Annotating Clinical Risk and Variation in Haitian Creole Medical Translation

Ludovic Mompelat
University of Miami
Miami, USA
ludovic.mompelat@miami.edu

David Tézil
University of Alabama
Tuscaloosa, USA
dtezil@ua.edu

Rose Flaure Accilien
University of Miami
Miami, USA
rxa1262@miami.edu

Abstract

We present an annotation schema for Haitian Creole medical translation that makes clinical risk and sociolinguistic variation explicit while remaining lightweight enough for small expert teams. The schema includes binary fields for overall acceptability, severity of potential misunderstanding, and foreign-influence cues, along with conditional error tags aligned with Multidimensional Quality Metrics (MQM), commonly used in the medical domain, for interoperability. Through three rounds of annotation and adjudication we achieve stable inter-annotator agreement and release a gold dataset of 152 EN→HC medical sentence pairs. A simple classifier–labeller baseline demonstrates that acceptability and severity are reliably learnable under data scarcity, while foreign-influence judgments remain limited by prevalence. These results show that clinically oriented, variety-sensitive annotation can both support immediate screening of patient-facing translations and provide reward-ready signals for future preference-based MT and LLM fine-tuning.

1 Introduction

Despite growing efforts in low-resource machine translation, and especially in medical contexts, Haitian Creole (HC) remains critically underrepresented (Mompelat, 2025). This further limits the availability and development of effective translation tools capable of adequately supporting the efforts of human interpreters and medical providers. In regions like Miami, which is a diasporic hub for HC speakers, healthcare disparities are exacerbated by language barriers, with existing translation systems failing to address intra-language variation, code-switching, and the absence of standardized medical terminology. The stakes in effective and adequate medical communication are particularly high since lexical choices and grammatical structures can directly affect patient comprehension, trust, adherence, and safety. Language barriers in healthcare have also been shown to reduce patient and provider satisfaction, compromise care quality, and increase miscommunication and associated costs (Pérez-Escamilla et al., 2010; Al Shamsi et al., 2020). In the particular case of HC, these risks are compounded by intra-linguistic variation across basilectal (Creole-exclusive) and mesolectal (French-influenced) varieties (Bickerton, 1973), and inter-linguistic contact with French, English and Spanish, particularly in Miami. Available HC medical data are uneven in format, size, and quality, making them difficult to aggregate or use directly for modeling. To address this, we began by normalizing and preprocessing the CMU Haitian Creole medical corpus[1], which revealed the need for a structured annotation framework. The schema we introduce in this paper guides both dataset cleaning and evaluation, capturing the sociolinguistic and clinical dimensions that matter most for safe and effective HC medical communication. This pilot is designed as a schema validation study under low-resource conditions and intended for interdisciplinary annotation efforts; simple enough for linguists, but precise and accurate enough to align with domain-specific standards of annotation such as Multidimensional Quality Metrics (MQM) used in the medical context. Annotation reliability and label learnability are therefore the primary objectives. To check practical utility, we train a small, interpretable classifier–labeller on the adjudicated EN→HC pairs with five-fold cross-validation. The model outputs a probability for each binary field and suggests error tags. In this setup, we find that (i) deciding whether a translation needs improvement is reliable; (ii) error detection is tractable once quality issues are identified but requires more instances of rarer terminology and orthographic errors; and (iii) detecting foreign-influence phe-

[1] http://www.speech.cs.cmu.edu/haitian/

Proceedings of the 20th Linguistic Annotation Workshop (LAW XX), pages 1–11

nomena remains data-limited at current coverage. Operationally, the *Severity* probabilities let reviewers look first at the most consequential cases, and the pre-filled error tags speed up adjudication.

2 Background and Motivation

2.1 Clinical Stakes of Language Choice and Access in Healthcare

When patients and providers do not share a language, communication breaks down and safety suffers. Recent reviews synthesize a consistent picture: language barriers drive misunderstandings, reduce satisfaction on both sides, and are linked to preventable errors and worse outcomes (Al Shamsi et al., 2020).

Professional interpreters remain the standard of care bridging the gap (often virtually) between patient and healthcare practitioners. Their use is associated with clearer communication, guideline-concordant treatment, and better outcomes (Karliner et al., 2007; Jacobs et al., 2004). Yet the availability of qualified interpreters is not guaranteed, particularly for low-resource languages where on-demand access is limited and scheduling delays are common. In these settings, clinicians often rely on ad-hoc interpreting, a practice shown to increase error rates and miscommunication, especially in pediatrics and emergency care; by contrast error rates fall significantly with trained interpreters (Flores et al., 2003, 2012; Divi et al., 2007).

Within that reality, technology often fills the gaps. A scoping review by Kreienbrinck et al. (2025) distinguishes two main tool types: fixed-phrase systems and Machine Translation apps. Phrasebooks and menu-driven tools can support brief, predictable exchanges but fail outside their limited inventories and require constant maintenance for domain coverage (Hudelson and Chappuis, 2024; Noack et al., 2021; Spechbach et al., 2019). MT systems, in turn, can handle arbitrary content and more languages, but accuracy remains inconsistent; studies commonly warn against unsupervised use in safety-critical communication, particularly beyond high-resource languages (Panayiotou et al., 2019; Halimi and Bouillon, 2019; Hwang et al., 2022). The consensus emerging from both research and clinical guidance is therefore clear: technology can mitigate access gaps when no interpreter is available, but it must operate under supervision and never replace qualified professionals.

Policy frameworks reinforce this balance between access and safety. The U.S. Department of Health and Human Services' National CLAS Standards mandate free language assistance, proactive notification of its availability, and interpreter competence (U.S. Department of Health and Human Services, 2013). The NIH Clear Communication Initiative and the CDC's "Everyday Words for Public Health" extend these principles to written materials, prescribing plain-language phrasing and comprehension testing with target audiences (National Institutes of Health, 2025; Centers for Disease Control and Prevention, 2016). Similarly, the World Health Organization emphasizes accessibility so that non-experts can act on health information (World Health Organization, 2021), and Translators without Borders (CLEAR Global) operationalize these ideals through multilingual glossaries and tools for crisis response (Translators without Borders, 2020). Together, these policies frame linguistic equity as both a legal and ethical duty rather than a technical convenience.

HC makes these tensions concrete. It exemplifies a language where interpreter availability is limited and technological coverage is uneven, forcing practitioners to rely on improvised, inconsistent solutions. In such settings, variety-appropriate phrasing becomes a matter of safety as well as clarity: what matters is not only that a translation is correct, but that it is intelligible and culturally legible to monolingual HC speakers who cannot fall back on French.

2.2 Haitian Creole and Medical Communication

HC is the native language of almost the entire population of Haiti; approximately 90–95% of Haitians speak it fluently, and for many, it is their only language (Dejean, 2010; Hebblethwaite, 2012). Globally, HC counts around 10–12 million speakers, making it the most widely spoken Creole (Valdman et al., 2017). In the United States, Haitian Americans number over 1.2 million, with substantial communities in South Florida (especially Miami and Orlando), New York City, and Boston (US Census 2024). In Florida specifically, HC ranks as the third most spoken language after English and Spanish, a fact reflected in local service provision and public communications.

Focusing on the city of Miami, one of the many problems faced by the HC community is access to appropriate care, and while many factors and vari-

ables come into play to explain the discrepancies in quality of and access to care, such as cultural literacy from the medical staff (Campbell, 2012), or financial/employment difficulties faced by the local community (in Miami, for example, and especially in Little Haiti) (Ryan et al., 2004; Kobetz et al., 2009, 2010; Menard et al., 2010), we are particularly interested in tackling the language barrier factor. Public agencies and health systems routinely provide language assistance and translated materials—including in HC—so that patients can access basic information and services. This is visible at the federal level (e.g., CDC language-assistance pages and multilingual health resources) and locally in Miami, where city and county plans specify notices and summaries in English, Spanish, and HC. In everyday practice, however, HC-specific tools remain patchy, so clinicians and interpreters often lean on improvised mixes of translation apps and glossaries. Additionally, during emergencies, agencies and partners have even circulated outdated or inaccurate purpose-built HC materials (e.g., cholera and earthquake response guides, triage glossaries), underscoring the need for up-to-date, domain resources. Finally, in the NLP space, HC MT is often routed via French as a pivot language, a standard low-resource strategy which at times exacerbates existing gaps in direct HC support or create new translation issues (Dholakia and Sarkar, 2014); major commercial systems like DeepL also do not list HC among supported languages.

2.3 Haitian Creole Linguistic Variation and Gaps in Current Models

HC exhibits systematic sociolinguistic variation often described along a continuum between *kreyòl swa* (prestige variety shaped by French contact) and *kreyòl rèk* (basilectal, Creole-like, monolingual variety) (Tezil, 2022; Valdman, 2015; DeGraff, 2005). This structural and lexical diversity matters particularly in the medical domain. For example, terms like *dyabèt* (from French *diabète*) are readily legible to bilinguals, but monolingual speakers may instead expect variants such as *maladi sik* ('sugar disease'). The linguistic continuum in HC yields multiple plausible translations for the same clinical concept, but with very different implications for comprehension and patient trust. For medication, *medikaman* suggests biomedical precision, whereas *remèd/renmèd* evokes a broader category that includes more-culturally tied folk or herbal remedies, requiring careful contextualization. The wrong lexical choice can introduce confusion or mistrust, even when the translation is literally accurate. (Valdman et al., 2017; DeGraff, 2005)

Other contrasts show the risks of outright misunderstanding. A provider's note about chest pain may be translated with *pwatrin* (chest) or *lestomak* (stomach), leading to clinically dangerous miscommunication. In substance-use screening, asking a patient if they are a *tafyatè* (literally 'rum/alcohol drinker') to ask if they drink alcohol (i.e. regularly, occasionally, or not) risks offense: the term is widely understood to mean 'alcoholic,' carrying stigma that can prevent disclosure. These examples underscore that in HC medical contexts, translation accuracy cannot be reduced to word-for-word fidelity. Audience design—deciding which variety and phrasing suit the patient population—becomes a clinical and critical safety issue. Our schema takes these linguistically-based observations into account by requiring annotators to score both correctness and record possible variety-specific variants of lexical items or phrases, producing potential signals, useful for language modeling, that surface when outputs are comprehensible but inappropriate for their intended audience.

Beyond variation, outdated lexicographic resources introduce practical constraints. Widely used dictionaries and phrasebooks predate post-2010 usage (e.g., not including the earthquake-era neologism *goudougoudou* for trauma/distress), and some entries reflect mesolectal/French bias or hybridized forms that are not transparent to younger or monolingual users; many sources also remain non-digital, complicating updates and integration into NLP pipelines (Freeman, 1997; Heuretelou et al., 2000).

Current MT systems for HC rarely capture the internal diversity of the language. They might default toward French-influenced lexical and syntactic patterns, which can alienate or confuse monolingual readers who are less familiar with acrolectal forms. Yet the inverse is also true: many "standard" basilectal forms used in official or corpus materials do not necessarily reflect how either monolingual or bilingual speakers actually speak in clinical or everyday settings. In practice, preferred usage often lies somewhere between the two poles. What counts as "plain" or "appropriate" HC depends on audience, region, and communicative setting. This fluidity makes it risky to treat HC as a homogeneous variety or to assume that basilectal vocabulary is always the most accessible (Lewis, 2010;

Mompelat, 2025).

These tensions are reflected in our annotated data, where fluency problems frequently co-occur with *foreign-influence* cues such as overly Frenchified function words, calques, or hybridized syntax—evidence that linguistic quality and sociolinguistic variation are intertwined. Our schema is designed to make this complexity explicit rather than to enforce a single normative variety. A sentence can display *foreign influence* without being wrong: in those cases, annotators may propose a basilectal variant when the translation is accurate but mesolectal or French-influenced, allowing for variety-sensitive alternatives. When *foreign influence* co-occurs with a *quality* issue, however, the influence often signals a deeper translation problem (e.g., forms that sound unnatural or implausible to any HC speaker).

3 Annotation Schema

We design a lightweight schema aimed at maximizing reliability in small expert teams while retaining the expressive power needed to capture three aspects central to HC medical communication: sensitivity to variety and contact-driven variety differences, explicit attention to clinical risk, and a minimal set of error tags aligned with widely used translation-quality taxonomies for downstream interoperability. Existing frameworks such as the Multimensional Quality Metrics (MQM) (Lommel et al., 2014) or the NCC MERP taxonomy (National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP), 2022) do not explicitly encode clinical risk *and* intralanguage variation, which are central in HC medical communication; our schema extends both minimally to capture these dimensions while preserving interoperability.

3.1 Guiding principles and rater instructions

Annotators worked from compact definitions, examples, and decision tests that privileged consistency over granularity. *Foreign influence* was judged by comparing outputs to reputable lexicographic sources (Valdman, 2015; Valdman et al., 2017) and by flagging recognizable contact-induced patterns in lexicon or morphosyntax (e.g., French-leaning terms, calques, complementizer *ke* in acrolectal styles). *Quality* was defined according to monolingual and bilingual HC speakers acceptability, grammaticality and understandability criteria. *Severity* was introduced as a separate, risk-oriented lens, grounded in clinical stakes rather than linguistic well-formedness (Flores et al., 2003, 2012).

Where a sentence was judged improvable (quality = 1), annotators had to provide a corrected version and could propose additional variants to the corrected forms that are as suitable depending on the audience. Raters also noted uncertainty, which we used to flag items for adjudication. Instructions were refined across rounds as systematic sources of disagreement emerged, with clarifications added for recurring borderline cases.

3.2 Labels and decision criteria

Foreign influence: binary field marking whether a HC output leans toward French- or English-influenced, or otherwise non-Creole, forms in its lexical choices or morphosyntax. The rationale is both sociolinguistic and diagnostic. On the one hand, HC exhibits socially conditioned contrasts often described along the *kreyòl swa–kreyòl rèk* continuum (Fattier-Thomas, 1984; Tézil, 2024), where the choice in the linguistic variety influences intelligibility and trust in patient-facing materials. On the other hand, *foreign influence* also captures translation-induced interference from source languages such as English negative polarity item behavior and negation structure as shown in example (1) and which corrected sentence is the negative concord sentence "men ou *pa* konnen *anyen ankò*".

(1) Men ou konnen okenn lòt bagay
but you know none other thing
'But you know nothing else.'

The label therefore serves a dual function: to flag variety alignment issues that may affect accessibility, and to detect morphosyntactic intrusions introduced by translators or machine systems.

Severity: binary risk label indicating whether the content of a sentence, if misunderstood or mistranslated, could plausibly alter a patient action or clinical decision (e.g., dosing, contraindications, symptom actionability, discharge instructions).

We draw on the NCC MERP Index, which organizes medication errors along a spectrum from potential risk (Category A) to patient harm and death (Categories E–I), but we do not attempt to reproduce these fine-grained distinctions (National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP), 2022). Instead,

we collapse this spectrum into a binary decision: sentences are marked as high-risk (*Severity*=1) if they involve information whose misinterpretation could have clinical consequences, regardless of whether that consequence would be minor or severe, and low/no risk (*Severity*=0) otherwise.

Crucially, this label does not assess whether a translation is erroneous, but whether the underlying *information type* carries potential clinical risk under miscommunication. We use, instead, the label *Error - terminology* for cases of clear erroneous domain-specific terminology. This design separates linguistic correctness from risk exposure and enables straightforward triage of patient-facing materials.

Quality: binary acceptability decision, clear and accurate for the intended audience vs. improvable. This criterion aligns with the clinical importance of plain-language communication and cultural fit. When a sentence is judged as needing improvement (i.e., *Quality* = 1), raters apply minimal *error tags* to diagnose the problem in MQM-compatible terms (Lommel et al., 2014): (i) terminology or accuracy problems, (ii) fluency or adequacy breakdown, and (iii) orthographic errors.

Example (2) is the proposed translation of a question on *casual* alcohol consumption illustrates how these dimensions interact but remain distinct. The translation uses *tafyatè* 'alcoholic', which is semantically inappropriate and socially negatively marked (i.e. someone diagnosed as suffering from alcoholism) in this context. This triggers *Quality* = 1 and a *terminology* error. At the same time, the sentence is marked *Severity* = 1 because misunderstanding or misframing alcohol use can affect patient disclosure and downstream care decisions. The example shows that *severity* is not reducible to linguistic error: it reflects the potential clinical consequences of miscommunication, even when the sentence is otherwise fluent.

(2) èske ou se yon tafyatè
Q 2SG COP DET alcoholic
'Are you an alcoholic?' (intended: 'Do you drink alcohol?')

Restricting error tags to improvable sentences reduces noise and mirrors clinical review workflows, where reviewers first decide whether the material is usable, and only then identify what needs fixing.

Finally, because multiple semantically equivalent lexical items and phrasings are possible and variety- or region-dependent in HC, we also record *variants*: a binary flag indicating whether a clearly viable alternative exists that preserves meaning but may better match the patient's idiolect (e.g., *bwason alkolize*, *tafya*, or *kleren* 'alcoholic drink').

In section 5, we apply the schema to a small, adjudicated EN→HC medical set and probe its modeling potential with a baseline classifier–labeller. Results show that acceptability detection is robust, error typing is tractable when quality issues are present, and foreign-influence detection is limited more by data coverage than by label design.

4 Annotation Workflow

4.1 Rounds and Sampling

Annotators were two bilingual Haitian Creole–English professionals: one certified medical interpreter with clinical experience in South Florida, and one trained linguist specializing in Haitian Creole variation and translation. Annotation was conducted in three successive rounds designed to iteratively refine the schema and improve inter-annotator agreement:

- **Phase 1**: 300 sentences were independently annotated by two annotators using the initial guidelines.
- **Phase 2**: 100 new sentences were introduced, with adjustments to field definitions and decision rules based on Phase 1 disagreements.
- **Phase 3**: 52 sentences from Phase 2 that were re-annotated after updating the schema, and 50 additional sentences were newly annotated to target categories that remained problematic, enabling focused reliability testing.

Adjudication followed each round. Disagreements and low-confidence cases were reviewed with reference to guidelines, lexicons, and audience assumptions. Recurring edge cases triggered refinements to instructions, reducing drift. The final round sampled additional material with an eye toward stress-testing categories that remained difficult such as foreign-influence phenomena and rare error types. All items were presented with the English source sentence, the HC hypothesis, and minimal discourse context. The pilot experiment later in the paper uses the 152 EN→HC medical pairs drawn from across Phases 2 and 3, after normalization and adjudication. The next annotation effort

will consist in re-annotating the 300 sentences from Phase 1.

4.2 Agreement and Adjudication

Reliability was assessed at the field level using Cohen's κ on raw labels prior to adjudication and Krippendorff's α for nominal data as a robustness check (Artstein and Poesio, 2008). Agreement is reported separately for each binary label; for error tags, we restrict computation to the subset where quality was judged improvable, consistent with the schema's conditional design.

Agreement improved over rounds. In Phase 2 we observe substantial reliability for severity ($\kappa = 0.709$) but only moderate to fair scores for quality (0.454), fluency (0.409), and foreign influence (0.251). By Phase 3, terminology rises sharply ($\kappa = 0.850$) after dictionary-backed clarifications, while quality (0.480) and typo (0.490) improve modestly. Severity declines somewhat (0.561) because Phase 3 targeted more ambiguous, higher-stakes content. Foreign influence remains difficult (0.203), reflecting both low prevalence and fuzzy sociolinguistic boundaries.

From the final adjudication pass, we compile a gold set with fully resolved labels across all schema dimensions. Disagreements were resolved collaboratively through guideline clarification and consensus review, ensuring that the dataset reflects adjudicated, reproducible judgments. The dataset is structured to support both sentence-level evaluation and span-based processing: corrected translations and audience-appropriate variants are stored as explicit token-level mappings, enabling fine-grained alignment and downstream use in sequence-to-sequence or correction modeling.

4.3 Toward Reward Modeling for Haitian Creole MT

Recent work in preference learning and reinforcement learning from human feedback (RLHF) shows how annotated judgments can steer large language models toward outputs that reflect human preferences and contextual appropriateness rather than surface likelihood alone (Wang et al., 2024; Dong et al., 2024; Ouyang et al., 2022). Within MT, reward modeling has been used to rerank candidate translations and fine-tune generation toward domain-specific criteria such as fluency, adequacy, and user-centered preferences (Kreutzer et al., 2018; Lyu et al., 2023).

The adjudicated labels—*quality* (acceptability), *error type* when problems occur, *foreign influence* as a signal of language contact influence and linguistic variation, and *severity* as a clinical-risk indicator—map naturally onto preference data and potential reward functions. In this framing, severity can serve as a cost-sensitive weight, quality as a binary accept/reject signal, and foreign influence as a dimension of idiolect appropriateness. Together, these labels provide the basis for both pairwise preference modeling and scalar reward training, and can be scaled through active learning once classifiers flag likely error cases.

In the pilot study that follows, we use the schema purely in a supervised setup with a simple classifier–labeller baseline. Results show that acceptability detection is reliable, error typing is tractable once quality issues are flagged, and foreign-influence judgments remain limited chiefly by data coverage. Future work will extend these same labels into reward modeling pipelines that bias generation toward patient-facing plain language and culturally appropriate renderings in HC medical communication.

5 Pilot Study

5.1 Setup and model

The baseline is intentionally simple and interpretable, serving as a lower bound on label learnability (Table 1).

Table 1: Classifier setup summary.

Component	Description
Features	TF–IDF (EN+HC), length, diacritics, overlap
Models	Logistic regression (per label), OVR for tags
Evaluation	5-fold Cross-validation, stratified
Outputs	Probabilities + binary labels

Each EN–HC pair is represented by a shared feature vector combining TF–IDF n-grams and lightweight numeric cues. We train one logistic-regression classifier per binary label (*quality*, *severity*, *foreign influence*, *variants*) and use a one-vs-rest setup for multi-label error tags. Evaluation uses five-fold cross-validation with stratification on the joint (*severity*, *quality*) label.

Feature design targets HC-specific signals while remaining interpretable: orthographic patterns (diacritics), length-based cues, and lexical overlap across source, hypothesis, correction, and variant capture alignment and variation without relying on contextual embeddings, which we leave for future

work.

5.2 Results

We compare against a prevalence-matched random baseline, which provides a lower bound under class imbalance. We report precision, recall, and F1 for each binary task, alongside the prevalence of positives (e.g., 18/152 for *foreign influence*).

Across labels, our model consistently exceeds this baseline, indicating that the annotation signals are recoverable rather than driven by label distribution. Because *severity* is intended to triage review, we compute a "limited-review recall": sentence pairs are ranked by their predicted probability of being high-risk, and we simulate reviewing only the top 10%, 20%, or 30% of that ranked list. The metric reports the proportion of all truly high-risk items recovered within each reviewed portion. This measures how well model scores prioritize scarce human attention. Finally, we compute macro-F1 and per-tag F1 for *fluency*, *terminology*, and *typo* on that subset.

Table 2 summarizes performance for the four binary labels. *Quality* is detected with high reliability (F1 = 0.94), and *variants* are captured well (F1 = 0.89). Performance on *severity* is solid (F1 = 0.77), above the random baseline (0.66), while *foreign influence* remains more limited (F1 = 0.36), reflecting both low prevalence (11.8%) and the difficulty of separating contact-induced forms from HC-internal variation.

For *severity*, probability outputs support ranking. Because high-risk cases represent 66.4% of the dataset, randomly inspecting 10% of sentences would recover about 10% of them. The model instead recovers 14.9%, rising to 28.7% and 42.6% at 20% and 30% inspection budgets, corresponding to a 1.42–1.49× improvement over random. High-risk cases are therefore concentrated near the top of the ranking, making the score effective for prioritization.

Error typing (Table 3) shows a clear dependence on label prevalence. *Fluency* errors are detected reliably (F1 = 0.85), while *terminology* is moderate (F1 = 0.35) and *typo* remains weak (F1 = 0.17), consistent with their relative scarcity. This pattern suggests that the model captures frequent structural issues more readily than rarer, lexically specific phenomena.

Variance across folds is low for the main tasks (e.g., *quality*: 0.941 ± 0.023; *severity*: 0.767 ± 0.044), indicating stable estimates. Higher variance for *foreign influence* (0.353 ± 0.216) reflects class imbalance and the limited number of positive instances.

Table 2: Binary Classification Tasks: positive counts and P/R/F1 (5-fold out-of-fold, n=152).

Label	Positives (#)	Prec.	Rec.	F1
Quality	53	0.980	0.906	0.941
Severity	101	0.837	0.713	0.770
Foreign infl.	18	0.296	0.444	0.356
Variants	54	0.920	0.852	0.885

Table 3: Error typing on the *quality*=1 subset (n=53): per-tag P/R/F1 and prevalence.

Tag	Positives (#)	Prec.	Rec.	F1
Fluency	45	0.864	0.844	0.854
Terminology	9	0.375	0.333	0.353
Typo	5	0.143	0.200	0.167
Macro-F1				0.458

6 Discussion

6.1 Linking Annotation, Sociolinguistics, and the Pilot

The schema was designed around two constraints distinctive to HC medical communication: contact-driven linguistic variation and clinical risk. The annotation rounds operationalized these with compact binary fields and conditional error tags, and the inter-annotator agreement profile revealed which judgments were inherently stable versus which demanded sharper guidance. The pilot then tested how learnable these labels are under data scarcity. Together, the findings line up: categories that humans applied consistently also supported robust automatic detection, while low-prevalence or conceptually fine-grained categories remain the main bottleneck.

Qualitatively, errors cluster into three recurring patterns. First, lexical misselection, including stigmatizing or overly specific terms (e.g., *tafyatè* 'alcoholic' instead of a neutral phrasing such as *èske w konn bwè* 'do you drink?'); these cases are typically captured under terminology errors but may also carry clinical risk. Second, morphosyntactic interference, often reflecting transfer from English or French (e.g., calqued terminology 'chest' #*pwatrin* vs. *lestomak* or compositional phrasing such as 'all three' **tout twa* vs. *touletwa*), which the model tends to capture under fluency. Third, register and variety mismatches, where otherwise

correct translations rely on forms that are not uniformly shared across speakers (e.g., 'how are you feeling' *èske sa va* vs. *èske ou anfòm*, or 'alcoholic beverage' *bwason alkolize* vs. *bwe lalkol*), frequently surfacing in the foreign-influence and variants fields. These patterns align with the quantitative results: frequent structural issues (fluency) are reliably detected, while lexically specific or contact-sensitive distinctions (terminology, foreign influence) remain more variable under current data coverage.

Quality is the clearest case. Moving to a binary acceptability decision improved rater agreement, and the baseline model achieved high F1 on the same task. This alignment is expected: the features (TF–IDF and simple text-shape cues) capture adequacy and fluency in exactly the way the rubric defines them.

Severity shows a complementary profile. Annotation agreement dipped on ambiguous, high-stakes content such as dosing or discharge instructions, reflecting real borderline cases. Yet the model's probability scores effectively ranked these cases: limited-review recall showed that inspecting only a small slice of the highest-probability items recovered a large share of true high-risk instances. This is well suited to clinical triage, where the goal is to prioritize review rather than to enforce a single threshold.

Foreign influence remains the most challenging dimension. Performance is modest (F1 = 0.356) and highly variable across folds (±0.216), reflecting both label scarcity (11.8%) and conceptual subtlety. This instability points to a data limitation rather than a flaw in the schema: expanding coverage with targeted examples will be necessary to stabilize this signal. Because French is HC's lexifier, many French-origin elements are part of the language's core system and appear even in basilectal varieties. Yet contemporary contact with French continues to shape new mesolectal forms that are not shared by all speakers.

Error tags show a similar division. Terminology agreement improved once definitions were tightened and dictionary attestations used, while fluency emerged as both common and learnable. Terminology and orthographic errors remain rare but matter precisely because they can undermine safety and credibility.

Finally, *variants* demonstrate the schema's core sociolinguistic insight: multiple renderings may be "correct" but not equally appropriate for monolingual versus bilingual audiences. Both annotators and the model handled this field well, and it provides a natural bridge to preference-based modeling.

6.2 From Pilot to Practice: Applications and Next Steps

The pilot turns the schema into a usable signal. Even with a compact model, the labels support screening of HC medical translations before human review. In practice, predicted *quality* flags likely problematic sentences, while *severity* probabilities rank them so that the riskiest material appears first. The limited-review analysis shows that this ordering makes scarce human attention more effective—a good match for clinical and public-health settings.

These same scores also help grow the dataset where it matters. Because error typing is meaningful once *quality*=1, the model can pre-label likely problems and defer to annotators where its confidence is low. This is classic active learning: propose, adjudicate, and retrain. *Foreign influence* is the clearest target. Our pilot quantified its scarcity, and now we can seek out additional examples from both Haiti and diaspora contexts, anchored in lexicographic attestations.

For evaluation and system steering, the schema becomes an actionable framework:

1. Require *quality*=0 for release,
2. Use *severity* to gate additional review,
3. Treat *variants* as a preference space for reranking toward monolingual or bilingual audiences,
4. Map error tags directly to edits—terminology errors suggest glossary fixes, fluency errors prompt rewrites, and typos trigger orthographic normalization.

As the dataset scales, these labels are reward-ready: they can drive pairwise preference modeling or scalar reward functions that bias MT toward plain, audience-appropriate HC while down-weighting foreign-influence intrusions.

A key extension will be the integration of curated glossaries and span-level mappings, already preserved in the normalized CSV. Grounding model suggestions in attested HC terminology rather than unconstrained paraphrase will make outputs more reliable and interpretable. In the longer term, the

same infrastructure can support dynamic adaptation—systems that learn to select variants aligned with a patient's idiolect or speech profile, bridging linguistic diversity and clinical safety in real time.

In short, the experiment shows that the schema's fields can serve as practical levers for screening, dataset growth, evaluation, and preference-driven MT for HC medical communication.

7 Conclusion and Future Work

We introduced an annotation framework for Haitian Creole medical translation that foregrounds audience-appropriate language varieties, overall acceptability, and clinical risk. Designed for compact expert teams, the schema emphasizes high inter-annotator agreement through binary decisions and conditional tags while remaining expressive enough to capture the sociolinguistic and clinical dimensions that matter in practice. Across three rounds of annotation and adjudication, we produced a gold set with fully resolved labels and a normalization pipeline that integrates directly into evaluation and modeling.

Using the adjudicated pool of 152 EN→HC pairs, a transparent classifier–labeller baseline showed that acceptability detection is already reliable, severity scores function as a practical triage signal with interpretable recall–effort trade-offs, and error typing is tractable once quality issues are flagged. By contrast, foreign-influence judgments and rare error types remain limited chiefly by data coverage rather than by schema design, underscoring where expansion is most needed.

These results carry methodological and applied implications. The schema separates acceptability from clinical risk and encodes audience design explicitly, avoiding the conflation of sociolinguistic variation with generic "fluency." The pilot further demonstrates that calibrated probabilities and threshold-free metrics can support workflows that require prioritization and ranking rather than binary calls. Because error tags align with MQM, the framework bridges annotation, evaluation, and downstream modeling in a reproducible way.

Future work will expand and rebalance the dataset through targeted sampling across Haiti and diaspora communities, reinforce lexicographic grounding for contact-sensitive forms, and add document-level cues where linguistic variation and meaning unfold across sentences. The same fields are reward-ready: they can be used to train preference models and RLHF systems that steer MT outputs toward plain-language, culturally appropriate HC while down-weighting risky or foreign-influence renderings.

References

Hilal Al Shamsi, Abdullah G Almutairi, Sulaiman Al Mashrafi, and Talib Al Kalbani. 2020. Implications of language barriers for healthcare: a systematic review. *Oman medical journal*, 35(2):e122.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Derek Bickerton. 1973. The nature of a creole continuum. *Language*, pages 640–669.

David Campbell. 2012. Cultural competency in haitian-serving community health centers in south florida. *McGill Journal of Medicine*, 14(1).

Centers for Disease Control and Prevention. 2016. Everyday words for public health communication. Report / plain language guidance, Centers for Disease Control and Prevention, Office of the Associate Director for Communication. Accessed: May 12, 2026.

Michel DeGraff. 2005. Linguists' most dangerous myth: The fallacy of creole exceptionalism. *Language in society*, 34(4):533–591.

Yves Dejean. 2010. Creole and education in haiti. *The Haitian Creole language: History, structure, use, and education*, pages 199–216.

Rohit Dholakia and Anoop Sarkar. 2014. Pivot-based triangulation for low-resource languages. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 315–328.

Chandrika Divi, Richard G Koss, Stephen P Schmaltz, and Jerod M Loeb. 2007. Language proficiency and adverse events in us hospitals: a pilot study. *International journal for quality in health care*, 19(2):60–67.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*.

Dominique Fattier-Thomas. 1984. De la variété rèk à la variété swa: Pratiques vivantes de la langue en haïti. *Conjonction161*, 162:39–51.

Glenn Flores, Milagros Abreu, Cara Pizzo Barone, Richard Bachur, and Hua Lin. 2012. Errors of medical interpretation and their potential clinical consequences: a comparison of professional versus ad hoc versus no interpreters. *Annals of emergency medicine*, 60(5):545–553.

Glenn Flores, M Barton Laws, Sandra J Mayo, Barry Zuckerman, Milagros Abreu, Leonardo Medina, and Eric J Hardt. 2003. Errors in medical interpretation and their potential clinical consequences in pediatric encounters. *Pediatrics*, 111(1):6–14.

Bryant C Freeman. 1997. *Haitian-English English-Haitian Medical Dictionary*. [Lawrence, Kan.]: Institute of Haitian Studies, University of Kansas.

Sonia Halimi and Pierrette Bouillon. 2019. Google translate and babeldr in community medical settings: Challenges of translating into arabic. In *Arabic translation across discourses*, pages 27–44. Routledge.

Benjamin Hebblethwaite. 2012. French and underdevelopment, haitian creole and development: Educational language policy problems and solutions in haiti. *Journal of Pidgin and Creole languages*, 27(2):255–302.

Maude Heuretelou, Féquière Vilsaint, Erst Mirville, Michel-Ange Hyppolite, and John D. Nickrosz. 2000. *English / Haitian Creole Medical Dictionary*. Educa Vision.

Patricia Hudelson and François Chappuis. 2024. Using voice-to-voice machine translation to overcome language barriers in clinical communication: an exploratory study. *Journal of General Internal Medicine*, 39(7):1095–1102.

Kerry Hwang, Sue Williams, Emiliano Zucchi, Terence WH Chong, Monita Mascitti-Meuter, Dina LoGiudice, Anita MY Goh, Anita Panayiotou, and Frances Batchelor. 2022. Testing the use of translation apps to overcome everyday healthcare communication in australian aged-care hospital wards—an exploratory study. *Nursing open*, 9(1):578–585.

Elizabeth A Jacobs, Donald S Shepard, Jose A Suaya, and Esta-Lee Stone. 2004. Overcoming language barriers in health care: costs and benefits of interpreter services. *American journal of public health*, 94(5):866–869.

Leah S Karliner, Elizabeth A Jacobs, Alice Hm Chen, and Sunita Mutha. 2007. Do professional interpreters improve clinical care for patients with limited english proficiency? a systematic review of the literature. *Health services research*, 42(2):727–754.

Erin Kobetz, Janelle Menard, Betsy Barton, Jennifer Cudris Maldonado, Joshua Diem, Pascale Denize Auguste, and Larry Pierre. 2010. Barriers to breast cancer screening among haitian immigrant women in little haiti, miami. *Journal of immigrant and minority health*, 12(4):520–526.

Erin Kobetz, Janelle Menard, Joshua Diem, Betsy Barton, Jenny Blanco, Larry Pierre, Pascale D Auguste, Marie Etienne, and Cheryl Brewster. 2009. Community-based participatory research in little haiti: challenges and lessons learned. *Progress in Community Health Partnerships: Research, Education, and Action*, 3(2):133–137.

Annika Kreienbrinck, Saskia Hanft-Robert, Alina Ioana Forray, Asithandile Nozewu, and Mike Mösko. 2025. Usability of technological tools to overcome language barriers in healthcare–a scoping review. *Archives of Public Health*, 83(1):52.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can neural machine translation be improved with user feedback? *arXiv preprint arXiv:1804.05958*.

William Lewis. 2010. Haitian creole: How to build and ship an mt engine from scratch in 4 days, 17 hours, & 30 minutes. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F Wong, Siyou Liu, and Longyue Wang. 2023. A paradigm shift: The future of machine translation lies with large language models. *arXiv preprint arXiv:2305.01181*.

Janelle Menard, Erin Kobetz, Joshua Diem, Martine Lifleur, Jenny Blanco, and Betsy Barton. 2010. The sociocultural context of gynecological health among haitian immigrant women in florida: applying ethnographic methods to public health inquiry. *Ethnicity & Health*, 15(3):253–267.

Ludovic Mompelat. 2025. Recommendations for overcoming linguistic barriers in healthcare: Challenges and innovations in nlp for haitian creole. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 20–31.

National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP). 2022. Ncc merp index for categorizing medication errors (2022 revision). Technical report, NCC MERP. Revised categorization of medication errors by severity.

National Institutes of Health. 2025. Nih clear communication initiative. https://www.nih.gov/institutes-nih/nih-office-director/office-communications-public-liaison/clear-communication. Accessed: May 12, 2026.

Eva Maria Noack, Jennifer Schulze, and Frank Müller. 2021. Designing an app to overcome language barriers in the delivery of emergency medical services: participatory development process. *JMIR mHealth and uHealth*, 9(4):e21586.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Anita Panayiotou, Anastasia Gardner, Sue Williams, Emiliano Zucchi, Monita Mascitti-Meuter, Anita MY Goh, Emily You, Terence WH Chong, Dina Logiudice, Xiaoping Lin, and 1 others. 2019. Language translation apps in health care settings: Expert opinion. *JMIR mHealth and uHealth*, 7(4):e11316.

Rafael Pérez-Escamilla, Jonathan Garcia, and David Song. 2010. Health care access among hispanic immigrants:¿ alguien está escuchando?[is anybody listening?]. *NAPA bulletin*, 34(1):47–67.

Ellen R Ryan, Wesley E Hawkins, Marilyn Parker, and Michele J Hawkins. 2004. Perceptions of access to us health care of haitian immigrants in south florida. *Florida Public Health Review*, 1:30–35.

Hervé Spechbach, Johanna Gerlach, Sanae Mazouri Karker, Nikos Tsourakis, Christophe Combescure, Pierrette Bouillon, and 1 others. 2019. A speech-enabled fixed-phrase translator for emergency settings: Crossover study. *JMIR medical informatics*, 7(2):e13167.

David Tezil. 2022. On the influence of kreyòl swa: Evidence from the nasalization of the haitian creole determiner/la/in non-nasal environments. *Journal of Pidgin and Creole Languages*, 37(2):291–320.

David Tézil. 2024. Sociolinguistic challenges and new perspectives on determining french speakers in creole communities: the case of haiti. *International Journal of the Sociology of Language*, 2024(288):177–207.

Translators without Borders. 2020. Twb glossary for covid-19 (multilingual plain-language glossary). https://translatorswithoutborders.org/covid-19/. Accessed: May 12, 2026.

U.S. Department of Health and Human Services. 2013. National standards for culturally and linguistically appropriate services in health and health care: A blueprint for advancing and sustaining clas policy and practice. Technical report, U.S. Department of Health and Human Services, Office of Minority Health, Washington, DC.

Albert Valdman. 2015. *Haitian Creole: structure, variation, status, origin*. Equinox Publishing Limited.

Albert Valdman, Marvin D Moody, and Thomas E Davies. 2017. *English-Haitian Creole Bilingual Dictionary*. iUniverse.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*.

World Health Organization. 2021. Use plain language: Communicating health information clearly. https://www.who.int/about/communications/understandable/plain-language. Accessed: May 12, 2026.