

# Ský hádegisfyrirlestur Vélnám

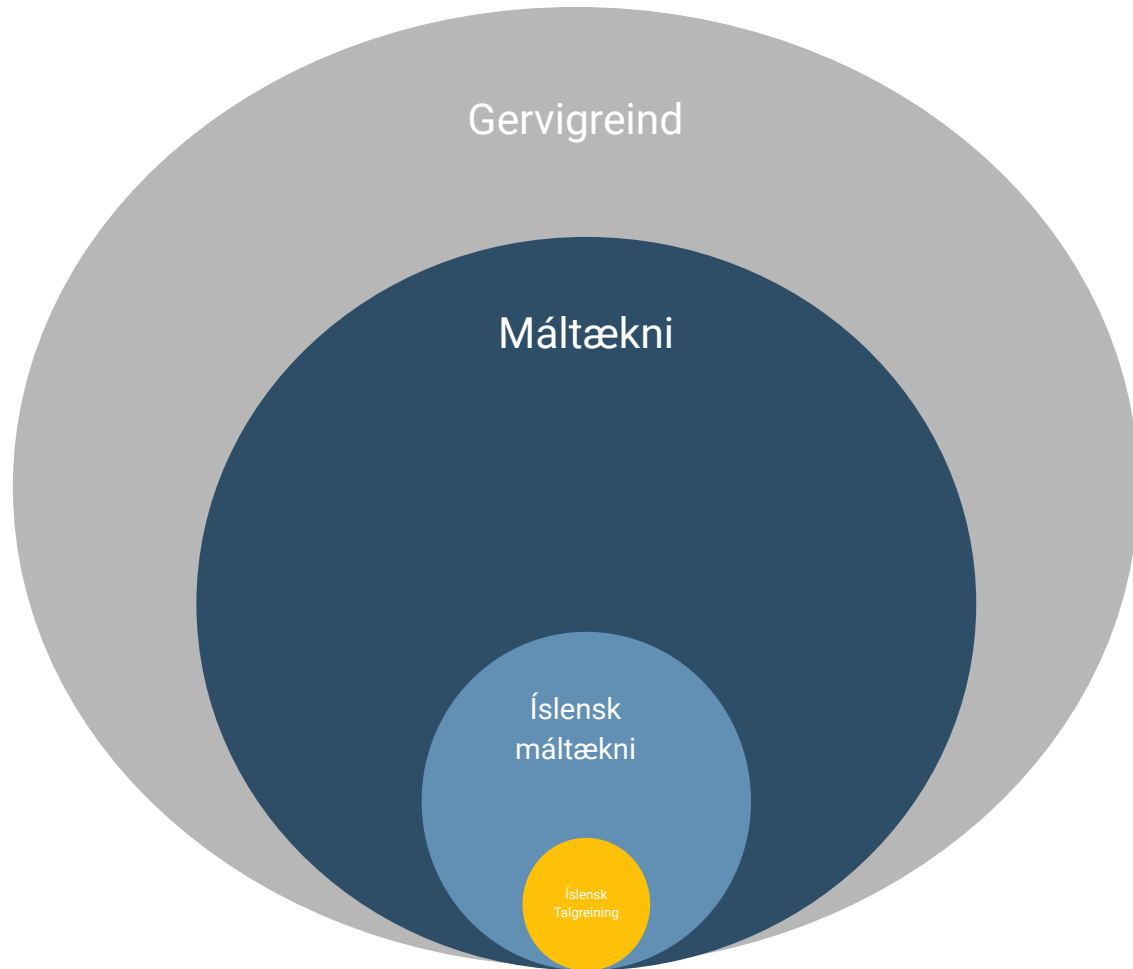
19.apríl 2023



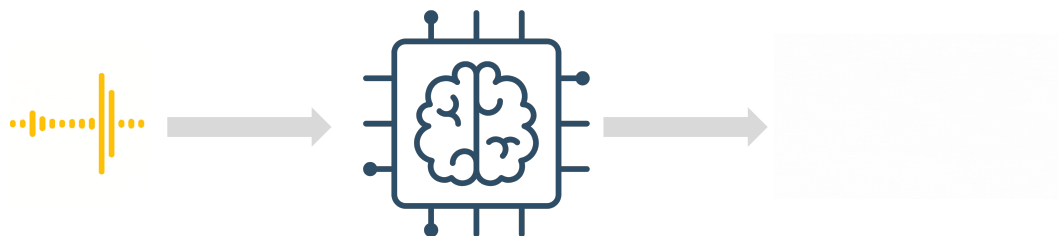
# Yfirlit

- Talgreining 101
- Hvað er hægt að gera við talgögn.
- Hvernig er talgreining að nýtast í dag á íslenskum markaði.





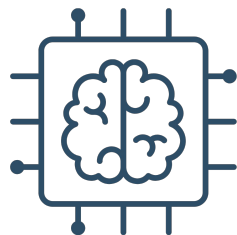
# Hvað er talgreining?



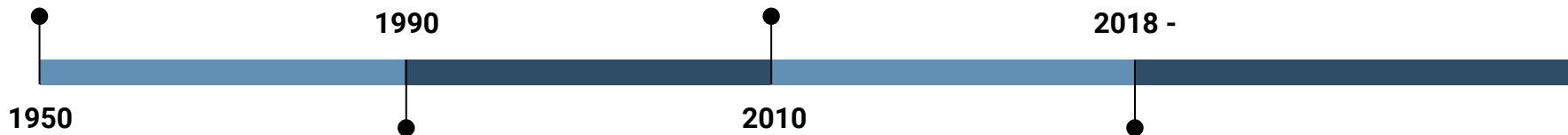
# Hvað er talgreining?



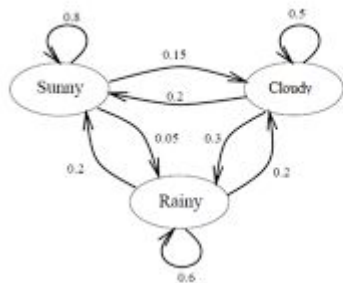
IBM shoebox



Djúp tauganet



HMM - GMM líkön



Transformer líkön



# Talgreining 101



# Talgreining 101

- Máltækniáætlun Stjórnvalda.
- Íslensk talgagnasöfn.
  - ~3000 klukkustundir.
  - Samrómur.
  - Alþingi.

← Velja fjölda 0 / 10

Á Vísindavefnum er að finna mörg önnur svör um krabbamein.

< 1 2 3 4 5 >

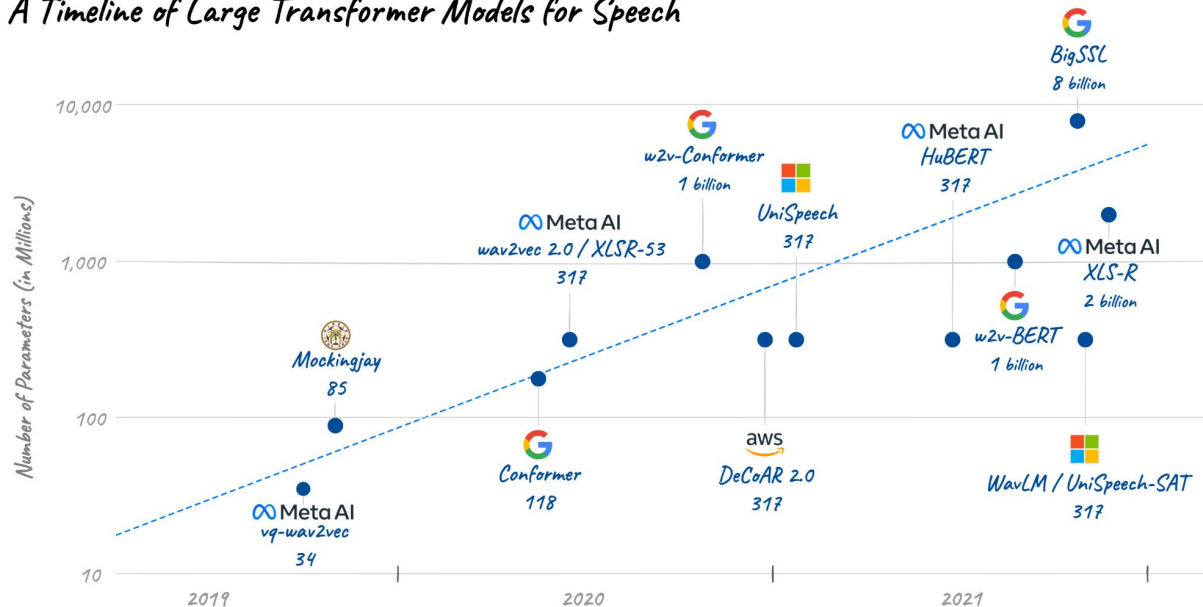
Smelltu á hljóðnemann og lestu setninguna upp



# Gagnagnótt

- GMM-líkön 10-100 klst.
- Fyrstu tauganetin 1000-3000 klst.
- Wav2vec2.0 54.000 klst.
- XLS-R 500.000 klst.
- Whisper 680.000 klst.

## A Timeline of Large Transformer Models for Speech

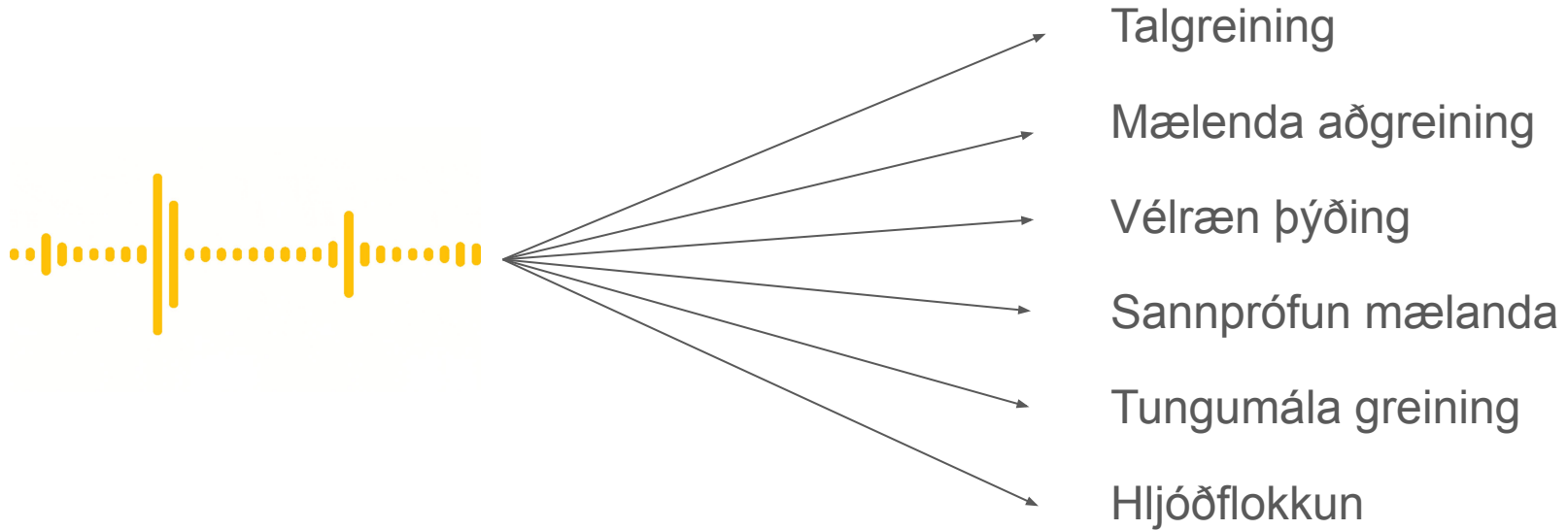


jonathanbgn.com





# Meira en talgreining



Dæmi um talgreining í noktun á Íslandi

# Einfalda störf

- Alþingisræður
  - Lögbundin skylda til að birta ræður þingfunda.
  - Áður fyrr störfuðu ritarar og yfirlesarar. Nú er einungis yfirlesarar.
- Diktering lækna
  - Það þarf að huga að öryggi vinnslunnar.
  - Erfitt að nálgast gögn.
  - Lækníska.



Dæmi:

*„seint væ positívt neer test, ei þreifi eymsl yfir a c liðnum, þreifi eymsl yfir löngu bicepssin“*



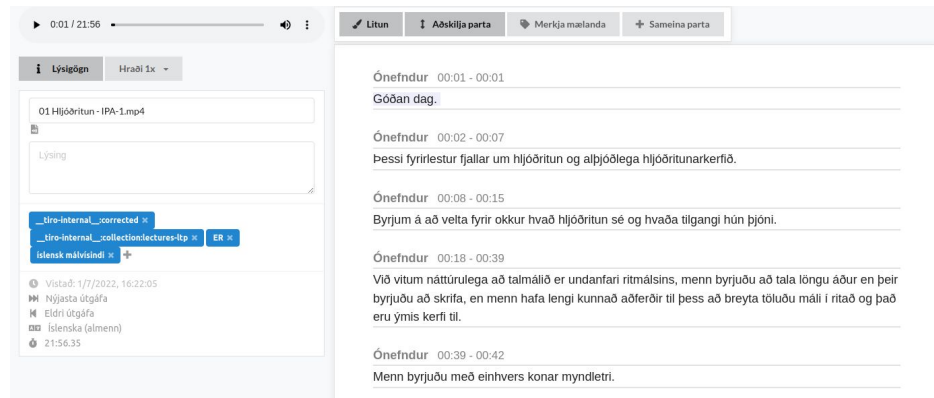
# Auka aðgengi að efni

- Rauntímatextun á myndefni.
- Mikilvægt fyrir heyrnalausar og heyrnaskerta.
- Áskoranir: SMS íslenska.
- Eykur textun á íslensku efni.



# Greining á talgögnum

- Greining og vinnsla á viðtölum.
- Ísmús - Sögulegar upptökur.
- Samræður í þjónustuverum.



The screenshot displays a software interface for audio transcription. At the top, there is a progress bar showing 0:01 / 21:56. Below it, a toolbar contains buttons for 'Litun', 'Aðskilja parta', 'Merkinga málalanda', and 'Sameina parta'. The main area is divided into two panels. The left panel, titled 'Lýsing', shows a file named '01.Hjódritun - IPA-1.mp4' with a 'Lýsing' field. Below this, there are several tabs: '...tiro-internal...corrected', '...tiro-internal...collectionlectures-1tp', and 'Íslensk málvisindi'. The right panel shows a list of transcription segments with their start and end times and the transcribed text.

Ónefndur	Texti
00:01 - 00:01	Góðan dag.
00:02 - 00:07	Þessi fyrirlestur fjallar um hjódritun og alþjóðlega hjódritunarkerfið.
00:08 - 00:15	Byrjum á að velta fyrir okkur hvað hjódritun sé og hvaða tilgangi hún þjóni.
00:18 - 00:39	Við vitum náttúrulega að talmálið er undanfari ritmálsins, menn byrjuðu að tala löngu áður en þeir byrjuðu að skrifa, en menn hafa lengi kunnað aðferðir til þess að breyta töluðu máli í ritað og það eru ýmis kerfi til.
00:39 - 00:42	Menn byrjuðu með einhvers konar myndletri.






---

[www.tiro.is](http://www.tiro.is)

[tiro@tiro.is](mailto:tiro@tiro.is)

[www.talgreinir.is](http://www.talgreinir.is)

A man with short brown hair, wearing a dark jacket over a blue shirt, is speaking into a microphone. He is gesturing with his left hand. The background is dark and out of focus. A subtitle is visible at the bottom of the frame.

fá upptöku af því að sjá hvernig það gengur að túlka þetta því ég  
tala oft ansi