

RESEARCH REPORT | 10 SEPTEMBER 2026

Preedy-29 Structural Null Tests

Structural organization in the 39-folio Herbal-A corpus

Juan Gabriel Molina
Independent Researcher

Framework by Christine Preedy | Independent testing, reconstruction, and report by Juan Gabriel Molina.

Abstract

This report documents the frozen structural-null experiment supplied in the Preedy-29 replication package. A deterministic reconstruction yields 16,575 units in 3,284 blocks across 39 Herbal-A folios, with 29 observable states. Real and simulated streams are mapped through the same fixed ten-role projection and evaluated using position-role normalized mutual information, adjacent role-transition normalized mutual information, and a pooled seven-relation ordering score. Each statistic exceeds all 500 simulated scores from each of three fitted generator families: first-order Markov, second-order Markov, and the package's finite-history CSSR implementation. Every comparison therefore returns the protocol's minimum empirical upper-tail probability, $1/501 = 0.001996$, satisfying its PASS criterion. The finding concerns these aggregate statistics under the specified generators and representation. It does not establish the literal meanings of the role labels, the uniqueness of the partition, or a manuscript-wide decipherment. Numerical tables below are drawn from the original package, not from external reviewer sensitivity analyses. [1-3]

Research question and scope

The question is whether the observed structural organization, measured after a fixed projection of the Preedy-29 stream, is reproduced by the fitted stochastic generators. This is an observable-sequence simulation comparison. It is not a folio-held-out prediction experiment and does not compare competing translations.

The package is treated as a frozen research-software record. This report explains its experiment and reproduces its recorded summaries; it introduces no new null family, revised role definition, additional corpus section, or independent per-relation significance test.

Keywords: Voynich Manuscript; Herbal-A; Preedy-29; structural null models; Markov surrogates; finite-history CSSR; reproducibility.

1. Corpus and deterministic representation

The occurrence export included in the package supplies the primary-eligible records. The selected folios run from f1r through f21r, with f12 absent, giving 39 folio sides. These counts describe this Herbal-A sample, not the full manuscript. The analysis uses the supplied transcription and block segmentation rather than independently establishing their palaeographic correctness. [1,2]

Property	Frozen value
Folio sides	39
Blocks	3,284
Reconstructed units	16,575
Observable states	29
In-corpus no-string r units	3
Average units per block	5.0472

1.1 Reconstruction rules

The builder keeps c and h separate, treats f+, k+, p+ and t+ as atomic forms, groups ordinary maximal e-runs into the specified e/ee/eee/eeee symbols, and maps unusual or starred forms to the special-symbol umbrella. Three identified occurrences are assigned r_no_string through the supplied location ledger; the known fourth occurrence on f22r is outside this corpus. The special r assignment is applied before generic handling of starred forms. Source-row provenance disambiguates a duplicated block identifier. [2]

1.2 The fixed scoring projection

Role name	Observable symbols
PROJECT / RECEIVE	c / h
SOLAR / HORIZON	2 / f+, k+, p+, t+
PLANET	f, k, p, q, t
GATE	d, n, o, y
STRING / DOSE	a, i / r, r_no_string
OTHER	e, ee, eee, eeee, g, l, m, s
SPECIAL	*

Role names identify the fixed scoring classes. Their English meanings are not separately tested. Although the generator retains the two r states, the projection maps both to DOSE; a pass therefore does not isolate the value of that distinction. This ten-role scoring projection is part of the declared experiment and is not a JEC12, neural, or hierarchical comparison.

2. Structural statistics and simulation protocol

All real and surrogate blocks pass through the same `structural_metrics` function. No transition crosses a block boundary. The three statistics are pooled over blocks; folio labels do not define a held-out assessment in this experiment. [2]

Normalized mutual information: $NMI(X,Y) = I(X;Y) / [(H(X) + H(Y))/2]$.

Position-role NMI compares the role with one of five normalized position bins. For a block of length $L > 1$ and zero-based position i , the bin is $\min(4, \text{floor}(5i/(L-1)))$. A singleton goes in the first bin. Role-transition NMI uses adjacent role pairs within blocks. The logarithm base cancels in the normalized ratio.

The seven-relation score is the total number of satisfied relation-block instances divided by the total number of applicable instances. Applicability requires all named roles to occur. Before/between rules use first occurrences only. The score weights relations by applicability; it is not the mean of seven equally weighted rates. A block can contribute to several relations.

Generator	Specification
M1	One-symbol context; additive smoothing 0.2; smoothed empirical block-start distribution.
M2	Two-symbol context; smoothing 0.2; unseen contexts back off to M1; starts and second symbols use the declared lower-order rules.
Finite-history CSSR	$L_{\max} = 2$; homogeneity threshold 0.001; minimum history count 5; emission smoothing 0.2; deterministic transition refinement.

The finite-history implementation groups histories by predictive-distribution homogeneity and refines symbol-labelled transitions. The recorded fit has 35 states and no nondeterministic transition pairs among 1,015 checked pairs. This is the package's explicitly bounded reconstruction, not a claim to have recovered the manuscript's unique minimal causal-state machine. CSSR methodology is described by Shalizi and Klinkner [4].

Every surrogate preserves the exact block-length vector, hard resets and 29-state vocabulary. Exact symbol totals are not fixed: they fluctuate under the fitted generators. Models are fitted to the observed stream. This is a plug-in simulation comparison, rather than an exact permutation test over exchangeable observations.

Seed base: 20260906. $N = 500$ per family. For each statistic, $p = (1 + \text{number of surrogate scores at least as large as real}) / (N + 1)$. The package passes only when all three statistics meet $p \leq 0.01$ under all three families.

3. Original aggregate results

Position-role NMI

Family	Real	Null mean	Null SD	Null max
M1	0.158061	0.048902	0.001942	0.055503
M2	0.158061	0.056219	0.002015	0.061600
CSSR	0.158061	0.052657	0.001929	0.059283

Role-transition NMI

Family	Real	Null mean	Null SD	Null max
M1	0.400985	0.358042	0.003690	0.368282
M2	0.400985	0.306262	0.004067	0.318898
CSSR	0.400985	0.371286	0.003583	0.381645

Seven-relation score

Family	Real	Null mean	Null SD	Null max
M1	0.816901	0.737362	0.005065	0.752429
M2	0.816901	0.751179	0.005248	0.771718
CSSR	0.816901	0.738385	0.005168	0.755340

Table 1. Original frozen results, 500 surrogates per family. All nine comparisons have zero exceeders and empirical upper-tail $p = 0.001996$. SD describes the simulated distribution, not uncertainty in the transcription or uncertainty in the real-corpus statistic. Values are rounded only for display. [3]

The observed position-role NMI is 0.158061, adjacent role-transition NMI is 0.400985, and pooled seven-relation score is 0.816901. The package therefore meets its declared aggregate PASS rule. The result is not a claim that $p = 0$; with 500 simulations the displayed procedure cannot resolve a smaller probability than 1/501.

The smallest raw separation between a real score and its family maximum occurs for role-transition NMI under CSSR: 0.400985 versus 0.381645, a difference of approximately 0.019340. This comparison is descriptive and is not a new test.

The nine cells reuse one corpus and three related generators. Their p-values are not multiplied together or presented as independent confirmations. Re-executing a frozen seed can check numerical reproducibility, but does not increase the number of independently sampled surrogates.

4. Descriptive relation accounting and interpretation

First-occurrence relation	Applicable	Satisfied	Rate
PROJECT before RECEIVE	1942	1942	1.000000
PROJECT < SOLAR < RECEIVE	422	414	0.981043
PROJECT < HORIZON < RECEIVE	358	281	0.784916
RECEIVE before GATE	1834	1332	0.726281
PLANET before GATE	1206	619	0.513267
GATE before DOSE	610	590	0.967213
STRING before DOSE	231	216	0.935065

Table 2. Descriptive counts already included in the original `real_metrics/summary` records. These rates are not individual-relation Monte Carlo p-values. No external reviewer calculations are incorporated into this table. [3]

The numerator sums to 5,394 satisfied instances and the denominator to 6,603 applicable instances, giving 0.816901. Differences in applicability mean that common relations contribute more to the pooled score. PROJECT before RECEIVE is satisfied in every applicable real block, but even a rate of 1.0 does not itself establish rarity under an appropriate null.

4.1 What the result supports

Under the frozen projection, the observed sequence has aggregate position, adjacency and first-occurrence ordering statistics exceeding all sampled outputs from the three declared generators. The comparison supports a reproducible mismatch between those generator families and these aspects of this Herbal-A stream.

4.2 Limits of inference

The pooled pass does not establish that each individual relation beats its own null distribution. Per-relation null significance is not part of the original frozen result reported here. Similarly, the experiment does not compare matched alternative role partitions and therefore does not establish that this partition is uniquely informative.

The generators share short-history assumptions: M1 uses one previous symbol, M2 uses two, and the finite-history reconstruction is limited to $L_{\max} = 2$. In particular, positional structure beyond the fitted start distribution is not explicitly matched by position-stratified transition models. The result does not reject every stochastic or nonsemantic explanation.

The package freezes the analysis for reproduction. Its contents alone do not establish that all relation choices preceded inspection of the corpus; this report makes no preregistration claim. Corpus selection, transcription decisions and representation choices condition the finding. Semantic labels, historical attribution and other manuscript sections require separate evidence.

5. Reproducibility, provenance and references

Replication packages are available by email request through the contact details accompanying this report on the hosting website. Request the Preedy-29 Structural Null Replication Package. The package contains source data, reconstruction and simulation code, frozen results, protocol, attribution and a SHA-256 manifest. [1]

From the extracted package directory:

```
python -m pip install -r requirements.txt
python code/build_preedy29_units.py
python code/verify_package.py
python code/run_structural_nulls.py --n 500
python code/verify_package.py
```

The seed is a module constant in this release; a different-seed sensitivity run requires editing that constant in a separately identified copy. Compare parsed numerical results if newline formatting differs. Preserve the original archive and record any environment or code changes. The package depends on Python, NumPy, pandas and SciPy; its requirements file specifies the accepted dependency floors.

Original replication ZIP SHA-256
0885f341196968d8c3e50b8b8bdb77ec
31f79af19adb32eda014ea5849cd893e

Frozen null_distributions.csv SHA-256
8e6d962fee63706ca314969bf73f9ab8
d447a1a325236b3fa2b7aaf8142ba71c

Report verification: the original 19-file manifest and package verifier passed inspection. The 1,500 stored surrogate-score rows were checked against the reported real values and reproduce all nine zero-exceeder counts. This report was assembled from those records; no new surrogate experiment was introduced during report preparation.

References and source records

[1] Preedy, C. and Molina, J. G. (2026). Preedy-29 Structural Null Replication Package. Research software and reproducibility dataset. Package identifier: Preedy29_Null_Replication_2026-09-07. README.md, PROTOCOL.md and CITATION.txt.

[2] Same package. code/build_preedy29_units.py; code/run_structural_nulls.py; data/split_r_locations.csv. These files define the reconstruction, projection and generator implementations.

[3] Same package. results/structural_null_ladder/summary.json; real_metrics.json; null_distributions.csv. Primary numerical sources for Tables 1 and 2.

[4] Shalizi, C. R. and Klinkner, K. L. (2004). Blind Construction of Optimal Nonlinear Recursive Predictors for Discrete Sequences. Proceedings of UAI, pp. 504-511. Author resource: <https://bactra.org/CSSR/> ; preprint: <https://arxiv.org/abs/cs/0406011>. This background reference does not independently validate the present implementation.

Contributions: Christine Preedy originated the framework and proposed interpretations. Juan Gabriel Molina contributed the testing architecture, computational analysis, reproducibility packaging and report.

External replication and technical review: Edwin Honeycutt reported reproducing the original PASS results across M1, M2 and finite-history CSSR. His additional analyses have not been independently reproduced by the report authors; proposed follow-up tests remain outside this release.