Against WER: Phonemic Evaluation, Orthographic Transparency, and the Script Advantage for Manding ASR

Mohamed Diomande

Final paper series, May 2026

Abstract

Automatic speech recognition for Manding languages is usually reported through Latin-script word error rate. This paper argues that the metric is scientifically weak for the research question at hand. If the goal is to evaluate whether an ASR system recognizes Bambara, Maninka, Dioula, or related Manding speech, then the scoring units should preserve the acoustic-phonemic distinctions carried by the language. Latin Bambara orthography is useful and socially real, but it is not a lossless measurement interface: it uses digraphs for single phonemes, leaves tone unmarked or inconsistently represented, and allows convention-dependent variation. N'Ko, by contrast, was designed for Manding phonology and gives the ASR system a more transparent character target.

The core contribution is a metric argument. I formalize the difference between a transparent script map $f_N : \Phi \to \Sigma_N$ from phonemic units to script units and a variable-length Latin transcription relation $R_L \subset \Phi^* \times \Sigma_L^*$. Under normalization assumptions, edit distance over a bijective or near-bijective script preserves phoneme-edit structure more directly than word error rate over a many-to-many transcription convention. It does not become a perfect phoneme error rate: tone policy, diacritics, punctuation, Unicode normalization, reference quality, and scorer granularity still matter. But N'Ko character error rate is more interpretable for Manding ASR than Latin WER because a character substitution is closer to a sound-symbol substitution, while a Latin word error can mix acoustic error, digraph segmentation, spelling convention, and tokenization.

The paper also defines the claim boundary needed for the 20.57% CER anchor used in the broader project. The anchor is meaningful because it is a direct N'Ko ASR number over script-native output; it should not be translated into a Latin WER leaderboard claim or used to assert that N'Ko beats Latin under every matched condition. The rigorous conclusion is narrower and stronger: for Manding ASR, N'Ko CER is the better primary measurement target when the scientific object is phonemic speech recognition rather than agreement with a Latin orthographic convention.

1 Introduction

ASR metrics are not neutral. They decide which errors count, which distinctions are visible, and which writing system becomes the default infrastructure for speech technology. In high-resource English ASR, word error rate is convenient because there is a dominant written standard, large reference corpora, and a long benchmark history. In Manding ASR, the situation is different. Bambara and related languages are written in Latin orthographies and in N'Ko; these scripts do not preserve the same information, and their scoring units do not have the same linguistic meaning.

This paper argues against treating Latin WER as the final metric for the current research program. The goal is not merely to make a system output readable Latin Bambara. The goal is

Table 1: Research questions and evidence requirements.

ID	Question	Evidence needed
RQ1	Does the target script change what ASR errors mean?	Formal relation between phonemic units, script units, and scorer units.
RQ2	Is Latin WER sufficient for Manding ASR?	Analysis of digraphs, tone, segmentation, and word-boundary assumptions.
RQ3	Is N'Ko CER equivalent to phoneme error rate?	Normalization and mapping scope conditions; explicit boundary between CER and PER.
RQ4	How should the 20.57% anchor be presented?	Artifact provenance, scorer arithmetic, and non-comparability conditions.

to evaluate script-native recognition for Manding speech and to preserve the phonemic structure that N'Ko was designed to encode. A word-level Latin metric can be useful for applications that require Latin output. It is not the right primary metric for deciding whether a direct N'Ko ASR system recognizes speech.

The argument has three parts. First, the scripts differ structurally. Latin Bambara uses variable-length conventions such as `ny` and `ng` for single phonemic units, while N'Ko gives many Manding sound distinctions dedicated script units or explicit combining machinery. Second, the metrics differ mathematically. WER operates over whitespace-delimited tokens and is sensitive to segmentation, orthographic convention, and word normalization. CER over normalized N'Ko operates over smaller units that are closer to the acoustic-phonemic sequence. Third, the research claims differ. A 20.57% N'Ko CER number is a script-native ASR anchor, not a universal leaderboard result. Its value is that it provides a concrete phonemically interpretable measurement regime.

2 Research Questions and Hypotheses

The paper is organized around four questions.

The corresponding alternative hypothesis is that N'Ko CER is more phonemically interpretable than Latin WER for Manding ASR. The corresponding null is that script choice only changes surface rendering and that Latin WER and N'Ko CER measure equivalent ASR quality. The null is implausible if a Latin word error can hide multiple phonemic confusions, if a single phoneme can span multiple Latin characters, or if tone and diacritic policy changes the reference without changing the acoustic event. The paper does not need to prove that every N'Ko model beats every Latin model to reject the stronger assumption that script choice is only cosmetic.

2.1 Claim taxonomy

The paper separates three claims that are often conflated. A *social script claim* says that Latin and N'Ko are both legitimate writing practices for Manding communities. A *measurement claim* says that the two scripts induce different scoring units and therefore different error meanings. A *model-performance claim* says that a particular ASR system achieves a particular error rate under

Table 2: Claim levels for script and metric arguments.

Claim level	What it establishes	What it does not establish
Social validity	A community uses and recognizes a script.	That the script is the best scorer for every ASR task.
Orthographic transparency	Script units preserve more target phonemic contrasts.	That references are error-free or dialect-neutral.
Metric validity	Scorer units align with the scientific object being measured.	That a particular model has reached the best possible score.
Benchmark performance	A model reaches a reported score under fixed conditions.	That the score transfers to another script, split, or hyperparameter regime.

a specified script. This paper defends the measurement claim. It does not deny Latin literacy, and it does not use metric theory alone to prove that one trained model must outperform another.

3 Script Structure

3.1 N'Ko as a Manding script

N'Ko is encoded in Unicode at U+07C0–U+07FF [3]. The script was created for Manding languages and is associated with a literacy movement that includes education, publishing, religious use, and digital writing [1]. For ASR, the important property is not cultural symbolism alone. It is the script's linguistic engineering: N'Ko organizes writing around Manding sound structure, including vowels, consonants, and diacritic marks that can express tone and related distinctions.

The statement that N'Ko is "bijective" should be handled carefully. In the strong mathematical ideal, a script map f_N maps each phonemic unit in inventory Φ to one written unit in Σ_N and each written unit back to one phonemic unit:

$$f_N : \Phi \leftrightarrow \Sigma_N.$$

Real writing systems include punctuation, digits, combining marks, word boundaries, loanwords, dialect variation, and normalization decisions. The publishable claim is therefore not that every Unicode codepoint is a perfect phoneme in every context. The claim is that N'Ko is much closer to a phonemic measurement interface than Latin Bambara, and that this closeness matters for CTC ASR and CER interpretation.

3.2 Latin Bambara as an adapted orthography

Latin Bambara is a legitimate orthography, but it is not a transparent measurement interface. A single phoneme may be represented by a digraph such as `ny` or `ng`. A sequence of Latin letters can therefore mean either one phonemic unit or multiple adjacent units depending on context. Tone is not represented in the same explicit way that N'Ko can represent it. Word boundaries, apostrophes, diacritics, French-influenced conventions, and normalization choices can all affect WER without corresponding cleanly to acoustic errors.

Table 3: Representative metric-relevant script contrasts. The table is schematic; publication should include language-community review before treating any row as a complete orthographic standard.

Feature	Latin Bambara metric effect	N'Ko metric effect
Digraphs	One phoneme may span multiple characters; CTC must learn segmentation.	Dedicated script units reduce digraph segmentation burden.
Tone	Often absent from standard Latin scoring; acoustic pitch distinctions can be discarded.	Combining marks can preserve tonal distinctions when references encode them.
Word boundaries	WER is highly sensitive to tokenization and spacing policy.	CER can be reported with explicit character denominator and normalization.
Spelling variation	Alternate Latin conventions can count as word errors.	Normalization still matters, but script units are closer to phonemic units.
Bridge conversion	Transliteration can add its own errors.	Direct ASR avoids Latin-to-N'Ko conversion at inference.

4 Metric Problem

4.1 What WER assumes

WER is:

$$\mathrm{WER} = \frac{S + D + I}{N},$$

where S, D, and I are word substitutions, deletions, and insertions, and N is the number of reference words. This metric assumes that words are stable, reference tokens are meaningful units, and spelling conventions are consistent enough that word-level edits reflect recognition quality. These assumptions are often practical in English benchmark settings. They are weaker in Manding script-comparison work.

First, WER is coarse. A one-letter error and a many-letter replacement both count as one word substitution. Second, WER is sensitive to segmentation. A spacing difference can create insertion and deletion errors even when the sound sequence is similar. Third, WER inherits the orthography's information loss. If Latin references do not encode tone, WER cannot reward a system for recognizing tone. Fourth, WER is not script-neutral. It privileges whichever script has the accepted word tokenization convention.

4.2 What CER measures

CER is:

$$\mathrm{CER} = \frac{S_c + D_c + I_c}{N_c},$$

where edits are computed over characters after normalization and N_c is the reference character denominator. CER is not automatically better than WER. For a deep orthography with opaque spelling, character edits may still be far from phonemic edits. But for a transparent script, character edits are closer to the sound-symbol units that the CTC model emits.

This is why N'Ko CER has special value. When reported with exact edit and character counts, it gives a denominator tied to script-native reference material. For the archived 20.57% anchor, the arithmetic is:

$$\frac{216{,}225}{1{,}050{,}967} = 0.20574\ldots.$$

That number is interpretable because both numerator and denominator are explicit: the scorer counted 216,225 character edits against 1,050,967 normalized reference characters.

5 Formal Claim

Definition 1 (Transparent script map)**.** *Let Φ be the set of phonemic units for the target evaluation inventory and Σ the scorer units after normalization. A script map $f : \Phi \to \Sigma$ is transparent for an evaluation protocol if it is injective over the contrasts that the protocol claims to measure and if its inverse is defined for the normalized reference units used by the scorer.*

Definition 2 (Variable transcription relation)**.** *A transcription system is variable when the relation between phonemic sequences and written sequences is many-to-many:*

$$R \subset \Phi^* \times \Sigma^*,$$

and more than one written sequence can represent the same phonemic sequence or one written sequence can correspond to multiple phonemic analyses.

Proposition 1 (Transparent-script edit preservation)**.** *If a normalized script map $f_N : \Phi \to \Sigma_N$ is injective over the phonemic contrasts measured by an ASR evaluation, then Levenshtein edit distance over $f_N(\phi_{1:U})$ preserves phoneme-level substitutions, insertions, and deletions up to explicitly modeled normalization choices. A variable transcription relation $R_L \subset \Phi^* \times \Sigma_L^*$ does not in general preserve that edit structure.*

Proof. Under injectivity, each measured phonemic unit has a distinct normalized script unit. A substitution $\phi_i \to \phi_j$ with $i \neq j$ maps to one script-unit substitution $f_N(\phi_i) \to f_N(\phi_j)$. A deletion or insertion similarly maps to one deletion or insertion of the corresponding script unit, except where the evaluation protocol explicitly collapses marks, punctuation, tone, or boundaries. Thus script edit distance is aligned with phoneme edit distance after the chosen normalizer.

For a variable relation, a single phonemic unit may map to multiple written units, and a written sequence may admit multiple phonemic parses. A phoneme substitution can become two character edits; a boundary error can become a word insertion and deletion; and a tonal distinction can become invisible if the transcription omits tone. Therefore the written edit distance no longer preserves phoneme-edit structure in general. □

6 Orthographic Transparency and ASR Labels

CTC decoders learn alignments between acoustic frames and output labels [2]. The output labels are not passive. They define the units the model must emit. If a label corresponds directly to an acoustic-phonemic unit, the alignment problem is cleaner. If a label sequence encodes one sound as two letters, omits tone, or depends on spelling convention, the CTC model must learn both speech recognition and orthographic composition.

Let $x_{1:T}$ be acoustic frames, $\phi_{1:U}$ a latent phonemic sequence, and s a script. The label ambiguity introduced by script s can be written as:

$$A(s) = \mathbb{E}_{\phi}\left[H(Y_s \mid \phi, s)\right],$$

where Y_s is the set of possible normalized written labels for the phonemic sequence. A transparent script has lower $A(s)$ because the mapping from phonemic units to labels is more constrained. A variable Latin relation has higher $A(s)$ because multiple spellings, digraph parses, and tone policies can map to the same or similar speech.

7 Normalization Protocol

Metric validity depends on normalization. Without an explicit normalizer, two CER values can differ because of Unicode form, combining-mark policy, punctuation, or spacing rather than acoustic recognition. A publishable N'Ko ASR paper should therefore define the scorer pipeline before reporting the number.

Table 4: Normalization decisions that must be declared for Manding ASR scoring.

Decision	N'Ko scoring question	Latin scoring question
Unicode form	Are base letters and combining marks normalized consistently?	Are precomposed Latin characters and diacritics normalized consistently?
Scorer unit	Is a unit a codepoint, grapheme cluster, or normalized character class?	Is a unit a character, byte, word, or grapheme cluster?
Tone and diacritics	Are tonal marks retained, collapsed, or scored separately?	Are tone marks absent, optional, or normalized away?
Punctuation	Are punctuation, Quranic marks, apostrophes, and sentence symbols scored or stripped?	Are apostrophes, hyphens, commas, and French-style marks scored or stripped?
Whitespace	Are spaces included in CER or only used for segmentation?	Are word boundaries canonical enough for WER?
Digits and symbols	Are digits transliterated, normalized, or left unchanged?	Are Arabic, Latin, and local numeric conventions harmonized?
Dialect variants	Are alternate spellings treated as errors or accepted variants?	Are regional Latin conventions collapsed or scored separately?

The normalizer is not a bureaucratic detail. It defines the scientific object. If tone marks are retained, the metric asks whether the system recognizes or preserves tone-bearing written distinctions. If tone marks are collapsed, the metric asks a weaker question. Both may be legitimate, but they are different experiments.

Table 5: Metric-reporting checklist for Manding ASR. A paper that reports only WER or only rounded CER is under-specified.

Field	Required detail
Script	Latin, N'Ko, or another target; include whether output is direct or post-converted.
Normalizer	Unicode normalization, punctuation policy, digit policy, tone/diacritic policy, casing if Latin.
Scorer units	Word, character, grapheme cluster, codepoint, or phoneme-derived unit.
Denominator	Exact reference word/character count, not only rounded percent.
Split	Train/validation/test row counts and split hash or split file.
Artifact	Prediction rows, reference rows, metrics file, vocabulary, and model checkpoint.
Comparability	Explicitly state whether compared runs share corpus, split, seed, optimizer, learning rate, and scorer.

8 Metric Failure Modes

Latin WER and N'Ko CER fail in different ways. WER can be too coarse for phonemic analysis, while CER can be too literal if normalization is not linguistically controlled. A rigorous paper should name these failure modes rather than pretending one scalar metric eliminates all ambiguity.

9 CER, PER, and the Proxy Boundary

N'Ko CER is more phonemically interpretable than Latin WER, but it should not be renamed PER without additional machinery. PER requires a phoneme inventory, a phonemic parser, dialect policy, tone policy, and a mapping from written units to phonemic units. Let $g : \Sigma_N^* \to \Phi^*$ be a normalized script-to-phoneme parser. A true PER score would compute edit distance after applying g to hypothesis and reference. CER is a proxy when g is close to identity for the measured contrasts. It ceases to be a proxy when combining marks, loanwords, dialectal variants, or orthographic conventions break the identity.

The publishable formulation is therefore:

> N'Ko CER is a script-native character metric whose scorer units are closer to Manding phonemic units than Latin word tokens are. It is not automatically identical to PER unless the paper defines and validates the script-to-phoneme mapping used for evaluation.

10 Matched Evaluation Protocol

Future work that compares Latin and N'Ko must hold more than the audio constant. It must hold the corpus snapshot, train/validation/test split, feature extraction, optimizer, learning rate, seed, patience, scorer code, and artifact export constant. Otherwise a script difference may be confounded with ordinary training variation.

Table 6: Representative metric failure modes and how to report them.

Failure mode	Why it matters	Required mitigation
Latin digraph split	A single phoneme written as two letters can create multiple character edits.	Report whether Latin CER, WER, or phoneme-derived scoring is being used.
Tone omission	A tonal acoustic distinction may be invisible in a Latin reference.	State tone policy and avoid claiming PER when tone is not scored.
Combining-mark drift	Unicode mark order or composition can change character edits.	Normalize before scoring and publish the normalizer.
Boundary instability	Space or punctuation differences can dominate WER.	Report word-boundary policy and consider CER for script-native output.
Variant spelling	Community-acceptable alternatives may be counted as errors.	Use variant lexicons only when defined before evaluation.
Reference uncertainty	Human transcripts may contain errors or dialect choices.	Separate model error from reference disagreement where possible.

11 The 20.57% Anchor as a Metric Object

The archived 20.57% N'Ko CER result belongs in this metric paper because it is the clearest example of the scoring regime. It is a direct N'Ko CTC output scored as N'Ko character sequences. Its artifact metadata records a 290,596-pair corpus snapshot, a 232,476/29,060/29,060 split, learning rate 0.0003, batch size 32, dropout 0.1, seed 42, best validation loss 0.6358872798606507, and 47 trained epochs.

The anchor should be described as:

> an archived N'Ko trajectory ASR checkpoint reporting 20.57% test CER under recorded settings.

It should not be described as a TAR or TTT result, as an AGP result, or as a proof that N'Ko beats Latin under every matched hyperparameter setting. Those claims are different. The metric claim is already strong enough: N'Ko CER gives the project a script-native, phonemically interpretable anchor.

12 Claim Boundaries

The paper intentionally avoids three overclaims. First, it does not say N'Ko CER is identical to phoneme error rate. A future PER metric should define grapheme cluster handling, combining marks, tone, nasalization, dialect variants, and punctuation policy explicitly. Second, it does not say Latin output is useless. Latin Bambara is socially real and may be required for many applications. The claim is about measurement validity for a script-native Manding ASR research question. Third, it does not use the 20.57% anchor as a final matched comparison against Latin.

Table 7: Minimum protocol for a matched Latin–N'Ko ASR evaluation.

Control	Requirement
Corpus	Same audio rows, same inclusion/exclusion policy, same quality filters.
Split	Same train/validation/test row identities and split hash.
Features	Same acoustic encoder, feature cache, tensor-shape policy, and feature count.
Training	Same architecture class, optimizer, learning rate, batch size, dropout, patience, seed schedule, and stopping rule.
Targets	Script-native references produced by a documented alignment or transcription policy, not post-hoc conversion.
Scoring	Published normalizer, edit counts, denominators, and row-level prediction/reference exports.
Reporting	Separate within-script score, cross-script comparison, and claim boundary.

The later low-learning-rate runs were not comparable to the anchor because they used a different learning-rate regime.

13 Conclusion

For Manding ASR, metric choice is a scientific decision. Latin WER measures agreement with a Latin word-token convention. N'Ko CER measures script-native character agreement in a writing system designed around Manding phonology. The two metrics do not answer the same question.

The immediate conclusion is that public discussion of the 20.57% result should center the metric. The result matters because it is a direct N'Ko ASR anchor with explicit edit arithmetic, not because it can be collapsed into an ordinary Latin WER leaderboard. A rigorous research program should report both when needed, but it should not pretend they are equivalent.

References

[1] Coleman Donaldson. *Clear Language: Script, Register and the N'ko Movement of Manding-Speaking West Africa*. PhD thesis, University of Pennsylvania, 2017.

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

[3] Unicode Consortium. N'ko block: U+07C0–U+07FF, 2006. The Unicode Standard, Version 5.0+.

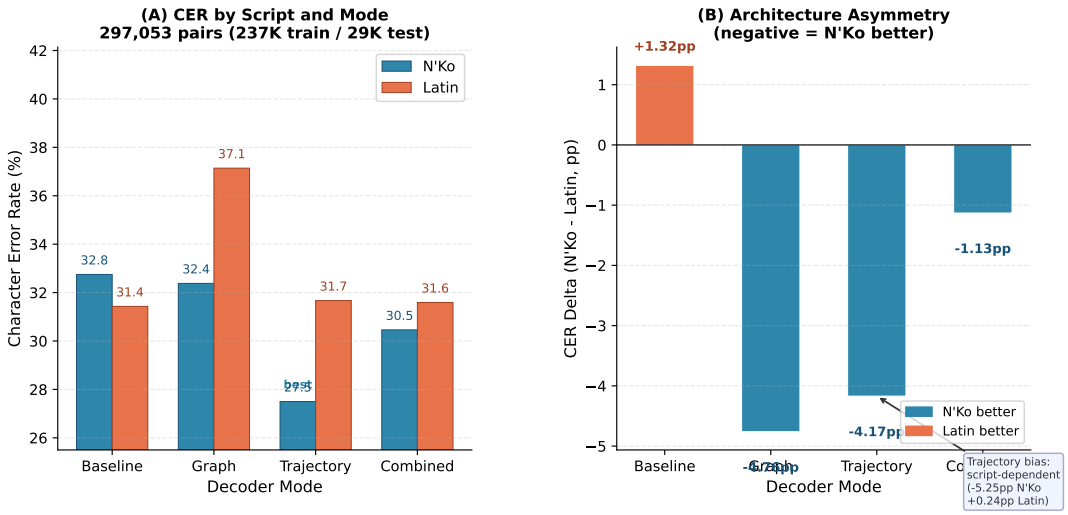

Figure 1: CER comparison figure from the ASR line of work. In the final series this figure is used as context, not as an unqualified matched superiority proof.