How to Start a Language on Mozilla Common Voice?

A case study for under-resourced Turkish Language

Bülent Özden

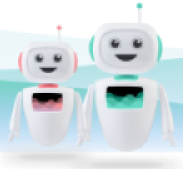
Computer Engineer (MSc), Harikalar Kutusu

Mozilla Common Voice Turkish Language Representative (2021-2022)

FOSDEM'22, 5th February 2022

Common Voice

moz://a Turkish Volunteers



We Teach Turkish to Technology

Goal

«A Crash Course for Dataset Caretakers»

Lessons learned…

Take care of your dataset!

...or it will get biased!

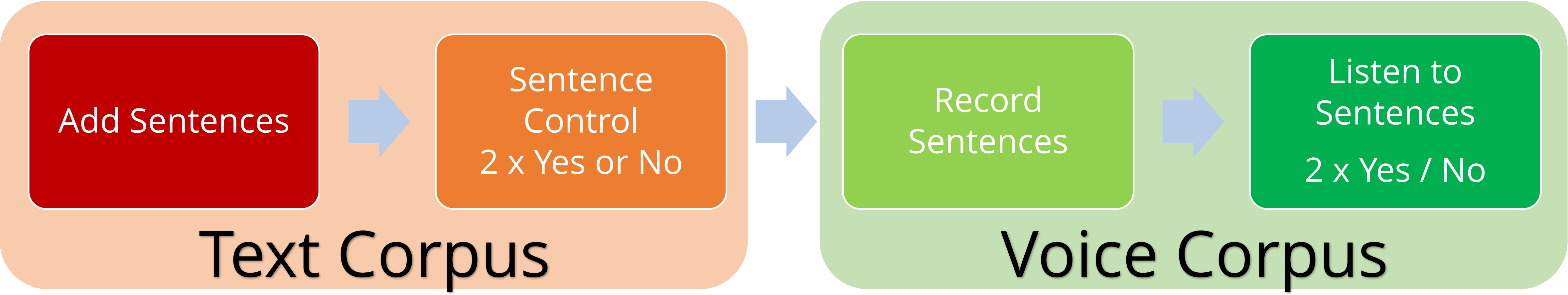
Bias => BAD

Introduction

General view

Important to know

CV processes are like a conveyor belt!



Add Sentences

Sentence Control
2 x Yes or No

Text Corpus

Record Sentences

Listen to Sentences
2 x Yes / No

Voice Corpus

If one fails, production rate/quality drops!

You need: Monitoring, Dedication, Time & a Crowd

It is a marathon !

3000 | ▶ 25000 BOZDEN

Important to know!

- Alphabet!
- Conversational text / speech
- CV is general purpose, not specific to application/method/model
- It is not a «clean» dataset => Already somewhat «augmented»

- Text Corpus: Max 14 words (default), keep it ~100 chars max
- Voice Corpus: Min 1.5 sec, max 14 sec recording

Bias

- Deep Learning is a black-box
    - It is mostly data driven
    - Resulting AI is least interpretable

- Bias => Near equal distribution / diversity!
    - Equal Gender & Age
    - Different Accents
    - Large # Speakers
    - Large # Sentences => Vocabulary

- «Contamination» or «unbalance» cannot be reversed easily
    - Postprocess
    - Re-balance

How much is needed?

This is for full training.

Depends on application

Fortunately we have **transfer learning**!

100 h ?

300 h ?

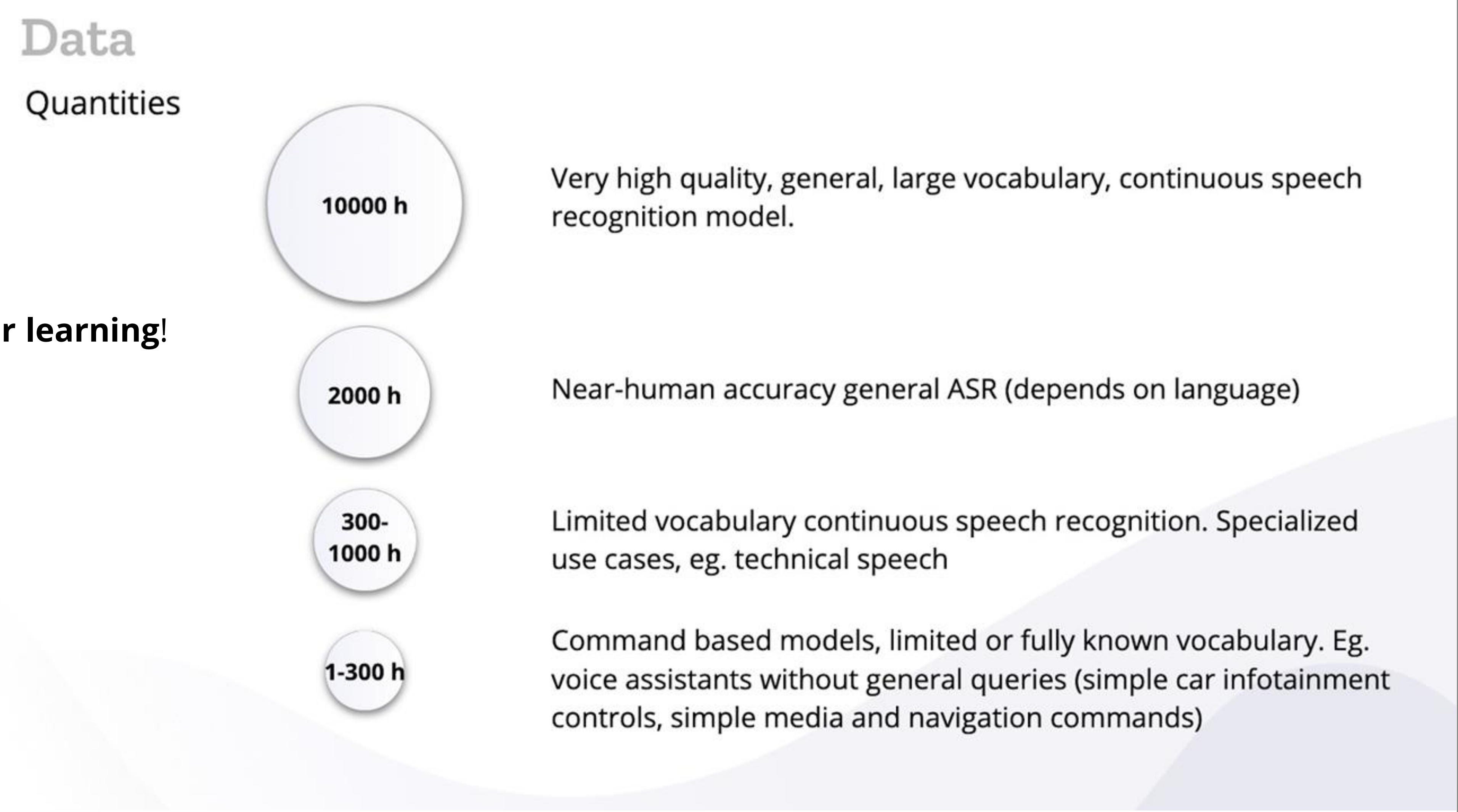
1000 h ?

2000 h ?

CER

WER

Real world performance



Data

Quantities

10000 h — Very high quality, general, large vocabulary, continuous speech recognition model.

2000 h — Near-human accuracy general ASR (depends on language)

300-1000 h — Limited vocabulary continuous speech recognition. Specialized use cases, eg. technical speech

1-300 h — Command based models, limited or fully known vocabulary. Eg. voice assistants without general queries (simple car infotainment controls, simple media and navigation commands)

Starting a language

- Requirements
- Resources on CV

Requirements to start a language

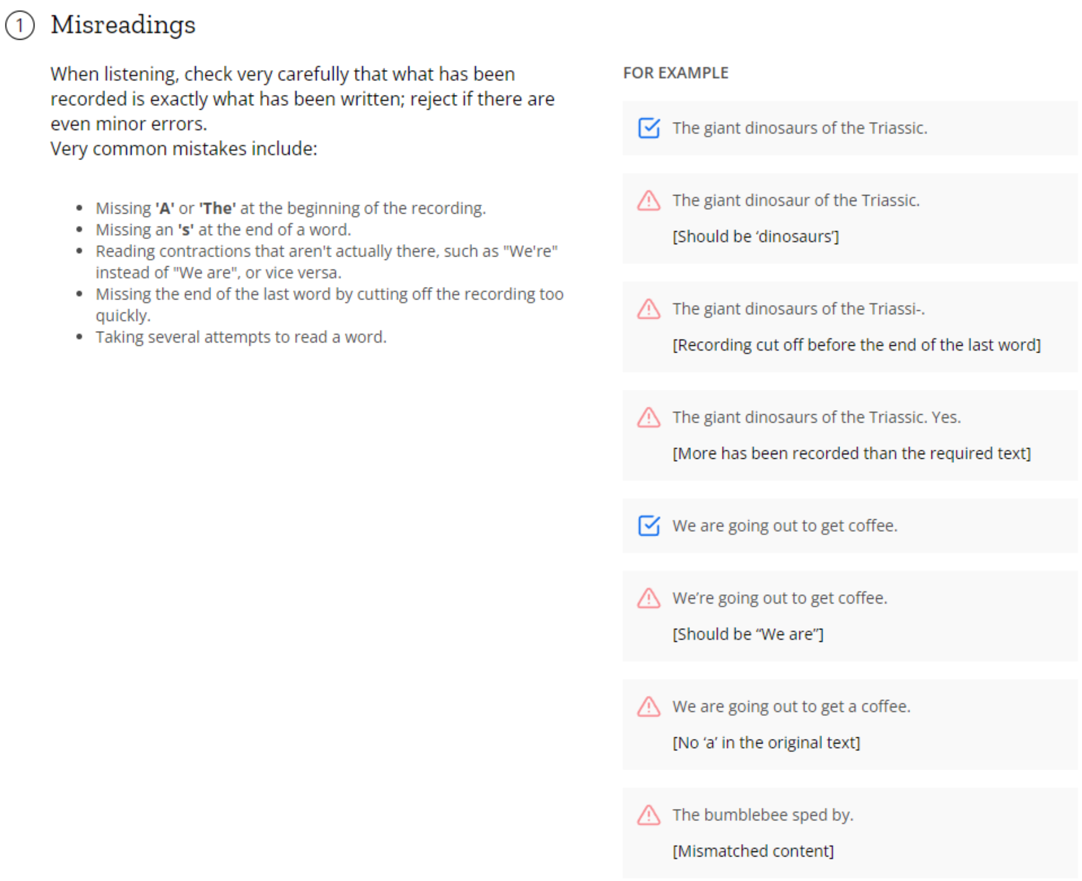
- Add 5000+ CC-0 / Public Domain sentences
- Translate 75% of the UI

- Request the language in Discourse/github

<u>For Turkish it was easier</u>

- We already had a working language/dataset
- UI was at 83% => translated the remaining
  - Localization of examples!!!
- Sentence Collector translations

Criteria page localization



① Misreadings

When listening, check very carefully that what has been recorded is exactly what has been written; reject if there are even minor errors.
Very common mistakes include:

- Missing **'A'** or **'The'** at the beginning of the recording.
- Missing an **'s'** at the end of a word.
- Reading contractions that aren't actually there, such as "We're" instead of "We are", or vice versa.
- Missing the end of the last word by cutting off the recording too quickly.
- Taking several attempts to read a word.

FOR EXAMPLE

☑ The giant dinosaurs of the Triassic.

⚠ The giant dinosaur of the Triassic.
[Should be 'dinosaurs']

⚠ The giant dinosaurs of the Triassi-.
[Recording cut off before the end of the last word]

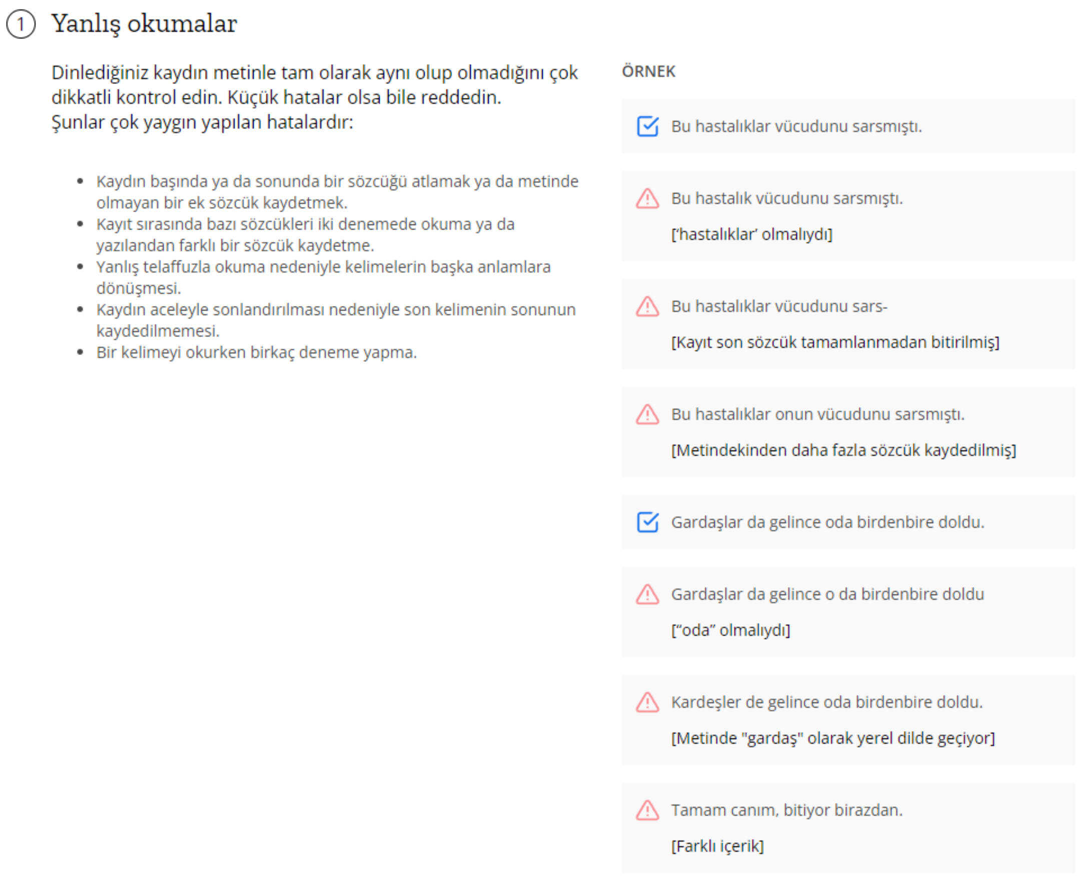
⚠ The giant dinosaurs of the Triassic. Yes.
[More has been recorded than the required text]

☑ We are going out to get coffee.

⚠ We're going out to get coffee.
[Should be "We are"]

⚠ We are going out to get a coffee.
[No 'a' in the original text]

⚠ The bumblebee sped by.
[Mismatched content]



① Yanlış okumalar

Dinlediğiniz kaydın metinle tam olarak aynı olup olmadığını çok dikkatli kontrol edin. Küçük hatalar olsa bile reddedin.
Şunlar çok yaygın yapılan hatalardır:

- Kaydın başında ya da sonunda bir sözcüğü atlamak ya da metinde olmayan bir ek sözcük kaydetmek.
- Kayıt sırasında bazı sözcükleri iki denemede okuma ya da yazılandan farklı bir sözcük kaydetme.
- Yanlış telaffuzla okuma nedeniyle kelimelerin başka anlamlara dönüşmesi.
- Kaydın aceleyle sonlandırılması nedeniyle son kelimenin sonunun kaydedilmemesi.
- Bir kelimeyi okurken birkaç deneme yapma.

ÖRNEK

☑ Bu hastalıklar vücudunu sarsmıştı.

⚠ Bu hastalık vücudunu sarsmıştı.
['hastalıklar' olmalıydı]

⚠ Bu hastalıklar vücudunu sars-
[Kayıt son sözcük tamamlanmadan bitirilmiş]

⚠ Bu hastalıklar onun vücudunu sarsmıştı.
[Metindekinden daha fazla sözcük kaydedilmiş]

☑ Gardaşlar da gelince oda birdenbire doldu.

⚠ Gardaşlar da gelince o da birdenbire doldu
["oda" olmalıydı]

⚠ Kardeşler de gelince oda birdenbire doldu.
[Metinde "gardaş" olarak yerel dilde geçiyor]

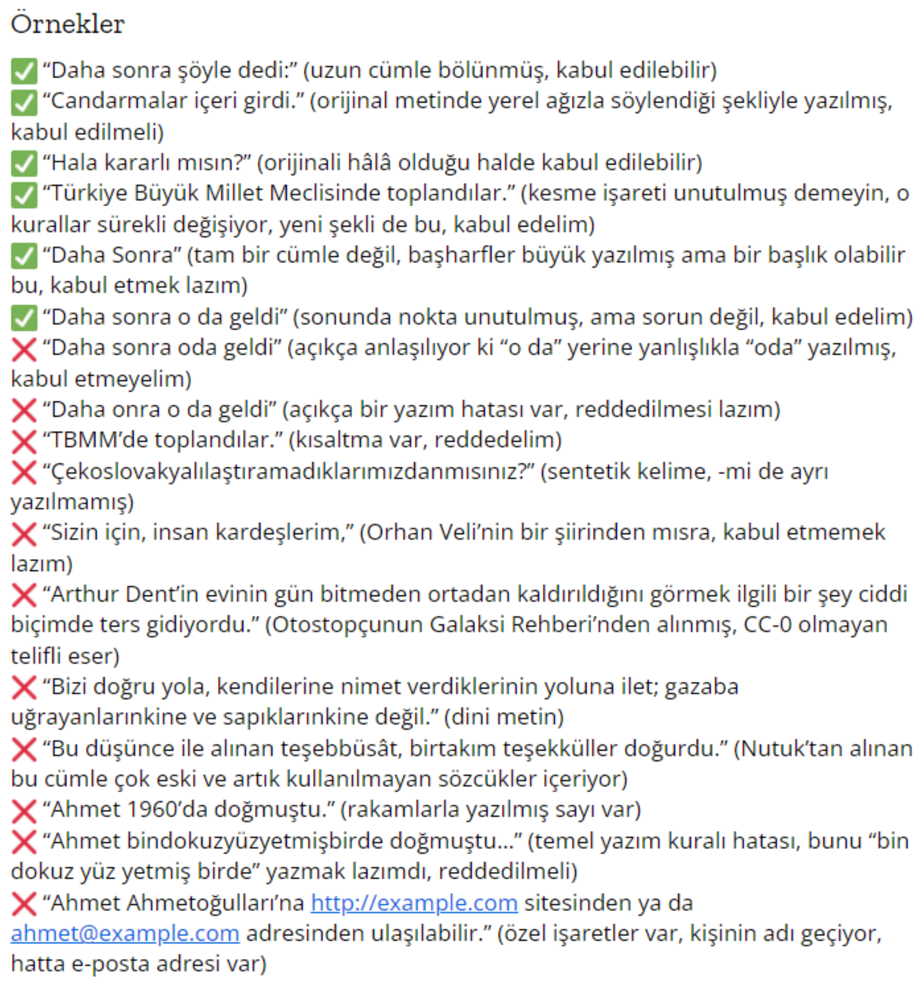
⚠ Tamam canım, bitiyor birazdan.
[Farklı içerik]

Common Voice

moz://a Turkish Volunteers

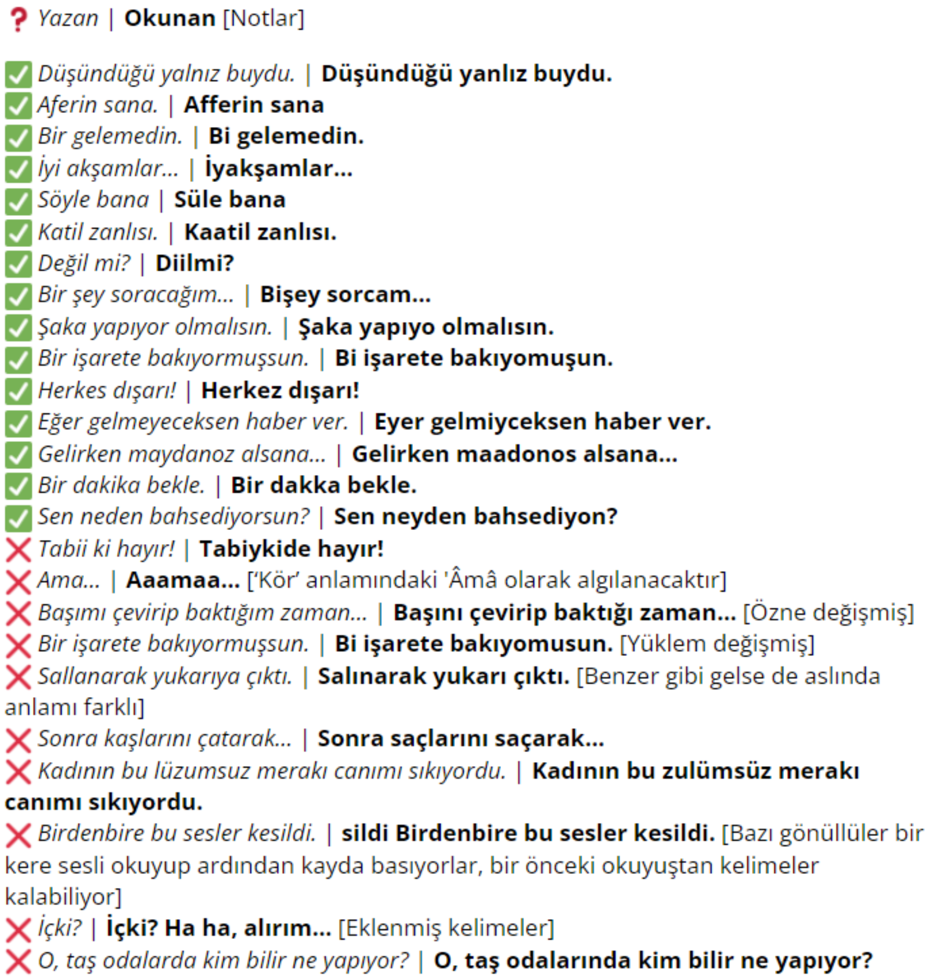
We Teach Turkish to Technology

Prepare more examples for the public



Örnekler

✅ "Daha sonra şöyle dedi:" (uzun cümle bölünmüş, kabul edilebilir)
✅ "Candarmalar içeri girdi." (orijinal metinde yerel ağızla söylendiği şekliyle yazılmış, kabul edilmeli)
✅ "Hala kararlı mısın?" (orijinali hâlâ olduğu halde kabul edilebilir)
✅ "Türkiye Büyük Millet Meclisinde toplandılar." (kesme işareti unutulmuş demeyin, o kurallar sürekli değişiyor, yeni şekli de bu, kabul edelim)
✅ "Daha Sonra" (tam bir cümle değil, başharfler büyük yazılmış ama bir başlık olabilir bu, kabul etmek lazım)
✅ "Daha sonra o da geldi" (sonunda nokta unutulmuş, ama sorun değil, kabul edelim)
❌ "Daha sonra oda geldi" (açıkça anlaşılıyor ki "o da" yerine yanlışlıkla "oda" yazılmış, kabul etmeyelim)
❌ "Daha onra o da geldi" (açıkça bir yazım hatası var, reddedilmesi lazım)
❌ "TBMM'de toplandılar." (kısaltma var, reddedelim)
❌ "Çekoslovakyalılaştıramadıklarımızdanmısınız?" (sentetik kelime, -mi de ayrı yazılmamış)
❌ "Sizin için, insan kardeşlerim," (Orhan Veli'nin bir şiirinden mısra, kabul etmemek lazım)
❌ "Arthur Dent'in evinin gün bitmeden ortadan kaldırıldığını görmek ilgili bir şey ciddi biçimde ters gidiyordu." (Otostopçunun Galaksi Rehberi'nden alınmış, CC-0 olmayan telifli eser)
❌ "Bizi doğru yola, kendilerine nimet verdiklerinin yoluna ilet; gazaba uğrayanlarınkine ve sapıklarınkine değil." (dini metin)
❌ "Bu düşünce ile alınan teşebbüsât, birtakım teşekküller doğurdu." (Nutuk'tan alınan bu cümle çok eski ve artık kullanılmayan sözcükler içeriyor)
❌ "Ahmet 1960'da doğmuştu." (rakamlarla yazılmış sayı var)
❌ "Ahmet bindokuzyüzyetmişbirde doğmuştu..." (temel yazım kuralı hatası, bunu "bin dokuz yüz yetmiş birde" yazmak lazımdı, reddedilmeli)
❌ "Ahmet Ahmetoğulları'na http://example.com sitesinden ya da ahmet@example.com adresinden ulaşılabilir." (özel işaretler var, kişinin adı geçiyor, hatta e-posta adresi var)

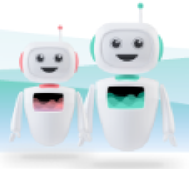


❓ *Yazan* | **Okunan** [Notlar]

✅ *Düşündüğü yalnız buydu.* | **Düşündüğü yanlız buydu.**
✅ *Aferin sana.* | **Afferin sana**
✅ *Bir gelemedin.* | **Bi gelemedin.**
✅ *İyi akşamlar...* | **İyakşamlar...**
✅ *Söyle bana* | **Süle bana**
✅ *Katil zanlısı.* | **Kaatil zanlısı.**
✅ *Değil mi?* | **Diilmi?**
✅ *Bir şey soracağım...* | **Bişey sorcam...**
✅ *Şaka yapıyor olmalısın.* | **Şaka yapıyo olmalısın.**
✅ *Bir işarete bakıyormuşsun.* | **Bi işarete bakıyomuşun.**
✅ *Herkes dışarı!* | **Herkez dışarı!**
✅ *Eğer gelmeyeceksen haber ver.* | **Eyer gelmiyceksen haber ver.**
✅ *Gelirken maydanoz alsana...* | **Gelirken maadonos alsana...**
✅ *Bir dakika bekle.* | **Bir dakka bekle.**
✅ *Sen neden bahsediyorsun?* | **Sen neyden bahsediyon?**
❌ *Tabii ki hayır!* | **Tabiykide hayır!**
❌ *Ama...* | **Aaamaa...** ['Kör' anlamındaki 'Âmâ olarak algılanacaktır]
❌ *Başımı çevirip baktığım zaman...* | **Başını çevirip baktığı zaman...** [Özne değişmiş]
❌ *Bir işarete bakıyormuşsun.* | **Bi işarete bakıyomusun.** [Yüklem değişmiş]
❌ *Sallanarak yukarıya çıktı.* | **Salınarak yukarı çıktı.** [Benzer gibi gelse de aslında anlamı farklı]
❌ *Sonra kaşlarını çatarak...* | **Sonra saçlarını saçarak...**
❌ *Kadının bu lüzumsuz merakı canımı sıkıyordu.* | **Kadının bu zulümsüz merakı canımı sıkıyordu.**
❌ *Birdenbire bu sesler kesildi.* | **sildi Birdenbire bu sesler kesildi.** [Bazı gönüllüler bir kere sesli okuyup ardından kayda basıyorlar, bir önceki okuyuştan kelimeler kalabiliyor]
❌ *İçki?* | **İçki? Ha ha, alırım...** [Eklenmiş kelimeler]
❌ *O, taş odalarda kim bilir ne yapıyor?* | **O, taş odalarında kim bilir ne yapıyor?**

Resources

- Main CV (About, FAQ, Terms, Criteria, Community Playbook, Campaign Guide) (commonvoice.mozilla.org)
- Sentence Collector (commonvoice.mozilla.org/sentence-collector/)
- Pontoon for UI translations (pontoon.mozilla.org)
- Discourse (forum - discourse.mozilla.org)
  - Common Voice
  - Check language sub forums
  - Deepspeech
- Matrix (chat)
  - Common Voice, Speech & Machine Learning, Coqui-ai/community
- Coqui website & docs & Gitter
- Github



Good to know!

- Native speaker – ask for a sub-Discourse, be translator on Pontoon
- Regular people will not come to Discourse/Matrix
- Regular people will not read technical jargon or long information
- Time / duration values on CV are approximations, no real statistics (yet)
- Volunteers must be >=20 years old or must have **consent**
- There is a «staging server» to check translations
  - *https://commonvoice.allizom.org/sentence-collector/#/tr/*
- Metadata for datasets are helpful
  - *https://github.com/common-voice/cv-dataset*
- Francis Morton Tyers' repos are very helpful
  - *https://github.com/ftyers/commonvoice-utils*
  - *https://github.com/ftyers/commonvoice-docker*

Dataset Analysis, Campaign & Monitoring

- Dataset analysis
- Set goals
- Design a social media campaign
- Monitoring
- Results for Turkish

Dataset analysis

- validated.tsv
- invalidated.tsv
- other.tsv

- train.tsv
- dev .tsv
- test.tsv

- reported.tsv

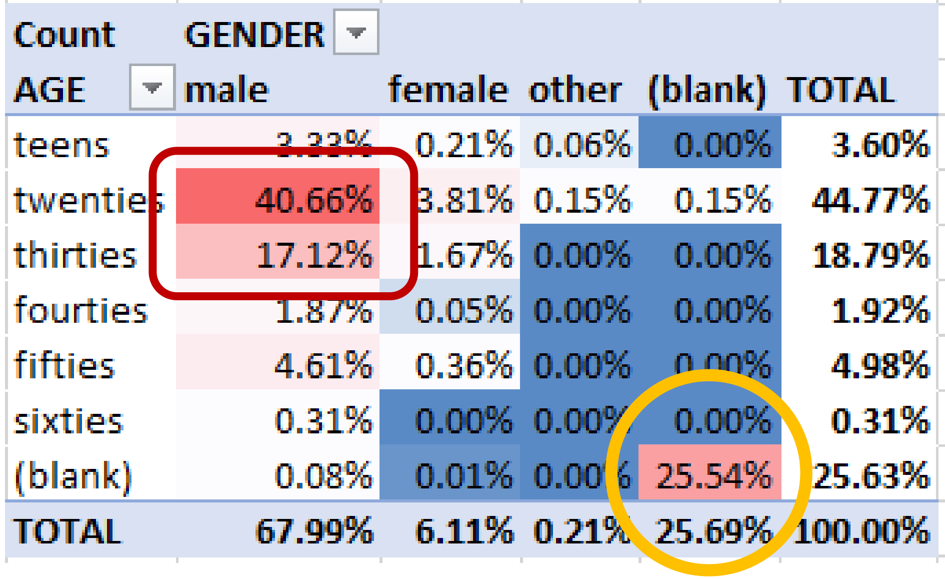
For all Validated / Train / Dev / Test

- Demografic bias/diversity?
  - Gender
  - Age
- Voice bias? (recs/person)
- Sentence bias? (recs/sentence)

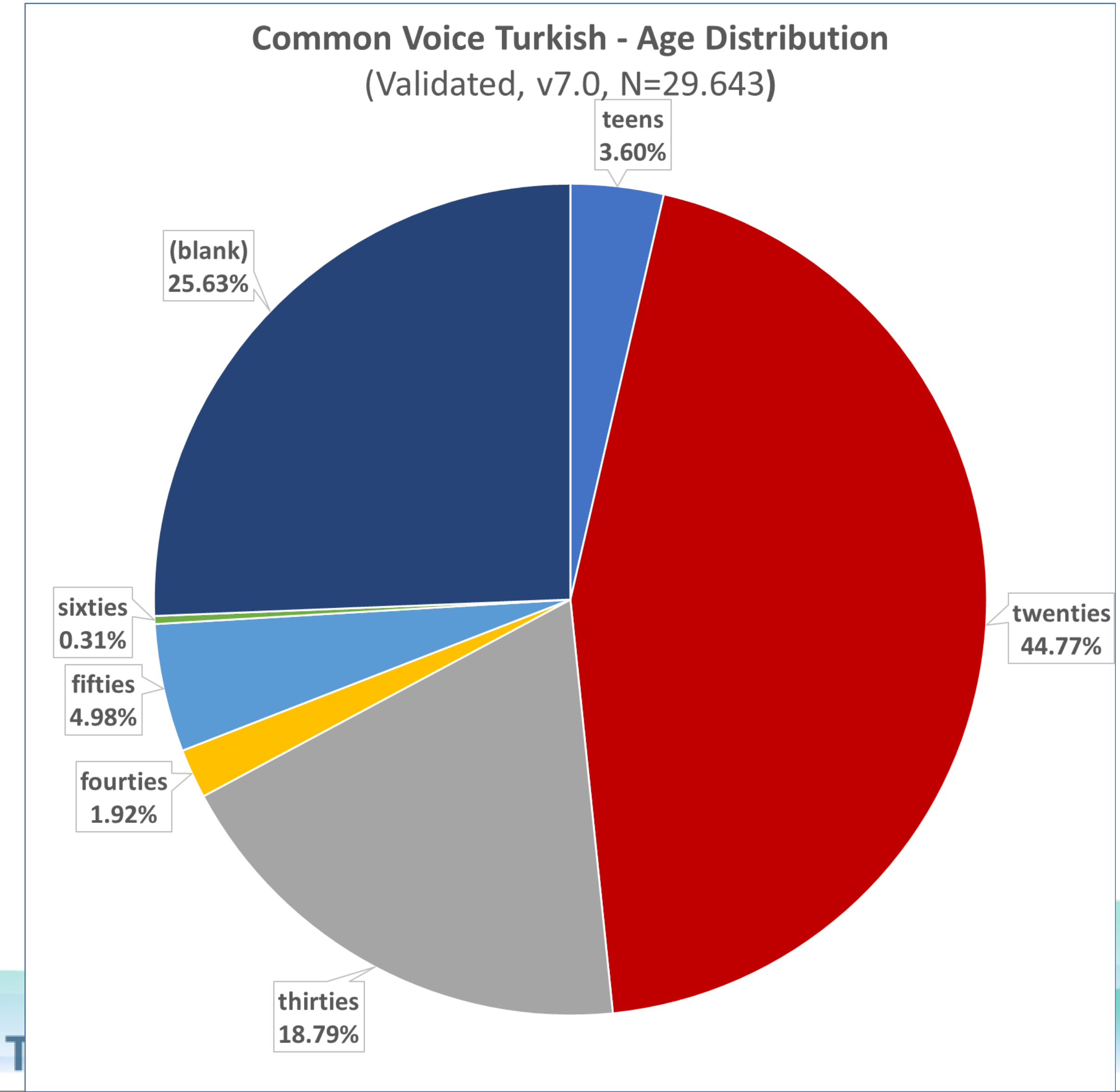
- Are rightfuly invalidated?
- 2 YES, 1 NO in validated?
- Reported rightfully?

TR v7.0 validated - Diversity

Gender & Age

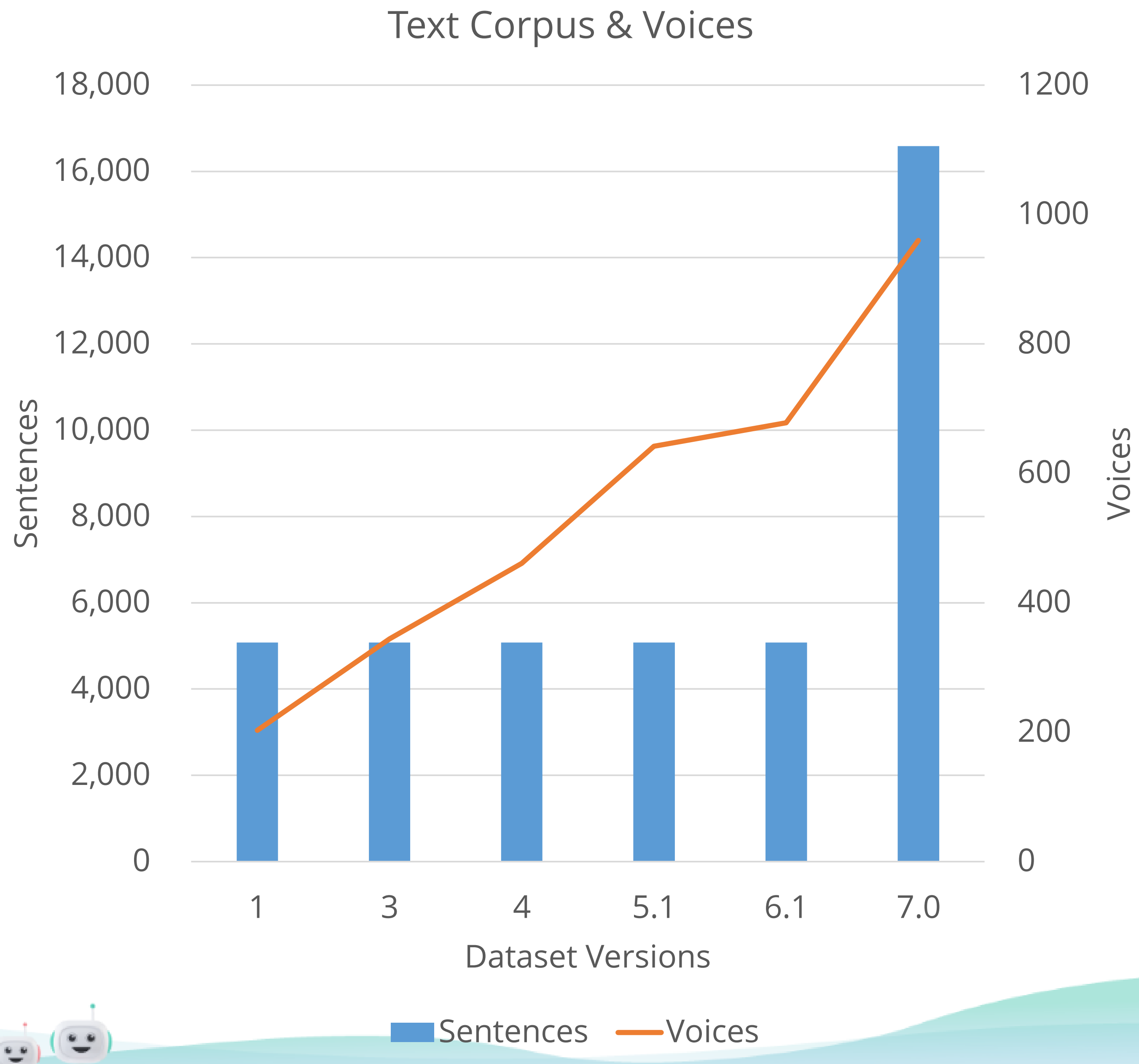


Count AGE	GENDER male	female	other	(blank)	TOTAL
teens	3.33%	0.21%	0.06%	0.00%	3.60%
twenties	40.66%	3.81%	0.15%	0.15%	44.77%
thirties	17.12%	1.67%	0.00%	0.00%	18.79%
fourties	1.87%	0.05%	0.00%	0.00%	1.92%
fifties	4.61%	0.36%	0.00%	0.00%	4.98%
sixties	0.31%	0.00%	0.00%	0.00%	0.31%
(blank)	0.08%	0.01%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


Common Voice Turkish - Age Distribution
(Validated, v7.0, N=29.643)

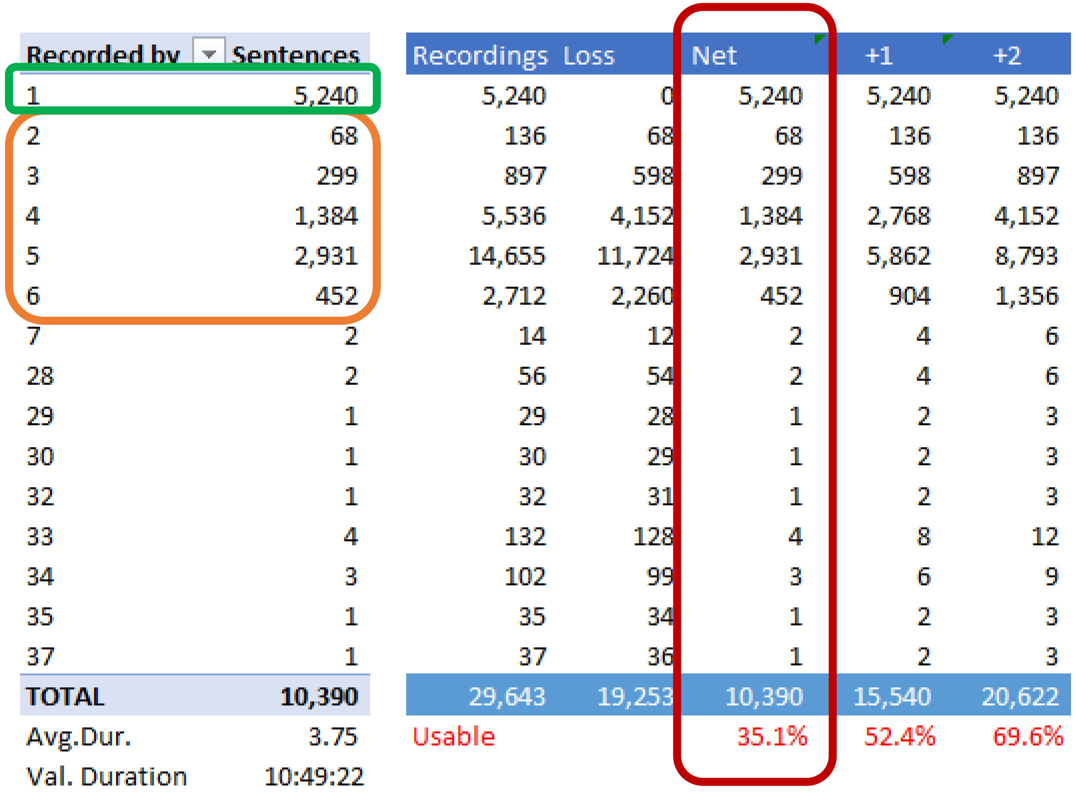
TR v1 – v7.0

- Text Corpus
- # Contributors



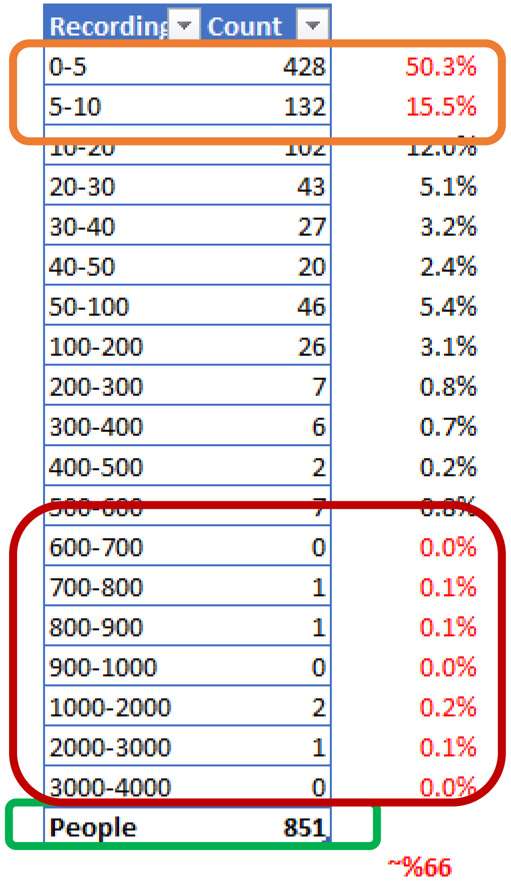
TR v7.0 validated - CorpusCreator

Recordings per sentence



Recorded by	Sentences	Recordings	Loss	Net	+1	+2
1	5,240	5,240	0	5,240	5,240	5,240
2	68	136	68	68	136	136
3	299	897	598	299	598	897
4	1,384	5,536	4,152	1,384	2,768	4,152
5	2,931	14,655	11,724	2,931	5,862	8,793
6	452	2,712	2,260	452	904	1,356
7	2	14	12	2	4	6
28	2	56	54	2	4	6
29	1	29	28	1	2	3
30	1	30	29	1	2	3
32	1	32	31	1	2	3
33	4	132	128	4	8	12
34	3	102	99	3	6	9
35	1	35	34	1	2	3
37	1	37	36	1	2	3
TOTAL	**10,390**	29,643	19,253	10,390	15,540	20,622
Avg.Dur.	3.75	Usable		35.1%	52.4%	69.6%
Val. Duration	10:49:22					

Recordings per person



Recording	Count	
0-5	428	50.3%
5-10	132	15.5%
10-20	102	12.0%
20-30	43	5.1%
30-40	27	3.2%
40-50	20	2.4%
50-100	46	5.4%
100-200	26	3.1%
200-300	7	0.8%
300-400	6	0.7%
400-500	2	0.2%
500-600	7	0.8%
600-700	0	0.0%
700-800	1	0.1%
800-900	1	0.1%
900-1000	0	0.0%
1000-2000	2	0.2%
2000-3000	1	0.1%
3000-4000	0	0.0%
People	**851**	

~%66

Set goals for the campaign

Current State

<u>Quantitative</u>

- Low Text Corpus
- Low number of recordings

<u>Qualitative</u>

- Low female/male Ratio
- Young male prominant

- (bias caused by single young male with accent)

Desired State / Goal

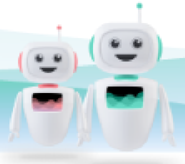
<u>Increase Amount</u>

- Add more Text Corpus
- Convince people to record
- (100-300 recordings/person)

<u>Reduce Bias</u>

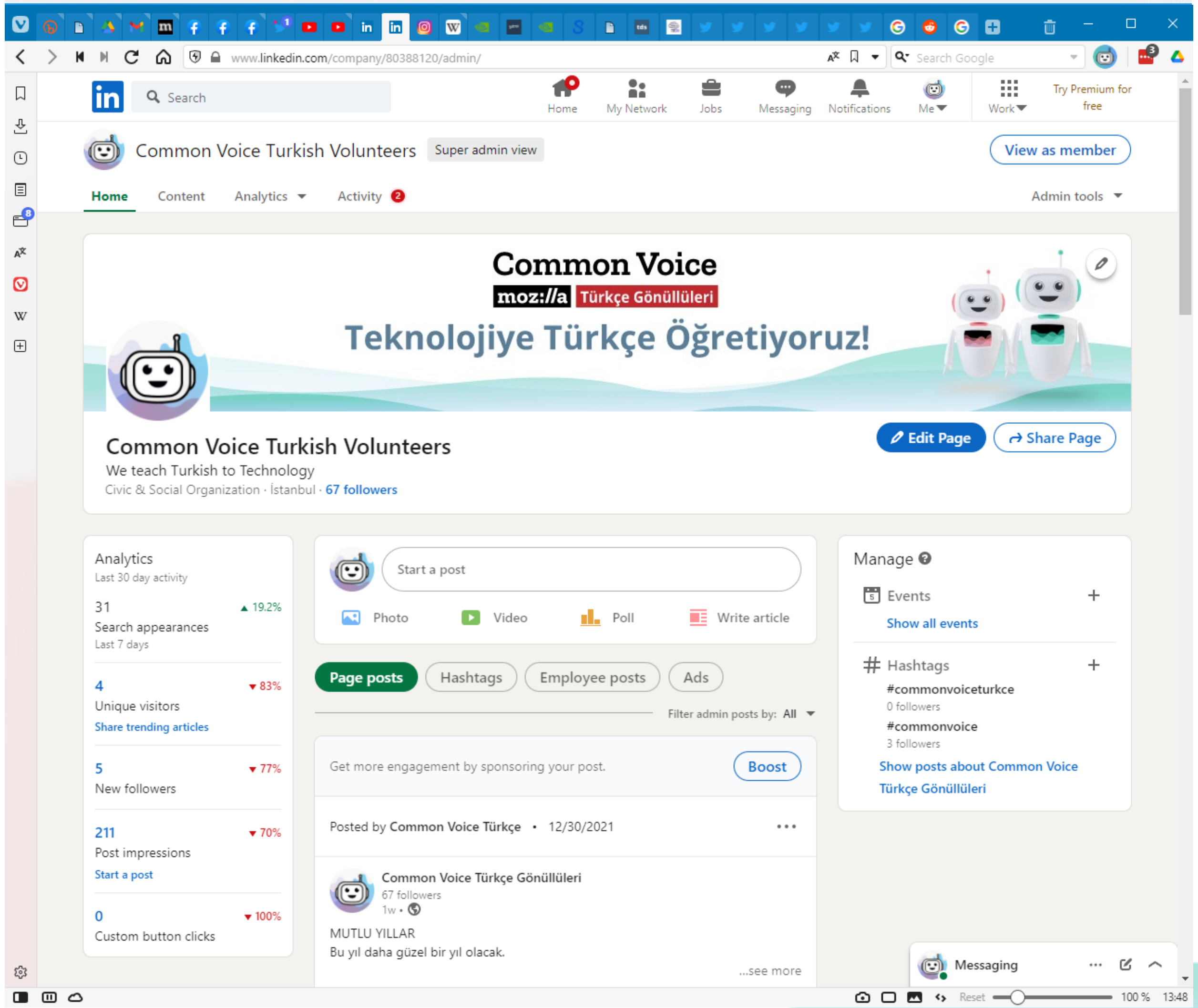
- More female voices
- More elderly (females)

- (> 1000 recordings for some)



Design a Social Media Campaign

- Two goals
  - Quantity & Quality
  - Build a community
- Multi Channel Campaign
  - Facebook (base & guides)
  - Youtube for guides
  - Telegram for instant support
  - + Twitter
  - + Instagram
  - + LinkedIn
- (Domain/Website)
- Bitly to get feedback



www.linkedin.com/company/80388120/admin/

Common Voice Turkish Volunteers Super admin view

Common Voice moz://a Türkçe Gönüllüleri

Teknolojiye Türkçe Öğretiyoruz!

Common Voice Turkish Volunteers

We teach Turkish to Technology

Civic & Social Organization · İstanbul · 67 followers

Analytics

Last 30 day activity

31 ▲ 19.2%
Search appearances
Last 7 days

4 ▼ 83%
Unique visitors
Share trending articles

5 ▼ 77%
New followers

211 ▼ 70%
Post impressions
Start a post

0 ▼ 100%
Custom button clicks

Get more engagement by sponsoring your post.

Posted by Common Voice Türkçe • 12/30/2021

Common Voice Türkçe Gönüllüleri
67 followers
1w •

MUTLU YILLAR
Bu yıl daha güzel bir yıl olacak.
...see more

Manage

Events
Show all events

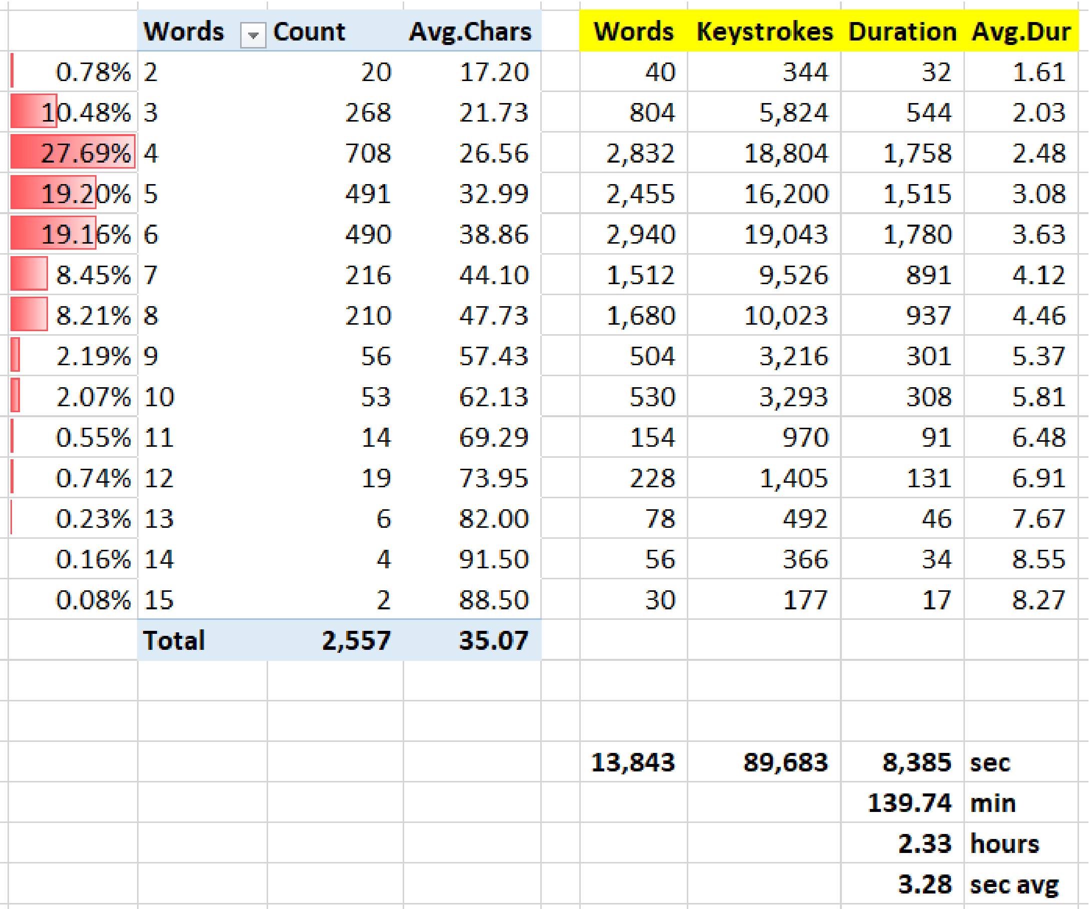
Hashtags
#commonvoiceturkce
0 followers
#commonvoice
3 followers
Show posts about Common Voice Türkçe Gönüllüleri

Prepare slogans, designs & messages

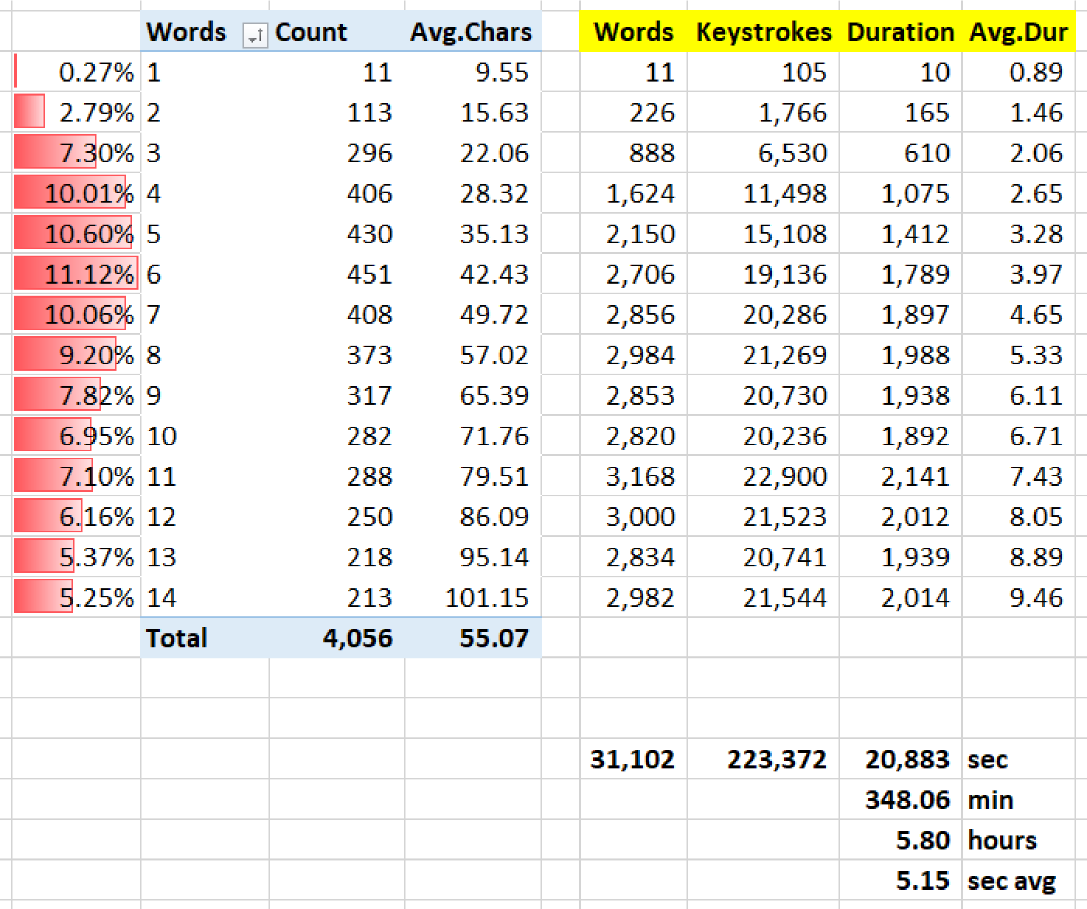
- Teasers
- Technical information
- Feedback from the campaign, dataset or trainings
- Call for female voices
- New year celebration

MUTLU YILLAR
Bu yıl daha güzel olacak
Common Voice
moz://a Türkçe Gönüllüleri
Teknolojiye Türkçe Öğretiyoruz!

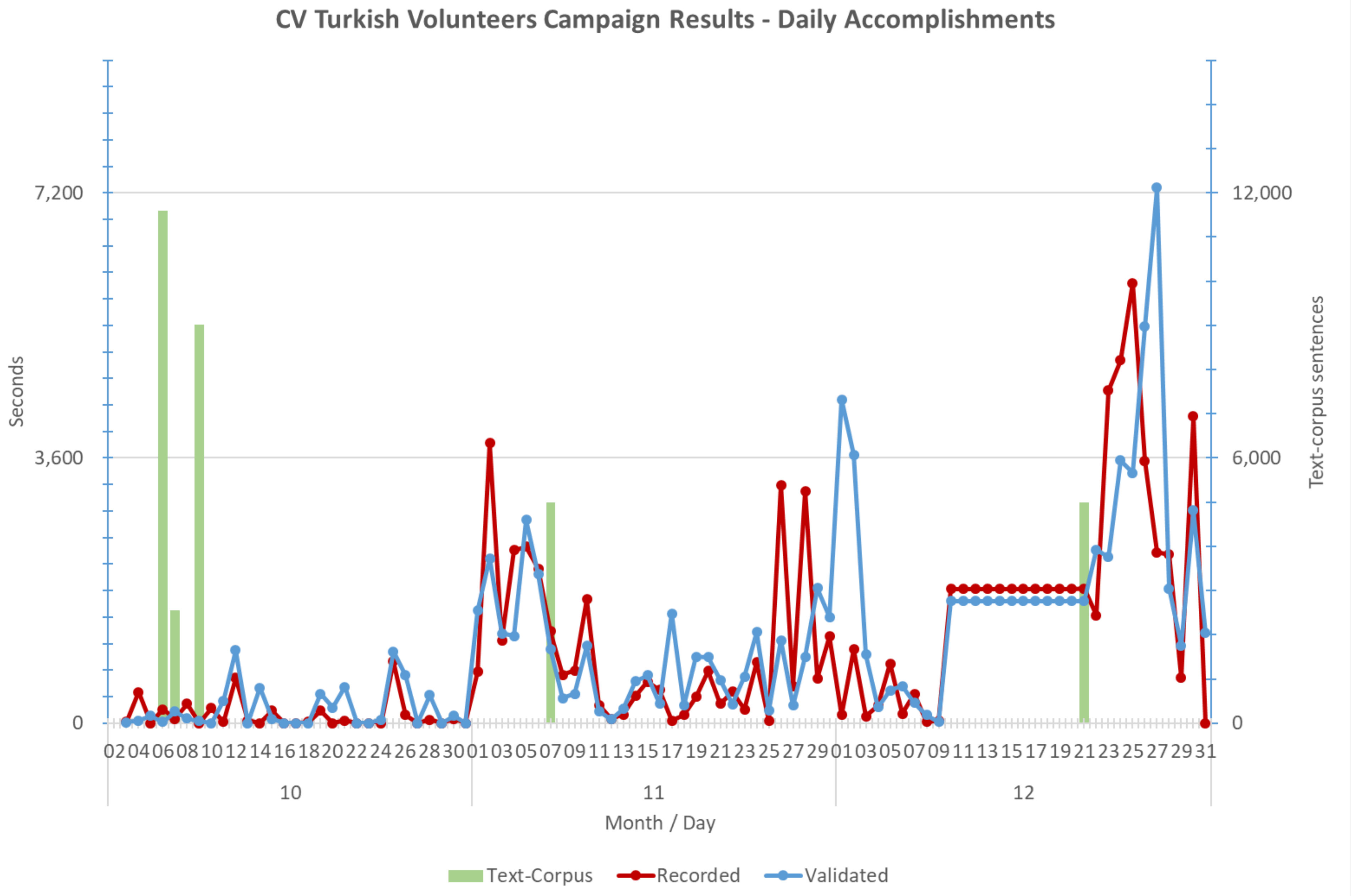
Prepare text-corpus!



	Words	Count	Avg.Chars	Words	Keystrokes	Duration	Avg.Dur
0.78%	2	20	17.20	40	344	32	1.61
10.48%	3	268	21.73	804	5,824	544	2.03
27.69%	4	708	26.56	2,832	18,804	1,758	2.48
19.20%	5	491	32.99	2,455	16,200	1,515	3.08
19.16%	6	490	38.86	2,940	19,043	1,780	3.63
8.45%	7	216	44.10	1,512	9,526	891	4.12
8.21%	8	210	47.73	1,680	10,023	937	4.46
2.19%	9	56	57.43	504	3,216	301	5.37
2.07%	10	53	62.13	530	3,293	308	5.81
0.55%	11	14	69.29	154	970	91	6.48
0.74%	12	19	73.95	228	1,405	131	6.91
0.23%	13	6	82.00	78	492	46	7.67
0.16%	14	4	91.50	56	366	34	8.55
0.08%	15	2	88.50	30	177	17	8.27
	Total	**2,557**	**35.07**				
				13,843	**89,683**	**8,385**	**sec**
						139.74	**min**
						2.33	**hours**
						3.28	**sec avg**



	Words	Count	Avg.Chars	Words	Keystrokes	Duration	Avg.Dur
0.27%	1	11	9.55	11	105	10	0.89
2.79%	2	113	15.63	226	1,766	165	1.46
7.30%	3	296	22.06	888	6,530	610	2.06
10.01%	4	406	28.32	1,624	11,498	1,075	2.65
10.60%	5	430	35.13	2,150	15,108	1,412	3.28
11.12%	6	451	42.43	2,706	19,136	1,789	3.97
10.06%	7	408	49.72	2,856	20,286	1,897	4.65
9.20%	8	373	57.02	2,984	21,269	1,988	5.33
7.82%	9	317	65.39	2,853	20,730	1,938	6.11
6.95%	10	282	71.76	2,820	20,236	1,892	6.71
7.10%	11	288	79.51	3,168	22,900	2,141	7.43
6.16%	12	250	86.09	3,000	21,523	2,012	8.05
5.37%	13	218	95.14	2,834	20,741	1,939	8.89
5.25%	14	213	101.15	2,982	21,544	2,014	9.46
	Total	**4,056**	**55.07**				
				31,102	**223,372**	**20,883**	**sec**
						348.06	**min**
						5.80	**hours**
						5.15	**sec avg**



Common Voice

moz://a Turkish Volunteers

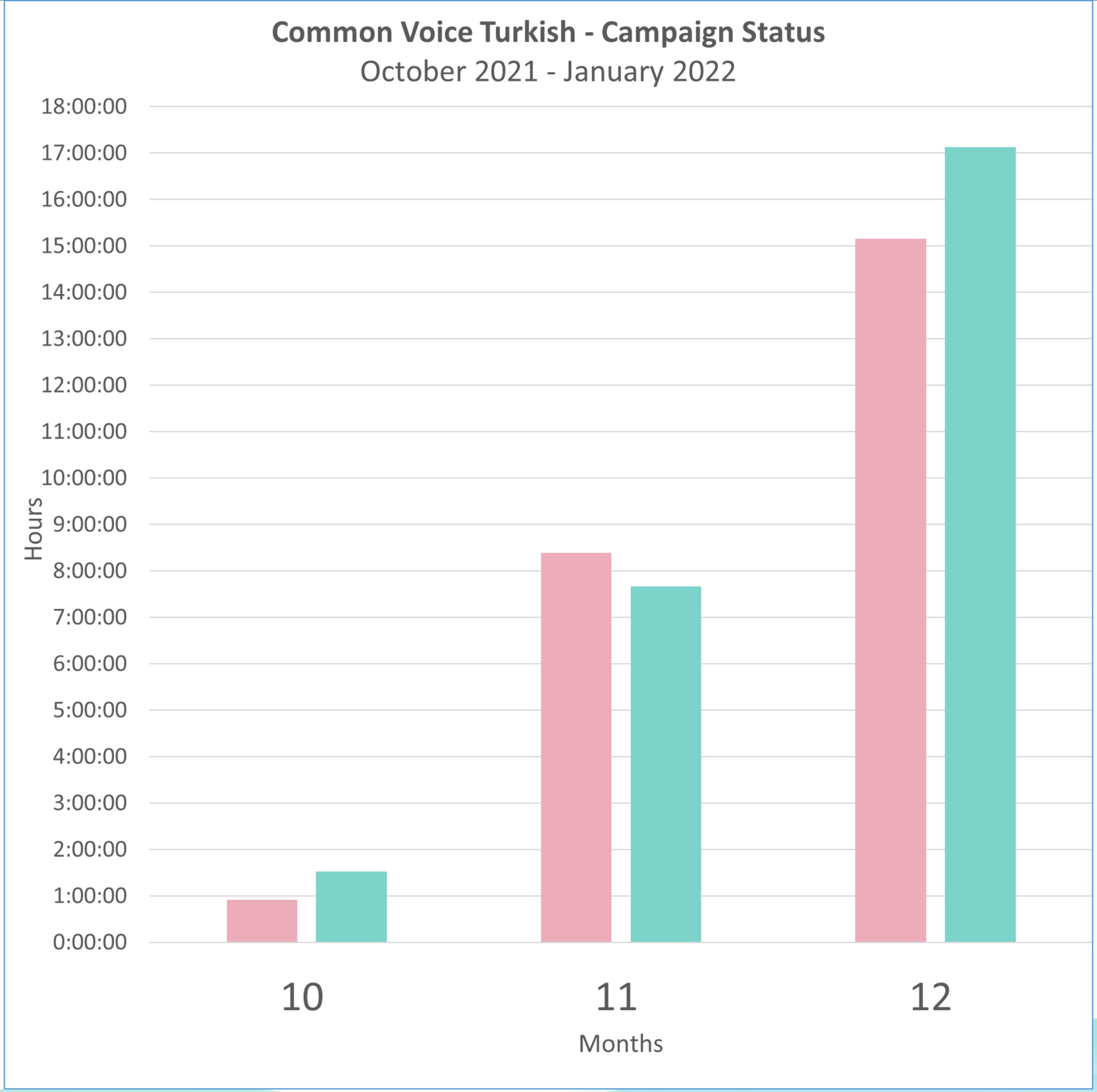
We Teach Turkish to Technology

As of 31/12/2021

Campaign started November 1st

November: Wide 100-300 recs/person

December: «Group work» (1000+ recs)



As of 31/12/2021

From v7

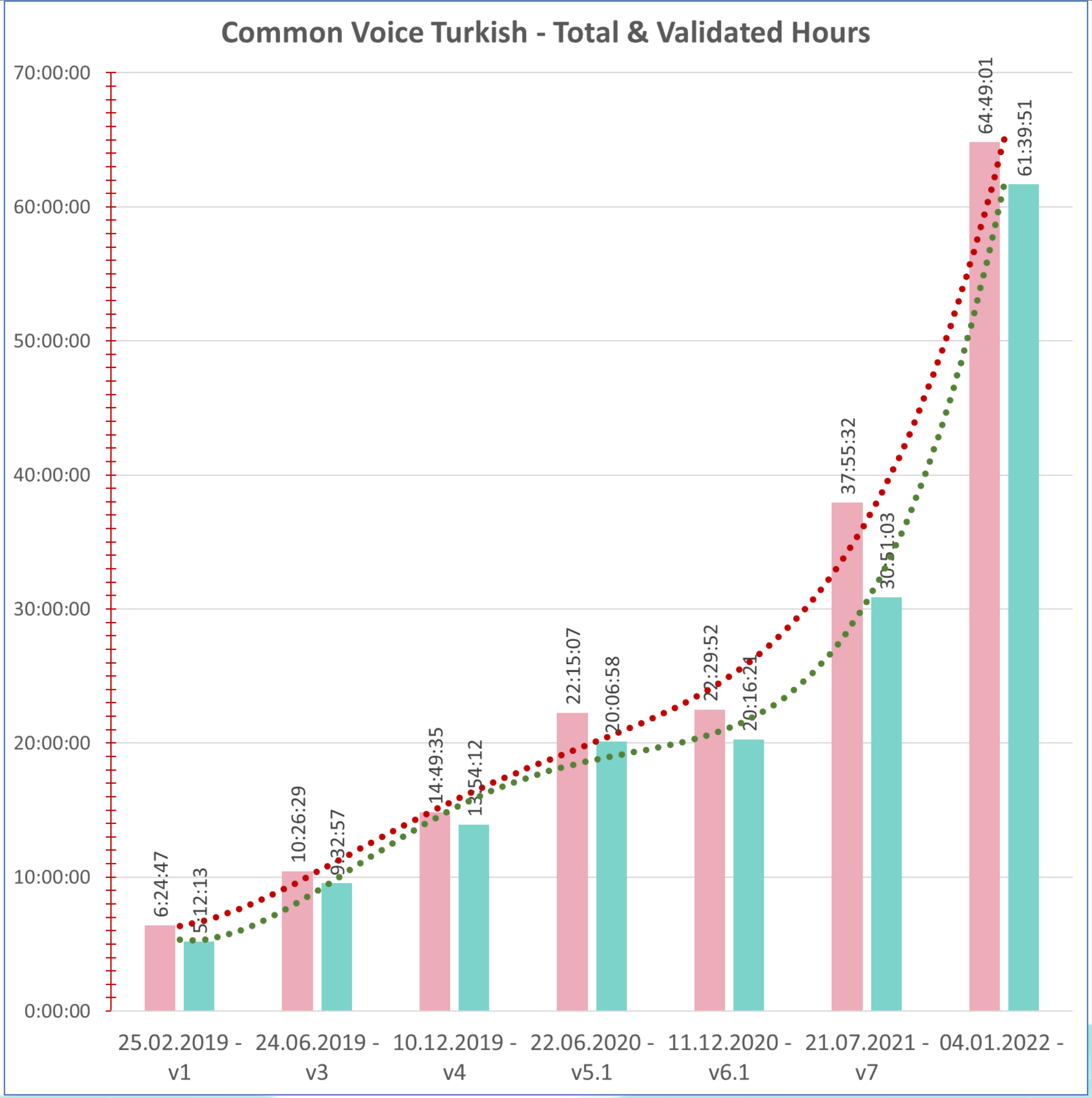
27+ hours total (71%+ increase)

31+ hours validated (100%+ increase)

From v6.1

>190% increase in total

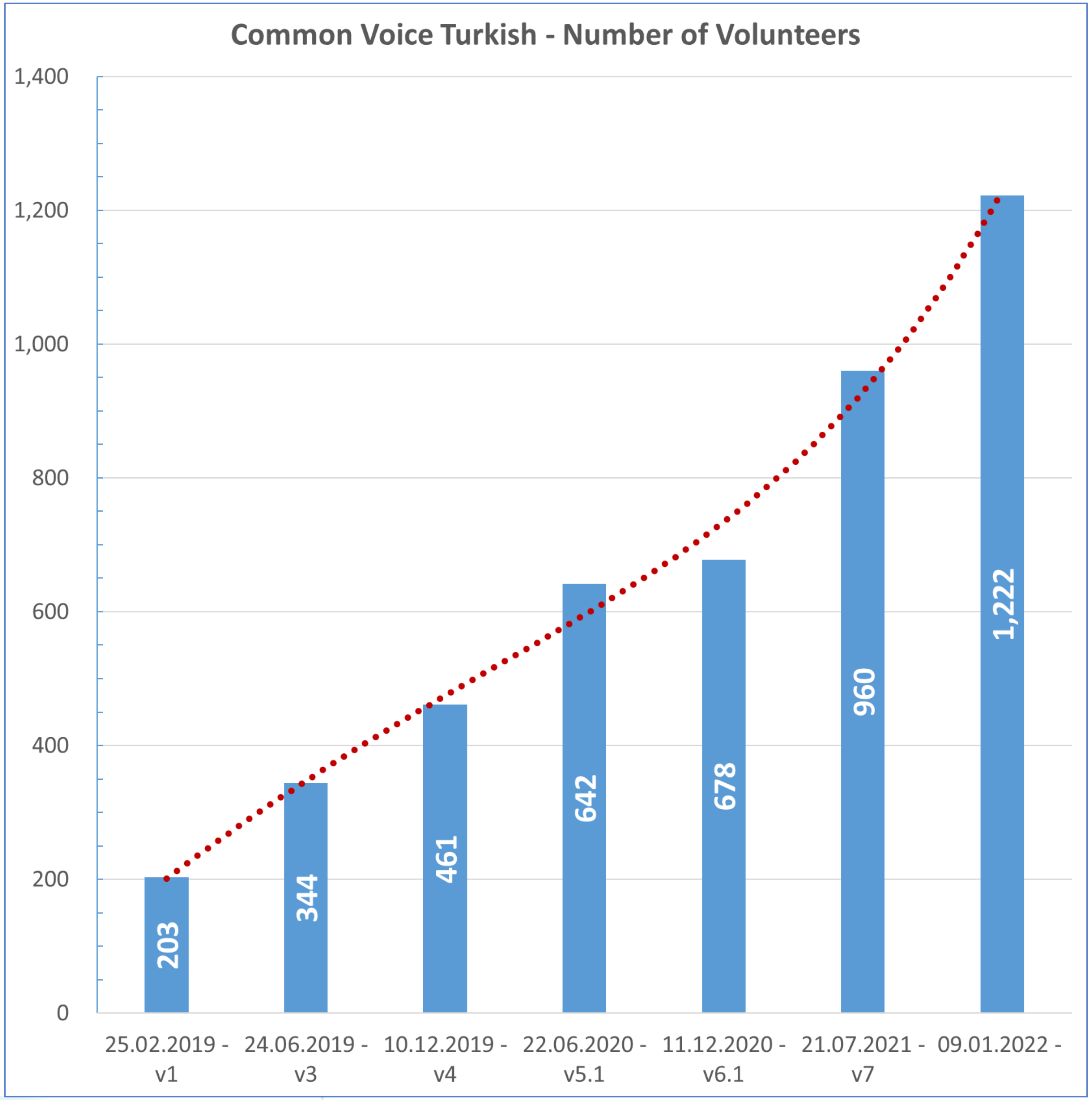
>205% increase in validated



Common Voice

moz://a Turkish Volunteers

We Teach Turkish to Technology

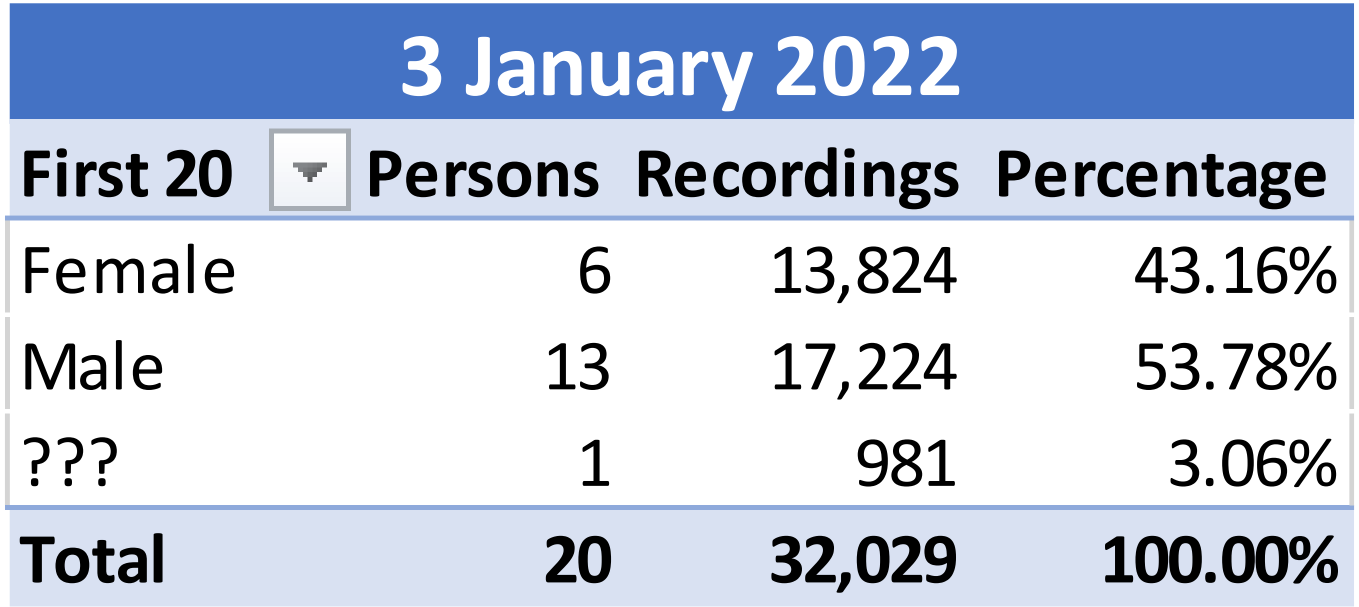


Common Voice

moz://a Turkish Volunteers

We Teach Turkish to Technology

Diversity?



3 January 2022			
First 20 ▾	Persons	Recordings	Percentage
Female	6	13,824	43.16%
Male	13	17,224	53.78%
???	1	981	3.06%
Total	**20**	**32,029**	**100.00%**

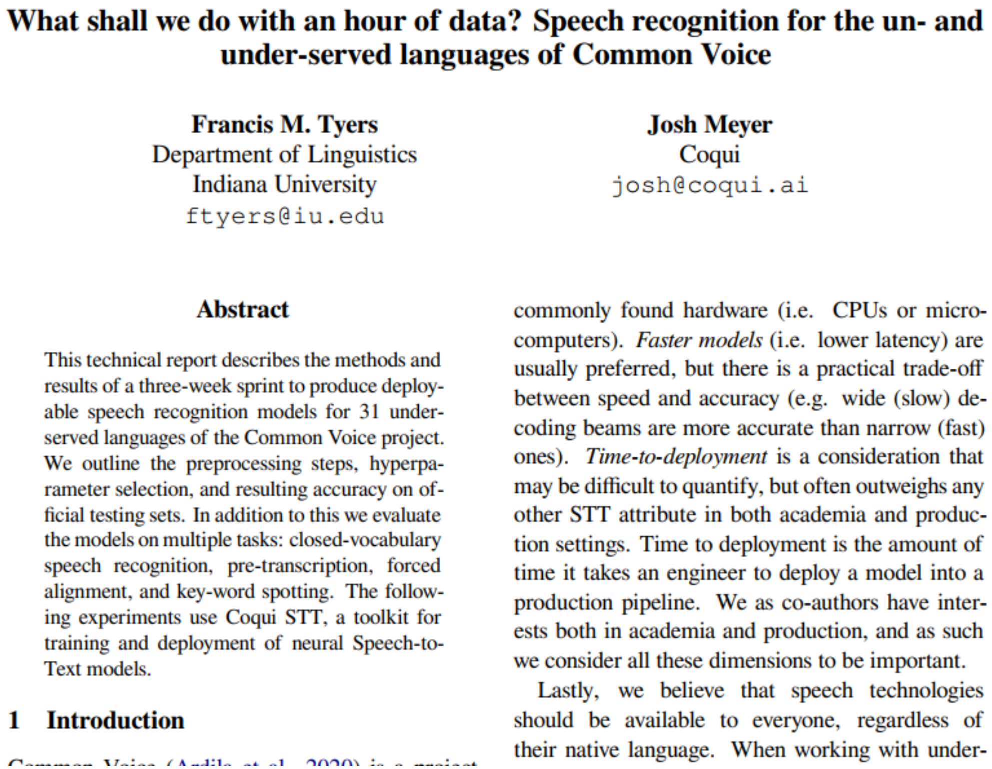
Lessons learned

- Widespread posting => low return
- Better: Targeted posts to relatives/friends
- Build 3-5 person groups and educate/support them
  - All record - All validate - One support them
- Social Media Algorithms are bad for your campaign.
- Best response from Twitter - young AI/Data Science experts
- Be prepared to give extensive technical support and spending time on validating recordings.
  - Mic connection problems
  - Login problems (e.g. elderly people)
  - Problems caused by mobile apps (e.g. WebView limitations)

Training a Voice AI

- Coqui + Google Colab
- Trained v1-v7
- Results for Turkish

Base: Tyers & Meyer 2021



What shall we do with an hour of data? Speech recognition for the un- and under-served languages of Common Voice

Francis M. Tyers
Department of Linguistics
Indiana University
ftyers@iu.edu

Josh Meyer
Coqui
josh@coqui.ai

Abstract

This technical report describes the methods and results of a three-week sprint to produce deployable speech recognition models for 31 under-served languages of the Common Voice project. We outline the preprocessing steps, hyperparameter selection, and resulting accuracy on official testing sets. In addition to this we evaluate the models on multiple tasks: closed-vocabulary speech recognition, pre-transcription, forced alignment, and key-word spotting. The following experiments use Coqui STT, a toolkit for training and deployment of neural Speech-to-Text models.

1 Introduction

Common Voice (Ardila et al., 2020) is a project

commonly found hardware (i.e. CPUs or micro-computers). *Faster models* (i.e. lower latency) are usually preferred, but there is a practical trade-off between speed and accuracy (e.g. wide (slow) decoding beams are more accurate than narrow (fast) ones). *Time-to-deployment* is a consideration that may be difficult to quantify, but often outweighs any other STT attribute in both academia and production settings. Time to deployment is the amount of time it takes an engineer to deploy a model into a production pipeline. We as co-authors have interests both in academia and production, and as such we consider all these dimensions to be important.

Lastly, we believe that speech technologies should be available to everyone, regardless of their native language. When working with under-

What shall we do with an hour of data?

Speech recognition for the un- and under-served languages of Common Voice

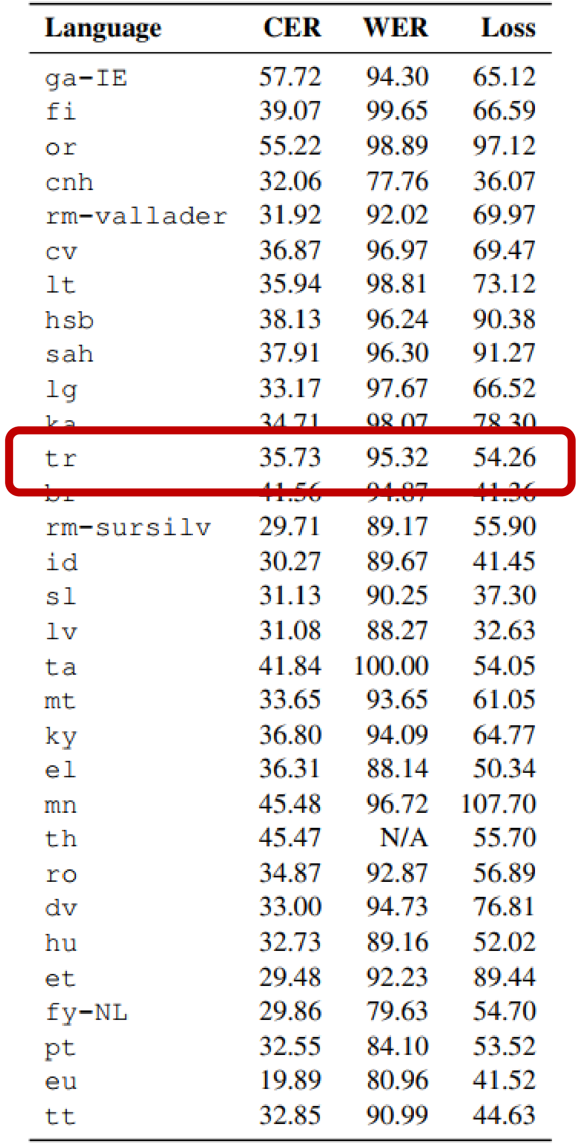
https://arxiv.org/abs/2105.04674

April 2021

- Based on Coqui 0.9.3
- Dataset v6.1 roundup

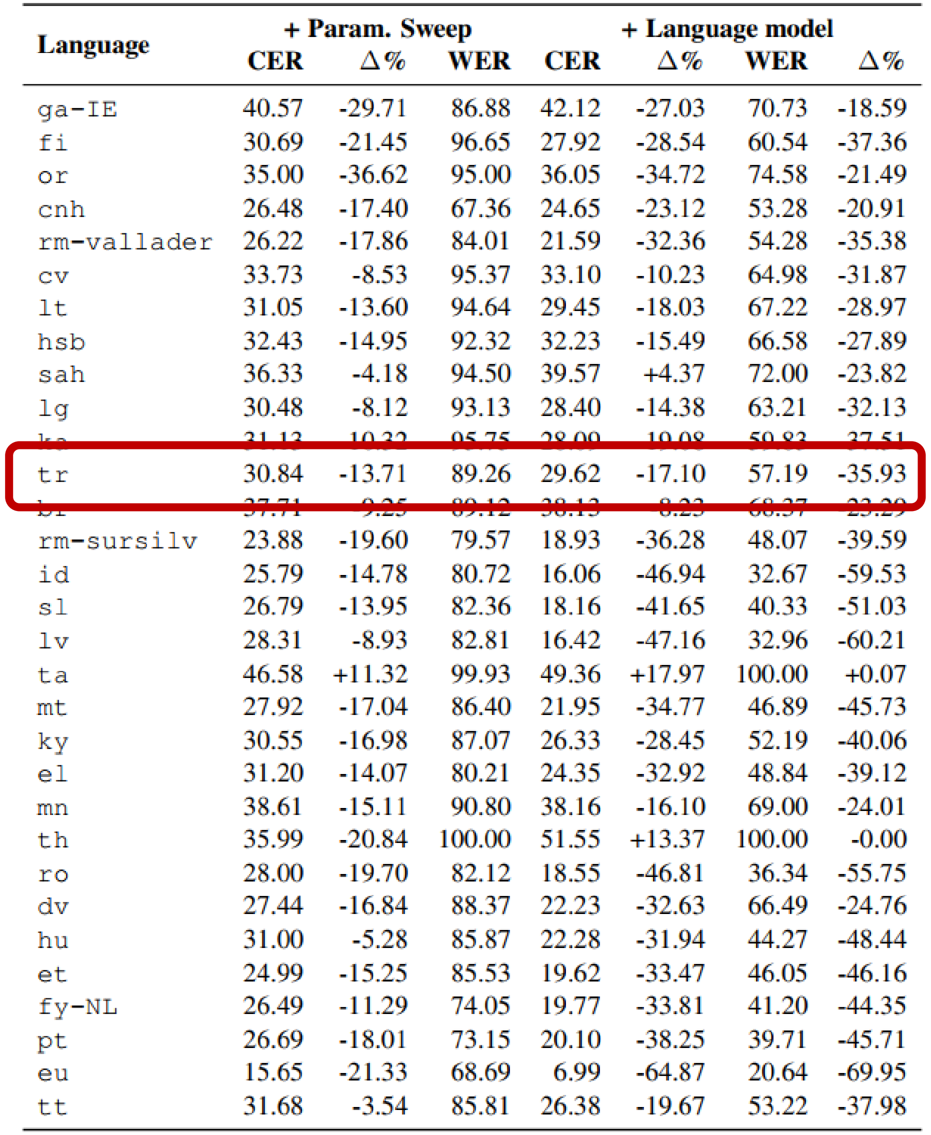
3-RUNS

- Baseline (short training)
- Parameter sweep => longer training
- Add Language Model



Language	CER	WER	Loss
ga-IE	57.72	94.30	65.12
fi	39.07	99.65	66.59
or	55.22	98.89	97.12
cnh	32.06	77.76	36.07
rm-vallader	31.92	92.02	69.97
cv	36.87	96.97	69.47
lt	35.94	98.81	73.12
hsb	38.13	96.24	90.38
sah	37.91	96.30	91.27
lg	33.17	97.67	66.52
ka	34.71	98.07	78.30
tr	35.73	95.32	54.26
bi	41.56	94.87	41.36
rm-sursilv	29.71	89.17	55.90
id	30.27	89.67	41.45
sl	31.13	90.25	37.30
lv	31.08	88.27	32.63
ta	41.84	100.00	54.05
mt	33.65	93.65	61.05
ky	36.80	94.09	64.77
el	36.31	88.14	50.34
mn	45.48	96.72	107.70
th	45.47	N/A	55.70
ro	34.87	92.87	56.89
dv	33.00	94.73	76.81
hu	32.73	89.16	52.02
et	29.48	92.23	89.44
fy-NL	29.86	79.63	54.70
pt	32.55	84.10	53.52
eu	19.89	80.96	41.52
tt	32.85	90.99	44.63

Table 2: **Baseline results**. Character Error Rate (CER), Word Error Rate (WER), and CTC Loss reported on



Language	+ Param. Sweep			+ Language model			
	CER	Δ%	WER	CER	Δ%	WER	Δ%
ga-IE	40.57	-29.71	86.88	42.12	-27.03	70.73	-18.59
fi	30.69	-21.45	96.65	27.92	-28.54	60.54	-37.36
or	35.00	-36.62	95.00	36.05	-34.72	74.58	-21.49
cnh	26.48	-17.40	67.36	24.65	-23.12	53.28	-20.91
rm-vallader	26.22	-17.86	84.01	21.59	-32.36	54.28	-35.38
cv	33.73	-8.53	95.37	33.10	-10.23	64.98	-31.87
lt	31.05	-13.60	94.64	29.45	-18.03	67.22	-28.97
hsb	32.43	-14.95	92.32	32.23	-15.49	66.58	-27.89
sah	36.33	-4.18	94.50	39.57	+4.37	72.00	-23.82
lg	30.48	-8.12	93.13	28.40	-14.38	63.21	-32.13
ka	31.13	-10.32	95.75	28.09	-19.08	59.83	-37.51
tr	30.84	-13.71	89.26	29.62	-17.10	57.19	-35.93
bi	37.71	-9.25	89.12	38.13	-8.23	68.37	-23.29
rm-sursilv	23.88	-19.60	79.57	18.93	-36.28	48.07	-39.59
id	25.79	-14.78	80.72	16.06	-46.94	32.67	-59.53
sl	26.79	-13.95	82.36	18.16	-41.65	40.33	-51.03
lv	28.31	-8.93	82.81	16.42	-47.16	32.96	-60.21
ta	46.58	+11.32	99.93	49.36	+17.97	100.00	+0.07
mt	27.92	-17.04	86.40	21.95	-34.77	46.89	-45.73
ky	30.55	-16.98	87.07	26.33	-28.45	52.19	-40.06
el	31.20	-14.07	80.21	24.35	-32.92	48.84	-39.12
mn	38.61	-15.11	90.80	38.16	-16.10	69.00	-24.01
th	35.99	-20.84	100.00	51.55	+13.37	100.00	-0.00
ro	28.00	-19.70	82.12	18.55	-46.81	36.34	-55.75
dv	27.44	-16.84	88.37	22.23	-32.63	66.49	-24.76
hu	31.00	-5.28	85.87	22.28	-31.94	44.27	-48.44
et	24.99	-15.25	85.53	19.62	-33.47	46.05	-46.16
fy-NL	26.49	-11.29	74.05	19.77	-33.81	41.20	-44.35
pt	26.69	-18.01	73.15	20.10	-38.25	39.71	-45.71
eu	15.65	-21.33	68.69	6.99	-64.87	20.64	-69.95
tt	31.68	-3.54	85.81	26.38	-19.67	53.22	-37.98

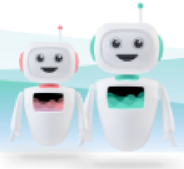
Table 5: **Results after parameter sweep and after adding a generic language model**. Character (CER) and Word Error Rate (WER) for each language when adding a language model.

WER

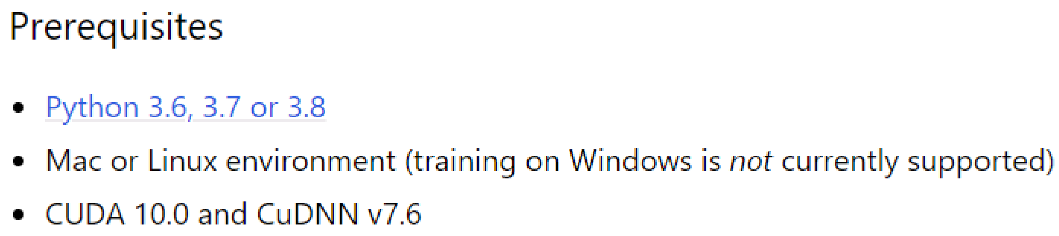
- **Baseline:** **95.32**
- **Long run:** **89.26**
- **+LM:** **57.19**

Decisions

- Use Coqui to replicate the experiment
- Alternatives (more costly)
  - Local Linux machine with powerful GPUs (chip shortage!)
  - Cloud instances & Docker
- Use Google Colab (low-cost)
  - Interactive notebooks (good for learning)
  - Powerful GPU's
  - No «that powerful» local linux machine, VM are no go
  - Free tier (K80 ~ 10x+ wrt CPU) => Colab Pro (P100 ~ 3x+)
- Roundup => Disk space! => 100GB Google Drive
  - Download problems/slow download



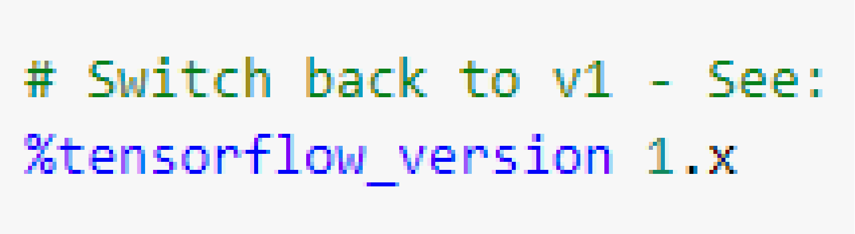
Coqui STT on Colab



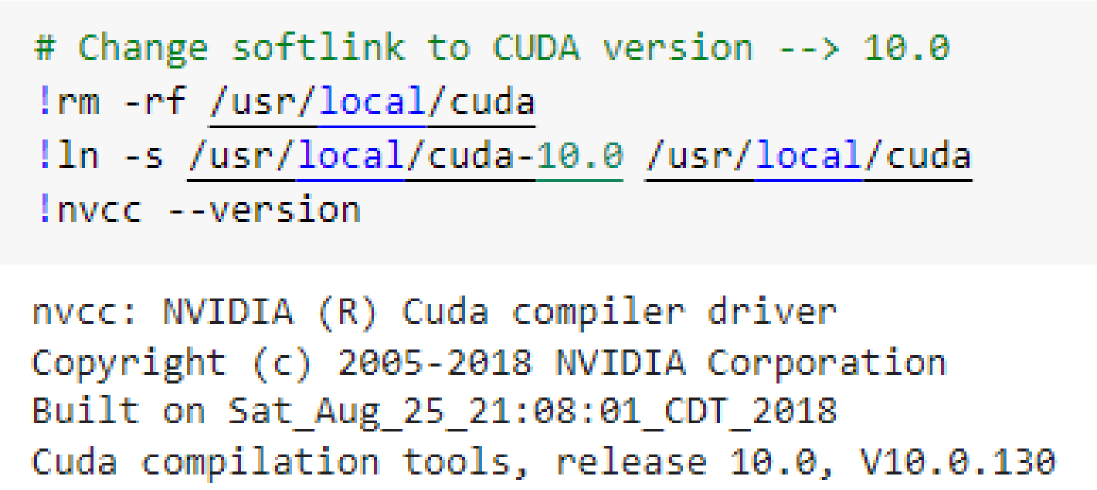
Prerequisites

- Python 3.6, 3.7 or 3.8
- Mac or Linux environment (training on Windows is *not* currently supported)
- CUDA 10.0 and CuDNN v7.6

- + Tensorflow 1.5.4 (GPU)



```
# Switch back to v1 - See:
%tensorflow_version 1.x
```

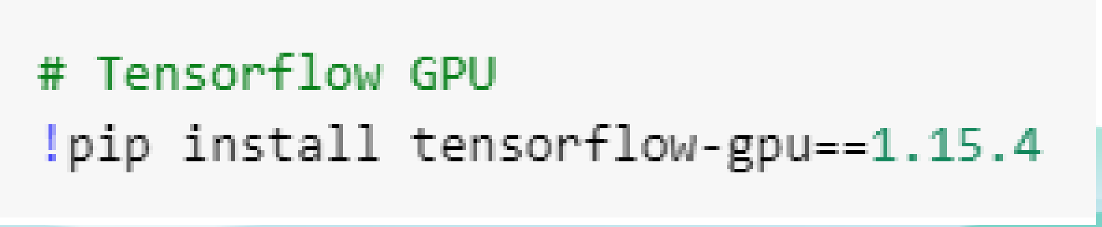


```
# Change softlink to CUDA version --> 10.0
!rm -rf /usr/local/cuda
!ln -s /usr/local/cuda-10.0 /usr/local/cuda
!nvcc --version
```

```
nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2018 NVIDIA Corporation
Built on Sat_Aug_25_21:08:01_CDT_2018
Cuda compilation tools, release 10.0, V10.0.130
```

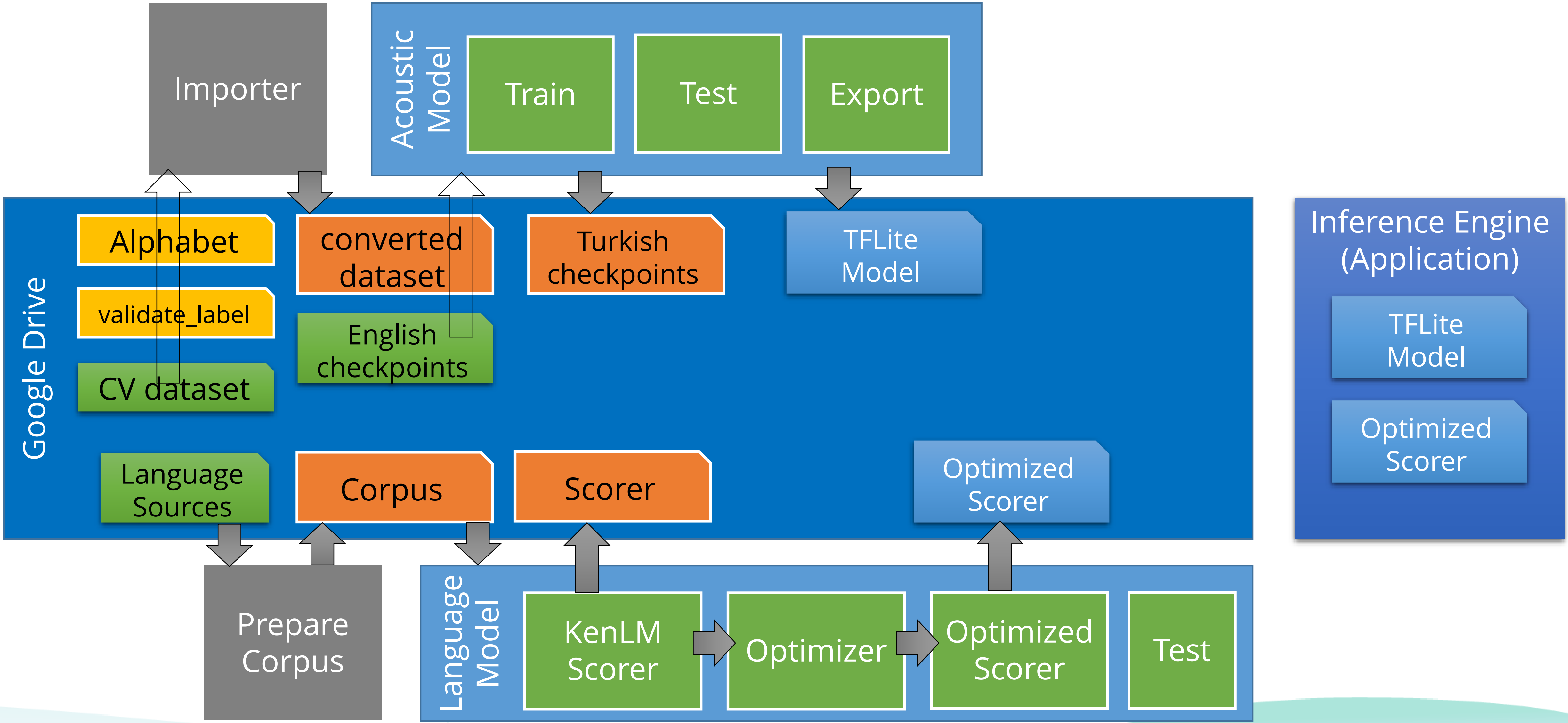
Current Default Colab

- Ubuntu 18.04.5 LTS
- Intel Xeon @2.30 GHz, 13G RAM
- Python 3.7.12
- Tensorflow CPU 2.x
- Cuda 11.2
- CuDNN 7.6.5



```
# Tensorflow GPU
!pip install tensorflow-gpu==1.15.4
```

Workflow



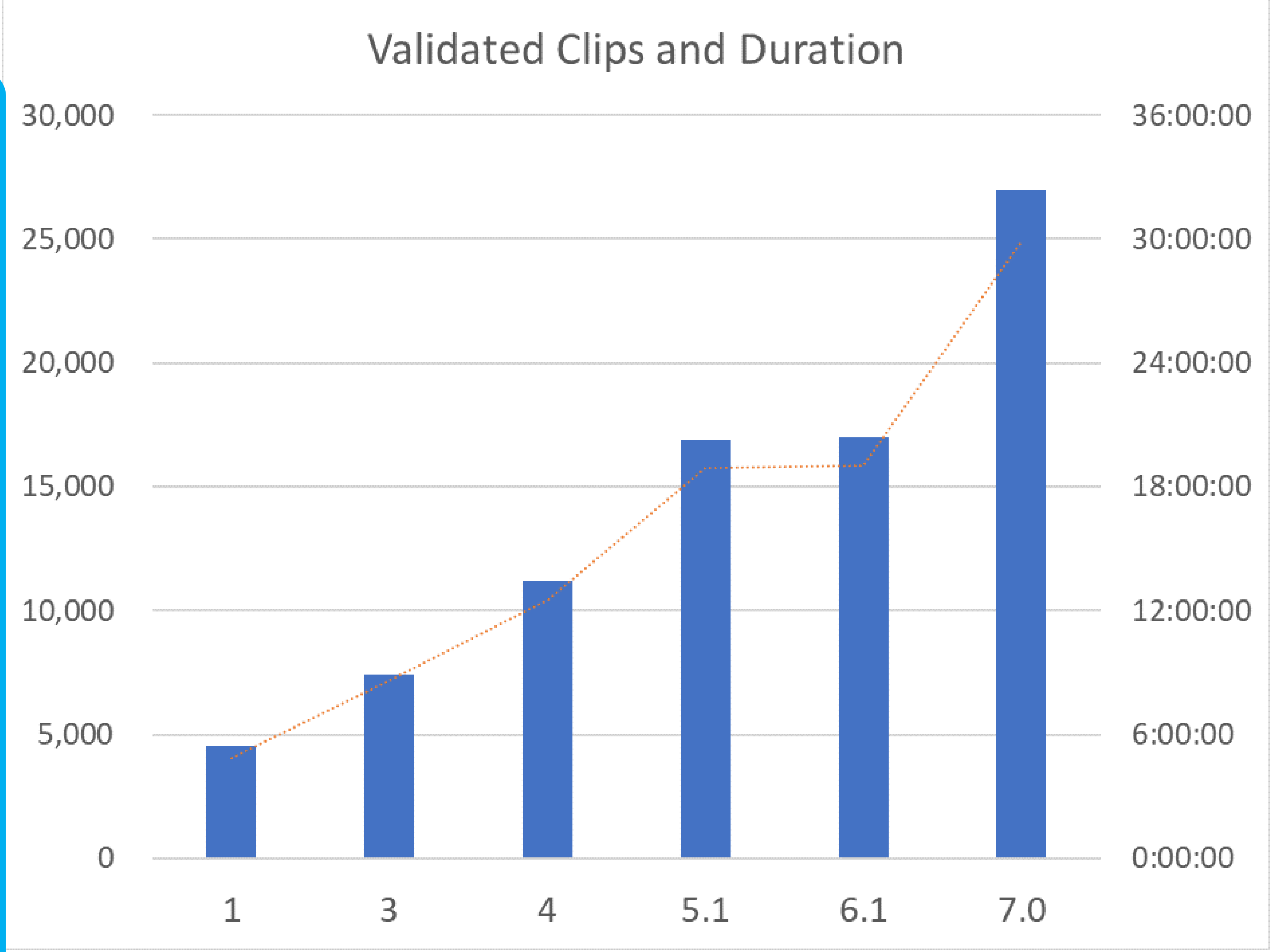
Common Voice

moz://a Turkish Volunteers

We Teach Turkish to Technology

Common Voice Turkish Coqui v1.0.0 Import Results

Datasets		Clips			Duration		
Ver	Set	Recs	Skip	Net	Initial	Final	Dur/Rec
1	Validated	4,807	245	4,562	5:06:21	4:51:08	3.829
	Train	1,209	56	1,153	1:17:34	1:14:16	3.865
	Dev	982	57	925	1:00:28	0:56:57	3.694
	Test	1,069	43	1,026	1:11:48	1:08:40	4.016
3	Validated	8,157	714	7,443	9:22:59	8:37:20	4.170
	Train	1,600	202	1,398	1:51:20	1:38:58	4.247
	Dev	1,463	157	1,306	1:38:56	1:28:48	4.080
	Test	1,521	100	1,421	1:54:25	1:47:00	4.518
4	Validated	11,787	577	11,210	13:08:12	12:31:00	4.020
	Train	1,729	94	1,635	1:52:27	1:46:54	3.923
	Dev	1,538	86	1,452	1:46:20	1:40:19	4.145
	Test	1,564	65	1,499	1:58:33	1:53:20	4.536
5.1	Validated	17,714	830	16,884	19:49:14	18:55:20	4.035
	Train	1,729	92	1,637	1:57:46	1:51:58	4.104
	Dev	1,548	90	1,458	1:51:48	1:45:28	4.340
	Test	1,576	63	1,513	2:02:06	1:56:54	4.636
6.1	Validated	17,851	834	17,017	19:57:42	19:03:23	4.031
	Train	1,738	93	1,645	1:58:23	1:52:31	4.104
	Dev	1,556	91	1,465	1:52:25	1:46:04	4.344
	Test	1,585	62	1,523	2:03:36	1:58:29	4.668
7.0	Validated	28,300	1,343	26,957	31:22:32	29:55:26	3.996
	Train	3,810	169	3,641	3:10:00	3:01:39	2.993
	Dev	3,045	164	2,881	3:37:29	3:25:58	4.289
	Test	3,060	149	2,911	3:59:49	3:48:12	4.704


Validated Clips and Duration

Acoustic Model - Baseline

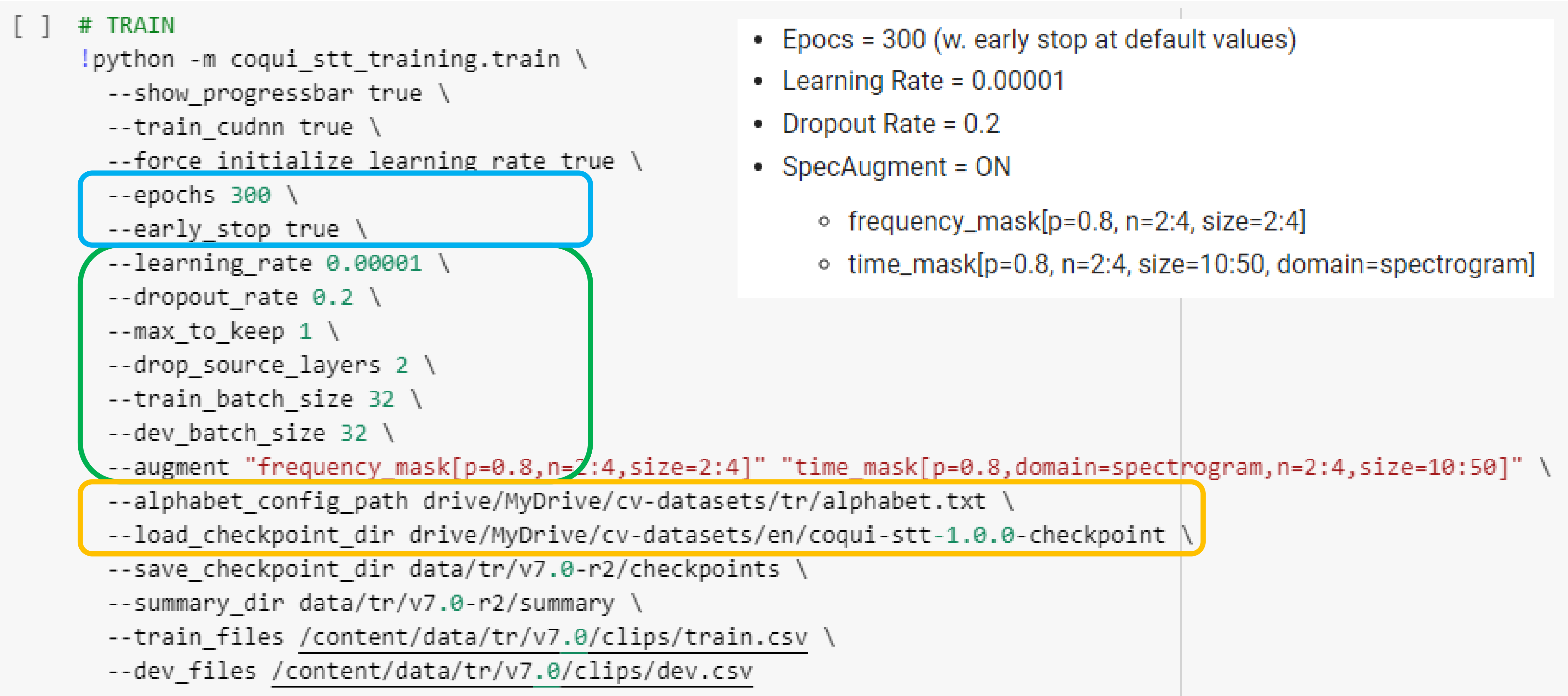
BASELINE: TYERS, MEYER, 2021 - CV v6.1 Turkish Dataset

Run	Batch	Epocs	Spk.	Clips	WER	CER	LOSS
Acoustic Model Baseline	32	25	77	1,739	**95.32**	35.73	54.26
AM w. Parameter Search	32	25	77	1,739	**89.26**	30.84	
AM + Language Model	32	100	77	1,739	**57.19**	29.62	

Common Voice - Coqui 1.0.0 Training Results (default splits, Acoustic Model Baseline)

TRAINING						PROCESS TIME			RESULTS		
Ver	**Run**	**GPU**	**Batch**	**Epocs**	**Best E.**	**T-TRAIN**	**T-TEST**	**T-TOTAL**	**WER**	**CER**	**LOSS**
1	1-BL	K80	32	25	10	0:20:22	0:26:56	0:47:18	0.915615	0.337848	57.549976
3	1-BL	K80	32	25	7	0:31:15	0:45:41	1:16:56	0.903781	0.325954	53.830456
4	1-BL	K80	32	25	9	0:28:36	0:50:31	1:19:07	0.864571	0.288360	46.893242
5.1	**1-BL**	**K80**	**32**	**25**	**9**	**0:31:28**	**0:52:18**	**1:23:46**	**0.864006**	**0.281902**	**43.230740**
6.1	**1-BL**	**K80**	**32**	**25**	**8**	**0:29:31**	**0:48:50**	**1:18:21**	**0.879326**	**0.287137**	**44.182713**
7.0	**1-BL**	**K80**	**32**	**25**	**3**	**0:58:31**	**1:36:56**	**2:35:27**	**0.910997**	**0.325844**	**53.041386**

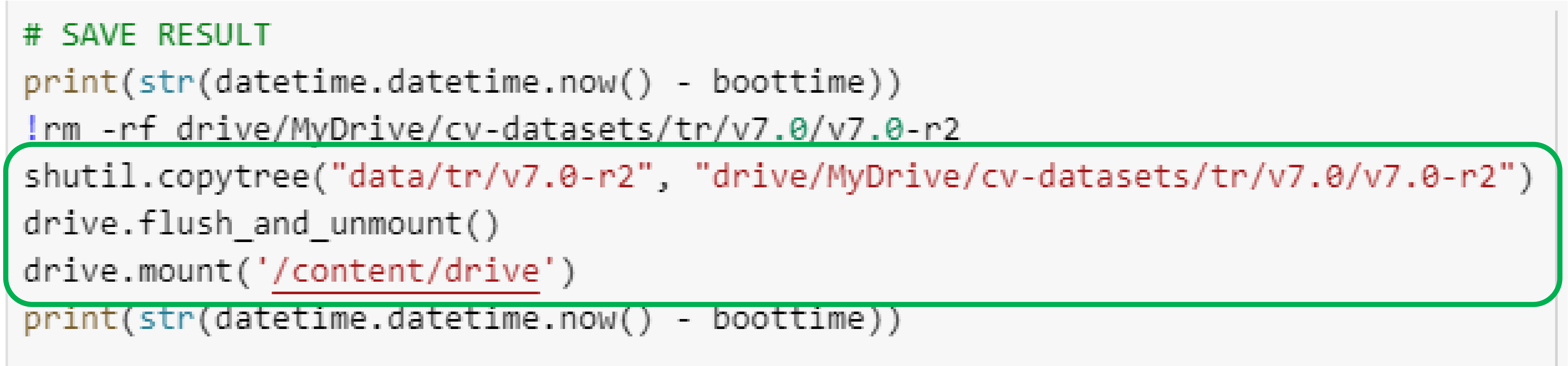
AM Fine Tuning & Augmentation



```
[ ]  # TRAIN
     !python -m coqui_stt_training.train \
       --show_progressbar true \
       --train_cudnn true \
       --force_initialize_learning_rate true \
       --epochs 300 \
       --early_stop true \
       --learning_rate 0.00001 \
       --dropout_rate 0.2 \
       --max_to_keep 1 \
       --drop_source_layers 2 \
       --train_batch_size 32 \
       --dev_batch_size 32 \
       --augment "frequency_mask[p=0.8,n=2:4,size=2:4]" "time_mask[p=0.8,domain=spectrogram,n=2:4,size=10:50]" \
       --alphabet_config_path drive/MyDrive/cv-datasets/tr/alphabet.txt \
       --load_checkpoint_dir drive/MyDrive/cv-datasets/en/coqui-stt-1.0.0-checkpoint \
       --save_checkpoint_dir data/tr/v7.0-r2/checkpoints \
       --summary_dir data/tr/v7.0-r2/summary \
       --train_files /content/data/tr/v7.0/clips/train.csv \
       --dev_files /content/data/tr/v7.0/clips/dev.csv
```

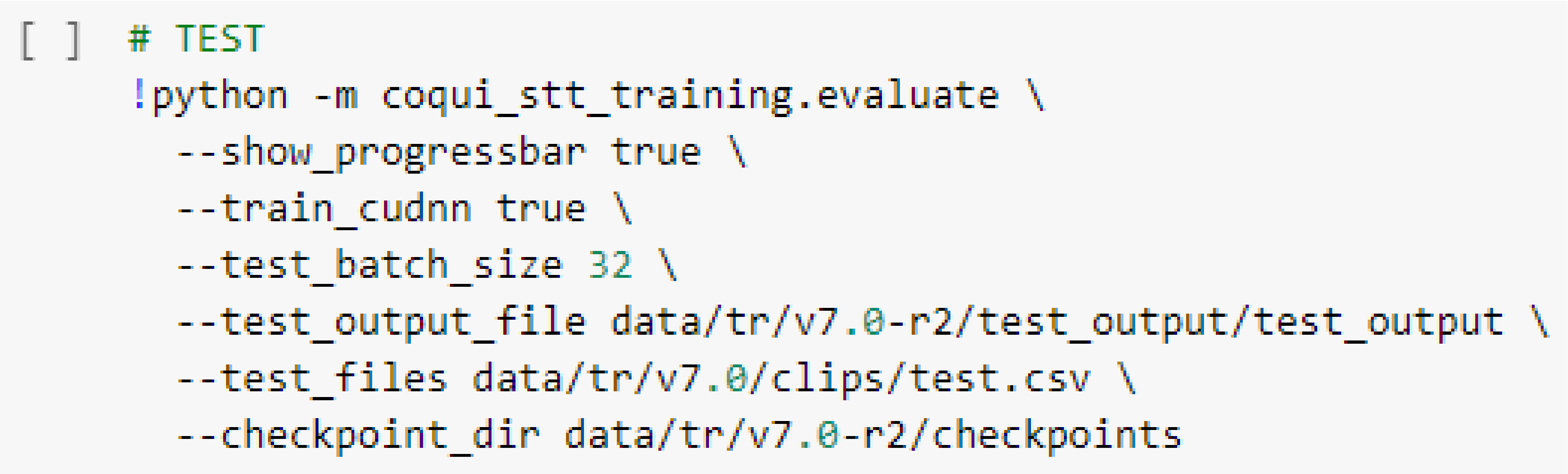
- Epocs = 300 (w. early stop at default values)
- Learning Rate = 0.00001
- Dropout Rate = 0.2
- SpecAugment = ON
  - frequency_mask[p=0.8, n=2:4, size=2:4]
  - time_mask[p=0.8, n=2:4, size=10:50, domain=spectrogram]

Saving intermediate results



```
# SAVE RESULT
print(str(datetime.datetime.now() - boottime))
!rm -rf drive/MyDrive/cv-datasets/tr/v7.0/v7.0-r2
shutil.copytree("data/tr/v7.0-r2", "drive/MyDrive/cv-datasets/tr/v7.0/v7.0-r2")
drive.flush_and_unmount()
drive.mount('/content/drive')
print(str(datetime.datetime.now() - boottime))
```

Test

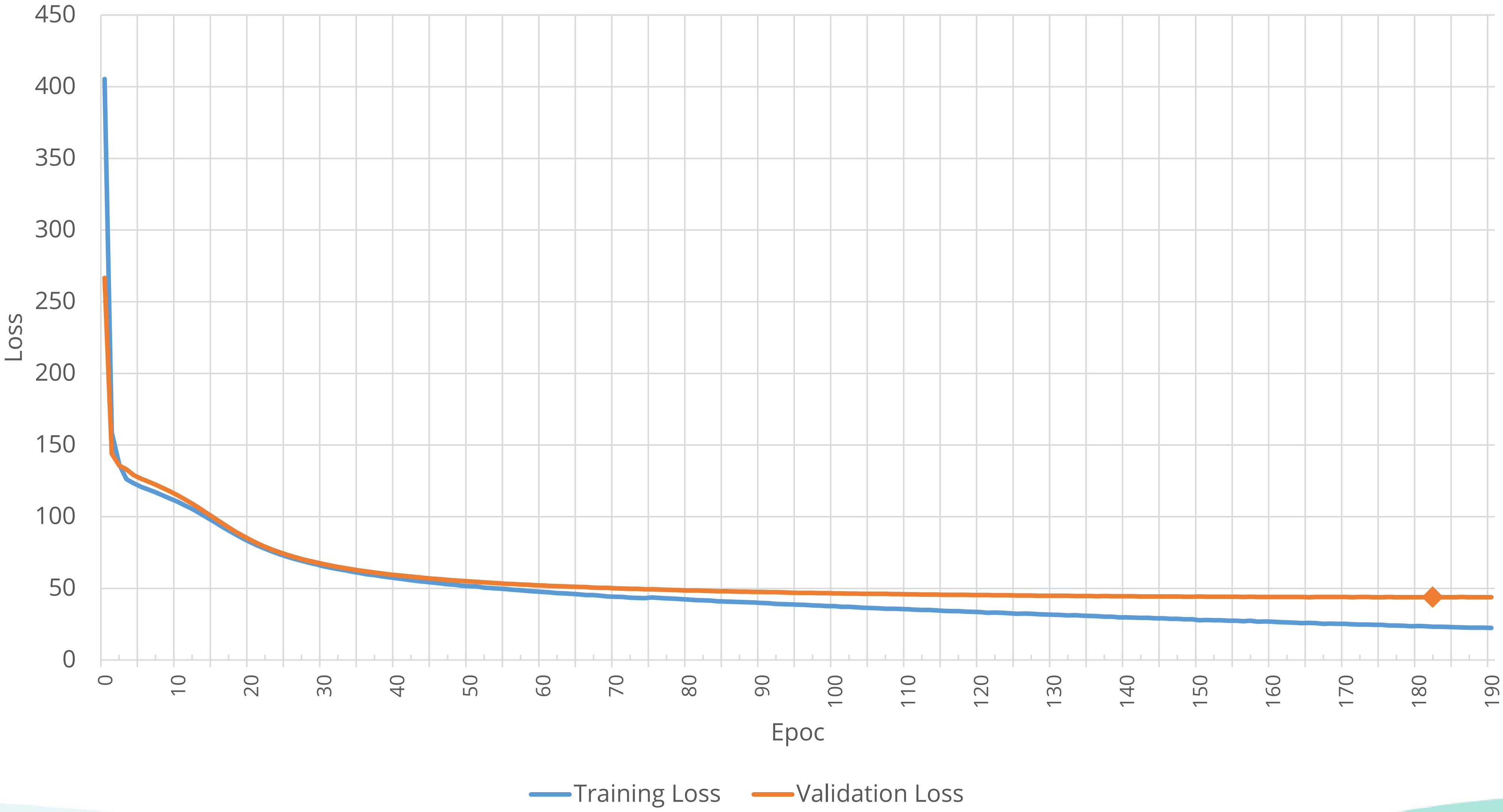


```
# TEST
!python -m coqui_stt_training.evaluate \
  --show_progressbar true \
  --train_cudnn true \
  --test_batch_size 32 \
  --test_output_file data/tr/v7.0-r2/test_output/test_output \
  --test_files data/tr/v7.0/clips/test.csv \
  --checkpoint_dir data/tr/v7.0-r2/checkpoints
```

Common Voice v6.1 Acoustic Model Training

Coqui 1.0.0 Transfer Learning w. batch size 128, LR=0.00001, DR=0.2 + SpecAugment

Google Colab Pro P100 GPU, 300 Epocs w. Early Stop - Best epoc on 182

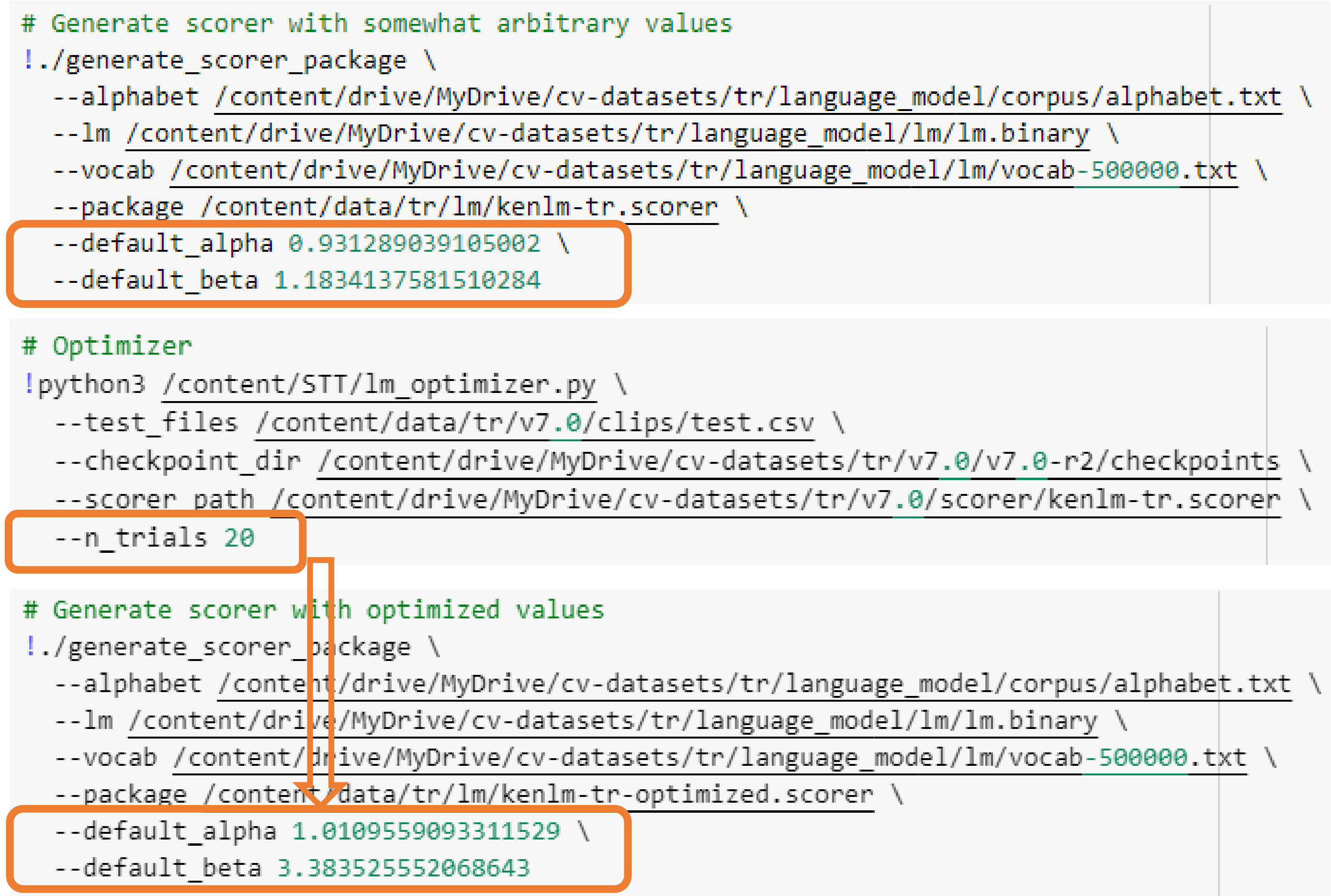


Adding a Language Model

Corpus	Lines	Words	Chars
Global Voices	6,118	75,201	615,984
OpenSubtitles	167,809,335	677,578,088	4,940,414,461
Tatoeba	730,933	3,606,127	26,358,543
TED2013	121,044	1,407,257	10,998,574
TED2020	371,225	3,851,340	30,451,350
TOTAL	**169,038,655**	**686,518,013**	**5,008,838,912**

- Added optimization!

Language Model Optimizer

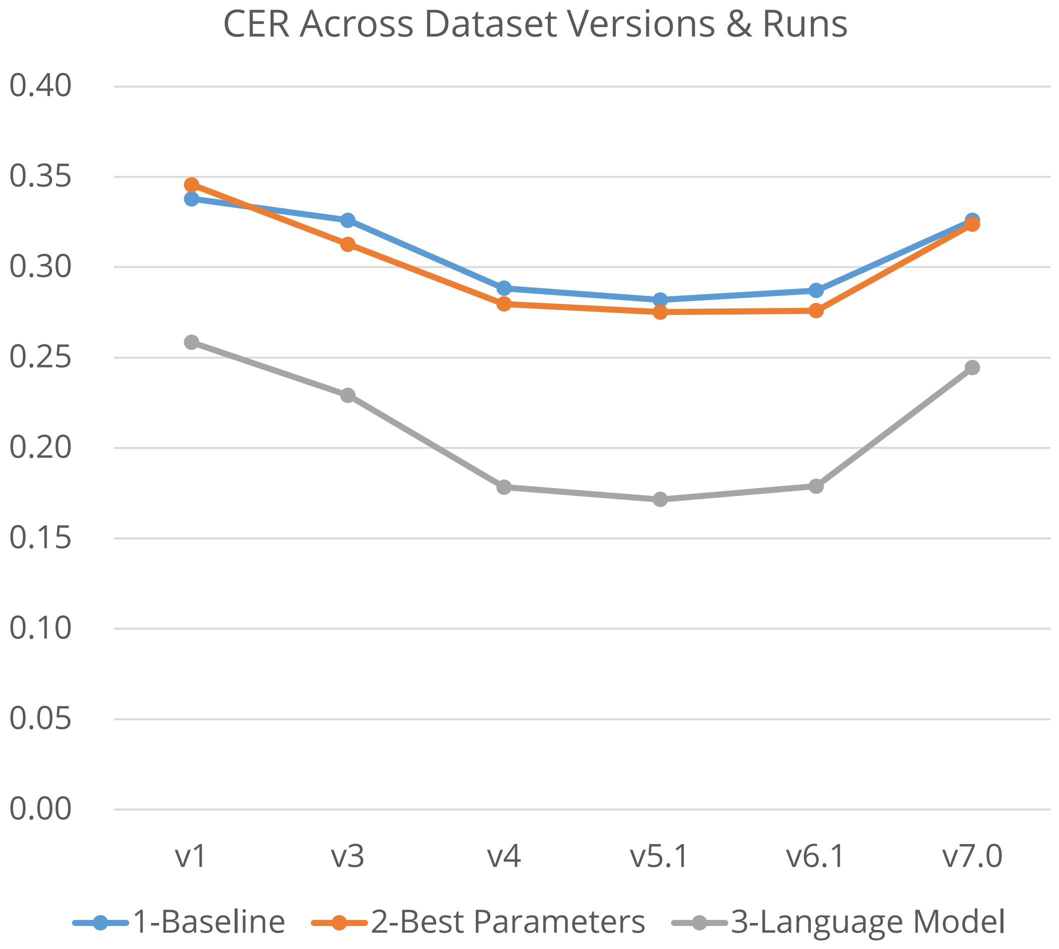


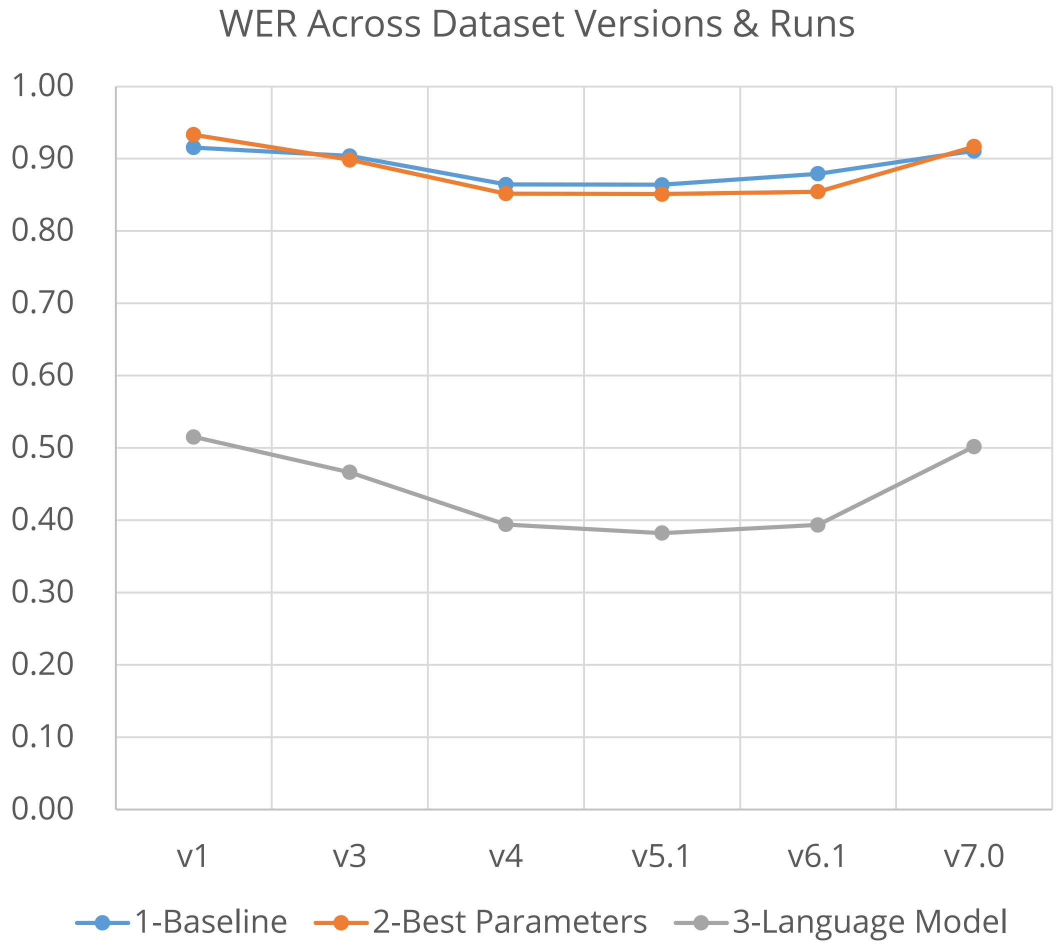
```
# Generate scorer with somewhat arbitrary values
!./generate_scorer_package \
  --alphabet /content/drive/MyDrive/cv-datasets/tr/language_model/corpus/alphabet.txt \
  --lm /content/drive/MyDrive/cv-datasets/tr/language_model/lm/lm.binary \
  --vocab /content/drive/MyDrive/cv-datasets/tr/language_model/lm/vocab-500000.txt \
  --package /content/data/tr/lm/kenlm-tr.scorer \
  --default_alpha 0.931289039105002 \
  --default_beta 1.1834137581510284
```

```
# Optimizer
!python3 /content/STT/lm_optimizer.py \
  --test_files /content/data/tr/v7.0/clips/test.csv \
  --checkpoint_dir /content/drive/MyDrive/cv-datasets/tr/v7.0/v7.0-r2/checkpoints \
  --scorer_path /content/drive/MyDrive/cv-datasets/tr/v7.0/scorer/kenlm-tr.scorer \
  --n_trials 20
```

```
# Generate scorer with optimized values
!./generate_scorer_package \
  --alphabet /content/drive/MyDrive/cv-datasets/tr/language_model/corpus/alphabet.txt \
  --lm /content/drive/MyDrive/cv-datasets/tr/language_model/lm/lm.binary \
  --vocab /content/drive/MyDrive/cv-datasets/tr/language_model/lm/vocab-500000.txt \
  --package /content/data/tr/lm/kenlm-tr-optimized.scorer \
  --default_alpha 1.0109559093311529 \
  --default_beta 3.383525552068643
```

Character & Word Error Rates





Back to Analysis

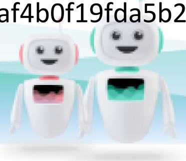
Number of Client IDs				
Ver	VAL	TRAIN	DEV	TEST
1	190	4	21	165
2	323	13	46	264
3	323	13	76	265
4	434	34	76	324
5.1	608	77	128	389
6.1	631	77	128	401
7.0	851	13	99	640

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%

Voices in Train & Dev

	Row Labels	Count
1	c3c204ffaebfc46c0265773376f9288a372694aa79f97afe224828033c28d6d2a90919aaf769bf14105cb4033650e10275760afe250490782eae15c1d1518799	2236
2	6c8743f86a41b2e1dd119942cf21633476a6185e3456098a6d4ad6c7849efb62733eb1c4c65040cae1fcaa26b5db66402cb3895c434b4f8ad4c911cecf5c7db7	579
3	9e752d5e672cb444e46093900db35c7ce913577ca5ee37202bd9e623ae47c00fcd8326862e5501836bd533857cef68cd6f904a80ccca3b392615783937dded32	358
4	4ea7534da9594c537caa422c93f0fc2e0fa0caaf75be3ae768de1369bb06765f246aa51e9161b7f0a9f4e21dffb1ceddfe27be1c81bcfed41343d5c3dc5d69d1	352
5	dde1f0f45fb4024cd66cebcc97f398a371be0c50f6133e4abebd240702b9fc9a2204bc2b461f9029fce7362f641de4105b847cba49c45e4b46ae49bc2d4a7b45	211
6	60cee2235d7ec4cdeb89d601b8c373955b303c712ec729ed0affdabda8819908f51cefa990163d2bc4ac04c93e6dac2909cca67829211df4a2a17af2507dd50a	142
7	25fdcdb28f13d4a31842aa3edb4da72e999080192e12df168123ac1012c88a96cc7db925618282f6d6c3e86d77ffcc4cb82b1ae7fa3e6791a0f6bc9277d15b63	51
8	33f649d48a5122b434291c85f109d9d7dac4d0486439d441a0b2979e54397bbd6e15c7515d44c5754377a83c9a1a88a3715d3382455f9150265ebafc8dba819d	16
9	6d78740a48bccbfa2f70f6a10bb81e57ceb29ade055e8dbac9b87d5bb1f4512b1df070f0b0d4f93f38328fd181bf229b47eb933690ec03cc12c39947dd740ceb	16
10	94c71f5af8e0dc3a94075137a518bb5fbf75c4818c2018610c6cefc0fcd62a3f2543727e74b0472683f11edd72d2032700479aaccc112c14f88395e5d3a1febc	7
11	6dcca4df178f49130349c5756eee44028f4eb99937a60117c6629b8613032383c092d77433b00df7df887aed05aa734ef858e48837c95fcde9790725acdf8bcb	4
12	0e03e4f6370c3e2952a880f4ffb50c08ce5932573d1bcb0bb83faeccb12cfdf420b7b37dbec04add639ede64c985d19652de332b31ad3ef892ab28678377f411	4
13	97ebc8329a0e86083942f65713a30826222ed10ab3db78f79979a2d9208865f678ddc297fef3225a6db99199aead9248a142530e3fde4915687765efd042a3d8	3
	Total	**3979**

	Row Labels	Count
1	c3c204ffaebfc46c0265773376f9288a372694aa79f97afe224828033c28d6d2a90919aaf769bf14105cb4033650e10275760afe250490782eae15c1d1518799	2417
2	e93011699a08bdc6a27a7b50da5a6ed312322138c10135bf7d2d1aa8316373c10aff0f2cf8ef038990eb18e0b3c989ee37591f7c7deea0041498d22fbc4c9432	427
3	f76ed0a17ffdb06fa37e68bad991b7eef235cdb2686fcec0d9a76a668c1ef1de13dadea046bfc0e30a69c8da3aa9e85cf5d0694a2e92d02dfcbf3bcbae6c725c	278
4	0178454bf2224e584e03086ed55af64e4286e8122f2f3444b44538d96c40a8d0939eda91f94fb665c388ab3aa061b334ffd41e0f31710f3e483622bcef5c28ee	121
5	0042924c9d0d1c87001e1720a1dbb8ef3954e4e3ca0fab2adba81d37f5d213f1a81ac8f8bbe95078b065cffd7c295453842629dcd9b742341874ffcce2f2484c	116
6	b168ff93eb790968fbe744a61f52432e011da19067af41ed27f274ea18f6e7b6a004db53087717c82728691e566520c807a1c42317b65977680226312c2c4087	104
7	bd2fe2a027fec3faee255825d444737536b1b52ec14c4adb01b4c30950b1d7d7ff07706629dafe91f73dc09e641b407890efc8411fbe645f7dbe2fd896c087aa	100
8	be57ee50010b2c7f108b81ac1e20c2cfde966a257d91b4df4afce5898c990c96045330627c9543fd988173966b1e1aead01e947211f35c3d75e5ca36cb564e96	69
9	3310fc5e0414c88fa9bd95366867748cd7a4d6ea0d3472a98c692a1e0f5a22d4fd1f4f4b935612950a4932fdd5b72f69a033d9e1126b4fd6f3ba17daf53cc6df	60
10	91e73150f71f02cc95d37175dbaad1580e9af4b0f19fda5b26c3fcbc1a4130dd63a6bda869349b53319c6747111b97bd9b6c6ce2982e36c7127c75d9c0128d45	58



Back to Corpora Creator

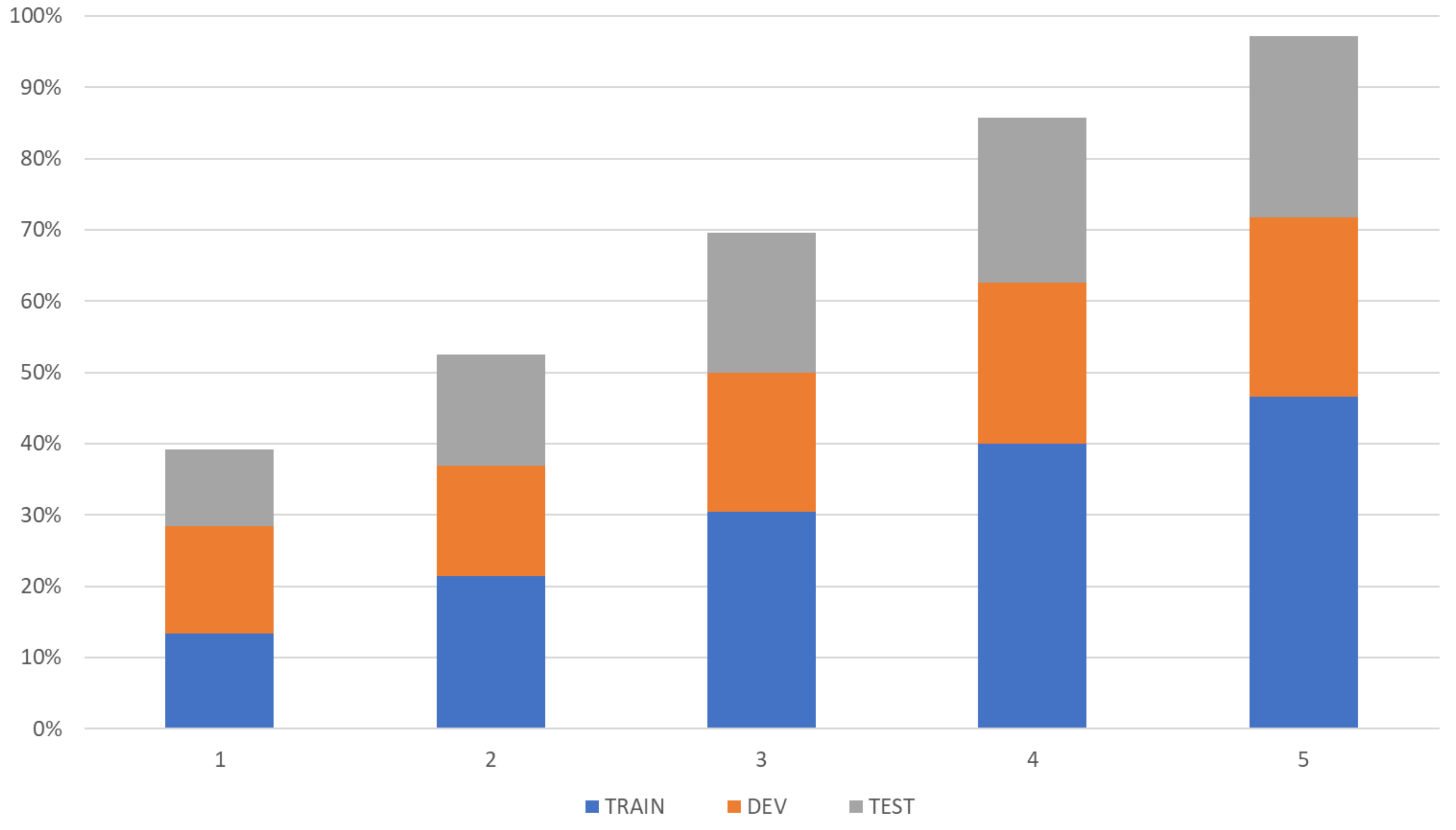
- --strict-speaker: One speaker only lives in one file => ON
- --strict-sentence: One sentence only lives in one file => ON
- --strict-audio: Only a single recording per sentence => OFF

We ran v6.1 & v7.0 with **–s 1** through **–s 5**

Coqui 1.1.0

We used batch size of 128 this time (~35% less time)

CV Turkish Dataset v7.0 Alternative Splits
Splits' Percentage if we take Multiple Recordings per Sentence

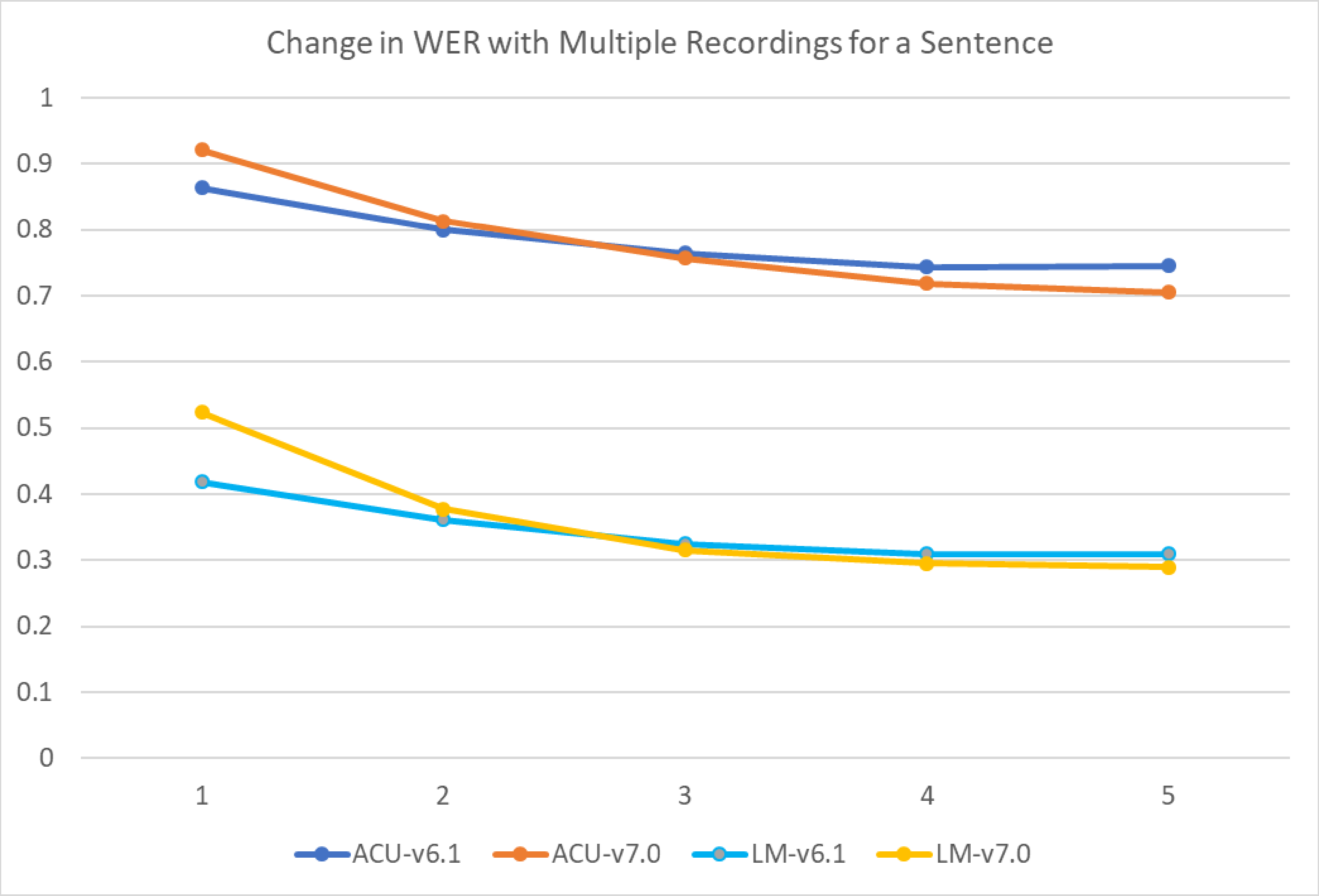


For v7.0...

Number of Client IDs				
-s	VAL	TRAIN	DEV	TEST
1	851	13	99	640
2	851	43	119	676
3	851	37	91	714
4	851	29	75	740
5	851	23	68	755

Gender - Female / Male Ratio				
-s	VAL	TRAIN	DEV	TEST
1	8.99%	0.00%	6.87%	10.29%
2	8.99%	10.28%	11.49%	8.23%
3	8.99%	9.39%	17.15%	7.32%
4	8.99%	3.22%	27.43%	8.82%
5	8.99%	2.96%	20.00%	12.69%

Results (Multiple Recordings/Sentence)

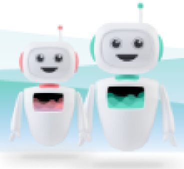


Remaining questions

- Will it be biased to Balkan News?

- We know the following is best for 100.000 recordings:
  - 1000 people 100 recordings/person
- But if you only have 100 volunteers? Which is better?
  - 100 people 100 recordings/person?
  - 100 people 1000 recordings/person?

- How many recordings per sentence is best?
  - Limited text corpus
  - Different voices speaking with different accents



Final words

Future Work

- Train next dataset

- More volunteers
- More text-corpus
- More voice-corpus

- Expand language model
- Measure real-world model performance
- Limited vocabulary models/applications
- Measure bias
- Measure effect of many recordings/person
- Other libraries/models: Transformers, Wav2Vec, etc

Acknowledgements

Common Voice is a great project!

Thank you Common Voice!

Thank you community!

- Special thanks to
  - Francis Morton Tyers
  - Hillary Juma
  - Michael Kohler

- CV Turkish Core
  - Tuğçe
  - Dilek
  - Mansur

- All CV Turkish Volunteers!

Thank you

Questions & Answers

But before, Common Voice January 2022 Dataset Results