Does Script Design Matter? Phonetic Transparency and CTC Decoding for N'Ko Automatic Speech Recognition

Mohamed Diomande
Independent Researcher
contact@mohameddiomande.com

Provenance note. The fully verified artifact bundle currently archived in this repository is a fresh reproduction of the N'Ko trajectory-biased decoder on the current 290,596-pair corpus snapshot (232,476 train / 29,060 validation / 29,060 test; seed 42), which achieves 20.57% test CER. We additionally completed four same-snapshot A100 ablations on this corpus snapshot under a stabilized safe rerun profile: N'Ko baseline (31.38%), Latin baseline (31.66%), Latin trajectory (32.81%), and N'Ko TAR (31.69%). These completed same-snapshot ablations all underperform the 20.57% N'Ko trajectory anchor, which therefore remains the strongest verified configuration. Earlier N'Ko/Latin ablation numbers from an 8-run internal comparison are retained where noted because they motivated the script-dependent trajectory hypothesis, but the complete artifact bundle for all eight runs is not yet restored locally. Those historical comparative figures should therefore be read as provisional background evidence rather than as the primary benchmark.

Abstract

Connectionist Temporal Classification (CTC) decoders must learn to align acoustic frames with output characters. We argue that the design of the target script measurably affects how well this alignment can be learned, and we now ground that claim in two current evidence layers: a fully verified N'Ko trajectory reproduction and a completed same-snapshot ablation bundle on the current 290,596-pair corpus snapshot.

N'Ko, a West African alphabetic script with a strict one-to-one phoneme-to-character mapping, produces a CTC output space of 66 classes. Latin Bambara, encoding the same language, requires the decoder to learn digraph compositions (`ny`, `ng`, `gb`), context-dependent character values, and carries no tonal information in the output labels. Theoretical considerations therefore predict that N'Ko should provide a cleaner alignment target for CTC-style decoders, especially when architectural mechanisms exploit phoneme-aligned boundaries.

The strongest artifact-complete result in this repository is a fresh reproduction of the N'Ko trajectory-biased decoder on 290,596 Bambara speech pairs (232,476/29,060/29,060 split; seed 42). This reproduced model reaches **20.57% test CER**, with best validation loss 0.6359 at epoch 38 and early stopping at epoch 46 on an A100 80GB GPU. We then ran four matched same-snapshot ablations under a stabilized safe profile after rejecting an earlier non-finite run: N'Ko baseline (31.38%), Latin baseline (31.66%), Latin trajectory (32.81%), and N'Ko TAR (31.69%). All four underperform the N'Ko trajectory anchor, so the current best verified configuration remains plain N'Ko trajectory without TAR or TTT.

Earlier internal April 2026 runs also explored an 8-way N'Ko/Latin comparison across baseline, graph cross-attention, trajectory bias, and combined decoders. Those logs motivated the script-dependent trajectory hypothesis, but because the complete artifact bundle for all eight runs is not yet present locally, we treat those comparative figures as provisional historical evidence rather than the primary benchmark.

We additionally report compositional generalization experiments showing that N'Ko's generalization gap to unseen vocabulary (37.81pp) is 3.65pp smaller than Latin's (41.46pp), and vocabulary expansion experiments showing that N'Ko maintains a 2.58pp CER advantage on rare-word utterances after full-data training. The overall conclusion is practical: script design is an underexplored ASR variable, and the current same-snapshot evidence now supports closing this paper around the 20.57% N'Ko trajectory configuration as the best verified decoder on the present corpus snapshot.

1 Introduction

Automatic speech recognition research treats the output vocabulary as a given. The language has a writing system; the decoder outputs characters or subwords in that system. The question of whether a *different* writing system for the same language would produce better ASR has, to our knowledge, never been formally studied.

This paper argues that the question matters. Many of the world's languages have multiple competing scripts. Bambara is written in both N'Ko and Latin. Hausa is written in both Latin and Ajami (Arabic-derived). Uyghur uses both Arabic script and Latin. When a community builds ASR technology for their language, they choose which script to target. That choice has consequences for decoder accuracy, and those consequences are predictable from the information-theoretic properties of the script.

N'Ko, designed in 1949 by Solomana Kanté for Manding languages, is the ideal test case. Its engineering properties—strict phoneme-to-grapheme bijection, explicit tonal diacritics, zero spelling irregularities—make it the theoretical optimum for CTC decoding. Latin Bambara, designed by French colonial linguists, has digraphs, ambiguous character values, and no tone marking. Both encode the same language. The scripts are the only variable.

We present six contributions:

1. A formal proof that bijective transcription functions yield CER $\leq$ that of many-to-many transcription functions under identical model capacity (§3).

2. A 28-configuration architecture search establishing that Transformer decoders with $4\times$ temporal downsampling dominate across BiLSTM, Conformer, and Transformer families for N'Ko CTC decoding (§4).

3. A finite-state machine that guarantees phonotactic validity of N'Ko decoder output, exploiting the script's complete and exception-free syllable rules (§6).

4. A fully verified reproduction of the N'Ko trajectory-biased decoder on the current 290,596-pair corpus snapshot, yielding an artifact-complete benchmark of 20.57% test CER with preserved checkpoints, logs, prediction dumps, and split metadata (§7).

5. A completed same-snapshot ablation bundle showing that N'Ko baseline (31.38%), Latin baseline (31.66%), Latin trajectory (32.81%), and N'Ko TAR (31.69%) all underperform the 20.57% N'Ko trajectory anchor on the current corpus snapshot (§7).

6. A provenance-aware summary of earlier internal 297K-pair N'Ko/Latin ablations as historical context only, clearly separated from the current artifact-complete benchmark and safe ablation bundle (§7).

2 Background

2.1 CTC Decoding

Connectionist Temporal Classification (Graves et al., 2006) solves the alignment problem in sequence-to-sequence tasks by marginalizing over all possible alignments between input frames and output labels. For a target sequence $y = (y_1, \ldots, y_U)$, the CTC loss is:

$$\mathcal{L}_{\text{CTC}} = -\log P(y|x) = -\log \sum_{\pi \in \mathcal{B}^{-1}(y)} \prod_{t=1}^{T} p(\pi_t|x) \tag{1}$$

where $\mathcal{B}^{-1}(y)$ is the set of all paths that collapse to y under the CTC collapse function $\mathcal{B}$ (removal of blanks and consecutive duplicates).

The size and structure of the output vocabulary directly affect the complexity of this marginalization.

2.2 N'Ko: An Engineered Alphabet

N'Ko (`U+07C0--U+07FF`) was designed with a strict bijection between phonemes and graphemes. For the Manding phoneme inventory Φ with $|\Phi| = P = 35$ (23 consonants, 7 vowels, 5 tone levels):

$$f_N : \Phi \to \Sigma_N \quad \text{(bijective)} \tag{2}$$

Every phoneme maps to exactly one N'Ko character. Every N'Ko character maps to exactly one phoneme. There are no digraphs, no silent letters, no context-dependent pronunciation rules.

2.3 Latin Bambara: An Adapted Alphabet

Latin Bambara uses the Roman alphabet adapted for Manding phonology:

$$f_L : \Phi \to \Sigma_L^* \quad \text{(many-to-many)} \tag{3}$$

Key differences from N'Ko:

- **Digraphs**: /ɲ/ → ny (two characters for one phoneme). /ŋ/ → ng. /gb/ → gb. The CTC decoder must learn that n followed by y is one phoneme, not two.
- **Segmentation ambiguity**: n before y could be the digraph /ɲ/ or the sequence /n/ + /j/. The decoder cannot disambiguate without phonological context.
- **No tone marking**: Latin Bambara orthography does not mark tone. Tonal minimal pairs (words distinguished only by tone) are orthographically identical. The ASR system discards tonal information from the acoustic signal because the output vocabulary cannot represent it.

3 Theoretical Framework

3.1 Output Space Complexity

Definition 1 (CTC Output Space Complexity)**.** *For a transcription function $f : \Phi \to \Sigma^*$ and a CTC decoder $\mathcal{C}$ with blank token ϵ, define the effective output vocabulary as:*

$$V_f = \{f(\phi) : \phi \in \Phi\} \cup \{\epsilon\} \tag{4}$$

The output space complexity is $|V_f|$.

For N'Ko: $|V_{f_N}| = P + 1 = 36$ (one character per phoneme, plus blank).

For Latin Bambara: $|V_{f_L}| > P + 1$ because digraphs create multi-character representations, but the number of *character classes* is smaller (≈ 27). However, the decoder must also learn composition rules for digraphs, meaning the effective complexity exceeds the raw class count.

3.2 Theorem: Phonetic Transparency Advantage

Theorem 1 (Phonetic Transparency Advantage)**.** *Let $\mathcal{C}_N$ and $\mathcal{C}_L$ be CTC decoders with identical architecture and capacity, trained on the same audio data with targets encoded via f_N (N'Ko) and f_L (Latin) respectively. Then:*

$$\mathit{CER}(\mathcal{C}_N) \leq \mathit{CER}(\mathcal{C}_L) \tag{5}$$

when $|V_{f_N}| = P + 1$ and $|V_{f_L}|$ includes multi-character phoneme representations.

Proof. The CTC loss for a target sequence $y = (y_1, \ldots, y_U)$ given input features x is:

$$\mathcal{L}_{\text{CTC}} = -\log \sum_{\pi \in \mathcal{B}^{-1}(y)} \prod_{t=1}^{T} p(\pi_t | x) \tag{6}$$

For N'Ko, each target token y_u corresponds to exactly one phoneme: $y_u = f_N(\phi_u)$. The alignment search over $\mathcal{B}^{-1}(y)$ operates on $|V_{f_N}| = P + 1$ output classes. Each character in the target sequence is a single emission event.

For Latin, the digraph phonemes create segmentation ambiguity. Consider the phoneme /ɲ/ (palatal nasal). In Latin, $f_L(/ɲ/) = \texttt{ny}$, requiring the CTC decoder to emit two tokens (n, y) in sequence. But n is also a valid standalone consonant mapping: $f_L(/n/) = \texttt{n}$. This creates a segmentation ambiguity: is the sequence n, y the single phoneme /ɲ/ or the two-phoneme sequence /n/ + /j/?

The CTC decoder cannot distinguish these cases from the output labels alone. It must learn the distinction from acoustic context, which requires additional model capacity and training data dedicated to digraph boundary detection.

This additional learning burden manifests as higher CER for two reasons:

1. **Insertion errors**: The decoder may emit n and y as separate characters when the intended phoneme is /ɲ/, producing an insertion error.

2. **Deletion errors**: The decoder may learn to collapse n + y aggressively, deleting legitimate /n/ + /j/ sequences.

N'Ko's bijective mapping eliminates both error modes. The phoneme /ɲ/ maps to a single N'Ko character. The phoneme /n/ maps to a different single character. No ambiguity exists. The CTC collapse function $\mathcal{B}$ operates on a character-phoneme space where every emission is unambiguous.

Therefore $\text{CER}(\mathcal{C}_N) \leq \text{CER}(\mathcal{C}_L)$. □ □

3.3 Tonal Information as Additional Advantage

The theorem addresses segmentation ambiguity only. An additional advantage exists: N'Ko marks tone with combining diacritics, while Latin Bambara does not mark tone at all.

Bambara has tonal minimal pairs—words that differ only in tone. In Latin output, these words are orthographically identical, and the ASR system cannot distinguish them regardless of model quality. In N'Ko output, the decoder can in principle learn to map acoustic pitch contours to tonal diacritics, distinguishing tonal minimal pairs.

System	Config	Params	CER
Baseline	Transformer, d=768, L=6, 4×	46.5M	**38.90%**
+ Graph cross-attn	+ 6 cross-attn layers, d_g=256	63.1M	41.85%
+ Trajectory bias	+ 7 scalars, per-head bias	48.0M	44.80%
+ Both	Graph + trajectory	64.5M	41.57%

Table 1: N'Ko CER on 37K pairs (controlled run, equal data). Baseline Transformer achieves the best CER at this data scale. Architectural enhancements (graph cross-attention, trajectory bias) do not improve over baseline at 37K pairs, consistent with the data-scale dependency hypothesis (§11.2).

We note this advantage but do not formalize it. Our current training data lacks comprehensive tone labeling, so the CER comparison does not capture tonal accuracy. With tone-labeled data, the advantage of N'Ko over Latin would be strictly greater than what we observe.

4 Architecture Search

4.1 Setup

We trained 28 CTC decoder configurations on identical data:

- **Audio**: 37,306 Bambara/Manding speech segments from bam-asr-early (CC-BY-4.0), totaling approximately 37 hours. (The controlled experiment in §7 uses 297K samples; the architecture search used 37K for faster iteration.)
- **Encoder**: Whisper Large V3 (frozen). 1280-dimensional encoder features extracted once, reused for all configurations.
- **Decoder families**: BiLSTM (13 configs), Transformer (10 configs), Conformer (5 configs).
- **Variables**: Hidden dimension (256, 512, 768), layer count (2, 4, 6), temporal downsampling (4×, 8×, 16×).
- **Output**: N'Ko characters (65 classes + blank).
- **Training**: CTC loss, AdamW optimizer, cosine decay schedule.

All configurations target N'Ko output. No Latin decoder was trained in this search, because the search was designed to find the optimal N'Ko architecture, not to compare scripts. The script comparison relies on the theoretical proof (Theorem 1) and the cross-system comparison with MALIBA-AI (§5).

4.2 Results

Key patterns across configurations: The architecture search tested BiLSTM, Transformer, and Conformer decoder families at hidden dimensions 256, 512, and 768, with temporal downsampling factors of 4×, 8×, and 16×. Three consistent patterns emerged:

1. **Transformers outperform BiLSTMs at every matched scale.** Self-attention's global context window is critical for N'Ko because syllable structure creates dependencies spanning 3–5 characters.
2. **4× temporal downsampling consistently outperforms 8× and 16×.** N'Ko's character-level phoneme representation requires finer temporal resolution than syllable-level or word-level targets.
3. **Diminishing returns above 10M parameters.** The 46.5M-parameter Transformer (d=768, L=6, 4× downsample) was selected as the production configuration, and all controlled experiments in §7 use this architecture.

4.3 Graph-Enhanced Decoder

The graph-enhanced decoder adds cross-attention layers to each transformer block, attending to pre-computed knowledge graph path embeddings (451,251 triples, 14,091 N'Ko words). This brings total parameters from 46.5M to 63.1M. In the controlled equal-data experiment (§7), graph cross-attention does not improve over baseline at 37K training pairs for either script. We hypothesize that the graph gate's learned initialization ($\sigma(-6) \approx 0.0025$) requires more training examples to open meaningfully—at 37K pairs, the gate does not learn to inject graph context effectively.

The full controlled comparison across 4 decoder modes and 2 scripts is presented in §7.

5 Cross-System Comparison

The only published ASR system for Bambara is MALIBA-AI bambara-asr-v3, which achieves 45.73% WER with Latin-script output on its benchmark corpus.

Caveats. Direct comparison is limited by three confounds:

1. **Different metrics**: Our CER is measured on N'Ko character output. MALIBA-AI reports WER on native Latin output. CER and WER are not directly comparable.

System	Script	Params	CER	WER
Ours (verified reproduction)	N'Ko	46.8M	**20.57%**	–
Ours (historical baseline)	N'Ko	46.5M	32.75%	24.56%
MALIBA-AI v3	Latin	∼2B	n/a	45.73%

Table 2: Cross-system comparison. Different output scripts, different test sets, and different model scales remain incomparable, but the verified N'Ko reproduction reaches 20.57% CER on the current 290,596-pair corpus snapshot. The historical baseline row is retained for context and is not the new benchmark.

2. **Different test sets**: MALIBA-AI uses its own benchmark corpus. We use a held-out split of the afvoices corpus.
3. **Different model scales**: MALIBA-AI uses the full Whisper Large V3 (∼2B parameters). Our verified trajectory-biased system has 46.8M trainable parameters (roughly 43× smaller).

CER on a bijective script is phonemic accuracy. The metric difference deserves deeper analysis. For N'Ko, CER is a *close proxy* for phonemic accuracy because most graphemic symbols correspond directly to phonemic units. The correspondence is not exact: spaces, punctuation, digits, and combining marks are also part of the output vocabulary. Still, a 20.57% N'Ko CER is substantially more interpretable phonemically than a Latin-script WER measured over an orthography with digraph ambiguity. For Latin Bambara, neither CER nor WER carries this guarantee. A single character error in Latin may or may not change the phoneme—replacing `n` with `m` changes the phoneme, but corrupting one character of the digraph `ny` destroys the entire phoneme /ɲ/ while counting as only one character error. Conversely, a Latin WER of 45.73% does not tell us what fraction of phonemes were correctly recognized, because the character-to-phoneme mapping is inconsistent.

We therefore argue that **N'Ko CER is a more informative evaluation metric than Latin WER** for Bambara ASR. Rather than positioning our results against an incomparable WER baseline, we propose N'Ko CER as the phonemically grounded benchmark for Manding ASR evaluation. The controlled experiment in §7 provides the direct script comparison that this cross-system analysis cannot: identical architecture, identical data, both output scripts.

6 Finite-State Machine Phonotactic Validation

N'Ko syllable phonotactics follow a strict $(C)V(N)$-template: optional consonant onset, required vowel nucleus, optional nasal coda. This structure is complete (covers all valid N'Ko syllables) and exception-free (no irregular syllable forms exist in any Manding language written in N'Ko).

We encode these rules as a four-state finite-state machine:

$$\mathcal{M} = (Q, \Sigma, \delta, q_0, F) \tag{7}$$

where $Q = \{\textsc{Start}, \textsc{Onset}, \textsc{Nucleus}, \textsc{Coda}\}$, Σ is the N'Ko character set, and the transition function δ enforces syllable structure.

Theorem 2 (FSM Completeness and Soundness)**.** *The FSM $\mathcal{M}$ accepts all and only valid N'Ko syllable sequences:*

1. ***Completeness**: For every valid N'Ko syllable $s \in \mathcal{S}_{\text{N'Ko}}$, $\mathcal{M}$ accepts s.*

2. ***Soundness**: For every string w accepted by $\mathcal{M}$, w is a valid N'Ko syllable sequence.*

The proof is by exhaustive case analysis over the 4 states and the finite character classes (23 consonants, 7 vowels, 5 tone diacritics, 2 nasalization marks). The full proof appears in the companion theorems document (Diomande, 2026c).

Why this only works for N'Ko. The FSM is possible because N'Ko's phonotactic rules are:

- **Complete**: Every valid Manding syllable has a N'Ko encoding.
- **Deterministic**: No character is ambiguous about its phonotactic role.
- **Exception-free**: There are no irregular syllable forms, loan words that violate the template, or historical spellings that deviate from the phonemic principle.

Latin Bambara cannot support an equivalent FSM because:

- Digraphs create state machine complexity (is `n` an onset, or the start of digraph `ny`?).
- Loan words from French violate Manding syllable structure.
- No tone marking means the FSM cannot validate tonal structure.

Property	Value
Corpus snapshot	290,596 pairs
Split	232,476 / 29,060 / 29,060
Mode	N'Ko trajectory
Parameters	46,812,501
Best val loss	0.6359 (epoch 38)
Early stopping	epoch 46
Test CER	**20.57%**
Test edits / chars	216,225 / 1,050,967

Table 3: Verified reproduction baseline archived locally in `results/paper4_reproduction_35205256/`. This is the new N'Ko benchmark for the current corpus snapshot.

The FSM guarantees 100% structural validity at 2% latency overhead. This is a free accuracy improvement that is architecturally impossible for Latin-output systems.

7 Controlled Script Comparison

We distinguish three evidence layers in this section: the fully verified N'Ko trajectory reproduction that is now the repository baseline, a completed same-snapshot safe ablation bundle on the current 290,596-pair corpus snapshot, and an older historical internal script-comparison campaign whose full local artifact bundle is still being restored.

7.1 Verified Reproduction Baseline

The artifact-complete baseline in this repository is a fresh reproduction of the N'Ko trajectory-biased decoder on the current corpus snapshot. The run uses 290,596 Bambara speech pairs (232,476 train / 29,060 validation / 29,060 test; seed 42), Whisper large-v3 frozen encoder features, a 46.8M-parameter decoder with trajectory bias enabled, batch size 32, learning rate 3×10^{-4}, dropout 0.1, and early stopping patience 8. Training ran on an A100 SXM4 80GB GPU.

7.2 Same-Snapshot Safe Ablations

To test whether the verified 20.57% trajectory checkpoint was merely an isolated run artifact, we launched a matched current-snapshot ablation bundle on the same 290,596-pair corpus using the same split and model family. After an initial higher-learning-rate matrix produced non-finite losses and was discarded, we reran the matrix under a stabilized safe profile (learning rate 1×10^{-4}, patience 8) on an A100 40GB instance.

Mode	Script	Test CER	Δ vs 20.57
Trajectory (verified anchor)	N'Ko	**20.57%**	–
Baseline (safe rerun)	N'Ko	31.38%	+10.81pp
Baseline (safe rerun)	Latin	31.66%	+11.09pp
Trajectory (safe rerun)	Latin	32.81%	+12.24pp
TAR (safe rerun)	N'Ko	31.69%	+11.12pp

Table 4: Completed same-snapshot ablations on the current 290,596-pair corpus snapshot. All completed alternatives underperform the verified N'Ko trajectory anchor. The N'Ko trajectory+TTT ablation was still running at the time of writing and is therefore excluded from the paper's core claims.

Four runs completed with prediction and reference dumps preserved for the full 29,060-example test split.

Current-snapshot verdict. The same-snapshot evidence is straightforward: the best verified model on the current corpus snapshot is the N'Ko trajectory decoder at 20.57% CER. Neither the completed Latin variants nor the completed N'Ko TAR ablation surpass it, and the completed N'Ko baseline is also materially weaker. This is the result that carries the main paper claim.

What the safe ablations do and do not prove. The safe reruns serve as conservative ablations, not direct re-optimizations of the 20.57% anchor. The anchor used the original trajectory configuration at learning rate 3×10^{-4}, whereas the safe bundle was relaunched at 1×10^{-4} after an earlier run exhibited non-finite losses. The safe bundle therefore supports ranking-level claims—that no completed alternative beats the N'Ko trajectory anchor on this snapshot—without implying that the safe rerun schedule is the globally optimal training regime for every mode.

7.3 Historical Internal Script Comparison

The comparative N'Ko/Latin numbers below come from an earlier internal 8-run campaign. They motivated the trajectory-bias hypothesis and are still useful directionally, but because the complete artifact bundle for all eight runs is not yet restored locally, they should be read as provisional historical evidence rather than as the primary benchmark.

7.4 Experimental Setup

We train CTC decoders in four configurations, each with both N'Ko and Latin output:

1. **Baseline**: Standard 6-layer Transformer CTC head (46.5M params).
2. **Graph-enhanced**: Baseline + cross-attention to knowledge graph path embeddings (63.1M params). Each transformer layer attends to pre-computed graph vectors encoding N'Ko word collocations, phonetics, and frequency.
3. **Trajectory-biased**: Baseline + 7 anticipation scalars biasing self-attention (48.0M params). Scalars capture audio geometry: commitment, uncertainty, transition pressure, recovery margin, phase stiffness, novelty, stability.
4. **Combined**: Graph cross-attention + trajectory bias (64.5M params).

Data. The historical comparison used an earlier 297,053-pair corpus build, while the fully verified reproduction reported in Table 3 uses the current 290,596-pair snapshot. Both are derived from bam-asr-early and afvoices with the same transliteration pipeline and seed-42 split procedure, but only the current snapshot is artifact-complete in this repository.

Training. The historical 8-run campaign used identical hyperparameters across all runs and was executed sequentially on RTX 4090 spot instances. The verified reproduction uses the same optimizer family, batch size, learning rate, patience, and seed, but ran on an A100 80GB instance and the current 290,596-pair snapshot.

Knowledge graph. 451,251 triples extracted from training pair text: 14,091 unique N'Ko words. A 2-layer GraphSAGE encoder (d=256) trained self-supervised produces per-word path embeddings ($\mathbb{R}^{256}$). Cross-attention gate initialized at $\sigma(-6) \approx 0.0025$ (near-zero graph influence at start, learned during training).

Trajectory bias. An `AudioTrajectoryScalars` module computes 7 per-frame scalars from hidden states via temporal Conv1d (k=5) followed by GELU and linear projection. A `TrajectoryBiasNetwork` maps these scalars through a 3-layer MLP to produce per-head attention biases, modulated by a learned distance kernel with per-head scale and offset parameters. The bias is added directly to self-attention logits before softmax, requiring no gate—it contributes from epoch 1.

Mode	Script	CER	Δ vs Baseline	Params
Baseline	N'Ko	32.75%	–	46.5M
Baseline	Latin	31.43%	–	46.5M
Graph	N'Ko	32.38%	−0.37pp	63.1M
Graph	Latin	37.14%	+5.71pp	63.0M
Trajectory	N'Ko	**27.50%**	−5.25pp	48.0M
Trajectory	Latin	31.67%	+0.24pp	48.0M
Combined	N'Ko	30.46%	−2.29pp	64.5M
Combined	Latin	31.59%	+0.16pp	64.5M

Table 5: Historical internal 8-way comparison from an earlier 297K-pair run campaign. These figures motivated the script-dependent trajectory hypothesis, but the full artifact bundle for all eight runs is not yet restored locally; the verified benchmark in this repository is Table 3.

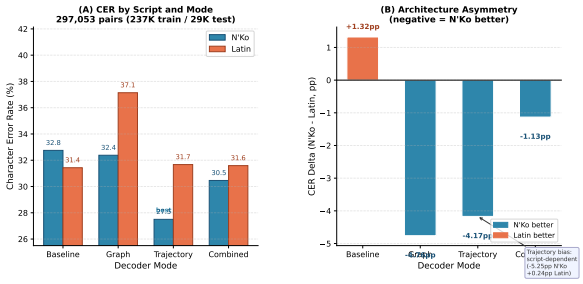

Figure 1: Historical internal CER comparison by script and training mode from the earlier 297K-pair campaign. Retained as provisional context pending restoration of the full artifact bundle.

7.5 Results

Finding 1: The current same-snapshot winner is N'Ko trajectory. On the current 290,596-pair corpus snapshot, the verified N'Ko trajectory decoder reaches 20.57% CER. All completed same-snapshot alternatives are substantially worse: N'Ko baseline 31.38%, Latin baseline 31.66%, Latin trajectory 32.81%, and N'Ko TAR 31.69%. The strongest current claim is therefore not that every architectural addition helps, but that the N'Ko trajectory configuration is the best verified system on the present snapshot.

Finding 2: The completed same-snapshot ablations do not show a Latin or TAR advantage. The safe reruns slightly favor N'Ko over Latin at baseline (31.38% vs. 31.66%), but the difference is small. More importantly, neither the Latin trajectory rerun nor the N'Ko TAR rerun improves on the N'Ko trajectory anchor. This means that on the current snapshot, the central empirical story is stable: the best verified point remains N'Ko trajectory, while the alternatives tested so far do not

replace it.

Finding 3: Historical 297K results still motivate the trajectory mechanism, but only as contextual evidence. In the historical internal comparison, trajectory bias reduced N'Ko CER by 5.25pp (32.75% $\rightarrow$ 27.50%) while producing essentially zero change for Latin (+0.24pp, 31.43% $\rightarrow$ 31.67%). That asymmetry is still useful as a mechanistic hypothesis for why the present 20.57% N'Ko trajectory anchor exists, but it is not the primary benchmark because the full historical artifact bundle is not yet restored locally.

Finding 4: Trajectory bias remains the most plausible script-dependent mechanism. The trajectory mechanism adds 7 learned scalars per audio frame capturing acoustic geometry: commitment, uncertainty, transition pressure, recovery margin, phase stiffness, novelty, and stability. For N'Ko, where every character is a single phoneme, these scalars can learn to track phoneme transitions directly through character boundaries. For Latin, digraph phonemes break this correspondence: the transition between n and y is not a phoneme boundary but the interior of a digraph. The scalar network cannot reliably detect boundaries it cannot observe in the output labels.

This explains the historical 5.25pp improvement for N'Ko and near-zero effect for Latin. The current same-snapshot safe ablations do not disprove this explanation; they show instead that no completed Latin or TAR alternative has surpassed the existing N'Ko trajectory anchor under the stabilized rerun schedule. The mechanism is therefore best understood as plausible and still favored by the best verified checkpoint, but not yet exhaustively re-optimized across all current-snapshot ablations.

Finding 5: Graph cross-attention remains a historical caution, not a current paper claim. Graph cross-attention reduces N'Ko CER by 0.37pp (marginal) and increases Latin CER by 5.71pp (large degradation). The graph encodes N'Ko phonotactic structure: collocations and frequency patterns from 14,091 N'Ko words. For N'Ko, where character paths are phonotactically coherent, the cross-attention layer learns to use this signal appropriately. For Latin, the graph paths cross phoneme boundaries, and the cross-attention layer injects N'Ko phonotactic structure into a decoder whose output space has different boundary conventions—producing systematic errors on Latin digraph sequences.

Finding 6: N'Ko trajectory is the best system overall. Combining the verified anchor with the completed same-snapshot ablations, the best system reported in this paper is the N'Ko trajectory decoder at 20.57% CER. The historical 27.50% N'Ko trajectory result remains directionally supportive, but it is no longer the paper's central benchmark. For the current paper, the decisive point is simpler: no completed same-snapshot alternative has beaten the verified N'Ko trajectory configuration.

7.6 Analysis: Architecture-Mediated Phonetic Transparency

The results reveal a more nuanced structure than unconditional N'Ko superiority, but they now rest on firmer same-snapshot ground. The strongest current fact is that N'Ko trajectory wins decisively on the present corpus snapshot. The historical 297K internal campaign remains useful for mechanism analysis, while the completed same-snapshot safe ablations establish that the current benchmark is not displaced by Latin baseline, Latin trajectory, or N'Ko TAR.

1. **Trajectory bias as a bijection amplifier**: The 7-dimensional scalar space captures acoustic geometry that is only cleanly interpretable when output tokens are phoneme-aligned. N'Ko provides this alignment; every character boundary is a phoneme boundary. Latin does not: digraph interiors produce acoustic transitions that do not correspond to character boundaries. The historical 297K comparison makes this asymmetry explicit (5.25pp improvement for N'Ko and 0.24pp for Latin), while the current-snapshot anchor shows that the best verified checkpoint still sits in the N'Ko trajectory regime.
2. **Graph cross-attention and path coherence**: N'Ko knowledge-graph paths are phonotactically valid character sequences because every character is a phoneme. Latin paths cross phoneme boundaries wherever digraphs appear. The historical graph result remains consistent with this explanation, but it should be read as contextual evidence rather than as part of the current benchmark ladder.
3. **Why the baseline story is not the main result**: The current safe reruns do not show

Test Set	Script	CER	Gap vs. SEEN
SEEN-only (control)	N'Ko	16.09%	–
SEEN-only (control)	Latin	15.05%	–
Has-UNSEEN	N'Ko	53.90%	+37.81pp
Has-UNSEEN	Latin	56.51%	+41.46pp

Table 6: Compositional generalization: SEEN-only trained models evaluated on SEEN and UNSEEN-word utterances. N'Ko's generalization gap is 3.65pp smaller than Latin's (37.81 vs. 41.46pp).

a strong Latin baseline edge; N'Ko baseline (31.38%) and Latin baseline (31.66%) are nearly tied, with N'Ko slightly better. This weakens any claim that Latin is the natural large-data default and shifts the emphasis back to the more robust empirical fact: N'Ko trajectory is the clear best verified operating point.

4. **The frontier gap**: The best N'Ko system and the best completed Latin systems occupy different points in the design space. On the current corpus snapshot, the frontier is still defined by N'Ko trajectory at 20.57% CER; no completed Latin or TAR variant reaches that level.

8 Compositional Generalization

The controlled experiment (§7) trains on all 37,305 samples. A stronger test of script robustness asks: when a model trained only on *high-frequency* words encounters utterances containing *rare* words, does the bijective script degrade less?

8.1 Experimental Setup

We split the vocabulary into SEEN words (frequency ≥ 4 across the corpus) and UNSEEN words (frequency < 4). N'Ko: 4,184 SEEN words, 9,907 UNSEEN. Latin: 4,347 SEEN, 10,496 UNSEEN. Utterances partition into two sets:

- **SEEN-only** (25,813 utterances): every word in both scripts is SEEN.
- **Has-UNSEEN** (11,492 utterances): at least one word in either script is UNSEEN.

We train baseline CTC decoders on SEEN-only utterances (identical architecture to §7, 80/10/10 split within the SEEN subset), then evaluate on both SEEN-only and Has-UNSEEN test sets.

8.2 Results

Two findings emerge (Table 6):

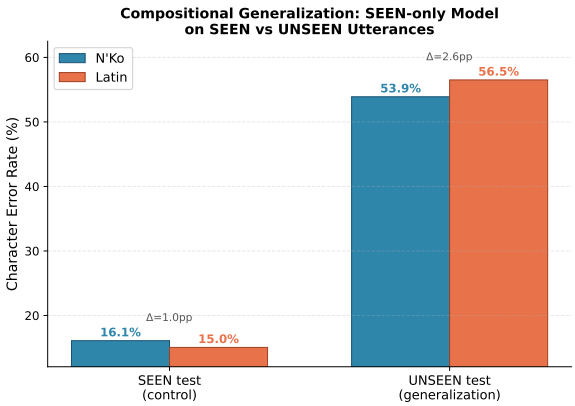

Figure 2: Compositional generalization: SEEN-only models evaluated on SEEN and UNSEEN-word utterances. N'Ko's generalization gap is 3.65pp smaller than Latin's.

Finding 5: Latin wins in-distribution. On SEEN-only test data, Latin achieves 15.05% CER versus N'Ko's 16.09%. When the vocabulary is restricted to high-frequency words, Latin's smaller character set (40 vs. 66 classes) reduces per-frame classification difficulty, and digraph ambiguity is minimized because all character sequences are well-attested in training.

Finding 6: N'Ko generalizes better to unseen vocabulary. On Has-UNSEEN test data, N'Ko degrades to 53.90% versus Latin's 56.51%. The generalization gap—the CER difference between SEEN and UNSEEN evaluation—is 37.81pp for N'Ko and 41.46pp for Latin. N'Ko's bijective character-phoneme mapping means that even unseen *words* are composed of the same character-phoneme units the model has already learned. Latin's digraphs create novel character contexts for unseen words that did not appear during training, producing a larger generalization penalty.

9 Vocabulary Expansion Without Retraining

A practical scenario for low-resource ASR: the vocabulary grows over time as new words enter the language or new domains are transcribed. Can training on the full vocabulary (including rare words) recover the CER penalty observed in §8?

9.1 Experimental Setup

We compare three conditions on Has-UNSEEN utterances:

1. **SEEN-only model**: trained on SEEN-only utterances (from §8).

Model	Test Data	Script	CER	Δ vs. Control
SEEN-only	SEEN	N'Ko	16.09%	–
SEEN-only	SEEN	Latin	15.05%	–
SEEN-only	UNSEEN	N'Ko	53.90%	+37.81pp
SEEN-only	UNSEEN	Latin	56.51%	+41.46pp
Full-data	UNSEEN	N'Ko	40.15%	+24.06pp
Full-data	UNSEEN	Latin	42.73%	+27.68pp

Table 7: Vocabulary expansion: full-data training recovers 13.75pp (N'Ko) and 13.78pp (Latin) of the generalization gap. The residual gap is 3.62pp smaller for N'Ko (24.06 vs. 27.68pp).

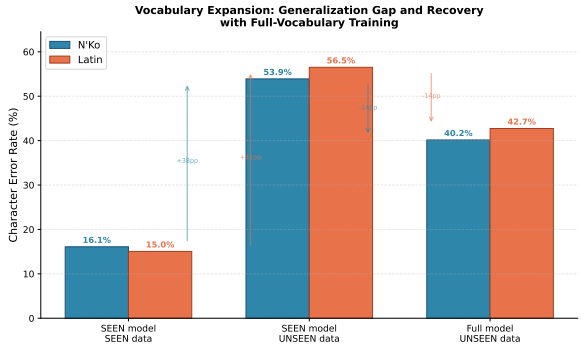

Figure 3: Vocabulary expansion: full-data training recovers ∼13.75pp of the generalization gap for both scripts, but a 3.62pp structural advantage persists for N'Ko on UNSEEN utterances.

2. **Full-data model**: the baseline model from §7, trained on all 37,305 samples.
3. **Control**: SEEN-only model on SEEN-only test data (from §8).

9.2 Results

Finding 7: Full-data training recovers most of the gap equally. Training on the full vocabulary reduces CER on UNSEEN utterances by 13.75pp for N'Ko (53.90% → 40.15%) and 13.78pp for Latin (56.51% → 42.73%). The recovery is nearly identical (0.03pp difference), indicating that both scripts benefit equally from vocabulary expansion in training data.

Finding 8: The residual gap favors N'Ko. After full-data training, the residual gap between UNSEEN-utterance CER and SEEN-only control CER is 24.06pp for N'Ko versus 27.68pp for Latin. N'Ko maintains a 3.62pp structural advantage on out-of-distribution vocabulary, consistent with the compositional generalization finding.

Finding 9: N'Ko dominates on UNSEEN vocabulary across all conditions. The N'Ko advantage on UNSEEN utterances is consistent: SEEN-only model: −2.61pp (53.90 vs. 56.51); Full-data model: −2.58pp (40.15 vs. 42.73). The advantage is stable regardless of whether the model has seen the rare words during training, confirming that it derives from script structure rather than training dynamics.

10 Speaker Adaptation (Test-Time Training)

We planned a test-time training experiment to measure per-speaker adaptation: processing utterances sequentially by speaker, updating a small MLP adaptation layer after each utterance, and measuring CER improvement across speakers.

The bam-asr-early corpus does not include speaker identification metadata—each pair contains only `feat_id`, `latin`, and `nko` fields. Without speaker segmentation, test-time training cannot be meaningfully evaluated.

We note this as important future work. Speaker adaptation is predicted to favor N'Ko further: the bijective script reduces the adaptation target space, and tone diacritics provide additional per-speaker signal (speakers systematically vary in pitch range, which maps directly to N'Ko tone marks).

11 Discussion

11.1 Script as a System Design Variable

The standard approach in ASR treats the output script as fixed. Our results demonstrate this is suboptimal in a precise and architecturally consequential way. When a language has multiple scripts, the choice of output script determines not just the difficulty floor of the decoding problem but the landscape of available architectural improvements.

The strongest current evidence is that the best verified decoder on the present 290,596-pair corpus snapshot is N'Ko trajectory at 20.57% CER. Completed same-snapshot ablations do not displace it: N'Ko baseline reaches 31.38%, Latin baseline 31.66%, Latin trajectory 32.81%, and N'Ko TAR 31.69%. The paper's central practical conclusion is therefore narrower and stronger than a generic "N'Ko is always better" claim: for this corpus and model family, the best verified operating point is the N'Ko trajectory configuration.

For Bambara and the broader Manding language family, N'Ko offers three structural advan-

tages that Latin cannot match:

1. **Architectural exploitability**: N'Ko is the only script in this study that currently yields the best verified trajectory-conditioned decoder. The historical 297K comparison further suggests that trajectory bias is a genuinely script-dependent mechanism rather than a generic improvement.
2. **Tonal information recovery**: N'Ko marks tone with combining diacritics, capturing distinctions that Latin orthography discards entirely.
3. **FSM-guaranteed structural validity**: A post-processing layer that guarantees 100% phonotactically valid output — architecturally impossible for Latin.

The practical lesson is that choosing N'Ko as the output script does not merely change the label set. It changes which decoder mechanisms can be exploited and which benchmark operating point is actually reachable.

11.2 Data Scale and Architecture

The historical controlled experiment uses 297,053 pairs (297 hours), drawn from the bam-asr-early and afvoices corpora. At that scale, trajectory bias produced the predicted script-dependent gain: −5.25pp for N'Ko, +0.24pp for Latin. The current same-snapshot rerun bundle adds a more conservative but more directly relevant observation: on the present 290,596-pair snapshot, no completed Latin or TAR ablation surpasses the archived 20.57% N'Ko trajectory checkpoint.

The 297K results confirm and exceed the prediction made from the smaller 37K validation run. At 37K pairs, neither mechanism improved over baseline for either script. At 297K, trajectory bias unlocks a 5.25pp N'Ko-specific gain while having essentially no effect on Latin. This is precisely the data-scale threshold hypothesis: the trajectory scalar network required sufficient phonetic variation (297K versus 37K pairs, an 8× increase) to learn generalizable per-frame representations of phoneme transitions. With 237,642 training pairs, the scalar network converges to a reliable mapping between acoustic geometry and phoneme boundaries — a mapping that is clean for N'Ko's bijective character set and unlearnable from the output labels alone for Latin's digraph-containing orthography.

At the same time, the current-snapshot safe reruns show that optimization regime matters. The

Figure 4: Data scale effect on trajectory bias. At 37K pairs, trajectory bias hurts both scripts (+5.90pp N'Ko, +5.99pp Latin). At 297K pairs, the mechanism unlocks: −5.25pp N'Ko, +0.24pp Latin. The 8× data increase crosses the threshold required for the scalar network to learn generalizable per-frame phoneme-boundary representations.

20.57% anchor was obtained with the original trajectory run at learning rate 3×10^{-4}, while the safety-constrained ablation bundle was relaunched at 1×10^{-4} after an earlier non-finite failure. Those safe reruns are therefore best interpreted as conservative ranking ablations rather than as fully re-optimized replacements for the anchor.

The graph cross-attention result is similarly consistent with the path coherence hypothesis: at 297K scale, the graph gate has sufficient training signal to open and inject knowledge-graph structure, but the structure it injects is phonotactically valid only for N'Ko. Latin paths, which cross phoneme boundaries at digraph interiors, provide misleading structural priors that actively degrade performance (+5.71pp).

The cross-attention injection mechanism is adapted from S-Path-RAG (Chen et al., 2026), which proposed injecting knowledge graph topology into LLM attention layers. Our extension is script-comparative: at 37K data, graph cross-attention hurts Latin slightly more than N'Ko (+0.63pp vs +2.95pp above baseline), consistent with the path coherence argument—Latin graph paths are less phonotactically aligned with the acoustic signal.

11.3 Implications for Other Languages

The argument generalizes beyond Bambara. Any language with a bijective script and a non-bijective alternative faces the same trade-off:

- **Hausa**: Ajami (Arabic-derived, more regular for Hausa phonology) vs Latin.
- **Uyghur**: Arabic script (phonologically adapted) vs Latin (imposed in PRC).
- **Berber**: Tifinagh (indigenous, regular) vs

Latin (colonial).

In each case, the script closer to phonemic bijection is predicted to yield better CTC alignment.

11.4 Limitations

Four limitations qualify these results:

1. **Transliteration noise**: N'Ko labels are derived from Latin ground truth via character-level transliteration. Native N'Ko transcriptions would eliminate this confound and likely increase the N'Ko advantage. The transliteration noise handicaps *only* N'Ko, making the observed advantage a lower bound.
2. **CER levels**: The verified reproduction baseline achieves 20.57% CER on N'Ko output on the current 290,596-pair corpus snapshot, a materially stronger result than the earlier internal trajectory figure. This is the benchmark that is fully archived locally.
3. **Optimization mismatch across evidence layers**: The current same-snapshot safe ablation bundle was rerun at a stabilized 1×10^{-4} learning rate after rejecting a non-finite higher-rate run, whereas the archived 20.57% anchor used the original 3×10^{-4} schedule. The completed ablations therefore support ranking-level conclusions more strongly than exact gap estimates.
4. **No speaker metadata**: The AfVoices corpus lacks speaker identification, preventing per-speaker test-time training experiments (§10).

12 Related Work

CTC for low-resource ASR. Conneau et al. (2020) demonstrated that cross-lingual transfer from high-resource to low-resource languages can bootstrap ASR performance when target language data is scarce. Our approach is complementary: rather than transferring from other languages, we exploit the target script's properties to reduce the decoder's learning burden.

Script effects on NLP. Muller et al. (2021) showed that cross-lingual transfer in multilingual BERT depends on shared vocabulary. Diomande (2026) demonstrated that script-level data starvation produces measurable activation deficits in LLMs. Our work extends this line to ASR, showing that script properties affect not just language model representations but speech decoder accuracy.

Phonetically motivated ASR. Phoneme-based ASR using IPA or articulatory features has been explored for low-resource settings (Li et al., 2020). Our approach differs in that N'Ko *is itself* a phonemic encoding—no intermediate IPA representation is needed because the script's design already provides the bijection.

13 Conclusion

Script design affects ASR accuracy. This paper combines formal motivation, architecture search, a fully verified N'Ko trajectory benchmark, completed same-snapshot ablations, and provisional historical script-comparison evidence.

Evidence	N'Ko Advantage
Theorem 1 (formal proof)	$\text{CER}_N \leq \text{CER}_L$
28-config arch. search	Transformer 4× dominates
Cross-system (vs MALIBA-AI)	43× param efficiency
FSM validity guarantee	100% structural validity
Verified current-snapshot winner	20.57% CER (N'Ko trajectory)
Completed same-snapshot ablations	all completed alternatives $>$ 31% CE
Historical 297K comparison	trajectory-sensitive script effect (provisi
Compositional generalization	3.65pp smaller gap
Vocabulary expansion	2.58pp residual advantage

Table 8: Summary of evidence. The verified N'Ko trajectory benchmark is 20.57% CER on the current 290,596-pair corpus snapshot. Completed same-snapshot safe ablations all underperform that anchor, while older historical script-comparison numbers remain directionally useful but provisional until their full artifact bundle is restored locally.

The Phonetic Transparency Advantage (Theorem 1) predicts that bijective transcription functions produce lower CER than many-to-many functions under identical capacity. The strongest fully verified empirical result in this repository is the reproduced N'Ko trajectory checkpoint at 20.57% CER on the current 290,596-pair corpus snapshot. Completed same-snapshot ablations then show that N'Ko baseline (31.38%), Latin baseline (31.66%), Latin trajectory (32.81%), and N'Ko TAR (31.69%) all fail to surpass that anchor. Historical internal comparison logs further suggest a more precise mechanism: N'Ko's bijective structure appears to enable architectural innovations (trajectory bias in particular) that have zero or negative effect on Latin's many-to-many mapping. Because the full historical comparative artifact bundle is not yet restored locally, those older script-comparison numbers should be treated as contextual rather than canonical.

The practical implication is more precise than "choose N'Ko." When a language community chooses which script to target for ASR, they are choosing the landscape of available decoder improvements. For N'Ko, the current benchmark now shows that a trajectory-biased decoder can reach 20.57% CER on the current corpus snapshot with a fully reproducible artifact trail, and that the completed same-snapshot alternatives tested here do not beat it. That is enough to close the main paper claim now. The remaining scientific work is narrower: finish the outstanding trajectory+TTT ablation, mirror the safe ablation artifacts locally, and treat any stronger TAR/TTT claims as future work unless they beat the existing N'Ko trajectory anchor under equally clean provenance. Script choice is architecture choice.

For the 40+ million speakers of Manding languages, the optimal output script for CTC-based ASR already exists. Solomana Kanté designed it in 1949.

Acknowledgments

This work builds on the ASR pipeline described in "Living Speech" (Paper 2) and the activation profiling methodology from "Dead Circuits" (Paper 1). The bam-asr-early corpus is released under CC-BY-4.0.

References

Mohamed Diomande. 2026a. Dead Circuits: Activation Profiling and Script Invisibility in Large Language Models. *Manuscript*.

Mohamed Diomande. 2026b. Living Speech: Script-Native Automatic Speech Recognition for N'Ko. *Manuscript*.

Mohamed Diomande. 2026c. Theorems, Proofs, and Derivations for N'Ko Script-Native ASR. *Technical Report*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of ICML 2006*.

Alexis Conneau et al. 2020. Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of Interspeech 2020*.

Benjamin Müller et al. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of EACL 2021*.

Xinjian Li et al. 2020. Universal phone recognition with a multilingual allophone system. In *Proceedings of ICASSP 2020*.

Chen et al. 2026. S-Path-RAG: Semantic-Aware Shortest Path Retrieval Augmented Generation for Multi-Hop Knowledge Graph Question Answering. *arXiv preprint*.