

AI for Computational History: The Algorithmic Challenges and Opportunities in Tracing Meaning at Scale

Filip Ginter

TurkuNLP, Department of Computing, University of Turku
ELLIS Institute Finland

Filip Ginter – who am I

- One of the co-founders of the TurkuNLP research group
- 25 years in NLP
 - **Finnish and multilingual NLP:** Turku neural parser pipeline, Universal Dependencies
 - **Large corpora + HPC:** Web crawled data and large scale parsing and filtering
 - **LLMs, especially encoders:** TurkuNLP is an important player in LLM pre-training
 - **Digital humanities:** Computational history, OCR-postcorrection, hand-written text recognition in historical documents



The outline of the talk

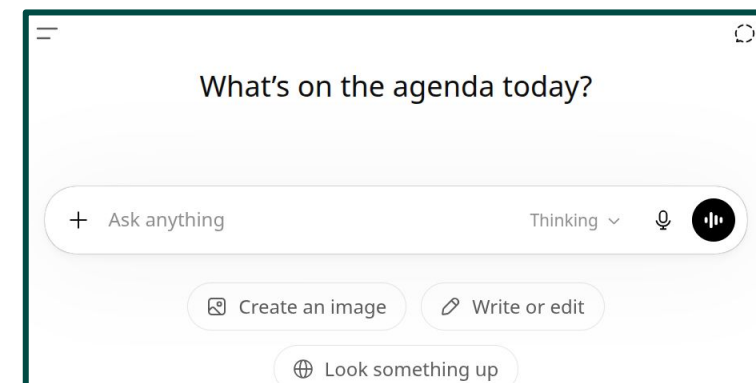
1. LLMs in digital humanities and specifically computational history
2. Tracing meaning as a new research direction enabled by latest encoder models
3. Practical solution to matching meaning at very large scales
4. LLM encoder training and future promise

TL;DR of the talk

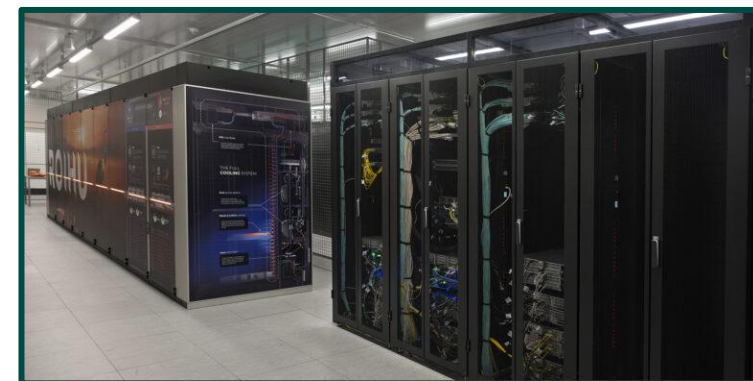
1. LLMs enable serious research with useful outcomes
2. Pushing data through LLM prompt is the usual low-hanging fruit, but exciting opportunities emerge when you go beyond
3. Tracing meaning in time and space, and across languages, can open new research questions in both fields
4. Collaboration is highly encouraged, but be mindful to preserve also own research agenda

LLMs are not a toy!

- LLMs usually evoke ChatGPT-like interactive interface
- Unsuitable for large dataset work
- Highly capable LLMs are available for local inference (Qwen, Llama, GPT OSS...)
 - Commercial inference API much cheaper than many think
 - Finnish national HPC infra (CSC) enables large-scale LLM inference runs



VS



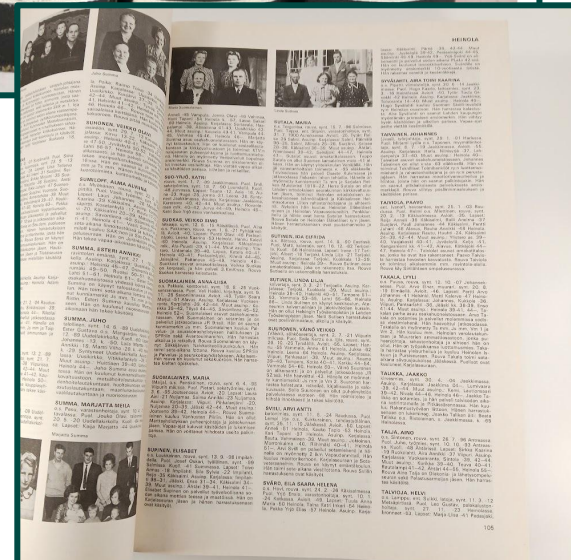
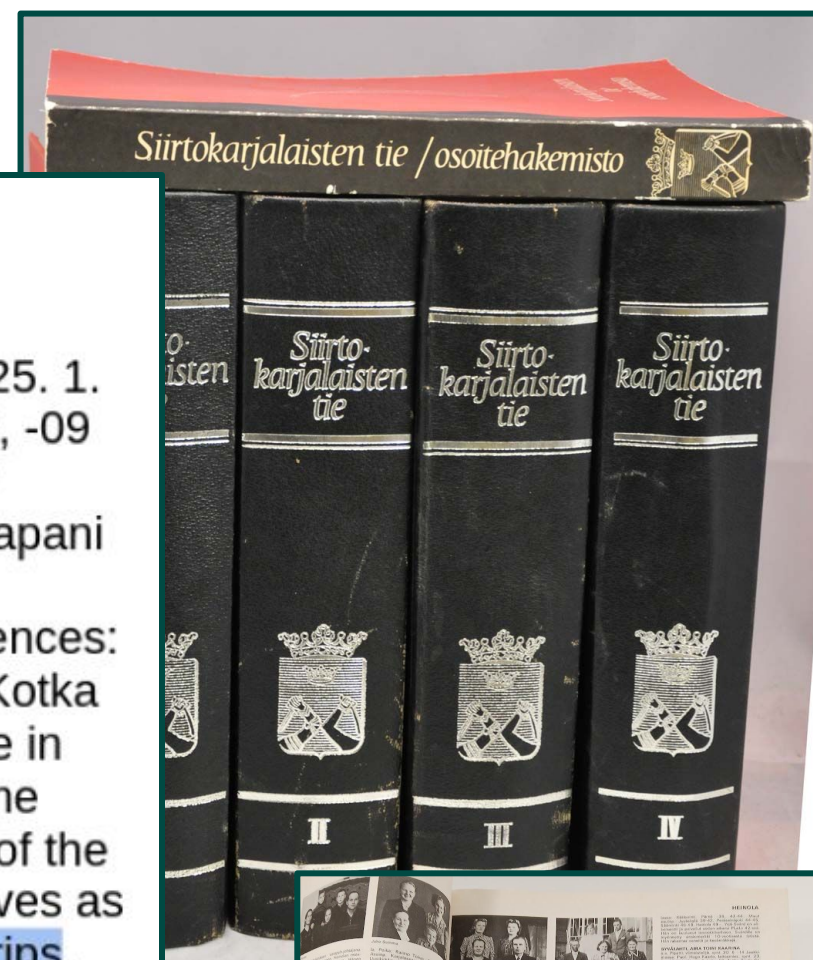
LLM inference

- Push 90,000 interviews through LLMs
- Look for free or low-cost data sources
- LLMs rough
- Total cost: ~400€ in API fees (in 2024)

Primary person: HILKKA NASKI

Spouse: Toivo

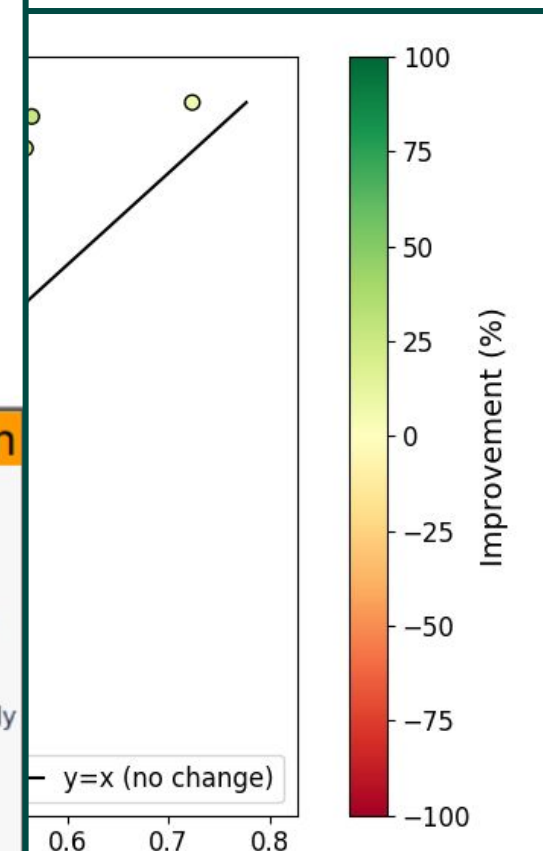
Sourcetext: née Kurki, telephone operator, born 25. 1. -19 Suursaari. Spouse: Toivo, first mate, born 1. 3, -09 Suursaari. Married. -39. Children: Alpo Antero -37 Suursaari. Leena Kaarina -39 Suursaari, Seppo Tapani -43 Loviisa, Eija Marjatta -51 Kotka. Residence in Karelia: Suursaari, Suurkylä 19— 39. Other residences: Loviisa -39, Heinola, Taipale 39— 40, Loviisa 40. Kotka 40—41, Loviisa 41 — 46, Kotka 46—, Naskies live in their own house. Toivo Naski has attended maritime school. He is a maritime pilot. Naski is a member of the Kotka Construction Workers' Union, where he serves as a workplace steward. His hobbies include island trips, outdoor activities, fishing, and traveling. Mrs. has been a member of the Suursaari Marttas. Her hobbies are outdoor activities, visiting islands, and handicrafts.



LLM inference in bulk

- Have LLM correct errors in old books (e.g. OCR errors)
- “Cheap” re-running pipeline
- Result: every word preserved
- Cost: CSC billing units

Original it ill, if I make use of that liberty, which our Friendship seems to allow. I am not so Ignorant either of your natural and acquired Accomplishments, or of my own Defects, as to take upon me to be your Director; for the highest of my ambition is to be your Remembrancer. It is not Improbable but you have already resolved upon most, if not all, the things, which I shall propose to you; and then I hope you will permit	OCR it yll,r I mleufe of that.liberiy, .which o ourantient-' Friendfolip feemsrrto allow.. I am. not flt'gldoihat eithe.i of your natural and acqll ired Acmomtli fhTiegrts, crit- my own DefeAs, as to take upon mi to be -your Di- recour; for the highell of my ambition is to:be your Remembrancer. It is not .Improballe but'Vou' have al- readyrefolvedupon most, if n'ot-all,; the'things,;,'which lhall propose to, .ou, andl. then" li!lope ydu'will permit
LLM correction "It is your, I make use of that liberty, which our ancient Friendship seems to allow. I am not so glad that either of your natural and acquired Accomplishments, or my own Defects, as to take upon me to be your Director; for the highest of my ambition is to be your Remembrancer. It is not improbable but you have already resolved upon most, if not all, the things which I shall propose to you, and then I hope you will permit me..."	

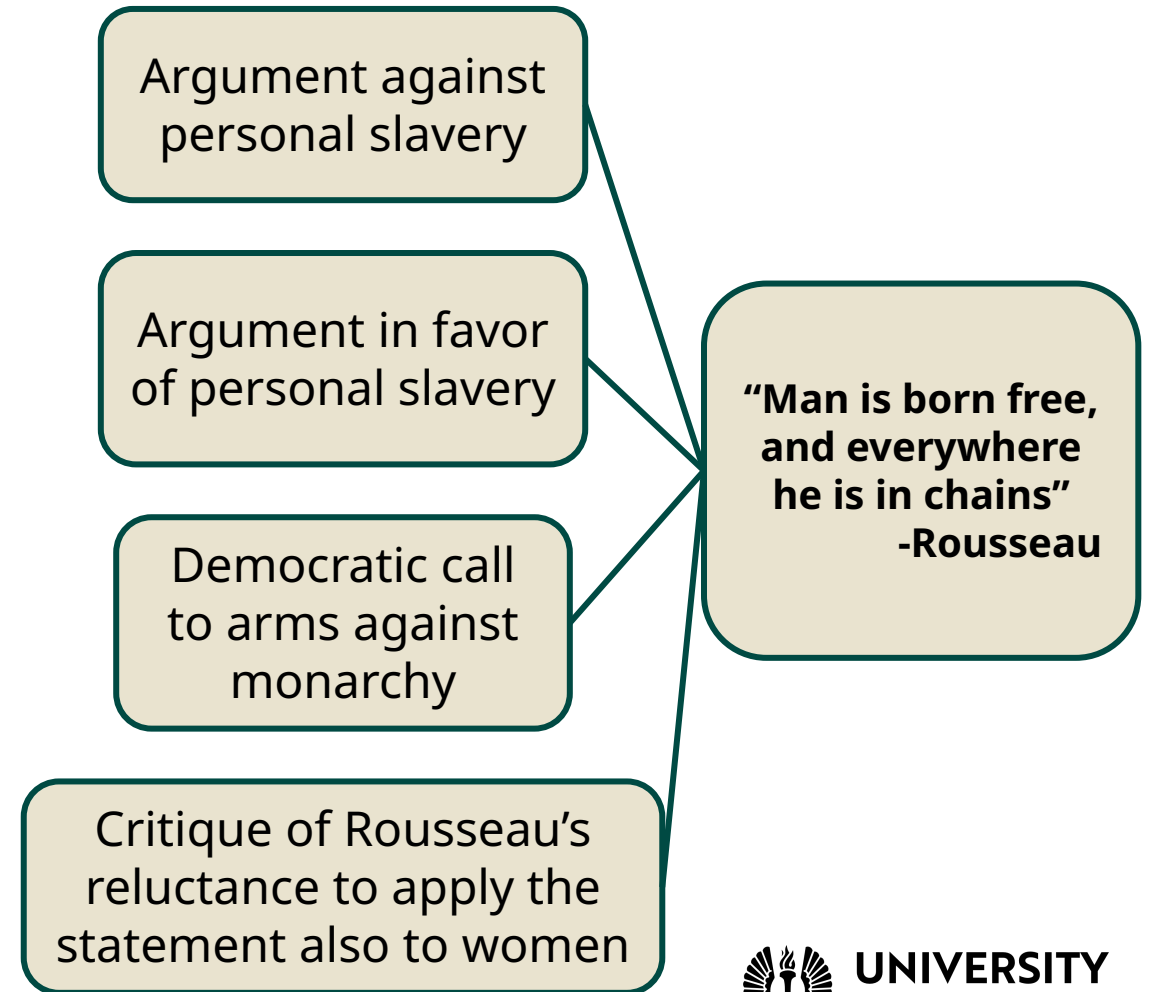


The incentives to collaborate

- **Running large quantities of text through an LLM is useful, but not super-exciting**
 - Why Karelians? Personal interest
 - Why OCR? Fascinated by seemingly superhuman performance of the LLMs on the task
- Is there more we can do with LLMs than simply run loads of data through a prompt on a supercomputer?
- Something we could get excited about also in the algorithmic sense?

Computational intellectual history

- Intellectual history deals with history of ideas
- How fundamental ideas emerge, get reused, misused, and disappear
- The objective is to **trace ideas and their reception in large bodies of historical literature**



Conceptualising Text Reuse and Semantic Similarity

From lexical overlap to meaning matching and translation mining

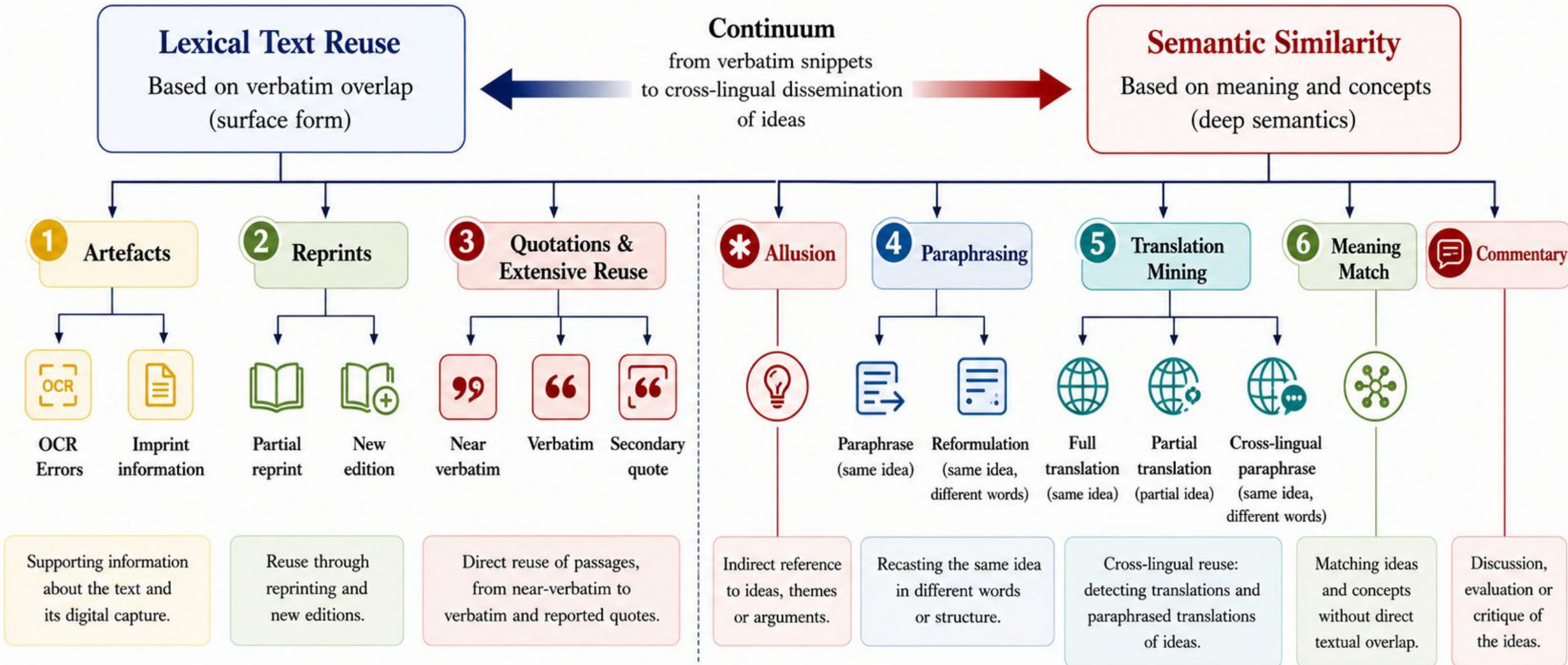


Fig: Mikko Tolonen / Uni. Helsinki

Verbatim text reuse

- Basically string matching complicated by OCR noise
- Can be successfully done e.g. with BLAST
- BLAST: protein database search tool

Multa t\ä@tä fyNlkÄsiii kchtalostu ,ct , Abouil Asi,3 wic!lä ticiun't>t ,mitää>«, »vaalii luiftti iloista M,m<iä Tshiragauissa, ©elä fi:föf3>i'öi että uiUfatfpäim -uhkaisiloui i Hviarat, miinto fu^tiaani 'fatifefi- fuffotai» IÄuja THi roinin, puutarhassa ja, ipici'ilitsi hwi'tt<iiöii fmmiamcrk^iUi ja anoo» » imilyMla,

Mutta tästä synkästä kohtalosta ei Abbul Asib »ielä tiennyt mitään, vaan »ietti iloista elämää TshiraganiSsa. Sekä sis»Stä «ttä ulkoapäin uhkasivat «aarat. mutta sulttaani katseli lukkotaisteluja Tfhiaaananin puutarhassa ja palkitsi voittajan lunnicnnerleillä ja ar° vonimityksillä.

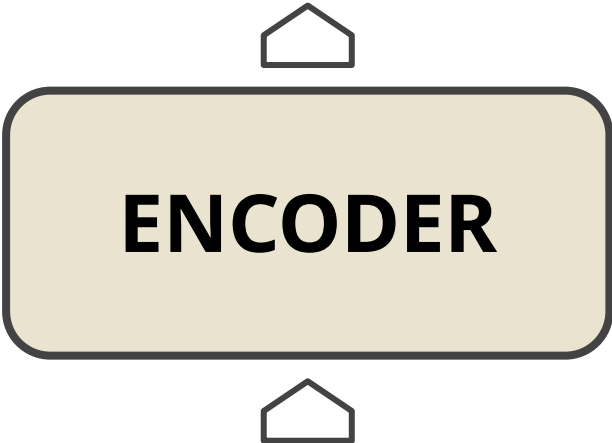
Modelling meaning for intellectual history

- Model and relate meaning of text passages
- **Non-verbatim and cross-lingual reuse:** cannot rely on string matching, need to model the meaning behind the text
- Paraphrase similarity - must focus on the meaning

Meaning is a vector

The two kinds of LLMs

(1.2, -0.7, 0.3, 1.5, ...)



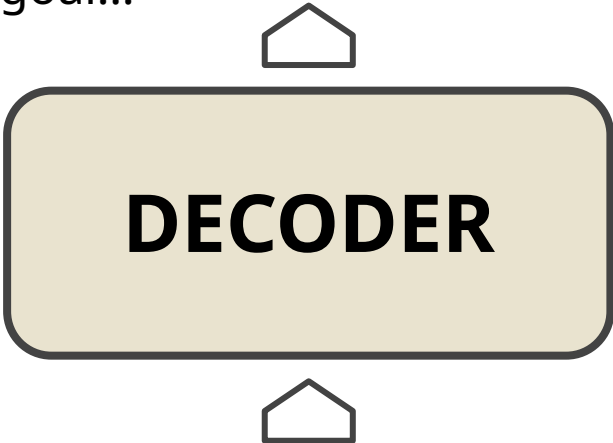
ENCODER

The diagram shows a light beige rounded rectangle labeled 'ENCODER'. Above it is a small house-shaped icon, and below it is another small house-shaped icon. The text '(1.2, -0.7, 0.3, 1.5, ...)' is positioned above the top icon, and 'Tell me how people landed on the Moon.' is positioned below the bottom icon.

Tell me how people landed on the Moon.

**Encodes text into a vector.
This allows texts to be
compared for similarity.**

In 1962, President Kennedy
inspired the nation by setting the
goal...



DECODER

The diagram shows a light beige rounded rectangle labeled 'DECODER'. Above it is a small house-shaped icon, and below it is another small house-shaped icon. The text 'In 1962, President Kennedy inspired the nation by setting the goal...' is positioned above the top icon, and 'Tell me how people landed on the Moon.' is positioned below the bottom icon.

Tell me how people landed on the Moon.

**Generates text based on a
prompt.**



TURKUNLP
.ORG



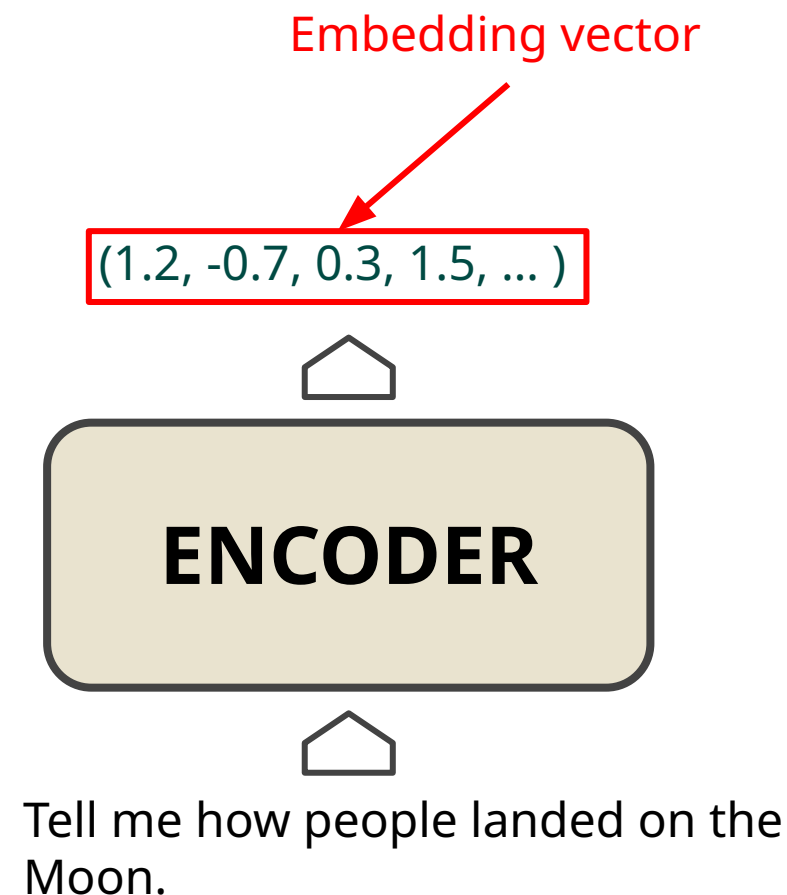
**UNIVERSITY
OF TURKU**

Encoders produce embeddings

Similar meaning $\Leftrightarrow$ Similar embedding

To compute similarity of text1 to text2:

1. Encode the texts into embeddings
2. Use e.g. the cosine similarity function, giving a score $[-1.0, 1.0]$



Encoders produce embeddings

$\text{sim}(\text{'I like cats'}, \text{'I like dogs'}) > \text{sim}(\text{'I like cats'}, \text{'I like hotdogs'})$

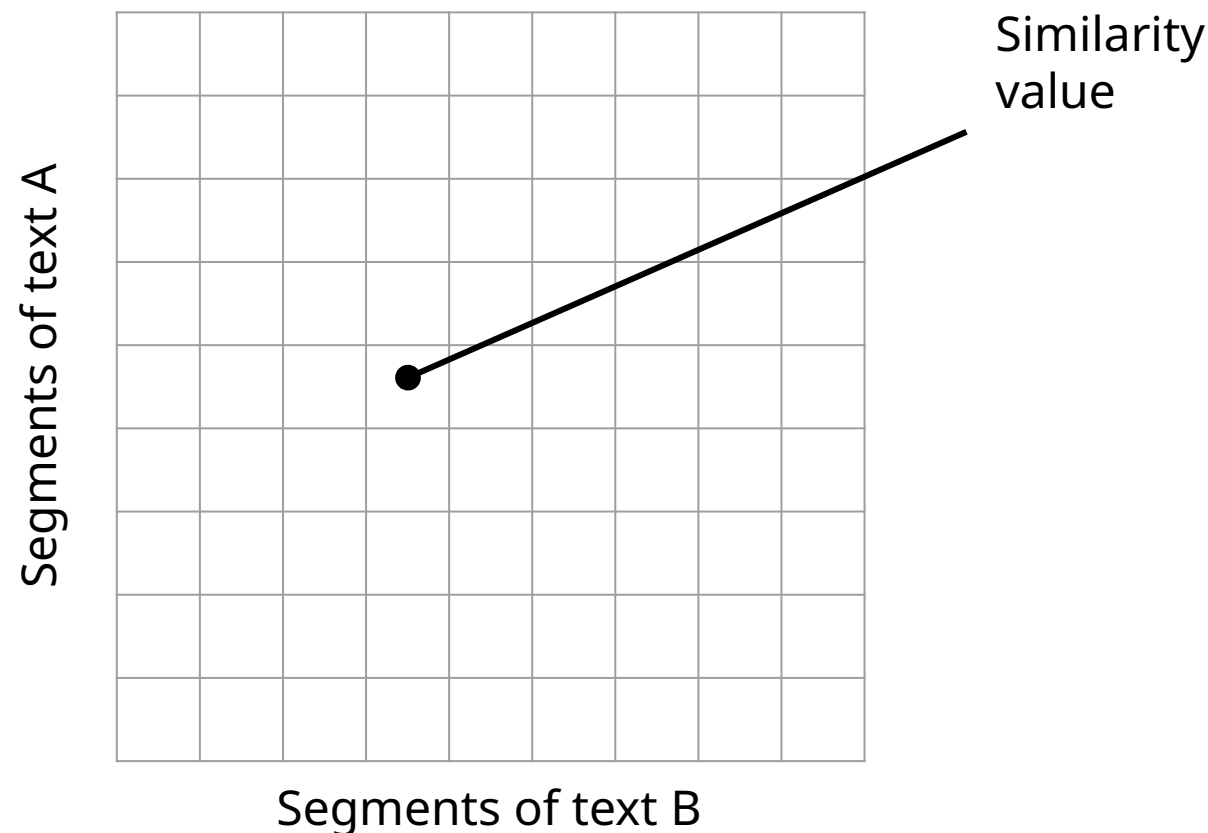
- Only makes sense for quite short texts
- Gives a number which is not properly calibrated
 $\text{sim}(\text{'I like cats'}, \text{'I like dogs'}) = 0.823$
- The value does not mean a whole lot
- Relative comparison of two values more meaningful

Pairwise comparison with sliding embeddings

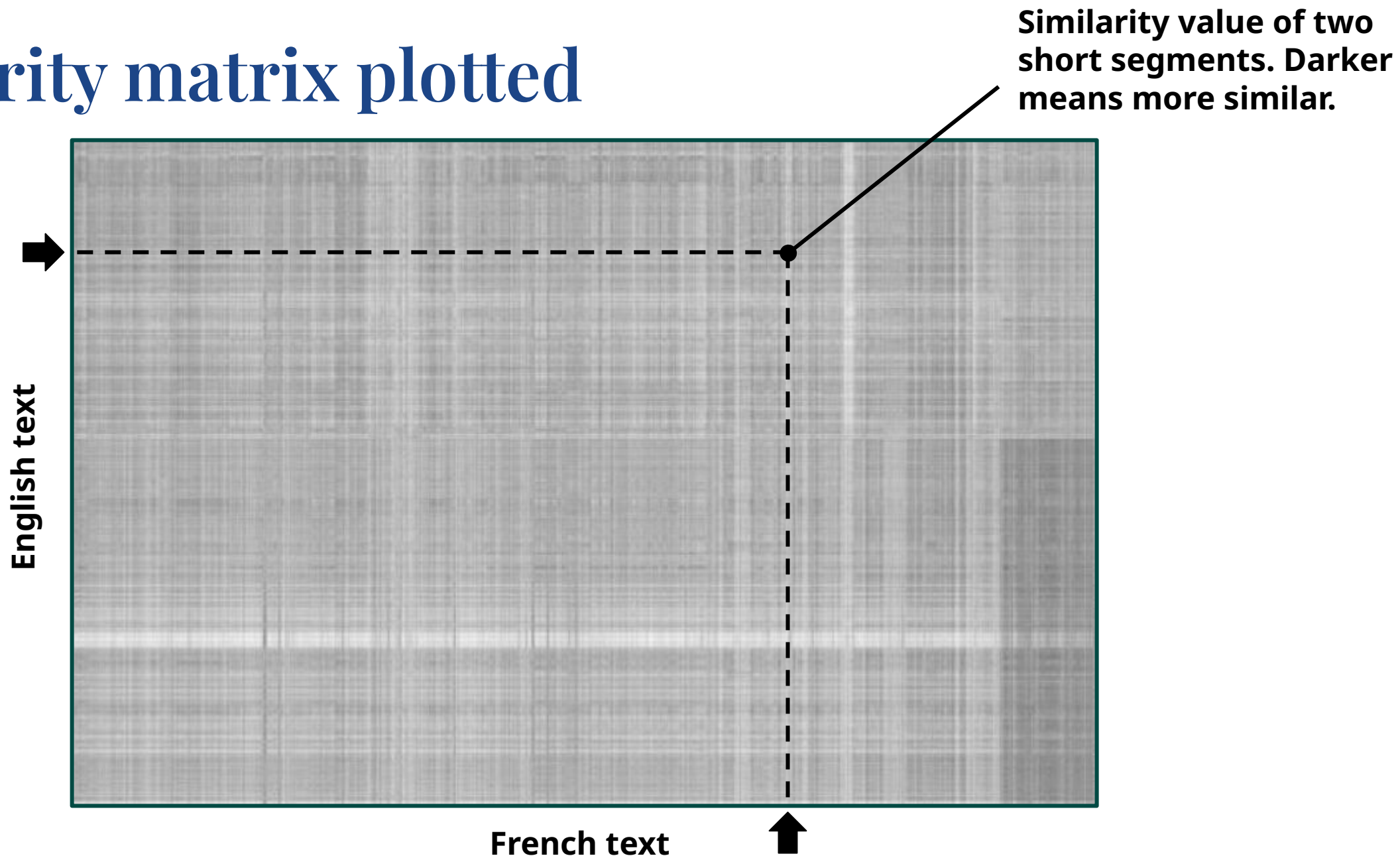
How about long texts then?

To compare two books:

1. Split each text into small segments
2. Compare all against all
3. Store the similarity values in a matrix

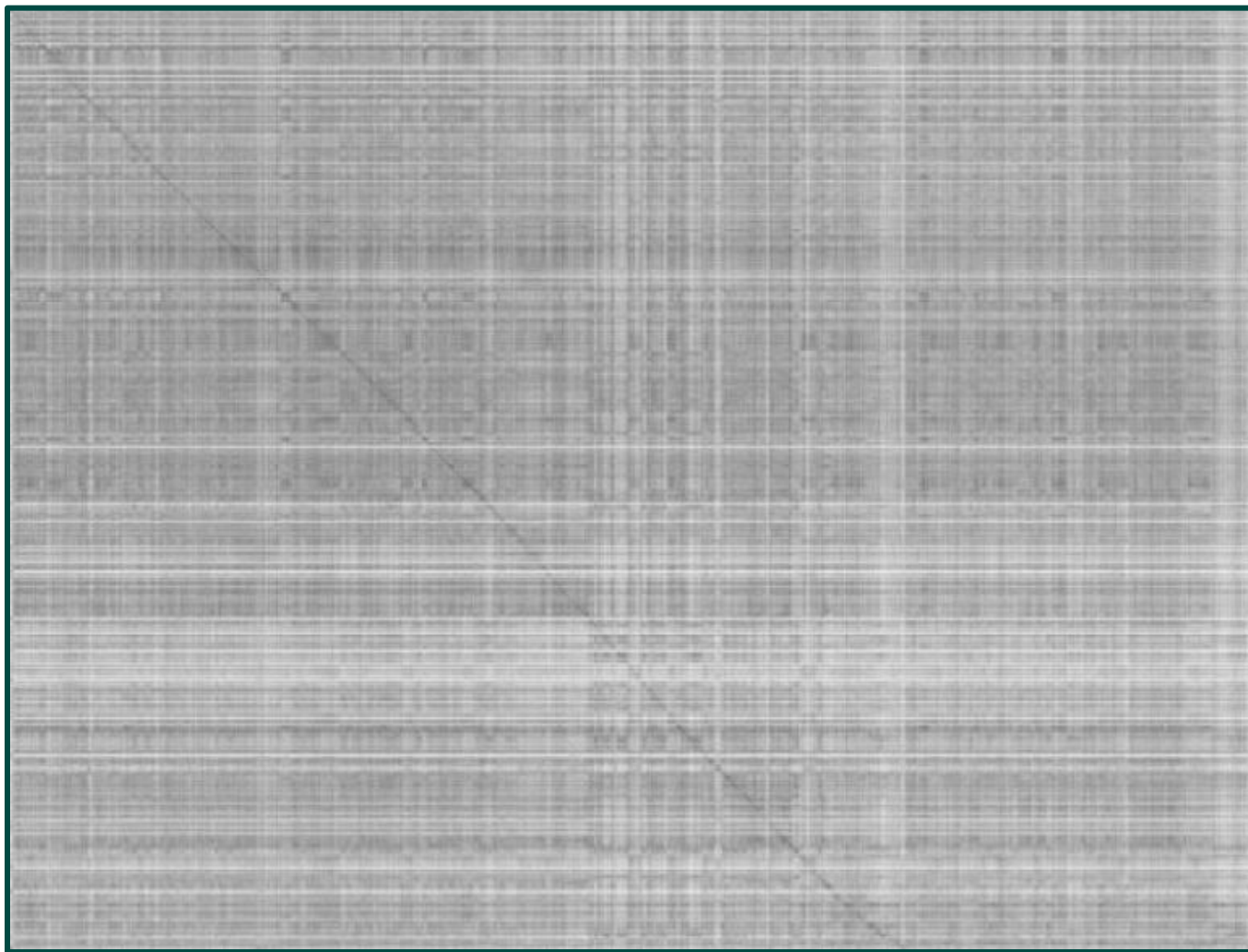


Similarity matrix plotted

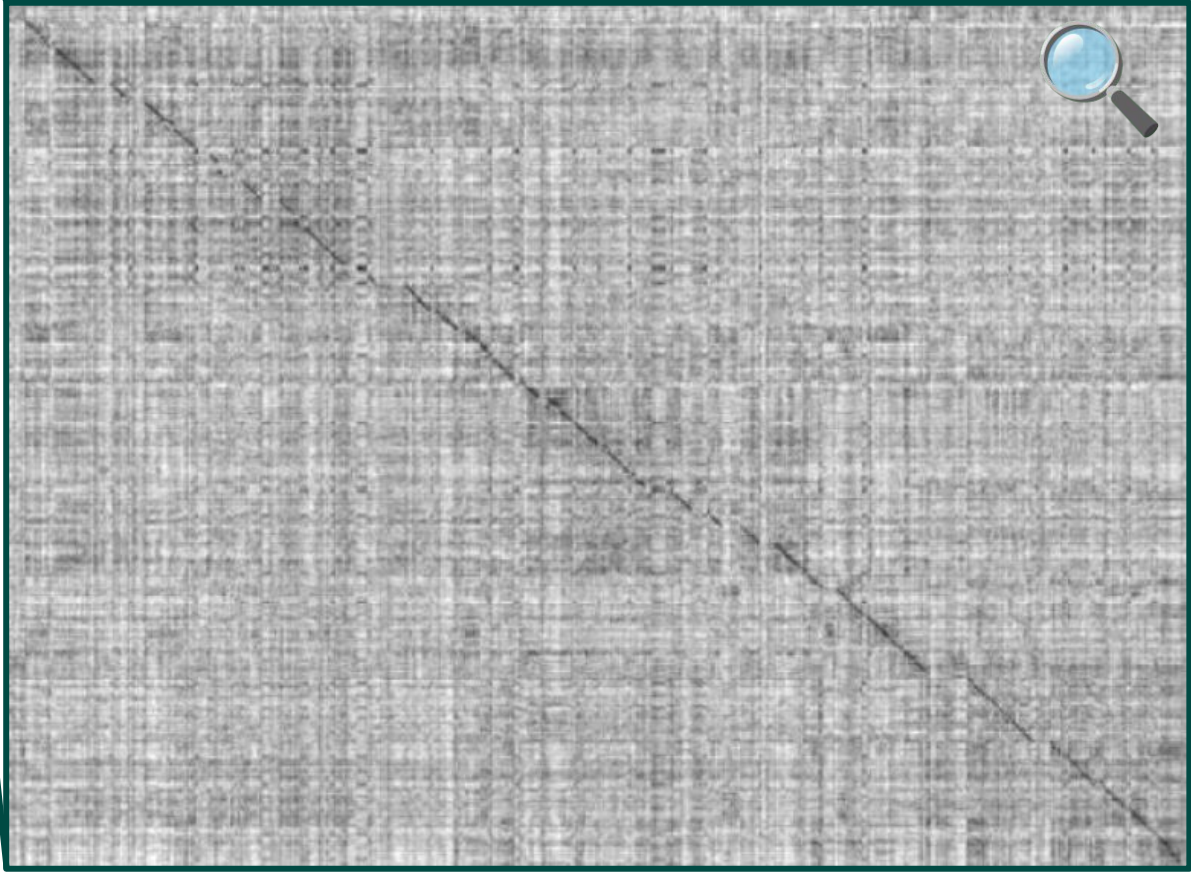
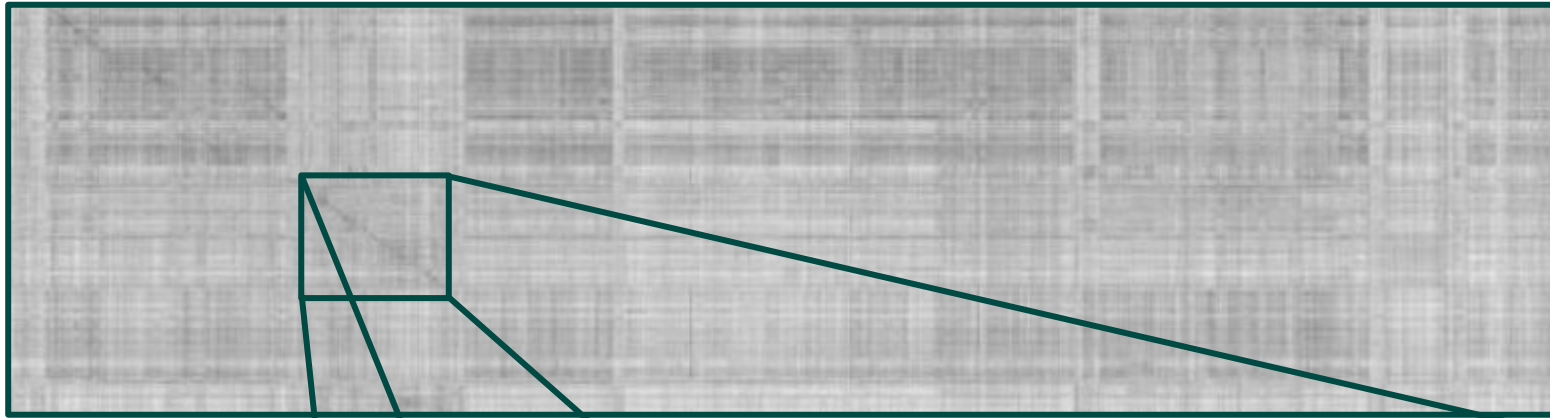


When would this happen?

English text



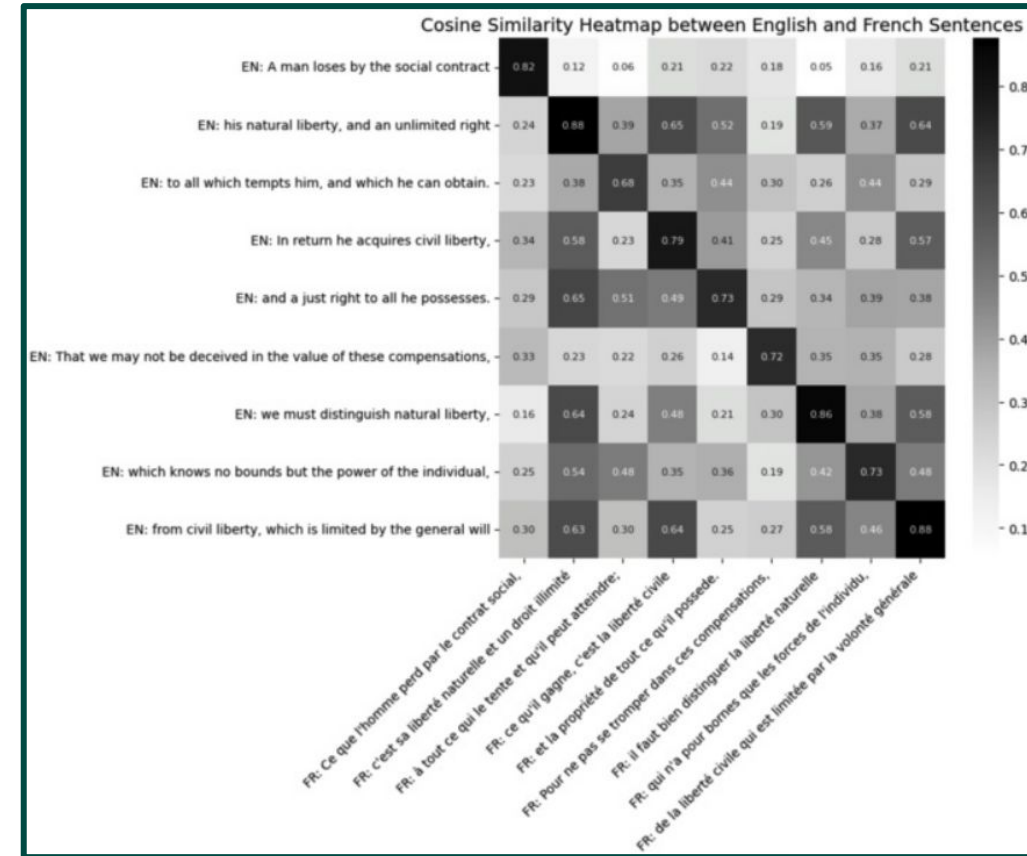
French text



A telltale diagonal trace

- A diagonal line in the image means that a sequence of segments are more similar than on average
- Can also be paraphrased as “say similar things in the same order”
- Texts in two languages + model reacting to semantic similarity
=>

we are likely seeing some form of translation



Scale up!

- Given two texts, we can identify potentially translated passages
- But where do we get these two texts from?
- Confirming known translations => uninteresting
- Finding something we did not know => a lot more interesting
- But then we should look in plenty of material!

Scale up!

Compare text pairs

ECCO: 206K English works

Gallica: 287K French works

59,122,000,000 pairs of works

Compare segment pairs

ECCO: 266,000,000 text segments

Gallica: 390,000,000 text segments

10^{17} pairs of individual segments

ECCO: High-coverage dataset of works printed in England 1700-1799

Gallica: French national library's digitized work collection covering 17th - early 20th century

Both result from a scan-and-OCR process - noise!

The best kind of crossdisciplinary collaboration

- 10^{17} comparisons ... how (on earth)?
- Detection of diagonals ... how?
- The similarities are not calibrated ... what now?
- Characterization of the different patterns ... how?

Exciting methodological challenges! Count me in!

The tools

- **Approximate K nearest neighbor search, FAISS library**
 - Retrieves nearest neighbors based on cosine similarity
 - Average constant retrieval time w.r.t. database size
 - An important component of modern web search => good libraries
- **Progressive Hough transform**
 - Fast detection of lines in images
 - Classic computer vision technique
- **Finnish national High Performance Computing infrastructure**
 - Cluster computing enables parallelization
 - EuroHPC allows access in EU

The solution

- 1. For every segment in the English data, retrieve the 100 nearest neighbors in the French data**
 - Does the 10^{17} comparisons implicitly
 - A segment in text *B* being on the top-100 list of a segment from text *A* constitutes a *hit* between *A* and *B*
- 2. Aggregate across texts and look for pairs of English-French texts which have some minimum number of hits**
- 3. Only if the hits seem to be somewhat aligned on a line, calculate the full similarity matrix and extract the alignment**
 - This is an expensive operation

1 Text segmentation,
segments embedded
as vectors

Text



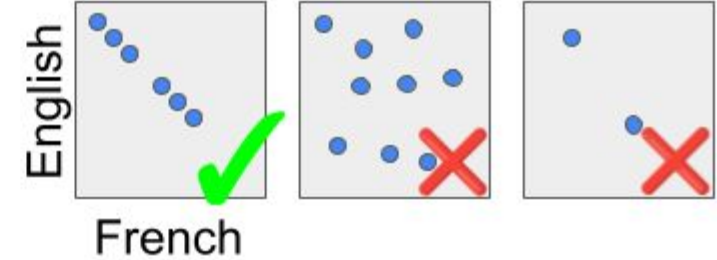
Embed
dings

2 Embeddings
indexed for
efficient retrieval

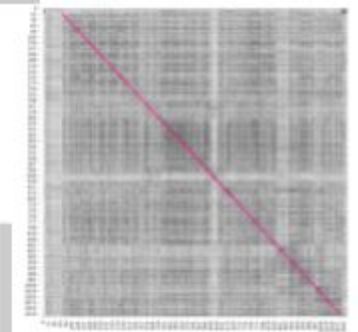


Vector
database

3 Fast but coarse filter based on vector
database retrieval of nearest neighbors



4 Full similarity evaluated for all
remaining book pairs, line
detection to identify candidate
translation segments



Scores:
1. Similarity strength
2. Line-to-noise ratio

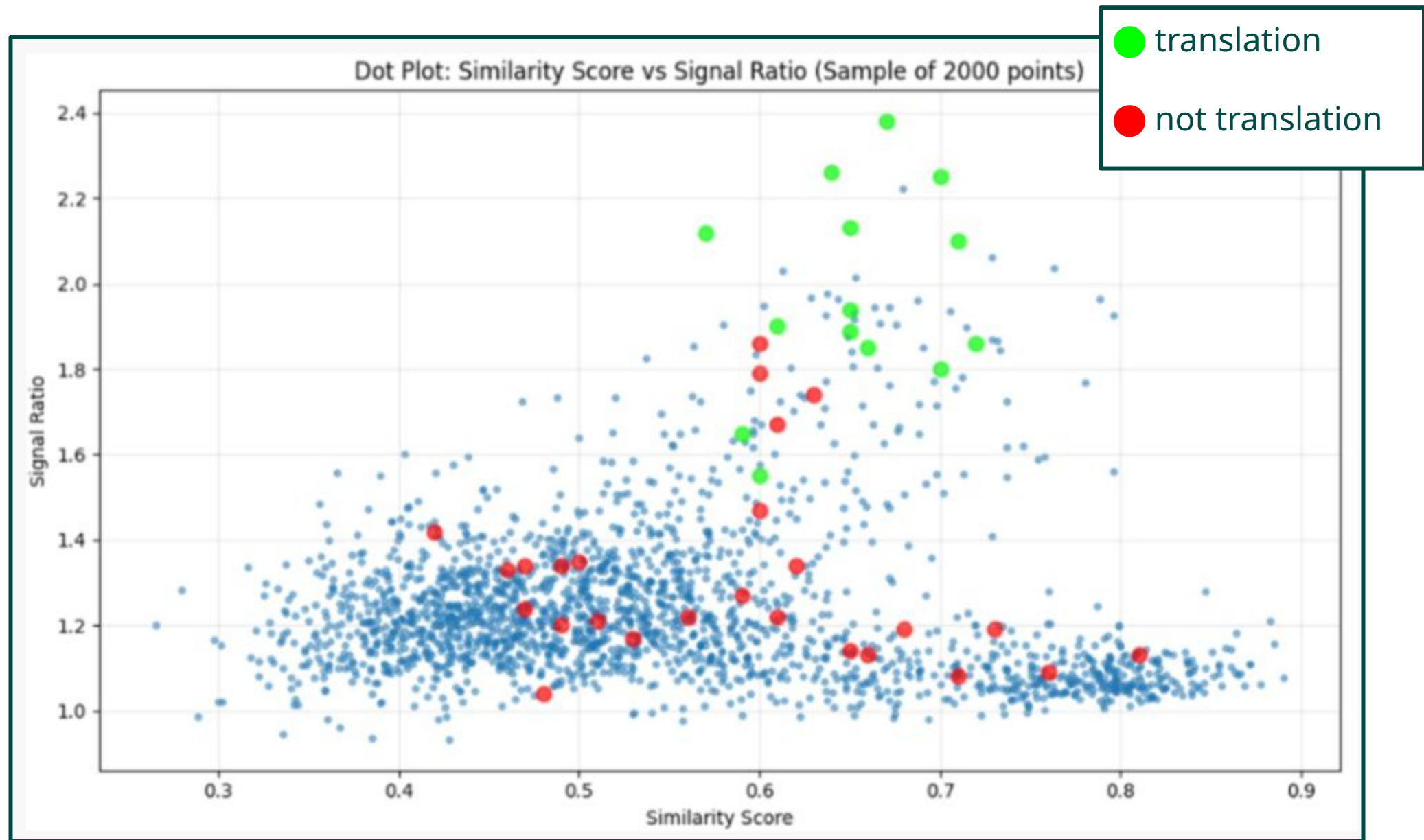
5 Extraction of
scores for
thresholding



Full output: 3M passages in 1M pairs
of works

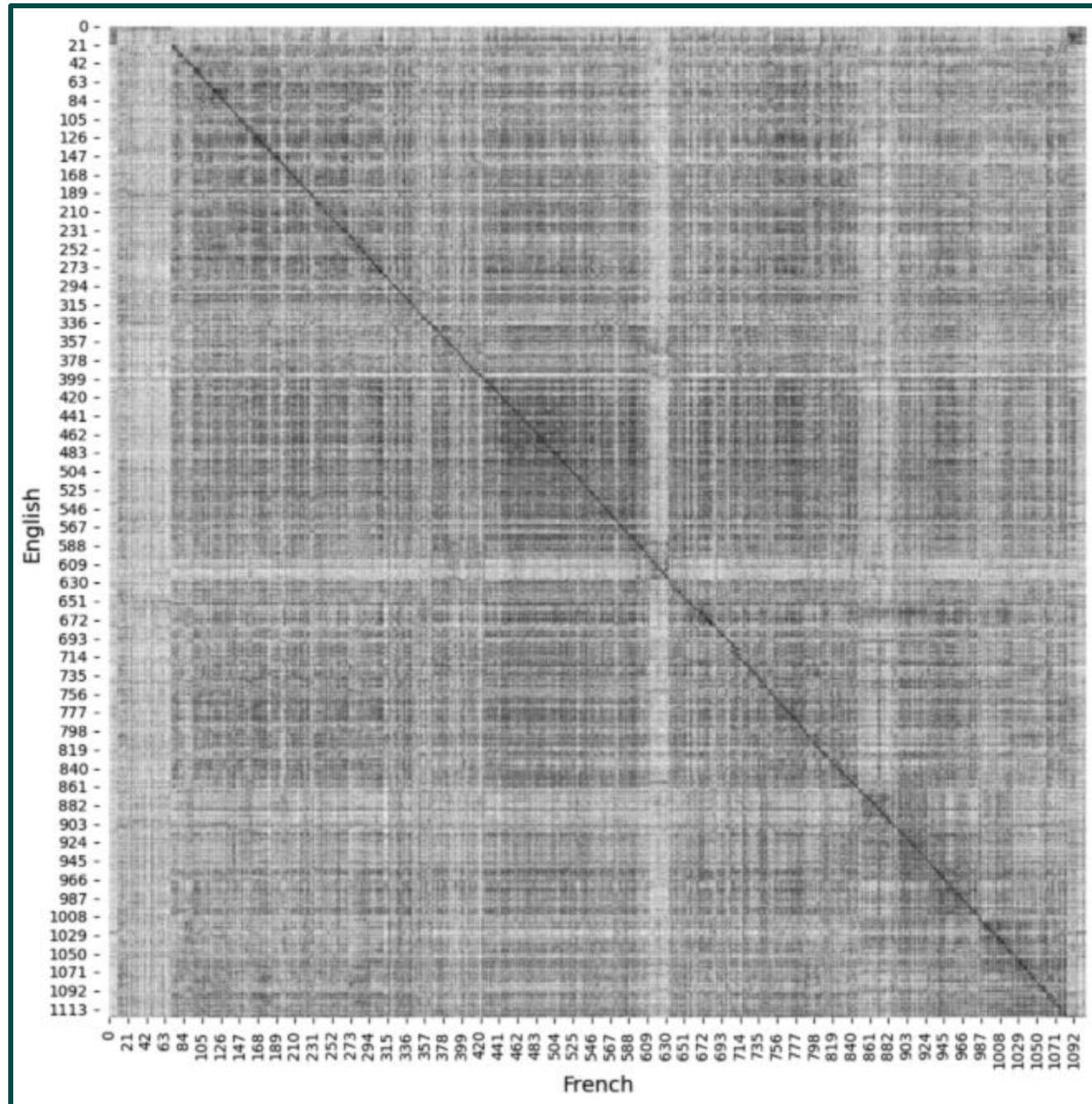
Loose filter: 720K passages in 400K
pairs of works

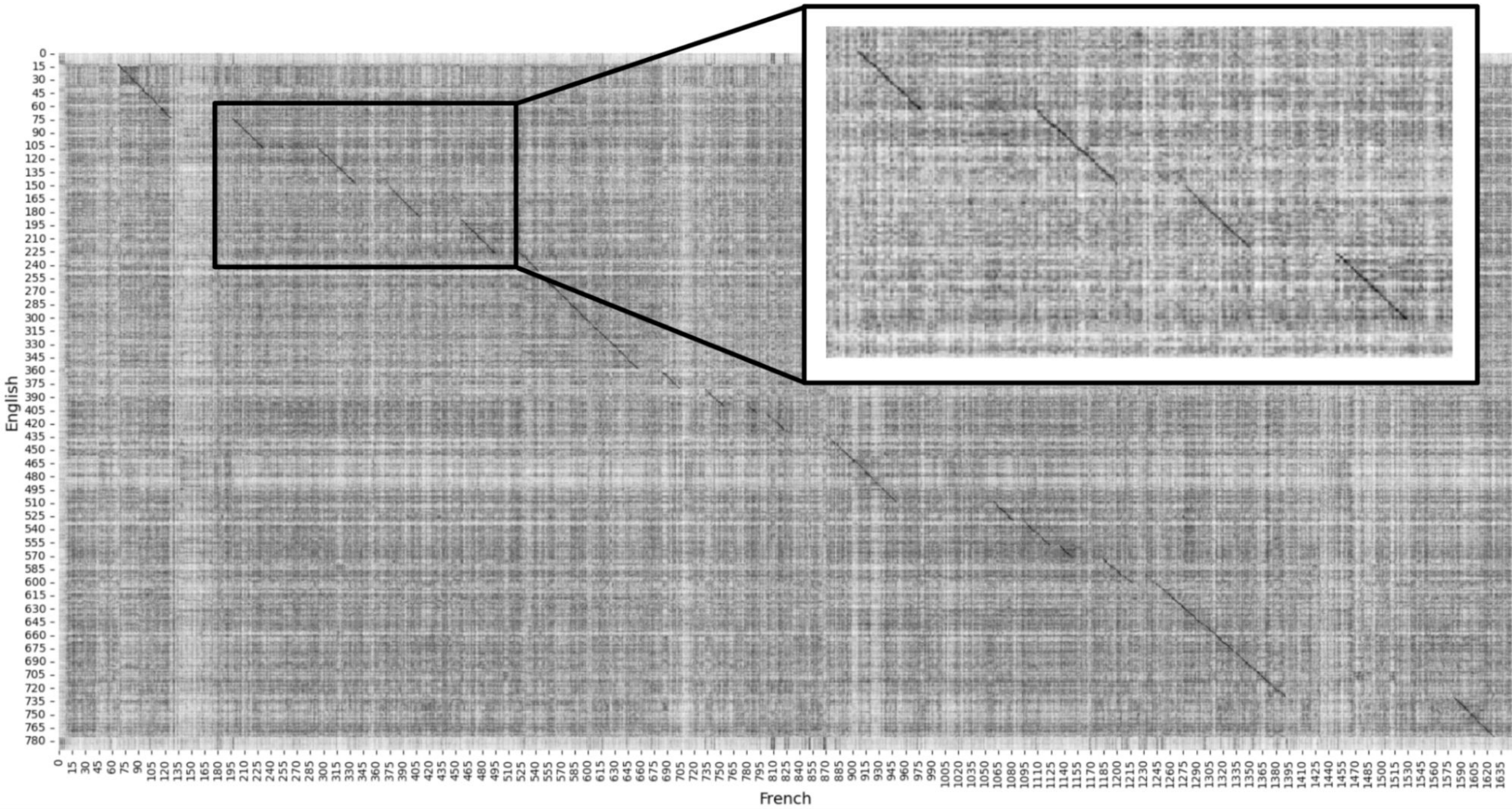
Tight filter: 200K passages in 130K
pairs of works



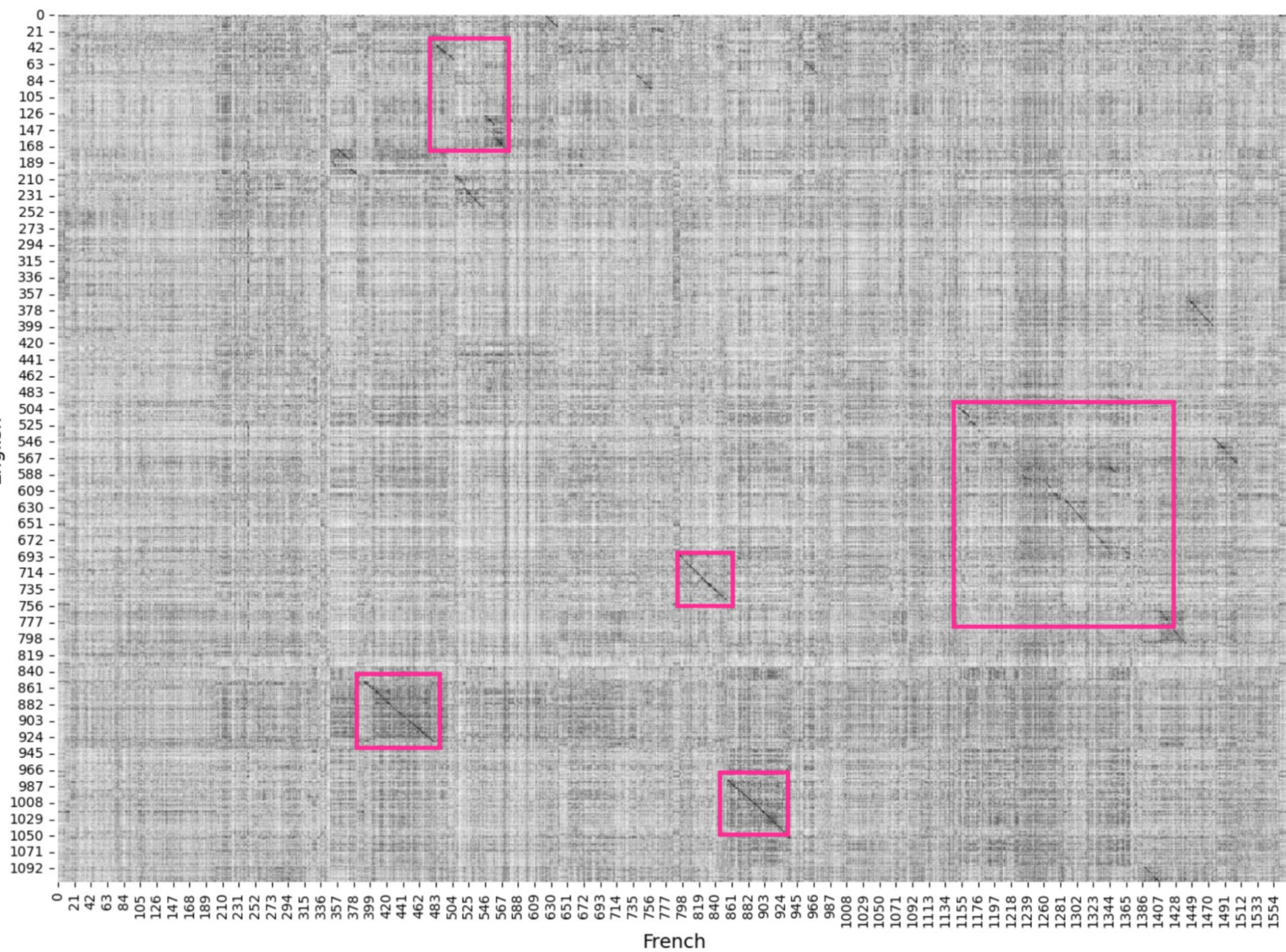
Signal-to-noise ratio: measure of how well the diagonal stands out from its surroundings

Similarity score: the average similarity value along the diagonal





English



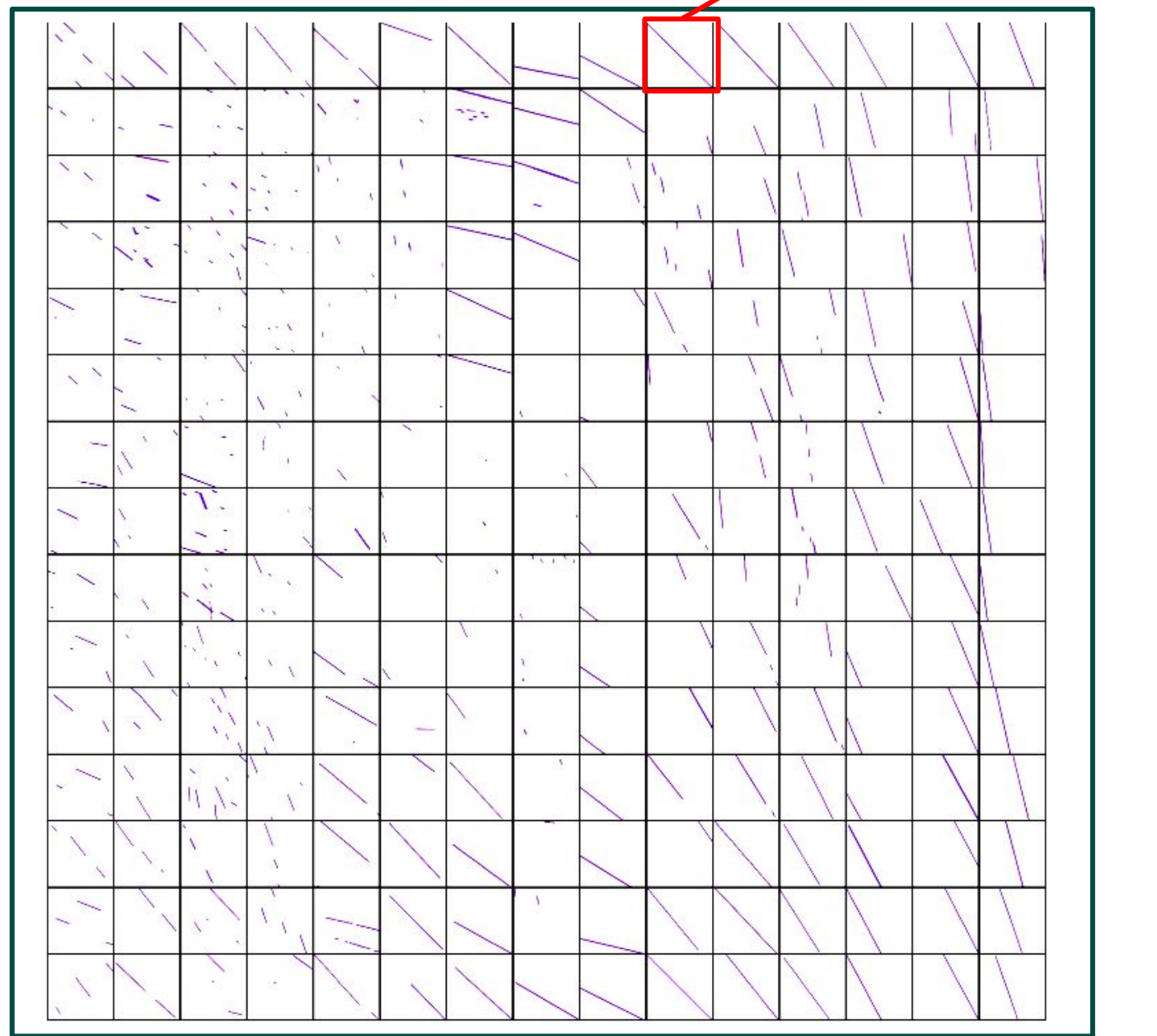
The big picture

- **What is a translation** becomes an interesting question in itself
- We can now revisit it in a data-driven manner
 - Discover different types of translations
 - Analyze their prevalence and historical context
- Types of translations correspond to distinct patterns in which the diagonals are arranged

The big picture

- Plot the identified diagonals
- Encode with a vision encoder
- Cluster the resulting image embeddings
- Build a classifier

Note: all matrices in the figure forced to square shape, which leads to diagonals under very different angles => plotting artefact, not actual



Translation mining mini-recap

- **A whole new kind of research in computational history:**
Compare centuries of literature all-vs-all on a fine resolution allowing even small, hidden reuse to be detected
 - This is a quadratic problem, people could never do this
- Hundreds of thousands of segments can be found across large bodies of literature when the process is scaled up on supercomputers
- The very question of **what constitutes a translation can be asked and answered in a data-driven manner**

On modern encoders

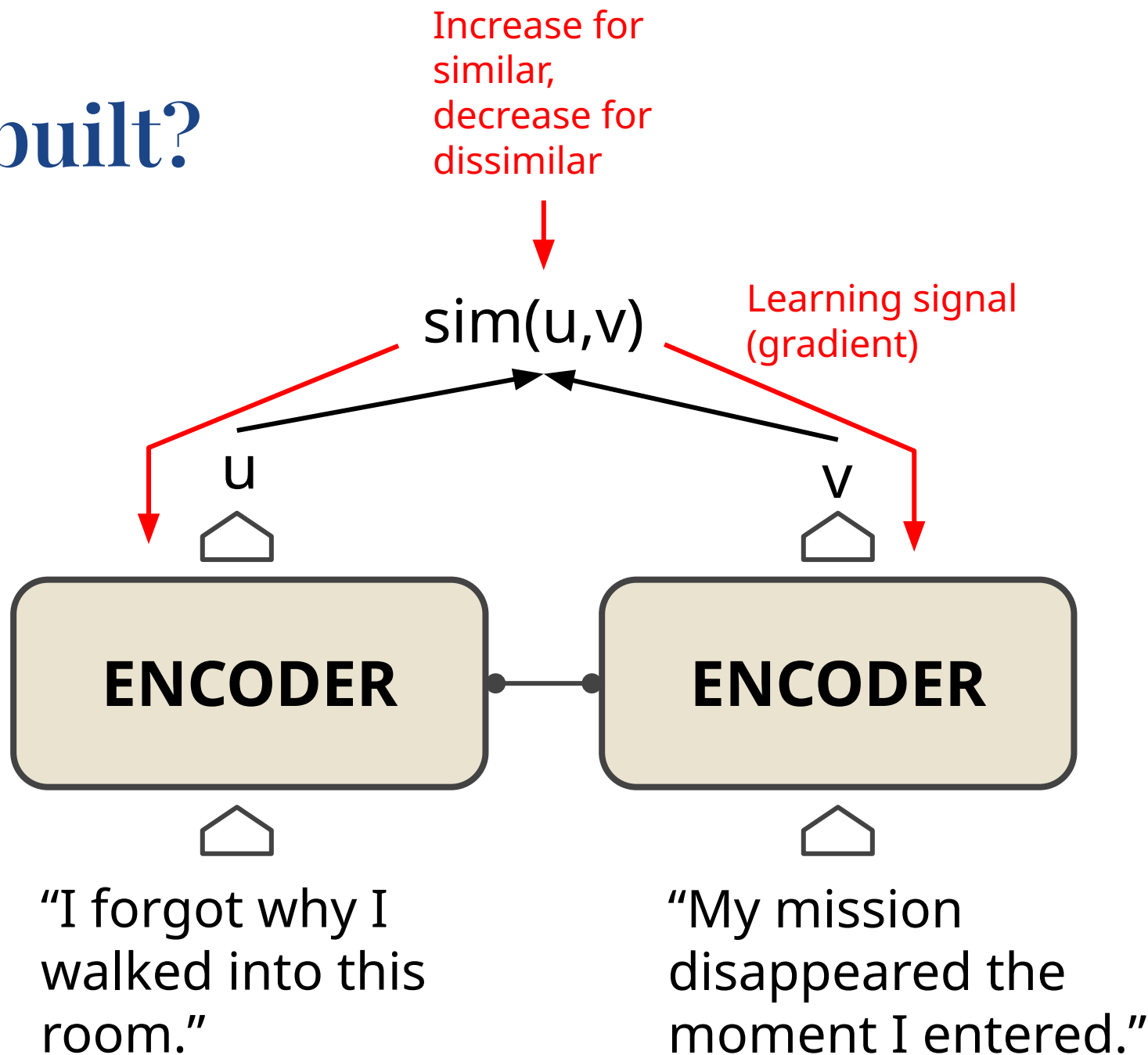
- Many available pre-trained
- Trained for “general” semantic similarity
- Many cross-lingual
- **But how do the encoders know how to encode?**

MTEB leaderboard

Rank (Bor...	Model	Zero-shot	Active Pa
1	harrier-oss-v1-27b	78%	25.6
2	KaLM-Embedding-Gemma3-12B-2511	73%	10.8
3	llama-embed-nemotron-8b	99%	7.0
4	Qwen3-Embedding-8B	99%	6.9
5	gemini-embedding-001	99%	
6	Qwen3-Embedding-4B	99%	3.6
7	Octen-Embedding-8B	99%	6.9
8	F2LLM-v2-14B	88%	13.2
9	F2LLM-v2-8B	88%	6.9
10	harrier-oss-v1-0.6b	78%	0.440
11	Seed1.6-embedding-1215	89%	

How is the encoder built?

- **Construct a dataset of pairs of texts which are similar**
- Random pairs can be assumed as dissimilar (on average)
- Train the encoder to produce embeddings which:
 - **Maximize similarity for known-similar texts**
 - **Minimize similarity for known/assumed-dissimilar texts**



But how do we know what is similar?

- Much better: **We get to decide!**
- The encoder learns the similarity from the training data - and we make that data

Translation pairs => cross-lingual encoder

Question-answer pairs => encoder useful for web search

Your pairs => encoder that serves your purpose

Present-day encoders

- Typically trained on a union of datasets with many varying tasks
 - Translation, paraphrase, entailment, question-answer, summary...
- React to some kind of “average notion of similarity”
- Prompt appended to the query text, controls which task the embedding should suit
 - Not entirely successful at that
 - Work in progress in NLP

The future opportunity

- Take more explicit control of the relation which the encoder implements
- Would enable us to model text engagement more broadly, rather than only translations and paraphrases
 - E.g. making a passage similar to a passage which reacts to it negatively
- It would be even better if we could reliably describe the desired relation as a prompt and have the encoder implement it on the fly

In summary

- Collaboration between NLP and digital history has lots to offer to both sides
 - Ability to trace connections among authors at a very large scale, opening up the “dark mass” of long-forgotten literature
 - Demands on embedding models spark serious methodological research - right now, we ask more than the models can provide

Large scale + multiple old languages + OCR noise is an exciting cocktail!

Thanks!

HISTORICAL METHODS: A JOURNAL OF QUANTITATIVE AND INTERDISCIPLINARY HISTORY
<https://doi.org/10.1080/01615440.2026.2675558>



OPEN ACCESS

Translation mining: An AI-driven taxonomy of eighteenth-century Anglo-French translation practices

Kira Hinderks^{a*}, Cassandra Ledins^{b*}, Filip Ginter^{b#} and Mikko Tolonen^a

^aHelsinki Computational History Group, Department of Digital Humanities, University of Helsinki, Helsinki, Finland; ^bTurkuNLP, Department of Computing, University of Turku, Turku, Finland

Available under open access.

Special thanks to co-authors: Kira Hinderks, Cassandra Ledins, Jenna Kanerva, Yu Wu, Anna Plassart and Mikko Tolonen

The Algorithmic Challenges and Opportunities in Tracing Meaning at Scale in Large Historical Document Collections

Filip Ginter
0000-0002-5484-6103
TurkuNLP
Department of Computing
University of Turku
Email: figint@utu.fi

Jenna Kanerva
0000-0003-4580-5366
TurkuNLP
Department of Computing
University of Turku
Email: jmnybl@utu.fi

Mikko Tolonen
0000-0003-2892-8911
COMHIS
Department of Digital Humanities
University of Helsinki
Email: mikko.tolonen@helsinki.fi

Kira Hinderks
0009-0004-9233-9315
COMHIS
Department of Digital Humanities
University of Helsinki
Email: kira.hinderks@helsinki.fi

Anna Plassart
0000-0002-3441-2332
History department
The Open University
Email: anna.plassart@open.ac.uk

Yu Wu
0009-0006-1634-5566
COMHIS
Department of Digital Humanities
University of Helsinki
Email: yu.wu@helsinki.fi



FedCSIS'26 proceedings



Get inspired.



UNIVERSITY
OF TURKU