INDUSVALLEY.ORG

From Computational Comparison to Decipherment

A reproducibility and evidence review of the Kee2u Indus-Tamil repository

IndusValley.org Research Desk
September 2026

Repository snapshot. Keerthana Jayakumar (Kee2u) and Soniaprogram, *Deciphering the Indus Valley Script*, GitHub commit `73b0b9fe2acb1106f8b1b017c7350d35fa798997`, 15 August 2021.

Overall assessment. The repository is a substantive exploratory digital-humanities project on Tamil morphology and sign-sequence statistics. It does not yet constitute a decipherment of the Indus script. Its strongest contribution is a documented computational pipeline and a set of testable ideas; its weakest point is the inferential jump from similarity in constructed statistics to language identification or reading.

Abstract

This paper independently reviews the public GitHub repository *Deciphering the Indus Valley Script*. The project converts a modern Tamil dependency-treebank corpus into an invented logosyllabic sign stream, develops a support-vector classifier for Tamil morpheme boundaries, measures sign-frequency and positional patterns, and explores whether positional distributions identify syllabic signs. We reconstruct the pipeline from nine principal notebooks, audit the committed feature table, and compare the project claims with established requirements for decipherment. The committed table contains 62,833 character-level rows, 3,583 word forms and 24,521 positive boundary labels. A saved SVM run reports 76.06 percent balanced accuracy, but the random row-level split permits characters from the same word to occur in both training and test sets, while prefix encoding is fitted before the split. These choices can inflate generalisation estimates. A separate positional-sign experiment reports weak minority-class performance, a result the repository itself acknowledges. Statistical resemblance between a deliberately constructed Tamil encoding and an undeciphered sign corpus cannot identify the language, prove a writing-system type, or validate readings. The repository is therefore best understood as an exploratory comparison. We propose a preregistered, leakage-resistant evaluation programme based on artefact-level splits, multiple sign lists, chronological and regional controls, typologically matched languages, synthetic non-linguistic baselines, ablation tests and held-out predictions.

Question and scope

The Indus or Harappan sign system occurs mainly on short inscriptions from the urban Indus tradition. Its brevity, the absence of an agreed bilingual text, uncertainty about the underlying language, and disagreements over sign segmentation make it an unusually hard decipherment problem. Computational work can quantify regularities, compare corpora and reject weak models. It cannot by itself supply missing phonetic values or meanings.

This review asks four narrower questions: What does the Kee2u repository actually compute? Are its published numerical results internally interpretable? Can its comparisons discriminate a Tamil or Dravidian hypothesis from alternatives? What changes would make the work reproducible and falsifiable? The review concerns the public repository at the commit identified on the title page. It does not assess private databases, later work, or the intentions of its authors.

Repository and method reconstruction

Materials

The repository contains preprocessing, machine-learning and statistical-analysis notebooks; a copy of Tamil Dependency Treebank v0.1; generated comma-separated tables and figures; SQL schemas; and a web visualisation. The top-level README credits the Indus Corpus of Indus Texts associated with Brian Wells and Andreas Fuls, the Tamil treebank of Ramasamy and Za˘bokrtský, and an AU-KBC tagged corpus. The committed Tamil treebank states a CC BY-NC-SA 3.0 licence. No top-level software licence or fixed computational environment was found in the audited snapshot.

The main pipeline can be reconstructed as follows:

1. Parse modern Tamil news text with forms, lemmas and part-of-speech tags.
2. Manually derive or curate noun and verb morphemes and reuse existing clitic and postposition segmentation.
3. Give lemmas and morphemes unique integer identifiers; encode residual characters as CV, VC, V or C units; treat proper names syllabically.
4. Join the units into an artificial sign sequence described as logosyllabic Tamil.
5. calculate sign-frequency, positional and pair-connectivity summaries.
6. train classifiers for Tamil morpheme boundaries and for the constructed category “syllable.”

This is a computational transformation of a known modern language. It is not a mapping from Indus signs to Tamil words, sounds or grammatical values. The distinction is decisive: showing that one can manufacture a sign stream with selected properties demonstrates the construction, not the historical identity of the original script.

Committed data audit

The file `Machine_Learning/features.csv` contains 62,833 rows and 17 columns. It represents character positions within 9,581 keyed tokens and 3,583 distinct surface forms. The target `morpheme boundary` contains 38,312 zeros and 24,521 ones. The feature set used by the saved SVM retains index position, vowel status, noun/verb flags, a label-encoded current character and a label-encoded prefix. There are 74 distinct character codes and 11,376 prefix codes (including the missing value). The column `FormWithoutLemma` is missing in 24,853 rows, although it is dropped before modelling.

Table 1: Audited quantities from the committed feature table

Quantity	Count
Character-position rows	62,833
Token keys	9,581
Distinct surface forms	3,583
Non-boundary labels	38,312
Boundary labels	24,521
Distinct character encodings	74
Distinct prefix encodings	11,376

Saved machine-learning results

The notebook `TestSVM.ipynb` uses a 75/25 stratified random split with `random_state=1`, standardises six numeric columns and fits an RBF support-vector classifier with $C = 1$ and $\gamma = 0.8$. Its saved output reports 76.055 percent balanced accuracy. The confusion matrix is 7,820 true class-0 predictions, 1,822 false class-1 predictions, 1,759 false class-0 predictions and 4,308 true class-1 predictions. The reported F1 scores are 0.81 for class 0 and 0.71 for class 1.

These numbers describe recovery of labels in the project's own modern-Tamil table. They do not measure Indus decipherment. More critically, the split occurs after expanding words into character rows. Because the same word and prefix patterns can recur across rows, a random row split is not a reliable test of transfer to unseen words, unseen texts, Old Tamil, or Indus inscriptions. The label encoders are also fitted on the full table before splitting. A defensible evaluation should group by document or at least by word type, fit every transformation on training data only, reserve validation data for tuning, and report repeated confidence intervals.

The positional experiment uses 482 held-out constructed signs (425 non-syllables and 57 syllables). A saved classifier output gives class-1 precision 0.31, recall 0.47 and F1 0.37; alternative tuned neural models also show unstable minority-class behaviour. The repository's README correctly concludes that positional distribution alone is insufficient to predict syllabic status. This negative result is useful and should be retained.

What the statistics can establish

Sequence length, sign frequencies, bigram restrictions, entropy and positional preferences can show that a corpus is structured. They can compare a sign system with controls and narrow the family of plausible generative processes. Rao and colleagues reported that conditional entropy of Indus sequences falls within a range occupied by several linguistic systems; subsequent debate

showed how results depend on corpus construction, controls and estimator choices. Structured sequence statistics are evidence of constraints, not a translation.

Likewise, a fitted rank-frequency exponent is not a unique classifier of writing-system type. Power-law-like distributions appear in many heterogeneous processes. The repository reports a modified-power-law exponent of -1.13 for its constructed Tamil and cites -1.35 for Indus. Even if both estimates are correct and comparable, two nearby scalar summaries discard sign inventory, genre, chronology, medium, directionality, allography and syntax. A single statistic cannot distinguish inheritance from convergence or deliberate construction.

The README also argues that a mean Indus inscription length near 4.5 signs establishes logo-syllabic writing because a subject-object-verb expression needs at least three signs. That conclusion does not follow. Artefact texts may be labels, names, titles, numerals, commodities, institutions, ownership marks or non-sentential formulae. Mean length neither reveals a sentence structure nor assigns units to linguistic functions.

Language inference and chronology

The Dravidian hypothesis is a serious research hypothesis with a long scholarly history. It is not established by the modern geographic distribution of languages or by genetic ancestry. Population genetics can reconstruct biological ancestry and mixture; it does not encode the language spoken by an archaeological population. The 2019 genome-wide study of South and Central Asia and the Rakhigarhi individual constrain population history, but any linguistic inference requires independent archaeological and historical arguments.

Modern Tamil is also not interchangeable with Proto-Dravidian or the unknown language or languages of the third millennium BCE Indus region. A valid comparison needs explicit historical reconstruction, regular sound correspondences, a defensible account of semantic change, and controls against other Dravidian branches and unrelated languages. Transforming modern Tamil news into integer signs builds in Tamil morphology and analyst choices; resemblance to the resulting distribution is therefore not independent confirmation of Tamil.

Decipherment criteria

A persuasive decipherment normally accumulates several independent successes:

- stable sign identification, direction and segmentation across an openly inspectable corpus;
- phonetic or semantic values derived by stated rules rather than fitted case by case;
- repeated readings that match archaeological context, grammar and chronology;
- novel predictions on inscriptions withheld from model construction;
- superiority over rival languages and non-linguistic or synthetic controls;
- robustness to alternative sign lists, damaged texts, variants and duplicate artefacts;
- expert scrutiny by epigraphers, archaeologists, historical linguists and statisticians.

The reviewed repository does not yet present sign-by-sign readings of held-out Indus texts, a bilingual anchor, or a likelihood comparison in which Tamil decisively outperforms plausible alternatives. Its title therefore describes an ambition rather than a demonstrated result.

Table 2: Evidence assessment of major repository claims

Claim	Assessment	Reason
Tamil can be converted into an integer sign stream	**Demonstrated**	The notebooks and output tables document a rule-based construction from a modern Tamil corpus.

Claim	Assessment	Reason
An SVM can recover curated Tamil boundary labels	**Preliminary**	Saved results are non-trivial, but row-level splitting and full-data encoding weaken the generalisation claim.
Constructed Tamil has a reported exponent near a logosyllabic range	**Reported**	The calculation is saved, but the classifier's uniqueness, uncertainty and cross-corpus comparability are not established.
Indus is necessarily logosyllabic	**Unresolved**	Short length and frequency shape do not uniquely identify a writing-system type.
Harappan language was Proto-Dravidian	**Plausible hypothesis**	Linguistic, archaeological and substrate arguments exist, but genetics cannot determine language and consensus is absent.
The project deciphers Indus as Tamil	**Not demonstrated**	There are no validated sign values, translations or held-out contextual predictions.

Reproducibility audit

The repository deserves credit for publishing code, generated tables, figures and explanatory READMEs. Nine main notebooks retain extensive outputs, which permits partial forensic reconstruction. The committed feature table allows the morpheme-boundary result to be examined without the former cloud database.

Full reproduction remains difficult. Many notebooks import a local `config` module and connect to a named PostgreSQL database on an AWS RDS endpoint. Several required tables are not rebuilt from a single scripted workflow. Paths and warning traces expose author-specific Windows environments. Dependency versions and random seeds are incomplete; a notebook contains a saved `NameError`; and there is no continuous test or machine-readable experiment manifest. The shallow public snapshot inspected here also lacks a top-level licence governing the repository code, while bundled data and visual assets have their own licences. Researchers should determine redistribution permissions for each source before repackaging a corpus.

The notebooks' state is itself a record, not proof that all cells execute in order on a clean machine. Reproducible publication would provide a frozen environment, a make-like entry point, immutable input checksums, schema migrations or local data extracts, an explicit sign-list version, an artefact deduplication procedure, and regenerated tables from a clean run.

A stronger experimental design

Preregistered hypotheses and controls

The first task should be phrased as discrimination rather than confirmation: does an explicitly reconstructed Dravidian model predict unseen Indus sequences or contextual labels better than matched alternatives? Candidate models should include multiple Dravidian reconstructions, other typological language families, known logosyllabic systems after matched truncation, heraldic or emblematic sequences, and synthetic Markov processes. All corpora should be equalised for sample size, sequence length and sign inventory where the statistic requires it.

Leakage-resistant partitions

Deduplicate artefacts and impressions before analysis. Split by archaeological object, site and chronological phase. For Tamil segmentation, group all characters from one word and all occurrences of the same word type in a single fold; a stronger test holds out entire documents. Fit

vocabulary encoders and scalers within each training fold. Select hyperparameters using nested cross-validation, then evaluate once on a sealed test set.

Uncertainty and robustness

Report bootstrap intervals for entropy, power-law exponents and classifier metrics. Repeat the analysis across alternative sign lists and segmentation rules. Use ablations to measure how much each analyst choice contributes: proper-name handling, lemma replacement, morpheme inventory, CV/VC decomposition, punctuation removal and inscription direction. Correct for multiple comparisons and publish null results.

From resemblance to reading

Any proposed sign value should make risky predictions. Before revealing a held-out inscription, specify its expected sign values, grammatical pattern and context-sensitive interpretation. A successful model should reuse values consistently and explain anomalies without proliferating exceptions. Evaluation must be performed by specialists who did not tune the model.

Conclusion

The Kee2u repository is valuable as an open exploratory project: it makes a complex chain of Tamil preprocessing and statistical experiments inspectable, reports a useful negative result, and invites better tests. The public evidence does not support the stronger claim that the Indus script has been deciphered as Tamil. Its synthetic Tamil representation is shaped by the target analogy; its machine-learning score concerns modern Tamil segmentation; and its aggregate statistics are not language-specific.

The productive next step is not a more elaborate pattern match on the same construction. It is a preregistered comparison in which competing language and non-language models face the same corpus, leakage controls and held-out archaeological predictions. Such a design could strengthen or weaken the Dravidian hypothesis. Either outcome would advance knowledge.

References

[1] Jayakumar, K. and Soniaprogram. *Deciphering the Indus Valley Script*. GitHub repository, audited commit `73b0b9fe2acb1106f8b1b017c7350d35fa798997` (2021). https://github.com/Kee2u/Deciphering_the_Indus_Valley_Script

[2] Rao, R. P. N. et al. Entropic evidence for linguistic structure in the Indus script. *Science* 324, 1165 (2009). doi:10.1126/science.1170391.

[3] Rao, R. P. N. et al. Entropy, the Indus script, and language: a reply to R. Sproat. *Computational Linguistics* 36, 795-805 (2010). https://aclanthology.org/J10-4016/

[4] Sproat, R. Ancient symbols, computational linguistics, and the reviewing practices of the general science journals. *Computational Linguistics* 36, 349-355 (2010). https://aclanthology.org/J10-3012/

[5] Farmer, S., Sproat, R. and Witzel, M. The collapse of the Indus-script thesis. *Electronic Journal of Vedic Studies* 11(2), 19-57 (2004). doi:10.11588/ejvs.2004.2.620.

[6] Yadav, N., Vahia, M. N., Mahadevan, I. and Joglekar, H. A statistical analysis of the Indus script using n-grams. arXiv:0901.3017 (2009). https://arxiv.org/abs/0901.3017

[7] Mahadevan, I. *The Indus Script: Texts, Concordance and Tables*. Archaeological Survey of India (1977). https://www.harappa.com/content/indus-script-texts-concordance-and-tables

[8] Parpola, A. *Deciphering the Indus Script*. Cambridge University Press (1994). https://researchportal.helsinki.fi/en/publications/deciphering-the-indus-script/

[9] Narasimhan, V. M. et al. The formation of human populations in South and Central Asia. *Science* 365, eaat7487 (2019). doi:10.1126/science.aat7487.

[10] Shinde, V. et al. An ancient Harappan genome lacks ancestry from Steppe pastoralists or Iranian farmers. *Cell* 179, 729-735.e10 (2019). doi:10.1016/j.cell.2019.08.048.

[11] Ramasamy, L. and Za ̌bokrtský, Z. Tamil dependency parsing: results using rule based and corpus based approaches. *CICLing 2011*, 82-95 (2011). Tamil Dependency Treebank v0.1, CC BY-NC-SA 3.0. https://ufal.mff.cuni.cz/tamiltb

[12] Sommerschield, T. et al. Machine learning for ancient languages: a survey. *Computational Linguistics* 49, 703-747 (2023). doi:10.1162/coli_a_00481.

[13] Luo, J. et al. Deciphering undersegmented ancient scripts using phonetic prior. *Transactions of the ACL* 9, 69-81 (2021). doi:10.1162/tacl_a_00354.

Method note: Counts and model metrics were independently extracted from the repository files and saved notebook outputs. This review did not use the former private cloud database and does not claim a clean-room rerun of every notebook. Links were checked during preparation in September 2026.