CVNSS4.0 AS AN EXPERIMENTAL ORTHOGRAPHIC-AUDIT WRITING MODEL FOR VIETNAMESE: VOWEL TRANSFORMATION, JESUIT MISSIONARY LINGUISTICS, AND DOMAIN-SPECIFIC AI/NLP/LLM INFRASTRUCTURE

NGO HOANG DAI LONG
Vietnam National University, Campus Ben Tre, Vinh Long, Ho Chi Minh City, Viet Nam
Corresponding author: nhdlong@vnuhcm.edu.vn

Abstract: This article proposes CVNSS4.0 as an experimental orthographic-audit writing model for Vietnamese rather than as a replacement for Quốc Ngữ. The model is framed as a machine-readable intermediate representation that decomposes Vietnamese syllables into onset, rhyme/nucleus, coda, tone, and audit metadata. The central argument is that the double transformation of the rhyme in CVNSS4.0—first from visually diacritized Quốc Ngữ vowels into structural vowel modules, and then into linearized glide-aware codes such as j/w plus tone and coda markers—belongs to a broader cross-linguistic pattern in which writing systems externalize phonological layers for literacy, missionization, administration, and now computational processing. The study places CVNSS4.0 in a four-century continuum beginning with Jesuit and missionary grammars in Vietnamese, Tupi, Guaraní, Quechua, and Aymara, and compares it with representative scripts and orthographies from Africa, the Americas, East Asia, and Southeast Asia. A conceptual research design is proposed for evaluating roundtrip accuracy, audit completeness, tone preservation, OCR/ASR robustness, and downstream AI/NLP/LLM utility. The article concludes that CVNSS4.0 should be investigated as an explainable preprocessing and audit layer for Vietnamese language infrastructure, especially for domain-specific corpora, dictionaries, administrative records, and low-resource model training.

Keywords: CVNSS4.0. Vietnamese orthography. Missionary linguistics. Writing-system typology. Audit trace. Low-resource NLP. Domain-specific LLM.

Resumo: Este artigo propõe o CVNSS4.0 como um modelo experimental de escrita ortográfica e de auditoria para o vietnamita, não como substituto do Quốc Ngữ. O modelo é entendido como uma representação intermediária legível por máquina, capaz de decompor a sílaba vietnamita em ataque, rima, coda, tom e metadados de auditoria. O argumento central é que a dupla transformação da rima no CVNSS4.0 pertence a um padrão tipológico mais amplo no qual os sistemas de escrita externalizam camadas fonológicas para fins de alfabetização, missão, administração e processamento computacional. O estudo relaciona o CVNSS4.0 a uma continuidade histórica de quatro séculos, desde gramáticas missionárias jesuíticas em vietnamita, tupi, guarani, quíchua e aimará, até modelos atuais de IA/NLP/LLM.

Palavras-chave: CVNSS4.0. Ortografia vietnamita. Linguística missionária. Tipologia dos sistemas de escrita. Auditoria textual. NLP de poucos recursos.

---

1. Introduction

Vietnamese writing is a layered historical object. In contemporary use, Quốc Ngữ appears as a Latin-based alphabet with diacritics, but its visible simplicity conceals several interacting levels: phonemic contrasts, vowel diacritics, tone marks, coda conventions, historical spellings, Sino-Vietnamese lexical strata, regional pronunciation, and the older ecological memory of chữ Hán and chữ Nôm. For human readers, these layers are normally integrated by literacy practice. For machines, however, the same layers create problems of normalization, ambiguity, reversibility, and provenance.

The present article develops a conceptual and methodological argument: CVNSS4.0 may be studied as an experimental orthographic-audit writing model for Vietnamese. The expression "orthographic-audit" is used deliberately. It indicates that the model is not proposed as a cultural replacement for Quốc Ngữ, but as an accountable computational layer that makes text transformation explicit, measurable, and reversible where possible. In a domain-specific AI infrastructure, the decisive question is no longer only whether a string can be converted, but whether the conversion can be explained, versioned, reproduced, and audited.

The historical motivation is equally important. The romanization of Vietnamese was not an isolated event; it emerged in the early modern contact zone among local Vietnamese speech communities, East Asian manuscript traditions, Portuguese orthographic habits, and Jesuit missionary linguistics. Francisco de Pina, Alexandre de Rhodes, Gaspar do Amaral, António Barbosa, and other missionaries participated in a complex codification process that culminated in the publication of the 1651 Dictionarium Annamiticum Lusitanum et Latinum and related grammatical materials. Fernandes and Assunção (2017) emphasize that early missionary descriptions of Vietnamese tones and Portuguese orthographic influence shaped the first codification of Vietnamese in Roman characters.

This four-century arc matters for AI. The first romanizers needed a practical inscription system for learning, teaching, preaching, and compiling dictionaries. Today, AI/NLP/LLM systems need a practical inscription layer for tokenization, corpus construction, OCR correction, ASR normalization, diacritic restoration, retrieval, and controlled generation. Both historical moments respond to the same structural challenge: how to make a tonal, syllable-centered language legible to an external technical system. In the seventeenth century, that system was missionary grammar and print. In the twenty-first century, it is machine learning, vector search, and model governance.

The research problem can therefore be formulated as follows: Can CVNSS4.0 be formalized as a Vietnamese intermediate representation that preserves cultural and orthographic specificity while improving computational auditability? The study answers this question through a comparative, typological, and architectural analysis. It examines the double transformation of Vietnamese rhymes in CVNSS4.0, situates it among comparable patterns in writing systems from Asia, Africa, and the Americas, and proposes a benchmark design for evaluating its practical usefulness.

The article makes three contributions. First, it offers a theoretical framing of CVNSS4.0 as a writing model, not merely a converter. Second, it builds a comparative bridge between Vietnamese

missionary codification and other missionary or Indigenous writing traditions, including Tupi, Guaraní, Quechua, Aymara, Yoruba, Ewe, N'Ko, Mapudungun, Vai, and Mende Kikakui. Third, it proposes mathematical metrics and an AI/NLP/LLM pipeline for testing CVNSS4.0 in domain-specific Vietnamese corpora. The goal is not to prove final superiority, but to define what would count as evidence.

Table 1. A 400-year continuum from missionary codification to computational audit writing

Period	Language / script	Historical event or typological function	Relevance to CVNSS4.0
1595	Old Tupi / Brazil	José de Anchieta, a Jesuit, produced a grammar of the language used on the Brazilian coast; the work represents early missionary grammaticalization of an Indigenous language.	Shows that missionary grammars converted oral/Indigenous systems into explicit analytical artifacts.
1603	Aymara / Andes	Ludovico Bertonio, S.J., published Arte y grammatica muy copiosa de la lengua aymara.	Demonstrates early formal description of a non-European language with complex morphology and vowel behavior.
1607–1608	Quechua / Andes	Diego González Holguín, S.J., produced grammar and dictionary work on the general language of Peru.	Shows the role of dictionary-grammar coupling as an infrastructure for linguistic standardization.
1639–1640	Guaraní / Paraguay	Antonio Ruiz de Montoya, S.J., published Guaraní grammar, vocabulary, catechetical and dictionary works.	Useful for comparing missionary codification, lexicography, and cultural-political language infrastructure.
1651	Vietnamese / Quốc Ngữ	Alexandre de Rhodes published a Vietnamese–Portuguese–Latin dictionary and grammatical description.	Provides the historical anchor for Vietnamese Roman orthography and tone description.
2026	Vietnamese / CVNSS4.0	CVNSS4.0 is framed as an audit-ready intermediate representation for NLP/LLM pipelines.	Reinterprets orthographic codification as explainable machine preprocessing.

2. Theoretical framework and literature review

2.1 Writing systems as layered technologies

A writing system is not merely a set of symbols. It is a technology that selects which linguistic contrasts are made visible, which are left implicit, and which are delegated to reader knowledge. Alphabetic systems tend to foreground segmental units; abugidas foreground consonant-vowel dependencies; syllabaries encode syllables as graphic units; abjads privilege consonantal skeletons; tonal orthographies may use diacritics, final letters, numerical marks, or separate tone signs. CVNSS4.0 can be located in this typological field as a secondary, computationally motivated layer over an existing alphabetic orthography.

The decisive feature of CVNSS4.0 is that it treats the Vietnamese syllable as a structured object rather than a sequence of characters. This interpretation is consistent with the fact that many writing systems encode syllable-internal organization directly. Hangul represents onset, nucleus, and optional coda in a square block. Brahmic scripts represent vowels as dependent signs around consonant bases. Hmong RPA and several African orthographies encode tone as a final or combining symbol. Canadian Aboriginal Syllabics modifies orientation or shape to express vocalic contrasts. These systems differ historically and graphically, but they share a common logic: the visible form of writing is an engineered compromise between phonology, pedagogy, material media, and social use.

For AI/NLP, this typological insight is practical. A model trained only on Unicode surface strings may learn statistical correlations without knowing whether a change is a tone error, a coda error, a vowel error, or an encoding artifact. A structured representation can expose these categories to

preprocessing, alignment, error diagnosis, and evaluation. CVNSS4.0 should therefore be assessed as a candidate for explainable preprocessing rather than as a mere romanized shorthand.

2.2 Vietnamese orthographic history and the Jesuit grammar-dictionary paradigm

The Vietnamese case is historically distinctive because Quốc Ngữ emerged from a contact zone involving Vietnamese speakers, East Asian manuscript traditions, Portuguese and Romance orthographic conventions, and Jesuit missionary education. The early missionaries did not invent Vietnamese language; they built a technical inscription system for representing it with Roman characters. That inscription became socially transformed over centuries, eventually becoming the modern national script. The distinction is important: historical codification began as an external technical layer, but later became a social writing system.

This history makes CVNSS4.0 culturally legible. CVNSS4.0 is not the first attempt to build a technical layer around Vietnamese. The seventeenth-century missionary project already transformed Vietnamese phonology into a grammar-dictionary apparatus. What differs today is the receiving system. The seventeenth-century receiver was Latin alphabetic print and missionary pedagogy; the twenty-first-century receiver is AI infrastructure. In both cases, the problem is the same: Vietnamese must be represented with sufficient granularity to preserve tones, vowels, and syllables.

The grammar-dictionary paradigm is also visible in other Jesuit and missionary contexts. Anchieta's Tupi grammar, Bertonio's Aymara grammar, González Holguín's Quechua works, and Ruiz de Montoya's Guaraní dictionary and grammar show that missionary linguistics functioned as an early global infrastructure for language description. These works were not neutral. They involved power, conversion, coloniality, pedagogy, and administration. Yet they also preserved lexical and grammatical records that remain crucial for historical linguistics. A contemporary CVNSS4.0 project must learn from both sides: the technical benefit of explicit codification and the ethical requirement to respect linguistic communities.

2.3 Diacritics, tone, nasalization, and low-resource NLP

Vietnamese is not alone in making diacritics and tone central to lexical meaning. Yorùbá orthography uses tonal marks and underdots; recent NLP literature emphasizes that the omission of diacritics affects lexical disambiguation, pronunciation, embeddings, and downstream tasks (Adewumi et al., 2020; Orife et al., 2020). Ewe combines tone marking with nasalized vowels. N'Ko has vowel letters, nasalization, tone marks, and length-related tonal distinctions, as described in script layout requirements by W3C. Guaraní exhibits nasal harmony, where nasality can spread through a phonological word, and Mapudungun orthographies reveal the relation between high vowels and glide-like segments.

These languages demonstrate that vowel letters are often hubs of multiple linguistic layers. A vowel may carry tone, nasality, length, stress, rounding, palatality, or syllabic status. In such contexts, removing diacritics is not a harmless normalization step; it may erase semantic, phonological, and morphological information. The same warning applies to Vietnamese data cleaning. A Vietnamese

NLP pipeline that strips diacritics for convenience may improve some rough matching operations but lose essential distinctions.

CVNSS4.0 responds to this problem by converting diacritic-rich surface writing into an explicit code in which tone and rhyme behavior can be traced. The model does not eliminate Vietnamese diacritics conceptually; it relocates their information into a linear, auditable channel. This is similar in spirit to systems that externalize tone using final letters or combining marks, but the computational aim is more explicit: traceability, roundtrip testing, and domain-specific data governance.

3. Research design and methods

This article is a theoretical and design-oriented study. It does not report a completed large-scale empirical benchmark; instead, it defines a testable research program. The method combines historical-comparative analysis, writing-system typology, formal modeling, and AI pipeline architecture. The expected result is not a claim that CVNSS4.0 is already superior, but a framework for deciding whether it adds measurable value.

The primary technical reference used for the model is a CVNSSConverter runtime identified as version 4.1.0-pro, whose ruleset is described as “CVNSS4.0 official formula + 56-rime patch + GI/i reverse fix.” The converter exposes three modes—CQN, CVN, and CVSS—and contains mappings for Vietnamese initials and vowels, including patches for longer rhymes. It also acknowledges that CQN to CVN/CVNSS is the deterministic direction, while reverse conversion can require dictionary or contextual support because CVNSS4.0 may contain ambiguous codes. This limitation is theoretically important: it makes audit flags and roundtrip metrics necessary rather than optional.

The comparative corpus of writing systems is not intended to enumerate every language in the world. Such enumeration would be impossible in a 17–20 page article. Instead, the study uses a global typological matrix of representative cases: Latin-based tonal orthographies, syllabaries, abugidas, abjads, missionary romanizations, Indigenous American languages, African scripts, and East/Southeast Asian systems. This design is appropriate because CVNSS4.0 is compared by function, not by genealogy.

3.1 Formal model of the Vietnamese syllable in CVNSS4.0

Let a Vietnamese syllable be represented as an ordered tuple consisting of surface form, normalized form, phonological slots, transformed code, and audit metadata. The model is written as follows:

$$S_i = \langle raw_i, clean_i, I_i, M_i, N_i, C_i, T_i, code_i, A_i \rangle \quad (1)$$

where I is onset, M is medial or glide behavior, N is nucleus, C is coda, T is tone, and A is audit metadata. The rhyme R can be decomposed as R = M + N + C + T, but CVNSS4.0 treats this decomposition operationally: a rhyme is not only an analytic category but also a conversion unit.

The double transformation of the rhyme can be expressed as a composition of functions:

$$CVNSS(x) = tauT \circ tauC \circ tauV2 \circ tauV1 \circ tauI\ (x) \quad (2)$$

Here, τI normalizes the onset, τV1 converts the visible Quốc Ngữ vowel/rhyme into a structural rhyme module, τV2 maps selected vowel behavior into glide-aware j/w-style codes, τC handles coda normalization, and τT linearizes tone. This order is conceptual: actual implementations may combine rules for efficiency, but the audit layer should preserve their logical separability.

$$RTR = (1/N) * SUM_i\ I[decode(encode(clean_i)) = clean_i] \quad (3)$$

Equation (3) defines the roundtrip rate. A system may be useful even when RTR is below 1.0, but every failure must be localizable. Therefore audit completeness is defined as:

$$AC = SUM_i\ SUM_k\ present(a_{ik}) / (N*K) \quad (4)$$

where K is the number of required audit fields, such as rule_id, input span, output span, confidence, converter version, ambiguity flag, and human-review status.

Table 2. Illustrative CVNSS4.0 conversions produced by the uploaded converter runtime

Quốc Ngữ input	CVNSS/CVSS output from converter	Structural observation
Việt Nam	Vidf Nam	Tone and coda are linearized; “Việt” becomes a compact audit code.
tiếng Việt	tizb Vidf	The rhyme involving iê/iêng is transformed into a patched code.
nguyễn	wylg	Initial ng/ngh maps to w; the uyên-rhyme and tone are encoded.
tuyết	tydb	The 56-rime patch behavior is visible in the uyêt/uyết family.
mượn	mulh	ươn + nặng is represented by a compact code sequence.
hướng	huzx	ương + sắc is linearized with coda and tone information.
xoay	xajp	OAY pattern and P-guard style behavior can be audited.
gì	jil	GI/i reverse-fix class is relevant for correct decoding.

3.2 Metrics for AI/NLP/LLM evaluation

For model training, CVNSS4.0 can be evaluated through preprocessing metrics and downstream metrics. A domain-specific LLM does not need CVNSS4.0 to replace Unicode. Instead, it may use CVNSS4.0 as a parallel feature channel. A hybrid representation can be written as:

$$z_i = concat(EmbUnicode(x_i), EmbCVNSS(c_i), Tone_i, Rhyme_i, Audit_i) \quad (5)$$

A training objective may add orthographic and audit constraints to the task loss:

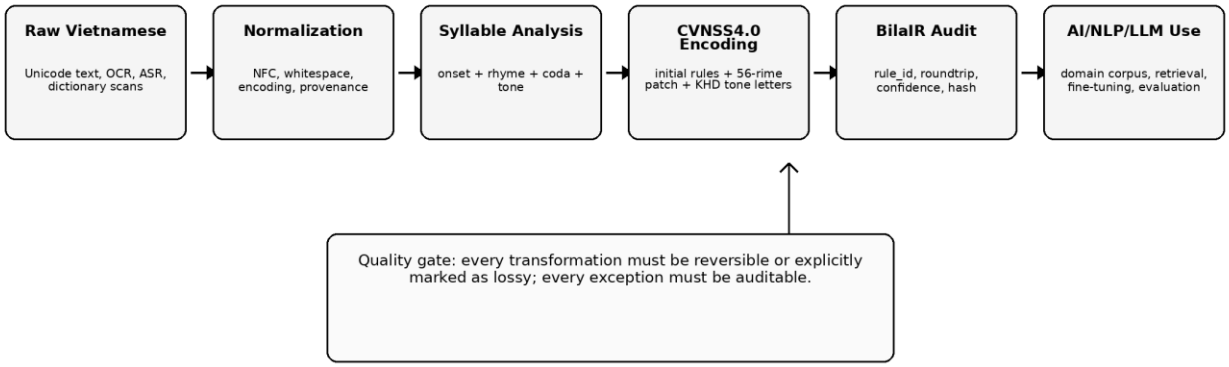
$$L = Ltask + lambda1*Lroundtrip + lambda2*Ltone + lambda3*Laudit + lambda4*Lcollision \quad (6)$$

This formulation is especially useful for domain-specific corpora: dictionaries, legal documents, administrative records, medical forms, educational materials, historical archives, OCR scans, and ASR transcripts. In these domains, a model must not only produce fluent text; it must preserve exact names, tonal distinctions, provenance, and document integrity.

4. Results: Proposed CVNSS4.0 orthographic-audit architecture

The result of the design analysis is an architecture in which CVNSS4.0 functions as a bridge between historical orthography and computational governance. The architecture contains three planes: a cultural-historical plane, a formal orthographic plane, and an AI/NLP/LLM plane. These planes should not be collapsed. A culturally legitimate model must respect Quốc Ngữ and Vietnamese linguistic history; a formally useful model must expose syllable structure; and a computationally useful model must support audit and evaluation.

Figure 1. CVNSS4.0 orthographic-audit pipeline for Vietnamese AI/NLP/LLM



Source: Author's conceptualization based on CVNSSConverter v4.1.0-pro and writing-system typology.

Figure 1. CVNSS4.0 orthographic-audit pipeline for Vietnamese AI/NLP/LLM.

4.1 The double transformation of the Vietnamese rhyme

The key result is the formalization of double rhyme transformation. In ordinary Quốc Ngữ, the vowel is a visual center: it carries diacritic shape, tone mark, and often the cue for syllable interpretation. In CVNSS4.0, that visual center is decomposed. First, the surface vowel/rhyme is recognized as a structural module. Second, the module is converted into a linear code that may include j/w-like behavior, coda information, and tone markers.

This double transformation is not arbitrary. In phonology, high front vowels are often related to palatal glides, while high back rounded vowels are often related to labiovelar glides. Pinyin uses y and w in spelling rules for i/u-initial finals; Mapudungun orthographies also show relations between high vowels and y/w in syllable positions. CVNSS4.0 uses this cross-linguistic intuition in a Vietnamese-specific way. The symbols j and w do not need to be read as literal IPA values in every context; they are internal code operators that make rhyme behavior auditable.

The model therefore shifts the status of the Vietnamese vowel. It is no longer merely the location of a tone mark. It becomes a computational unit that can be transformed, hashed, compared, and traced. This is the central orthographic innovation of the proposed model.

Table 3. Global typological matrix for comparing CVNSS4.0 with representative writing systems

Region / family	Language or script	Relevant feature	Analogy to CVNSS4.0	Difference from CVNSS4.0
East Asia	Pinyin	i/u finals may surface with y/w; tone can be marked over vowels or by numbers.	Shows positional transformation of finals and the use of glide-like spelling behavior.	Pinyin is a public romanization for Mandarin, not an audit IR for Vietnamese.
Korea	Hangul	Syllable blocks combine onset, nucleus, and coda.	Supports the view that syllables can be structured as formal units.	Hangul is a primary national script with geometric blocks.
Southeast Asia	Hmong RPA	Final letters represent tone categories.	Similar to linearizing tone away from diacritics.	It is a community orthography, not a secondary audit layer.
South Asia	Brahmic abugidas	Dependent vowel signs attach to consonants.	Vowels function as structural modifications, not mere independent letters.	CVNSS4.0 remains Latin-linear rather than abugida-based.
Middle East	Arabic/Hebrew abjads	Vowels may be absent or diacriticized.	Shows separation between consonantal skeleton and vocalic layer.	CVNSS4.0 preserves Vietnamese vowel information explicitly.
West Africa	Yorùbá	Tone and underdot diacritics are essential for meaning and NLP.	Confirms the computational importance of diacritics in tonal languages.	Yorùbá uses diacritics in the public orthography.
West Africa	Ewe	Tone and nasalization may combine on vowels.	Vowels serve as hubs for several phonological layers.	CVNSS4.0 relocates rather than stacks these layers.
West Africa	N'Ko	Tone marks, vowel length and nasalization are encoded with script-specific signs.	Treats tone and vowel features as first-class data.	N'Ko is an independent script for Mande languages.
West Africa	Vai	Syllabary encodes CV syllables.	Shows syllable-centered encoding.	CVNSS4.0 uses analytical linear codes rather than glyphic syllables.
West Africa	Mende Kikakui	Unicode describes 185 CV syllabic characters and 12 vowels.	Supports syllable modules as legitimate script units.	Mende Kikakui is right-to-left and glyph-based.
South America	Mapudungun	w/y are related to high-vowel/glide behavior in orthographies.	Supports the i/u ↔ y/w analogy for rhyme transformation.	Mapudungun is unrelated to Vietnamese typologically and historically.
South America	Quechua	Vowel systems may be represented by reduced phonemic cores.	Supports the idea of normalizing surface variation to a structural layer.	Quechua does not use Vietnamese tone or CVNSS-style audit.
South America	Aymara	Vowel deletion occurs under phonological/morphological conditions.	Shows that vowels can be dynamic rather than fixed characters.	Aymara morphology differs sharply from Vietnamese.
South America	Guaraní	Nasal harmony spreads from nasal vowels or consonants.	Shows vowels carrying suprasegmental layers that must be modeled.	Nasal harmony is not equivalent to Vietnamese tone.
North America	Canadian Aboriginal Syllabics	Glyph orientation or form encodes vowel quality in syllabic signs.	Demonstrates graphic transformation of vowel information.	CVNSS4.0 transforms by code rather than glyph rotation.
North America	Cherokee	Syllabary encodes syllables rather than alphabetic segments.	Supports syllable-level writing units.	Cherokee is an independent community script.

4.2 Orthographic audit stack

The audit stack proposed here consists of six layers. Layer 1 is the historical multiscript ecology of Vietnamese. Layer 2 is Quốc Ngữ, the socially dominant script. Layer 3 is phonological orthography, where onset, nucleus, coda, tone, and glide behavior are analyzed. Layer 4 is CVNSS4.0/CVSS, where these components become compact linear codes. Layer 5 is provenance and audit. Layer 6 is domain AI/LLM use. The stack is shown in Figure 2.

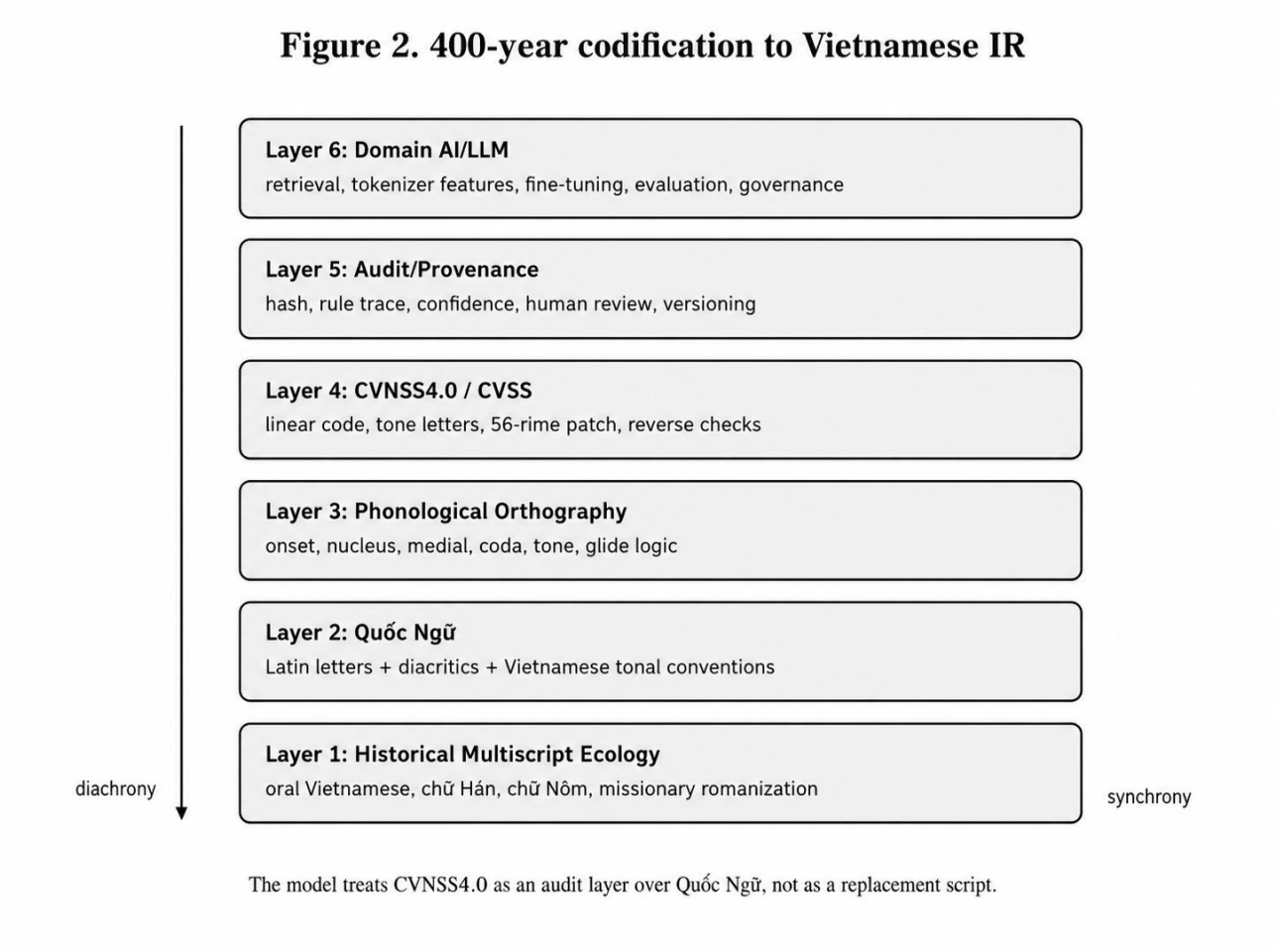


Figure 2. Historical and computational stack of CVNSS4.0 as an audit layer over Quốc Ngữ.

Table 4. Proposed CVNSS4.0 data pipeline as an auditable IR stack

Layer	Input	Transformation	Output	Audit requirement
Raw	Original user/document string	No destructive editing	raw_text	Preserve exact source and provenance.
Clean	Raw text	NFC, whitespace, encoding repair	clean_text	Record normalization version and changed spans.
Syllable	Clean text	Segment into onset/rhyme/coda/tone	syllable_IR	Mark ambiguity and segmentation confidence.
CVNSS4.0	syllable_IR	Apply initial, rhyme, coda, tone rules	cvnss_code	Store rule_id, converter version, and exceptions.
Roundtrip	cvnss_code	Decode and compare with clean_text	pass/fail/partial	Localize loss or ambiguity.
AI layer	Unicode + CVNSS + audit	Embedding, retrieval, fine-tuning, evaluation	domain model features	Maintain dataset card and reproducibility metadata.

4.3 Benchmark design

A benchmark for CVNSS4.0 should not be limited to conversion accuracy. The architecture must be evaluated along five axes: accuracy, reversibility, auditability, robustness, and downstream utility. A converter that produces compact strings but cannot explain failures is insufficient for scientific or administrative use. Conversely, a converter that exposes failures honestly may be valuable even when some reverse decoding requires dictionaries or contextual models.

The benchmark should include clean dictionary entries, modern news text, administrative records, OCR outputs from historical documents, ASR transcripts, social media text, domain-specific corpora, and multilingual mixed strings. It should also include adversarial cases: names, abbreviations, code-switched text, emoji, URLs, mathematical notation, obsolete spelling, regional words, and scanned dictionary artifacts.

Table 5. Proposed research questions, baselines, and metrics for CVNSS4.0 evaluation

Research question	Task	Baseline	CVNSS4.0 condition	Metrics
RQ1	Does CVNSS4.0 preserve Vietnamese orthographic information?	Unicode-only normalization	CVNSS encode/decode with audit	RTR, tone preservation, coda preservation
RQ2	Does the audit layer improve error localization?	Regex pipeline without trace	rule_id + span trace + confidence	Audit completeness, error localization score
RQ3	Does CVNSS4.0 help noisy text processing?	OCR/ASR correction without CVNSS	CVNSS-assisted diagnosis	Precision, recall, false positives, human-review reduction
RQ4	Does CVNSS4.0 improve domain corpora?	Unicode-only tokenizer features	Unicode + CVNSS feature channel	NER/F1, retrieval MRR, terminology consistency
RQ5	Does it help LLM governance?	Unlogged preprocessing	logged preprocessing with hashes	Reproducibility score, dataset-card completeness

5. Discussion

5.1 CVNSS4.0 and the 400-year continuity of Vietnamese linguistic engineering

The most important interpretive result is cultural: CVNSS4.0 is best understood as the next technical layer in a long history of Vietnamese linguistic engineering. The seventeenth-century Jesuit codifiers transformed Vietnamese speech into Roman orthographic and lexicographic form. Their work was embedded in missionary and colonial power relations, but it also generated a durable technology of literacy. Four hundred years later, the question is not whether Vietnamese should again be transformed by an external system, but whether Vietnamese can control its own computational representation.

This distinction changes the normative interpretation of CVNSS4.0. If CVNSS4.0 were presented as a replacement for Quốc Ngữ, it would invite cultural resistance and methodological confusion. If it is presented as an audit IR, it becomes a tool for strengthening Quốc Ngữ in digital environments. It allows Vietnamese text to travel through OCR, ASR, tokenizers, vector stores, and LLM fine-tuning pipelines without losing a trace of how it was normalized. In this sense, CVNSS4.0 is not anti-historical; it is historically aware.

The same logic applies to other language communities. The Jesuit grammars of Tupi, Guaraní, Quechua, Aymara, and Vietnamese were early examples of cross-cultural linguistic modeling. Modern AI must avoid repeating the extractive logic of early codification. A responsible CVNSS4.0 project should therefore include open documentation, versioned rules, reproducible benchmarks, community review, and explicit acknowledgment of the historical source layers of Vietnamese language.

5.2 Comparative implications for Indigenous and minority languages

The comparative evidence suggests that CVNSS4.0 can contribute to a broader theory of minority-language AI. Many languages do not fit cleanly into English-centered tokenization assumptions. They may be tonal, syllabic, agglutinative, polysynthetic, nasal-harmony based, or written in scripts where diacritics are essential. A model architecture that treats diacritics as expendable noise will damage

such languages. A model architecture that treats orthographic features as structured metadata can help preserve them.

For African languages such as Yorùbá and Ewe, the problem is not merely the existence of diacritics but the computational habit of omitting them. For N'Ko, tone, nasalization, and vowel length are encoded within a script system that requires appropriate rendering and normalization. For Mende Kikakui and Vai, syllabic encoding challenges alphabet-centric preprocessing. For Guaraní, nasal harmony requires modeling a suprasegmental feature. For Mapudungun, Quechua, and Aymara, orthographic debates show the tension between phonemic abstraction, colonial spelling, and community usage. CVNSS4.0 belongs to this global field of language technologies that must respect internal structure rather than flatten it.

The claim, therefore, is not that Vietnamese is identical to these languages. The claim is that CVNSS4.0 operationalizes a universal design principle: when a language carries meaning through tone, vowel quality, coda, glide, nasalization, or syllable structure, preprocessing must preserve those features as first-class data.

5.3 Domain-specific AI/NLP/LLM infrastructure

CVNSS4.0 is most promising for domain-specific models rather than general-purpose chat systems alone. In legal, medical, administrative, educational, historical, and dictionary domains, exact orthography matters. A wrong tone mark in a personal name, place name, legal clause, or dictionary headword can produce retrieval failure or semantic error. A model trained with parallel Unicode and CVNSS features may learn a more robust association between surface orthography and syllable structure.

A practical training pipeline can be organized as follows: raw corpus collection; Unicode normalization; syllable segmentation; CVNSS4.0 encoding; audit log generation; dictionary and gazetteer alignment; noisy-channel augmentation; tokenization ablation; model training; evaluation; and error analysis. The ablation should compare Unicode-only, CVNSS-only, and hybrid Unicode+CVNSS conditions. The expected benefit is not always higher benchmark accuracy; it may be better explainability, improved retrieval for noisy inputs, lower human review cost, and stronger dataset governance.

The model also supports domain-specific retrieval-augmented generation. A search index may store Unicode strings for display, CVNSS codes for normalized matching, phonological features for fuzzy search, and audit hashes for provenance. When a user asks a question over a scanned historical dictionary, the system can retrieve entries not only by surface spelling but also by structurally similar CVNSS forms, while preserving the original scanned string and correction history.

Table 6. Domain-specific uses of CVNSS4.0 in Vietnamese AI/NLP/LLM infrastructure

Domain	Typical risk	CVNSS4.0 contribution	Evaluation signal
Historical dictionaries	OCR errors, archaic spelling, merged entries	Preserve raw scan, clean entry, CVNSS headword, rule trace	Headword recovery, audit recall, duplicate detection
Administrative records	Name/place-name tone errors	Tone-aware normalization and roundtrip checks	Entity consistency, false match reduction
Legal texts	Exact citation and clause integrity	Hash and provenance at paragraph/sentence level	Reproducibility score, citation fidelity
Medical forms	Abbreviations and names with diacritics	Human-review flags for ambiguous conversions	Safety review reduction, precision
ASR transcripts	Tone and coda confusion	Syllable-level diagnosis and repair candidates	WER/CER by tone/coda class
OCR archives	Broken diacritics, Unicode composition mismatch	NFC + CVNSS audit alignment	Normalization accuracy, collision analysis
Education	Learner errors in tone/vowel/coda	Explainable feedback by rule_id	Error-type classification accuracy
LLM fine-tuning	Hallucinated spellings and unstable tokenization	Parallel Unicode/CVNSS features and audit loss	Downstream F1, hallucination rate, audit completeness

5.4 Limitations and risks

The proposed model has several limitations. First, CVNSS4.0 requires a stable, public, versioned specification. If different runtimes implement different rule tables, the audit layer will become a source of inconsistency rather than trust. Second, reverse conversion is not always deterministic; homography and code collisions require dictionaries, context, or probabilistic disambiguation. Third, dialectal variation cannot be solved by orthographic transformation alone. Fourth, historical and Indigenous language comparisons must not be used superficially; they should inform design principles without erasing the specificity of each community.

There is also a cultural risk. Because Quốc Ngữ is tied to modern Vietnamese literacy and identity, any technical proposal that appears to replace it may be misunderstood. The article therefore insists on a strict distinction: Quốc Ngữ is the public writing system; CVNSS4.0 is a machine-facing audit representation. This distinction should be built into documentation, software interfaces, educational materials, and research claims.

Finally, the strongest scientific risk is overclaiming. CVNSS4.0 cannot by itself make AI understand Vietnamese pragmatics, metaphor, omitted subjects, politeness, kinship hierarchy, satire, or socio-cultural implication. It can only improve the lower layers of the pipeline: orthographic traceability, syllable structure, tone preservation, and preprocessing audit. These lower layers are necessary but not sufficient for deep language understanding.

5.5 Jesuit missionary linguistics as computational prehistory

A central comparative argument of this paper is that missionary linguistics can be interpreted as a pre-digital form of language engineering. Early modern Jesuits did not merely translate religious doctrine; they built grammars, dictionaries, catechisms, alphabets, spelling conventions, and pedagogical protocols. These artifacts functioned as infrastructure. They converted speech into teachable, printable, searchable, and administratively usable forms. In contemporary terminology, they produced data models, annotation categories, lexical databases, and interface conventions.

The Vietnamese case is especially revealing because the later success of Quốc Ngữ was not inherent in the first missionary transcriptions. It required centuries of social adoption, educational policy,

printing, journalism, nationalism, and institutional standardization. This means that a technical representation becomes culturally powerful only when it is embedded in social practice. CVNSS4.0 should therefore avoid the illusion that a code table alone can transform language technology. It needs tools, documentation, public test cases, governance, and community interpretation.

The Americas provide an instructive comparison. Anchieta's Tupi grammar, Montoya's Guaraní works, Bertonio's Aymara grammar, and González Holguín's Quechua grammar and vocabulary show how missionary linguistics made Indigenous languages visible to European scholastic categories. The result was double-edged: it preserved records and enabled literacy, but it also imposed external categories. CVNSS4.0 must learn from this history by making its categories explicit, revisable, and accountable to Vietnamese linguistic reality rather than merely convenient for machines.

From an AI perspective, the missionary grammar-dictionary model resembles the modern dataset-model loop. A grammar defines permissible structure; a dictionary maps lexical items; examples establish usage; correction practices create feedback; and the resulting corpus shapes future learning. This is precisely what domain-specific LLMs require. Without a grammar-like constraint layer and dictionary-like lexical layer, a model may hallucinate plausible-looking Vietnamese while failing on names, tones, historical spellings, and technical terminology.

The novelty of CVNSS4.0 is not that it repeats missionary romanization. Its novelty is that it reverses the direction of control. The seventeenth-century model translated Vietnamese into European scholarly media. A contemporary Vietnamese AI infrastructure can translate Vietnamese orthographic knowledge into machine media while retaining local control over rules, validation, and purpose.

Table 7. Missionary grammar-dictionary artifacts and their computational analogies

Missionary / historical artifact	Language or region	Infrastructure function	Computational analogy for CVNSS4.0
Anchieta grammar	Old Tupi / Brazil	Converted oral language knowledge into a teachable grammar.	Rule specification and pedagogical grammar for preprocessing.
Bertonio grammar	Aymara / Andes	Mapped non-European grammar into explicit categories.	Feature schema and category inventory.
González Holguín grammar and vocabulary	Quechua / Peru	Linked grammar with lexicographic infrastructure.	Dictionary-core plus morphological/phonological annotation.
Ruiz de Montoya grammar and dictionary	Guaraní / Paraguay	Stabilized a missionary-linguistic corpus and lexical reference.	Domain corpus with controlled lexicon and examples.
Rhodes dictionary and Brevis Declaratio	Vietnamese / Rome 1651	Codified Vietnamese Roman spelling, tones, and lexical entries.	Historical precedent for tone-aware orthographic modeling.
CVNSS4.0 runtime and audit schema	Vietnamese / AI era	Transforms Quốc Ngữ into auditable machine representation.	Intermediate representation for NLP/LLM training, retrieval, and validation.

5.6 Audit schema, provenance, and data governance

For CVNSS4.0 to be scientifically useful, every conversion must carry provenance. Provenance is not an administrative afterthought; it is the condition of reproducibility. A Vietnamese string that has passed through OCR, Unicode normalization, CVNSS conversion, dictionary correction, and LLM augmentation should not be treated as a single undifferentiated text. It should be treated as a chain of transformations.

A minimal provenance tuple can be expressed as follows:

$$P_i = \langle source_i, raw_i, clean_i, rule_i, version_i, confidence_i, reviewer_i, hash_i \rangle \quad (7)$$

The purpose of this tuple is to prevent silent corruption. If a scanned dictionary entry changes from one form to another, the system should know whether the change came from Unicode normalization, OCR correction, manual review, CVNSS conversion, or automatic LLM suggestion. Each transformation should be independently reversible or explicitly marked as lossy.

In LLM pipelines, this provenance layer becomes a safety layer. It allows a model builder to audit whether training data came from reliable dictionaries, noisy social media, administrative records, or historical scans. It also supports red-team evaluation: when a model fails on a tone-sensitive term, the audit record can identify whether the failure was introduced in OCR, tokenization, conversion, retrieval, or generation.

Table 8. Minimal audit schema for CVNSS4.0 as a machine-facing IR

Field	Type	Required?	Purpose	Failure mode if absent
raw	string	Yes	Preserve the exact source string.	Irreversible data loss.
clean	string	Yes	Store normalized Unicode form.	Hash instability and duplicate confusion.
cvnss	string	Yes	Store CVNSS4.0/CVSS representation.	No structural matching or audit code.
rule_id	array	Yes	List rules applied to each span.	Transformation becomes unexplainable.
version	string	Yes	Lock converter and rule version.	Results cannot be reproduced.
confidence	float	Yes	Indicate automatic certainty.	No human-review prioritization.
roundtrip	enum	Yes	Pass, fail, partial, unsupported.	Loss cannot be detected.
ambiguity	array	Conditional	Record collisions or multiple readings.	Ambiguous forms appear falsely resolved.
hash	string	Yes	Integrity and deduplication.	No stable identity across pipeline stages.
review_status	enum	Conditional	Track human validation.	Unverified data enters gold corpora.

5.7 Architecture for domain-specific Vietnamese LLMs

A domain-specific Vietnamese LLM can use CVNSS4.0 in three ways. The first is as a preprocessing validator: before training, the corpus is checked for Unicode stability, tone integrity, and roundtrip behavior. The second is as a parallel representation: the model sees both Quốc Ngữ and CVNSS features. The third is as an evaluation lens: errors are classified by onset, rhyme, coda, tone, dictionary status, and provenance category rather than by surface character mismatch alone.

The strongest near-term use is retrieval and dictionary-grounded generation. In Vietnamese legal, administrative, medical, and historical datasets, retrieval often fails because surface forms vary by diacritics, OCR errors, old spelling, or inconsistent segmentation. CVNSS4.0 can provide an auxiliary index that groups structurally related forms without erasing the original display form. This supports human-readable output while improving machine matching.

For deep training, the hybrid representation should be tested through ablation. Unicode-only models should be compared with CVNSS-only and Unicode+CVNSS models. The expected effect may vary by task. CVNSS-only may not be ideal for semantic tasks requiring ordinary lexical distribution, but Unicode+CVNSS may improve robustness in noisy spelling, tone restoration, dictionary linking, and syllable-level explainability.

```
Index(q) = TopK(alpha*simUnicode(q,d) + beta*simCVNSS(q,d) +
                gamma*simAudit(q,d))    (8)
```

Equation (8) expresses a hybrid retrieval score. The weights α, β, and γ can be learned or tuned by domain. In historical dictionaries, β and γ may be high because structural matching and provenance are important. In ordinary news retrieval, α may dominate. This tunability is central to treating CVNSS4.0 as infrastructure rather than ideology.

Table 9. CVNSS4.0 augmentation in the lifecycle of domain-specific Vietnamese LLMs

Pipeline phase	Unicode-only weakness	CVNSS4.0 augmentation	Expected benefit
Corpus ingestion	Different Unicode forms may look identical.	NFC + hash + rule trace.	Stable corpus identity.
OCR correction	Diacritic errors are hard to classify.	Tone/rhyme/coda error labels.	Better diagnosis and review.
Tokenization	Subword tokens may split Vietnamese syllables unpredictably.	Syllable-aware auxiliary features.	More interpretable segmentation.
NER and gazetteers	Names with missing tones collide.	CVNSS structural index with original display preserved.	Higher recall without losing exact form.
RAG retrieval	Surface spelling variation reduces matching.	Hybrid Unicode + CVNSS retrieval.	Robust noisy-query retrieval.
Fine-tuning	Model learns surface correlations only.	Parallel code and audit losses.	Better tone and orthographic consistency.
Evaluation	Character error hides linguistic type.	Error by onset/rhyme/coda/tone.	Actionable error analysis.

5.8 Scopus-oriented contribution and falsifiability

For publication in an indexed venue, the proposal must be framed as a falsifiable research program rather than an advocacy text. The strongest version of the article is therefore not "CVNSS4.0 is better than Quốc Ngữ," which would be culturally and scientifically misleading, but "CVNSS4.0 increases the auditability and robustness of selected Vietnamese NLP preprocessing tasks under measurable conditions." This statement can be tested, rejected, or refined.

A Scopus-oriented contribution should also distinguish theoretical novelty from engineering novelty. The theoretical novelty is the formulation of Vietnamese rhyme transformation as a two-step orthographic-audit operation. The engineering novelty is the conversion of that operation into versioned rules, roundtrip tests, and feature channels for AI systems. The cultural novelty is the reinterpretation of Vietnamese romanization history as a long continuum of language infrastructure rather than a completed historical episode.

The model would be falsified if it fails to preserve tone and coda information, produces excessive collisions without explainable ambiguity flags, does not improve audit completeness over simpler baselines, or creates more human-review burden than it removes. These falsification criteria are essential. Without them, CVNSS4.0 would remain an interesting notation, but not a scientific artifact.

The paper therefore recommends that future work publish the converter rules, a gold test set, a noisy test set, and an error taxonomy. Each release should include a changelog and model card. This practice would align CVNSS4.0 with contemporary expectations in reproducible NLP and with the ethical need to avoid opaque language technology for low-resource and culturally significant languages.

Table 10. Falsifiability conditions for CVNSS4.0 as a scientific artifact

Claim	Evidence needed	How to falsify	Minimum artifact
CVNSS4.0 preserves Vietnamese syllable structure.	High roundtrip rate and low tone/coda loss on gold data.	Show systematic loss of tone, coda, or rhyme distinctions.	gold_tests.jsonl + converter report
CVNSS4.0 improves auditability.	Higher audit completeness than Unicode-only or regex baselines.	Show that rule trace does not localize errors better than baseline.	audit_schema.json + error logs
CVNSS4.0 helps noisy OCR/ASR data.	Better error localization or correction suggestions on noisy corpora.	Show no gain or more false positives in noisy data.	noisy_benchmark.jsonl
Hybrid Unicode+CVNSS helps domain models.	Ablation gains in retrieval, NER, spelling correction, or dictionary linking.	Show equal or worse performance under controlled ablation.	training/eval scripts + dataset card
CVNSS4.0 is culturally responsible.	Clear statement that it is an IR, not a replacement script; community review.	Show social confusion or harmful replacement claims in documentation.	governance note + usage guide

6. Conclusion

This article has proposed CVNSS4.0 as an experimental orthographic-audit writing model for Vietnamese. The central argument is that CVNSS4.0 should not be evaluated as an alternative public script, but as a machine-readable intermediate representation that exposes the structure of Vietnamese syllables and makes text transformation auditable. Its core operation is the double transformation of the rhyme: from visually diacritized Quốc Ngữ vowels into structural modules, and from those modules into linear codes that preserve onset, rhyme, coda, tone, and rule trace.

The comparative analysis shows that this design is not isolated. Across the world, writing systems have repeatedly externalized vowel, tone, nasalization, length, syllable, and glide information. Pinyin, Hangul, Hmong RPA, Yorùbá, Ewe, N'Ko, Vai, Mende Kikakui, Mapudungun, Quechua, Aymara, Guaraní, and Canadian Aboriginal Syllabics all demonstrate different solutions to the same general problem: how to make phonological structure visible enough for a given social or technical purpose. CVNSS4.0 applies that principle to Vietnamese in the age of AI.

The historical discussion also shows that Vietnamese orthographic engineering has a long duration. From the seventeenth-century Jesuit grammar-dictionary paradigm to twenty-first-century AI infrastructure, Vietnamese has repeatedly entered new media through technical representation. The task now is to make that representation accountable, reproducible, and culturally respectful. CVNSS4.0 can contribute to this task if it is formalized through open specifications, versioned converters, benchmark datasets, audit schemas, and community review.

Future research should implement a gold-standard benchmark with parallel Unicode and CVNSS layers; measure roundtrip rate, collision rate, tone preservation, OCR/ASR robustness, and downstream domain performance; and test hybrid Unicode+CVNSS representations in retrieval, NER, spelling correction, dictionary alignment, and LLM fine-tuning. If these experiments show consistent gains in auditability and domain robustness, CVNSS4.0 may become a scientific artifact for Vietnamese language AI infrastructure: not a new alphabet for society, but a rigorous audit layer for machines.

References

Adewumi, T. P., Liwicki, F., & Liwicki, M. (2020). The challenge of diacritics in Yoruba embeddings. arXiv. https://arxiv.org/abs/2011.07605

Anchieta, J. de. (1595). Arte de grammatica da lingoa mais usada na costa do Brasil. Coimbra: Antonio Mariz.

Bertonio, L. (1603). Arte y grammatica muy copiosa de la lengua aymara. Roma: Luis Zannetti.

Boidin, C. (2020). Guarani language in the missions of Paraguay: Beyond linguistic description. In Cultural worlds of the Jesuits in colonial Latin America. University of London Press.

Fernandes, G., & Assunção, C. (2017). First codification of Vietnamese by 17th-century missionaries: The description of tones and the influence of Portuguese on Vietnamese orthography. Histoire Épistémologie Langage, 39(1), 155–176. https://doi.org/10.1051/hel/2017390108

Ganson, B. (2016). Antonio Ruiz de Montoya, Apostle of the Guaraní. Jesuit Higher Education: A Journal, 5(1), 69–76.

González Holguín, D. (1607). Gramatica y arte nueva de la lengua general de todo el Peru, llamada lengua Quichua, o lengua del Inca. Lima: Francisco del Canto.

Jimoh, T. Á., et al. (2025). Bridging gaps in natural language processing for Yorùbá: A systematic review. Natural Language Processing Journal, 11, 100153.

Orife, I., Adelani, D. I., Fasubaa, T., Williamson, V., Oyewusi, W. F., Wahab, O., & Tubosun, K. (2020). Improving Yorùbá diacritic restoration. arXiv. https://arxiv.org/abs/2003.10564

Orife, I. (2018). Attentive sequence-to-sequence learning for diacritic restoration of Yorùbá language text. Proceedings of Interspeech 2018.

Rhodes, A. de. (1651). Dictionarium Annamiticum Lusitanum et Latinum. Rome: Propaganda Fide.

Ruiz de Montoya, A. (1639). Tesoro de la lengua guaraní. Madrid: Juan Sánchez.

Ruiz de Montoya, A. (1640). Arte de la lengua guaraní. Madrid: Juan Sánchez.

Russell, K. R. (2022). Interactions of nasal harmony and word-internal prosody in Paraguayan Guaraní. Languages, 7(1), 67. https://doi.org/10.3390/languages7010067

The Unicode Consortium. (2026). The Unicode Standard, Version 17.0: Chapter 19, Africa-Ideographic and Syllabic Scripts. https://www.unicode.org/versions/Unicode17.0.0/core-spec/chapter-19/

Tran, A. N. H. Q. (2022). Catholicism and the development of the Vietnamese alphabet. Journal of Vietnamese Studies, 17(2–3), 60–88.

World Wide Web Consortium. (2026). N’Ko script resources. https://www.w3.org/TR/nkoo-lreq/

Zwartjes, O. (2011). Portuguese missionary grammars in Asia, Africa and Brazil, 1550–1800. John Benjamins.