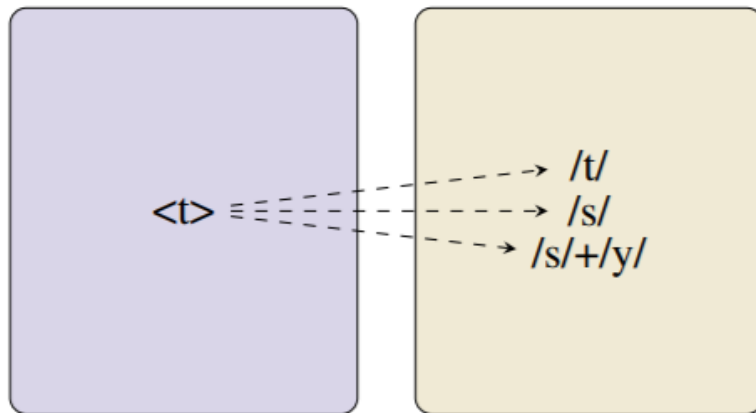


# Comparing Writing Systems with G2P and P2G

University of the Basque Country  
Language Analysis and Processing  
Computational Morphology  
Julen Etxaniz

# 1. Introduction

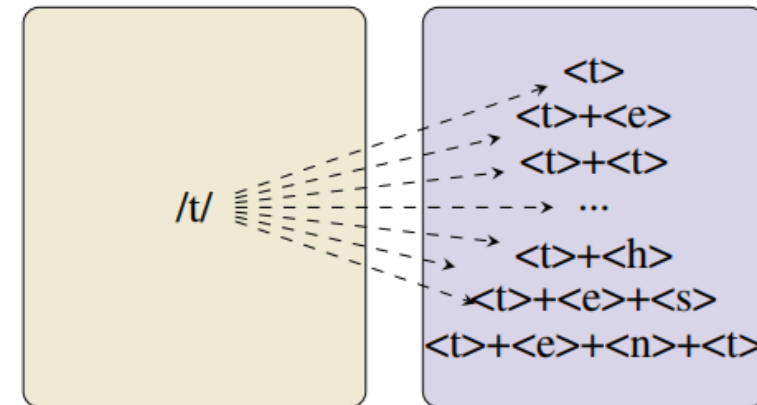
- G2P (Reading)



ablatif    a b l a t i f

abolition a b o l i s j ɔ̃

- P2G (Writing)



acquitter a k i t e

auguste o g y s t

athlète a t l ɛ t

actes a k t

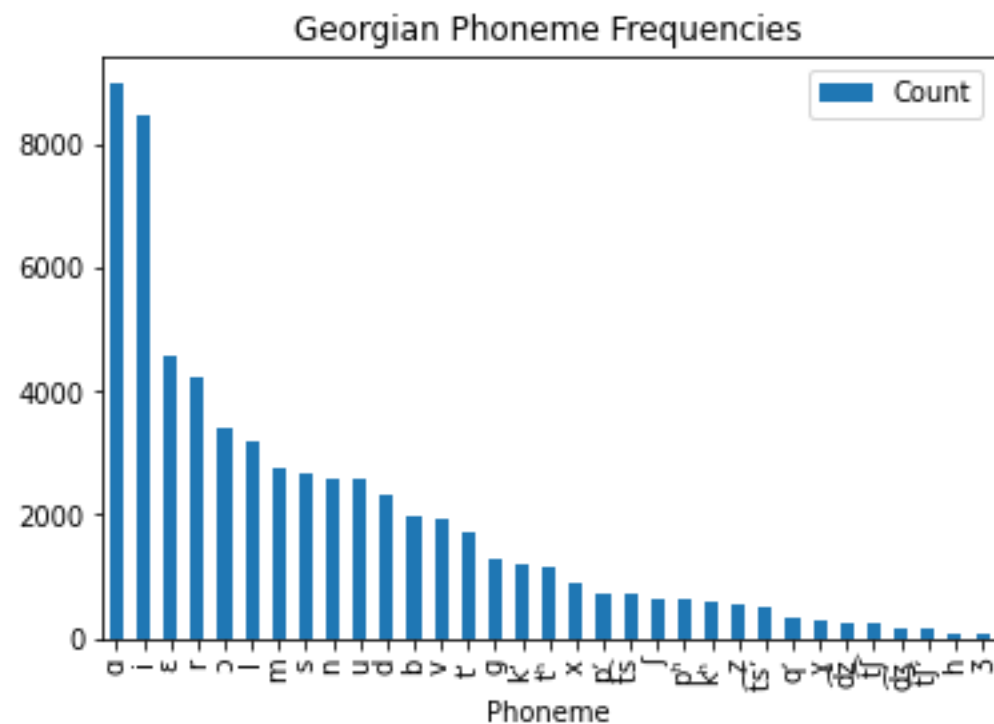
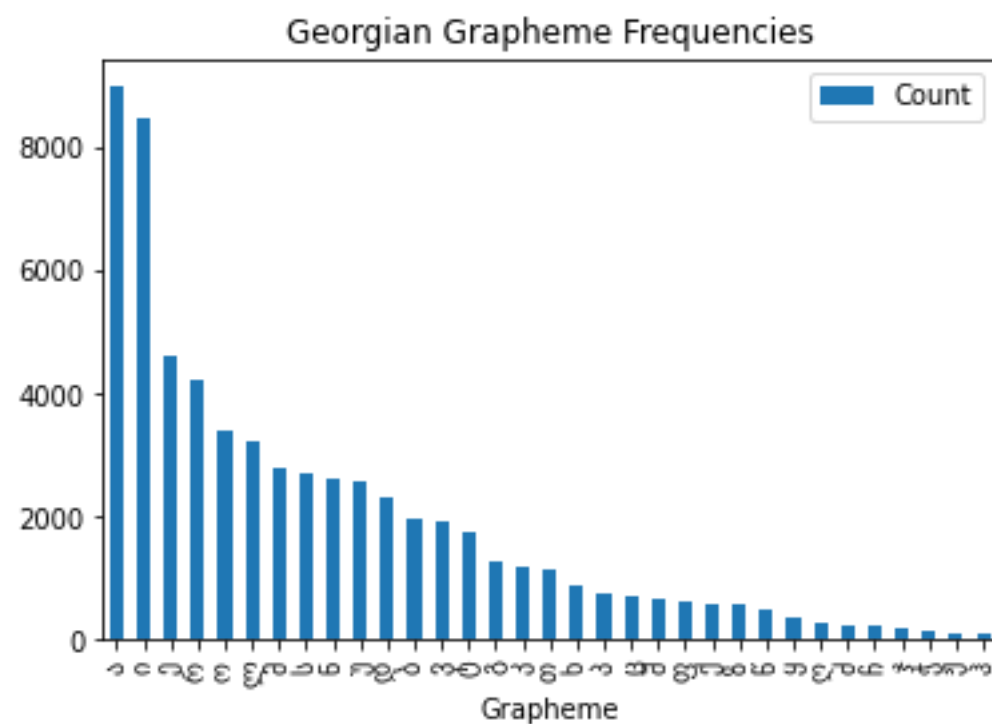
assistant a s i s t

Data

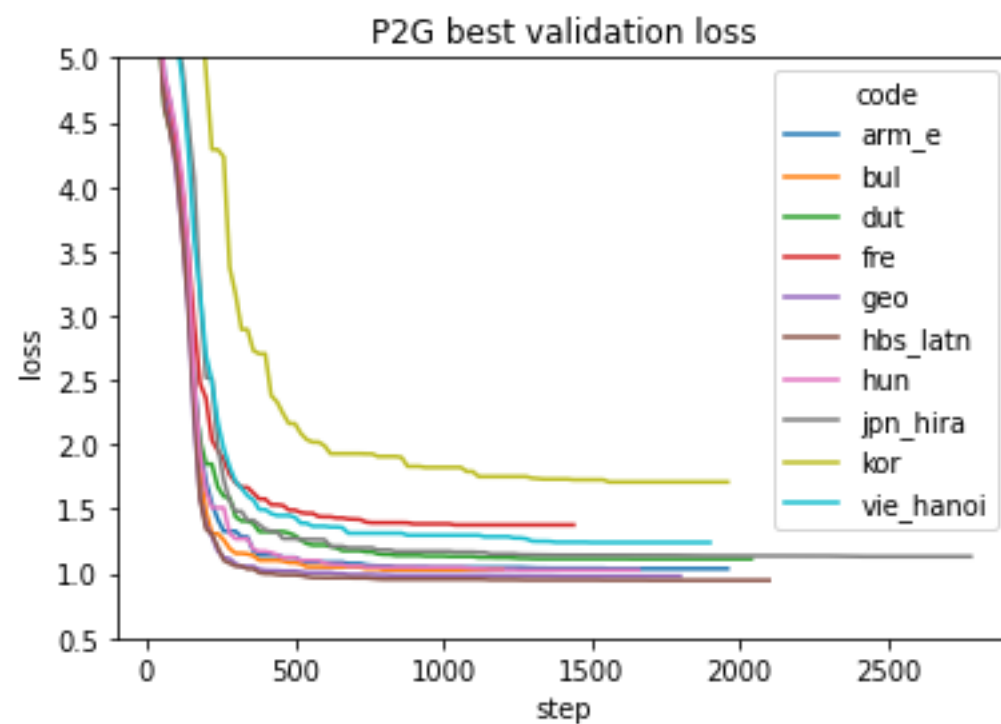
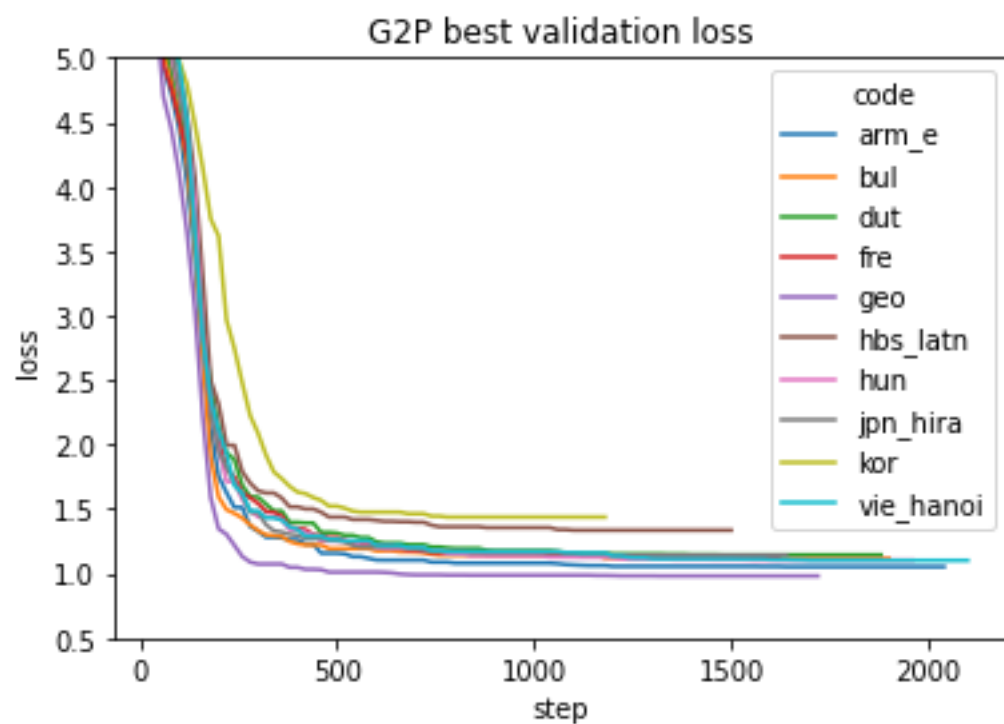
# Preprocessing

Language	Code	Graphemes	Phonemes	Average Graphemes	Average Phonemes
Armenian (Eastern)	arm_e	38	41	7.09	7
Bulgarian	bul	29	45	8.44	8.64
Dutch	dut	30	45	7.71	6.99
French	fre	37	37	7.52	5.75
Georgian	geo	33	33	7.74	7.74
Serbo-Croatian (Latin)	hbs_latn	27	61	7.47	7.36
Hungarian	hun	34	61	7.65	7.18
Japanese (Hiragana)	jpn_hira	76	64	4.21	6.56
Korean	kor	559	60	2.58	6.54
Vietnamese (Hanoi)	vie_hanoi	89	49	5.81	7.51
Average	average	95.2	49.6	6.62	7.13

# Preprocessing



# Training



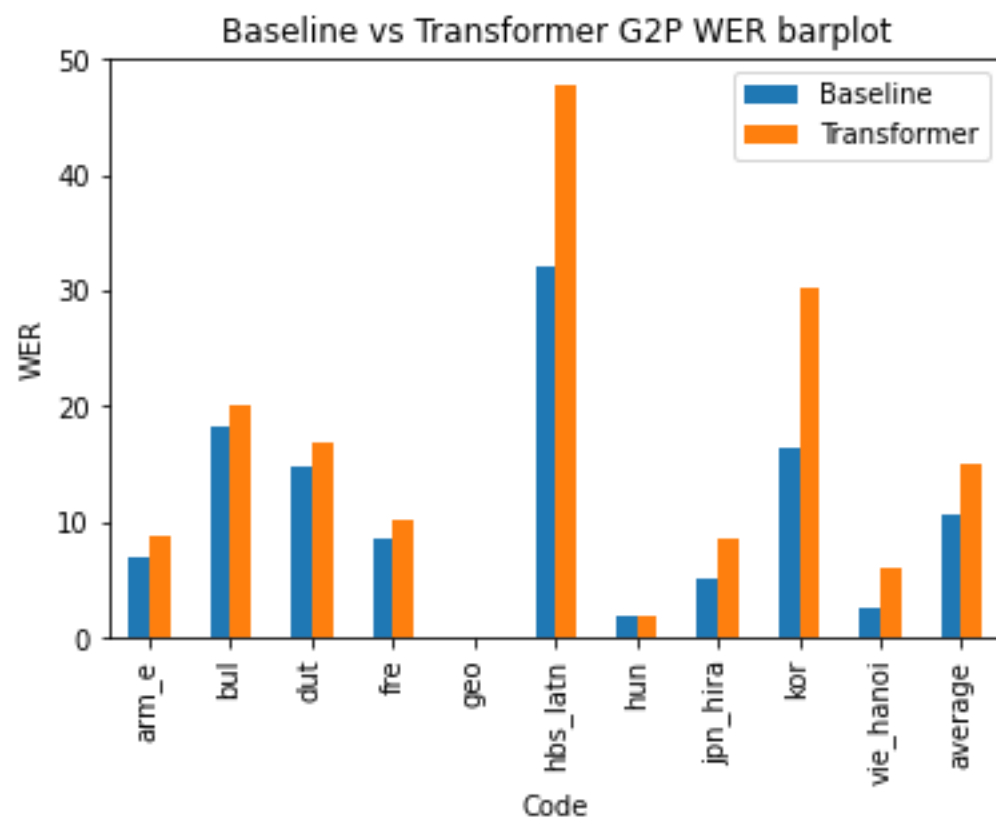
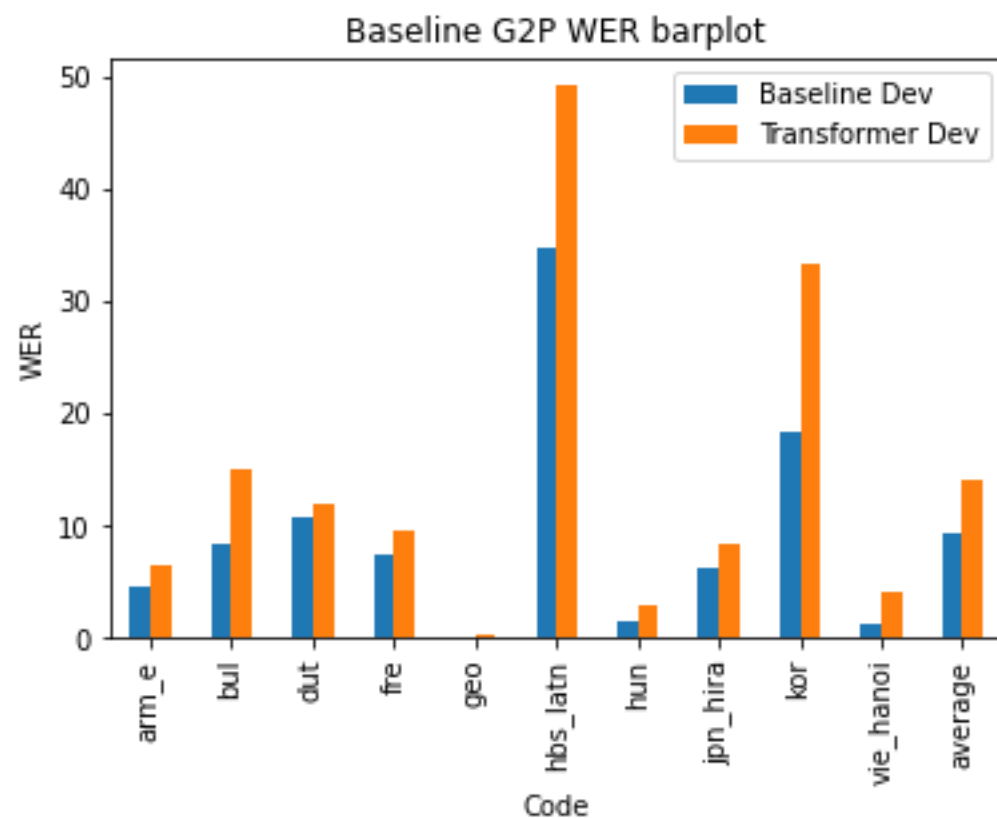
# Testing

# Results – G2P Baseline

Code	Baseline Dev	Transformer Dev	Baseline Test	Transformer Test
arm_e	4.50	6.5	7.00	8.90
bul	8.30	14.9	18.30	20.10
dut	10.80	12.0	14.70	16.90
fre	7.40	9.5	8.50	10.20
geo	0.00	0.3	0.00	0.10
hbs_latn	34.70	49.2	32.10	47.70
hun	1.50	2.8	1.80	1.90
jpn_hira	6.20	8.4	5.20	8.50
kor	18.40	33.4	16.30	30.20
vie_hanoi	1.30	4.0	2.50	6.00
average	9.31	14.1	10.64	15.05



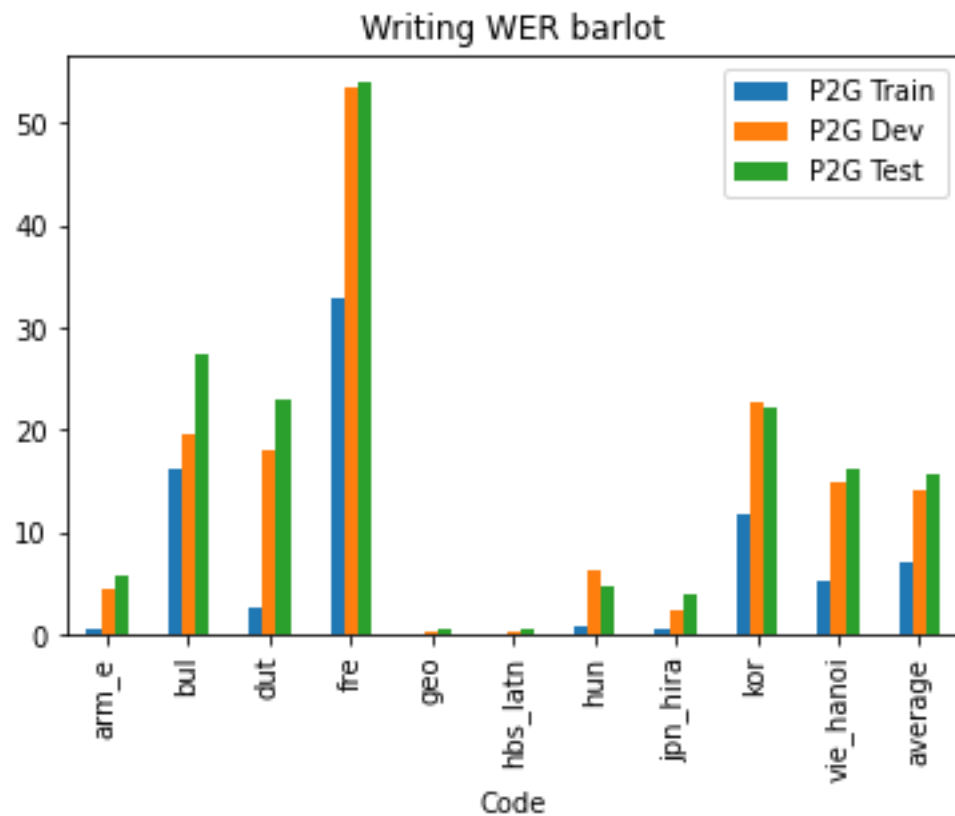
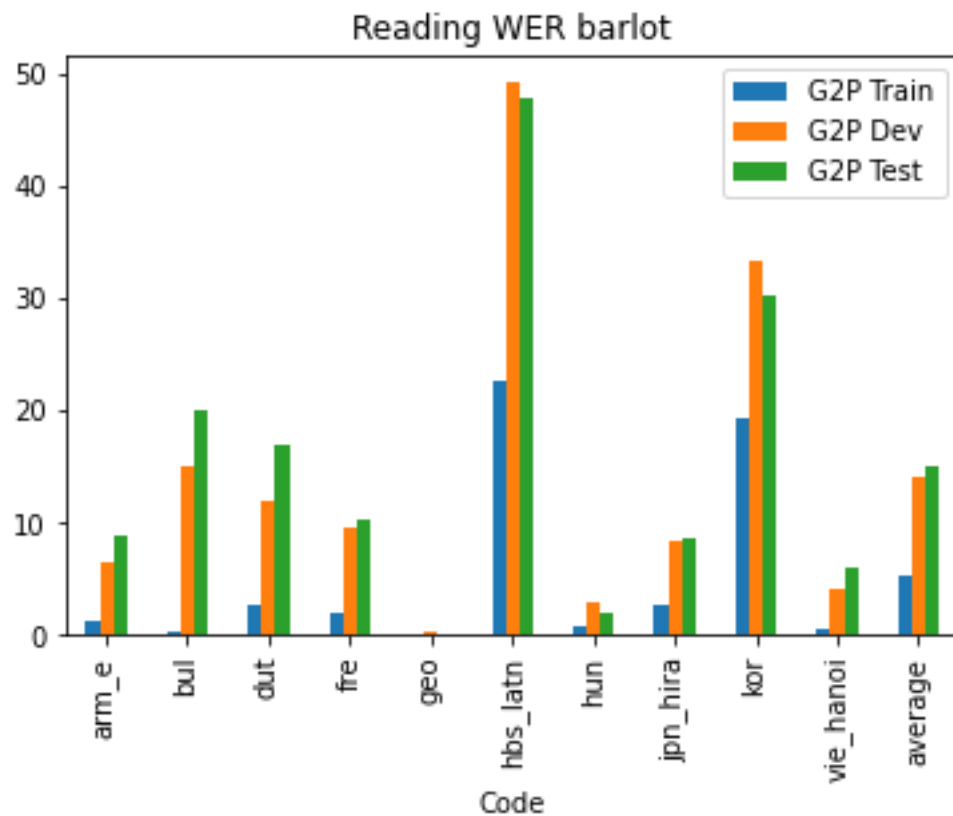
# Results – G2P Baseline



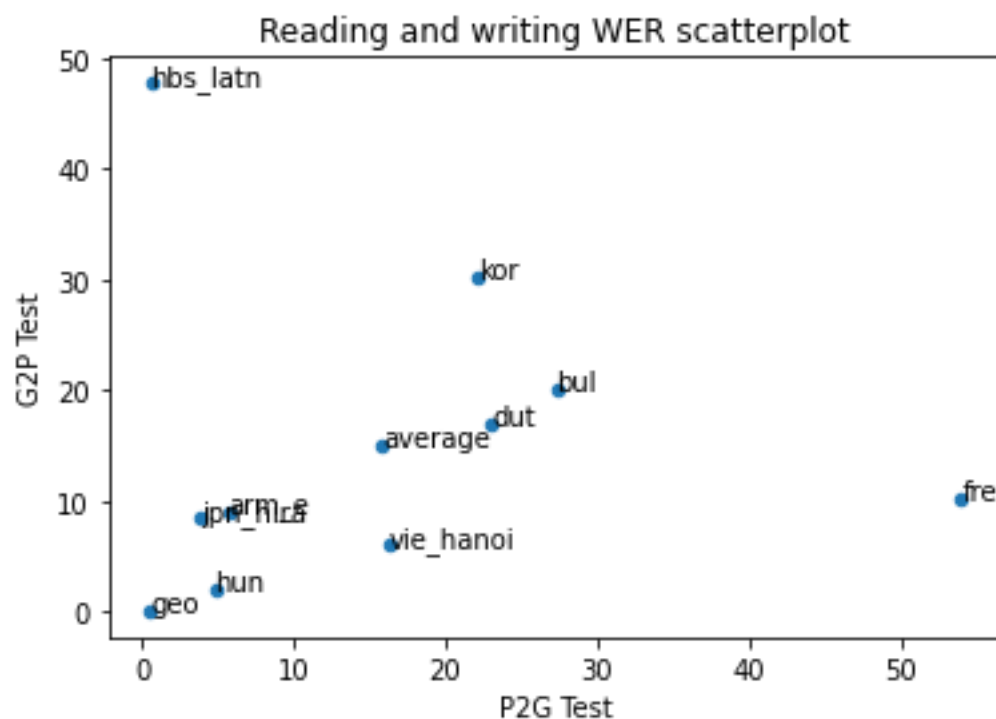
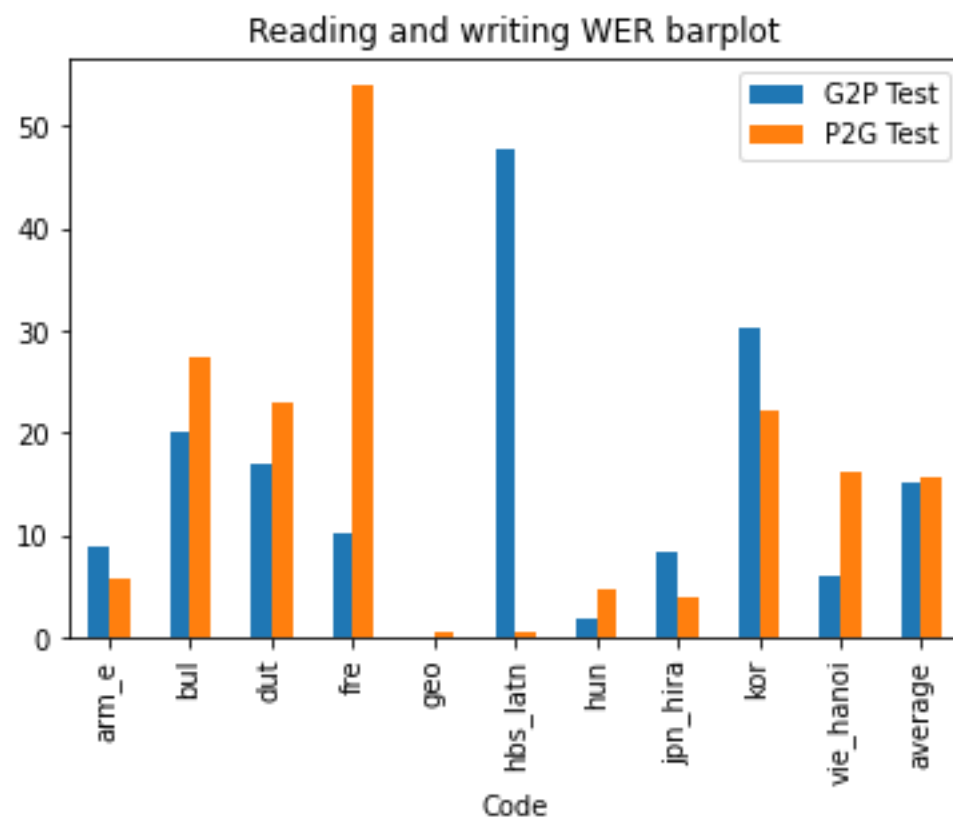
# Results – G2P and P2G

Code	G2P Train	G2P Dev	G2P Test	P2G Train	P2G Dev	P2G Test
arm_e	1.16	6.5	8.90	0.57	4.40	5.70
bul	0.34	14.9	20.10	16.25	19.70	27.30
dut	2.59	12.0	16.90	2.67	18.00	23.00
fre	2.04	9.5	10.20	32.96	53.50	54.00
geo	0.06	0.3	0.10	0.12	0.30	0.50
hbs_latn	22.56	49.2	47.70	0.09	0.20	0.60
hun	0.64	2.8	1.90	0.84	6.20	4.80
jpn_hira	2.74	8.4	8.50	0.59	2.40	3.90
kor	19.27	33.4	30.20	11.84	22.80	22.10
vie_hanoi	0.44	4.0	6.00	5.35	15.00	16.30
average	5.18	14.1	15.05	7.13	14.25	15.82

# Results – G2P and P2G



# Results – G2P and P2G



# Conclusions