UPDATE FOR V8.0 DATASET

How to Start a Language on Mozilla Common Voice?

A case study for under-resourced Turkish Language

Bülent Özden

Computer Engineer (MSc), Harikalar Kutusu

Mozilla Common Voice Turkish Language Representative (2021-2022)

FOSDEM'22, 5th February 2022

CorporaCreator default splits

Number of Client IDs				
Ver	VAL	TRAIN	DEV	TEST
1	190	4	21	165
2	323	13	46	264
3	323	13	76	265
4	434	34	76	324
5.1	608	77	128	389
6.1	631	77	128	401
7.0	851	13	99	640
8.0	1,159	13	110	1,009

Gender - Female / Male Ratio				
Ver	VAL	TRAIN	DEV	TEST
1	9.89%	0.00%	40.69%	9.13%
2	16.24%	22.02%	5.87%	8.51%
3	16.04%	16.63%	12.50%	8.12%
4	17.28%	16.35%	7.40%	5.51%
5.1	8.57%	16.96%	2.61%	9.61%
6.1	8.57%	16.81%	2.82%	10.33%
7.0	8.99%	0.00%	6.87%	10.29%
8.0	47.11%	103.63%	58.29%	17.06%

Dataset diversity

Gender & Age – v7.0

Count	GENDER				
AGE	male	female	other	(blank)	TOTAL
teens	3.33%	0.21%	0.06%	0.00%	**3.60%**
twenties	40.66%	3.81%	0.15%	0.15%	**44.77%**
thirties	17.12%	1.67%	0.00%	0.00%	**18.79%**
fourties	1.87%	0.05%	0.00%	0.00%	**1.92%**
fifties	4.61%	0.36%	0.00%	0.00%	**4.98%**
sixties	0.31%	0.00%	0.00%	0.00%	**0.31%**
(blank)	0.08%	0.01%	0.00%	25.54%	**25.63%**
TOTAL	**67.99%**	**6.11%**	**0.21%**	**25.69%**	**100.00%**

Gender & Age – v8.0

Count	GENDER				
AGE	male	female	other	(blank)	TOTAL
teens	1.71%	0.10%	0.03%	0.03%	**1.86%**
twenties	25.63%	2.01%	0.15%	0.07%	**27.86%**
thirties	9.43%	1.15%	0.05%	0.00%	**10.63%**
fourties	1.79%	2.69%	0.00%	0.00%	**4.48%**
fifties	5.97%	4.02%	0.00%	0.00%	**9.99%**
sixties	1.25%	7.15%	0.00%	0.00%	**8.40%**
seventies	0.00%	4.20%	0.00%	0.00%	**4.20%**
eighties	0.00%	0.28%	0.00%	0.00%	**0.28%**
(blank)	0.04%	0.00%	0.00%	32.26%	**32.30%**
TOTAL	**45.82%**	**21.58%**	**0.24%**	**32.36%**	**100.00%**

Validated

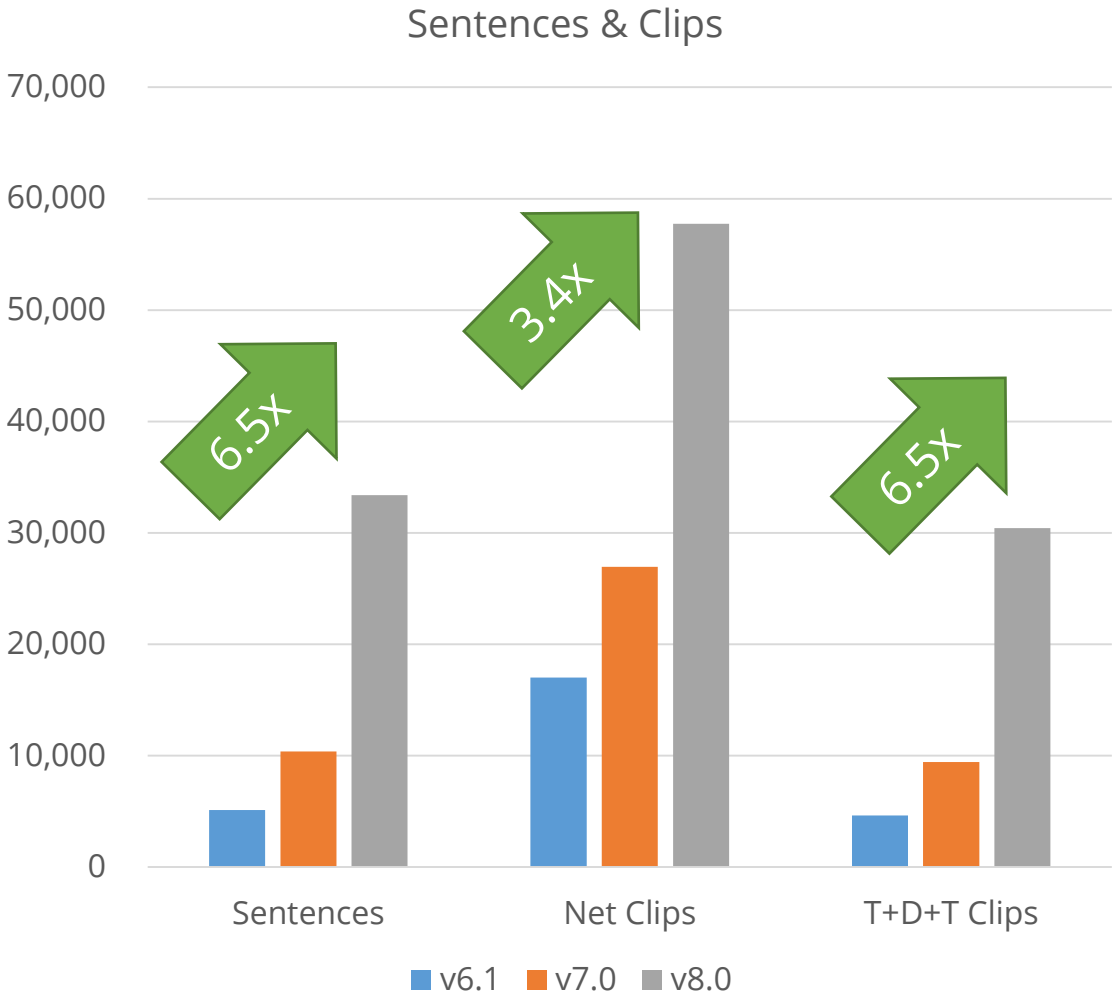
Recordings per sentence

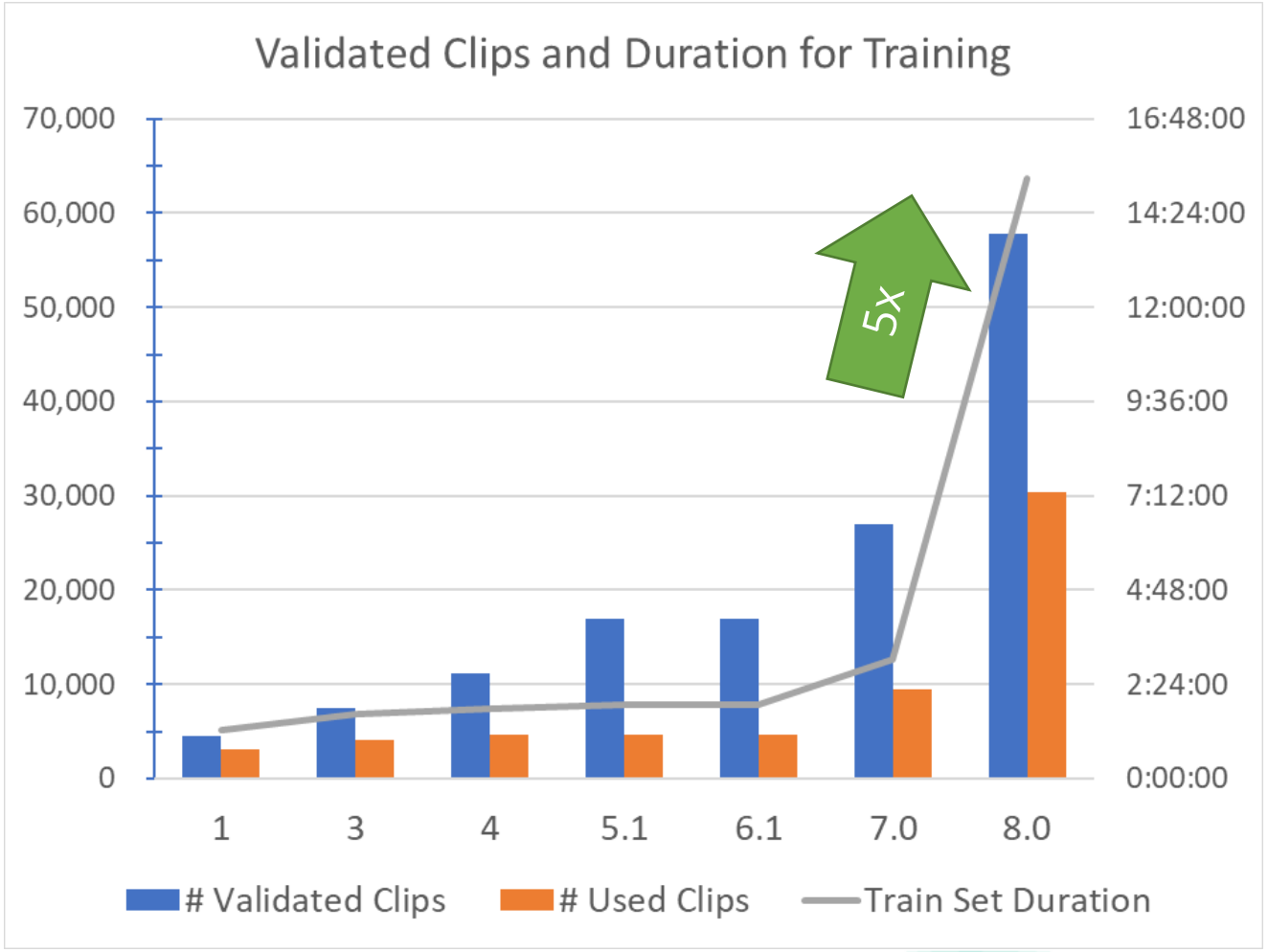
Recorded by	v6.1	v7.0	v8.0
1	43	5,240	17,723
2	302	68	10,472
3	1,548	299	397
4	3,145	1,384	1,384
5	66	2,931	2,930
6	5	452	453
7	1	2	2
28	3	2	1
29		1	2
30	1	1	1
32	3	1	1
33	5	4	1
34		3	5
35	1	1	1
36			1
37	1	1	
38			1
TOTAL	**5,124**	**10,390**	**33,375**
6mo Increase		2.03	3.21
Usable with -s 1	27.4%	35.1%	52.8%

Recordings per person

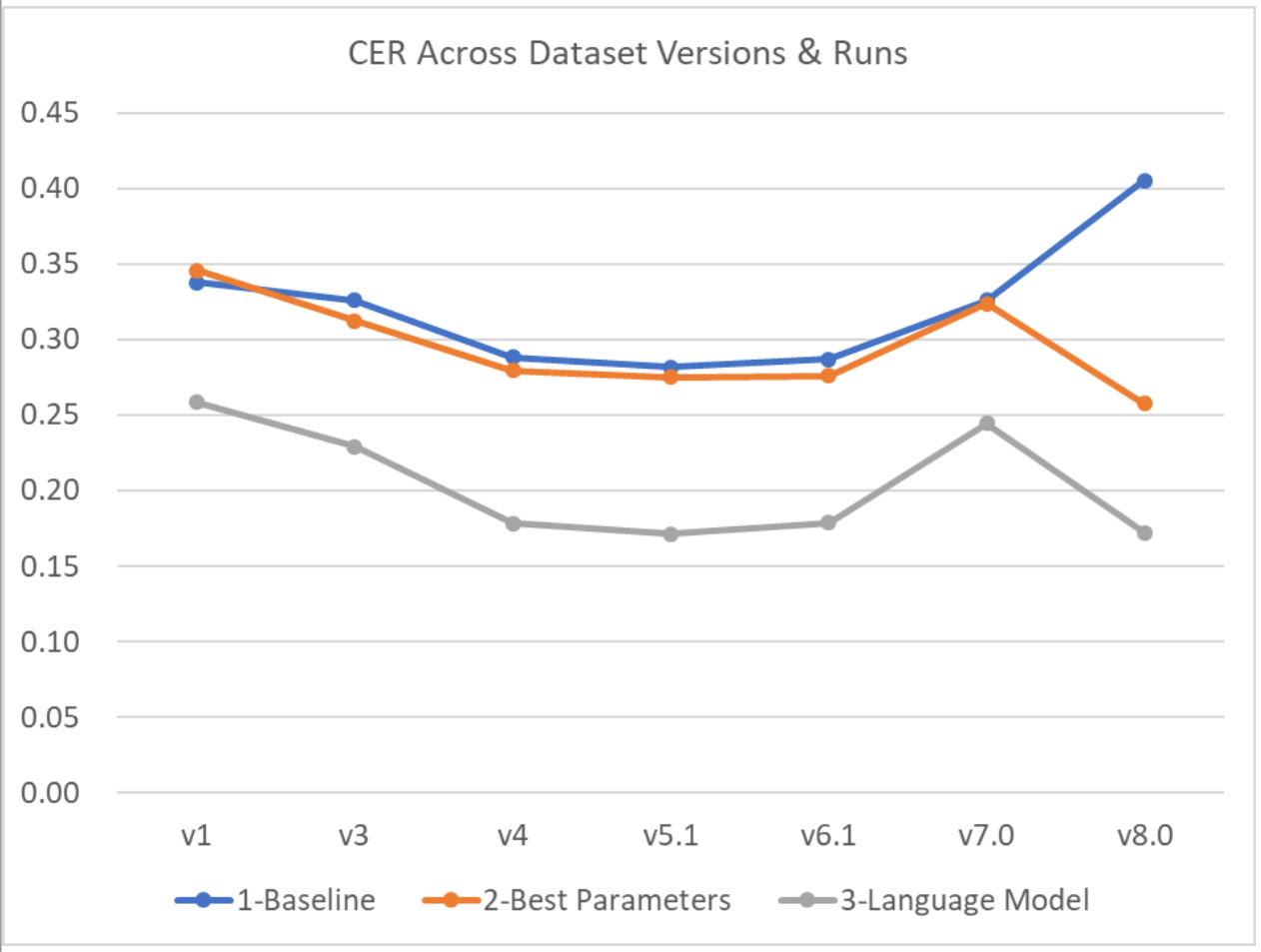
Rec/Person	v6.1	v7.0	v8.0	INC	%
0-5	322	428	556	128	11.0%
5-10	94	132	177	45	3.9%
10-20	79	102	149	47	4.1%
20-30	30	43	61	18	1.6%
30-40	24	27	35	8	0.7%
40-50	13	20	26	6	0.5%
50-100	34	46	72	26	2.2%
100-200	18	26	34	8	0.7%
200-300	5	7	16	9	0.8%
300-400	5	6	8	2	0.2%
400-500	2	2	3	1	0.1%
500-600	3	7	8	1	0.1%
600-700	0	0	1	1	0.1%
700-800	0	1	1	0	0.0%
800-900	0	1	1	0	0.0%
900-1000	0	0	2	2	0.2%
1000-2000	2	2	4	2	0.2%
2000-3000	0	1	0	-1	-0.1%
3000-4000	0	0	2	2	0.2%
4000-5000	0	0	2	2	0.2%
5000-99999	0	0	1	1	0.1%
People	**631**	**851**	**1,159**		
Increase		**220**	**308**		

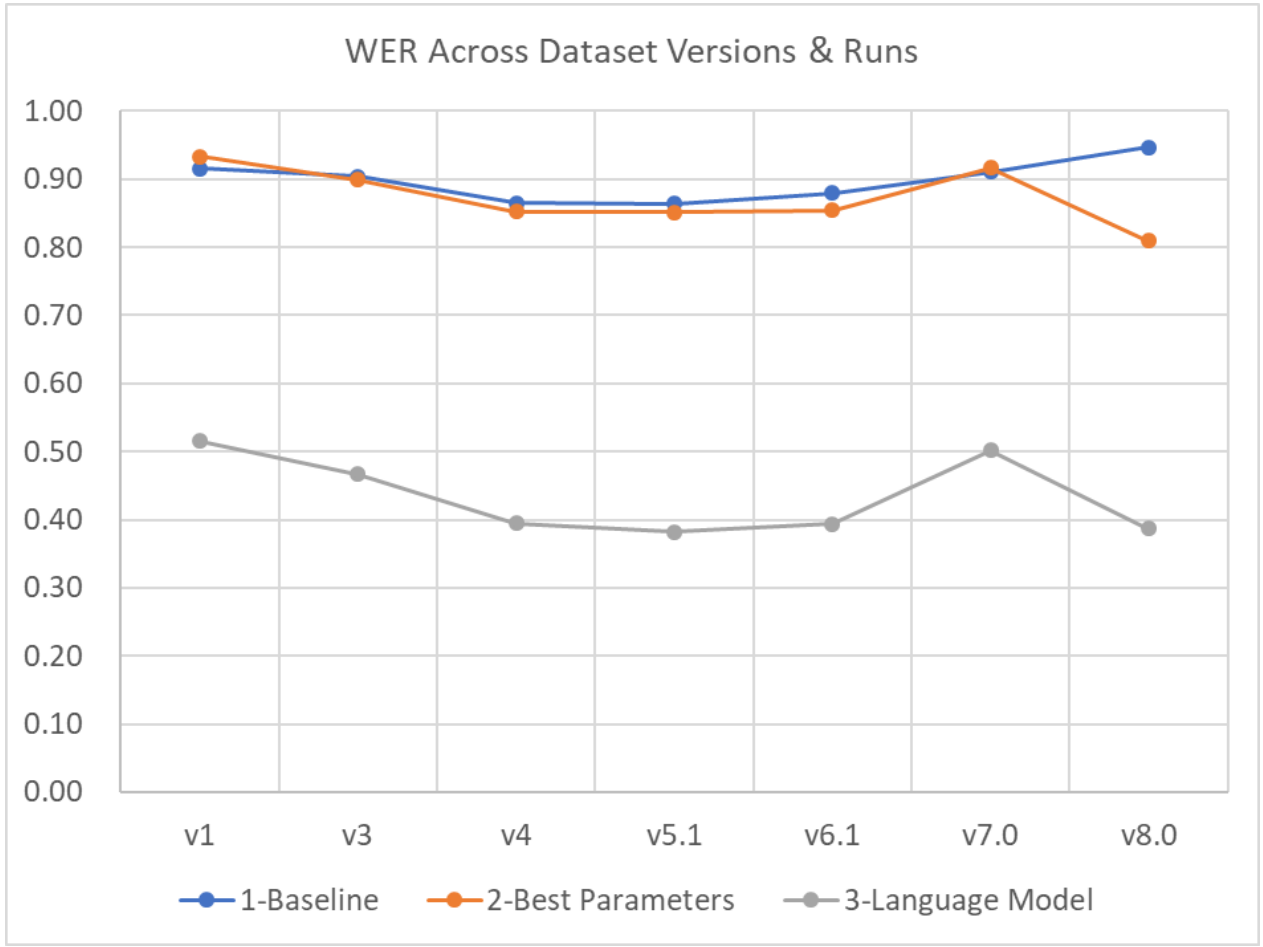
Increase in data

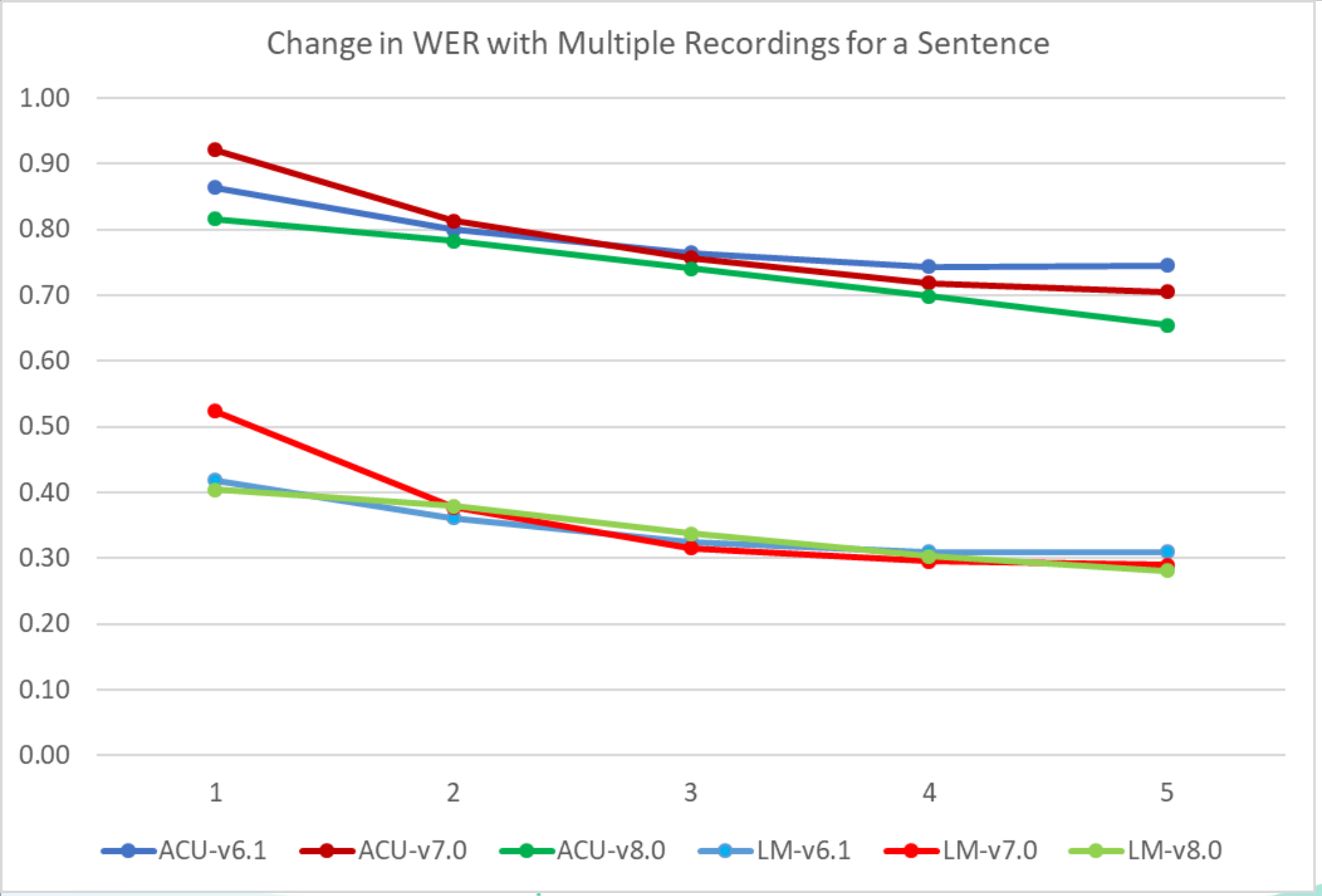




Character & Word Error Rates







Best Values

For –s 5

Train: 36:17:40

AM Only

- CER: 18.69%
- WER: 65.43%

AM+LM

- CER: 11.71%
- WER: 28.11%

Better results!

-15% for AM only

-10% for AM+AM