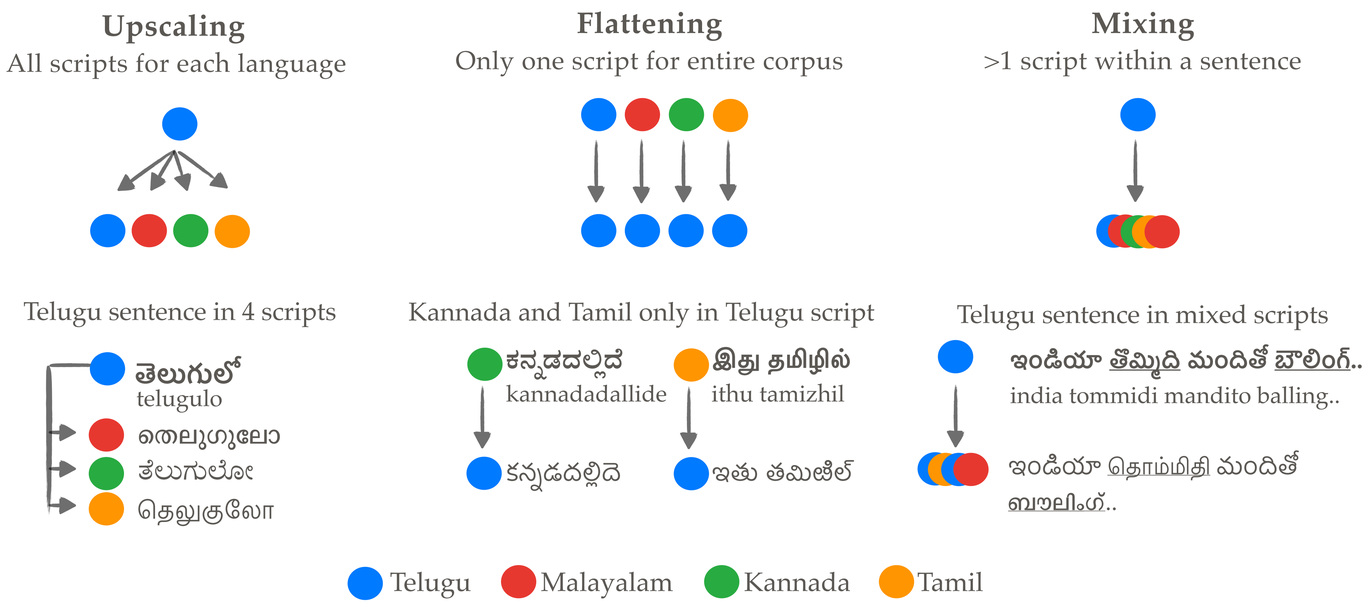
We can identify languages no matter what **script** they are written in

Script-Agnostic Language Identification

Milind Agarwal, Joshua Otten, Antonios Anastasopoulos

NLP GEORGE MASON



Script-Agnosticism can be modeled under different strategies, with intra-sentence script-mixing offering the most robustness

	FLORES200	GLOT	UDHR	MCS350	AVERAGE
Test Set Size	4048	3934	285	15000	5817
BASELINE (FLORES200)	95.26	82.41	79.00	45.34	**75.50**
FASTTEXT (WIKI)	100.00	99.96	100.00	71.75	**92.93**
UPSCALE (16K)	96.35	81.67	77.54	44.79	**75.09**
FLATTEN (TELU)	91.28	43.18	44.56	33.95	53.24
NOISE (ALL)	95.41	80.19	76.14	43.41	**73.79**

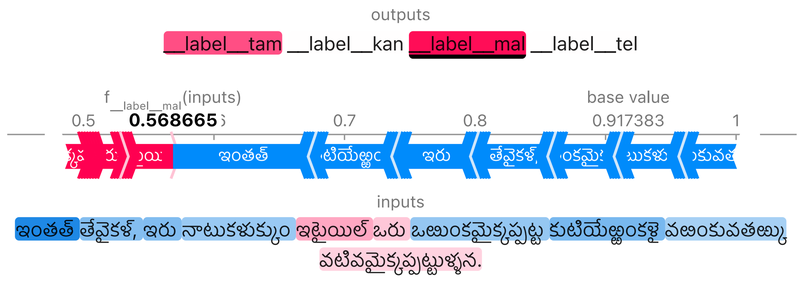
Multi-script training setups lead to similar performance as FLORES200 baseline i.e. no performance loss due to more scripts

Some More Results

* No particular script is best suited for projection, despite differing letter counts

Script:	Tamil		Kannada		Malayalam		Telugu	
	Baseline	Flatten	Baseline	Flatten	Baseline	Flatten	Baseline	Flatten
TAMIL	94.37	80.43	-	80.63	-	80.93	-	80.73
KANNADA	-	91.60	92.59	92.19	-	91.60	-	91.70
MALAYALAM	69.27	99.31	88.93	98.32	100.00	98.42	88.93	98.91
TELUGU	-	93.68	-	93.77	-	93.08	94.07	93.77
AVERAGE	40.91	91.25	45.28	91.23	25.00	91.01	45.75	91.28

* Interpretability analysis reveals Mal-Tam lexical overlap affects langID



* Negligible differences between parallel and non-parallel data

Model	FLORES (transliterated)	FLORES (clean)	GLOT	UDHR	MCS350	average
BASELINE (FLORES200)	39.26	95.26	82.41	79.00	45.34	68.25
4-WAY PARALLEL	**96.32**	96.35	81.67	77.54	44.79	**79.33**
NON-PARALLEL	**94.39**	94.37	84.61	83.86	51.76	**81.80**

Email ⟶ magarwa@gmu.edu, jotten4@gmu.edu

Poster Design Inspiration ⟶ Anjishnu Mukherjee and Chahat Raj

George Mason NLP