Rabat, Morocco · EACL 2026

HarfoSokhan

A Comprehensive Parallel Dataset for Transitions between Persian Colloquial and Formal Variations

H.J. Sarvestani[1], **V. Ramezanian**[1], **S. Saadat**[2], **N.T. Serajeh**[2], **M.S. Razavi**[1], **Sh. Kasaei**[1], **M.A. Fazli**[1], **E. Asgari**[3]

[1] Sharif University of Technology · [2] University of Bonn · [3] QCRI-HBKU

Best Resource Award EACL 2026

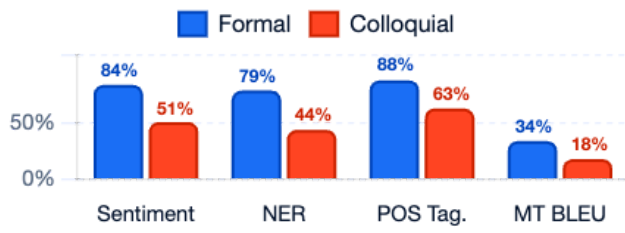
PERSIAN: A GLOBAL LANGUAGE

Official (Iran, Afghanistan, Tajikistan) · Significant minority · Historical / diaspora

100M+	24th	10th	3+
Native speakers worldwide	Most spoken language globally	Largest internet presence	Official forms: Farsi / Dari / Tajik

▲ MOTIVATION

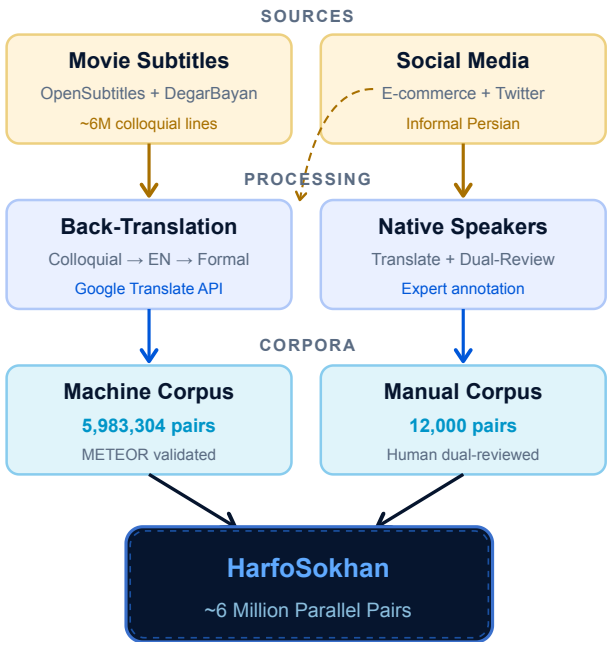
NLP models trained on **formal text** fail on **colloquial input** — which dominates social media and daily speech.



- ↓33% Sentiment
- ↓35% NER
- ↓25% POS Tag
- ↓16pt MT BLEU

💡 **Solution:** Normalise colloquial → formal *before* the NLP pipeline

⬡ DATASET ARCHITECTURE



✦ FORMAL VS. COLLOQUIAL PERSIAN

Shekaste-nevisi spans **syntactic**, **morphological**, and **phonological** levels — far deeper than word substitution.

Verb conjugation	Vocabulary shift
-ید → -ین, -ند → -ن	*bozorg* → *gondeh*
Pronunciation	**Stem reduction**
آن → اون	رفتن: رو → ر

Transformation Example

COLLOQUIAL: داره میره → FORMAL: در حال رفتن است

"He is going" — full syntactic restructuring

Colloquial Persian → English bridge → Formal Persian

METEOR evaluation: machine quality ≈ human quality

◈ DATASET STATISTICS

6M	12K	533K	2.8×
parallel pairs total	manual expert pairs	colloquial unique tokens	more unique tokens vs formal

Statistic	Formal	Colloquial
# Sentences	5,995,304	5,995,304
# Tokens	67.2M	62.8M
# Unique tokens	191K	**533K**
Avg. length	48.37	47.18

Colloquial Persian has **2.8× more unique tokens** — reflecting high lexical creativity of informal speech.

⚙ MODELS

GPT2-Based · 117M params
· GPT2-Manual — 12K pairs
★ **GPT2-HarfoSokhan — 6M**

T5-Based · 275M params
· T5-Manual — 12K pairs
· T5-HarfoSokhan — 6M

Baselines
· FarsiYar (rule-based)
· GPT-3.5-turbo

🏆 RESULTS: HUMAN EVALUATION

Top-rank frequency score (%) — 10 native Persian speakers · 200 sentences each · 6 models ranked blind

Model	top@1	top@2	top@3
GPT2-Manual	8.3	25.6	45.4
T5-Manual	4.4	12.6	23.0
T5-HarfoSokhan	5.4	19.1	38.4
FarsiYar	18.0	43.5	65.4
GPT-3.5-turbo	20.5	37.5	53.9
GPT2-HarfoSokhan	**43.0**	**61.4**	**73.5**

◆ KEY CONTRIBUTIONS

- **First large-scale** parallel corpus for Persian colloquial ↔ formal normalization
- **6M sentence pairs** via hybrid pipeline: 12K expert-annotated + auto-scaled
- **Recovers up to 35%** accuracy loss in downstream NLP caused by colloquial drift
- Fine-tuned **GPT2-HS outperforms GPT-3.5-turbo by 2.1×** in native-speaker ranking
- **Fully open:** dataset, models & evaluation suite on HuggingFace

⚠ BLEU VS. HUMAN EVALUATION: A KEY INSIGHT

FarsiYar · BLEU **0.697** ▼ #4 human rank ≠ GPT2-HS · BLEU **0.338** ▲ #1 human rank

BLEU rewards **token overlap** even when formalization is *incomplete*. FarsiYar keeps colloquial structure — high BLEU, last in human ranking. Human eval reveals true quality.

Input: داره میره *("He is going")*

FarsiYar · #1 BLEU · #4 human
داره می رود
Partial — colloquial structure remains

GPT2-HS · #2 BLEU · #1 human
در حال رفتن است
Fully formal — complete restructuring

Contact: asgari@hbku.edu.qa

huggingface.co/datasets/llm-lab/HarfoSokhan

EACL 2026 · Long Paper #346