Beyond Controlled Comparison: Deployment Properties of Script-Aware ASR for N'Ko

Mohamed Diomande
Independent Researcher
contact@mohameddiomande.com

Provenance note. The fully verified checkpoint currently archived in this repository is the N'Ko trajectory-biased decoder trained 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. This deployment manuscript currently mixes two evidence layers. The compositional generalization and vocabulary-expansion figures are inherited from the smaller controlled experiments in the companion script-comparison work. The Djoko domain-transfer and speaker-level TTT figures come from earlier internal deployment experiments that have not yet been rerun as a fully artifact-complete current-snapshot bundle. Where this draft refers to those deployment figures, they should therefore be read as historical/provisional deployment evidence anchored by the verified 20.57% N'Ko checkpoint, not as already-rerun current-snapshot benchmarks.

Abstract

Controlled experiments show that phonetically transparent scripts yield lower CER for CTC-based ASR. But ASR systems are not evaluated in controlled conditions—they encounter unseen vocabulary, new speakers, and domain shift. This paper assembles deployment-relevant evidence for Bambara ASR systems using N'Ko (bijective script) and Latin (many-to-many script), anchored by the verified 20.57% N'Ko trajectory checkpoint but drawing on both current and historical experiments.

First, **compositional generalization**: models trained only on high-frequency words are evaluated on utterances containing rare words. N'Ko's generalization gap is 3.65pp smaller than Latin's (37.81pp vs 41.46pp), confirming that character-phoneme bijection enables composition of known units into unknown words.

Second, **vocabulary expansion**: full-data training recovers 13.75pp of the generalization gap equally for both scripts, but N'Ko retains a 2.58pp structural advantage on rare-word utterances—stable across training conditions.

Third, **test-time training**: in an earlier internal deployment experiment, we transcribe 32,826 segments from the Djoko soap opera (an out-of-domain Malinke broadcast), apply consensus filtering to extract 5,492 candidate pairs, and measure per-speaker adaptation via online weight updates. Those historical deployment experiments suggest that the N'Ko model produces usable transcriptions on 99.4% of out-of-domain segments versus Latin's distribution collapse, and that speaker-level TTT can reduce loss sharply for the best-adapting speaker.

Taken together, these results suggest that N'Ko's phonetic transparency advantage is not limited to static benchmarks. It may extend to the deployment properties that determine whether an ASR system improves with use: generalization to new words, expansion without retraining, and adaptation to new speakers and domains. Total compute cost: under $6 across all experiments.

1 Introduction

CTC-based ASR systems are typically evaluated on held-out test sets drawn from the same distribution as training data. This evaluation protocol answers one question: how well does the model transcribe speech similar to what it trained on? It does not answer the questions that matter for deployment: What happens when the model encounters words it has never seen? Can the vocabulary grow without retraining? Does the model improve as it processes more speech from a new speaker?

These questions are especially pressing for under-resourced languages like Bambara, where training data is scarce and new domains (broadcast media, conversational speech, religious text) introduce vocabulary and speaker variation that no fixed training set can anticipate.

In a companion paper (Diomande, 2026), we argue that N'Ko—a phonetically transparent script for Manding languages with strict phoneme-to-character bijection—is especially well-matched to trajectory-biased CTC decoding. The strongest fully verified checkpoint in this repository is the N'Ko trajectory model at 20.57% CER on the current 290,596-pair corpus snapshot. Earlier comparative Latin numbers from internal runs remain informative, but they are not yet artifact-complete locally.

This paper asks whether that advantage extends beyond controlled benchmarks to the properties that determine operational lifetime. We test three scenarios:

1. **Compositional generalization** (§4): Can the model transcribe words it has never seen in training?
2. **Vocabulary expansion** (§5): Does adding rare words to training data benefit both scripts equally?
3. **Domain transfer and speaker adaptation** (§6, §7): Can a model trained on read-speech corpora handle broadcast media, and does it improve per-speaker via test-time training?

The current evidence suggests that N'Ko's advantage is present in all three scenarios, and is strongest in the domain transfer setting where Latin's many-to-many character mapping appears to produce distribution collapse on out-of-domain audio.

2 Related Work

Compositional generalization in sequence models. The ability to generalize to novel combinations of known primitives is well-studied in semantic parsing (Lake and Baroni, 2018) and machine translation, but rarely examined in ASR. Character-level CTC decoders should, in principle, compose known character sequences into novel words. We test whether this composition is script-dependent.

Zero-shot vocabulary expansion. Knowledge-graph-augmented decoding has been explored for named entity recognition in ASR (Zhao et al., 2019). Our approach is simpler: we test whether training on a fuller vocabulary transfers differently to the two scripts, without requiring external knowledge injection.

Test-time training and speaker adaptation. Test-time training (TTT) updates model parameters during inference to adapt to the test distribution (Sun et al., 2020). Speaker adaptation via fine-tuning on adaptation data has a long history in ASR (Liao et al., 2013). We apply TTT at the speaker level, updating CTC decoder weights after each utterance to measure online adaptation rate.

Script effects on ASR. To our knowledge, no prior work systematically compares deployment properties across scripts for the same language. Diomande (2026) showed that trajectory-biased CTC is script-dependent at training time; we extend the comparison to inference-time adaptation.

3 Experimental Setup

This manuscript does not yet rest on one homogeneous artifact bundle. Instead it combines two evidence layers from (Diomande, 2026) and related internal deployment work.

Evidence layer A: verified benchmark anchor. The strongest fully verified checkpoint is the N'Ko trajectory model on the current 290,596-pair corpus snapshot (232,476 train / 29,060 validation / 29,060 test; seed 42), with 20.57% test CER.

Evidence layer B: controlled robustness experiments. The compositional generalization and vocabulary-expansion results reported below come from the smaller controlled experiments in the companion manuscript, where SEEN/UNSEEN splits were constructed over the development corpus and baseline decoders were compared under identical conditions.

Evidence layer C: historical deployment experiments. The Djoko domain-transfer and speaker-level TTT results come from earlier internal deployment experiments using pre-anchor checkpoints. Those results are directionally informative but have not yet been rerun as a current-snapshot artifact-complete bundle.

Architecture. Across these experiments we use Whisper large-v3 frozen encoder features with CTC decoders over N'Ko or Latin output. The current verified anchor is trajectory-biased. The historical deployment experiments compare N'Ko and Latin checkpoints from the earlier internal campaign.

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

Table 1: Compositional generalization. N'Ko's generalization gap (37.81pp) is 3.65pp smaller than Latin's (41.46pp).

Out-of-domain data. 32,826 audio segments extracted from 1,124 episodes of *Djoko*, a Bambara-language soap opera broadcast in Mali. Segments are 2–30 seconds, extracted via voice activity detection. Speaker diarization identifies 5 speakers with 27–100 segments each (speakers with $\geq$5 segments). This data was not used in training and represents a significant domain shift: conversational broadcast speech with background music, overlapping speakers, and Malinke dialect variation.

4 Experiment F: Compositional Generalization

4.1 Setup

We split the vocabulary by frequency. SEEN words appear ≥ 4 times across the corpus; UNSEEN words appear < 4 times. N'Ko: 4,184 SEEN words, 9,907 UNSEEN. Latin: 4,347 SEEN, 10,496 UNSEEN.

Utterances partition into SEEN-only (25,813 utterances where every word in both scripts is SEEN) and Has-UNSEEN (11,492 utterances with at least one UNSEEN word).

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

4.2 Results

Latin wins in-distribution (15.05% vs 16.09%), reflecting the smaller output vocabulary when all character sequences are well-attested. But on Has-UNSEEN data, N'Ko degrades less: 53.90% versus 56.51%. The generalization gap is 37.81pp for N'Ko and 41.46pp for Latin, a 3.65pp difference.

4.3 Analysis

N'Ko's bijective character-phoneme mapping means that unseen *words* are composed of the same character-phoneme units the model has already learned. The CTC decoder has seen every character in every phonemic context during training; a novel word simply arranges known characters in a new sequence.

Latin's digraphs (`ny`, `ng`, `gb`) create novel character contexts for unseen words. A word containing `ngb` requires the decoder to disambiguate whether `ng` is a digraph followed by `b` or `n` followed by `gb`—a disambiguation it may not have encountered in training. This produces the larger generalization penalty.

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

Table 2: Vocabulary expansion. Full-data training recovers 13.75pp (N'Ko) and 13.78pp (Latin). Residual gap: 24.06pp N'Ko vs 27.68pp Latin.

5 Experiment H: Vocabulary Expansion

5.1 Setup

Can training on the full vocabulary recover the generalization gap? We compare three conditions on Has-UNSEEN utterances: (1) SEEN-only model from §4; (2) Full-data model trained on all 290,596 samples; (3) Control: SEEN-only model on SEEN-only test data.

5.2 Results

Full-data training reduces Has-UNSEEN CER by 13.75pp for N'Ko (53.90% $\rightarrow$ 40.15%) and 13.78pp for Latin (56.51% $\rightarrow$ 42.73%). The recovery is nearly identical (0.03pp difference), indicating both scripts benefit equally from vocabulary expansion.

The N'Ko advantage on UNSEEN utterances is stable across conditions: SEEN-only model: -2.61pp (53.90 vs 56.51); Full-data model: -2.58pp (40.15 vs 42.73). This stability confirms the advantage derives from script structure, not training dynamics.

6 Domain Transfer: Djoko Soap Opera

Status of this section. The Djoko results in this section come from the earlier internal deployment

Script	Segments	Non-empty	Total chars	Avg len
N'Ko	32,826	99.4%	2,651,260	81.2
Latin	32,826	100%	–	–

Table 3: Djoko transcription coverage. N'Ko produces usable N'Ko-script output on 99.4% of segments. Latin produces output on all segments but with distribution collapse (repeated character patterns).

Threshold	Pairs	Pass rate	Mean conf.
≥ 0.7	258	0.8%	0.96
≥ 0.5	269	0.8%	0.96
≥ 0.3	5,492	16.7%	0.53

Table 4: Consensus filtering at three thresholds. The sharp cliff between 0.5 and 0.3 reflects the CTC confidence distribution on out-of-domain audio: most segments fall below 0.5 confidence.

experiment and should be read as provisional deployment evidence. They have not yet been rerun on the current 290,596-pair artifact-complete benchmark bundle.

The most demanding deployment scenario is domain shift: applying a model trained on read-speech corpora to conversational broadcast media. We transcribe 32,826 segments from the Djoko soap opera using both the N'Ko and Latin checkpoints, then apply multi-signal consensus filtering to identify reliable training pairs.

6.1 Transcription Results

In the historical deployment experiment, the N'Ko model produces non-empty transcriptions on 99.4% of segments (32,640/32,826), with an average of 81.2 characters per segment. The Latin model produces output on 100% of segments but exhibits **distribution collapse**: repeated character patterns (e.g., `òèqèqèqèq...`) that indicate the decoder has fallen into a low-entropy attractor state. This is a direct consequence of domain mismatch: the Latin model was trained on AfVoices read-speech data with clean articulation; Djoko's conversational Malinke dialect produces acoustic distributions that the Latin decoder maps to degenerate character sequences.

The N'Ko model, by contrast, produces diverse character output with recognizable N'Ko syllable structure. This qualitative difference motivated a one-sided consensus approach.

6.2 Consensus Filtering Pipeline

We score each N'Ko transcription using three quality signals:

1. **CTC confidence**: mean posterior probability across output frames (range 0–1).
2. **Text quality**: fraction of characters that are valid N'Ko syllable-initial consonants or vowels, weighted by length normalization.
3. **Character diversity**: ratio of unique characters to total characters, penalizing repetitive output.

These three signals are combined into a consensus score (weighted mean). Latin CTC and Whisper cross-script agreement were planned as additional signals but produced no useful signal: Latin collapses on this domain, and Whisper agreement is null for all pairs.

At threshold ≥ 0.3, the consensus pipeline produces **5,492 pairs** (16.7% of input segments) with mean CTC confidence 0.530 and mean consensus score 0.368. At threshold ≥ 0.7, only 258 pairs survive (0.8%). We use the ≥ 0.3 threshold for downstream experiments, reasoning that the expanded dataset benefits from diversity even at lower individual confidence.

The consensus pipeline is one-sided by necessity: it filters N'Ko CTC output quality because the Latin model provides no usable signal on this domain. This is itself an empirical finding—the Latin decoder's distribution collapse means it cannot contribute to consensus scoring on out-of-domain Malinke audio.

7 Experiment G: Test-Time Training

Status of this section. Like the Djoko transcription experiment, the TTT results are historical deployment results from the earlier internal checkpoint family. They remain useful as evidence about possible deployment behavior, but they are not yet current-snapshot reruns.

7.1 Setup

Test-time training (TTT) updates model weights during inference to adapt to the test distribution. We apply TTT at the speaker level: for each diarized speaker in the Djoko corpus with ≥ 5 segments, we process utterances sequentially, computing CTC loss and updating the last two MLP layers of the decoder after each utterance.

Speakers. 5 speakers with segment counts: 27, 100, 42, 58, 100 (total 327 segments).

Script	Spkrs	Avg Δloss	Improved	Best Δ	Valid
N'Ko	5	-16.18	20%	-5.22	312/327
Latin	5	$+16.30$	40%	$+94.89$	53/327

Table 5: Test-time training summary. N'Ko: lower starting loss, more valid predictions, one speaker adapts well (-5.22 loss improvement). Latin: higher starting loss, sparse valid predictions, apparent improvements are regression from catastrophic starting points.

	N'Ko			Latin		
Speaker	n	First	Last	n	First	Last
000	27	7.50	8.25	27	148.3	53.4
001	100	2.81	2.85	100	4.44	20.6
002	42	6.93	1.70	42	70.1	60.4
003	58	1.87	4.17	58	4.44	4.44
006	100	3.27	86.3	100	6.94	13.9

Table 6: Per-speaker TTT loss trajectories (first and last CTC loss). N'Ko speaker 002 shows clear adaptation (6.93 $\rightarrow$ 1.70, 75% reduction). N'Ko speaker 006 diverges catastrophically. Latin speakers start from much higher loss, and apparent improvements (speaker 000: 148 $\rightarrow$ 53) represent regression toward the mean from degenerate initialization.

TTT protocol. Learning rate: 10^{-5}. Updated parameters: last 2 MLP layers of the CTC decoder. Loss: CTC loss computed using the model's own predictions as pseudo-labels (self-training signal). Each utterance is processed once, sequentially within each speaker.

7.2 Results

The results reveal an asymmetry between the two scripts that operates at a different level than the controlled CER comparison.

N'Ko TTT. The N'Ko model starts from low loss on most speakers (1.87–7.50), indicating it can produce reasonable pseudo-labels for Djoko audio even without adaptation. Speaker 002 shows clear adaptation: loss drops from 6.93 to 1.70 across 42 utterances, a 75% reduction. The model produces valid (non-NaN) predictions on 312 of 327 segments (95.4%). Speaker 006 diverges catastrophically (3.27 $\rightarrow$ 86.32), likely due to segments with heavy background music or overlapping speech that produce unstable gradients. Excluding speaker 006, the average improvement is $+0.54$ (slight improvement), with 1 of 4 remaining speakers showing clear adaptation.

Latin TTT. The Latin model starts from much higher loss on 3 of 5 speakers (70–148), reflecting the distribution collapse observed in §6. It produces valid predictions on only 53 of 327 segments (16.2%). The apparent improvements (speaker 000: 148 $\rightarrow$ 53, speaker 002: 70 $\rightarrow$ 60) represent regression from catastrophically high starting points, not genuine speaker adaptation. Speaker 003 produces only 1 valid prediction out of 58 segments.

7.3 Analysis

The historical TTT experiment suggests that the Phonetic Transparency Advantage may operate at the domain-transfer level, not just the within-distribution level.

Starting point matters. The N'Ko model's ability to produce low-loss, valid predictions on 95% of out-of-domain segments means TTT has a viable signal to work with. The Latin model's distribution collapse on 84% of segments means TTT cannot begin—there is no gradient signal to adapt from.

Adaptation requires a stable base. Speaker 002's N'Ko adaptation (75% loss reduction over 42 utterances) demonstrates that online speaker adaptation is achievable for CTC decoders when the base model provides reasonable initial transcriptions. The failure mode (speaker 006) suggests that TTT should be gated on per-utterance loss: if a single update produces loss $> \tau$, the weight update should be rolled back.

Latin's structural problem. The Latin model's inability to produce valid pseudo-labels on Djoko audio is not simply a training-data problem within the historical experiment. It appears to be a script problem: Latin's many-to-many character mapping produces a decoder that is tightly fit to the acoustic distribution of its training data. When that distribution shifts (read speech $\rightarrow$ broadcast conversation), the decoder collapses to repeated high-frequency character patterns. N'Ko's bijective mapping produces a decoder that generalizes across acoustic distributions because each character always corresponds to the same phoneme, regardless of speaking style, dialect, or recording conditions.

8 Discussion

8.1 Operational Lifetime Properties

We define three properties that determine an ASR system's operational lifetime:

1. **Compositional robustness**: degradation on unseen words relative to in-distribution performance.
2. **Expansion efficiency**: how much vocabulary expansion improves unseen-word performance.
3. **Adaptation rate**: per-speaker improvement via online weight updates on new-domain audio.

N'Ko currently looks stronger on all three:

Property	N'Ko	Latin
Generalization gap	37.81pp	41.46pp
UNSEEN CER (full-data)	40.15%	42.73%
Djoko non-empty rate	99.4%	100%*
Valid TTT predictions	95.4%	16.2%
Best speaker adaptation	−75% loss	regression

Table 7: Operational lifetime comparison. *Latin's 100% non-empty rate reflects distribution collapse (repeated patterns), not usable transcription.

The pattern is consistent, but the provenance differs by subsection: N'Ko's bijective script appears to produce decoders that degrade more gracefully under distribution shift, while Latin's many-to-many mapping appears more brittle outside its training distribution.

8.2 Consensus Labeling for Low-Resource Domain Adaptation

The consensus pipeline extracts 5,492 labeled pairs from 32,826 unlabeled broadcast segments using only CTC self-confidence as the quality signal. This is a practical methodology for expanding training data in low-resource settings:

1. Transcribe unlabeled domain audio with the existing model.
2. Score transcriptions by CTC confidence, character diversity, and text quality.
3. Filter at an appropriate threshold to produce pseudo-labeled training pairs.
4. Retrain on the expanded dataset (original + pseudo-labeled pairs).

This pipeline is available only for the N'Ko model—the Latin model's distribution collapse means it cannot contribute pseudo-labels for Djoko audio. The ability to generate pseudo-training data from new domains is itself a deployment advantage of the bijective script.

8.3 CER as Phonemic Accuracy: A Case for Script-Native Evaluation

The standard Bambara ASR benchmark (MALIBA-AI) reports word error rate (WER) on Latin-script output. Our results are reported as character error rate (CER) on N'Ko-script output. These metrics are not directly comparable, but the incomparability is itself informative.

For a bijective script like N'Ko, CER is a closer proxy for phonemic accuracy than Latin WER: most N'Ko character substitutions correspond to phonemic errors, although spaces, punctuation, digits, and combining marks mean the correspondence is not exact. For Latin Bambara, this correspondence does not hold. A single Latin character error may or may not change the phoneme: corrupting one character of the digraph `ny` destroys the phoneme /ɲ/ entirely, while substituting `n` for `m` changes a different phoneme. Latin WER is even further removed from phonemic accuracy, as word boundaries do not align with phonemic boundaries.

We therefore propose that **N'Ko CER should serve as the phonemically grounded benchmark for Manding ASR**, rather than attempting to compare against Latin WER. This is not merely a notational convenience. The deployment experiments in this paper show that N'Ko's bijective structure enables capabilities (domain transfer, self-training, speaker adaptation) that Latin's metric opacity does not. A system whose error metric directly measures phonemic accuracy is more interpretable, more auditable, and more useful for downstream applications such as pronunciation feedback and language pedagogy.

8.4 Limitations

1. **Mixed evidence layers**: this paper combines a verified current-snapshot benchmark anchor, smaller controlled generalization experiments, and historical deployment experiments that have not yet been rerun as one artifact-complete bundle.
2. **No ground truth for Djoko**: consensus quality is measured by model self-confidence, not against human transcription. The 5,492 pairs may contain systematic errors.
3. **Speaker diarization quality**: speaker boundaries from automated diarization may

be imperfect, affecting TTT trajectories.

4. **TTT instability**: speaker 006's catastrophic divergence shows that online adaptation requires safeguards (loss thresholds, weight rollback) not yet implemented.
5. **Latin collapse may be recoverable**: domain-adaptive pre-training or Latin-specific fine-tuning on Djoko audio might recover the Latin model's performance. We test only zero-shot transfer.
6. **Single language**: all experiments use Bambara. Generalization to other bijective-script languages (e.g., Hausa in Ajami, Uyghur in Arabic script) requires separate validation.

9 Conclusion

Phonetic transparency is not just a static CER advantage. It appears to be an operational property that can determine how an ASR system behaves under deployment conditions.

The current verified anchor is the 20.57% N'Ko trajectory checkpoint on the 290,596-pair corpus snapshot. Around that anchor, the controlled robustness experiments show that N'Ko generalizes better to unseen vocabulary (3.65pp smaller gap) and retains a structural advantage after vocabulary expansion (2.58pp residual). The historical deployment experiments further suggest that N'Ko can produce valid transcriptions on 99.4% of out-of-domain broadcast audio where Latin collapses, and may enable per-speaker adaptation via test-time training, with the best speaker showing 75% loss reduction.

The consensus labeling pipeline suggests that N'Ko models may be able to bootstrap their own training data from unlabeled broadcast audio—a self-improvement loop unavailable to the Latin decoder in the historical deployment experiment.

For the Manding language community choosing a script for ASR technology, the choice is not between marginally different CER numbers. It is between a system that already has a verified high-performing N'Ko operating point and a set of deployment experiments that increasingly favor the script designed for the language. The next step is straightforward: rerun the Djoko and TTT studies on the current artifact-complete checkpoint family and promote only the claims that survive that provenance upgrade. Script design is not just architecture choice—it is lifecycle choice.

Acknowledgments

This work extends the controlled comparison in "Does Script Design Matter?" (Diomande, 2026). The Djoko audio corpus was extracted from publicly available YouTube broadcasts. All compute ran on RTX 4090 spot instances at a total cost under $6.

References

Mohamed Diomande. 2026. Does Script Design Matter? Phonetic Transparency and CTC Decoding for N'Ko Automatic Speech Recognition. *Manuscript*.

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*.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of ICML 2018*.

Hank Liao, Erik McDermott, and Andrew Senior. 2013. Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. In *Proceedings of ASRU 2013*.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of ICML 2020*.

Jing Zhao et al. 2019. Shallow-fusion end-to-end contextual biasing. In *Proceedings of Interspeech 2019*.