Detecting Experiential Intertextuality Across Migration Routes: Beyond Surface Similarity in French Narratives

Sakayo Toadoum Sari[1], **Nelly Robin**[2], **Michelle Auzanneau**[2], **Lakhdar Sais**[1],

Véronique Petit[2], **Marie Veniard**[3], **Said Jabbour**[1], **Fabien Delorme**[1]

[1]CRIL, CNRS – Université d'Artois, France, [2]CEPED, Université Paris Cité, France,
[3]EDA, Université Paris Cité, France

[1]{sakayo,sais,jabbour,delorme}@cril.fr, [2]nelly.robin@ird.fr,
[2]michelle.auzanneau@u-pariscité.fr, [2]veronique.petit@u-paris.fr,
[3]marie.veniard@u-paris.fr

Abstract

Migrants traversing geographically distinct routes such as the Trans-Saharan and Balkan corridors often recount strikingly parallel lived experiences: police violence, smuggler exploitation, dangerous crossings, and family separation. We introduce the task of *experiential intertextuality detection*: automatically identifying shared experiential echoes across migration narratives without requiring annotated training data. From 108 French migration narratives spanning both corridors, we automatically generate sentence pairs and score them using annotation-free methods: lexical baselines, sentence embeddings, POS-based structural features, a migration-specific theme lexicon, context-aware narrative features, and zero-shot LLM scoring with Qwen2.5-7B and Mistral-7B under three prompting strategies. We validate all methods against 816 expert-annotated intertextuality judgments (inter-annotator Krippendorff's α=0.27). Our results reveal that all surface, structural, and embedding methods correlate only weakly with expert judgments ($r \leq 0.30$); Qwen2.5-7B zero-shot achieves the best single-method correlation (r=0.38); few-shot examples degrade Qwen but dramatically improve Mistral; narrative position significantly predicts intertextuality, with departure-phase pairs showing the highest experiential echoes; and a supervised hybrid combining all 31 features achieves r=0.45, a 21% improvement over the best individual method.

1 Introduction

Migration is one of the defining phenomena of our era, and the narratives produced by migrants during their journeys are an invaluable yet underexploited source of knowledge for the Humanities and Social Sciences (HSS). These narratives collected through interviews at transit points along migration routes capture lived experiences in the migrants' own words: the dangers faced, the resources mobilized, and the complex decision-making that shapes each journey (Robin, 2014; Bacon, 2022). A recurring observation among HSS researchers is that migrants following completely different geographical routes often describe strikingly similar experiences (Robin, 2014; Bacon, 2022). A minor from Côte d'Ivoire crossing the Sahara and a Congolese refugee traversing the Balkans may both recount police violence at borders, exploitation by smugglers, perilous crossings, and the anguish of family separation. This phenomenon where distinct narratives echo shared experiential content despite originating from different geographical and cultural contexts is what we term *experiential intertextuality*, a concept we introduce in this paper to distinguish from traditional literary intertextuality. We recognize experiential parallelism at three levels: (i) *thematic*, when both sentences describe the same category of experience (smuggler exploitation, police violence, dangerous crossing); (ii) *functional*, when both occupy the same role in the narrative arc of the journey (departure motivation, transit danger, arrival); and (iii) *pragmatic*, when both carry the same speech act or stance (self-motivation to leave, testimony of suffering, expression of hope). Two sentences may share all three levels while sharing no vocabulary whatsoever.

Unlike traditional intertextuality in literary studies, which concerns textual references and allusions between works (Kristeva, 1969), experiential intertextuality captures parallels in *lived experience* as articulated through narrative. Detecting such parallels automatically is valuable for HSS researchers seeking to identify universal patterns in migration, understand which experiences

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 139–150
August 2–5, 2026.

transcend specific routes, and ultimately support policy-making with evidence-based insights. Prior work on these narratives (Ing et al., 2025) focused on extracting domain terms and recognizing locations answering *what* is mentioned. We address the complementary and harder question: *do two sentences from different routes describe the same kind of experience?* This requires moving beyond entity extraction to experiential comparison, and from supervised evaluation against term lists to continuous correlation against expert judgments. We formalize this as a scoring task. Given a sentence pair (s_i, s_j) drawn from narratives on different routes, we seek a function $f(s_i, s_j) \rightarrow [0, 1]$ that approximates expert-assessed experiential intertextuality, without requiring annotated training data. Figure 1 illustrates our approach.

Our contributions are: we formalize experiential intertextuality detection as a new NLP task with an annotation-free pipeline; we validate against 816 expert judgments with formal IAA; we introduce context-aware narrative features revealing journey-phase effects. Our code is publicly available.[1]

2 Related Work

Text mining has been increasingly applied to migration-related texts, though primarily on public discourse. Öztürk and Ayvaz (2018) use sentiment analysis on Twitter data to investigate public opinion toward the Syrian refugee crisis, while Hussain et al. (2018) study shifts in blogosphere narratives during the European migrant crisis using named-entity extraction and targeted sentiment analysis. Both analyze *public reactions* to migration, whereas we analyze *migrants' own narratives*. Most closely related, Ing et al. (2025) present a text mining framework for migration narrative corpus focusing on domain term extraction via a modified set expansion algorithm (MultiWidthExpan) and location recognition/disambiguation using NER and BELA (Plekhanov et al., 2023). Their evaluation uses P/R/F1 against expert term lists with no embedding or LLM baselines. Our work addresses a fundamentally different question on the same corpus: rather than extracting what entities are mentioned, we detect whether two sentences describe the same *kind of experience*, evaluated via continuous correlation with expert intertextuality scores across 16 methods.

Semantic Textual Similarity (STS) is a well-established NLP benchmark (Cer et al., 2017; Agirre et al., 2016). Modern approaches leverage sentence embeddings from pretrained transformers (Reimers and Gurevych, 2019; Conneau et al., 2020), achieving strong performance on English STS benchmarks. However, experiential intertextuality is fundamentally different from semantic similarity: two sentences can be semantically dissimilar yet experientially parallel. For instance, *On a pris le pickup pour aller à Gao* ("We took the pickup to go to Gao") and *On a pris le bus pour aller à Belgrade* ("We took the bus to go to Belgrade") share the *experience* of collective transit to a city, yet a standard STS model would score them low due to different locations and vehicle types. Conversely, sentences sharing keywords like "police" may receive high STS scores while describing entirely different experiences (routine checkpoint vs. violent refoulement).

Discourse relation frameworks such as RST (Mann and Thompson, 1988) and PDTB (Prasad et al., 2008) identify structural and semantic relations between text segments. Our task differs in that we compare segments *across* documents rather than within a single document, and our relations are experiential rather than rhetorical. Computational narrative analysis has explored story similarity through emotional arcs (Reagan et al., 2016), narrative event chains (Chambers and Jurafsky, 2008), and commonsense story understanding (Mostafazadeh et al., 2016). Bamman et al. (2013) learn narrative schemas from text, while Caselli and Vossen (2017) propose event-centric approaches. Cross-document event coreference (Cybulska and Vossen, 2014) is close in spirit, comparing "the same kind of event" across documents, though it targets event identity rather than experiential parallelism. In digital humanities, quantitative intertextuality detection employs text reuse and sequence alignment for literary echoes (Forstall et al., 2015); our work extends this paradigm from textual allusion to experiential resonance. Our work adds a new dimension: identifying experiential parallels across narratives from different speakers in different geographical contexts, where the "events" are real lived experiences.

Recent work demonstrates that LLMs can perform nuanced discourse tasks, including stance

[1] https://github.com/Toadoum/IntertextMigra

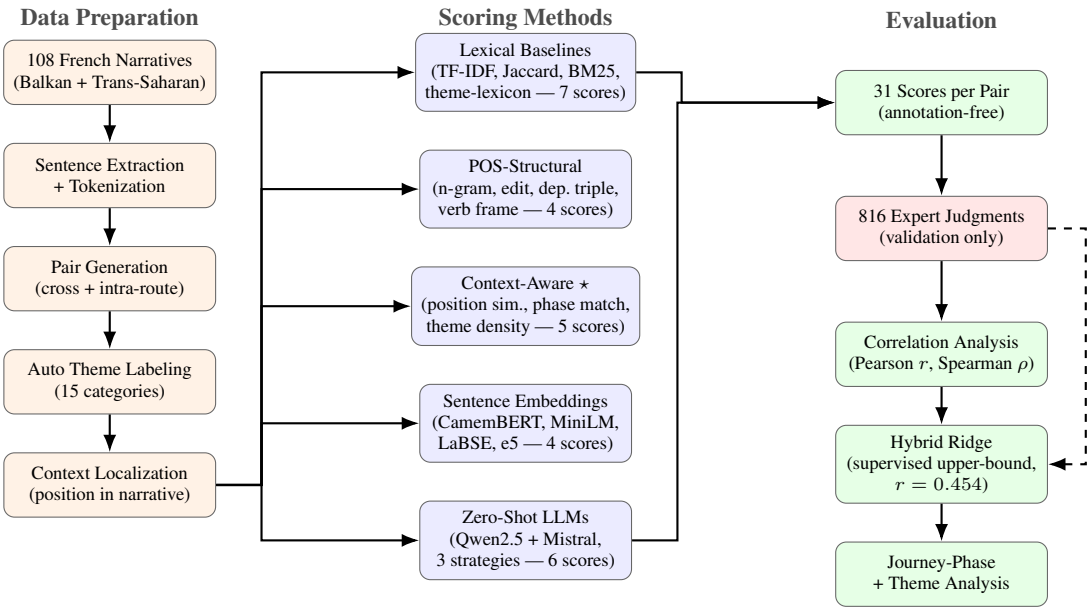

Figure 1: Pipeline overview. **Left**: sentence pairs are automatically generated from raw narratives with context localization. **Middle**: five annotation-free method families produce 31 scores per pair. **Right**: scores validated against expert judgments; supervised hybrid as upper-bound. ⋆ = novel.

detection, argumentation mining, and pragmatic interpretation (Gilardi et al., 2023). Qwen Team (2024) show strong multilingual capabilities for Qwen2.5, while the Mistral family (Jiang et al., 2023) demonstrates competitive performance with efficient architectures. We evaluate both model families in a novel annotation-free setting, revealing substantial differences in how they handle experiential comparison.

3 Data

Our corpus consists of 108 French-language migration narratives collected through semi-structured interviews at transit points along two corridors:

- **Trans-Saharan corridor** (99 narratives): Collected in Niger (Agadez, Arlit), Algeria (Adrar, Tamanrasset, Maghnia), Senegal (Dakar, Mbour, Ziguinchor), and Morocco (Oujda, Rabat) from sub-Saharan minors. Countries of origin: Côte d'Ivoire, Mali, DRC, Guinea, Senegal, Nigeria, Gambia, Burkina Faso.
- **Balkan corridor** (9 narratives): Collected along transit routes through Serbia, North Macedonia, and Bosnia from migrants originating from Congo-Brazzaville, Algeria, Guinea, Côte d'Ivoire, Somalia, Mali, and Senegal.

Table 1: Corpus statistics by corridor. Countries = countries of origin of migrants.

	Balkan	**Trans-Sah.**	**Total**
Narratives	9	99	108
Avg. words	1,672	787	876
Countries	8	8	13
Unique sentences	3,592	2,330	5,922

Table 1 reports corpus statistics. The narratives average 876 words, with Trans-Saharan narratives being shorter (avg. 787 words) due to the younger age of minors, while Balkan narratives are substantially longer (avg. 1,672 words) reflecting more complex multi-country journeys. The corpus was collected between 2015 and 2022 by trained researchers and local associations within the ANR HYCI project (Robin, 2014) and related fieldwork, and was first used for computational analysis by Ing et al. (2025). Interviews are conducted in French or local languages and transcribed into French. A relationship of trust is established through time and the guarantee of anonymity. Due to the sensitive nature of this data, we do not release it publicly (§7).

To illustrate the nature of the data, consider these two excerpts from different corridors describing parallel experiences of smuggler exploitation:

Trans-Saharan: "J'ai payé 75.000 FCFA pour

passer en Algérie. J'ai emprunté le pickup avec une dizaine de migrants. On était très serré dans le pickup." *(I paid 75,000 FCFA to cross into Algeria. I took the pickup with about ten migrants. We were very cramped in the pickup.)*

Balkan: "Il m'avait vendu un passeport congolais avec un visa de la Turquie à l'intérieur [...] quand je suis arrivé en Turquie, ils m'ont dit 'Monsieur ce n'est pas bon', et ils m'ont refoulé." *(He had sold me a Congolese passport with a Turkish visa inside [...] when I arrived in Turkey, they told me 'Sir, this is not valid', and they deported me.)*

Both describe exploitation, financial and documentary, yet share almost no vocabulary, precisely the challenge our task addresses. From the narratives, we automatically extract sentences and generate pairs using stratified sampling across three configurations: cross-route (one Balkan, one Trans-Saharan), intra-Balkan, and intra-Trans-Saharan. Each pair is automatically assigned thematic labels via keyword matching against a lexicon of 15 expert-defined categories, including *violence_police*, *passeur_exploitation*, *traversée_dangereuse*, *solidarité*, *famille_séparation*, *exploitation_travail*, *document_fraude*, *détention_camp*, *motivation_départ*, *mineur_seul*, *rêve_football*, *discrimination_racisme*, *mort_danger_vital*, *attente_stagnation*, and *autre*. No manual annotation is involved in pair generation.

A stratified sample of 816 pairs was annotated by two HSS migration experts on a continuous scale from 0.0 (no experiential link) to 1.0 (nearly identical experiences). Among these, 283 received independent dual annotations, yielding moderate agreement (Table 2). For dual-annotated pairs, the consensus score is the arithmetic mean; for the remaining 533 single-annotated pairs, the single annotator's score is used directly. No per-annotator normalization was applied, as both annotators' means are similar (0.389 vs. 0.405). This agreement level ($\alpha = 0.273$) is comparable to other subjective discourse annotation tasks (Prasad et al., 2008) and reflects genuine disagreement about what constitutes "shared experience."

Table 2: Inter-annotator agreement (AnnotatorA × AnnotatorB). The noise ceiling estimates the maximum attainable r given annotator disagreement.

Metric	Value
Dual-annotated pairs	283
Krippendorff's α (interval)	0.273
Cohen's κ (weighted, 3-bin)	0.259
Pearson r	0.281***
Mean |diff|	0.306
Spearman-Brown reliability ($\hat{\rho}$)	0.438
Noise ceiling ($r_{\max} = \sqrt{\hat{\rho}}$)	**0.662**

To contextualize our results, we estimate the maximum correlation any method can achieve given annotator disagreement. Using the Spearman-Brown prophecy formula, the reliability of the averaged score from two annotators is $\hat{\rho} = 2r_{12}/(1 + r_{12}) = 0.438$, where $r_{12} = 0.281$ is the inter-annotator Pearson r. The noise ceiling, the theoretical maximum r between a perfect method and the consensus score, is $r_{\max} = \sqrt{\hat{\rho}} = 0.662$. Our best single method (Qwen2.5 zero-shot, $r = 0.375$) achieves 56.6% of this ceiling, and the supervised hybrid ($r = 0.454$) achieves 68.6%, indicating substantial room for improvement but also that a significant portion of the remaining gap is attributable to irreducible annotator noise.

4 Methods

Let $\mathcal{N} = \{n_1, \ldots, n_K\}$ denote a corpus of K narratives, each consisting of sentences $n_k = (s_1^k, \ldots, s_{m_k}^k)$. Given a pair (s_i, s_j) from different narratives, we define scoring functions $f : \mathcal{S} \times \mathcal{S} \to [0, 1]$ that estimate experiential intertextuality without supervision.

4.1 Lexical Baselines

We compute seven lexical scores. Let $W(s)$ denote the word set of sentence s. Beyond standard measures (Jaccard similarity, ROUGE-1 F-score, BM25-approximate, character 3-gram overlap, and overlap coefficient) we employ TF-IDF cosine similarity fitted on the full 108-narrative corpus rather than sentence pairs alone, for better inverse document frequency estimation. We additionally define a domain-specific **theme lexicon score**. We distinguish between the 15 thematic *labels* used for pair categorization and a subset of $L = 8$ thematic *word sets* used for computing the lexicon feature; the 8 sets correspond to experiential categories with sufficient distinctive

vocabulary (*violence*, *passeur*, *transport*, *famille*, *danger*, *travail*, *document*, *hébergement*), totaling 130 terms curated from the narratives. Categories like *autre* and *rêve_football* were excluded from the lexicon as they lack stable keyword indicators. The score is:

$$f_{\text{thm}}(s_i, s_j) = \frac{1}{L} \sum_{l=1}^{L} \mathbb{1}[W_i \cap T_l \neq \emptyset] \cdot \mathbb{1}[W_j \cap T_l \neq \emptyset] \tag{1}$$

where $W_i = W(s_i)$, counting the fraction of thematic categories activated in *both* sentences. This captures topical co-occurrence at the experiential category level rather than the word level.

4.2 POS-Structural Features

We hypothesize that migrants describing parallel experiences may use similar grammatical structures even with entirely different vocabulary. Let $P(s) = (p_1, \ldots, p_n)$ denote the POS tag sequence of s (punctuation removed), obtained via spaCy's French model (`fr_core_news_sm`). We compute:

POS n-gram Jaccard. Let $G(s)$ be the multiset of POS n-grams ($n \in [2, 5]$):

$$f_{\text{png}}(s_i, s_j) = \frac{\sum_g \min(G_i[g],\, G_j[g])}{\sum_g \max(G_i[g],\, G_j[g])} \tag{2}$$

where $G_i = G(s_i)$ and $G_j = G(s_j)$.

POS edit similarity. Let $P_i = P(s_i)$:

$$f_{\text{ped}}(s_i, s_j) = 1 - \frac{\text{Lev}(P_i,\, P_j)}{\max(|P_i|,\, |P_j|)} \tag{3}$$

Dependency triple overlap. Let $D(s) = \{(r, p_h, p_c)\}$ be the set of dependency triples (relation, head POS, child POS). We compute Jaccard over $D(s_i)$ and $D(s_j)$.

Verb-frame overlap. Using the dependency parse, for each verb in a sentence we extract the set of dependency labels of its children (ignoring the verb's lemma). This captures action argument structures cross-vocabulary: "prendre [obj, obl]" matches whether the object is "pickup" or "bus".

4.3 Context-Aware Narrative Features

Since each sentence s_i originates from a specific narrative n_k at a known position, we exploit source context, a signal unavailable to sentence-level methods.

Position similarity. Let $\pi(s_i) \in [0, 1]$ be the normalized character offset of s_i within its source narrative (0 = beginning, 1 = end):

$$f_{\text{pos}}(s_i, s_j) = 1 - |\pi(s_i) - \pi(s_j)| \tag{4}$$

The intuition is that migration narratives follow a natural temporal arc (departure, transit, arrival), and sentences at similar positions describe experiences from similar journey phases.

Journey-phase match. Let $\phi : [0, 1] \to \{1, 2, 3, 4\}$ map positions to four phases (quartiles): departure (1), early transit (2), late transit (3), arrival (4):

$$f_{\text{ph}}(s_i, s_j) = \begin{cases} 1.0 & \text{if } \phi_i = \phi_j \\ 0.5 & \text{if } |\phi_i - \phi_j| = 1 \\ 0.0 & \text{otherwise} \end{cases} \tag{5}$$

where $\phi_i = \phi(\pi(s_i))$.

Context theme density. For each sentence, we extract a 500-character window from the source narrative centered on the sentence's position, and compute the fraction of theme-lexicon words. The similarity of densities between two sentences captures whether both originate from thematically rich narrative passages.

4.4 Sentence Embeddings

We evaluate four multilingual sentence embedding models, listed with HuggingFace identifiers for reproducibility: `sentence-camembert-large` (CamemBERT-STS) (Martin et al., 2020); `paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2019); `LaBSE`; and `multilingual-e5-large` (Conneau et al., 2020). We use multilingual models to enable future extension to English narratives from the same project. For each pair, we compute cosine similarity between L2-normalized embeddings.

4.5 LLM Scoring

We evaluate two open-source 7B-parameter LLMs, `Qwen2.5-7B-Instruct` (Qwen Team, 2024) and `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023), both NF4-quantized via `bitsandbytes`, with temperature = 0.1 and `max_new_tokens` = 256, under three prompting strategies:

Zero-shot. The system prompt defines experiential intertextuality in French and provides the 0–1

rating scale. The model receives only the two sentences and returns a JSON score. No examples are provided.

Few-shot. Three expert-annotated pairs are prepended as demonstrations, selected to cover low (∼0.1), medium (∼0.4), and high (∼0.8) intertextuality scores. This tests whether calibration examples help the model understand the scale.

Chain-of-thought (CoT). The prompt requests structured reasoning: (1) identify the theme of each sentence, (2) compare whether the experiences are parallel, (3) produce a score. This tests whether explicit reasoning improves scoring quality.

An example of the zero-shot prompt structure:

> `System`: *Tu es un expert en études migratoires (You are an expert in migration studies). Évalue le degré d'intertextualité expérientielle entre deux phrases (Assess the degree of experiential intertextuality between two sentences) [...] Échelle: 0.0 à 1.0 (Scale: 0.0 to 1.0). Réponds avec un JSON (Respond with a JSON object): {"score": <float>}*
>
> `User`: *Phrase 1 (Sentence 1): "[sentence 1]" Phrase 2 (Sentence 2): "[sentence 2]"*

4.6 Hybrid Model (Supervised Upper-Bound)

We combine all $d = 31$ features into $\mathbf{x}_{ij} \in \mathbb{R}^d$ for each pair and train Ridge regression with narrative-level grouped 5-fold CV (GroupKFold; pairs are grouped by narrative-pair ID so that no narrative appears in both train and test folds, preventing data leakage from shared sentences; see Appendix B for the annotation sampling protocol):

$$\hat{y}_{ij} = \mathbf{w}^\top \mathbf{x}_{ij} + b \qquad (6)$$
$$\mathcal{L} = \textstyle\sum_{(i,j)} (y_{ij} - \hat{y}_{ij})^2 + \lambda \|\mathbf{w}\|^2$$

where y_{ij} is the expert score. The 31 features comprise: 7 lexical, 4 POS, 5 context-aware, 4 embedding, 6 LLM scores, plus `pos_combined` (weighted POS combination), raw narrative positions ($\pi(s_i)$, $\pi(s_j)$), and two metadata indicators (`same_country`, `cross_route`). All 16 annotation-free methods (lexical, POS, context, embedding, LLM) are computed without access to expert labels; TF-IDF weights are fitted on the full 108-narrative corpus (analogous to using a pretrained model), not on CV folds. Unlike other methods, the hybrid *uses expert labels* and serves as an upper-bound.

5 Experiments and Results

All annotation-free methods are validated by Pearson r and Spearman ρ against the expert sample ($n = 816$). No method uses expert scores for training except the supervised hybrid. Experiments were conducted on a GPU machine equipped with single NVIDIA Quadro RTX 8000 (48 GB VRAM) using a virtual environment with PyTorch, Transformers, and sentence-transformers. Table 3 presents the complete results ranked by Pearson r.

Table 3: Methods vs. expert validation (n=816). 95% CIs for Pearson r via Fisher z-transform. ⋆ novel contribution of this paper. † uses expert labels (narrative-grouped CV). ** p<0.01, *** p<0.001. Note: 815/816 expert pairs are cross-route.

Method	r	95% CI	ρ
LLM (zero-shot / few-shot / CoT)			
Qwen2.5 zero-shot	**.375*****	[.314, .432]	.304***
Qwen2.5 CoT	.365***	[.304, .423]	**.331*****
Mistral few-shot	.327***	[.264, .387]	.272***
Qwen2.5 few-shot	.289***	[.224, .351]	.239***
Mistral CoT	.243***	[.177, .307]	.219***
Mistral zero-shot	.134***	[.066, .201]	.124***
Sentence Embeddings			
multi-MiniLM	.295***	[.231, .356]	.254***
CamemBERT-STS	.284***	[.220, .346]	.236***
LaBSE	.249***	[.184, .312]	.204***
e5-large	.219***	[.153, .283]	.193***
Lexical			
Theme-lexicon	.287***	[.223, .349]	.270***
TF-IDF (original)	.254***	[.189, .317]	.286***
Char-3gram	.230***	[.164, .294]	.184***
Context-aware ⋆			
Position sim.	.131***	[.063, .198]	.119***
Phase match	.107**	[.039, .174]	.102**
Theme density (avg)	.089*	[.020, .157]	.074*
Theme density (sim)	.049	[−.020, .118]	.035
Window overlap	.011	[−.058, .080]	.015
POS-structural			
POS n-gram	.119***	[.051, .186]	.076
Dep. triple	.091**	[.023, .159]	.061
Supervised upper-bound†			
Hybrid Ridge (31)	**.454*****	[.398, .507]	**.383*****
Noise ceiling			
$r_{\max}$	.662	—	—

No annotation-free method exceeds $r = 0.38$, but contextualized against the noise ceiling of $r_{\max} = 0.662$ (§3), Qwen2.5 achieves 56.6% of the theoretical maximum. The **theme lexicon** (Eq. 1, $r = 0.287$) outperforms three of four neural embedding models; Qwen2.5 significantly outperforms it (Williams test: $t = 3.03$, $p = 0.003$). **Qwen2.5-7B** strongly outperforms Mistral-7B in zero-shot (0.375 vs. 0.134), a gap larger than any prompting strategy effect. The **hybrid** ($r = 0.454$, 68.6% of ceiling) significantly outperforms the

best single method (Williams test: $t = 4.61$, $p < 0.0001$), demonstrating strong complementarity. **POS features are weak** ($r \leq 0.119$), discussed in §6. Note: 815 of 816 expert pairs are cross-route, so results directly measure inter-corridor experiential echoes. Figure 2 reveals a striking asymmetry between the two LLMs:

- **Qwen2.5-7B**: Zero-shot is strongest ($r = 0.375$); few-shot *degrades* performance to $r = 0.289$ (-23%); CoT maintains strength ($r = 0.365$) and yields the best ranking quality ($\rho = 0.331$).

- **Mistral-7B**: Zero-shot is weakest ($r = 0.134$); few-shot *dramatically improves* to $r = 0.327$; CoT is intermediate ($r = 0.243$).

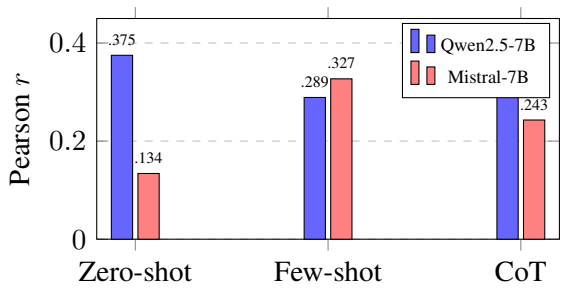

Figure 2: Prompting strategy comparison. Few-shot degrades Qwen but dramatically improves Mistral.

This asymmetry suggests that Qwen has stronger zero-shot French comprehension and migration-relevant world knowledge, while Mistral requires calibration to understand the task. The few-shot degradation for Qwen may reflect *score anchoring*: the three demonstration examples bias the model toward their score distribution, overriding its own superior zero-shot judgment. This finding has practical implications for annotation-free deployment: the choice of prompting strategy must be model-specific.

Figure 3 shows expert intertextuality stratified by journey phase. Departure × departure ($\mu = 0.280$) is highest, while mismatched phases (e.g., arrival × departure, $\mu = 0.129$) are lowest. The position similarity feature (Eq. 4) captures this effect ($r = 0.131$, $p < 0.001$).

	Dep.	E-Tr.	L-Tr.	Arr.
Dep.	.280	.227	.218	.194
E-Tr.	.129	.263	.224	.189
L-Tr.	.128	.169	.175	.243
Arr.	.129	.143	.275	.240

Figure 3: Mean expert intertextuality by journey-phase pair (row = sentence 1, column = sentence 2). Darker red = higher. The diagonal and near-diagonal show strongest echoes.

The diagonal pattern confirms that matching journey phases yield stronger intertextuality, with departure being the most universal. This aligns with HSS research observing that the motivations for leaving (family pressure, economic hardship, conflict) are shared across African migration contexts (Robin, 2014), while transit and arrival experiences are shaped by route-specific factors (desert vs. sea crossings, different border policies). Notably, the off-diagonal pair arrival × late_transit ($\mu = 0.275$) also scores high, suggesting that experiences near the end of the journey converge regardless of exact phase boundaries: migrants describe similar exhaustion, hope, and encounters with authorities. Table 4 reports mean expert scores by thematic category. Life-threatening danger ($\mu = 0.340$) and labor exploitation ($\mu = 0.326$) show the strongest cross-route intertextuality, suggesting that these experiences are systemic to irregular migration regardless of corridor. Document fraud ($\mu = 0.286$) and the football dream ($\mu = 0.254$) also show strong echoes: forged papers and aspirations of an athletic career are remarkably consistent themes. In contrast, unaccompanied minor experiences ($\mu = 0.117$) and waiting/stagnation ($\mu = 0.125$) are the weakest, indicating route-specific variation. Waiting experiences depend heavily on local transit infrastructure (desert oases vs. Balkan refugee camps), while unaccompanied minor narratives reflect different legal frameworks across countries.

Table 4: Mean expert intertextuality by theme ($n \geq 30$).

Theme	n	**Expert** μ
mort_danger_vital	30	0.340
exploitation_travail	69	0.326
document_fraude	57	0.286
rêve_football	49	0.254
traversée_dangereuse	79	0.242
passeur_exploitation	93	0.196
violence_police	82	0.186
détention_camp	42	0.182
solidarité	66	0.177
attente_stagnation	123	0.125
mineur_seul	46	0.117

Table 5 decomposes the supervised hybrid (Eq. 6) into feature groups.

Table 5: Ablation (narrative-grouped 5-fold CV Ridge). LLM features contribute most ($\Delta r = -0.059$). Removing embeddings slightly *improves* the model.

Configuration	r	ρ	Δr
Full (31 feat.)	.454	.383	—
Each group alone:			
LLM (6)	.416	.335	—
Lexical (7)	.338	.268	—
Embedding (4)	.295	.249	—
POS (4)	.100	.059	—
Leave-one-group-out:			
w/o LLM	.395	—	−.059
w/o Lexical	.430	—	−.024
w/o POS	.453	—	−.001
w/o Embedding	.463	—	+.010

Three findings emerge. First, **LLM features are the most valuable group**: removing them causes the largest drop ($\Delta r = -0.059$), and LLM features alone achieve $r = 0.416$, already higher than any non-LLM method. Second, **embeddings are redundant**: removing all four embedding models actually *improves* the hybrid from $r = 0.454$ to 0.463, indicating that the semantic similarity signal captured by embeddings is entirely subsumed by the LLM scores. Third, **POS features are negligible**: removal changes r by only 0.001, confirming that syntactic structure provides almost no independent signal for experiential intertextuality. The feature importance analysis (Ridge coefficients) shows that the top-5 most influential features are: ROUGE-1 ($\beta = -0.111$, a suppressor, since controlling for unigram overlap lets other features capture genuine experiential similarity), *jaccard* ($\beta = +0.077$), *theme_lexicon* ($\beta = +0.045$), *Qwen2.5-CoT* ($\beta = +0.041$), and *ctx_position_sim* ($\beta = +0.036$).

To verify that LLMs capture signal beyond surface overlap, we compute partial correlations controlling for Jaccard similarity. The Qwen2.5-7B zero-shot correlation with expert scores drops only marginally when partialing out Jaccard ($r = 0.375 \rightarrow r_{\text{partial}} = 0.351$), confirming substantial beyond-surface signal. Similarly, partialing out the theme lexicon yields $r_{\text{partial}} = 0.336$, indicating that Qwen captures experiential parallels not reducible to thematic keyword overlap. For the theme lexicon itself, controlling for Jaccard yields $r_{\text{partial}} = 0.253$ (vs. raw $r = 0.287$), showing that approximately 12% of its signal is explained by simple word overlap, with the remainder reflecting genuine thematic co-activation. We examine the pairs where all methods disagree most with experts to understand the limits of current approaches.

High expert, low predicted. Experts rate these pairs as experientially parallel, but no method detects it. A representative example: "Mon père dit que je suis courageux et que je peux réussir en Europe" (*My father says I am brave and can succeed in Europe*; Trans-Saharan) and "Je me suis dit que je ne peux pas rester comme ça" (*I told myself I cannot stay like this*; Balkan). Both express the moment of deciding to migrate, sharing no vocabulary, no syntactic structure, and no thematic keywords. This is the hardest case: experiential intertextuality encoded purely in *pragmatic intent*, the speech act of self-motivation before departure.

Low expert, high predicted. Methods score these pairs highly, but experts disagree. For example, sentences sharing "police" or "frontière" in different experiential contexts, such as a routine identity check versus a violent *refoulement*. The same vocabulary describes fundamentally different experiences. This confirms that lexical overlap, and even embedding similarity, can mislead when identical words carry different experiential weight.

Annotator disagreement cases. Among the 283 doubly-annotated pairs, those with the largest AnnotatorA–AnnotatorB disagreement ($|\text{diff}| > 0.6$) tend to involve implicit experiential links, such as one sentence describing a cause ("J'ai payé le passeur") and another describing a consequence ("On était 20 dans le pickup"), where recognizing the link requires domain knowledge about smuggling logistics.

6 Discussion

Experiential intertextuality $\neq$ similarity. No standard similarity method achieves strong correlation with expert judgments. The best single method (Qwen2.5-7B, $r = 0.375$) explains only 14% of variance ($r^2 = 0.141$). This is a property of the phenomenon: experiential parallels are expressed through different vocabulary, syntax, and discourse structure.

Domain knowledge $>$ neural embeddings. The theme lexicon ($r = 0.287$) outperforms CamemBERT-STS ($r = 0.284$), LaBSE ($r = 0.249$), and e5-large ($r = 0.219$). For domain-specific experiential analysis, 130 curated terms can rival neural models with millions of parameters.

LLMs: model identity matters more than prompting. The Qwen–Mistral gap in zero-shot (0.375 vs. 0.134) is 2.8$\times$ larger than any within-model prompting effect. Pretraining data composition is more important than prompt engineering. The few-shot asymmetry shows that calibration examples serve different functions depending on model capability.

Journey phase as a structural predictor. The significant correlation of position similarity with expert scores ($r = 0.131$, $p < 0.001$) shows that migration narratives have an inherent temporal structure conditioning the universality of experiences (Robin, 2014).

Why POS features underperform. POS features yielded only weak correlations ($r \leq 0.119$) because POS patterns are too generic (e.g., `PRON VERB PREP NOUN` matches both relevant and irrelevant sentences) and varying French proficiency among migrants produces diverse syntax for identical experiences.

Toward event-structure representations. Our POS and dependency features capture only shallow syntax. Richer event-centric representations (semantic role labeling, frame-semantic parsing, or predicate-argument structures) could better capture the experiential content of sentences by abstracting over vocabulary while preserving "who did what to whom." For instance, an SRL-based feature could match the agent-action-patient structure of "Le passeur nous a abandonnés" and "The smuggler left us behind" despite lexical divergence. However, robust French SRL tools remain limited compared to English, and available frame-semantic resources (e.g., French FrameNet) offer incomplete coverage for migration-specific events such as *refoulement*, smuggling, or border crossing. We consider event-structure baselines an important direction for future work, particularly as multilingual SRL models improve.

Implications for HSS research. The universality of life-threatening danger ($\mu = 0.340$) and labor exploitation ($\mu = 0.326$) across routes suggests *systemic* patterns in migration risk. The route-specificity of waiting ($\mu = 0.125$) points to differences in transit infrastructure and border policy.

7 Conclusion

We introduced experiential intertextuality detection in migration narratives, a novel NLP task at the intersection of computational social science and discourse analysis. Our annotation-free pipeline, validated against expert judgments, shows that: (1) experiential parallels across migration routes are real but hard to detect, with no single method exceeding $r = 0.38$ (56.6% of the noise ceiling); (2) a hybrid combining 31 features reaches $r = 0.45$ (68.6% of ceiling); (3) departure experiences are the most universally shared; (4) LLM scores subsume embeddings; and (5) prompting effects are model-dependent. Future work will explore fine-tuned models, richer discourse features, and experiential graphs linking shared experiences across the full corpus.

Limitations

The corpus is small (108 narratives) with a strong route imbalance (99 Trans-Saharan vs. 9 Balkan), which may bias cross-route comparisons. The inter-annotator agreement ($\alpha = 0.273$), while consistent with task subjectivity, limits gold standard reliability; the noise ceiling analysis (§3) shows that 31.4% of the hybrid's gap to perfection is attributable to irreducible annotator noise. We evaluated only 7B-class LLMs at 4-bit quantization; larger models (70B+) or full-precision inference may yield different conclusions, and the effect of quantization on score distributions was not isolated. We did not fine-tune models because the annotation-free framing precludes training on expert labels by design; fine-tuned approaches are an important future direction. Our

lexical baselines do not apply French lemmatization or morphological normalization, which may understate their potential given French's rich inflectional morphology. The 533 single-annotated pairs may carry annotator-specific bias; however, the correlation of automated methods on dual-annotated vs. single-annotated subsets shows consistent patterns, and both annotators' score distributions are similar (means: 0.389 vs. 0.405). The POS features are limited to shallow patterns; richer event representations (semantic role labels, predicate-argument structures, or frame-semantic features) could better capture experiential content and deserve exploration in future work. The context-aware features exploit only positional information; discourse-level features (coreference, causal chains) remain unexplored. For the hybrid model, we mitigate data leakage via narrative-level GroupKFold (§4); TF-IDF weights are fitted on the full 108-narrative corpus (not on CV folds), analogous to using a pretrained model, and all embedding/LLM scores are computed without access to expert labels. Finally, the theme lexicon is used both as a feature and for pair labeling; while the two serve different purposes (8 word-set scoring vs. 15-category labeling) and labeling was not used to select annotation pairs, we acknowledge this dual role. Empirically, the number of theme labels per pair is *negatively* correlated with expert scores ($r = -0.148$, $p < 0.001$), ruling out the concern that thematic stratification biases toward high-scoring pairs. The partial correlation analysis (§5) further confirms that LLM and lexicon signals are not reducible to theme-based distributional effects.

Ethical Considerations

The narratives were collected from minors and young adults at transit points during their migratory journeys, a context of extreme vulnerability. Interviews were conducted by trained researchers from transit countries and local associations, with a relationship of trust established through time and the use of the language spoken by the minors. All life stories are fully anonymized: names are replaced with codes, and identifying details removed. **The dataset will not be publicly released** due to the sensitive and personal nature of the content. Access may be granted to qualified researchers upon ethical review, in coordination with the ANR HYCI project partners. We release the evaluation pipeline code to support methodological reproducibility.

Acknowledgements

This work has benefited from the support of the Hauts-de-France region, ANR HYCI Project (ANR-22-CE55-0010) of the French National Research Agency, CRIL-Lab CNRS, and Artois University.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of SemEval*, pages 497–511. ACL.

Lucie Bacon. 2022. *La fabrique du parcours migratoire sur la route des Balkans: Co-construction des récits et écritures (carto)graphiques*. Ph.D. thesis, Université de Poitiers.

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of ACL*, pages 352–361. ACL.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86. ACL.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of SemEval*, pages 1–14. ACL.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL*, pages 789–797. ACL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451. ACL.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of LREC*, pages 4545–4552.

Christopher Forstall, Walter Scheirer, David Bamman, and Gregory Crane. 2015. Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching. *Digital Scholarship in the Humanities*, 30(4):503–515.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Muhammad Nihal Hussain, Kevin K Bandeli, Samer Al-khateeb, and Nitin Agarwal. 2018. Analyzing shift in narratives regarding migrants in Europe via blogosphere. In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 181–190. Springer.

David Ing, Fabien Delorme, Said Jabbour, Nelly Robin, and Lakhdar Sais. 2025. Text mining from migration narratives. In *Proceedings of ECML-PKDD*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Julia Kristeva. 1969. *Séméiotikè: Recherches pour une sémanalyse*. Seuil, Paris.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of ACL*, pages 7203–7219. ACL.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pages 839–849. ACL.

Nihan Öztürk and Serkan Ayvaz. 2018. Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics*, 35(1):136–147.

Mikhail Plekhanov, Nora Kassner, Kashyap Popat, Louis Martin, Simone Merello, Boris Kozlovskii, Fabio A Dreyer, and Nicola Cancedda. 2023. Multilingual end to end entity linking. In *Proceedings of ACL*, pages 3512–3527. ACL.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse TreeBank 2.0. In *Proceedings of LREC*, pages 2961–2968.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):1–12.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992. ACL.

Nelly Robin. 2014. Migrations, observatoire et droit: Complexité du système migratoire ouest-africain. In *HdR*. Université de Poitiers.

A Prompt Templates

Zero-shot system prompt (French):

> *Tu es un expert en études migratoires et analyse de discours. Évalue le degré d'intertextualité expérientielle entre deux phrases issues de récits de migration français provenant de routes migratoires différentes. L'intertextualité expérientielle = expériences vécues partagées, échos thématiques, situations parallèles. Échelle: 0.0 (aucun lien) à 1.0 (expériences quasi-identiques). Réponds UNIQUEMENT avec un JSON: {"score": <float>}*

CoT system prompt (French):

> *[. . .] Raisonne étape par étape: 1. Identifie le thème de chaque phrase. 2. Compare: s'agit-il d'expériences parallèles? 3. Évalue la force du lien. Réponds avec un JSON: {"theme1": ". . . ", "theme2": ". . . ", "shared_experience": ". . . ", "score": <float>}*

Few-shot examples (3 demonstrations): Selected from expert-annotated pairs to cover low ($\sim$0.1), medium ($\sim$0.4), and high ($\sim$0.8) scores, drawn from pairs *not* in the evaluation set. Each example shows the two sentences and the expert consensus score.

B Annotation Sampling Protocol

The 816 expert-annotated pairs were sampled from the full auto-generated pool with stratification by thematic category to ensure coverage across all 15 themes (annotation was conducted via a multi-user Streamlit application). In practice, 815 of 816 annotated pairs are cross-route (one Balkan, one Trans-Saharan), reflecting the primary research question of cross-corridor experiential echoes. The single intra-route pair entered through an edge case in the sampling procedure. The thematic distribution of annotated pairs is: *attente_stagnation* (123), *autre* (115), *passeur_exploitation* (93), *violence_police* (82), *traversée_dangereuse* (79), *exploitation_travail* (69), *solidarité* (66), *document_fraude* (57),

famille_séparation (52), *rêve_football* (49), *mineur_seul* (46), *discrimination_racisme* (43), *détention_camp* (42), *motivation_départ* (39), *mort_danger_vital* (30).