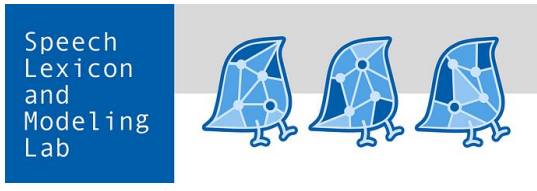
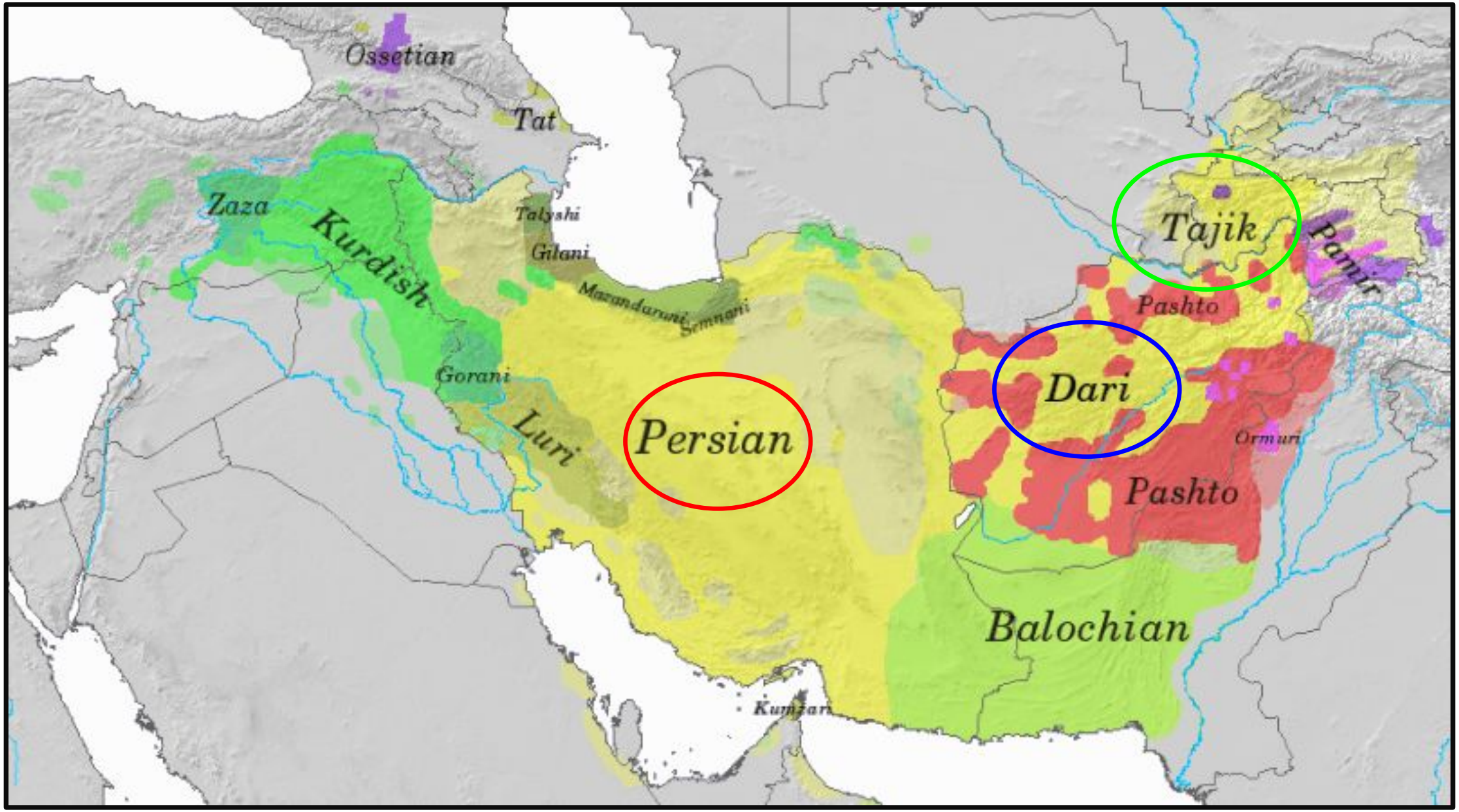
ParsTranslit: Truly Versatile Tajik-Farsi Transliteration

Rayyan Merchant[1] & Kevin Tang[23]
[1]Zoorna Institute
[2]Heinrich Heine University Düsseldorf
[3]University of Florida



Map of modern Iranian languages.

The Problem With Digraphia

Mutual intelligibility between standard varieties is extremely high in spoken form, but falls to zero in written form

Tajikistan has ~10 million Persian speakers who cannot access any written media from the rest of the Persian-speaking world (~100 million)

Tajik

Қаноъат Одил: Самарқанду Бухоро – Кримеи билқуввa аст

3 Октябр, 2014

Анҷумани Осиёи Миёна “омили тоҷикӣ” дар Узбакистонро чашми Исфандиёр ё нуктаи заъфи минтақаи Осиёи Миёна дар дасти “душманон” медонанд

Коршиносони Анҷумани Осиёи Миёна, созмони ғайридавлатие, ки дар Суэд сабти ном шуда, бар ин бовараанд, ки пас аз Укройн аз миёни кишварҳои собиқ шӯравӣ Узбакистон метавонад, саҳнаи даргириҳои баъдӣ шавад.

ТОҶИКИСТОН

Farsi

سمرقند و بخارا، 'کریمه بالقوه است'

۱۲ مهر ۱۳۹۳ - ۴ اکتبر ۲۰۱۴

انجمن آسیای میانه "عامل تاجیکی" در ازبکستان را چشم اسفندیار یا نقطه ضعف منطقه آسیای میانه در دست "دشمنان" می‌دانند

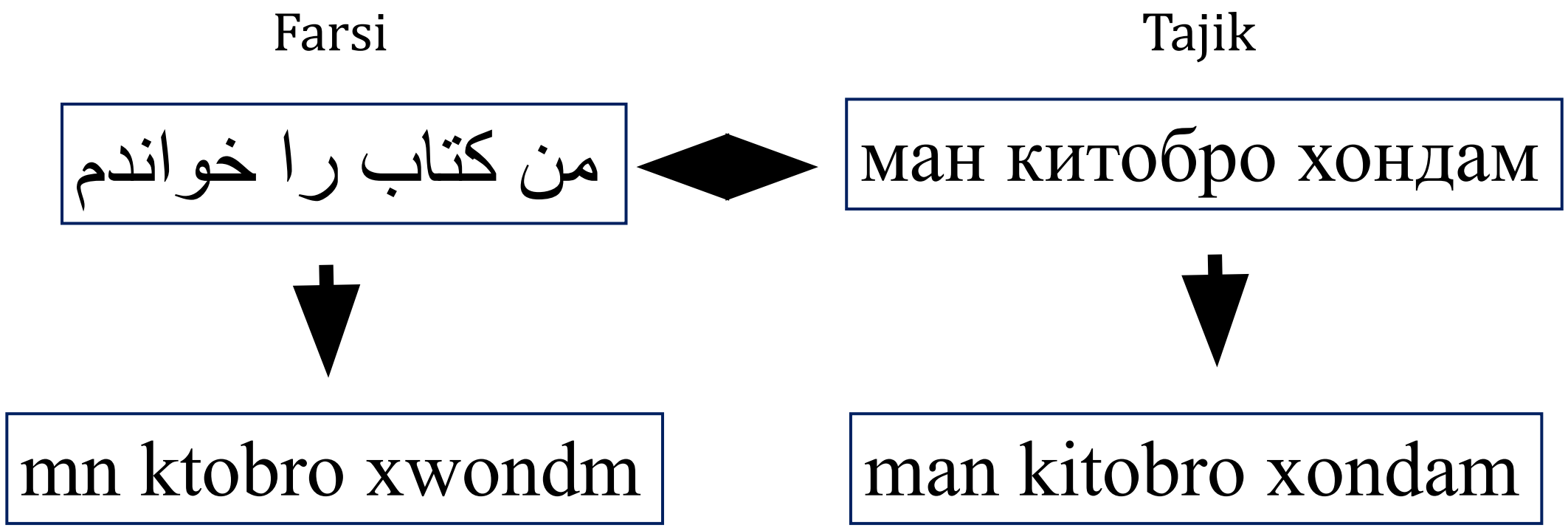
کارشناسان انجمن آسیای میانه، یک سازمان غیردولتی که در سوئد ثبت نام شده، بر این باورند که پس از اوکراین از میان کشورهای سابق شوروی ازبکستان می‌تواند صحنه درگیریهای بعدی شود.

BBC NEWS فارسی

Challenges in Transliteration

Perso-Arabic script is an (impure) abjad	•Vowels are often unwritten and ambiguous
Tajik-Cyrillic script is an Alphabet	•Phonetic

Example: “I read the book”



Challenges in Transliteration

Perso-Arabic script does not implement case, Cyrillic does

Several syllables and letters have one rendering in Tajik, but several in Farsi

Previous Work

3.1- FST (Megerdoomian and Parvaz, 2008)

3.2 - Math-based (Graschenko et al., 2008)

3.3 - Statistical (Davis, 2008)

3.4 - Transformer (Seredkina, 2024)

3.5 - Transformer (Merchant and Tang, 2024)

3.6 - Transformer (SadraeiJavaheri et al., 2024)

All have at least one issue!

- one direction only (Tajik-to-Farsi)
- very different datasets (Ex: poetry vs. news)
- lack of comprehensive and comparable metrics (Ex: only BLEU, only CER, etc.)

Datasets	Previous Systems						Ours
	3.1	3.2	3.3	3.4	3.5	3.6	5
Word List			✓				
Dictionary				✓	✓		✓
Shahnameh						✓	✓
Masnavi (either version)				✓	✓		✓
Assorted Poetry				✓	✓		✓
BBC News				✓	✓		✓
Blogs					✓		✓
ParaNames							✓
Direction: Tajik→Farsi	✓	✓	✓	✓	✓	✓	✓
Direction: Farsi→Tajik	✗	✗	✓	✗	✓	✓	✓
Digraphic Training Data	✗	✗	✓	✓	✓	✓	✓
Digraphic Test Data	✗	✗	✗	✓	✓	✓	✓
Model/Data Availability	✗	✗	✗	✓	✓	✓	✓

Table 1: Overview of previous systems and our system in terms of a breakdown of the training datasets (top), and the system properties (bottom) – transliteration direction, data type and model/data availability

ParsTranslit (2025)

Resolves all of these issues!

- 1st model to use ALL available datasets
- Directly comparable metrics across models, text genres, indiv. datasets

Also contributes 2 new datasets:

- Entity names (Sälevä and Lignos, 2024)
  - Never been evaluated before
- New version of the Masnavi
  - Replaced version by Seredkina (2024)
  - Misalignments manually checked by L2 speaker (Presenting Author)

Utilizes contextual markers as seen in Arabizi transliteration (Shazal et al., 2020)

Datasets	Previous Systems						Ours
	3.1	3.2	3.3	3.4	3.5	3.6	5
Word List			✓				
Dictionary				✓	✓		✓
Shahnameh						✓	✓
Masnavi (either version)				✓	✓		✓
Assorted Poetry				✓	✓		✓
BBC News				✓	✓		✓
Blogs					✓		✓
ParaNames							✓
Direction: Tajik→Farsi	✓	✓	✓	✓	✓	✓	✓
Direction: Farsi→Tajik	✗	✗	✓	✗	✓	✓	✓
Digraphic Training Data	✗	✗	✓	✓	✓	✓	✓
Digraphic Test Data	✗	✗	✗	✓	✓	✓	✓
Model/Data Availability	✗	✗	✗	✓	✓	✓	✓

Table 1: Overview of previous systems and our system in terms of a breakdown of the training datasets (top), and the system properties (bottom) – transliteration direction, data type and model/data availability

Domains & Datasets

- Poetry
  - Shahnameh (SadraeiJavaheri et al., 2024)
  - Masnavi (ParsTranslit)
  - Assorted Poetry (Seredkina, 2024)
- Prose
  - Blogs from ParsText (Merchant and Tang, 2024)
  - Assorted Prose (Seredkina, 2024)
    - includes BBC articles
- Proper Names
  - Subset of ParaNames (Sälevä and Lignos, 2024)
- Dictionary(Seredkina, 2024)

Domain / Datasets		# of Pairs	Farsi		Tajik	
			Avg. # Tokens	Avg. # Char.	Avg. # Tokens	Avg. # Char.
Poetry	Shahnameh	68,206	5.68	24.77	5.47	29.25
	Masnavi	39,011	6.12	25.81	5.75	31.43
	Assorted Poetry	156,576	6.31	27.20	6.01	33.10
Prose	Dr Blog	1,554	12.33	62.40	11.84	72.02
	Jamujam Blog	819	15.79	82.21	15.34	94.34
	Assorted Prose	23,992	19.56	102.45	18.62	118.95
Names	Places	11,179	1.48	9.46	1.45	10.46
	Organizations	4,185	1.39	9.05	1.31	9.72
	People	23,988	1.77	9.96	1.76	10.93
Dictionary		49,758	1.38	6.70	1.00	7.63

Table 2: Overview of the datasets used as training data. Text domain in bold are described in Section 6.2.

Model Training Setup

- Fairseq transformer model
- 80/10/10 split on each dataset
- Addition of contextual tokens
  - inspired by Arabizi transliteration (Shazal et. al, 2020)

шоҳи мо → @ ш о ҳ и $ _ @ м о $

شاه ما → $ ا م @ _ $ ه ا ش @

Hyperparameter	Value
Batch Size	128
Learning Rate	0.0007
Dropout	0.1
Layers	2
Heads	4
Embedding Dimension	256
Epochs	20
Learning Rate Scheduler	Reduce on Plateau
Optimizer	Adam

Evaluation Method

Compared ParsTranslit output with:

Model	**Model Description**
(Merchant and Tang, 2024) **Label: DP**	Our previous bidirectional model grapheme-to-phoneme model
(Seredkina, 2024)) **Label: TG2FA**	Unidirectional (Tajik-to-Farsi)
SadraeiJavaheri et al. (2024)	Bidirectional model only trained on Shahnameh (poetry) dataset (metrics as reported in their paper)

Model Results (Farsi → Tajik)

Subset	chrF		chrF++		CER		Normalized CER		Acc%		Acc% (No WS)	
	DP	ParsTranslit	DP	ParsTranslit	DP	ParsTranslit	DP	ParsTranslit	DP	ParsTranslit	DP	ParsTranslit
Poetry	65.24	**92.52**	58.14	**90.36**	4.30	**0.90**	0.13	**0.03**	1.84	**53.74**	3.22	**58.52**
Prose	64.40	**86.56**	56.44	**83.14**	13.47	**7.55**	0.13	**0.06**	4.88	**18.25**	5.40	**20.91**
Dictionary	66.27	**89.33**	60.16	**87.55**	1.56	**0.36**	0.20	**0.05**	24.70	**76.91**	36.34	**77.00**
Names	27.63	**71.78**	21.81	**66.83**	4.29	**1.50**	0.40	**0.15**	4.05	**40.82**	4.14	**41.44**
Overall	63.83	**90.34**	56.57	**87.91**	4.57	**1.36**	0.17	**0.05**	5.28	**52.98**	7.81	**56.57**

Table 3: Farsi→Tajik model performance across different data subsets and across metrics. The score in bold indicates the best performing model for each metric and subset.

- ParsTranslit outperforms across all metrics and genres

Model Results (Tajik → Farsi)

Subset	chrF			chrF++			CER			Normalized CER			Acc%			Acc% (No WS)		
	DP	TG2FA	ParsTranslit	DP	TG2FA	ParsTranslit	DP	TG2FA	ParsTranslit	DP	TG2FA	ParsTranslit	DP	TG2FA	ParsTranslit	DP	TG2FA	ParsTranslit
Poetry	84.73	93.21	**95.63**	78.69	91.19	**93.98**	2.16	0.86	**0.60**	0.08	0.03	**0.02**	15.64	55.75	**66.35**	33.22	65.56	**76.50**
Prose	86.25	**95.59**	91.64	80.54	**94.16**	89.85	5.97	**2.09**	6.39	0.07	**0.03**	0.04	11.11	**41.96**	38.25	19.17	**51.21**	47.11
Dict.	**94.14**	85.14	91.22	**86.86**	78.82	85.73	0.54	0.55	**0.38**	0.08	0.08	**0.06**	59.13	62.75	**72.52**	**86.82**	74.41	82.69
Names	37.41	44.57	**80.08**	31.05	38.28	**75.82**	3.40	2.75	**1.02**	0.34	0.29	**0.11**	7.43	13.17	**53.68**	9.48	14.15	**54.48**
Overall	83.72	92.02	**94.00**	77.80	90.08	**92.28**	2.34	1.1	**1.01**	0.11	0.06	**0.04**	20.18	51.30	**63.90**	36.82	60.41	**72.99**

Table 4: Tajik→Farsi model performance across different data subsets and across metrics. The score in bold indicates the best performing model for each metric and subset.

- ParsTranslit outperforms but other models show train-test contamination

Model Results (Shahnameh)

Model	CER	
	Farsi to Tajik	Tajik to Farsi
ParsTranslit	**0.93**	0.90
SadraeiJavaheri et al. (2024)	1.05	**0.88**
DP	1.56	2.55
TG2FA	N/A	1.09

Table 5: Model performance on Shahnameh dataset in both directions. The score in bold indicates the best performing model for each direction.

- ParsTranslit exhibits similar Tajik-to-Farsi performance despite more data, suggesting gains from addtl. data plateau

Key Trends

- Entity names are the biggest challenge for all models
  - particularly for those that have never seen this data
- Prose is harder than Poetry
- Tajik-to-Farsi is easier than Farsi-to-Tajik
- Removing whitespace improves Acc% in both directions

Main Takeaways

- Tajik-Farsi transliteration is not a uniform task
  - Some genres of text are easier to transliterate than others
- ParsTranslit's performance proves is currently the **only** available model able to handle a wide variety of texts in both directions
- Readily able to produce understandable, if not perfect, transliterations in all contexts

Example Model Output (Farsi to Tajik)

Script	Example Sentence (errors highlighted)
Farsi (input)	زوندرمان معتقد است که این از نخستین موارد کاربرد واژه تاجیک به معنای ایرانی مسلمان است
Tajik (output)	заванdармони муътақид аст ки ин аз нахустин мавориди корбурди вожаи тоҷик ба маънои эронӣ мусалмон аст
Tajik (reference)	зундерманн муътақид аст ки ин аз нахустин мавориди корбурди вожаи тоҷик ба маънои эронии мусалмон аст

Example Model Output (Tajik to Farsi)

Script	Example Sentence (errors highlighted)
Tajik (input)	зундерманн муътақид аст ки ин аз нахустин мавориди корбурди вожаи тоҷик ба маънои эронии мусалмон аст
Farsi (output)	زندرمن معتقد است که این از نخستین موارد کاربرد واژه تاجیک به معنای ایرانی مسلمان است
Farsi (reference)	زوندرمان معتقد است که این از نخستین موارد کاربرد واژه تاجیک به معنای ایرانی مسلمان است

Example Model Output (Farsi to Tajik)

Script	Example Sentence (errors highlighted)
Farsi (input)	ما در حوزه‌ی ادبی و فرهنگی بسیار کوشیده‌ایم
Tajik (reference) (alternative transliteration highlighted)	мо дар ҳавзаи адабиву фарҳангӣ бисёр кӯшидаем
Tajik (output) (alternative transliteration highlighted)	мо дар ҳавзаи адабӣ ва фарҳангӣ бисёр кӯшидаем

Future Work

- Major:
  - Develop task-specific testing methods
    - Stop penalizing valid alternative transliterations
      - Ex: ‘و’ can become ‘y’, ‘ва’, ‘ву’, all meaning ‘and’
    - Measure model’s ability to detect Ezafe
  - Investigate model errors to find deeper patterns, if any
  - Make tool available! Receive user feedback
- Minor:
  - Refine entity names dataset

Github

https://github.com/merchantrayyan/ParsTranslit

References

- Karine Megerdoomian and Dan Parvaz. 2008. Low-Density Language Bootstrapping: The Case of Tajiki Persian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 3293–3298, Marrakech, Morocco. European Language Resources Association (ELRA).
- L. A. Graschenko, Z. J. Usmanov, and A. Yu. Fomin. 2009. Tajik-Persian converter of graphic writing systems. Intellectual product registered by the National Patent Information Center of Ministry of Economic Development and Trade of the Republic of Tajikistan
- Chris Irwin Davis. 2012. Tajik-Farsi Persian Transliteration Using Statistical Machine Translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3988–3995, Istanbul, Turkey. European Language Resources Association (ELRA).
- Rayyan Merchant, Akhilesh Kakolu Ramarao, and Kevin Tang. 2025. Connecting the Persian-speaking world through transliteration. Preprint, arXiv:2502.20047.
- MohammadAli SadraeiJavaheri, Ehsaneddin Asgari, and Hamid Reza Rabiee. 2024. Transformers for bridging Persian dialects: Transliteration model for Tajiki and Iranian scripts. In Proceedings of the 2024 Joint International Conference on Computational Lin- guistics, Language Resources and Evaluation (LREC- COLING 2024), pages 16770–16775, Torino, Italia. ELRA and ICCL.
- Jonne Sälevä and Constantine Lignos. 2024. ParaNames 1.0: Creating an entity name corpus for 400+ languages using Wikidata. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC- COLING 2024), pages 12599–12610, Torino, Italia. ELRA and ICCL.
- Ali Shazal, Aiza Usman, and Nizar Habash. 2020. A unified model for Arabizi detection and transliteration using sequence-to-sequence models. In Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 167–177, Barcelona, Spain (Online). Association for Computational Linguistics.