Script Normalization for Unconventional Writing of Under-Resourced Languages

Sina Ahmadi and Antonios Anastasopoulos
George Mason University

The 61st Annual Meeting of the
Association for Computational Linguistics

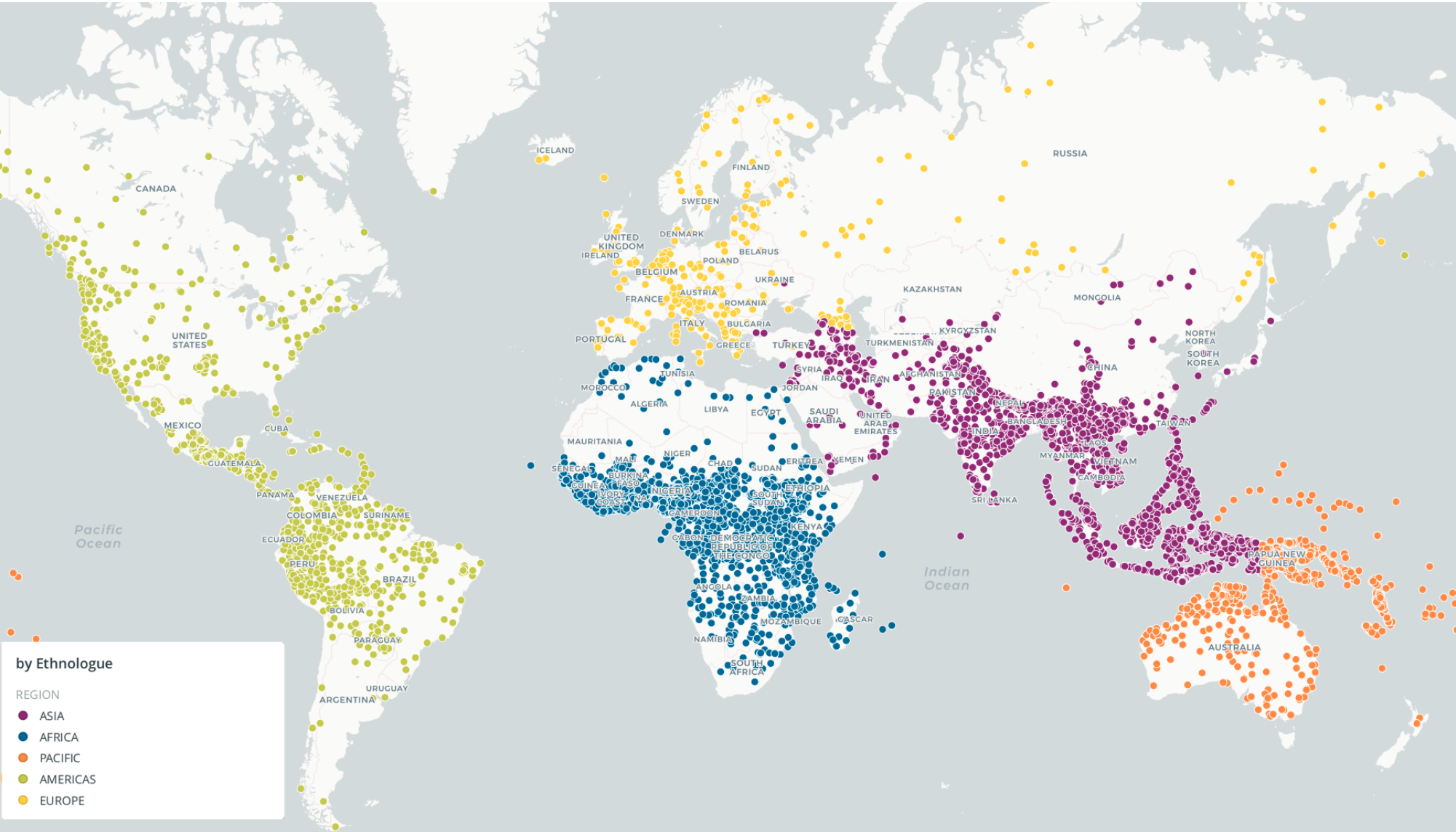
July 10, 2023

GEORGE MASON UNIVERSITY

Context

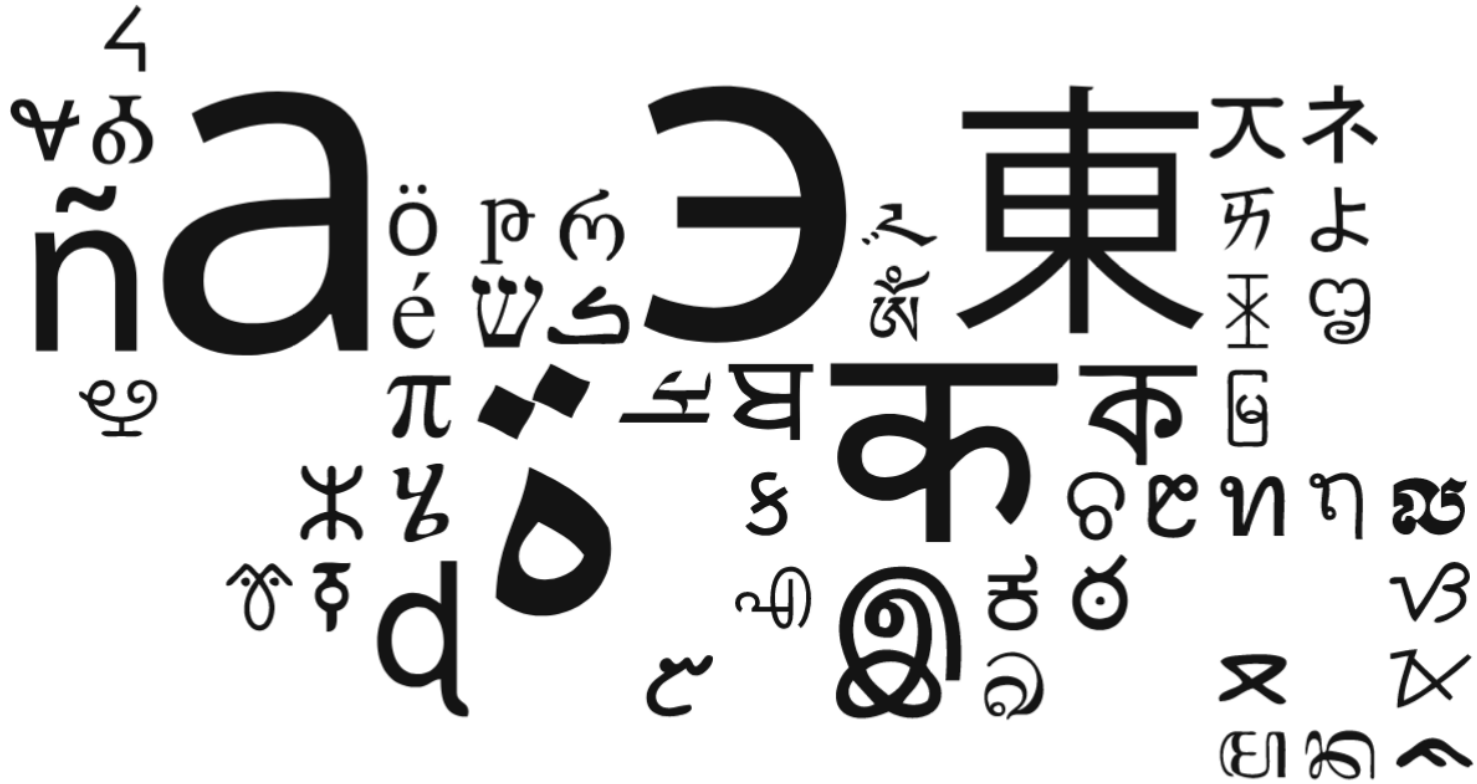
Context: Languages and Writing Systems

- More than 7,000 “languages” are spoken (Ethnologue, 2023).



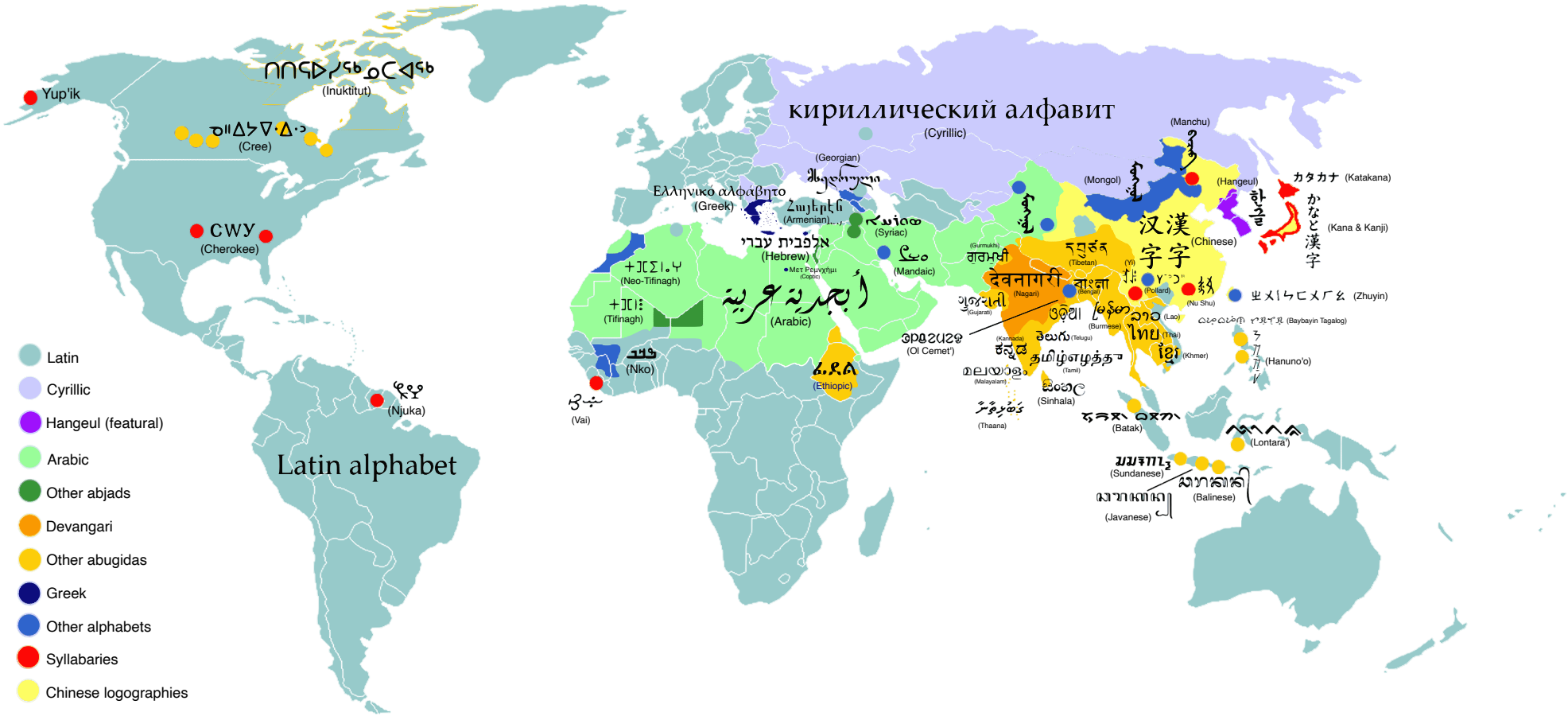
Context: Languages and Writing Systems

- More than 7,000 “languages” are spoken (Ethnologue, 2023).
- Almost 300 writing systems exist (and many adopted ones)



Context: Languages and Writing Systems

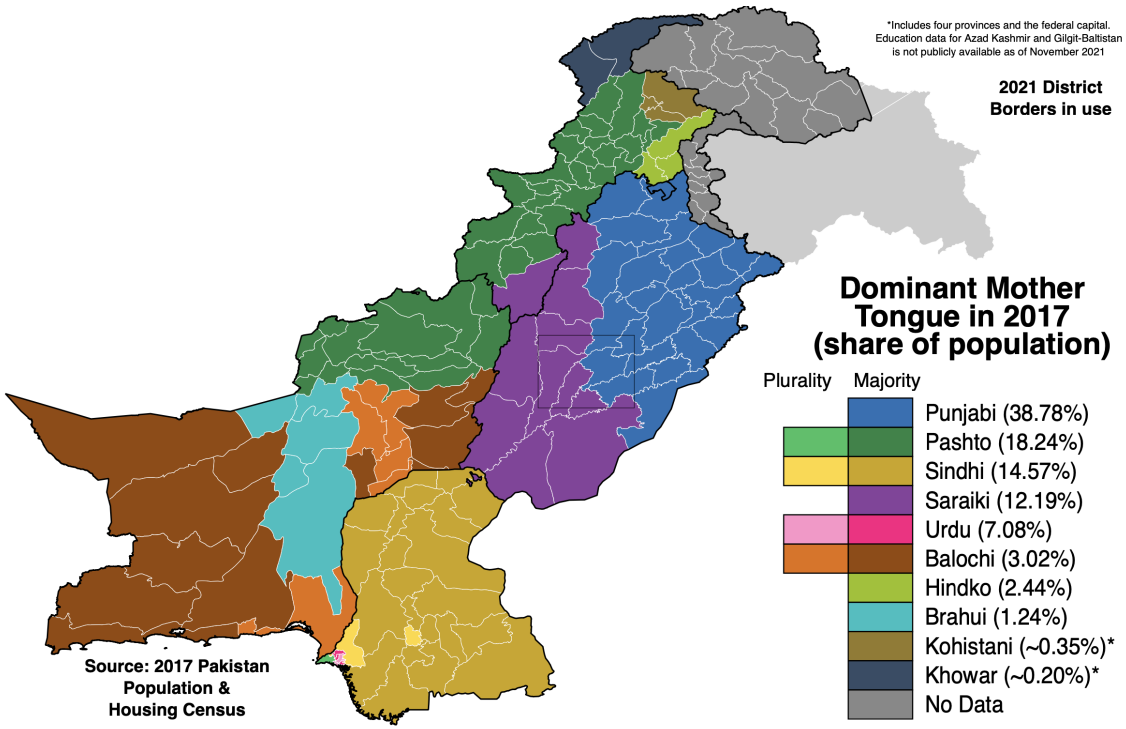
- More than 7,000 “languages” are spoken (Ethnologue, 2023).
- Almost 300 writing systems exist (and many adopted ones)
- Less than 4,000 languages have a written form



Context: Language Communities

Most countries are multi-lingual, but not all officially!

- Pakistan:
  $\rightarrow$ Urdu and English



Context: Language Communities

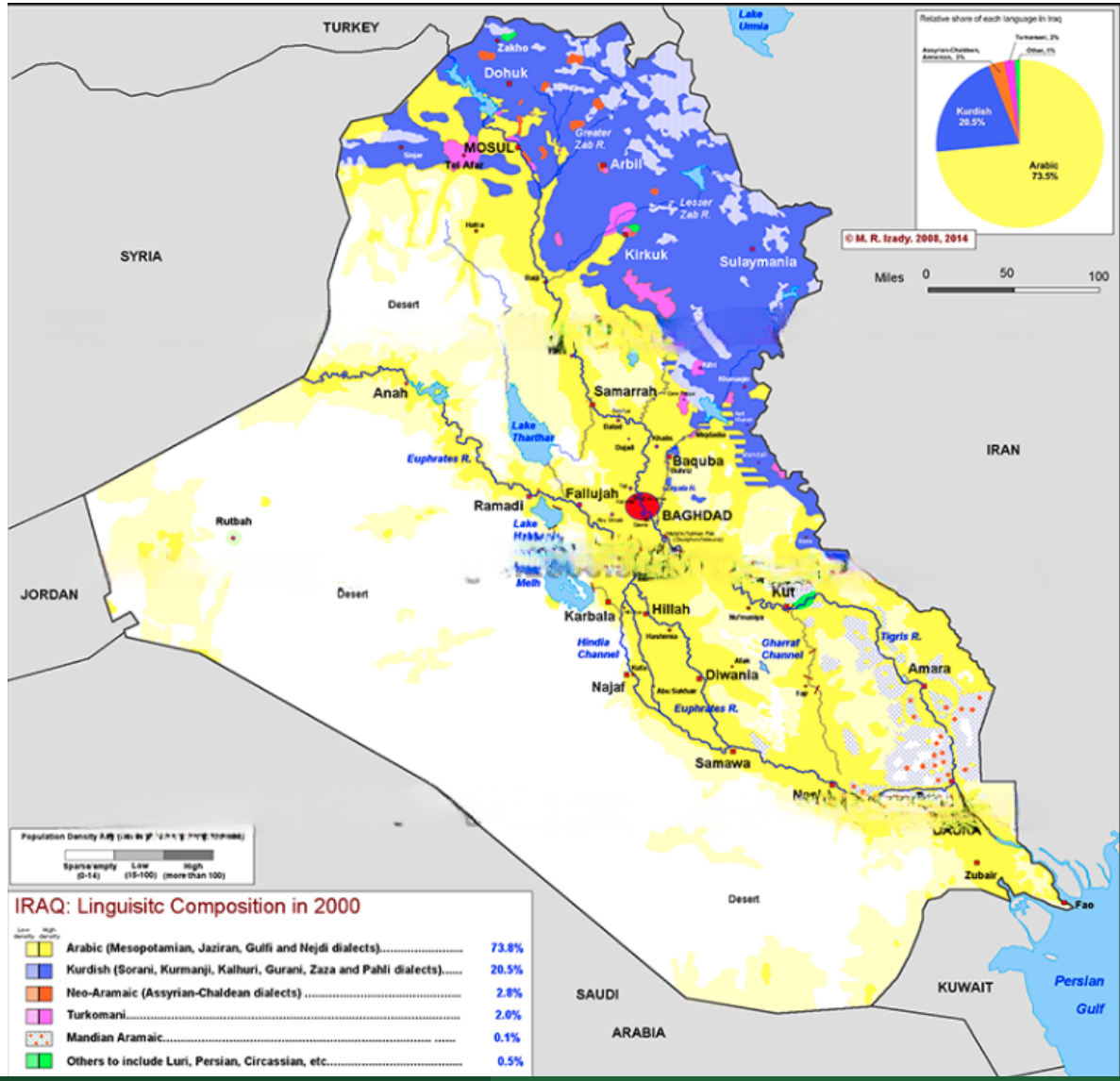
Most countries are multi-lingual, but not all officially!

- Pakistan:
  $\rightarrow$ Urdu and English
- India:
  $\rightarrow$ Hindi, Kashmiri, Sindhi and 20 more

Context: Language Communities

Most countries are multi-lingual, but not all officially!

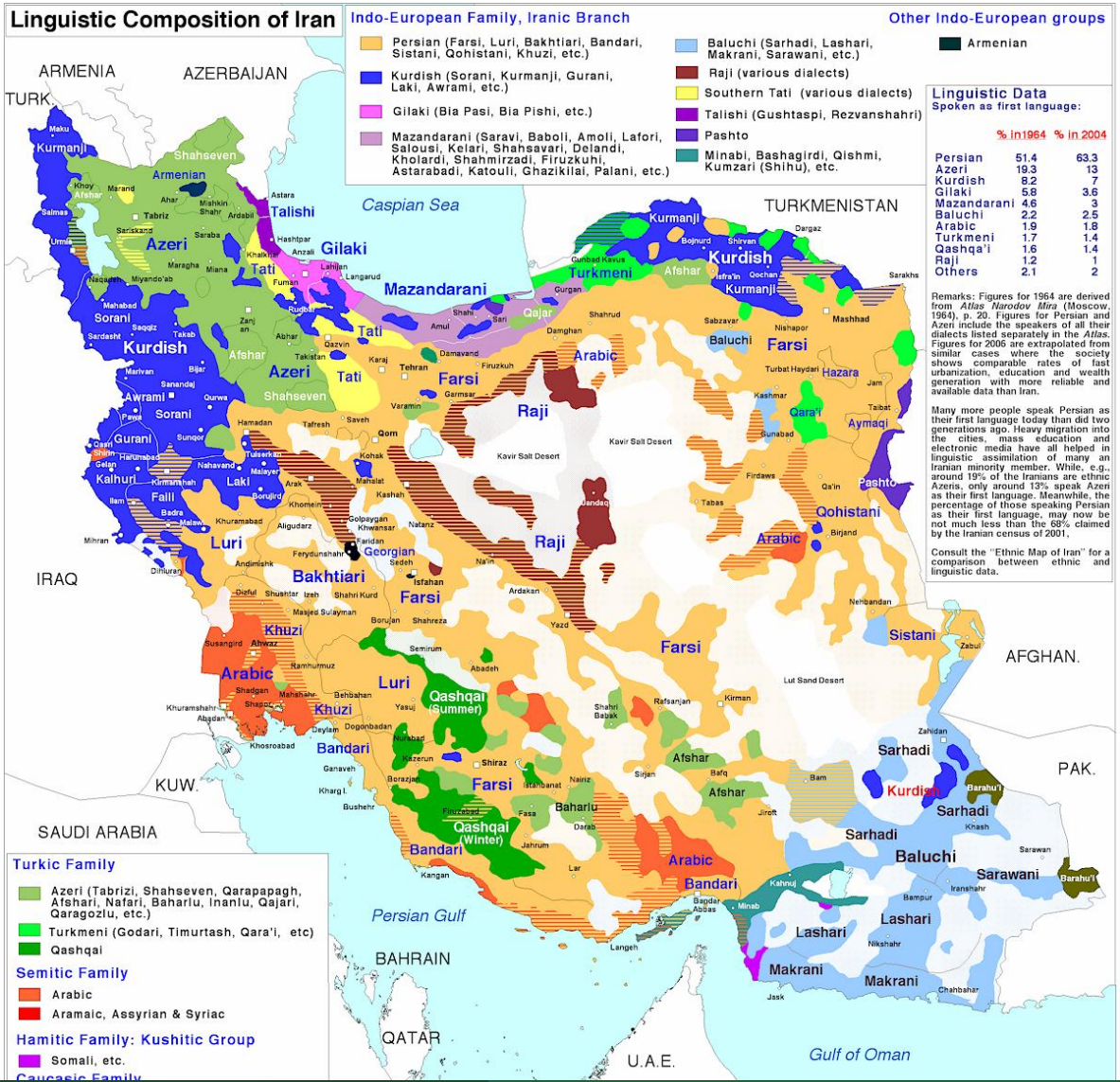
- Pakistan:
  → Urdu and English
- India:
  → Hindi, Kashmiri, Sindhi and 20 more
- Iraq:
  → Arabic and Kurdish



Context: Language Communities

Most countries are multi-lingual, but not all officially!

- Pakistan:
  → Urdu and English
- India:
  → Hindi, Kashmiri, Sindhi and 20 more
- Iraq:
  → Arabic and Kurdish
- Iran:
  → Persian!



Unconventional Writing

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching
- No specific rule for mapping graphemes-phonemes

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching
- No specific rule for mapping graphemes-phonemes

A few examples:

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching
- No specific rule for mapping graphemes-phonemes

A few examples:

- *ana raye7 el gam3a el sa3a 3 el 3asr.* (Ararbic in Latin script, aka Arabizi)

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching
- No specific rule for mapping graphemes-phonemes

A few examples:

- *ana raye7 el gam3a el sa3a 3 el 3asr.* (Ararbic in Latin script, aka Arabizi)
- *Tora ti na sou po* (Greek in Latin, aka Greeklish)

Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching
- No specific rule for mapping graphemes-phonemes

A few examples:

- *ana raye7 el gam3a el sa3a 3 el 3asr.* (Ararbic in Latin script, aka Arabizi)
- *Tora ti na sou po* (Greek in Latin, aka Greeklish)
- *mer6 pr tn mess pr mn anif* (French SMS language)

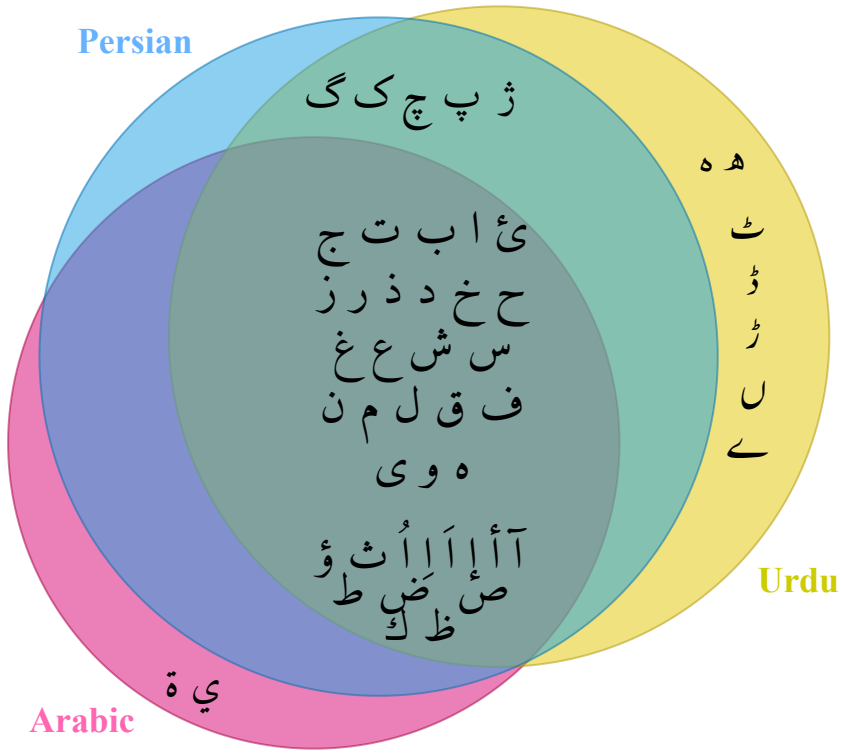
Unconventional Writing: Perso-Arabic scripts

Unconventional Writing: Perso-Arabic scripts

- Traditionally used for Arabic
- Used for over a millennium
- A *Reichssprache* for centuries

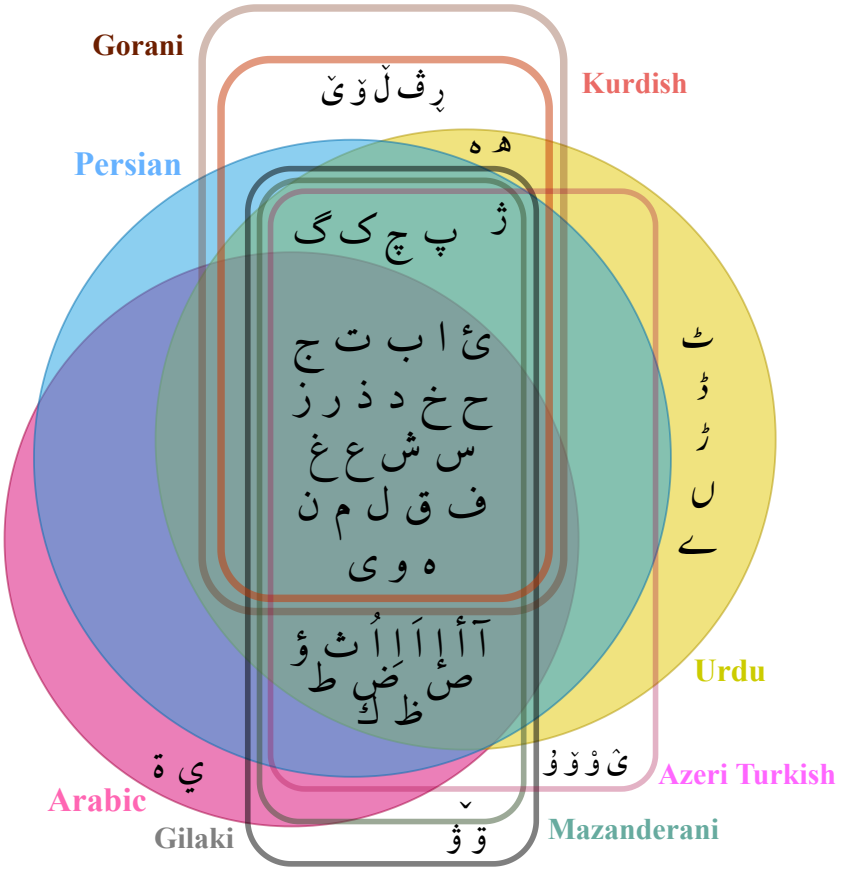
Unconventional Writing: Perso-Arabic scripts

- Traditionally used for Arabic
- Used for over a millennium
- A *Reichssprache* for centuries
- Have been adopted:
  Ajami, Wolofal, Pegon, Jawi, etc.
- **Perso-Arabic scripts:**



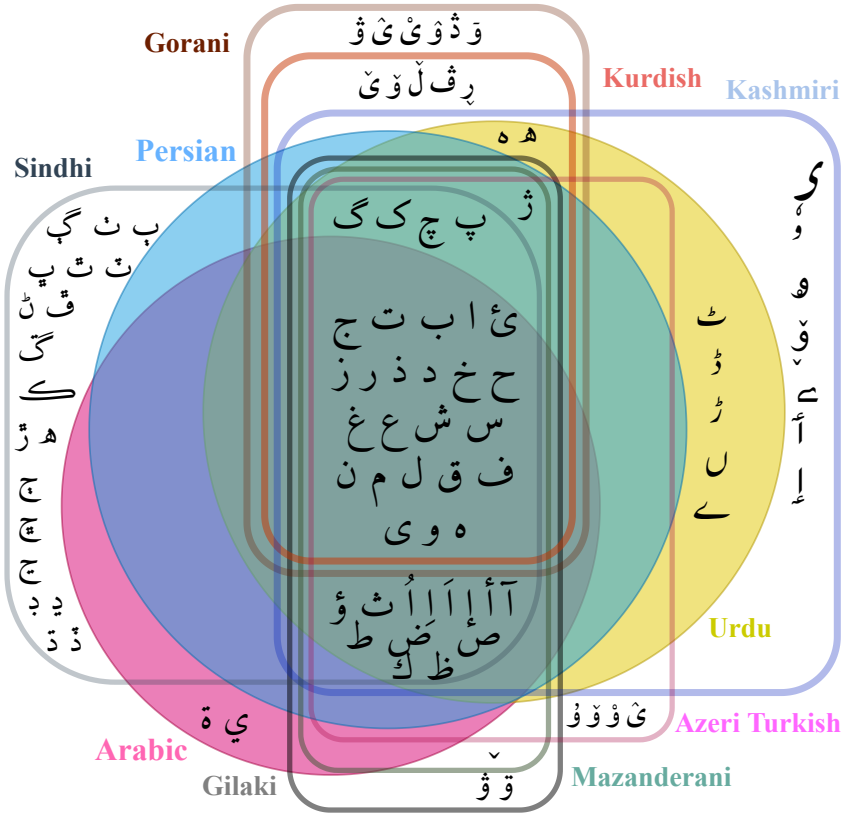
Unconventional Writing: Perso-Arabic scripts

- Traditionally used for Arabic
- Used for over a millennium
- A *Reichssprache* for centuries
- Have been adopted:
  Ajami, Wolofal, Pegon, Jawi, etc.
- **Perso-Arabic scripts:**
  - Languages in Iran and Iraq



Unconventional Writing: Perso-Arabic scripts

- Traditionally used for Arabic
- Used for over a millennium
- A *Reichssprache* for centuries
- Have been adopted:
  Ajami, Wolofal, Pegon, Jawi, etc.
- **Perso-Arabic scripts:**
  - Languages in Iran and Iraq
  - Languages in Pakistan and India



Unconventional Writing: the Pandora's Box

Script Normalization

Script Normalization

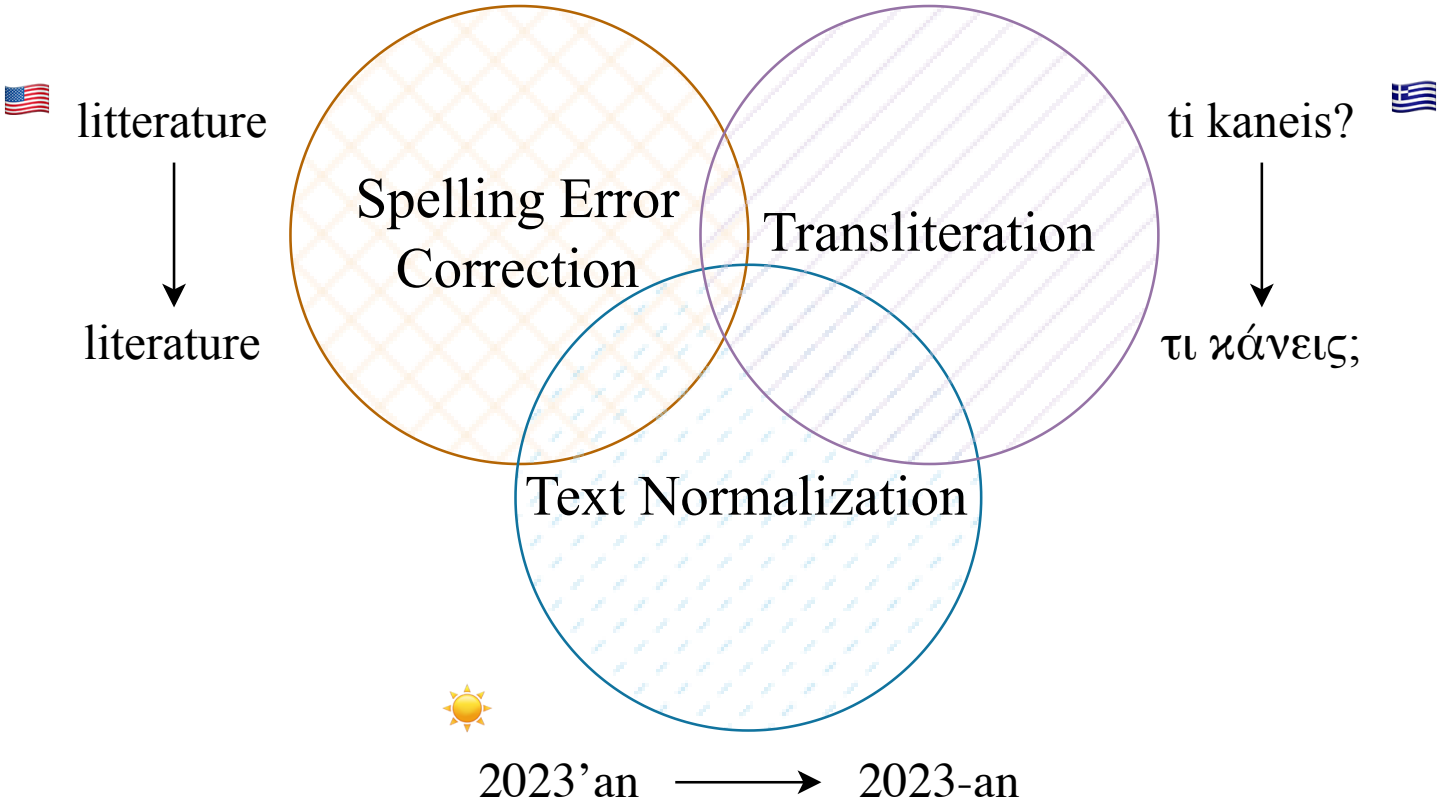
Script Normalization

Normalization of a text written in an unconventional script based on the conventional script and orthography

Script Normalization

Script Normalization

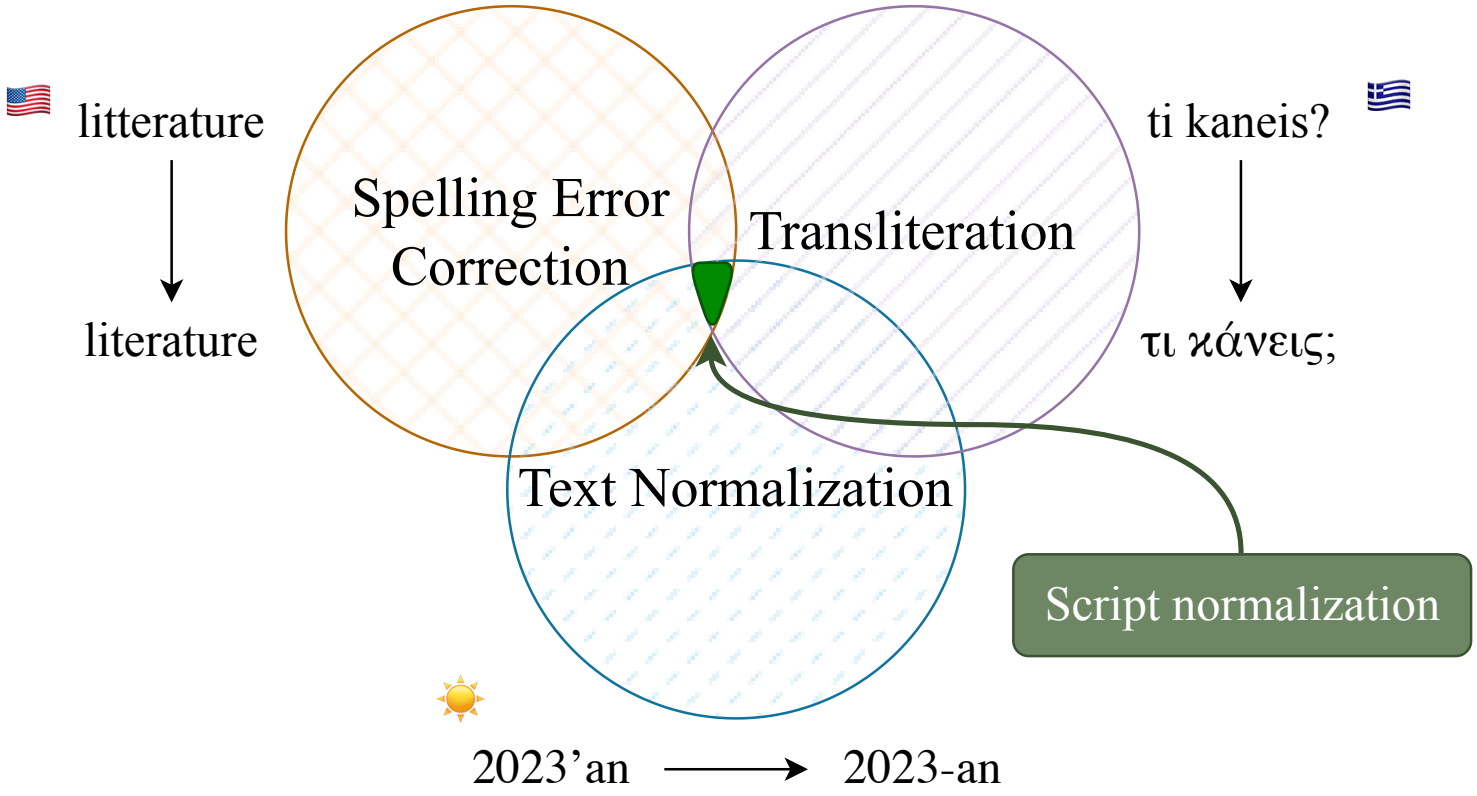
Normalization of a text written in an unconventional script based on the conventional script and orthography



Script Normalization

Script Normalization

Normalization of a text written in an unconventional script based on the conventional script and orthography



Script Normalization: Approach

1. Data collection – Not easy!

Language	639-3	WP	script type	diacritics	ZWNJ	Dominant
Azeri Turkish	azb	azb	Abjad	✓	✓	Persian
Kashmiri	kas	ks	Alphabet	✓	✗	Urdu
Gilaki	glk	glk	Abjad	✓	✓	Persian
Gorani	hac	-	Alphabet	✗	✗	Persian, Arabic, Sorani
Kurmanji	kmr	-	Alphabet	✗	✗	Persian, Arabic
Sorani	ckb	ckb	Alphabet	✗	✗	Persian, Arabic
Mazanderani	mzn	mzn	Abjad	✓	✓	Persian
Sindhi	snd	sd	Abjad	✓	✗	Urdu
Persian	fas	fa	Abjad	✓	✓	-
Arabic	arb	ar	Abjad	✓	✗	-
Urdu	urd	ur	Abjad	✓	✓	-

Script Normalization: Approach

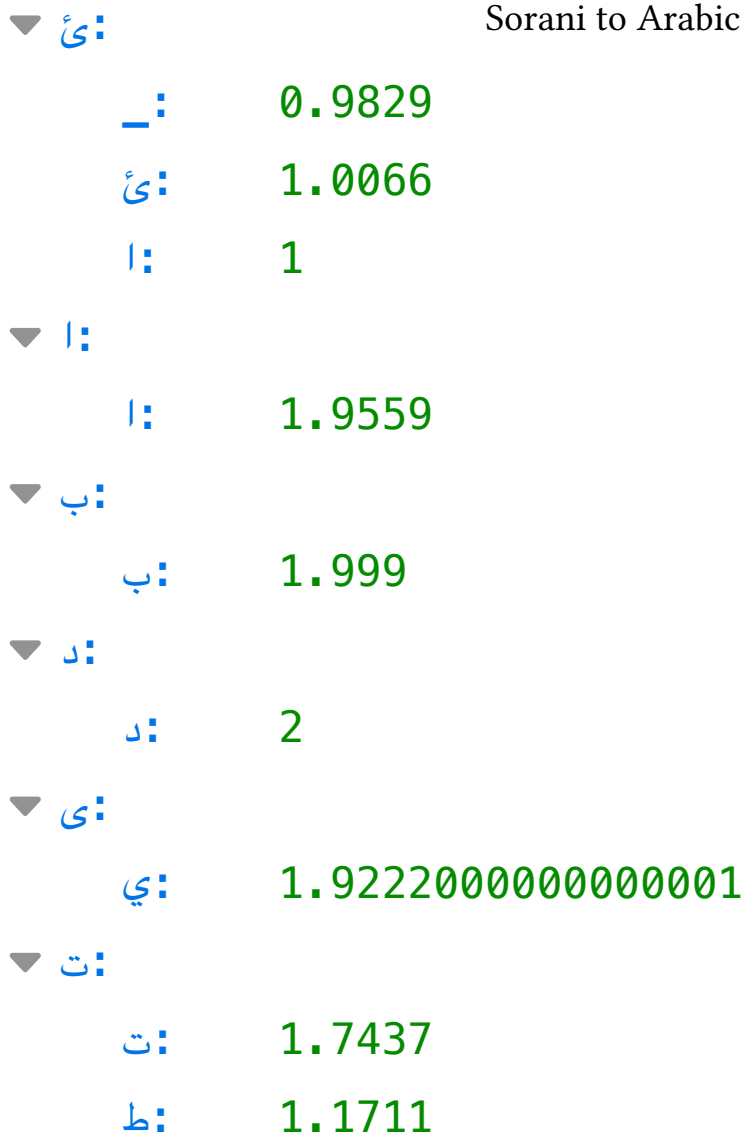
1. Data collection – Not easy!
2. Script mapping
   - Common characters
   - Visual resemblance
   - Orthographic rules

Language	**Unconventional script**	**Source**	**Target**
Azeri Turkish	Persian	چ	چ
Sorani	Arabic	ز	ذ / ض / ظ / ز
Kashmiri	Urdu	أ	ا / اُ
Sindhi	Urdu	ي	ے / ي / ى

Script Normalization: Approach

1. Data collection – Not easy!
2. Script mapping
   - Common characters
   - Visual resemblance
   - Orthographic rules
3. Character-alignment matrix
   → sequence alignment based on dictionaries

Sorani to Arabic



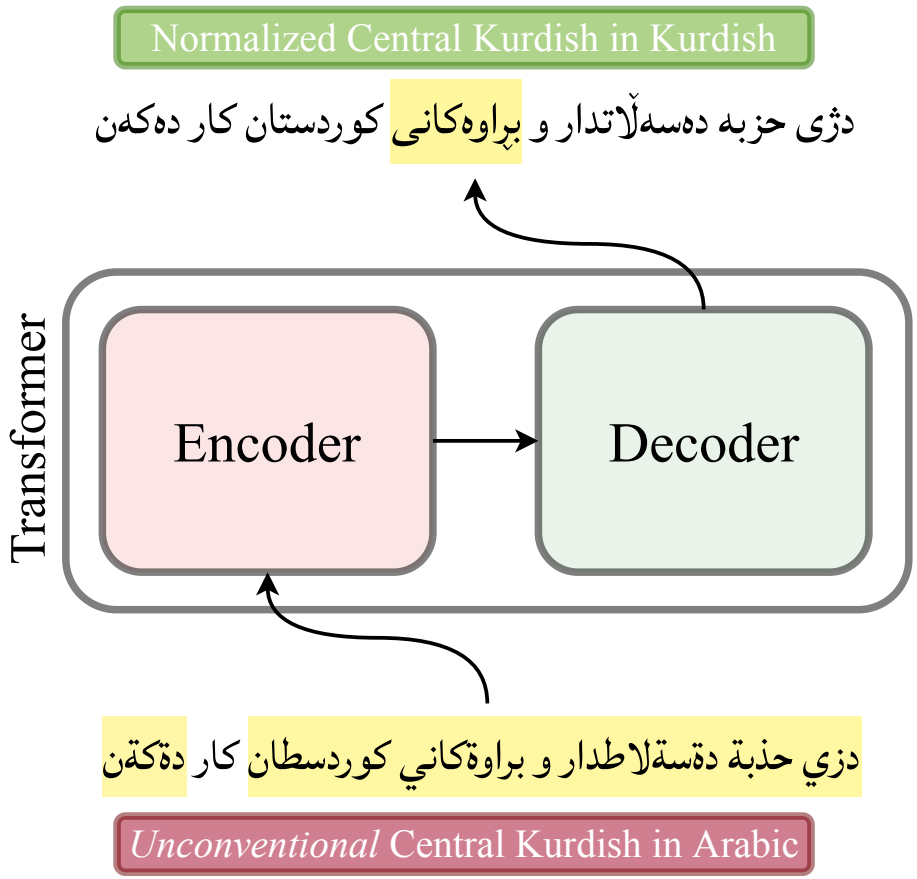
Script Normalization: Approach

1. Data collection – Not easy!
2. Script mapping
   - Common characters
   - Visual resemblance
   - Orthographic rules
3. Character-alignment matrix
   $\rightarrow$ sequence alignment based on dictionaries
4. Synthetic data generation
   $\rightarrow$ randomly generate pairs
   $\rightarrow$ inject noise

Noise %	Sentence
Clean	دووەمین پێشانگەها فۆتۆگرافەرێن کورد ل بەلجیکا Second Kurdish photographers' exhibition in Belgium
20	دووهمین پێشانکهها فۆتۆکرافهرێن کورد ل بهلجیکا
40	دووه مین بشانکه ها فطکرافه رن کورد ل به لجیکا
60	دووة مین بشانکة ها فوتوکرافة رن کورد ل بة لجیکا
80	دووةمین بیشانکةها فؤتؤکرافةرین کورد ل بةلجیکا
100	دووهمین بیشانکهها فوتوکرافهرین کورد ل بهلجیکا

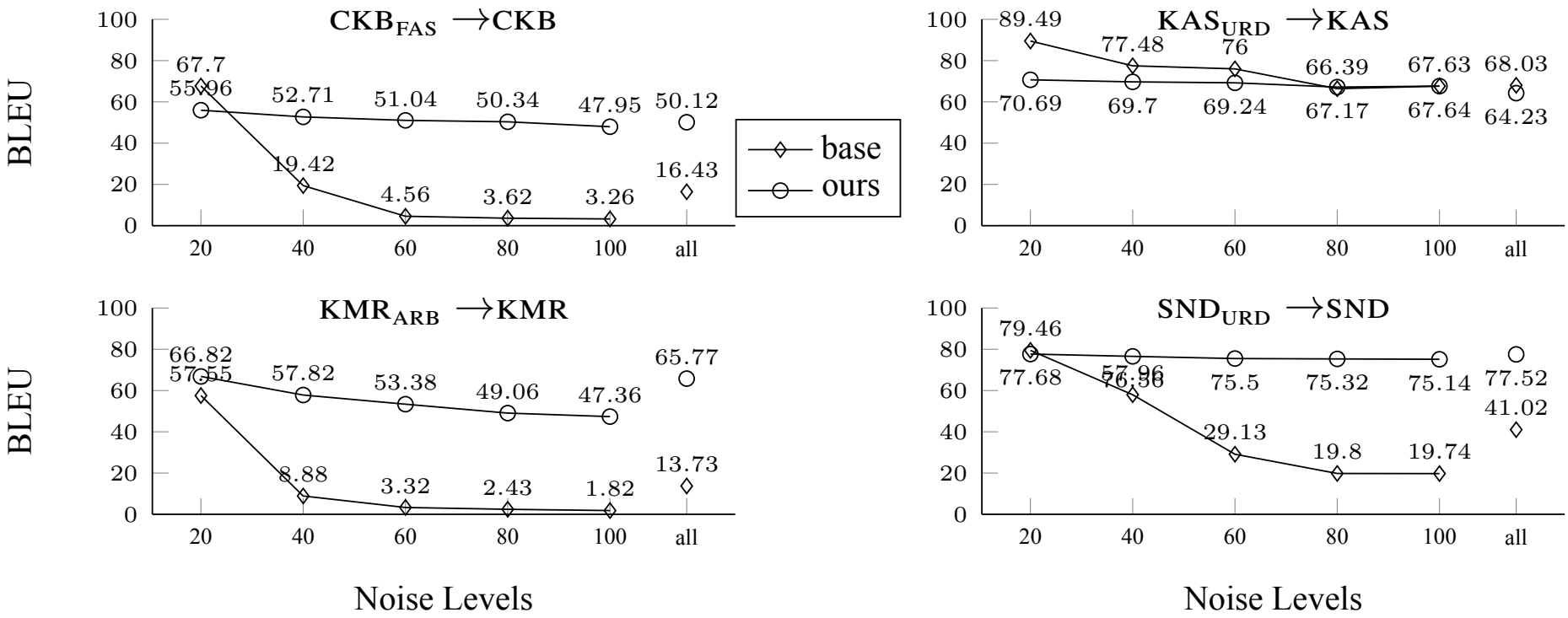
Script Normalization: Approach

1. Data collection – Not easy!
2. Script mapping
   - Common characters
   - Visual resemblance
   - Orthographic rules
3. Character-alignment matrix
   → sequence alignment based on dictionaries
4. Synthetic data generation
   → randomly generate pairs
   → inject noise
5. Model
   → encoder-decoder with self-attention



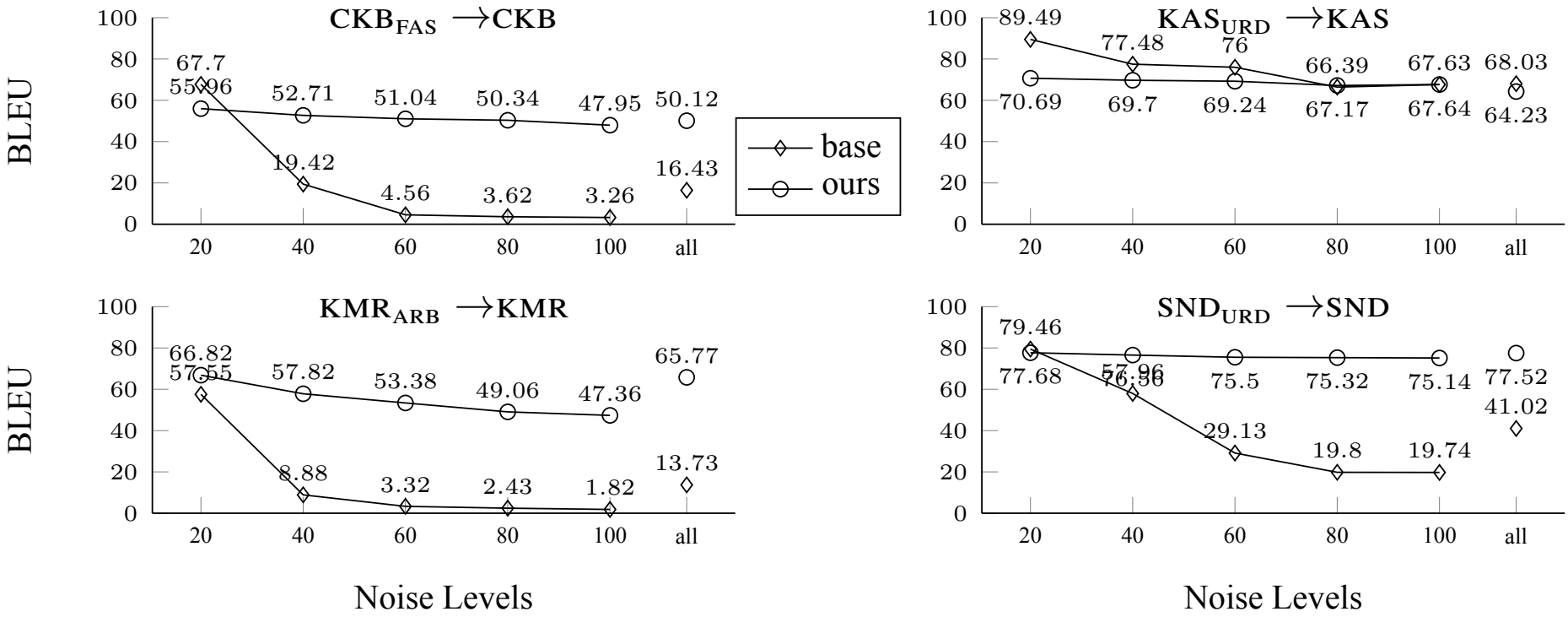
Script Normalization: Intrinsic Experiments

- Baseline: a naive "copy" system



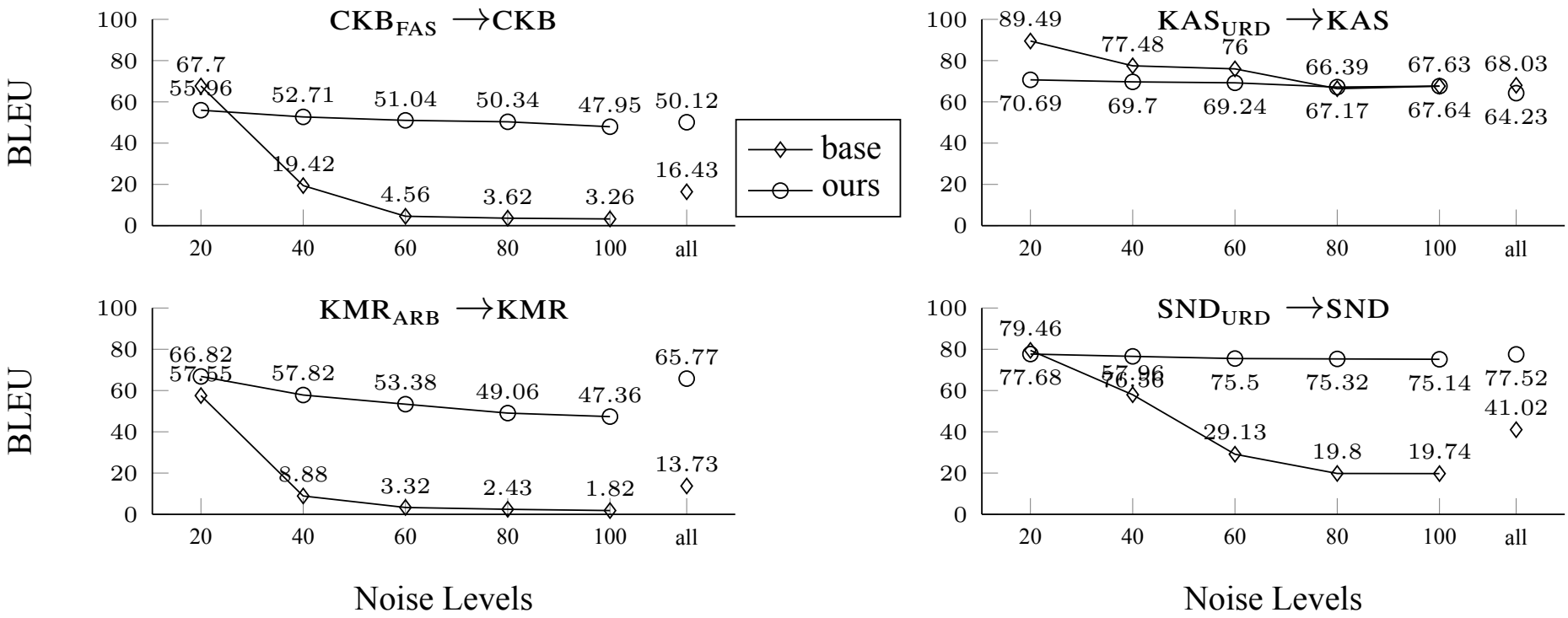
Script Normalization: Intrinsic Experiments

- Baseline: a naive "copy" system
- Ours: trained models on different levels of noise



Script Normalization: Intrinsic Experiments

- Baseline: a naive "copy" system
- Ours: trained models on different levels of noise
- **Our models dramatically improve over the baseline**



Script Normalization: Intrinsic Experiments

- Baseline: a naive "copy" system
- Ours: trained models on different levels of noise
- **Our models dramatically improve over the baseline**
- Including when evaluated on real data

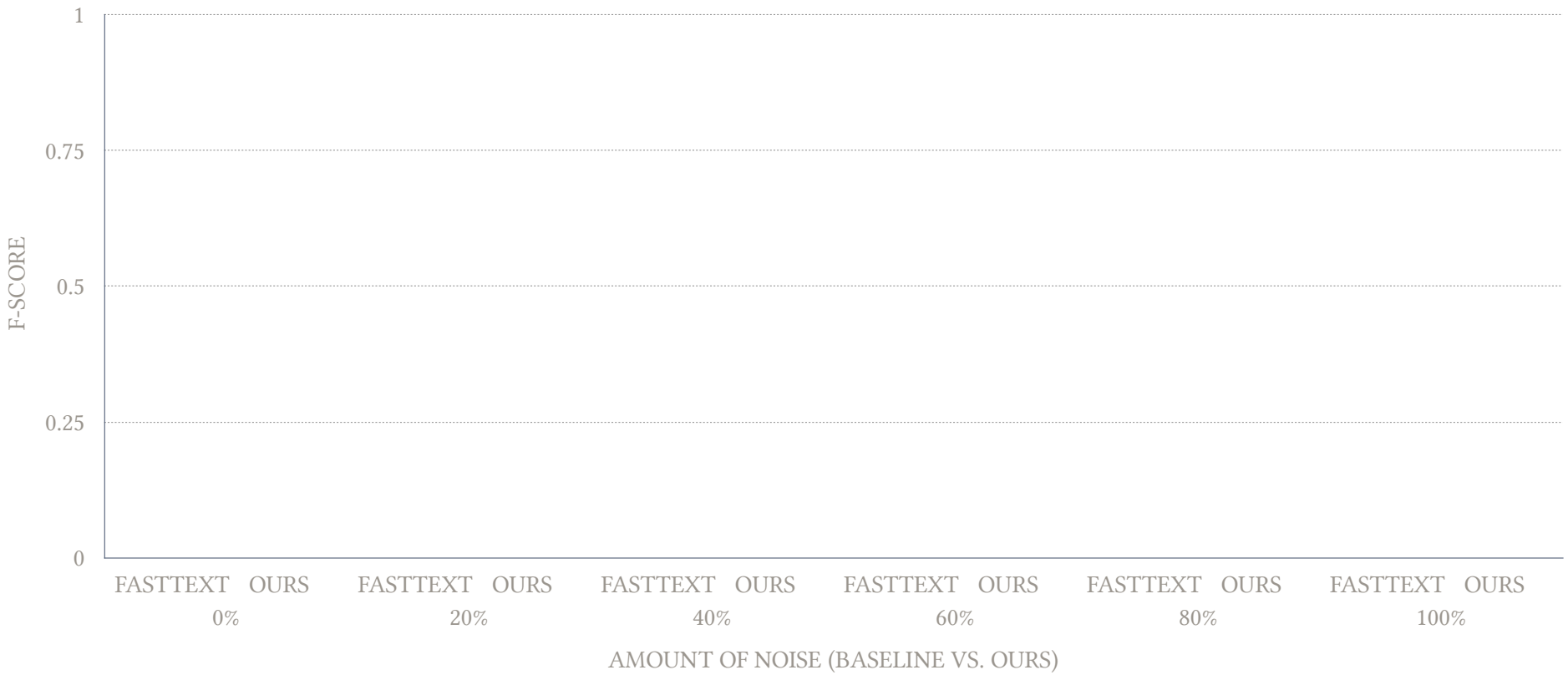
Sorani Eval Set	Original (Unconventional)		Normalized	
	BLEU	chrF	BLEU	chrF
$\text{CKB}_{\text{FAS}} \rightarrow \text{CKB}$	1.2	38.4	20.1	69.6
$\text{CKB}_{\text{ARB}} \rightarrow \text{CKB}$	0.4	19.4	12.7	65.2

Script Normalization: Extrinsic Experiments

1. **Language identification (LID)**

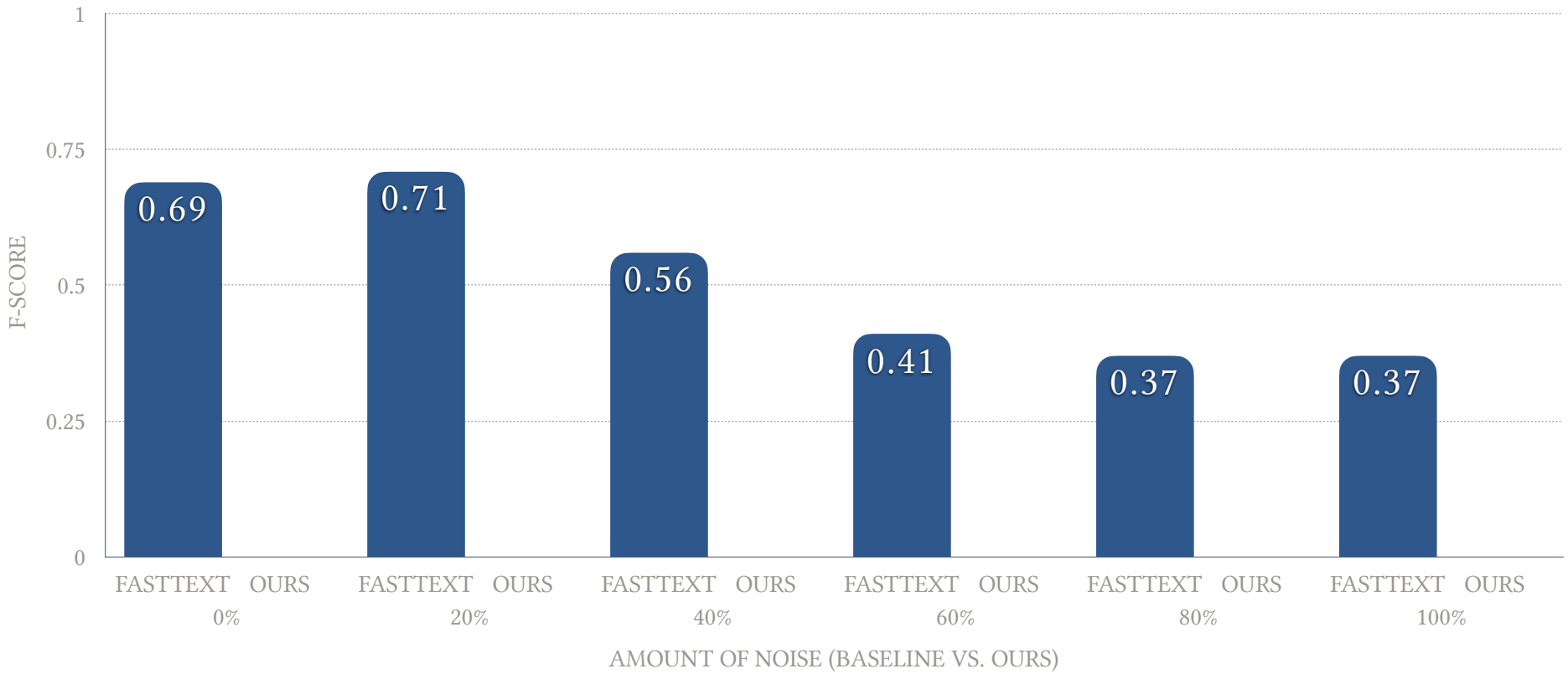
Script Normalization: Extrinsic Experiments

1. **Language identification (LID)**
   - Compare LID with and without normalization



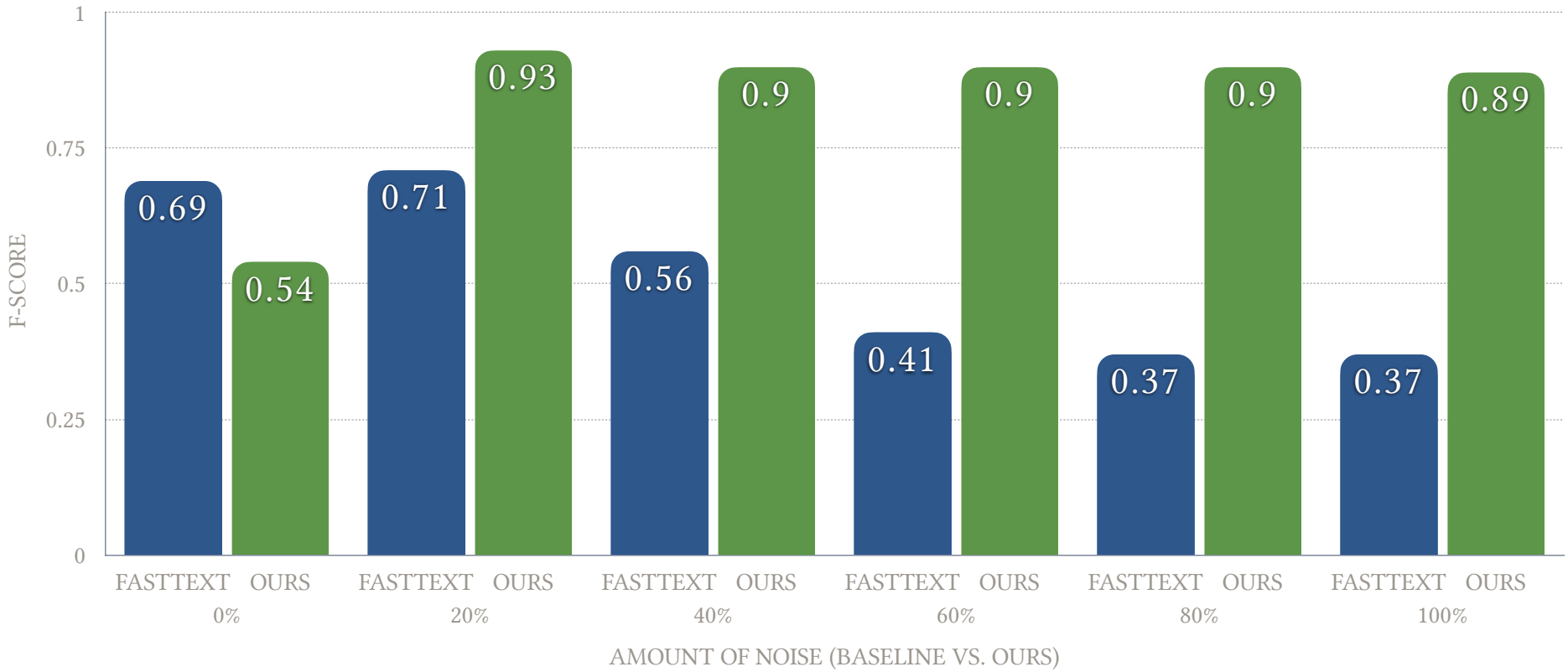
Script Normalization: Extrinsic Experiments

1. **Language identification (LID)**
   - Compare LID with and without normalization
   - Terrible performance by any existing model



Script Normalization: Extrinsic Experiments

1. **Language identification (LID)**
   - Compare LID with and without normalization
   - Terrible performance by any existing model
   - Models trained on normalized datasets improve the F-scores

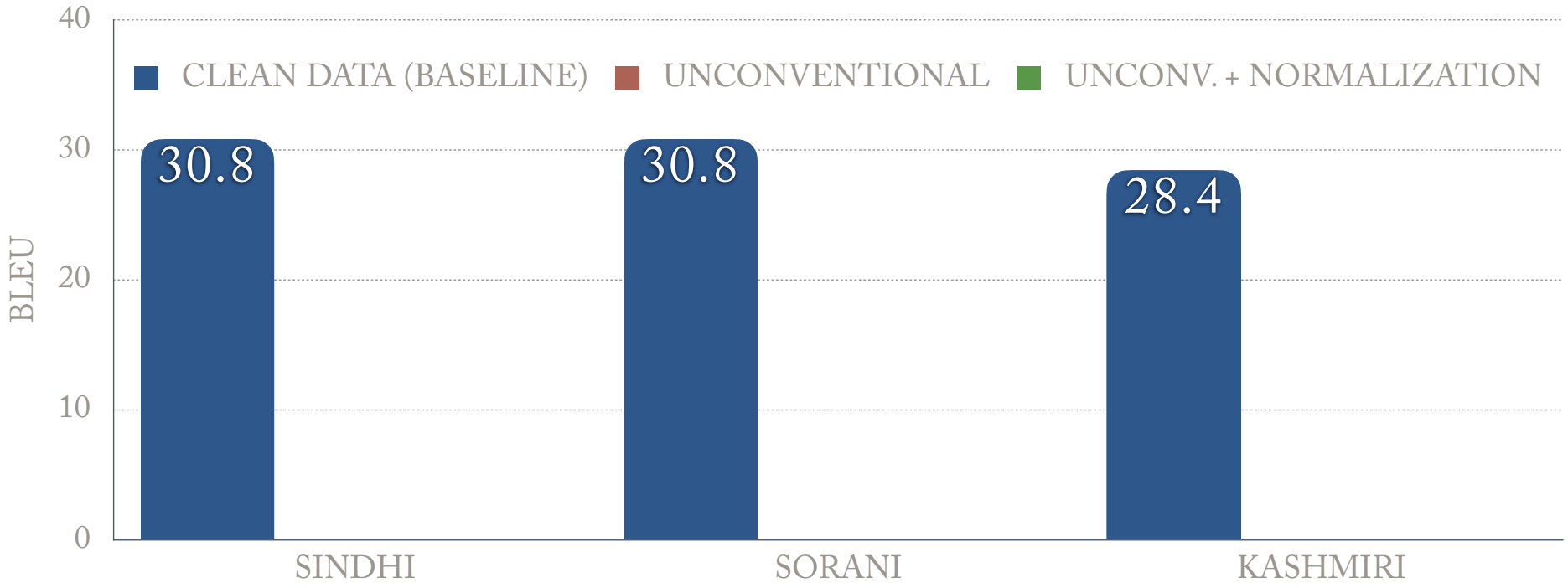


Script Normalization: Extrinsic Experiments

1. **Language identification (LID)**
   - Compare LID with and without normalization
   - Terrible performance by any existing model
   - Models trained on normalized datasets improve the F-scores
   - Closely-related languages (scripts) are confused!

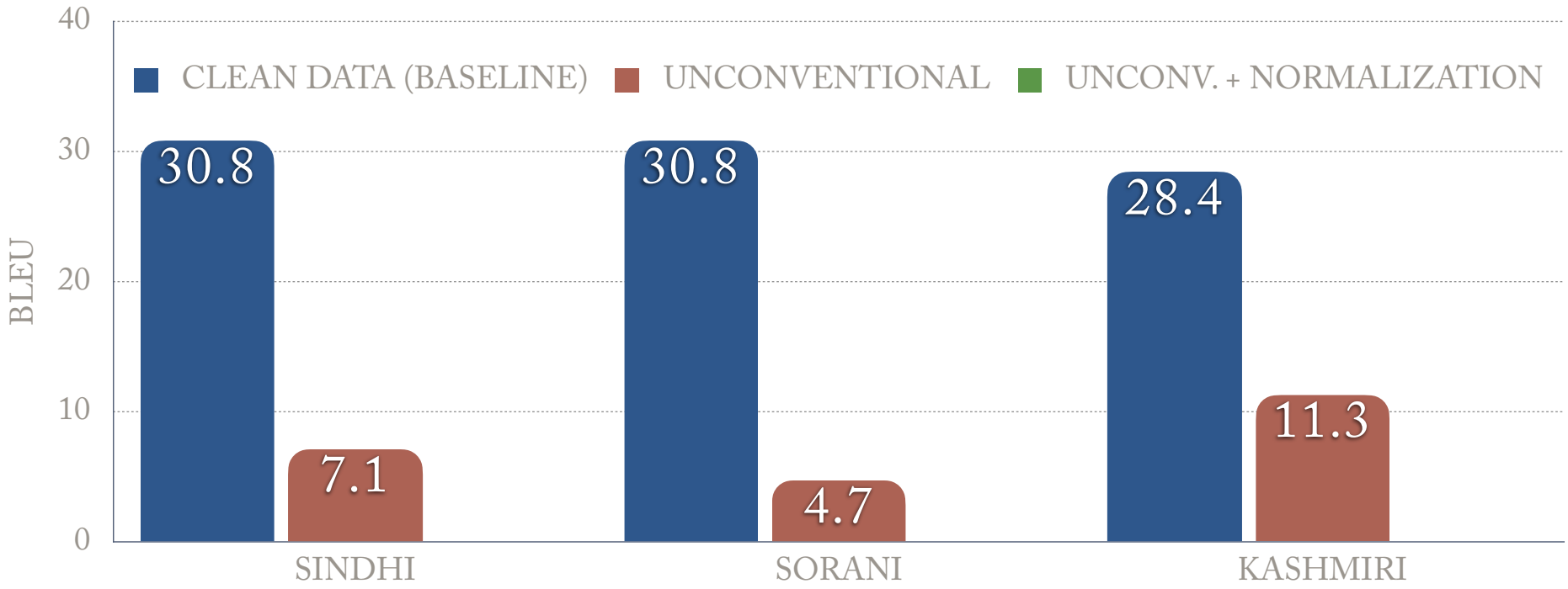
Script Normalization: Extrinsic Experiments

1. Language identification (LID)
2. **Machine Translation (MT)**
   - Evaluate MT with and without normalization



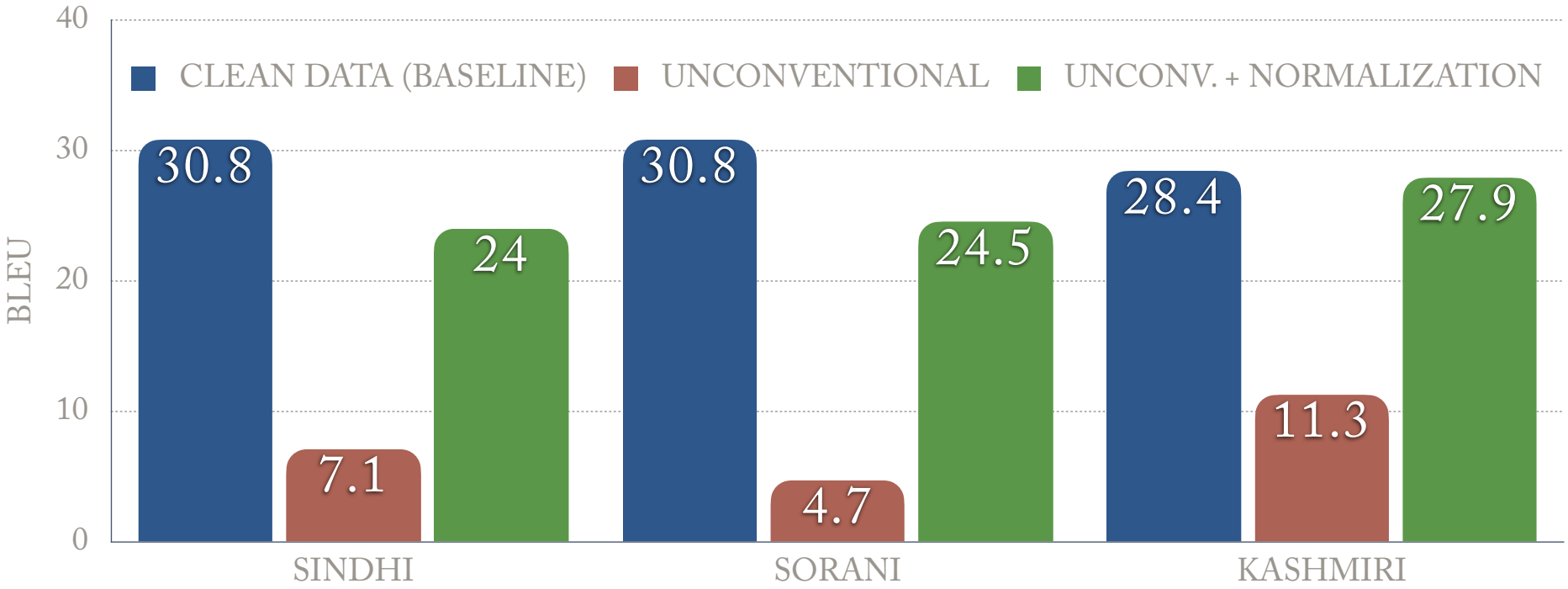
Script Normalization: Extrinsic Experiments

1. Language identification (LID)
2. **Machine Translation (MT)**
   - Evaluate MT with and without normalization
   - Terrible performance on noisy data (NLLB as baseline)



Script Normalization: Extrinsic Experiments

1. Language identification (LID)
2. **Machine Translation (MT)**
   - Evaluate MT with and without normalization
   - Terrible performance on noisy data (NLLB as baseline)
   - Models trained on normalized datasets improve the F-scores



Conclusion

Key Takeaways

1. Unconventional writing is more widespread than thought of.

Key Takeaways

1. Unconventional writing is more widespread than thought of.
2. It is an open problem and non-trivial to solve.

Key Takeaways

1. Unconventional writing is more widespread than thought of.
2. It is an open problem and non-trivial to solve.
3. It negatively affects NLP for low-resourced languages.

Key Takeaways

1. Unconventional writing is more widespread than thought of.
2. It is an open problem and non-trivial to solve.
3. It negatively affects NLP for low-resourced languages.
4. We can effectively remediate it, *but only to some extent*...

Key Takeaways

1. Unconventional writing is more widespread than thought of.
2. It is an open problem and non-trivial to solve.
3. It negatively affects NLP for low-resourced languages.
4. We can effectively remediate it, *but only to some extent*...
5. Do we always need to write a language?
   - Multi-modal NLP
   - Multi-lingual NLP
   - Multi-task NLP
   - Better adaptation in NLP

Key Takeaways

1. Unconventional writing is more widespread than thought of.
2. It is an open problem and non-trivial to solve.
3. It negatively affects NLP for low-resourced languages.
4. We can effectively remediate it, *but only to some extent*...
5. Do we always need to write a language?
   - Multi-modal NLP
   - Multi-lingual NLP
   - Multi-task NLP
   - Better adaptation in NLP
6. Models and codes:
   https://github.com/sinaahmadi/ScriptNormalization

Any questions?

References I

- Ahmadi, Sina, Milind Agarwal, and Antonios Anastasopoulos (May 2023). "PALI: A Language Identification Benchmark for Perso-Arabic Scripts". In: *Proceedings of the 10th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Dubrovnik, Croatia: The 17th Conference of the European Chapter of the Association for Computational Linguistics.
- Ahmadi, Sina and Antonios Anastasopoulos (2023). "Script Normalization for Unconventional Writing of Under-Resourced Languages in Bilingual Communities". In: Tornonto, Cananda: The 61st Annual Meeting of the Association for Computational Linguistics.
- Sheyholislami, Jaffer (2012). "Kurdish in Iran: A case of restricted and controlled tolerance". In: *International Journal of the Sociology of Language* 2012.217, pp. 19–47.