SIGDIAL 2026 · EMORY UNIVERSITY, ATLANTA

Detecting Experiential Intertextuality Across Migration Routes

Beyond Surface Similarity in French Narratives

Sakayo Toadoum Sari Nelly Robin Michelle Auzanneau Lakhdar Sais
Véronique Petit
Marie Veniard Said Jabbour Fabien Delorme

CRIL, CNRS & Univ. d'Artois · CEPED, Univ. Paris Cité · EDA, Univ. Paris Cité

github.com/Toadoum/IntertextMigra

1

The problem

Migrants on different routes tell the same story

TRANS-SAHARAN ROUTE

> *“J’ai payé 75.000 FCFA pour passer en Algérie. On était très serré dans le pickup.”*
>
> (I paid 75,000 FCFA to cross into Algeria. We were crammed in the pickup.)

BALKAN ROUTE

> *“Il m’avait vendu un passeport congolais [. . .] ils m’ont refoulé.”*
>
> (He had sold me a Congolese passport [. . .] they deported me.)

Both describe paying to be exploited. They share **almost no vocabulary**.

Semantic similarity answers the wrong question

SAME EXPERIENCE, DIFFERENT WORDS

> *“On a pris le **pickup** pour aller à **Gao**”*
> *“On a pris le **bus** pour aller à **Belgrade**”*

Similarity models score this **low**.

SAME WORDS, DIFFERENT EXPERIENCE

> *“La **police** a contrôlé mes papiers”*
> *“La **police** nous a frappés et refoulés”*

Similarity models score this **high**.

> We want **shared lived experience**, not word or embedding overlap.

The task

Score how strongly two sentences from **different routes** describe the same lived experience:

$$f(s_i, s_j) \longrightarrow [0,\ 1]$$

WE CALL IT

Experiential intertextuality

Kristeva links texts. This links lives.

HARD CONSTRAINT

Annotation-free

Expert labels only validate. Nothing trains on them.

What makes two sentences “parallel”?

1. Thematic

Same category of experience

smuggler exploitation, police violence

2. Functional

Same role in the journey arc

departure, transit danger, arrival

3. Pragmatic

Same speech act or stance

self-motivation, testimony, hope

A pair can match on all three while sharing **no words at all**.

2

Data and evaluation

108 French narratives, two corridors

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

Interviews at transit points, 2015–2022, within ANR HYCI.

Speakers are **minors and young adults** interviewed mid-journey. The corpus stays closed; we release the pipeline.

Two experts scored **816** pairs from 0.0 to 1.0; 815 are cross-route.

The experts barely agree with each other

Dual-annotated pairs	283
Krippendorff's α	**0.273**
Cohen's κ (weighted)	0.259
Annotator–annotator r	0.281

The obvious objection: if experts agree this weakly, how can any method be judged against them?

Disagreement about what counts as a shared experience is part of the phenomenon.

So we measure how much agreement is **even reachable**.

The noise ceiling

1. Reliability of the averaged score

$$\hat{\rho} = \frac{2r_{12}}{1 + r_{12}} = \frac{2(0.281)}{1.281} = 0.438$$

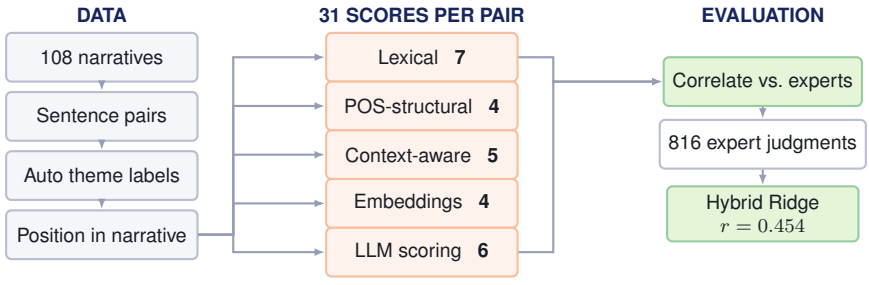
2. Attenuation bound

$$r_{\max} = \sqrt{\hat{\rho}} = \mathbf{0.662}$$

A perfect method tops out at $r = 0.662$ against this gold standard.

So $r = 0.375$ is not 37% of the way there. It is **56.6%** of what is obtainable.

Pipeline



Left of the expert box, nothing sees a label.

16 methods, five families

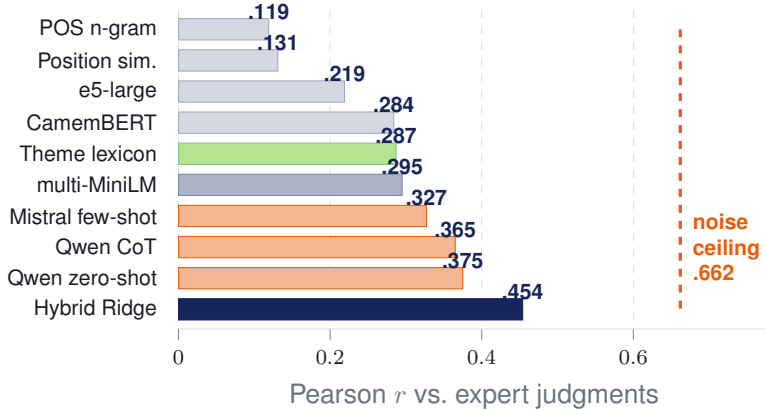
Family	What it measures	Scores
Lexical	Jaccard, TF-IDF, ROUGE-1, BM25, theme lexicon	7
POS-structural	POS n-grams, dependency triples, verb frames	4
Context-aware ⋆	Narrative position, journey-phase match	5
Embeddings	CamemBERT, MiniLM, LaBSE, e5-large	4
LLM scoring	Qwen2.5-7B, Mistral-7B $\times$ 3 prompts	6
Hybrid	Ridge over all 31, narrative-grouped CV	bound

⋆ New: where a sentence sits in its narrative, invisible to sentence-level methods.

3

Results

Nothing gets close to the ceiling



Qwen zero-shot: **56.6%** of the ceiling. Hybrid: **68.6%**.

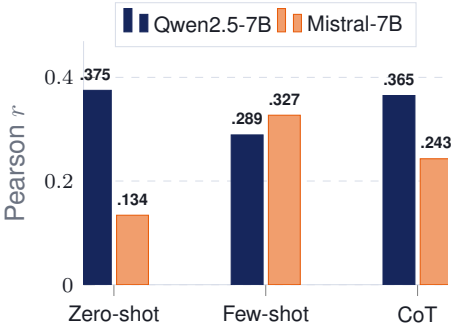
130 hand-picked words beat three neural models

Method	r
Qwen2.5-7B zero-shot	**.375**
multi-MiniLM	.295
Theme lexicon	**.287**
CamemBERT-STS	.284
LaBSE	.249
e5-large	.219

A 130-term expert word list rivals models with hundreds of millions of parameters.

Qwen still beats it: Williams $t = 3.03$, $p = .003$.

Few-shot helps one model and hurts the other



Qwen drops **23%** with examples.

Mistral gains **144%**.

Demonstrations anchor Qwen to their scores; Mistral needs them to grasp the scale.

> Pick the prompt per model. There is no default.

Departure is the most universal moment

	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

Mean expert score by journey-phase pair

Departure $\times$ departure: **.280**, the top cell.

Reasons to leave repeat everywhere: family pressure, hardship, conflict.

Transit and arrival split by route.

> Position alone predicts experts: $r = .131$, $p < .001$

The hybrid reaches .454, and embeddings are dead weight

Configuration	r	Δr
Full (31 features)	**.454**	—
w/o LLM	.395	**−.059**
w/o Lexical	.430	−.024
w/o POS	.453	−.001
w/o Embedding	.463	**+.010**

Ridge, narrative-grouped 5-fold CV: no narrative in both train and test.

LLMs carry the model.

Embeddings add nothing. Dropping all four *improves* the fit.

POS is noise.

> Hybrid vs. Qwen: Williams $t = 4.61$, $p < .0001$

Where every method fails

EXPERTS SCORED THIS PAIR HIGH. NO METHOD DID.

> *“Mon père dit que je suis courageux et que je peux réussir en Europe”*
> (My father says I can succeed in Europe)

> *“Je me suis dit que je ne peux pas rester comme ça”*
> (I told myself I cannot stay like this)

No shared words. No shared syntax. No shared theme keywords.

> Both perform the same act: **talking yourself into leaving**. The parallel is purely pragmatic.

4

Closing

What this study cannot tell you

- **Small, lopsided corpus.** 99 Trans-Saharan vs. 9 Balkan narratives.
- **Noisy gold standard.** $\alpha = 0.273$; 533 of 816 pairs single-annotated.
- **Only 7B models, 4-bit.** Larger models may behave differently.
- **No fine-tuning**, by design of the annotation-free framing.
- **Closed data.** Speakers are minors; we release the pipeline instead.

Takeaways

A new task. Experiential intertextuality, scored without annotation.

Hard by nature. Best single method: 56.6% of ceiling. Hybrid: 68.6%.

LLMs subsume embeddings. Dropping encoders helps.

Prompting is model-specific. Qwen -23%, Mistral $+144\%$.

Departure travels. Leaving echoes across routes.

Domain knowledge holds up. 130 terms beat 3 of 4 encoders.

Thank you

Questions welcome

Sakayo Toadoum Sari · sakayo@cril.fr

github.com/Toadoum/IntertextMigra

ANR HYCI (ANR-22-CE55-0010) · Région Hauts-de-France