them a second Offer of the fame Conditions; which they were so far from accepting, that they flon'd Lycidas a Senator, the fame whom ,DemorfhLenes call Cyfili/s, for only moving that it might be taken into Consideration; and his Wife and Chil- dren met with the fame Treatment from the Women. Then they sent prefling Inflances to Sparta to haften their Supplies: But the Lacednemonians being intent upon their old Method of fortifying the IJfhmws, put them off with dilatory Excuses, till at last the Athenianls told 'em plainly, The little Regard they ex- prefs'd for the common Itterefi, would oblige them to fol- low their Example, and provide for themSelves; and that thei- Defence cf the Isthmus would be very little Securi- ty to Peloponnefus in general, if they, who were /tfi Jers rs the Seas about it, Jfouldjoin with the Enemy. These M\enaces had so good an Effet, that when they rent next to the Ephori to know their final Resolution, they told the Messengers that five thousand Men, attended with seven thousand of the Helots, were ac- tually on their March towards Attica; and gave 'em leave to levy five thousand more in the Spartan Ter- ritories, and follow them, These Forces were join'd at the]Jihzhus by the other Peloponnejianss; which Marlidonius having notice of, thought fit to retire into Beotia, as being a more Champaign Country. But before his Departure, finding the Athenians would hearken to no Terms, he fct Fire to their City, and burnt and demolifll'd every thing that had elcap'd his Mafler's Fury. At Elelfis the Athenians from 5al(mis, with the

Histoire ancienne des Égyptiens, des Carthaginois, des Assyriens, des Babyloniens, des Medes et des Perses, des Macedoniens, des Grecs. Tome 2 / . Par M. Rollin,...

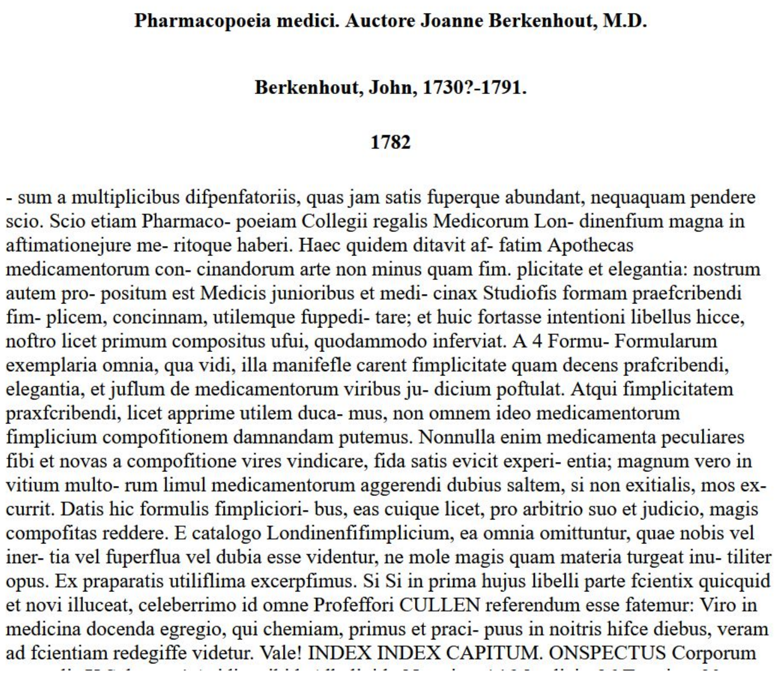
Rollin, Charles (1661-1741). Auteur du texte

1740

état de résister à ce torrent, s'étoient retirés à Salamine , & avoient une seconde fois abandonné leur ville. Mardonius ne perdant pas encore toute espérance d'accommodement avec eux, leur envoia un Député pour leur faire les mêmes proportions qu'auparavant. Un Athénien , nommé Lycidas, étant d'avis qu'on l'écoutât , fut lapidé sur le champ 5 & les femmes Athéniennes, coururent en même tems à sa maisonlapidèrent aussi sa femme & ses enfans : tant la paix avoc le Barbare paroissoit un crime détestable > On respeda néanmoins dans le Député le caractère dont il étoit revêtu, & on le renvoia sans lui faire aucun mauvais traitement. Mardonius connut alors qu'il n'y avoit point de paix à attendre. Il entra dans Athènes , brula & démolit tout ce qui avoit échapé au saccagement de l'année précédente. * Pausanias nous apprend que dans la suite on laissa exprès quelques temples dans l'état où les Perses les avoient a Pofteaqtiam nullo pretio liberta-tem his videt ven.alem- 3 &ç. :Jufl. lib. 2. -fai. 14. mis XERXES. Jlerod. lib. 9. cap. i-t;. Plut, in Arifi. pttg. 324. Diod. li,&. il. f"3 .W. i o. p. 679. mis, sans les rétablir, afin que ces ruines saçrées fussent des motifs toujours subsistans de la haine irréconciliable ' qui devoit être entre les Grecs & les Barbares. Les Lacédémoniens, au lieu de conduire leurs troupes dans l'Attique comme ils s'y etoient engagés, songeoient à se renfermer dans le Péloponnése pour s'y défendre, & dans cette vue avoient commencé à élever un mur sur Hsthme pour en fermer l'entrée à l'ennemi, & par là ils complotent qu'ils seroient en sureté, & n'auroient plus besoin des Athéniens. Ceux-ci députérent à Sparte , pour se plaindre de la lenteur & de la négligence de leurs alliés. Les Ephores ne parurent pas fort touchés de leurs remontrances 5 & comme ce jour étoit la fête * d'Hyacinthe ,,,,,ils le passerent en festins & en réjouiiïances, remettant leur reponse au

Is talking about the same topic in a different way enough?

Noise: “It’s all Latin to me”

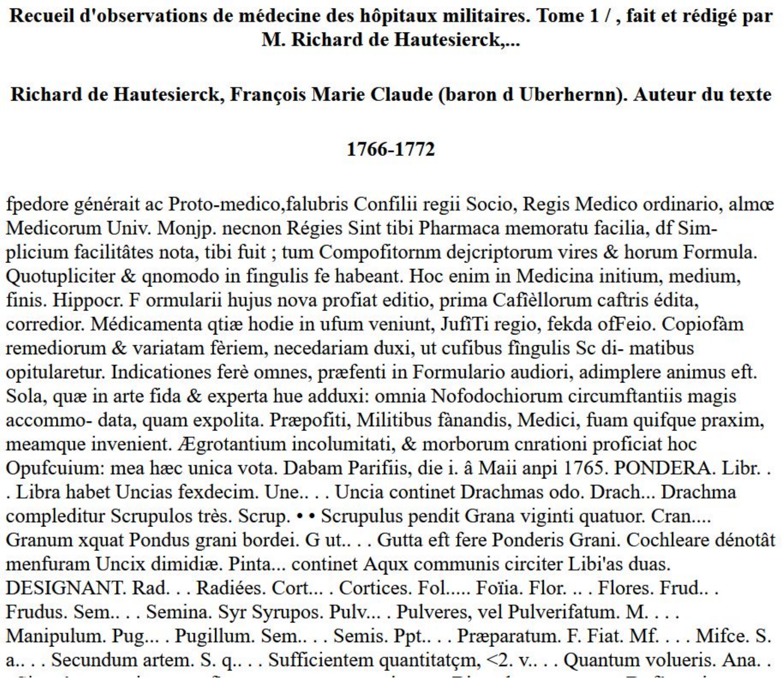


Pharmacopoeia medici. Auctore Joanne Berkenhout, M.D.

Berkenhout, John, 1730?-1791.

1782

- sum a multiplicibus difpenfatoriis, quas jam satis fuperque abundant, nequaquam pendere scio. Scio etiam Pharmaco- poeiam Collegii regalis Medicorum Lon- dinenfium magna in aftimationejure me- ritoque haberi. Haec quidem ditavit af- fatim Apothecas medicamentorum con- cinandorum arte non minus quam fim. plicitate et elegantia: nostrum autem pro- positum est Medicis junioribus et medi- cinax Studiofis formam praefcribendi fim- plicem, concinnam, utilemque fuppedi- tare; et huic fortasse intentioni libellus hicce, noftro licet primum compositus ufui, quodammodo inferviat. A 4 Formu- Formularum exemplaria omnia, qua vidi, illa manifefle carent fimplicitate quam decens prafcribendi, elegantia, et juflum de medicamentorum viribus ju- dicium poftulat. Atqui fimplicitatem praxfcribendi, licet apprime utilem duca- mus, non omnem ideo medicamentorum fimplicium compofitionem damnandam putemus. Nonnulla enim medicamenta peculiares fibi et novas a compofitione vires vindicare, fida satis evicit experi- entia; magnum vero in vitium multo- rum limul medicamentorum aggerendi dubius saltem, si non exitialis, mos excurrit. Datis hic formulis fimpliciori- bus, eas cuique licet, pro arbitrio suo et judicio, magis compofitas reddere. E catalogo Londinenfifimplicium, ea omnia omittuntur, quae nobis vel iner- tia vel fuperflua vel dubia esse videntur, ne mole magis quam materia turgeat inu- tiliter opus. Ex praparatis utiliflima excerpfimus. Si Si in prima hujus libelli parte fcientix quicquid et novi illuceat, celeberrimo id omne Profeffori CULLEN referendum esse fatemur: Viro in medicina docenda egregio, qui chemiam, primus et praci- puus in noitris hifce diebus, veram ad fcientiam redegiffe videtur. Vale! INDEX INDEX CAPITUM. ONSPECTUS Corporum



Recueil d'observations de médecine des hôpitaux militaires. Tome 1 / , fait et rédigé par M. Richard de Hautesierck,...

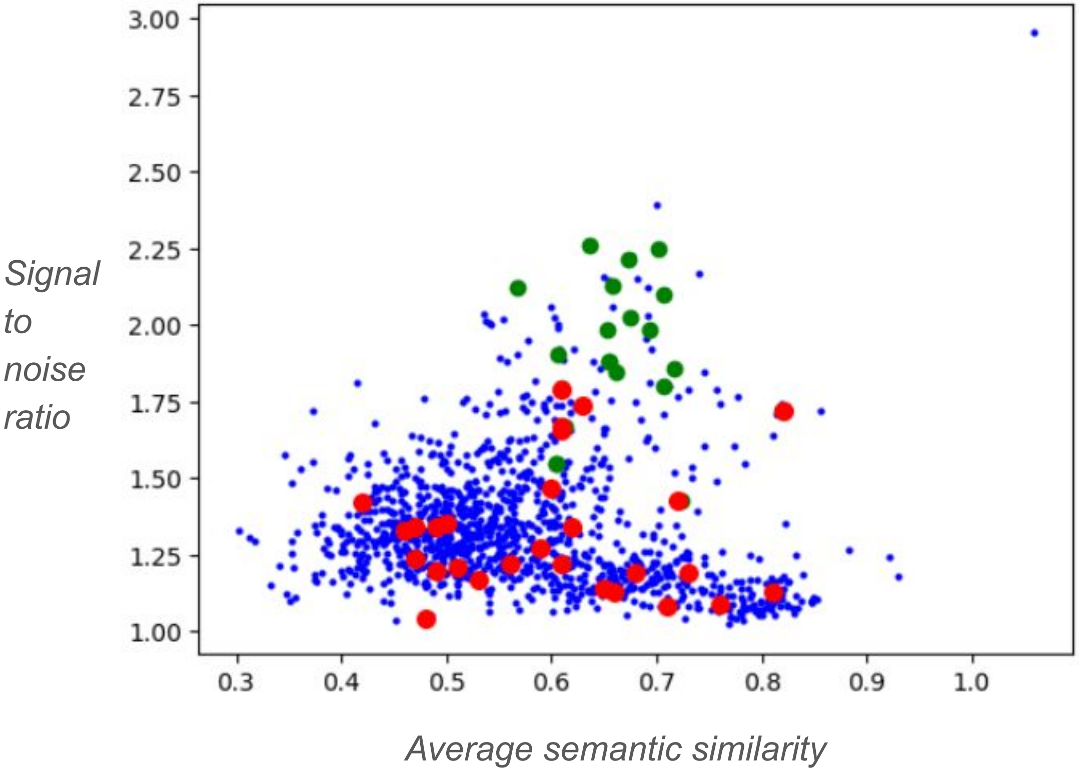
Richard de Hautesierck, François Marie Claude (baron d Uberhernn). Auteur du texte

1766-1772

fpedore générait ac Proto-medico,falubris Confilii regii Socio, Regis Medico ordinario, almœ Medicorum Univ. Monjp. necnon Régies Sint tibi Pharmaca memoratu facilia, df Simplicium facilitâtes nota, tibi fuit ; tum Compofitornm dejcriptorum vires & horum Formula. Quotupliciter & qnomodo in fingulis fe habeant. Hoc enim in Medicina initium, medium, finis. Hippocr. F ormularii hujus nova profiat editio, prima Cafièllorum caftris édita, corredior. Médicamenta qtiæ hodie in ufum veniunt, JufîTi regio, fekda ofFeio. Copiofàm remediorum & variatam fèriem, necedariam duxi, ut cufibus fîngulis Sc di- matibus opitularetur. Indicationes ferè omnes, præfenti in Formulario audiori, adimplere animus eft. Sola, quæ in arte fida & experta hue adduxi: omnia Nofodochiorum circumftantiis magis accommo- data, quam expolita. Præpofiti, Militibus fànandis, Medici, fuam quifque praxim, meamque invenient. Ægrotantium incolumitati, & morborum cnrationi proficiat hoc Opufcuium: mea hæc unica vota. Dabam Parifiis, die i. â Maii anpi 1765. PONDERA. Libr. . . Libra habet Uncias fexdecim. Une.. . . Uncia continet Drachmas odo. Drach... Drachma completitur Scrupulos très. Scrup. • • Scrupulus pendit Grana viginti quatuor. Cran.... Granum xquat Pondus grani bordei. G ut.. . . Gutta eft fere Ponderis Grani. Cochleare dénotât menfuram Uncix dimidiæ. Pinta... continet Aqux communis circiter Libi'as duas. DESIGNANT. Rad. . . Radiées. Cort... . Cortices. Fol..... Foïia. Flor. .. . Flores. Frud.. . Frudus. Sem.. . . Semina. Syr Syrupos. Pulv... . Pulveres, vel Pulverifatum. M. . . . Manipulum. Pug... . Pugillum. Sem.. . . Semis. Ppt.. . . Præparatum. F. Fiat. Mf. . . . Mifce. S. a.. . . Secundum artem. S. q.. . . Sufficientem quantitatçm, <2. v.. . . Quantum volueris. Ana. .

Other noise: OCR artifacts (e.g. tables, formulae), Bible commentaries

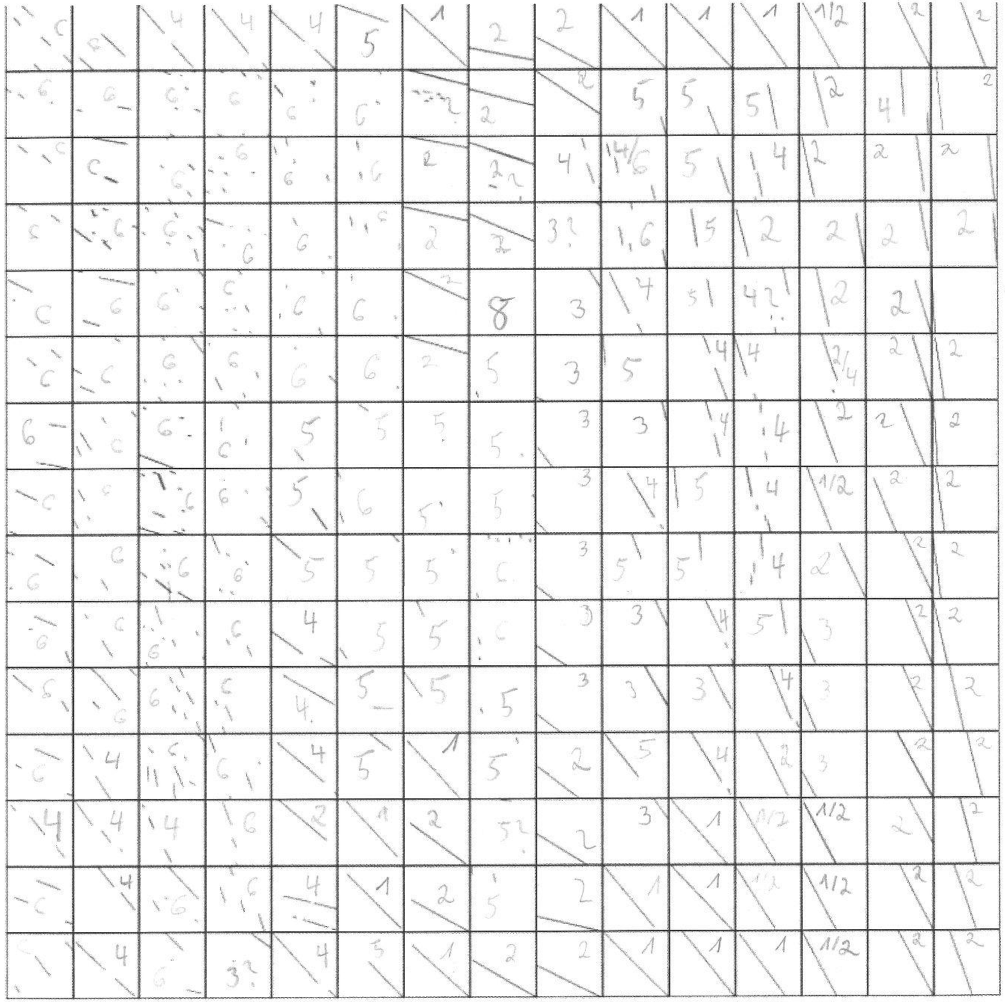
What is a translation?: Evaluation at scale



Blue : line scores for a random 1000 pairs

Green : pairs manually classified as translations

Red : pairs manually rejected



6	6	4	4	4	5	1	2	2	1	1	1	1/2	2	2
6	6	6	6	6	6	2	2	2	5	5	5	2	4	2
6	6	6	6	6	6	2	2	4	4/6	5	4	2	2	2
6	6	6	6	6	5	2	2	3?	1,6	5	2	2	2	2
6	6	6	6	6	6	2	8	3	4	5	4?	2	2	
6	6	6	6	6	6	2	5	3	5	4	4	2/4	2	2
6	6	6	6	5	5	5	5	3	3	4	4	2	2	2
6	6	6	6	5	6	5	5	3	4	5	4	1/2	2	2
6	6	6	6	5	5	5	6	3	5	5	4	2	2	2
6	6	6	6	4	5	5	6	3	3	4	5	3	2	2
6	6	6	6	4	5	5	5	3	3	3	4	3	2	2
6	4	6	6	4	5	1	5	2	5	4	2	3	2	2
4	4	4	6	2	1	2	5?	2	3	1	1/2	1/2	2	2
6	4	6	6	4	1	2	5	2	1	1	1/2	1/2	2	2
6	4	6	3?	4	5	1	2	2	1	1	1	1/2	2	2

Human(ist)-Computer (Scientist) Interaction: Developing a translation taxonomy

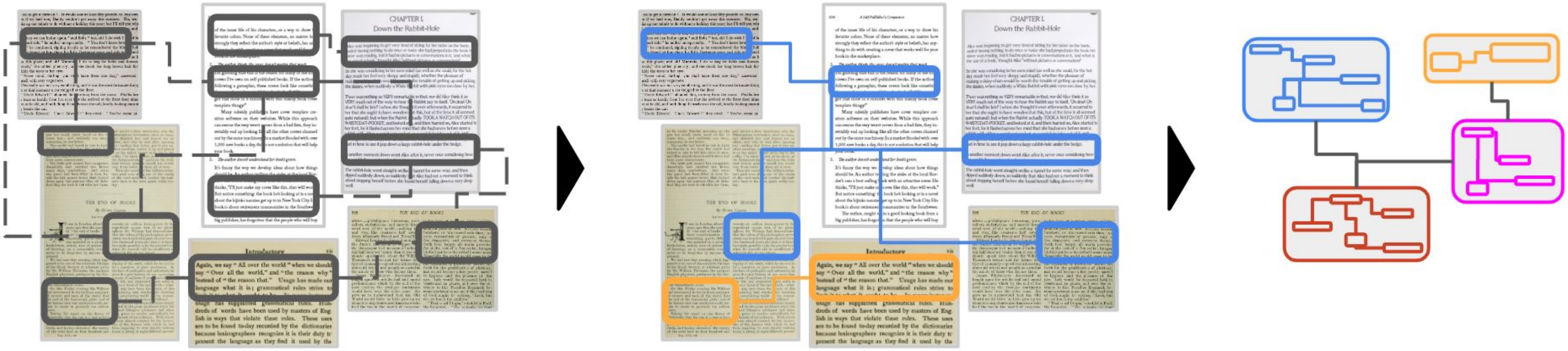
Use of translation matching in intellectual history

Conceptually not that far from paraphrasing, but leads to fundamental issues:

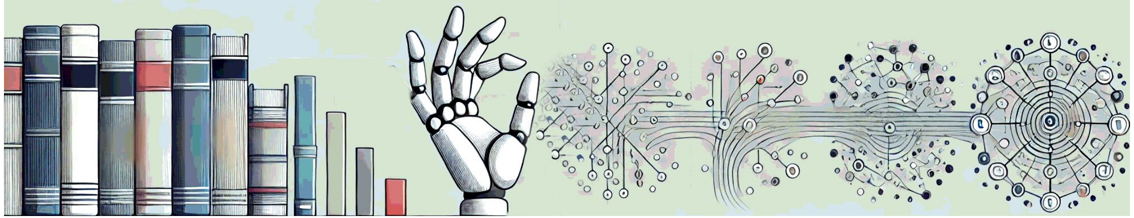
- *What is a translation? At what point does a text stop being a translation?*

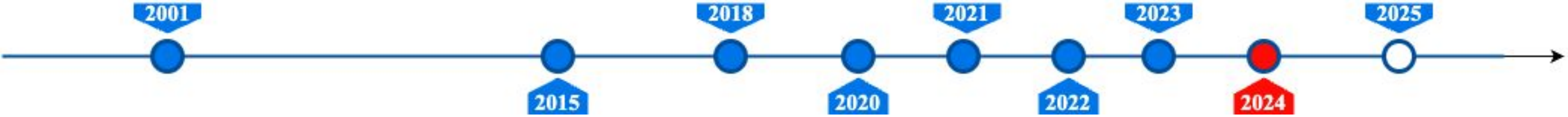
Cross-lingual elements change the setting

- Cultural influences, more comprehensive possibilities to think about cross-border dissemination of ideas.



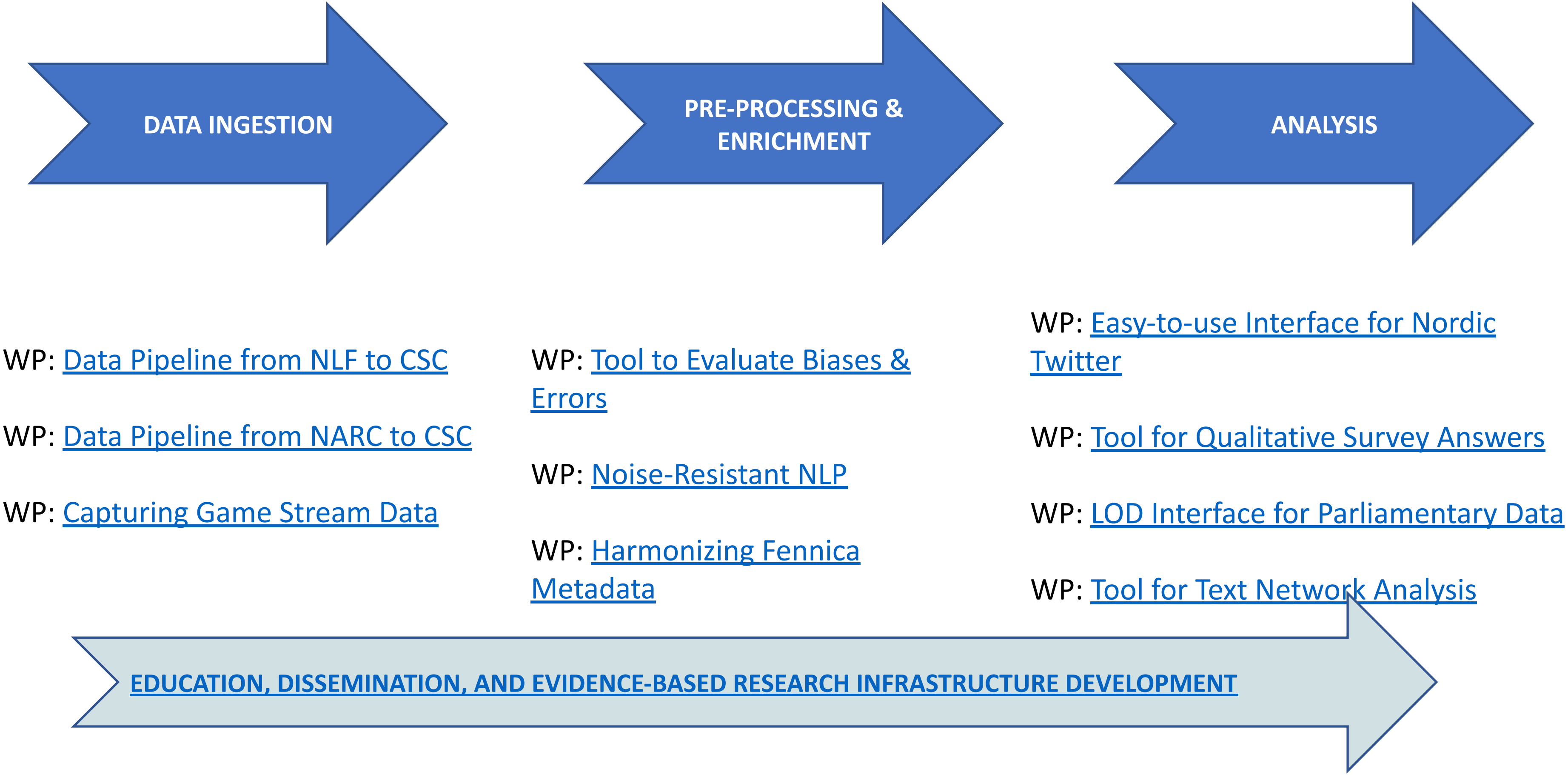
Crossing linguistic boundaries; combining bibliodata; use of semantic similarity; **NLP and transformers** are needed ***as one set of tools***.





A bit more on DARIAH-FI

DARIAH-FI WORK PACKAGES



DATA INGESTION

- WP: Data Pipeline from NLF to CSC
- WP: Data Pipeline from NARC to CSC
- WP: Capturing Game Stream Data

PRE-PROCESSING & ENRICHMENT

- WP: Tool to Evaluate Biases & Errors
- WP: Noise-Resistant NLP
- WP: Harmonizing Fennica Metadata

ANALYSIS

- WP: Easy-to-use Interface for Nordic Twitter
- WP: Tool for Qualitative Survey Answers
- WP: LOD Interface for Parliamentary Data
- WP: Tool for Text Network Analysis

EDUCATION, DISSEMINATION, AND EVIDENCE-BASED RESEARCH INFRASTRUCTURE DEVELOPMENT

DIGITAL HUMANITIES HACKATHON
MAY 11-15 2015 HELSINKI

DIGITAL HUMANITIES HACKATHON
MAY 2016 HELSINKI

DIGITAL HUMANITIES HACKATHON
MAY 2017 HELSINKI

DIGITAL HUMANITIES HACKATHON
MAY 2018 HELSINKI

DIGITAL HUMANITIES HACKATHON
HELSINKI

DIGITAL HUMANITIES HACKATHON
MAY 2019 HELSINKI

DIGITAL HUMANITIES HACKATHON

DIGITAL HUMANITIES HACKATHON

DIGITAL HUMANITIES HACKATHON
HELSINKI

#DHH

- The Helsinki Digital Humanities Hackathon is a chance to experience an interdisciplinary research project from start to finish within the span of 10 days.
    - For researchers and students from computer science and data science, the hackathon gives the opportunity to test their abstract knowledge against complex real-life problems.
    - For people from the humanities and social sciences, it shows what is possible to achieve with such collaboration.
    - For both, the hackathon gives the experience of intensely working with people from different backgrounds as part of an interdisciplinary team.

FIN-CLARIAH MODULES

THEMES	MODULES	FOCUS AREAS
Text and speech processing and annotation	NLP	Standard and colloquial Finnish
Licensing and storing language data	Language Research Infrastructure	Specialized needs in language-based research
Data access and documentation	Social Sciences and Humanities Big Data	Digitised and born-digital data
Research methods and tool development	Analytica	Computational techniques and environments
Dissemination of best practices	Information Interaction	Researcher support

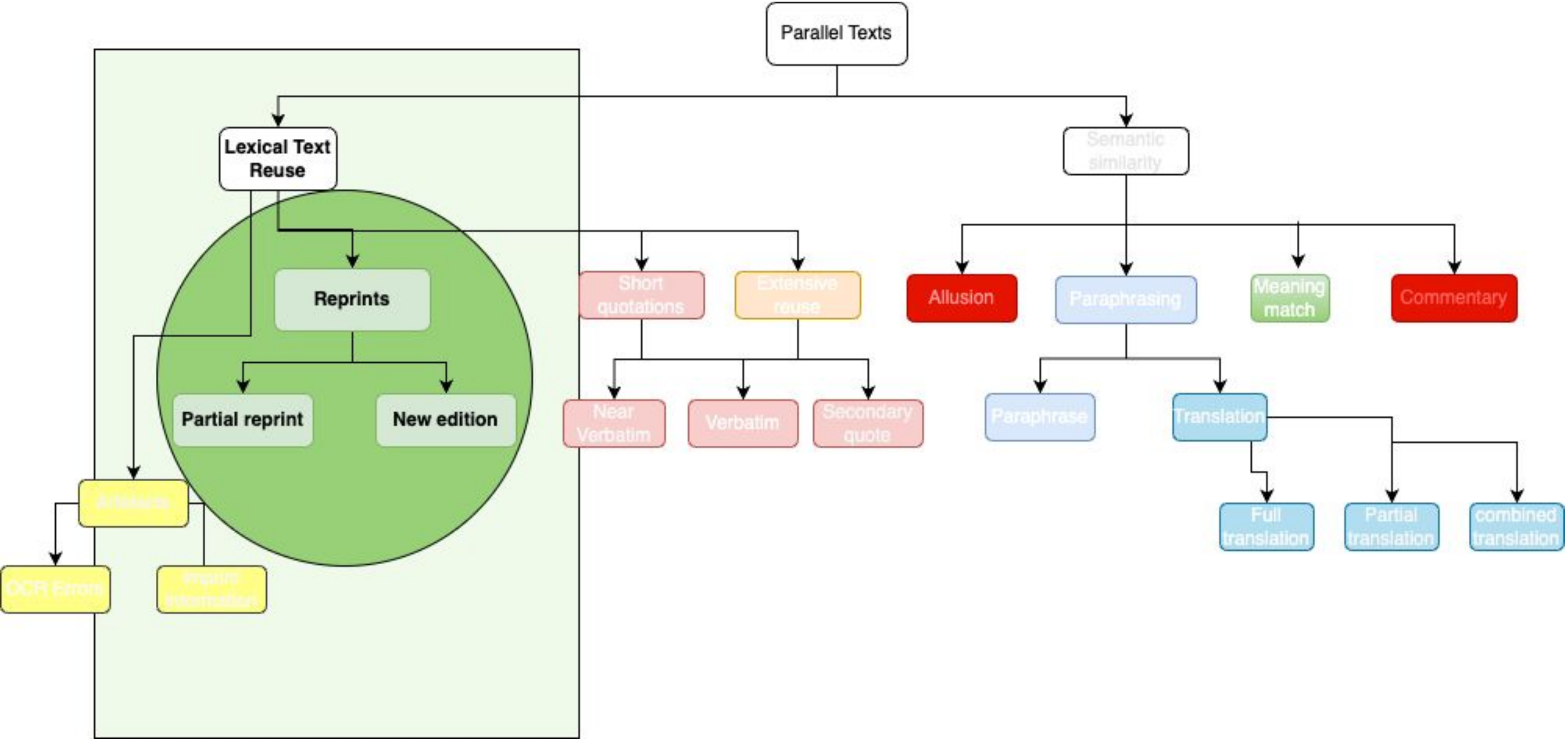
Low resource languages and LLMs

- Transformer models for most languages trained mainly on English - what to do, is this the best way forward, consequences?
- In practice people are already implementing things like this - do they work?
  - Some things work and something most likely don’t. Some things are easier when we detach from “rule based past”, that’s for sure.
  - https://turkunlp.org/gpt3-finnish -> this is now tested in many different contexts; but the fine tuning question might be better for many.
- Nevertheless, there is a move from language-specific to more general, the issue currently is that there’s no real evaluation results yet of what works and why
- Perhaps here is the good niche for EU funding for non-indo-European languages combined to historical language?

DARIAH-FI 2.0

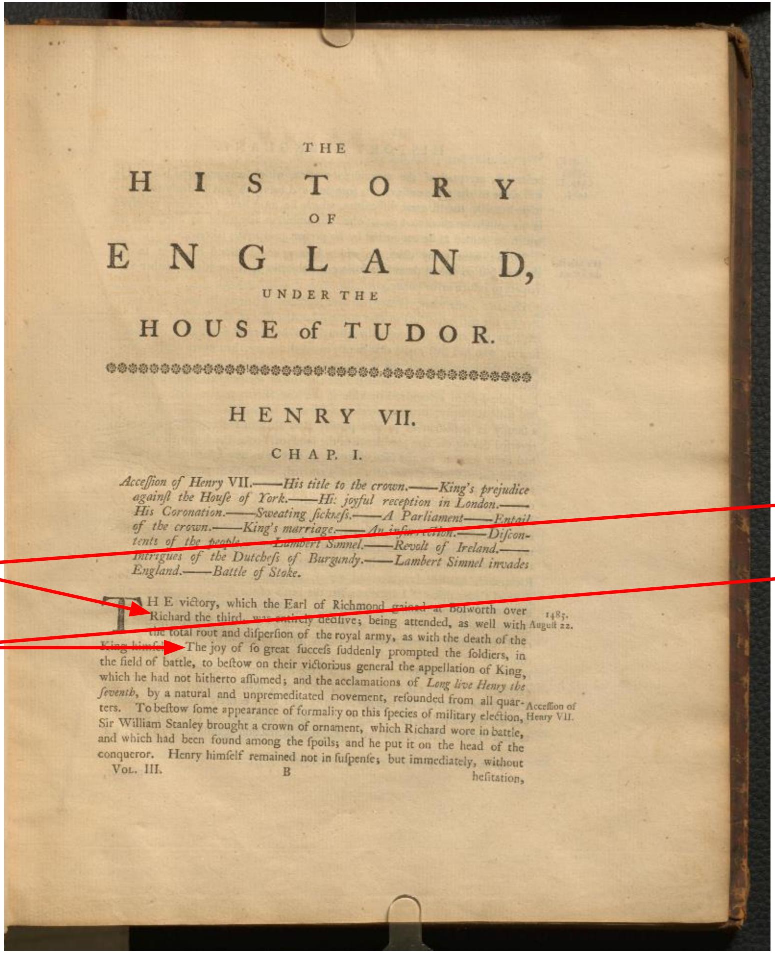
STRAIGHT OUTTA NORTH POLE

Edition and emblem mapping: Extensions from text reuse



Same chapter in two different editions

- Different layouts
- Minor changes in text



THE

HISTORY OF ENGLAND,

UNDER THE

HOUSE of TUDOR.

HENRY VII.

CHAP. I.

Acceſſion of Henry VII.——*His title to the crown.*——*King's prejudice againſt the Houſe of York.*——*Hi: joyful reception in London.*——*His Coronation.*——*Sweating ſickneſs.*——*A Parliament*——*Entail of the crown.*——*King's marriage.*——*An inſurrection.*——*Diſcontents of the people.*——*Lambert Simnel.*——*Revolt of Ireland.*——*Intrigues of the Dutcheſs of Burgundy.*——*Lambert Simnel invades England.*——*Battle of Stoke.*

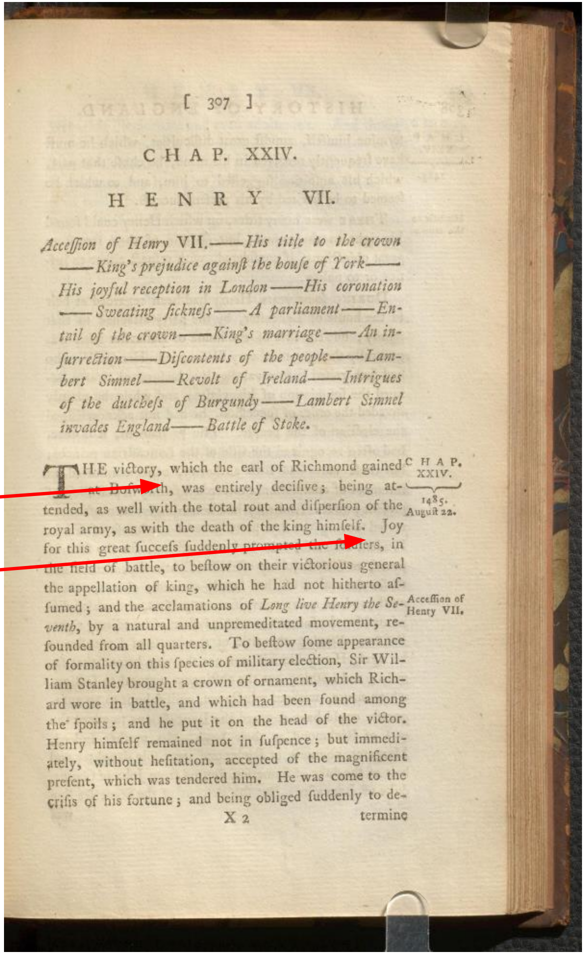
THE victory, which the Earl of Richmond gained at Boſworth over Richard the third, was entirely deciſive; being attended, as well with the total rout and diſperſion of the royal army, as with the death of the King himſelf. The joy of ſo great ſucceſs ſuddenly prompted the ſoldiers, in the field of battle, to beſtow on their victorious general the appellation of King, which he had not hitherto aſſumed; and the acclamations of *Long live Henry the ſeventh*, by a natural and unpremeditated movement, reſounded from all quarters. To beſtow ſome appearance of formality on this ſpecies of military election, Sir William Stanley brought a crown of ornament, which Richard wore in battle, and which had been found among the ſpoils; and he put it on the head of the conqueror. Henry himſelf remained not in ſuſpenſe; but immediately, without heſitation,

1485. August 22.

Acceſſion of Henry VII.

VOL. III. B

History of England, vol 3, 1762 page 1



[307]

CHAP. XXIV.

HENRY VII.

Acceſſion of Henry VII.——*His title to the crown*——*King's prejudice againſt the houſe of York*——*His joyful reception in London*——*His coronation*——*Sweating ſickneſs*——*A parliament*——*Entail of the crown*——*King's marriage*——*An inſurrection*——*Diſcontents of the people*——*Lambert Simnel*——*Revolt of Ireland*——*Intrigues of the dutcheſs of Burgundy*——*Lambert Simnel invades England*——*Battle of Stoke.*

THE victory, which the earl of Richmond gained at Boſworth, was entirely deciſive; being attended, as well with the total rout and diſperſion of the royal army, as with the death of the king himſelf. Joy for this great ſucceſs ſuddenly prompted the ſoldiers, in the field of battle, to beſtow on their victorious general the appellation of king, which he had not hitherto aſſumed; and the acclamations of *Long live Henry the Seventh*, by a natural and unpremeditated movement, reſounded from all quarters. To beſtow ſome appearance of formality on this ſpecies of military election, Sir William Stanley brought a crown of ornament, which Richard wore in battle, and which had been found among the ſpoils; and he put it on the head of the victor. Henry himſelf remained not in ſuſpence; but immediately, without heſitation, accepted of the magnificent preſent, which was tendered him. He was come to the criſis of his fortune; and being obliged ſuddenly to determine

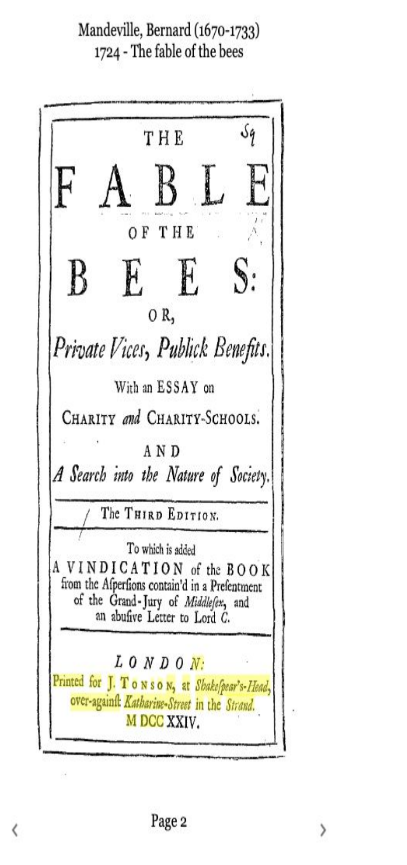
CHAP. XXIV. 1485. Auguſt 22.

Acceſſion of Henry VII.

X 2

History of England, vol 3, 1778 page 307

Reuse of printer information



Mandeville, Bernard (1670-1733)
1724 - The fable of the bees

59

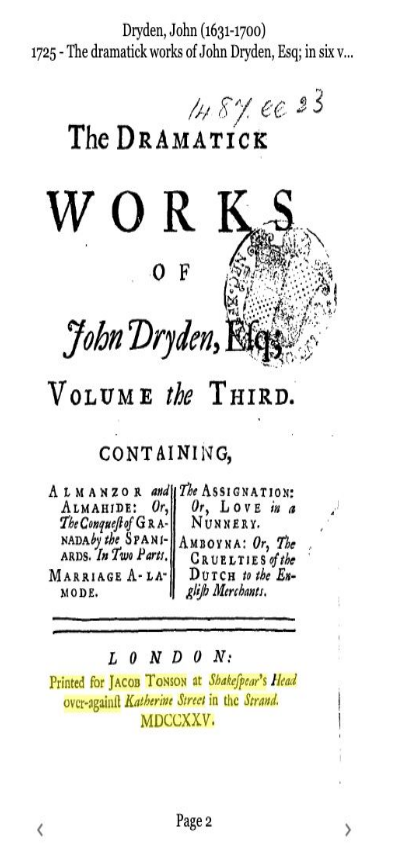
THE
FABLE
OF THE
BEES:
OR,
Private Vices, Publick Benefits.
With an ESSAY on
CHARITY *and* CHARITY-SCHOOLS.
AND
A Search into the Nature of Society.

The THIRD EDITION.

To which is added
A VINDICATION of the BOOK
from the Aſperſions contain'd in a Preſentment
of the Grand-Jury of *Middleſex*, and
an abuſive Letter to Lord *C.*

LONDON:
Printed for J. TONSON, at *Shakeſpear's-Head*,
over-againſt *Katharine-Street* in the *Strand*.
M DCC XXIV.

Page 2



Dryden, John (1631-1700)
1725 - The dramatick works of John Dryden, Esq; in six v...

1487.ee 23

The DRAMATICK
WORKS
OF
John Dryden, Eſq;
VOLUME *the* THIRD.

CONTAINING,

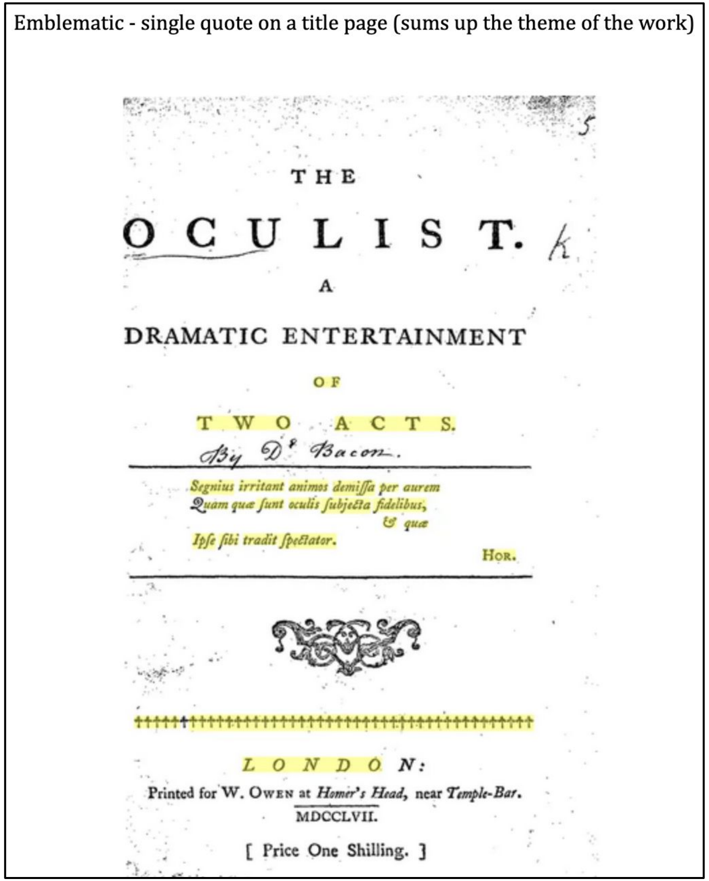
ALMANZOR *and* ALMAHIDE: *Or*, *The Conqueſt of* GRANADA *by the* SPANIARDS. *In Two Parts.*
MARRIAGE A-LA-MODE.

The ASSIGNATION: *Or*, LOVE *in a* NUNNERY.
AMBOYNA: *Or*, *The* CRUELTIES *of the* DUTCH *to the Engliſh Merchants.*

LONDON:
Printed for JACOB TONSON at *Shakeſpear's Head*
over-againſt *Katherine Street* in the *Strand*.
MDCCXXV.

Page 2

Reuse of emblem on title-page



Emblematic - single quote on a title page (sums up the theme of the work)

5

THE
OCULIST. k
A
DRAMATIC ENTERTAINMENT
OF
TWO ACTS.
By Dr Bacon.

Segnius irritant animos demiſſa per aurem
Quam quæ ſunt oculis ſubjecta fidelibus,
& quæ
Ipſe ſibi tradit ſpectator.
HOR.

LONDON:
Printed for W. OWEN at *Homer's Head*, near *Temple-Bar*.
MDCCLVII.

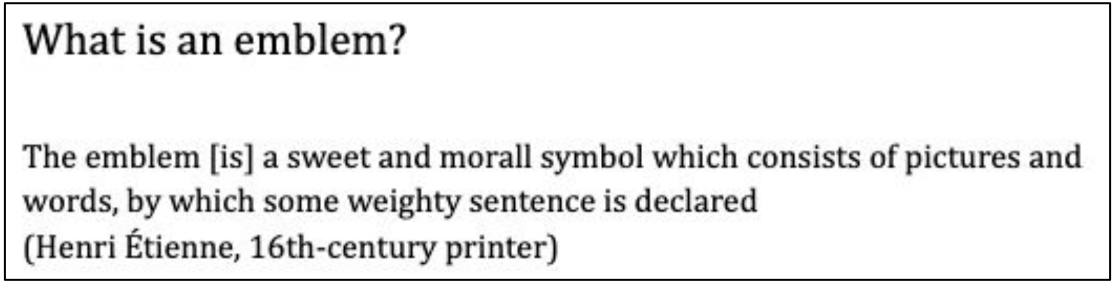
[Price One Shilling.]

Use of artefact detection in intellectual history

- Often dismissed as noise for common research questions.
  - But, reprinting of printer information can be important when studying dissemination and advertising of works.

Study of title-pages can be very useful: Quantifying the Presence of Ancient Greek and Latin Classics in Early Modern Britain

- Latin or English emblematic uses; Diachronic change.



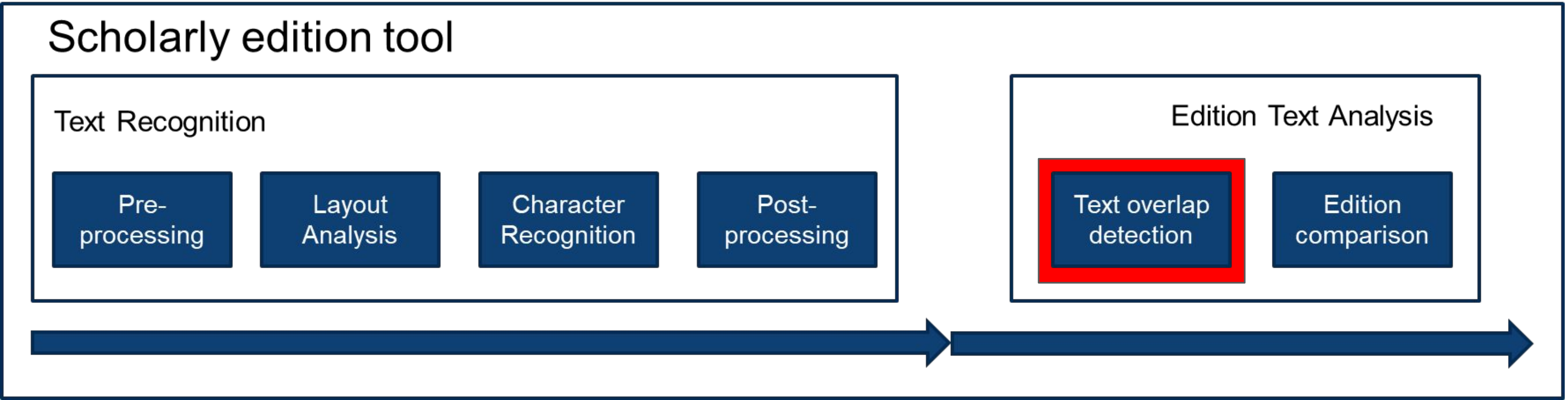
What is an emblem?

The emblem [is] a sweet and morall symbol which consists of pictures and words, by which some weighty sentence is declared
(Henri Étienne, 16th-century printer)

Use of reprint matching in intellectual history

Newspapers as intellectual platform for printing of essays and chapters of different works.

Critical edition of Hume’s History of England for OUP: detecting the changes from edition to another.



Scholarly edition tool

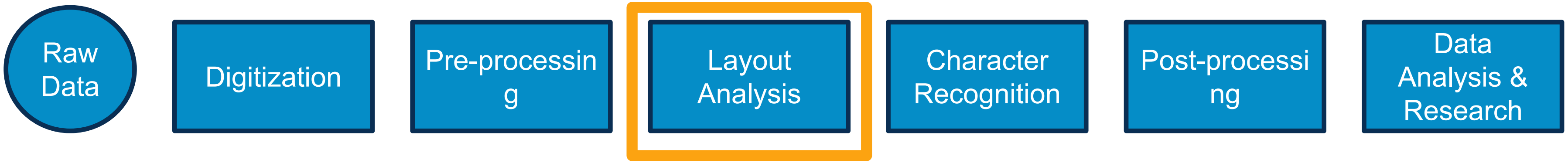
Text Recognition

- Pre-processing
- Layout Analysis
- Character Recognition
- Post-processing

Edition Text Analysis

- Text overlap detection
- Edition comparison

Low resource “Do It Yourself Digitization” -pipeline (simplified)

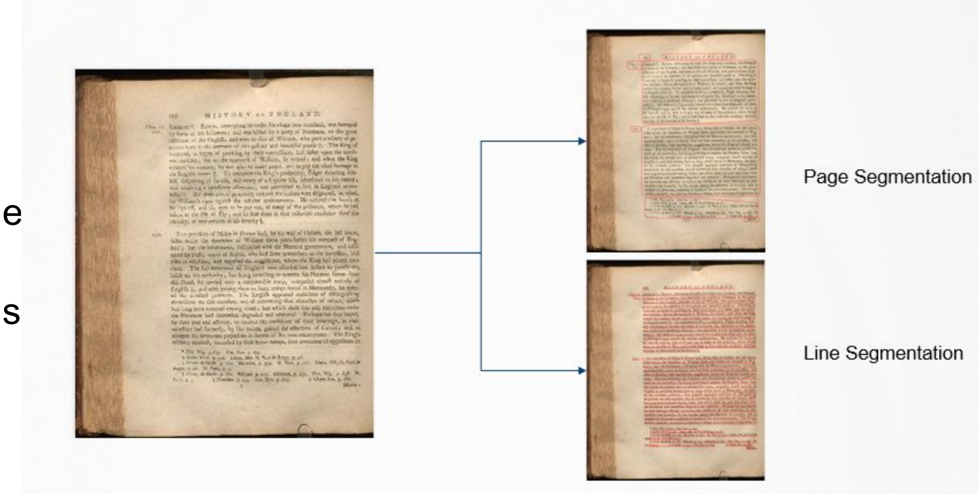


DOCUMENT LAYOUT ANALYSIS

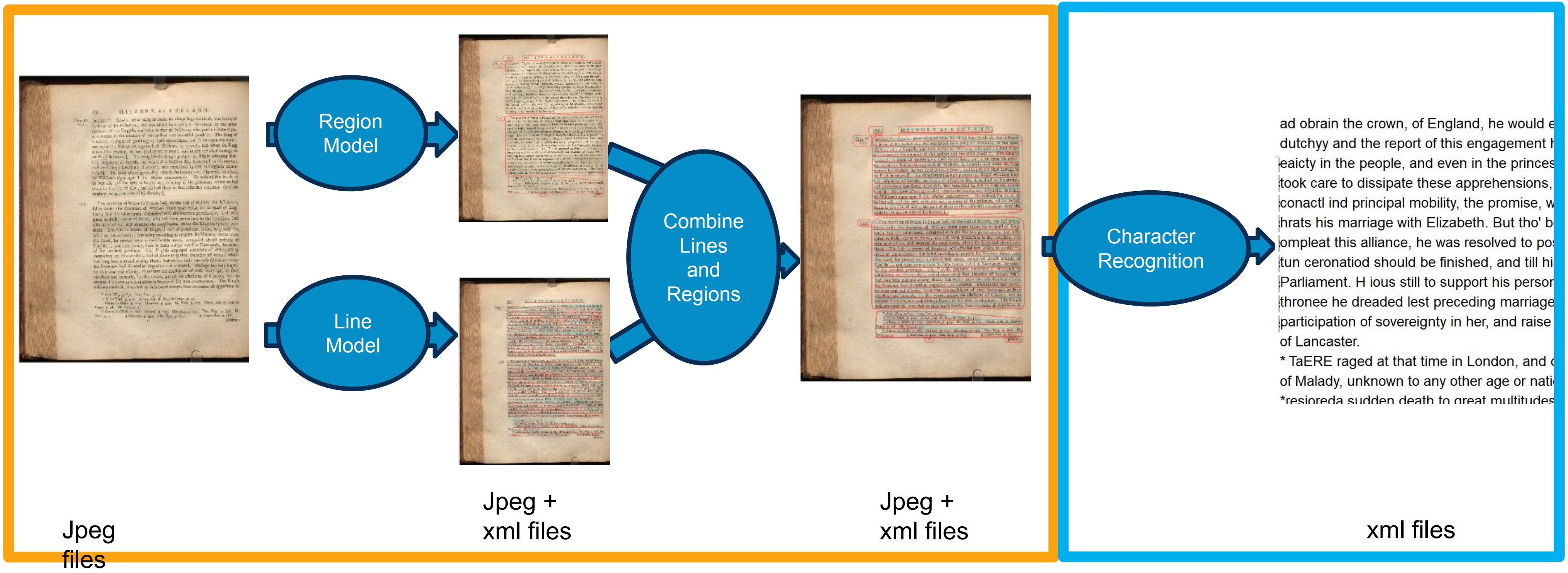
Three main tasks

- Page segmentation: identifies single text regions
- Line segmentation: text lines are extracted from the input image containing only a single text line
- Region classification: classifies each regions

Same tasks applicable for both printed and handwritten documents



TEXT RECOGNITION PIPELINE



ad obrain the crown, of England, he would e
dutchyy and the report of this engagement h
eaicty in the people, and even in the princes
took care to dissipate these apprehensions,
conactl ind principal mobility, the promise, w
hrats his marriage with Elizabeth. But tho' b
ompleat this alliance, he was resolved to po
tun ceronatiod should be finished, and till hi
Parliament. H ious still to support his person
thronee he dreaded lest preceding marriage
participation of sovereignty in her, and raise
of Lancaster.
* TaERE raged at that time in London, and c
of Malady, unknown to any other age or nati
*resioreda sudden death to great multitudes