DatasetA	En Ko	Ko En	Model information	DatasetA		Evaluation Matrix
Transformer	13.1(100K) (validset-15.9)	20.6(85K) (validset-22.4)	Number of model parameters: 93326081	30,264 test.tok.en	30,264 test.tok.ko	**Bleu**
	13.3(avg-last 8 ckpts)	**21.0(avg-last 8 ckpts)**	Model Size 1.1G	544,759 train.tok.en	544,759 train.tok.ko	
TransformerRelative	12.8 (95K)	21.0 (90K)	Number of model parameters: 93389057	30, 264 valid.tok.en	30,264 valid.tok.ko	
	13.0(avg-last 8 ckpts)	**21.2(avg-last 8 ckpts)**	Model Size 1.1G	605,287 total	605,287 total	

DatasetA+B	En Ko	Ko En	Model information	DatasetB	
Transformer	**14.4 (validset-best score15.9)**		Number of model parameters: 93326081	**Tatoeba** dataset - after cleaning	
	(avg-last 8 ckpts - 235K steps)		Model Size 1.1G	4,672,270 tatoeba.ko	4,672,270 tatoeba.en
				233,613 test.ko	233,613 test.en
				4,205,044 train.ko	4,205,044 train.en
				233,613 valid.ko	233,613 valid.en

DatasetA+B remove HTTP stuff	En Ko	Ko En	Model information	Dataset A + B	
Transformer	**14.4 (validset-best score16.03)**	25.0 (validset-best score 25.26)	Number of model parameters: 93326081	263877 test.ko	263877 test.en
	(avg-last 8 ckpts - 255K steps)	(avg-last 8 ckpts - 230K steps)	Model Size 1.1G	4749802 train.ko	4749802 train.en
				263877 valid.ko	263877 valid.en
TransformerRelative	14.3 (validset-besr score 15.8852	25.2 (validaset-best score 25.48)	Number of model parameters: 93389057	5277556 total	5277556 total
	(avg-last 8 ckpts - 205K steps)	(avg-last 8 ckpts - 320K)	Model Size 1.1G		
TransformerBig	**19.4** (validset-best sore 20.5)	**27.5** (validset-best score 27.55)	Number of model parameters: 274700545		
	avg-last 8 ckpts - 220K steps	avg-last 8 ckpts - 160K steps	Model Size 3.1G		

additional dataset		**mono dataset**		
5000 inference_5000_samples_ko_to_en.csv		1. download - WMT En-De dataset		
3297 inference_first_5000.xlsx		4562102 wmt.en 4520346 wmtclean.en. (remove HTTP stuff)		
14543 lyrics_save100.csv		https://github.com/QuoQA-NLP/T5_Translation		
49590 lyrics_save500.csv		HuggingFace EnKo model - inference to make synthetic dataset		
Evaluation Matrix		2. download Korean dataset		
Bleu				

Dataset D	En Ko	Ko En	**Dataset D**
	280000 (best bleu so far: 38.425109)	200000 (best bleu so far: 40.677246)	(base) ➜ data wc -l *.ko
	AVG-300K steps	AVG-210K steps	557778 test.tok.ko
Transformer	**"score": 34.4,**	**"score": 40.2,**	10041133 train.tok.ko
			557778 valid.tok.ko