Converging Truths

LANGUAGE-SPECIFIC BIAS AND HISTORICAL NARRATIVES FROM ENCYCLOPEDIAS TO WIKIPEDIA

DIOGO PINTO

NOVA SBE

CONTEXT

Battle Name	Language	Source type	Winner	Loser	Number of Swedish Soldiers	Number of Russian Soldiers
Battle of Narva (1700)	russian	encyclopedia	sweden	russia	**23000.0**	**35000.0**
Battle of Narva (1700)	french	encyclopedia	sweden	russia	**10000.0**	**80000.0**

Research Question

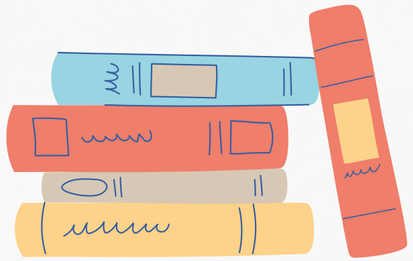
1. Do historical battle accounts from the 20th century differ systematically across languages?
2. Are these differences inherited by early Wikipedia?
3. Do records converge over time?

Hypotheses

- H1: Pre-digital encyclopedias show strong cross-lingual divergence
- H2: Winner–loser asymmetries shape historical reporting
- H3: Early Wikipedia inherits encyclopedic bias
- H4: Narrative bias persists longer than numeric bias
- H5: Overall bias declines across the digital transition

DATA STRUCTURE

Physical Encyclopedias (1930–2000)



- 7 languages
- Digitized & OCR-processed
- Pre-digital baseline

Wikipedia Pages

- Earliest version (2001–2005)
- Mid version (2006–2008)
- Latest version

~50 major historical battles

Data Collection & Database Construction

Encyclopedia Only

Collection and scanning of physical and online encyclopedias (1930–2000)

1

Detection of Pages with battle information.
OCR Optimization.

(Pages where city and battle year were mentioned.)

2

AI-based extraction of general battle related information.

3

Data Collection & Database Construction

Encyclopedia and Wikipedia

Usage of an AI pipeline to normalize information into key values.

(Winner, Loser, Number of soldiers...)

4

Normalization of key values

(language, battle id, display of troops, among others...)

5

What Is being Measured:

Numerical Dimension

- Troop counts
- Casualty figures

Narrative Dimension

- Causes of the battle
- Consequences
- Key actors & framing

Identical key variables extracted for both encyclopedias and Wikipedia pages

Methodology

Numeric Analysis

- Max / Min ratios
- Log deviations from median

Narrative Analysis

- TF-IDF (lexical framing)
- LLM embeddings (semantic framing)

Statistical Tests

- Mann–Whitney U
- Wilcoxon signed-rank

Pre-Digital Encyclopedic Bias (H1)

Large cross-language disagreement

I **Large cross-lingual numeric disagreement:**

Troop counts often differ by 2–3×

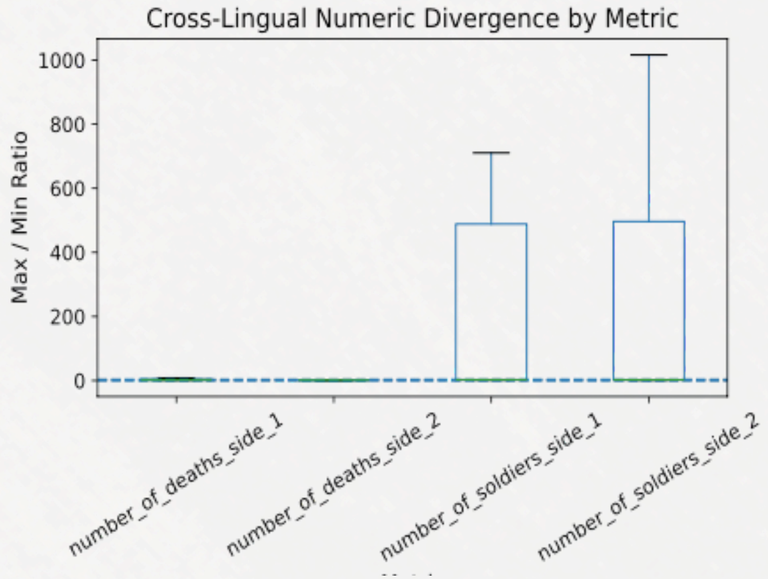
II **Narrative divergence is substantial:**

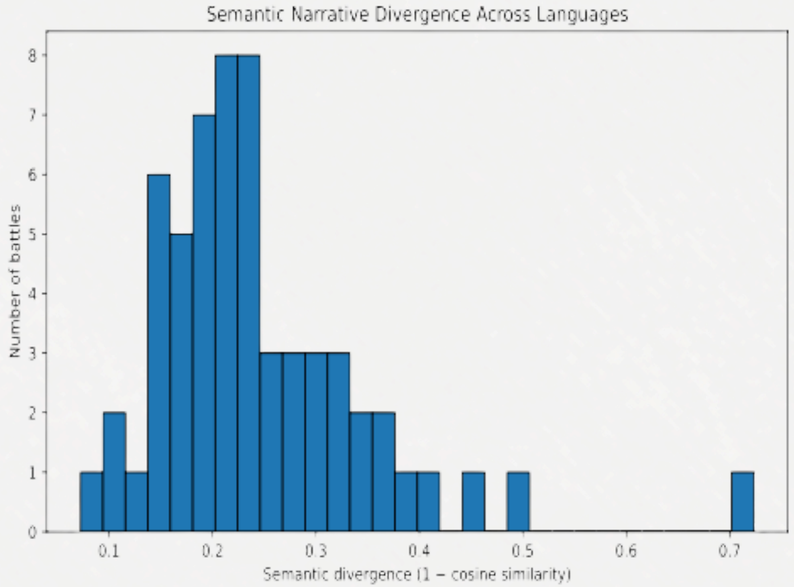
Low TF-IDF and semantic similarity

III **Bias can be seen as systematic, not random.**

Pre-Digital Encyclopedic Bias (H1)

Large cross-language disagreement





Metric	Battles	Median Ratio	Share Ratio Gt 1 5	Share Ratio Gt 2	Median Cv
number_of_deaths_side_1	3	1.500	0.667	0.333	0.283
number_of_deaths_side_2	5	1.146	0.200	0.000	0.109
number_of_soldiers_side_1	7	2.875	1.000	0.571	0.590
number_of_soldiers_side_2	6	2.574	0.667	0.667	0.438

Outcome-Conditioned Bias (H2)

I

Loser-aligned languages report on average higher enemy troop counts.

II

Effect is directional but weaker than H1.

Table 5.4: Winner–Loser Numeric Asymmetry in Pre-Digital Encyclopedias

Metric	Winner Mean Log Bias	Loser Mean Log Bias	Difference Loser Minus Winner	P Value	N Winner	N Loser
winner_troops_log_bias	-0.01	0.187	0.197	0.215	111	113

Table 5.5: Winner–Loser Numeric Asymmetry (Wikipedia v2)

metric	winner_language_mean	loser_language_mean	loser_minus_winner	p_value	n_winner	n_loser
winner_troops_log_bias	-0.232	0.056	0.287	0.08	23	24

Bias Inheritance & Persistence (H3 + H4)

H3 — Bias Inheritance

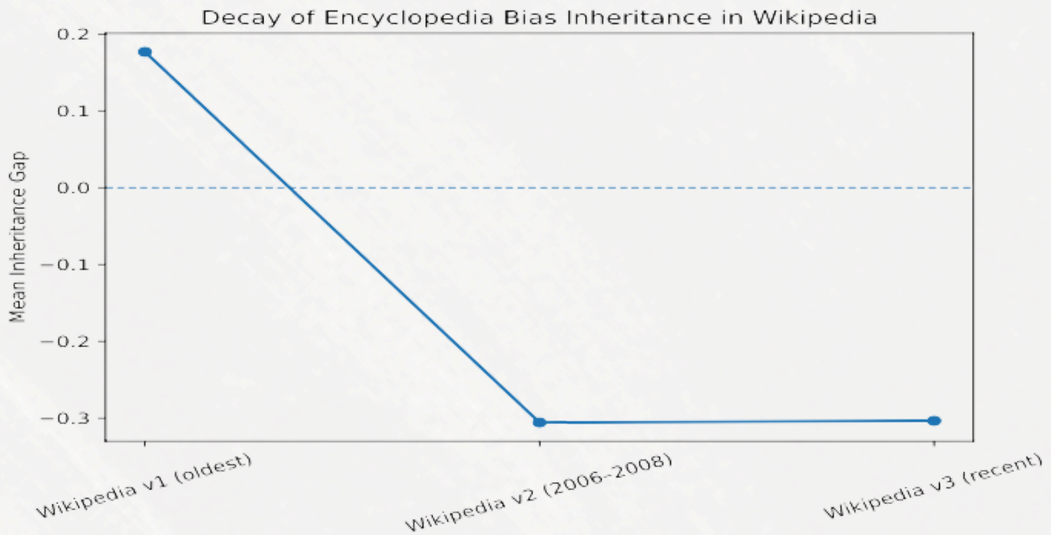
- Early Wikipedia mirrors encyclopedic bias
- Strong inheritance in earliest versions
- Bias weakens over time

H4 — Narrative Persistence

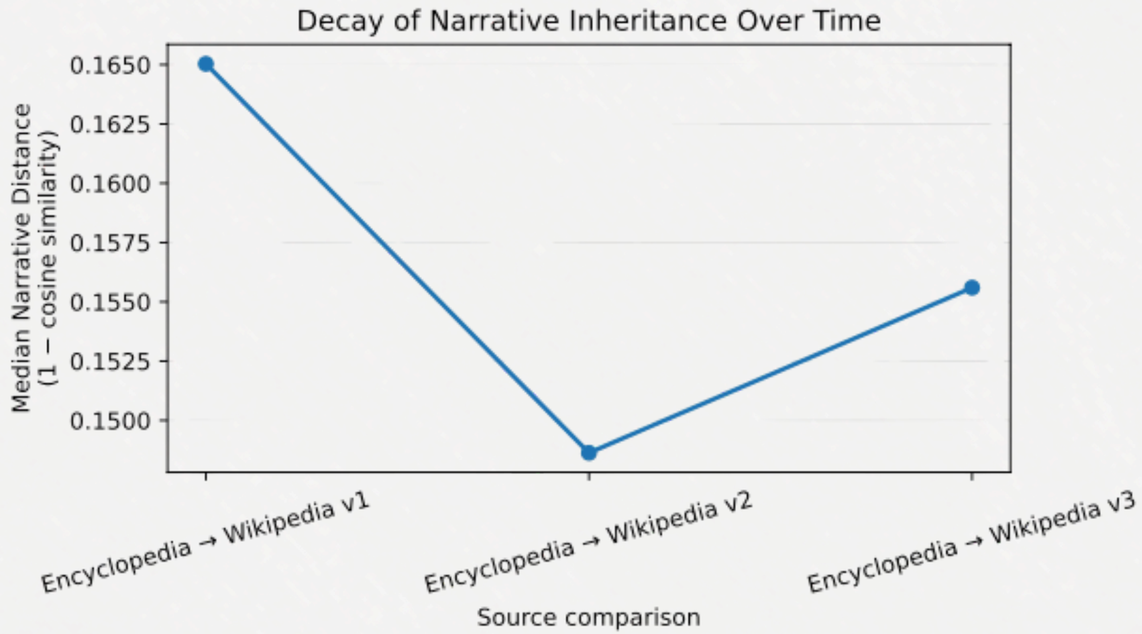
- Numeric bias converges quickly
- Narrative framing converges slowly
- Winner-aligned narratives persist longer

Bias Inheritance & Persistence (H3 + H4)

H3 — Bias Inheritance



H4 — Narrative Persistence

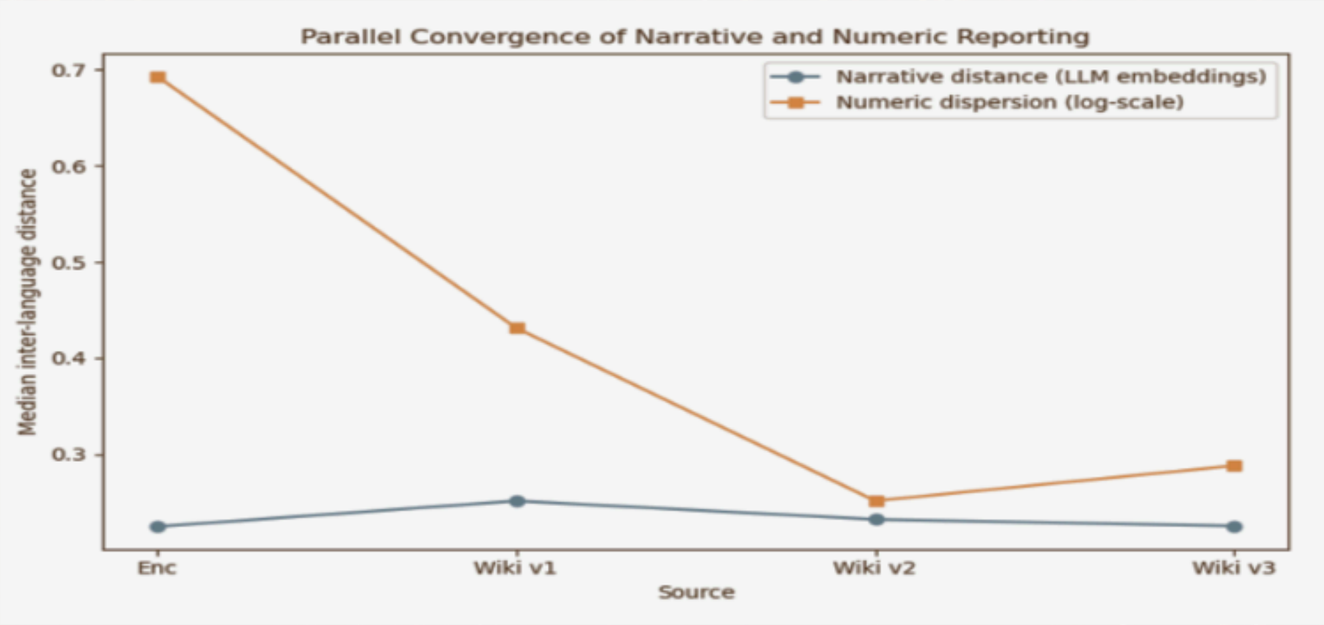


Overall Trajectory of Bias (H5)

I **Bias is highest in pre-digital encyclopedias.**

II **Numeric disagreement converges rapidly on Wikipedia.**

III **Narrative framing converges slowly and unevenly.**



Limitations & Scope

Uneven cross-lingual coverage limits statistical power.

No truly neutral languages: neutrality is relative.

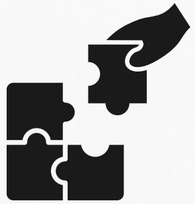
AI-based extraction and embeddings introduce noise.

Wikipedia sampled only at three discrete time points.

Time constraints limited scale and language coverage.

Contributions & Conclusion

Key Contributions



- Likely first large-scale pre-digital baseline of historical narratives
- Evidence of bias inheritance into early Wikipedia.
- Layered convergence of numeric vs narrative bias.

Conclusion

- Bias is historically embedded, inherited, and partially correctable
- Digital platforms constrain bias, rather than erase it

The End

THANK YOU FOR LISTENING

Diogo Pinto

NOVA SBE

APPENDIX

Table 5.1: Numeric Dispersion Across Languages

Metric	Battles	Median Ratio	Share Ratio Gt 1 5	Share Ratio Gt 2	Median Cv
number_of_deaths_side_1	3	1.500	0.667	0.333	0.283
number_of_deaths_side_2	5	1.146	0.200	0.000	0.109
number_of_soldiers_side_1	7	2.875	1.000	0.571	0.590
number_of_soldiers_side_2	6	2.574	0.667	0.667	0.438

Figure 5.1

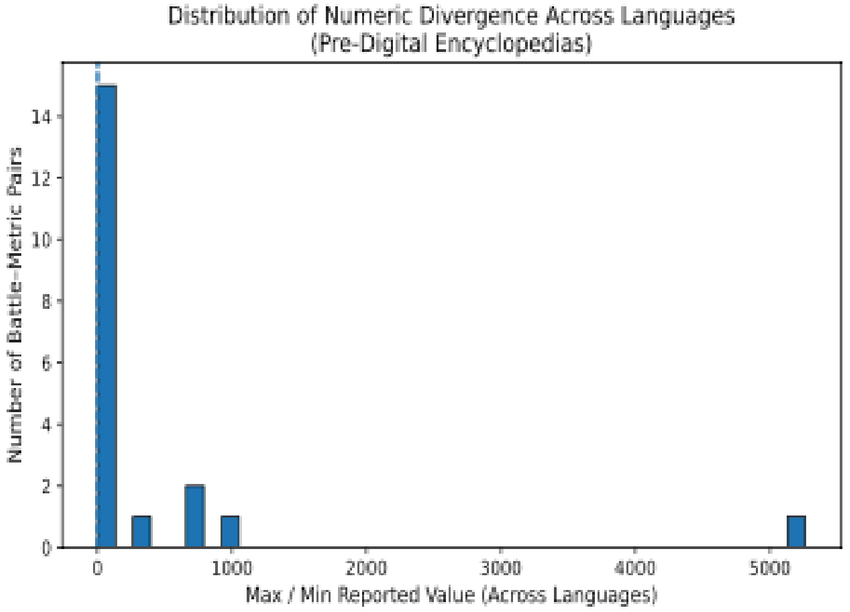
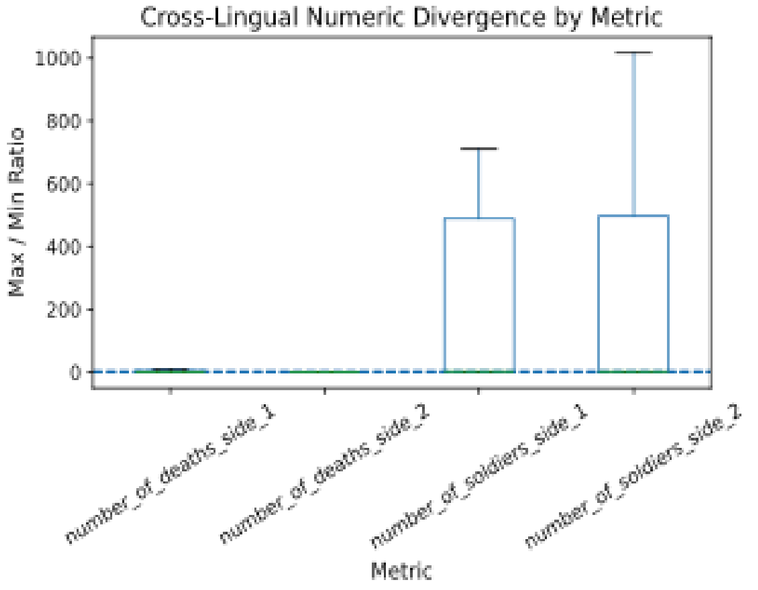


Figure 5.2



APPENDI

Table 5.2: Language-Specific Numeric Bias (Encyclopedia Corpus)

Metric	Language	N	Mean log bias	Median log bias	p-value
number_of_deaths_side_1	russian	6	0.0	0.0	nan
number_of_soldiers_side_1	french	8	0.115	0.0	0.1625
number_of_soldiers_side_1	russian	10	-0.722	0.0	0.3907
number_of_soldiers_side_1	spanish	6	-0.751	0.0	0.5003
number_of_soldiers_side_1	english	6	-0.833	0.0	0.3522
number_of_soldiers_side_2	french	5	0.09	0.0	0.1985
number_of_soldiers_side_2	russian	11	-0.594	0.0	0.2825

Figure 5.3

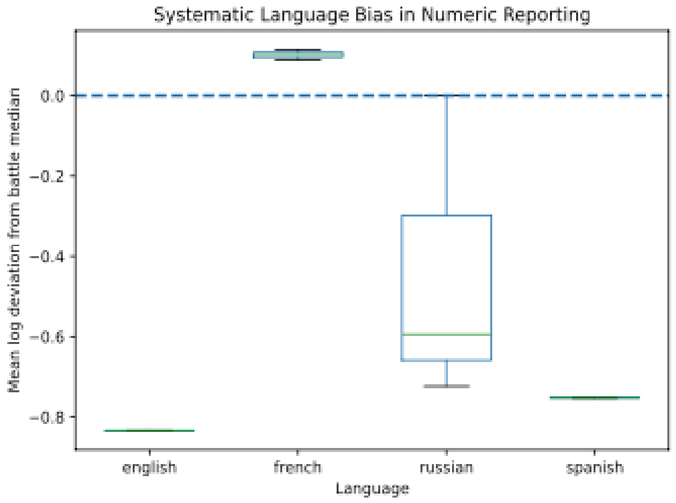
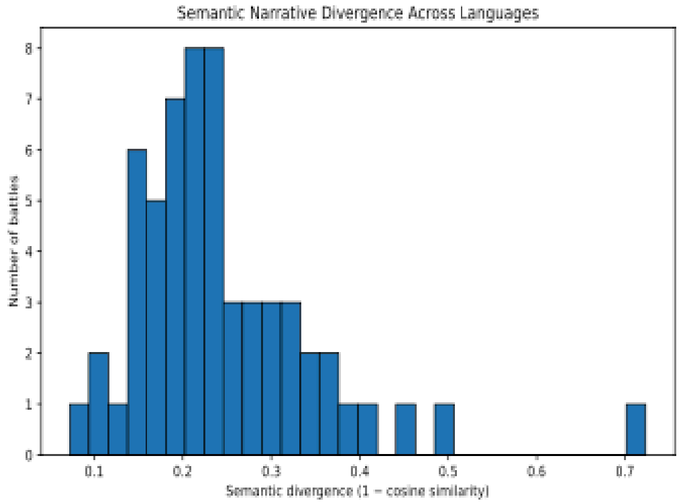


Figure 5.4



APPENDIX

Figure 5.5

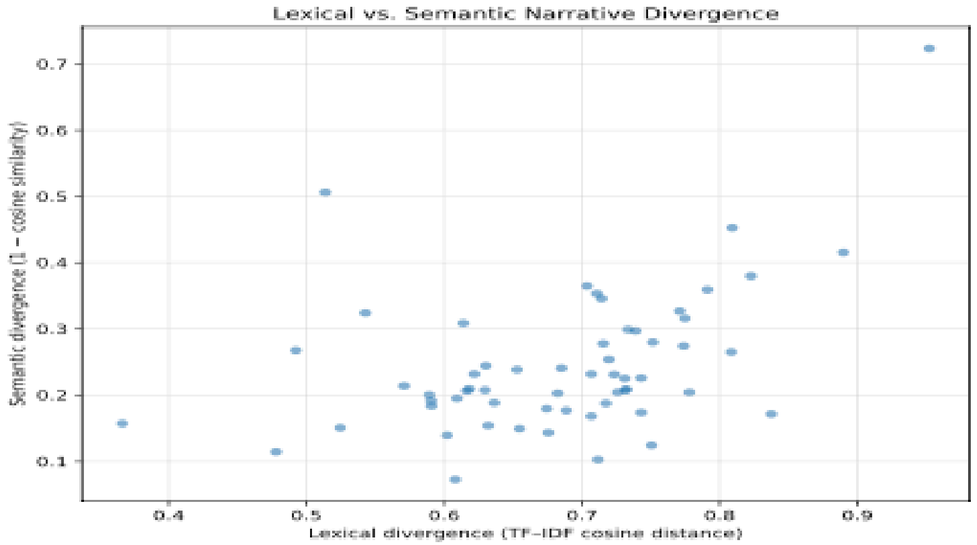


Table 5.3: Battles with Highest Numeric and Narrative Divergence

Battle	Max / Min Numeric Ratio	Semantic Divergence
Tsushima_1905	5263.16	0.204
Trafalgar_1805	1016.13	0.143
Lepanto_1571	266.67	0.184
Narva_1700	2.88	0.209
Waterloo_1815	2.86	0.225
Culloden_1746	1.8	0.114

APPENDIX

Table 5.4: Winner–Loser Numeric Asymmetry in Pre-Digital Encyclopedias

Metric	Winner Mean Log Bias	Loser Mean Log Bias	Difference Loser Minus Winner	P Value	N Winner	N Loser
soldiers_side_1	-0.01	0.187	0.197	0.215	111	113
soldiers_side_2	-0.214	-inf	-inf	nan	106	98
deaths_side_1	-inf	-inf	nan	nan	103	97
deaths_side_2	-0.045	-inf	-inf	nan	105	89

Table 5.5: Winner–Loser Numeric Asymmetry (Wikipedia v2)

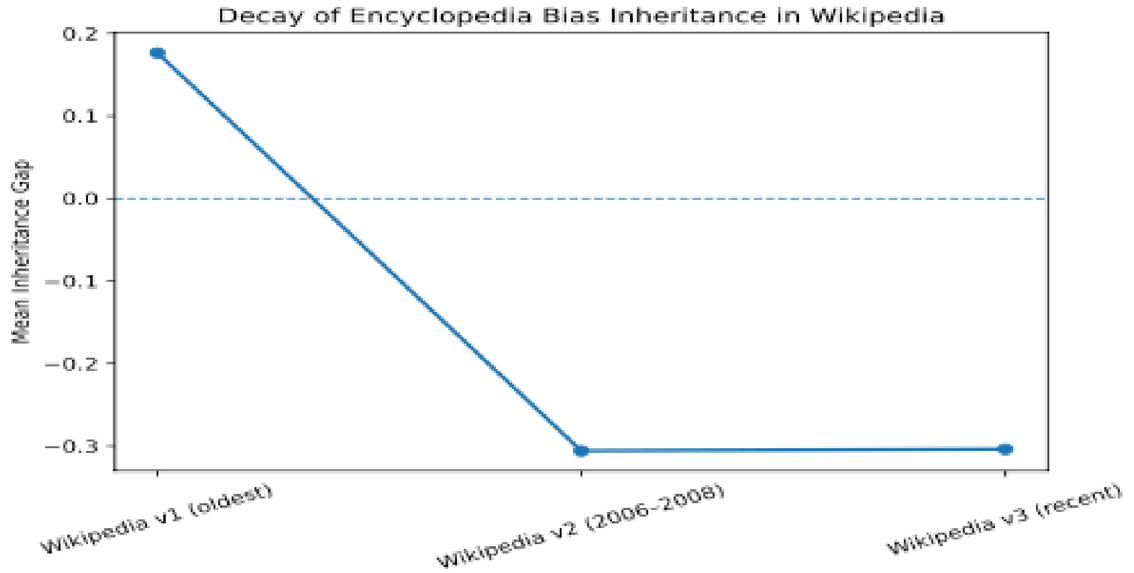
metric	winner_language_mean	loser_language_mean	loser_minus_winner	p_value	n_winner	n_loser
winner_troops_log_bias	-0.232	0.056	0.287	0.08	23	24
winner_casualties_log_bias	-0.168	-inf	-inf	nan	23	16
loser_troops_log_bias	-0.151	-0.023	0.128	0.513	22	18
loser_casualties_log_bias	0.065	-inf	-inf	nan	25	14

APPENDIX

Table 5.7: Numeric Bias Inheritance from Encyclopedias to Wikipedia

Wikipedia Era	Number of Observations	Mean Inheritance Gap	Median Inheritance Gap	Share Positive Gaps	Wilcoxon p-value
Wikipedia v1 (oldest)	11	0.177	0.144	0.727	0.007
Wikipedia v2 (2006–2008)	14	-0.306	0.002	0.5	0.643
Wikipedia v3 (most recent)	20	-0.304	-0.044	0.4	0.785

Figure 5.6

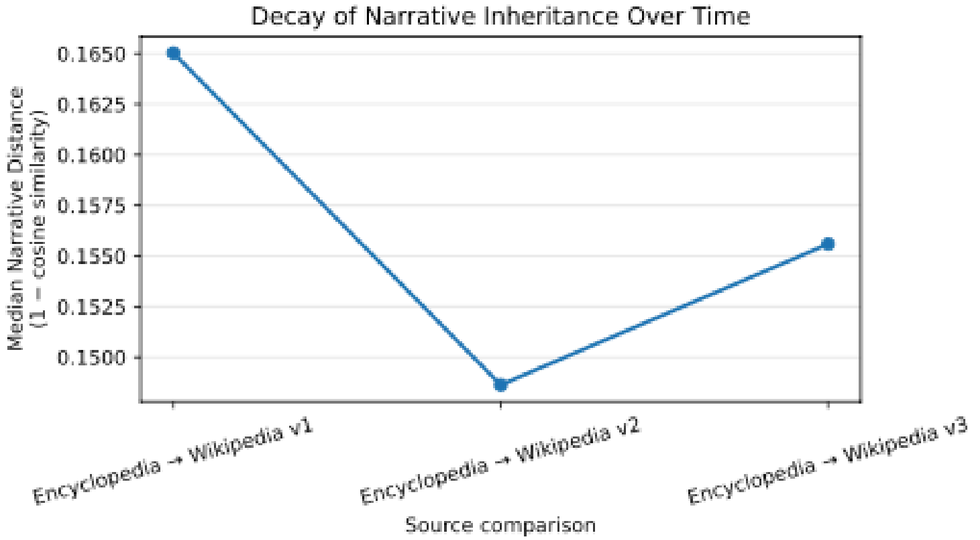


APPENDIX

Table 5.9: Narrative Distance Between Encyclopedias and Wikipedia Versions

Source comparison	Median semantic distance	Mean semantic distance	Battle–language pairs	Unique battles
Encyclopedia → Wikipedia v1	0.171	0.232	274	68
Encyclopedia → Wikipedia v2	0.152	0.193	179	57
Encyclopedia → Wikipedia v3	0.159	0.211	276	72

Figure 5.7



APPENDIX

Figure 5.8