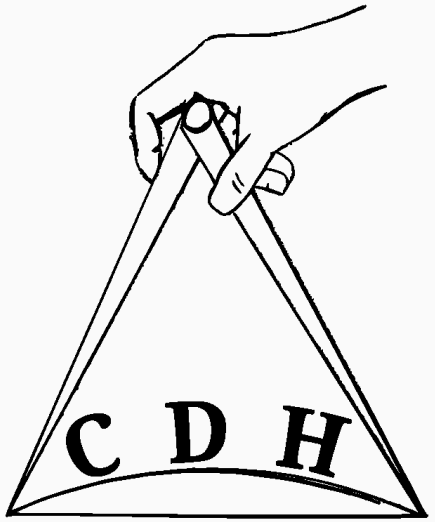
What does Computational Humanities *look like*?

Tom Lippincott

Center for Digital Humanities
Department of Computer Science



This talk

Quickly outline recent publications

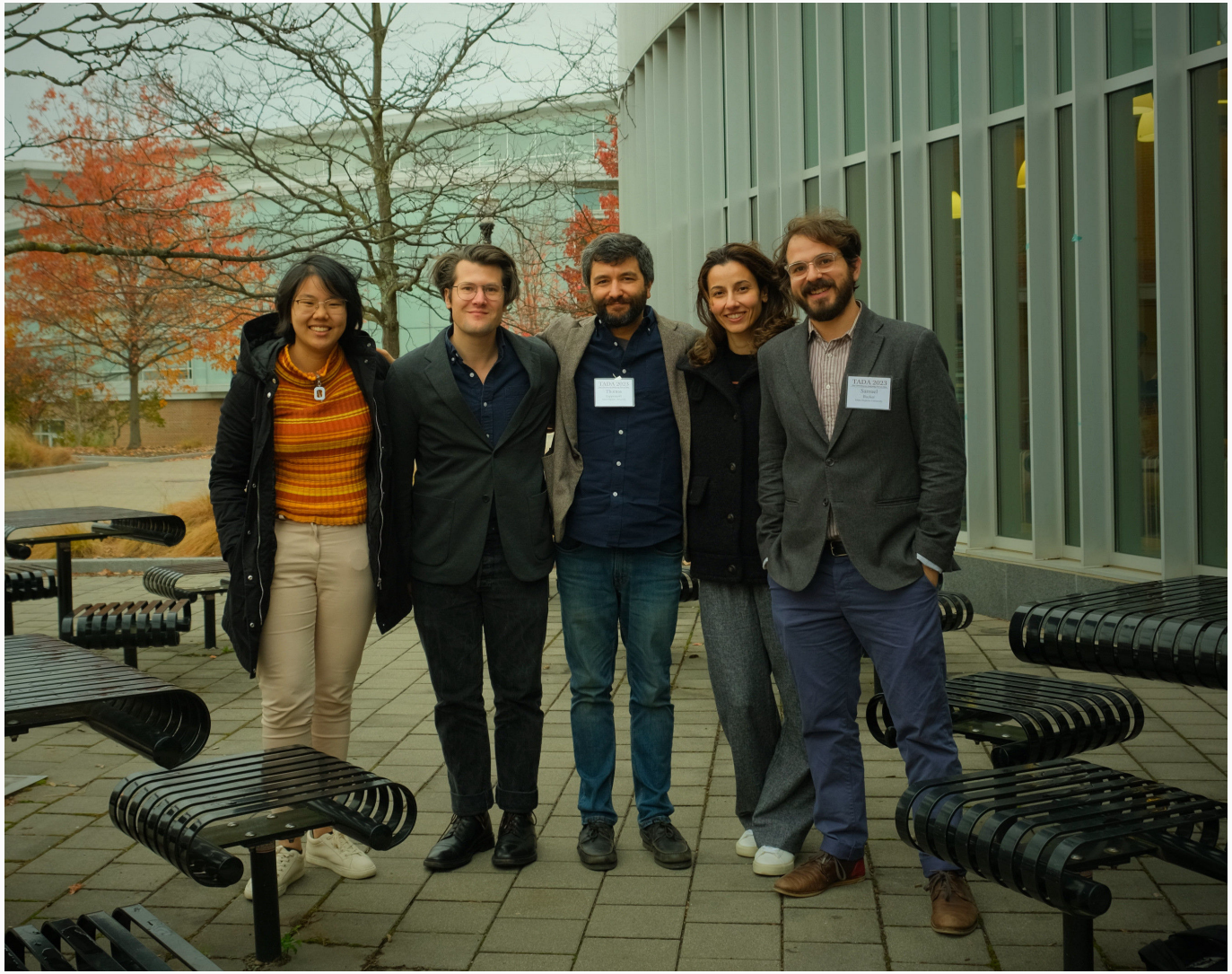
- "Nearest-neighbor" matching
- Some technical details in footnotes
- In-depth conversations . . .

Basic research pattern

- Identify something of humanistic importance
- Translate into minimally-supervised ML method
- Interpret output as dense, aggregated document

Such research ***fundamentally*** depends on cross-training

CDH researchers



Sabrina Li, Craig Messner, me, Hale Sirin, Sam Backer

Shifts of semantic modality in Latin

Foundational scholarship from comparative literature:

- Certain (sets of) words underwent a sequence of shifts
- Particular authors drove such change

Train and inspect a temporal topic model[1]

- Measure *novelty* of authors w.r.t. preceding time-window[2]
- Measure *bi-modality shift* of each word[3]
- Aggregate these measurements in various ways

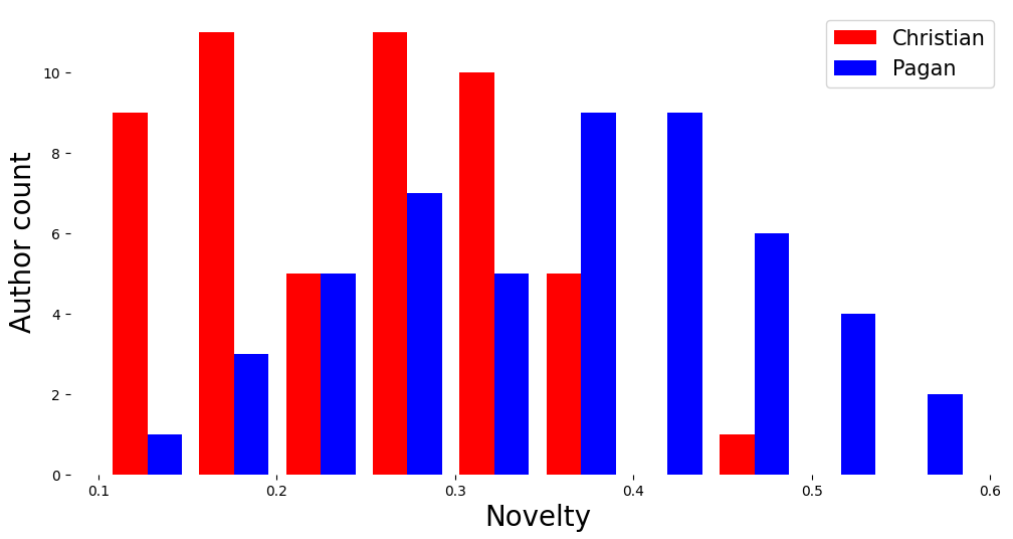
---

[1] Dynamic embedded topic model

[2] Jensen-Shannon divergence

[3] Function of first and second strongest topics and change-point detection

Shifts of semantic modality in Latin



Word	Year	Delta
cathedra (*chair*)	-175	0.947
cicatrix (*scar*)	425	0.944
conlatio (*bring together*)	350	0.939
auster (*south wind*)	350	0.927
recte (*upright*)	350	0.915

Literary orthographic variation and LLMs

Representations of dialect in fiction

- Differs from variation "in the wild"
- Aesthetic, culture, narrative . . .
- Non-linguistic regularities

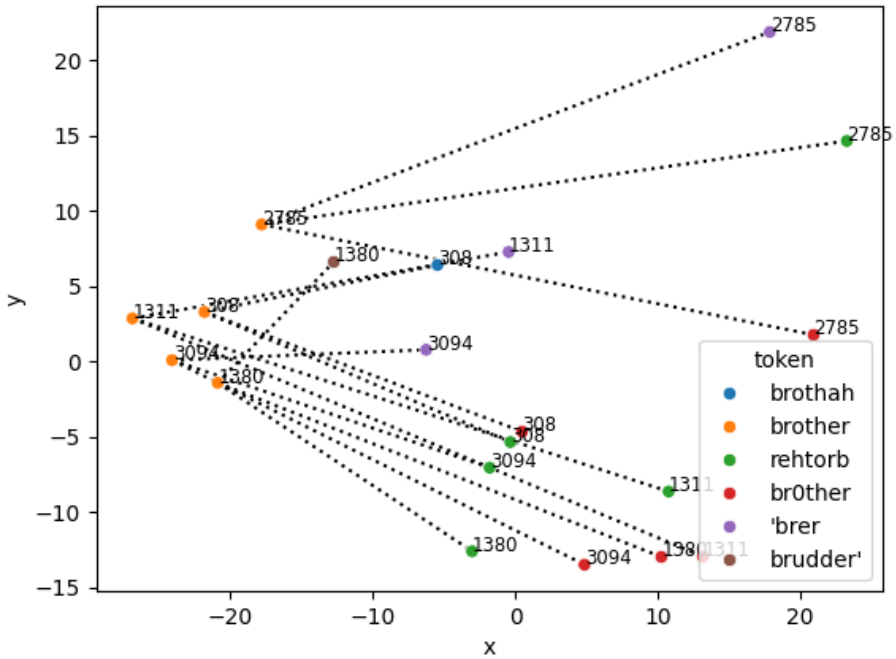
Apply models[4] to variety of orthographic realizations

- Build corpus of annotated variants
- Generate perturbations
- Inspect relative positions in semantic space

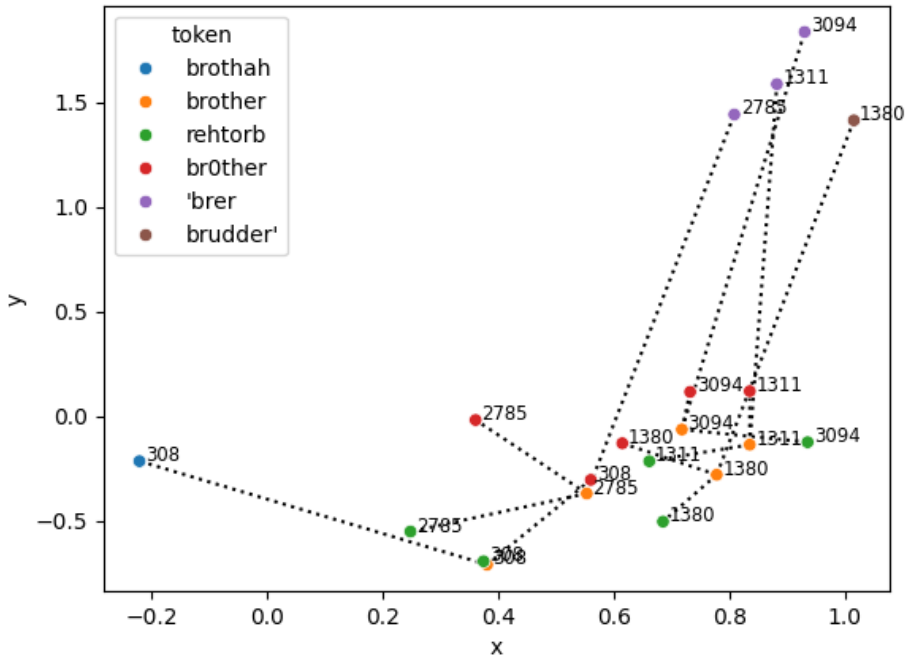
[4] BERT and CANINE

Literary orthographic variation and LLMs

"Brother" and variants



Embedded using BERT (subwords)



Embedded using CANINE (characters)

Armeno-Turkish, large corpora, and language alternations

Historical population with invisible record

- Turkish in Armenian script, not well-catalogued
- Known to be present in library collections
- Types of multi-lingual documents "oscillate" between languages

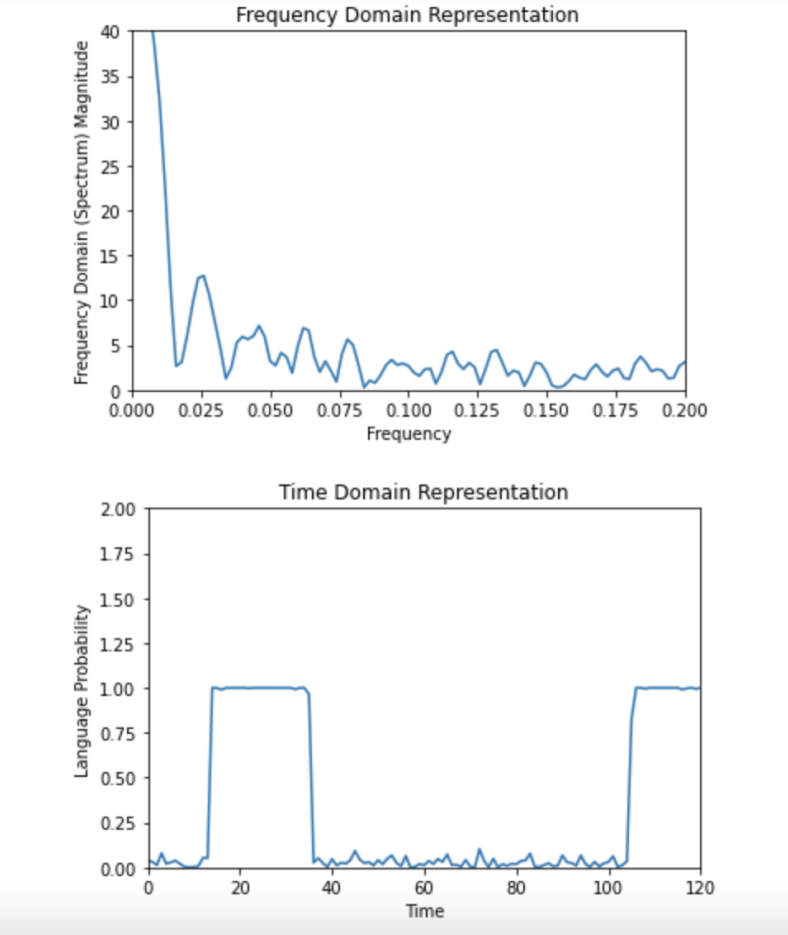
Bootstrap language ID model[5], apply to HathiTrust

- Assemble and train on a (noisy) corpus
- Find more Armeno-Turkish documents
- Characterize **periodicity** of language alternation[6]

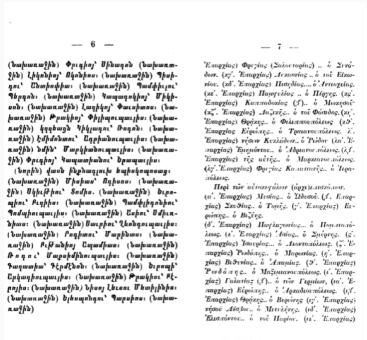
[5] FastText trained from scratch

[6] Fourier analysis to frequency domain

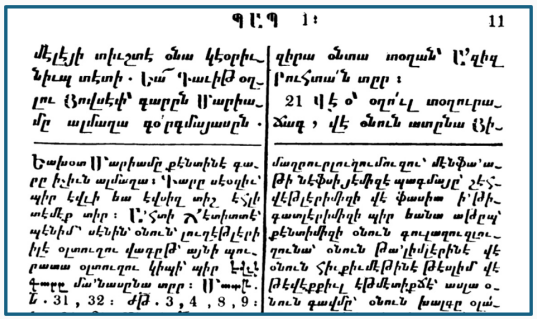
Armeno-Turkish, large corpora, and language alternations



Time series and frequency



թուրքճէ	ինկլիզճէ	թէլէֆֆիւզ
Սիզէ տէմէյէ կէլ-տիմ։	I have come to tell you.	Այ հէ՛վ քըմ թու թէլ եու։
Ո խուսուստա չօգ տիւշիւնմէմ։	I don't think much of it.	Այ սո՛նք թհինք մըշ ավ իք։
Տէտիյի իլէ իք-տիյի պիր օլտու։	No sooner said than done.	Նօ ոուՆըր սէտ ատէ՛ն տըն։
Ակթըզ տայանամամ	I can bear it no longer.	Այ քէ՛ն պէէր իք Նօ լօ՛նկըր։
Եափապիլտիյիմ գա-տար։	As much as I can.	Է՛զ մըշ է՛զ այքէ՛ն։



Exemplars from three clusters

Publications

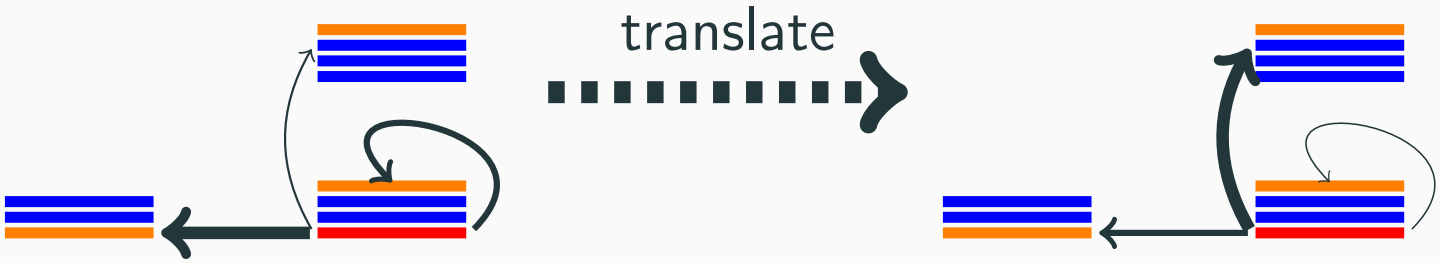
Craig Messner and Tom Lippincott. 2024.
Pairing orthographically variant literary words to standard equivalents using neural edit distance models.
In *LaTeCH@EACL*.

Hale Sirin, Sabrina Li, and Tom Lippincott. 2024.
Detecting structured language alternations in historical documents by combining language identification with Fourier analysis.
In *LaTeCH@EACL*.

Hale Sirin and Tom Lippincott. 2024.
Dynamic embedded topic models and change-point detection for exploring literary-historical hypotheses.
In *LaTeCH@EACL*.

Nascent research

Preserving inter-textual geometry across translations[7]



Cognitive salience of poetic structure[8]

$$\acute{A}rm\breve{a}\ v\breve{i}r|\acute{u}mq\breve{u}e\ c\breve{a}|n\acute{o}\ ||Tr\acute{o}i|\acute{a}e\ q\acute{u}i\ |pr\acute{i}m\breve{u}s\ \breve{a}b\ |\acute{o}r\acute{i}s$$

```
Poem -> ?
? -> Line+
Line -> ?
? -> (Stress|Unstress)+
```

[7] MBART shared embedding space

[8] Hierarchical non-parametric processes