Computational Humanities, by Example

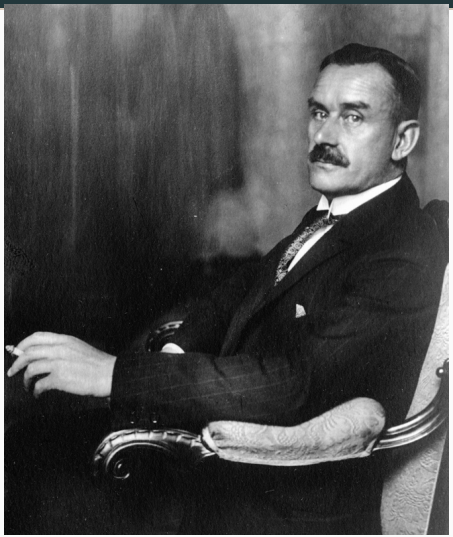
Texts, records, minds, and populations

Tom Lippincott

Computer Science & Center for Digital Humanities, Johns Hopkins University

Framing

The writer as instrument



That the writer . . . is a reporting instrument, seismograph, medium of sensitivity, though lacking clear knowledge of his organic function and therefore quite capable of wrong judgments also—this seems to me the only proper perspective on writing.

— Thomas Mann, *The Story of a Novel*

Cultural artifacts are measurements. What instruments can we build to read them?

Why might computation help humanists?

Scholars reason deeply and narrowly, but...

- inherit centuries of received wisdom
- are often invested in a conclusion
- cannot attend evenly to all the data and all the hypotheses

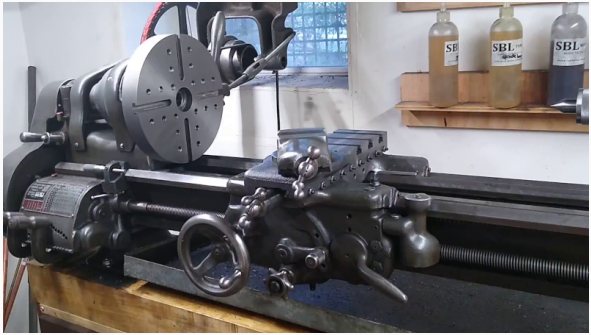
Models ...

- are indifferent to the conclusion
- process data and hypothesis-spaces uniformly
- yield explicit representations and probabilities

Major advantages

Independence from prior beliefs • Assumptions made explicit • Guaranteed connection to observation • Stepping into the same river twice

Not a black box



We are not (mostly) applying off-the-shelf NLP to old texts.

We build instruments: models whose structure, data and training *encode a scholarly hypothesis*, so that their behavior bears on it.

Two roles for AI in humanistic research

AI as research assistant

"I can't transcribe/label/link all this!"

- OCR, transcription, translation
- Entity extraction and normalization
- Agentic collation of archives
- Requires careful task definition and verification

AI as empirical model

"What does the data itself say?"

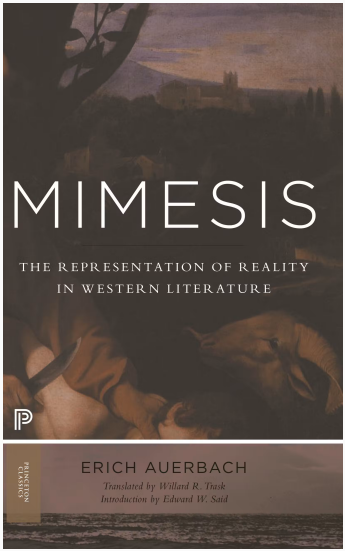
- The model *is* the hypothesis
- Self-supervised, minimal annotation
- Constraints and affordances designed to mirror the domain
- Ablate, compare, and inspect

Five studies

	Study	Motivating monograph	
1	Auerbach's history of style	*Mimesis*	COMPLETE
2	Biblical intertextuality & chiasmus	*The Art of Biblical Narrative*	COMPLETE
3	People in the archive	*Soul by Soul*	COMPLETE IN PROGRESS
4	Hierarchy and memory in poetry	*The Chapter*	IN PROGRESS
5	Simulating Appalachian folkways	*Albion's Seed*	SPECULATIVE

Texts → records → minds → populations

Study 1: Auerbach's history of style



Study 1

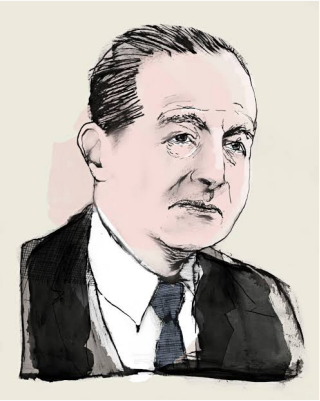
Auerbach's history of style

Erich Auerbach, *Mimesis: The Representation of Reality in Western Literature* (1946)

COMPLETE

Can a model trained on small amounts of historical text recover the stylistic change Auerbach describes, without inheriting his conclusions?

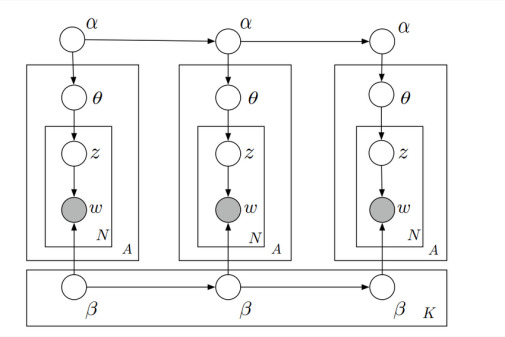
Auerbach's claims about late Latin



- Declining syntactic complexity:
  hypotaxis → parataxis
- Semantic shift, e.g. *figura*:
  abstract → concrete → bi-modal
- Particularly influential authors
  - Lucretius, Varro, Cicero
  - Tertullian, Augustine

Chapter 1, "Odysseus' Scar", sets Homer against Genesis.

Dynamic embedded topic model



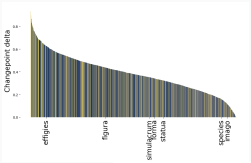
- Topics drift smoothly over time
- Words and topics share an embedding space
- Define measurements over the learned parameters, e.g.

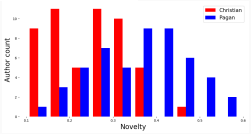
$$\textit{bimodality} = \tfrac{1}{2}\left(\textit{even} + \textit{exhaustive}\right)$$

where a word's top two topics are evenly weighted and together cover most of its mass

[Sirin & Lippincott 2024; Fittschen et al. 2025]

Interpretation: words that become bi-modal



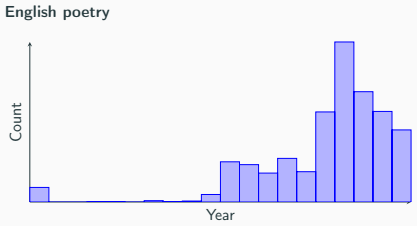


Word	Year	Delta
cicatrix (*scar*)	425	0.944
conlatio (*bring together*)	350	0.939
auster (*south wind*)	350	0.927
recte (*upright*)	350	0.915

Word	Year	Δ
cicatrix (scar)	425	.944
conlatio (bring together)	350	.939
auster (south wind)	350	.927
recte (upright)	350	.915

The model's top word is *scar*: the subject of Auerbach's first chapter, reached without reading him.

Problem: historical data is small and uneven



[Fittschen et al. 2026, EACL]

Alternative: train generative LMs per period

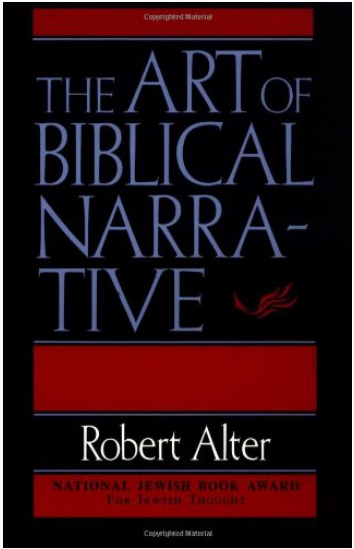
- Arrange materials into temporal buckets
- Train small LMs *de novo* on each
- Compare with a large pretrained LM fine-tuned on the same buckets
- Challenge tasks that should (not) be solvable at a given date

Completion ranking: what should a 1960 reader know?

Sentence	Definition	Year	De novo	Finetune
I'm going to sell my car... No more police towing [it] ..to a car **pound**.	A place in which vehicles impounded by the police or other authorities are kept...	1970	101	0
Hill ... which won three gold and a **silver**.	Elliptical for silver medal n.	1960	7	0

A pretrained model "knows" words before they were coined; a period-trained model does not. Avoiding that leakage is the point.

Study 2: Biblical intertextuality



Study 2

Biblical intertextuality and chiasmus

Robert Alter, *The Art of Biblical Narrative* (1981)

COMPLETE

What do allusion and literary structure look like inside pretrained multilingual models, and does translation preserve them?

Inter-textuality



The best writing of our contemporaries is not an act of creation, but an act of evocation, peculiarly saturated with reminiscences.

— Harry Levin, *James Joyce*

Simple starting point: how do known allusions appear in pretrained embedding spaces?

An allusion across testaments

Exodus 14:21 *Hebrew Bible*	Revelation 16:12 *New Testament*
20 [...] and it was a cloud and darkness to them, but it gave light by night to these: so that the one came not near the other all the night.	11 And blasphemed the God of heaven because of their pains and their sores, and repented not of their deeds.
21 And Moses stretched out his hand and the LORD caused the sea to go back by a strong east wind all that night, and made the sea dry land, and the waters were divided.	12 And the sixth angel poured his vial upon the great river Euphrates; and the water thereof was dried up, that the way of the kings of the east might be prepared.
22 And the children of Israel went into the midst of the sea upon the dry ground [...]	13 And I saw three unclean spirits like frogs come out of the mouth of the dragon [...]

Does translation change intertextuality?

		Within		Across
		Jewish (OT)	Christian (NT)	
Orig.	Ancient Hebrew	0.98± 0.14	–	–
	Ancient Greek	1.27± 0.21	1.30±0.19	1.31± 0.20
Human	English	1.66± 0.21	1.70± 0.30	1.69± 0.27
	Finnish	1.42± 0.41	1.36± 0.53	1.48± 0.22
	Turkish	1.50± 0.18	1.43± 0.29	1.51± 0.12
	Swedish	1.33± 0.15	1.39± 0.12	1.37± 0.09
	Marathi	1.35± 0.12	1.42± 0.12	1.44± 0.10
NMT	English	1.32± 0.20	1.50± 0.17	1.48± 0.20
	Finnish	1.24± 0.22	1.28± 0.18	1.26± 0.11
	Turkish	1.60± 0.15	1.71 ± 0.12	1.52± 0.32
	Swedish	1.31± 0.22	1.29± 0.31	1.36± 0.30
	Marathi	1.02 ± 0.10	1.22± 0.25	1.30± 0.18

[McGovern, Sirin & Lippincott 2025, NAACL]

- Same allusions, measured in original languages, human translations, and machine translations
- Within vs. across the Jewish and Christian canons
- Translation reshapes the intertextual signal, and not uniformly by language

Chiasmus

A Ask not what your country
B can do for you,
B′ but what you can do
A′ for your country.

Given similarities S_{ij} between segments:

$$\mu_{\text{consistent}} = \frac{1}{|\text{matched}|} \sum_{i,j \in \text{matched}} S_{ij}$$

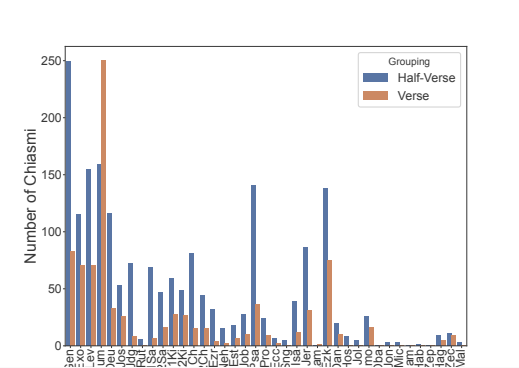
$$\mu_{\text{inconsistent}} = \frac{1}{|\text{unmatched}|} \sum_{i,j \in \text{unmatched}} S_{ij}$$

$$\text{chiasm-ness} = \mu_{\text{consistent}} - \mu_{\text{inconsistent}}$$

[McGovern, Sirin & Lippincott 2025, NAACL]

Structure across religious text

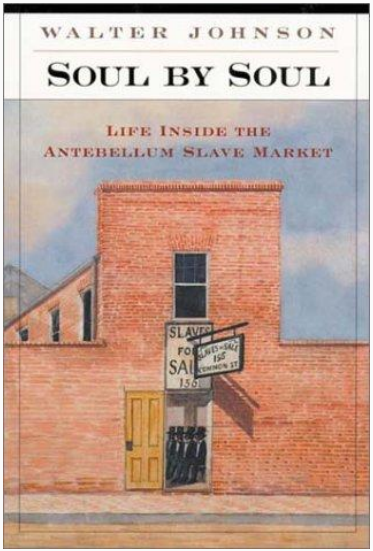
Chiasmi



Latin	English
panis esuriret	so that **bread might hunger**,
satietas sitiret	**satiety might thirst**,
uirtus infirmaretur	**strength might be weakened**,
sanitas uulneraretur	**health might be wounded**,
uita moreretur	**life might die**.

Rhetorical devices Augustine: the structure survives in Latin, and in translation.

Study 3: People in the archive



Study 3

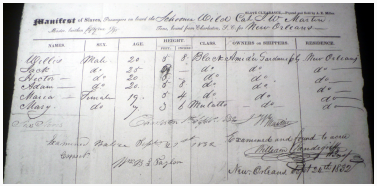
People in the archive: from manifests to clay tablets

Walter Johnson, *Soul by Soul: Life Inside the Antebellum Slave Market* (1999)

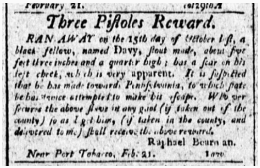
COMPLETE

Can a general model of structured records surface the people, roles and patterns that historians reconstruct by hand?

Primary sources: the domestic slave trade



Shipping manifests: 45k entries, five ports



Fugitive notices: 11k from 70 gazettes

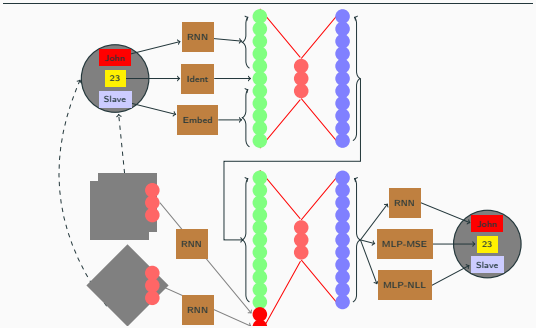
Data in the wild is

- *unnormalized*: one person recorded many times
- *noisy*: Bark, Barke, BArque, Barque...
- *underspecified*: most enslaved people have no surname; are two "John"s the same person?

Historians want to

- follow one person through their life
- reconstruct families without surnames
- map trade ecosystems

StarCoder: an autoencoder for entity-relational data



[Lippincott 2020, 2024]

1. Scholar describes the domain as a schema: entity types, typed fields, relationships
2. Schema generates a model: per-type encoders/decoders, linked by graph convolutions
3. Train to reconstruct entities with randomly masked fields

Use: embedding similarity, clustering, filling missing fields, flagging improbable values

What the model notices

Mistranscriptions

Baltiomre	$\Longleftrightarrow$	Baltimore
William Wiliams	$\Longleftrightarrow$	William Williams

Semantically-equivalent variants

George Y. Kelso	$\Leftrightarrow$	Kelso & Ferguson
New Orleans	$\Leftrightarrow$	Louisiana

Same slave transported multiple times[1]

Louisa, F, 16yo	$\Leftrightarrow$	Louisa, F, 17yo
Waters, F, 14yo	$\Leftrightarrow$	Waters, F, 15yo
Kesiah, F, 20yo	$\Leftrightarrow$	Kesiah, F, 22yo
Taylor, F, 15yo	$\Leftrightarrow$	Taylor, F, 16yo

[1] M

Now: officials on the Hittite frontier (Emar, c. 1300 BCE)



EA 286 (P271088)

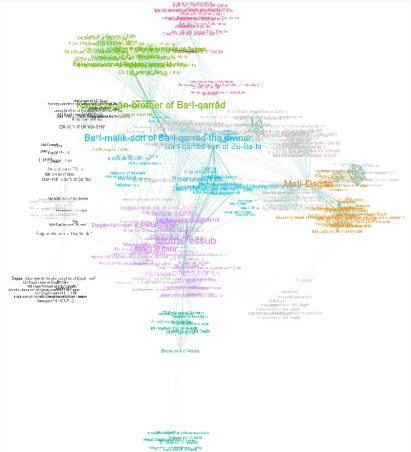
Primary Publication: Knudtzon, Jørgen A. (1915) EA EA 286
Collection: Vorderasiatisches Museum, Berlin, Germany
Museum no.: VAT 01642
Provenience: Akhetaten (mod. el-Amarna)
Period: Middle Babylonian (ca. 1400-1100 BC)
Object Type: tablet or envelope > tablet
Material: clay

Transliteration:
atf: lang akk
tablet
obverse
1. [a]-na {disz}lugal _en_-ia qi2-bi-ma
2. um-ma {disz}ARAD2-he-ba _ARAD2_-ka-ma
3. a-na 2(disz) _giri3-mesz en_-ia _lugal_-ri
4. 7(disz)-ta-a-an u3 7(disz)-ta-a-an am-qut-mi
5. ma-an-na ep-sza-ti a-na _lugal en_-ia
6. i-ka-lu ka-ar-s,i-ya :" u2-sza-a-ru

[with J. Lauinger & L. Freire; Freire 2026]

- Akkadian legal and economic tablets from Tell Meskene, Syria
- Scholarly debate: what separated a "king's son" from an "overseer of the land"?
- The Prosopography of the Neo-Assyrian Empire: 20 years, 70 people

Both roles for AI in one project



IN PROGRESS

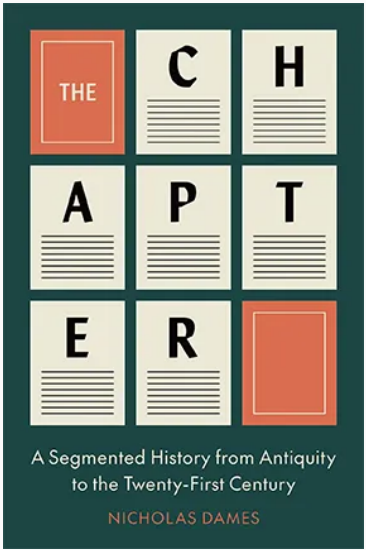
Research assistant

- An agentic LLM extends a dissertation's hand-built entity-relational data across the archive
- Checked against held-out human annotation

Empirical model

- GCN trained on uncontested structure only (who appears with whom, in what transaction)
- Do the learned roles align with the *titles*? Where do they diverge?

Study 4: Hierarchy and memory in cultural artifacts



Study 4

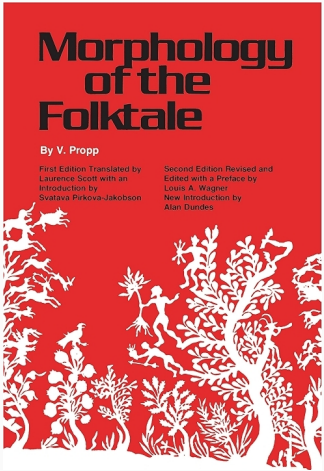
Hierarchy and memory in cultural artifacts

Nicholas Dames, *The Chapter: A Segmented History from Antiquity to the Twenty-First Century* (2023)

IN PROGRESS

Can a model with human-like memory constraints discover the segments and hierarchies of cultural artifacts on its own?

Segmentation is cultural and cognitive



- Chapters, stanzas, verses, scenes: readers and writers segment continuous experience
- Propp (1928): every Russian fairy tale draws from 31 functions in a fixed order, an early theory of schemas
- Cognitive science: we remember experience as events and schemas ("eating at a restaurant")

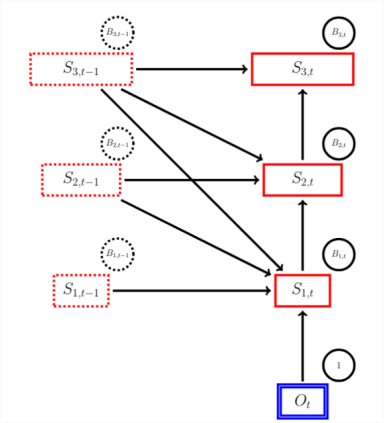
Hypothesis: structure emerges from *limited memory* that must decide what to keep, and at what level.

Sequential cultural artifacts, natural and synthetic

Domain	Source	Scope	Structure
Poetry	Chadwyck Poetry Corpus	∼300k English poems, 1400–present	Typographic
Narrative	Multilingual Bibles	∼30k verses in ∼1k languages	Book/chapter/verse
	US Novels	∼1,300 pre-1920 texts	Typographic
Music	PDMX	∼250k scores / 6,250 hours	
	Distant Listening Corpus	∼1.2k scores	Expert annotated
Synthetic	Long-distance	recall after local noise	Known
	Hierarchical	linearized boolean equations	Known

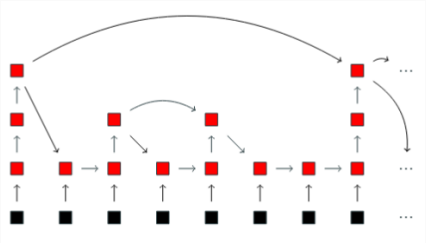
[with R. Lieck, J. Hale, M. Martin]

A hierarchical recurrent memory cell (HRNN)



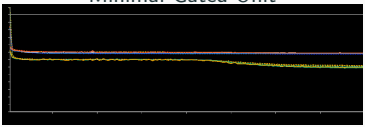
- L memory levels; only the lowest sees the input and makes predictions
- Each level gates how much to update vs. copy forward
- The model must learn to route information up for safekeeping and back down when needed
- Gates lie in $[0, 1]$: directly interpretable, and near-identity paths can be skipped at test time
- Trained only on next-token prediction

Long-distance recall amid local patterns

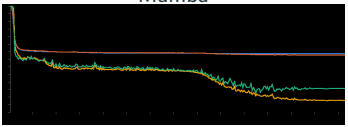


Remember an early token across many locally-predictable ones; same tiny parameter budget for all models.

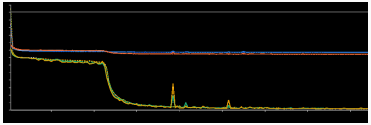
Minimal Gated Unit



Mamba

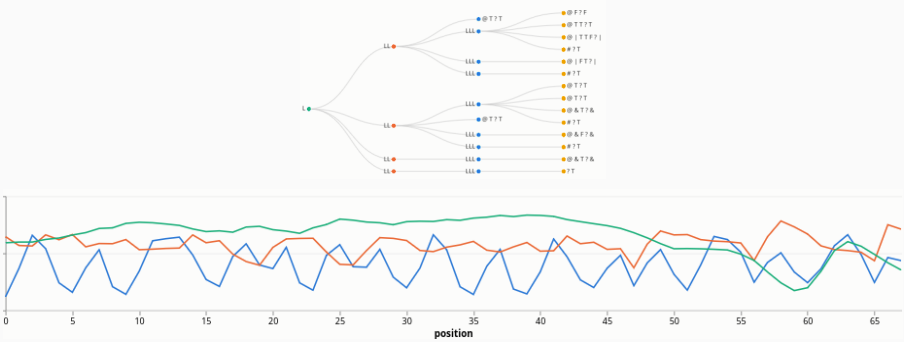


HRNN



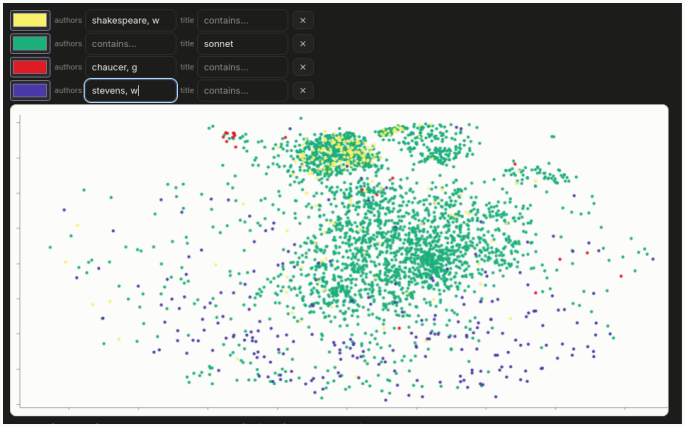
Interpretable memory activity

$$((T \mid (F \mid T)) \,\&\, ((T \& T) \& T))$$



Lower levels track literal values and subclauses; the top level waits until information about the whole equation arrives.

A humanistic example: poems

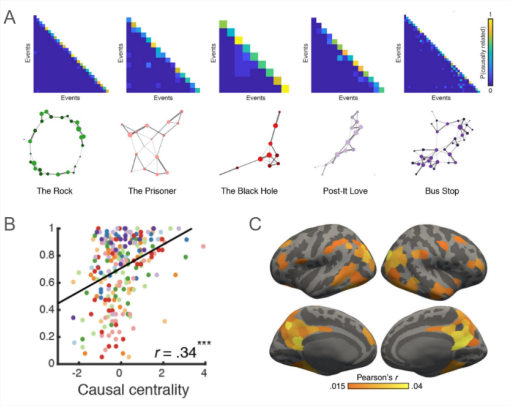


t-SNE of poems, using only 9 numbers each: means/variances of the memory gates and their strongest frequency component

- Shakespeare sits among the sonnets
- Stevens disperses

No words, no meter labels: just *how memory was used*.

Connecting to human memory

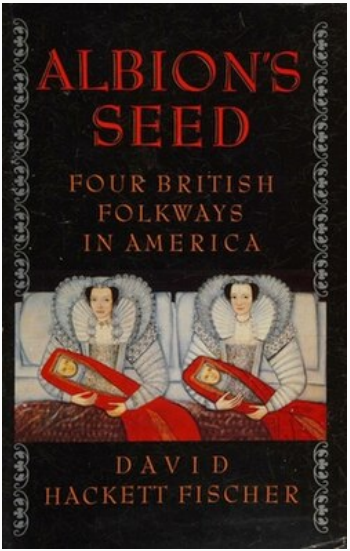


Lee & Chen 2022, *Nature Communications*

- Events whose causal centrality in a narrative is high are better remembered
- Effect is stronger when the subject has agency over the narrative
- Hippocampal activity marks event boundaries and registration to long-term memory

Does the HRNN show the same salience effects? Where does it diverge?

Study 5: Simulating Appalachian folkways



Study 5

Simulating Appalachian folkways

David Hackett Fischer, *Albion's Seed: Four British Folkways in America* (1989)

SPECULATIVE

Can we simulate the constraints and population of the early American frontier to test which proposed factors explain how it was settled?

Fischer's thesis

Four waves of British migrants carried distinct, durable folkways:

Origin	Destination	Period
East Anglia (Puritans)	Massachusetts	1629–1640
South England (Cavaliers, servants)	Virginia	1642–1675
North Midlands (Quakers)	Delaware Valley	1675–1725
Borderlands (Scots, Irish, English)	Backcountry / Appalachia	1717–1775

Appalachia as a cultural flashpoint of geography, climate, and settlers' religious and social backgrounds. Since challenged and revised, e.g. Rehder, *Appalachian Folkways* (2004).

Data in hand

Structural constraints

- Elevation (3DEP), soils (SSURGO), hydrography (NHD), climate
- Historic topographic maps (USGS)

Speculative inputs

- Native populations and languages
- Arrival dates, origins and quantities from *Albion's Seed*

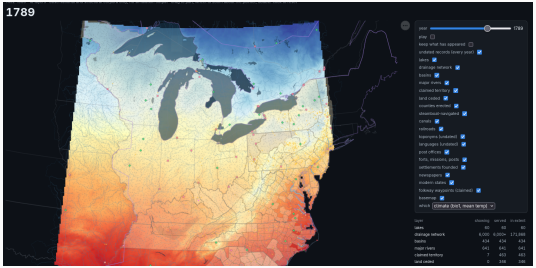
[with A. Maxwell (WVU)]

Systematic observations

Event type	Count
County definitions	7,938
Settlement foundings	38,843
Post office creations	44,875
Railroad construction	40,336
Territory claimed/ceded	809
Newspapers established	1,704

Model: many small agents with memory

- Multi-agent RL and/or flow-based trajectory models
- Each settler is an agent built from Study 4's HRNN: tiny, individual memory
- Group- and global-level parameters for shared cultural priorities
- Rewards from terrain, soil, water, resources, existing settlement



Prototype interface: empirical and simulated data over time

The scholarly payoff is in the ablations

Example
Does giving agents a preference for land that resembles their *European origin* produce a settlement distribution closer to the historical record?

- Each factor proposed by the scholarship becomes a switch
- Compare simulated and observed foundings, post offices, counties
- Same framework could serve historical linguistics, spread of religious thought, disease, conflict...

Closing

Beyond: challenge tasks for cultural competence

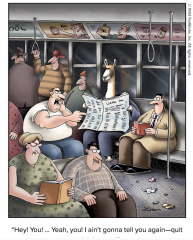
Natural contingency

hog	ham	turkey
clown	gull	mac
tit	jay	beef

Detect introduced errors

HURRY UP PLEASE ITS TIME

HURRY UP PLEASE ITS CLOSING





Takeaways

- Start from a scholarly claim, ideally a book someone spent a decade on
- Build models that encode the hypothesis, not just predict
- Small data, small models, constraints as features: what a model *can't* know or do is often the point
- Interpretability falls out of good architectural choices
- The same machinery transfers: poems $\leftrightarrow$ game NPCs $\leftrightarrow$ frontier settlers

References i

E. Fittschen, S. Li, T. Lippincott, L. Choshen, C. Messner. Pretraining Language Models for Diachronic Linguistic Change Discovery. *EACL* 2026.

E. Fittschen et al. Dynamic Embedded Topic Models: properties and recommendations based on diverse corpora. arXiv:2504.19209, 2025.

H. Sirin, T. Lippincott. Dynamic embedded topic models and change-point detection for exploring literary-historical hypotheses. *LaTeCH-CLfL* 2024.

H. McGovern, H. Sirin, T. Lippincott. Characterizing the Effects of Translation on Intertextuality using Multilingual Embedding Spaces. *NAACL* 2025.

H. McGovern, H. Sirin, T. Lippincott. Computational Discovery of Chiasmus in Ancient Religious Text. *NAACL* 2025.

T. Lippincott. StarCoder: A general neural ensemble technique to support traditional scholarship, illustrated with a study of the post-Atlantic slave trade. *Digital Humanities* 2020.

T. Lippincott. Graph-Convolutional Autoencoder Ensembles for the Humanities, Illustrated with a Study of the American Slave Trade. arXiv:2401.00824, 2024.

References ii

L. G. Freire. Social Networks and the Communities of King's Sons and Overseers of the Land in Late Bronze Age Emar. PhD thesis, JHU, 2026.

H. Lee, J. Chen. Predicting memory from the network structure of naturalistic events. *Nature Communications* 13:4235, 2022.

E. Auerbach. *Mimesis*. 1946. R. Alter. *The Art of Biblical Narrative*. 1981.

W. Johnson. *Soul by Soul*. 1999. N. Dames. *The Chapter*. 2023.

V. Propp. *Morphology of the Folktale*. 1928. D. H. Fischer. *Albion's Seed*. 1989.

J. B. Rehder. *Appalachian Folkways*. 2004.