Rabat, Morocco | March 26, 2026 | Long Paper #346

EACL 2026
RABAT · MOROCCO
Mars · March 24-29, 2026 · مارس

HarfoSokhan

حرف و سخن

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

Best Resource Award

↔ Text Style Transfer | 6M Sentence Pairs | Persian NLP | NLG








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

Outline

Motivation & Problem

- Persian language landscape
- Performance drop: formal → colloquial
- Resource scarcity

Models

- GPT2-based (decoder-only)
- T5-based (encoder-decoder)
- Baselines

HarfoSokhan Dataset

- Manual corpus — 12K pairs
- Machine-generated — 5.98M pairs
- Dataset statistics

Evaluation & Results

- Human evaluation (ranking-based)
- BLEU scores
- LLM-as-a-Judge

Motivation: Persian NLP in the Real World

Persian Language

- **100M+** native speakers worldwide
- Ranked **24th** most spoken globally
- **10th** in internet content
- Spoken in Iran, Afghanistan, Tajikistan, and Central Asia

The Core Problem

- NLP models trained on **formal** text fail on **colloquial** input
- Social media & messaging = colloquial Persian
- No large-scale **parallel** dataset existed

Our Approach

HarfoSokhan

First large-scale parallel corpus

6 Million

colloquial ↔ formal Persian pairs

Four Types of Changes:

- **Verb conjugation** — suffix changes
  -ید → -ین, -ند → -ن, -د → -ه
- **Stem changes** — shorter colloquial forms
  رفتن: stem رو → ر (colloquial)
- **Vocabulary substitution**
  bozorg → *gondeh* (big), *sar* → *kalleh* (head)
- **Pronunciation shifts**
  اونا → آنها, اون → آن

Shekaste-nevisi

Changes span **syntactic**, **morphological**, and **phonological** levels — far deeper than word substitution.

↔ Transformation Example

Colloquial:
می‌ره داره او

↓

Formal:
است رفتن حال در او

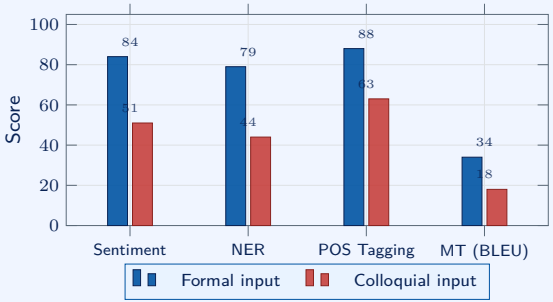
(He is going)

Example of changes

Feature	Example
Deletion	کتابا → کتاب‌ها
Substitution	بارون → باران
Insertion	اون → آن
Interchange	قلف → قفل

Motivation: Significant Performance Drop on Colloquial Text

NLP Task Performance: Formal vs. Colloquial Persian



Drop Summary

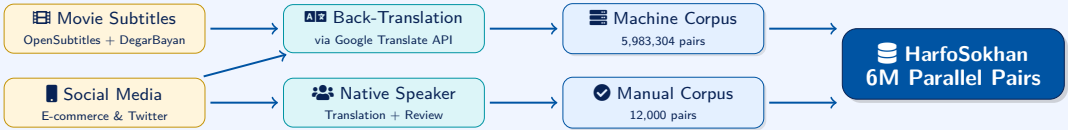
Task	Formal	Drop
Sentiment	84%	↓33%
NER	79%	↓35%
POS Tagging	88%	↓25%
MT	34 pt	↓16 pt

Root Cause

- Models trained on formal text only
- Colloquial has **2.8**× more unique tokens
- Language drifts unseen at train time

A Solution: normalise colloquial → formal *before* the NLP pipeline

HarfoSokhan: Dataset Architecture



Manual Corpus

Human-translated & **dual-reviewed** by native Persian speakers with formal education

Machine Corpus

Back-translation: Colloquial → English → Formal Persian, validated by METEOR scores

Manual Dataset: Quality over Quantity

Data Sources

- **3K lines** from DegarBayan subtitle dataset
- **9K sentences** crawled from Persian e-commerce platforms and social media

Quality Pipeline

1. Native speaker translates colloquial → formal
2. **Second independent reviewer** verifies
3. Covers full spectrum of colloquial phenomena

Source Selection Rationale

Rich colloquial phenomena captured from:

- Movie & TV dialogue
- Product reviews
- Social media posts
- Informal messaging

12K

verified sentence pairs

dual human-annotated

Large-Scale Machine-Generated Dataset

Back-Translation Pipeline

6M Colloquial Persian Subtitles

↓

Persian → English

wash out informal style

↓

English → Formal Persian

regenerate in formal register

↓

Formal Persian Output

Intuition

Translation through English reduces colloquial surface forms and encourages regeneration in a more **formal written register**.

Translation Quality

METEOR-based evaluation compares Google Translate output against two independent human annotators.

Machine quality ≈ Human quality

Data Sources

- **OpenSubtitles** (Tiedemann, 2012)
- **DegarBayan** dataset
- Predominantly spoken-language input

5.98M

machine-generated pairs

validated via METEOR

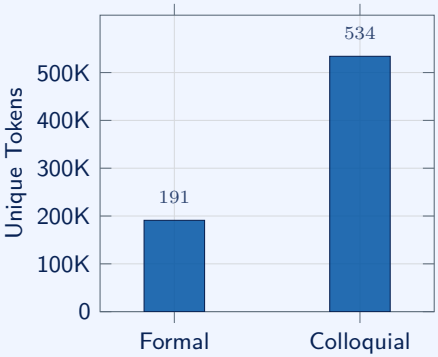
Table 1: HarfoSokhan at a Glance

Statistic	Formal	Colloquial
# Sentences	5,995,304	5,995,304
# Tokens	67,224,318	62,815,506
# Unique tokens	191,058	**533,773**
Avg. length	48.37	47.18

Key Observation

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

Unique Tokens: Formal vs. Colloquial



HuggingFace:
llm-lab/HarfoSokhan

Models for Colloquial-to-Formal Conversion

GPT2-Based (Decoder-Only)

Base: ParsGPT2 (Hooshvare Team, 2021)

- **117M parameters**
- Pre-trained on diverse Farsi corpora

Fine-tuned variants:

- **GPT2-Manual** — manual 12K pairs
- **GPT2-HarfoSokhan** — full 6M pairs ★

T5-Based (Encoder-Decoder)

Base: ParsT5 (Pouramini, 2021)

- **275M parameters**
- Trained on Farsi OSCAR corpus

Fine-tuned variants:

- **T5-Manual** — manual 12K pairs
- **T5-HarfoSokhan** — full 6M pairs

Comparison Baselines

FarsiYar (rule-based)
Replaces informal words with formal equivalents

GPT-3.5-turbo (OpenAI, 2022)
Large commercial LLM as a strong baseline

Three-Pronged Evaluation Framework

1 Human Evaluation

- **Gold standard**
- 10 native Persian speakers
- Each evaluates 200 sentences
- 6 models shown per sentence (sources concealed)
- **Ranking-based scoring**

Metric: top-rank frequency score

2 LLM-as-a-Judge

- Scalable alternative
- Contextual & semantic scoring
- Point-wise grading
- Multiple orderings tested (position bias mitigation)
- Reference-free metric

No reference translations needed

3 BLEU Score

- Traditional baseline
- n-gram overlap
- References: top-rated human translations
- 1,000 test sentences

Included for completeness

Why not BLEU alone? A model can score high on BLEU by making only *superficial word substitutions* without achieving true formality, human judgment captures deeper linguistic quality.

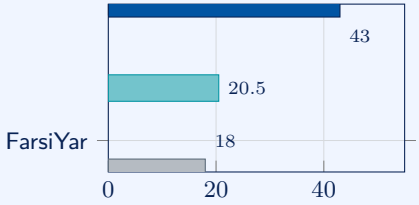
Table 2: Top-Rank Frequency Score (%)

Higher = more often ranked as the best formalization

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 Takeaways

- **GPT2-HarfoSokhan** outperforms GPT-3.5-turbo by **2.1**× on top@1
- Full 6M training ≫ 12K-only training
- Back-translation data is highly effective



Results: BLEU Scores & Qualitative Analysis

Table 3: BLEU Scores

Model	BLEU
GPT2-Manual	0.185
T5-Manual	0.038
T5-HarfoSokhan	0.164
FarsiYar	**0.697**
GPT-3.5-turbo	0.243
GPT2-HarfoSokhan	0.338

BLEU Can Mislead

FarsiYar is **#1 in BLEU** but only **#4 in human evaluation**. BLEU rewards token overlap, even when formalization is incomplete.

LLM-as-a-Judge

Colloquial input:
میره داره او

FarsiYar (#1 BLEU, #4 Human):
رود می داره او
Partial normalization; colloquial structure remains.

GPT2-HarfoSokhan (#2 BLEU, #1 Human):
است رفتن حال در او
Fully formal and syntactically restructured.

Length-Based Analysis

GPT2-HarfoSokhan ranks **#1** on **short** (<50 chars) and **medium** (50–150 chars) inputs. Models trained on the full **6M-pair** corpus outperform **12K-only** variants.

Thank You

Main Contributions

1. **HarfoSokhan**: the first large-scale Persian colloquial ↔ formal parallel dataset
2. A corpus of nearly **6 million sentence pairs**, combining manual and machine-generated data
3. A practical normalization framework for improving downstream NLP on colloquial Persian
4. Public release of the dataset and trained models

Main Results

- **GPT2-HarfoSokhan** outperformed stronger baselines, including **GPT-3.5-turbo**, in human evaluation
- Best human ranking result: **43.0% top@1** vs. **20.5%** for GPT-3.5-turbo
- Large-scale training data was more effective than small manual-only fine-tuning
- BLEU alone was insufficient; human evaluation and LLM-as-a-Judge gave better insight

EACL 2026
RABAT • MOROCCO
Mars • March 24-29, 2026 • مارس

Resources: huggingface.co/datasets/llm-lab/HarfoSokhan