PRODUKTION AF
DANSKE TALEDATA

Sprogteknologisk konference
30-11-22

ALEXANDRA INSTITUTTET

Alexandra Instituttet A/S

er non-profit og et af de

syv Danske **GTS** instituttet

ALEXANDRA
INSTITUTTET

Alexandra Instituttet A/S

er non-profit og et af de

syv Danske **GTS** instituttet

Vi er **100% ejet** af Aarhus

Universitets Forskningsfond

Alexandra Instituttet A/S

er non-profit og et af de

syv Danske **GTS** instituttet

ALEXANDRA INSTITUTTET

Vi er **100% ejet** af Aarhus

Universitets Forskningsfond

AARHUS UNIVERSITETS FORSKNINGSFOND

Vi er ca. **80 IT konsulenter** i

Aarhus (IT-byen Katrinebjerg)

og **København** (ITU)

Institut for Datalogi
Aarhus Universitet

IT-Universitetet i
København

ANVENDT FORSKNING

Roskilde University

Danmarks Tekniske Universitet

DTU

AARHUS UNIVERSITET

SDU

Syddansk Universitet

KØBENHAVNS UNIVERSITET

AALBORG UNIVERSITET

IT-UNIVERSITETET I KØBENHAVN

ANVENDT FORSKNING

Roskilde University

Danmarks Tekniske Universitet

AARHUS UNIVERSITET

SDU Syddansk Universitet

KØBENHAVNS UNIVERSITET

AALBORG UNIVERSITET

IT-UNIVERSITETET I KØBENHAVN

KOMMERCIELLE AKTIVITETER

GRUNDFOS

novo nordisk®

LEGO®

Topdanmark Forsikring • Pension

Sund≋Bælt

ERHVERVSSTYRELSEN

ANVENDT FORSKNING

KOMMERCIELLE AKTIVITETER

Roskilde University

Danmarks Tekniske Universitet

AARHUS UNIVERSITET

SDU Syddansk Universitet

KØBENHAVNS UNIVERSITET

AALBORG UNIVERSITET

IT-UNIVERSITETET I KØBENHAVN

GRUNDFOS

novo nordisk

LEGO

Topdanmark Forsikring • Pension

Sund≋Bælt

ERHVERVSSTYRELSEN

DaNLP er en NLP **platform**, med fokus på det danske sprog

1. Netværk for Data Scientists og virksomheder
2. Open source datasæt og AI modeller
   github.com/alexandrainst/danlp
3. Vidensdeling og inspiration
   danlp.alexandra.dk

Hvorfor danske taledata?

UDFORDRINGER

Omkostningsfulde at producere

Dansk er "fragmentarisk" understøttet

↓ ↓

KONSEKVENSER

Høj barriere for at komme igang

Mange dialekter og sprogstile understøttes ikke

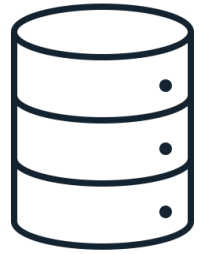
EFFEKTER

Danske virksomheders konkurrenceevne vil blive yderligere udfordret i takt med at udviklingen på de store sprog accelererer.

Danske forbrugere og borgere får ikke gavn af taleteknologiske services (til trods for at Danmark er et af de mest digitaliserede lande).

MÅL

Producere et åbent dansk taledatasæt i høj kvalitet der kan bruges til udvikling af taleteknologiske services



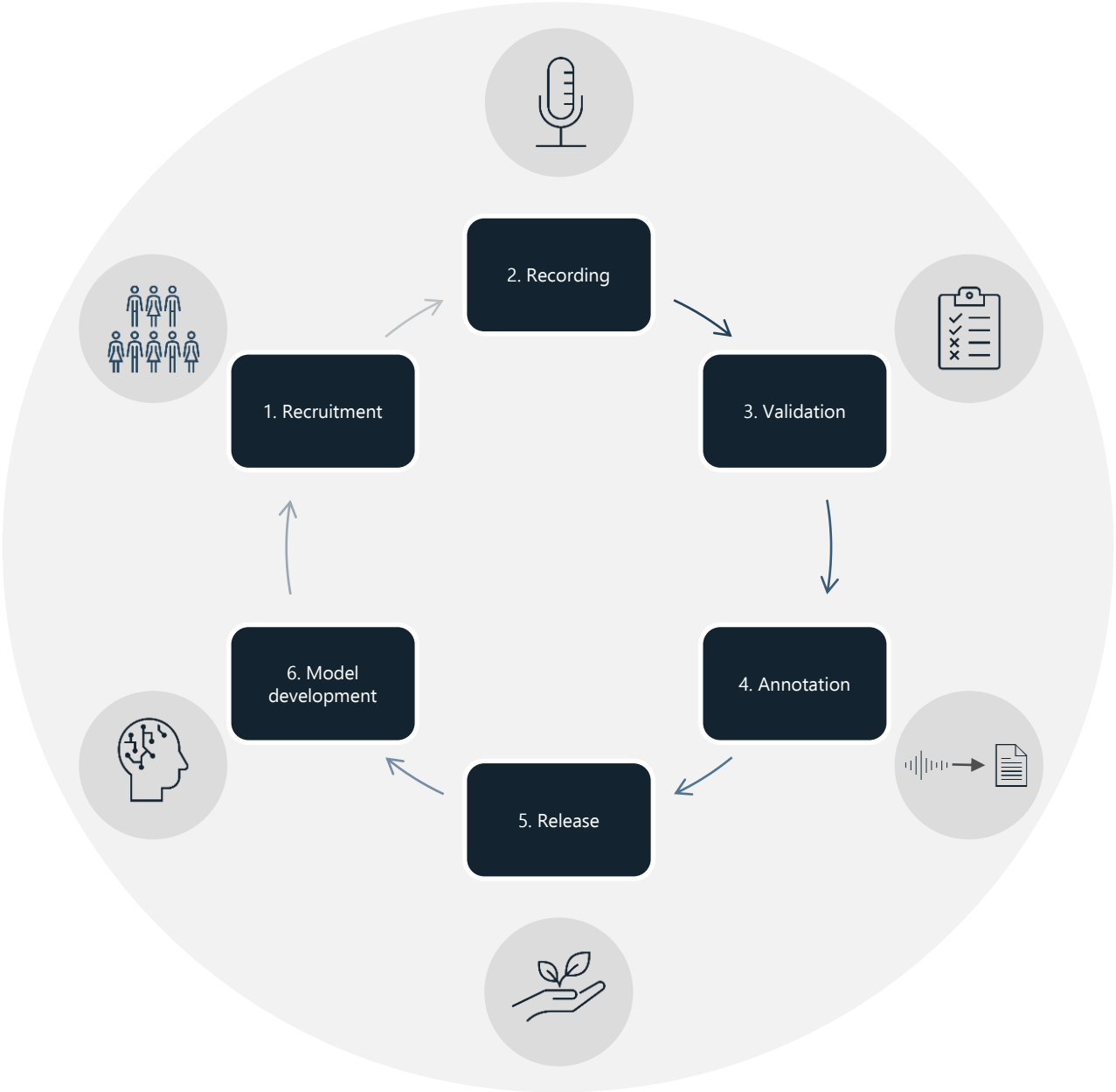
+1500 timer

- **Samtaler**
- **Diversitet**
- **Annotering**

OPEN SOURCE

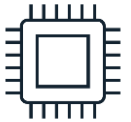
PIPELINE FOR PRODUKTION AF TALEDATA

PRELIMNARY PLAN



FAIR SPEECH DATA

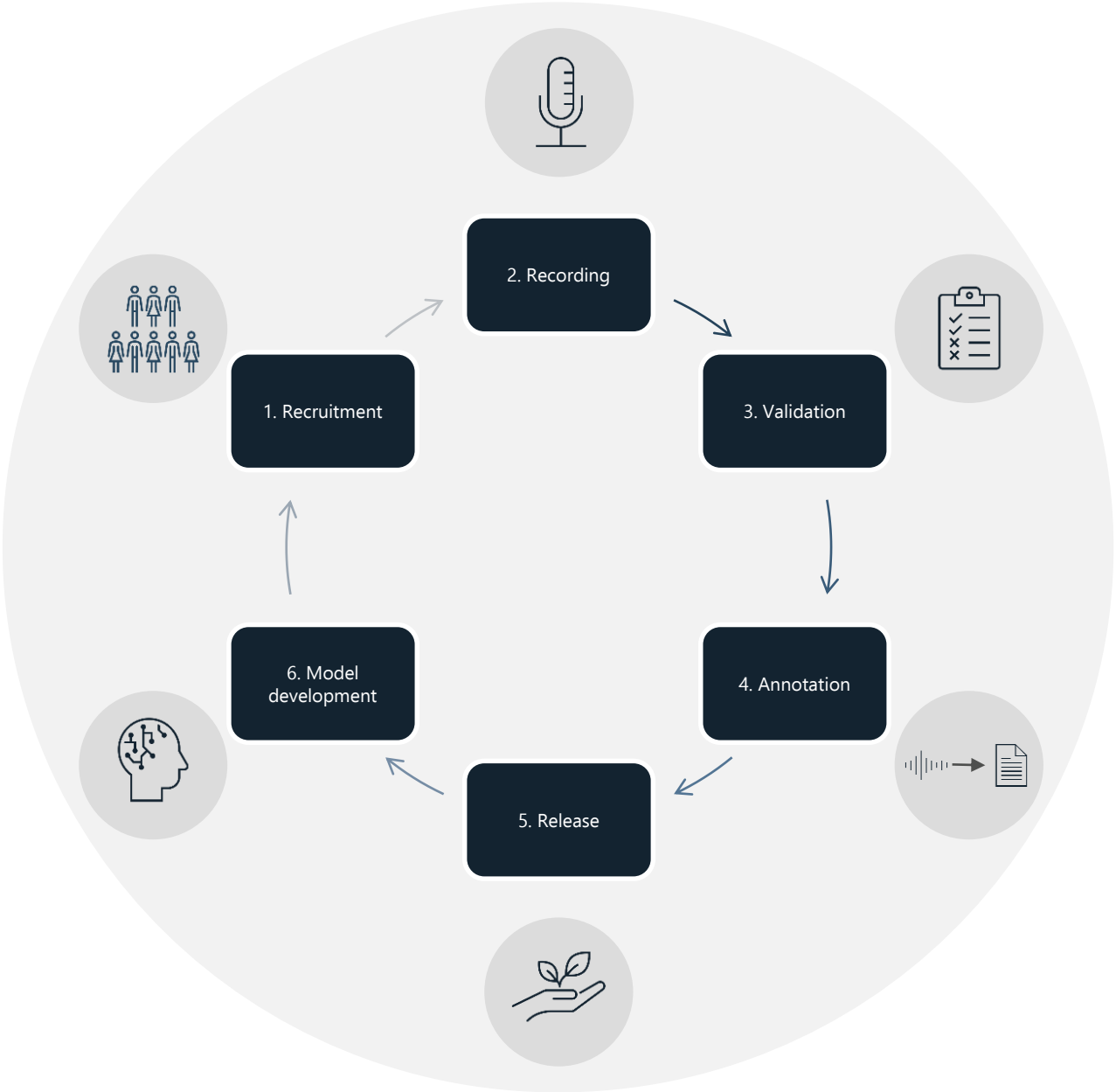
SPEECH MODELS



ONLINE RESOURCES

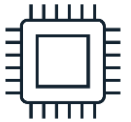
PIPELINE FOR PRODUKTION AF TALEDATA

PRELIMNARY PLAN



FAIR SPEECH DATA

SPEECH MODELS

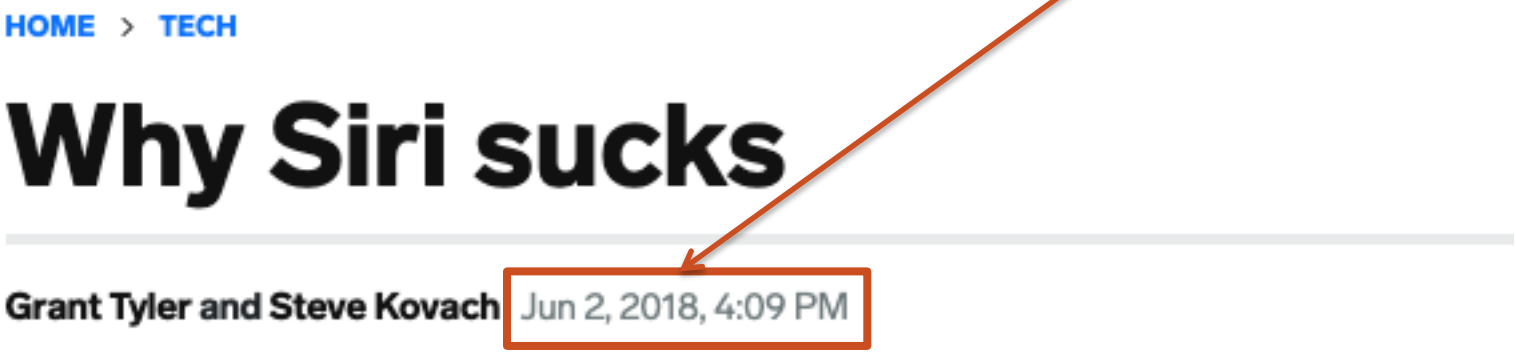


ONLINE RESOURCES

The Alexandra Institute
and
the Talk of Not-So-Fun-Facts

Why do we need more data?

Not-so-fun-fact: We have not become so much better since then…



Why do we need more data?

LibriSpeech ASR corpus

Identifier: SLR12
Summary: Large-scale (1000 hours) corpus of read English speech
Category: Speech
License: CC BY 4.0
Downloads (use a mirror closer to you):
dev-clean.tar.gz [337M] (development set, "clean" speech) Mirrors: [US] [EU] [CN]
dev-other.tar.gz [314M] (development set, "other", more challenging, speech) Mirrors: [US] [EU] [CN]
test-clean.tar.gz [346M] (test set, "clean" speech) Mirrors: [US] [EU] [CN]
test-other.tar.gz [328M] (test set, "other" speech) Mirrors: [US] [EU] [CN]
train-clean-100.tar.gz [6.3G] (training set of 100 hours "clean" speech) Mirrors: [US] [EU] [CN]
train-clean-360.tar.gz [23G] (training set of 360 hours "clean" speech) Mirrors: [US] [EU] [CN]
train-other-500.tar.gz [30G] (training set of 500 hours "other" speech) Mirrors: [US] [EU] [CN]
intro-disclaimers.tar.gz [695M] (extracted LibriVox announcements for some of the speakers) Mirrors: [US] [EU] [CN]
original-mp3.tar.gz [87G] (LibriVox mp3 files, from which corpus' audio was extracted) Mirrors: [US] [EU] [CN]
original-books.tar.gz [297M] (Project Gutenberg texts, against which the audio in the corpus was aligned) Mirrors: [US] [EU] [CN]
raw-metadata.tar.gz [33M] (Some extra meta-data produced during the creation of the corpus) Mirrors: [US] [EU] [CN]
md5sum.txt [600 bytes] (MD5 checksums for the archive files) Mirrors: [US] [EU] [CN]

About this resource:
LibriSpeech is a corpus of approximately 1000 hours of 16kHz read English speech, prepared by Vassil Panayotov with the assistance of Daniel Povey. The data is derived from read audiobooks from the LibriVox project, and has been carefully segmented and aligned.

Acoustic models, trained on this data set, are available at kaldi-asr.org and language models, suitable for evaluation can be found at http://www.openslr.org/11/.

For more information, see the paper "LibriSpeech: an ASR corpus based on public domain audio books", Vassil Panayotov, Guoguo Chen, Daniel Povey and Sanjeev Khudanpur, ICASSP 2015 (submitted) (pdf)

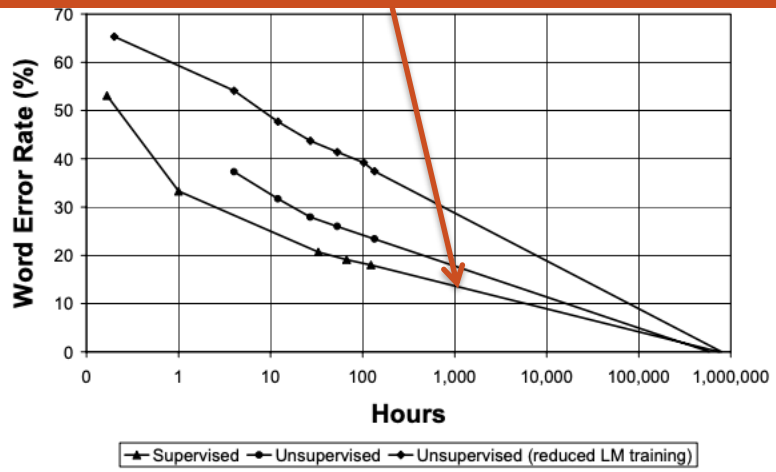
Why do we need more data?

LibriSpeech ASR corpus

Identifier: SLR12
Summary: Large-scale (1000 hours) corpus of read English speech
Category: Speech
License: CC BY 4.0

	WER	Annotated data (Hours)	Unannotated (Hours)
Zhang et al. (2021) (BigSSL)	**1.4%**	34,000	**~1,000,000**
Radford et al. (2022) (Whisper)	2.7%	**680,000**	NA
Panayotov et al. (2015)	5.5%	*1000*	*60,000*
Amodei et al. (2015) (Humans)	5.8%	A lot	***Even more***

Not-so-fun-fact: Danish is hovering around this point

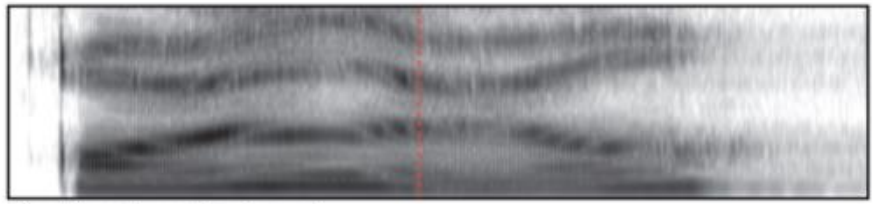


Moore (2003): Extrapolated word error rates for increasing quantities of training data.

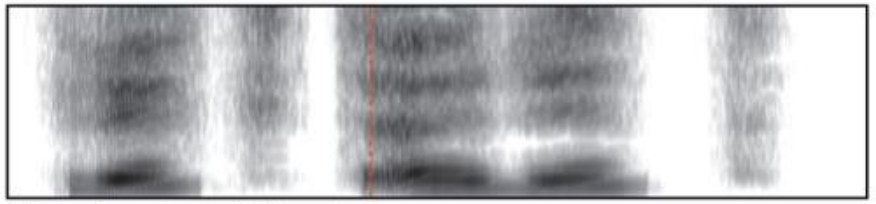
Why do we need more data?

Danish is goddamn hard!

Pronunciation



Danish sentence: Røget ørred



Norwegian sentence: Røkt ørret

State of Danish NLP. (Source: DDSA Conference 2022, Kenneth Enevoldsen, Lasse Hansen. Image: Trecca et al. (2018))

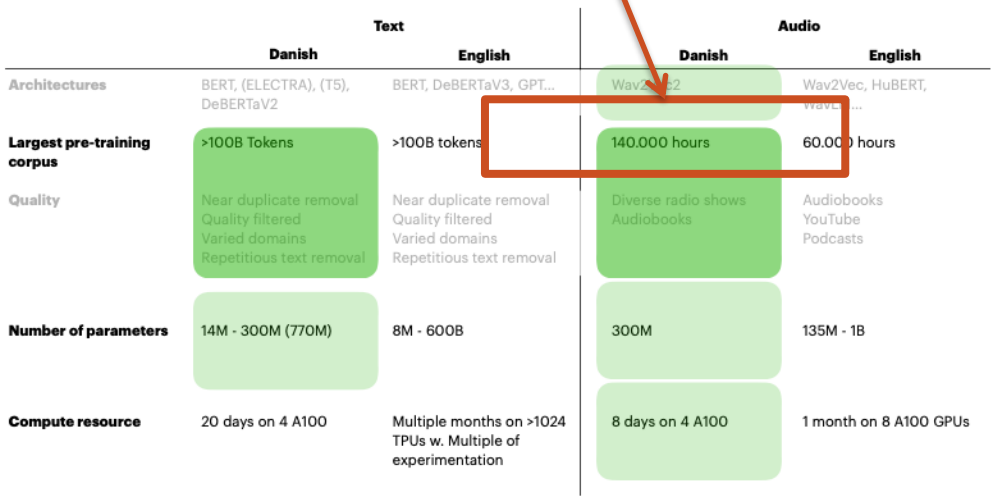
	NST (Norwegian)	NST (Danish)
XLR pretrained	**WER ~3%**	**WER ~20%**

Not-so-fun-fact: Its not just hard for toddlers…

We are lacking behind on open data

Name	Hours	Problems
FTSpeech	1800h	• **License**
NST	390h	
DanPass	9h 46min	• **License**

Not-so-fun-fact: Not open for everybody. Actually almost closed to everyone

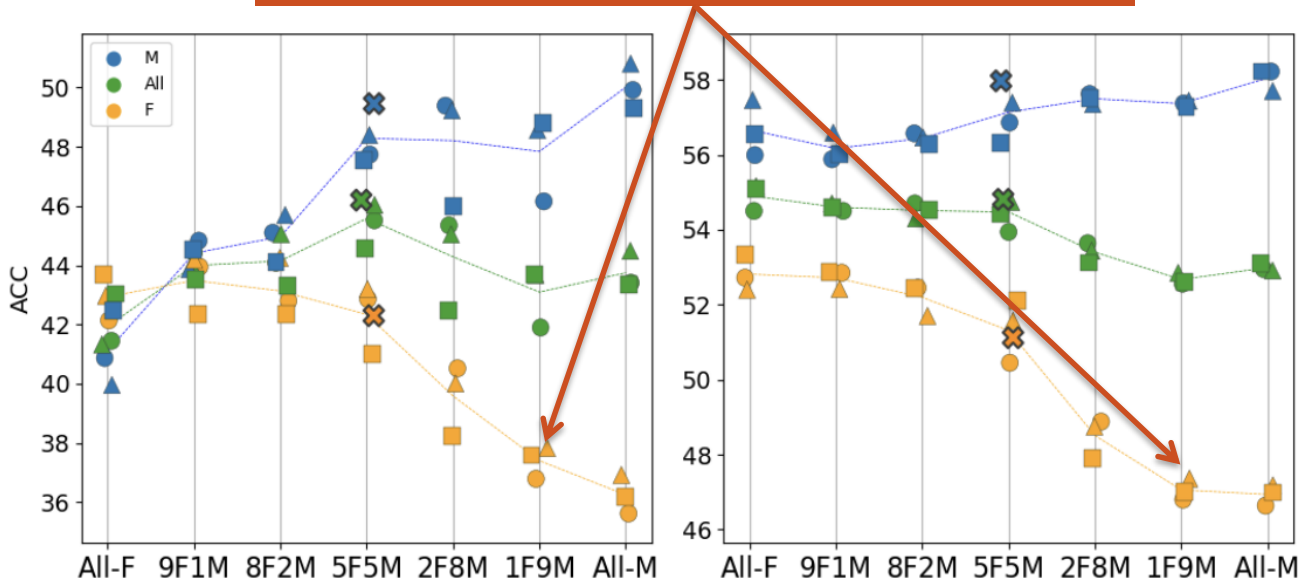


State of Danish NLP. (Source: DDSA Conference 2022, Kenneth Enevoldsen, Lasse Hansen)

We are lacking behind on data with gender equality

Name	Hours	Problems
FTSpeech	1800h	• License • **Gender**
NST	390h	• **Gender**
DanPass	9h 46min	• License • **Gender**

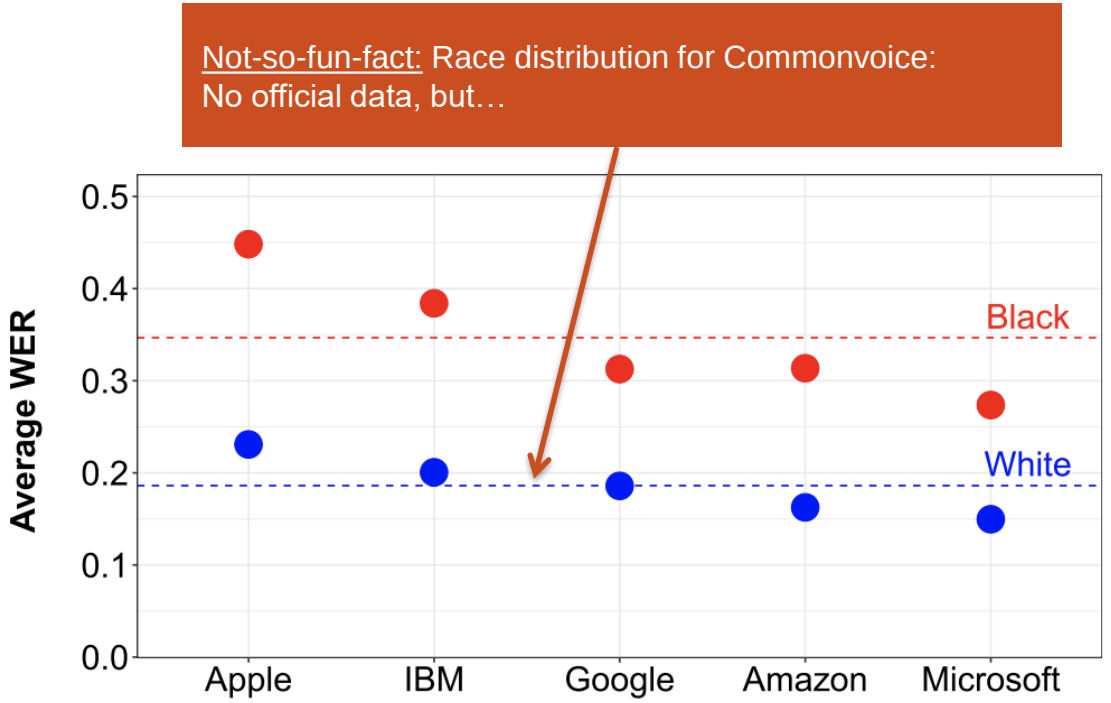
Not-so-fun-fact: Gender distribution for Commonvoice:
Men: 90%
Females: 10%



Meng et al. (2022): Self-supervised Speech models, pretrained on different gender distribuitions, and evaluated on data with different gender distributions. Task: Speaker identification, metric: accuracy.

We are lacking behind on ethnically varied data

Name	Hours	Problems
FTSpeech	1800h	• License • Gender • **Race**
NST	390h	• Gender • **Race**
DanPass	9h 46min	• License • Gender • **Race**



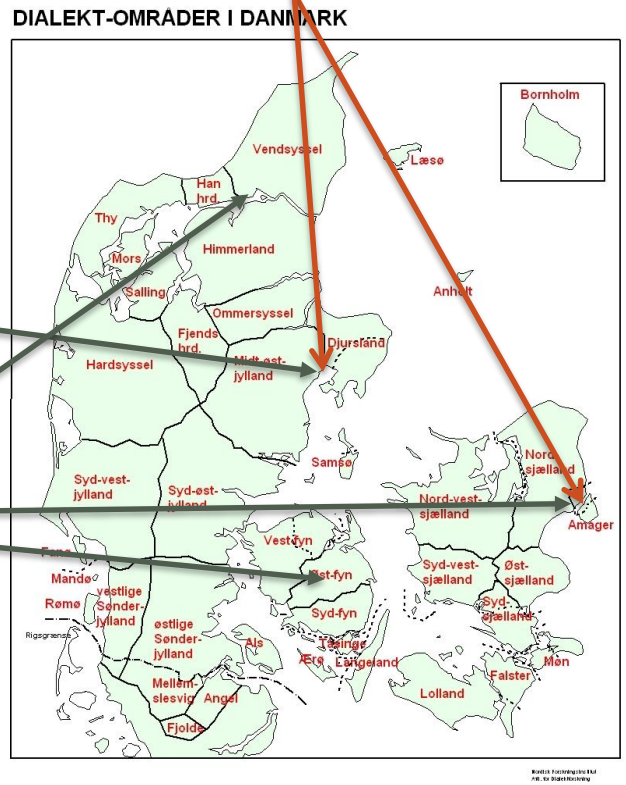
Tatman et al. (2017) - Average WER for different ASR services as a function of race.

We are lacking behind on data which represents different dialects

Name	Hours	Problems
FTSpeech	1800h	• License • Gender • Race • **Dialects**
NST	390h	• Gender • Race • **Dialects**
DanPass	9h 46min	• License • Gender • Race • **Dialects**

Not-so-fun-fact: Dialect distribution for Commonvoice:
No official data, but…

Region	Population
Midt-Øst-Jylland, Fjends hrd., Samsø, Syd-Øst-Jylland	764738
Vendsyssel, Han hrd., Thy, Mors, Salling, Himmerland, Ommersyssel	578839
Nordsjælland	465887
Østfyn, Sydfyn	412427
...	...
Bornholm	39499
Langeland	12367
Møn	9226
Tåsinge	6146
Ærø	5951
Fanø, Mandø, Rømø	3300
Læsø	1807
Anholt	136

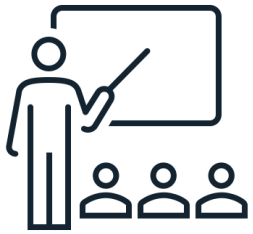


Number of danish dialects: ~32. (source: dialekt.dk)

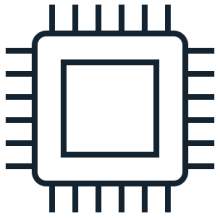
So what do we need?

FAIR SPEECH DATA

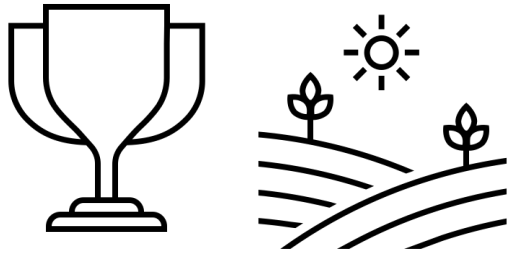
ONLINE RESOURCES



SPEECH MODELS



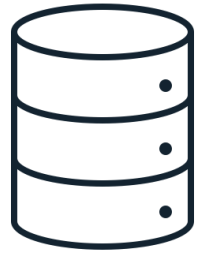
BUSINESS POTENTIAL
&
MORE FAIR SOCIETY



Not-so-fun-fact: All of these steps are time-consuming, hard and expensive

MÅL

Producere et åbent dansk taledatasæt i høj kvalitet der kan bruges til udvikling af taleteknologiske services



+1500 timer

- **Samtaler**
- **Diversitet**
- **Annotering**

OPEN SOURCE

THANK YOU
FOR YOUR TIME

Anders Jess Pedersen · anders.j.pedersen@alexandra.dk

Kasper Fænø Bay Noer · kasper.noer@alexandra.dk