Automatic transcriptions of folklore audio recordings

Trausti Dagsson, Rósa Þorsteinsdóttir, Finnur Ágúst Ingimundarson (The Árni Magnusson Institute for Icelandic Studies)

Luke O'Brien (Tiro)

STOFNUN ÁRNA MAGNÚSSONAR
í íslenskum fræðum

Ísmús

- Digitized audio archive
- The Árni Magnússon Institute for Icelandic Studies
- More than 2000 hours
- Mostly collected 1960 - 1980

Main collectors

Hallfreður Örn Eiríksson (1932-2005)

Helga Jóhannsdóttir (1935-2006)

Jón Samsonarson (1931-2010)

2700 informants

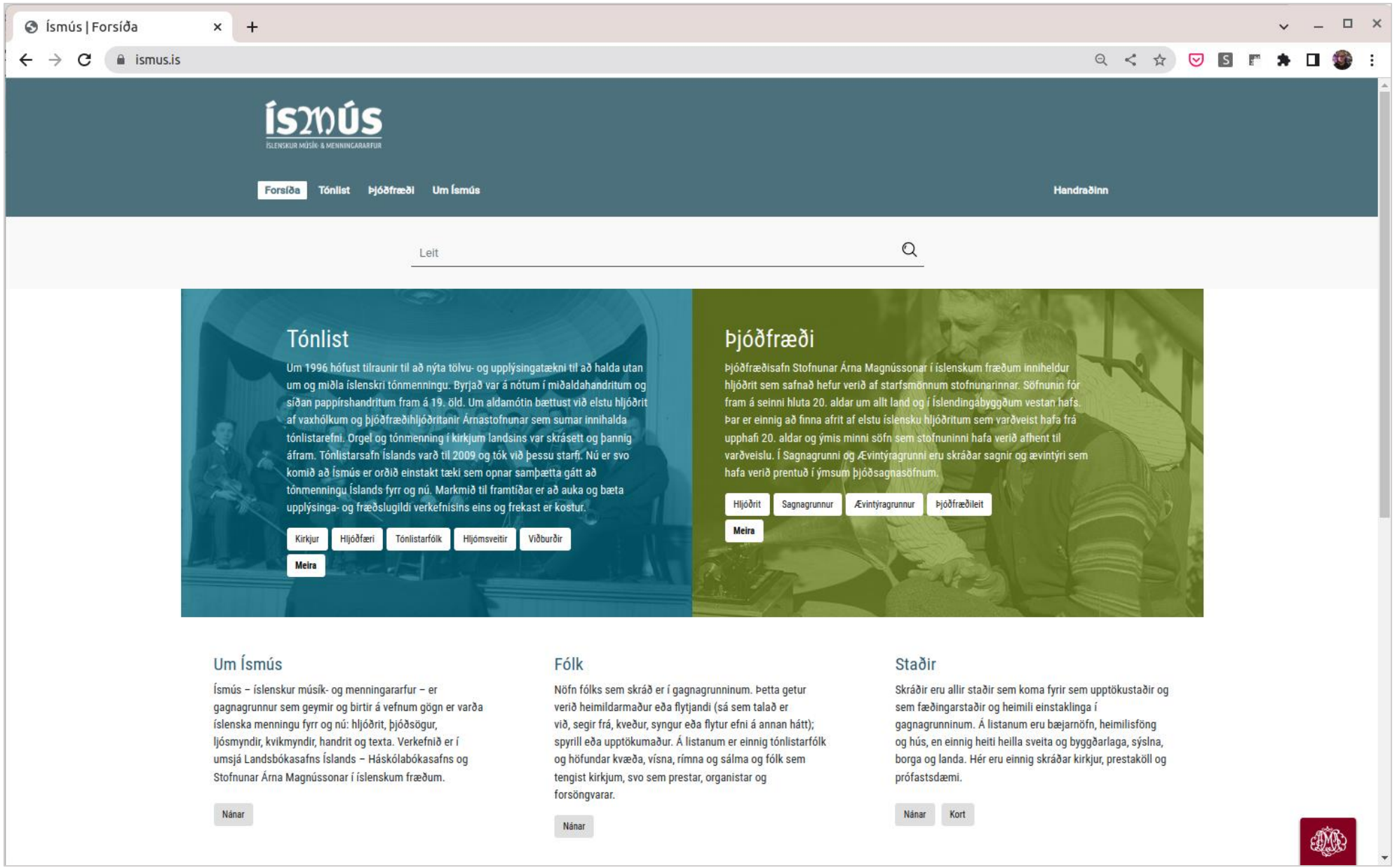
Ísmús

- Folklore and musicology platform
  - Ethnographic interviews
  - Legends database
  - Fairy tales
  - Letter collections
  - Musicology
  - and more...

- www.ismus.is

Ísmús

- Personal connections
- Artistic inspiration

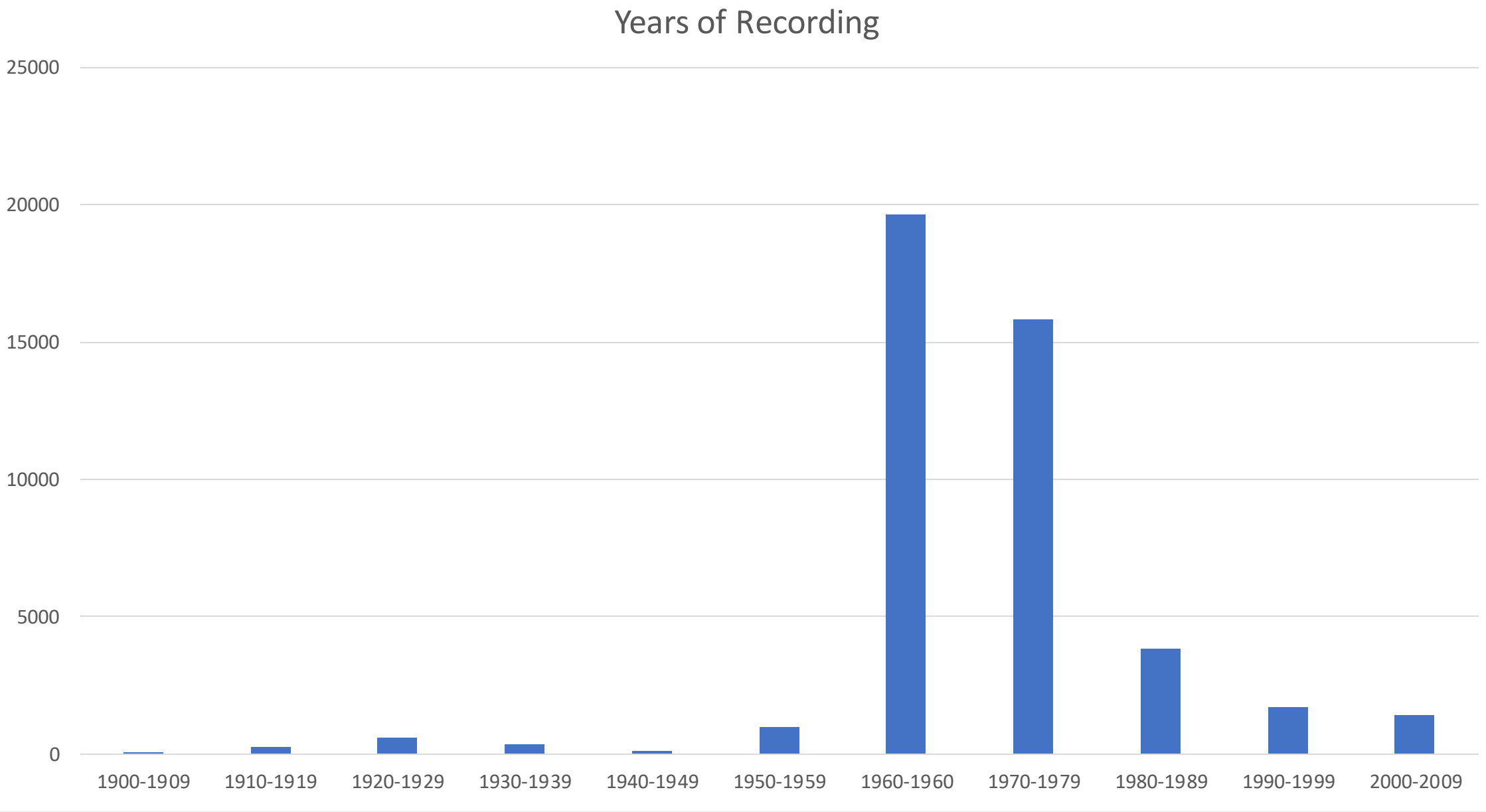

Ísmús | Forsíða
ismus.is
ÍSMÚS
ÍSLENSKUR MÚSÍK- & MENNINGARARFUR
Forsíða
Tónlist
Þjóðfræði
Um Ísmús
Handraðinn
Leit
Tónlist
Um 1996 hófust tilraunir til að nýta tölvu- og upplýsingatækni til að halda utan um og miðla íslenskri tónmenningu. Byrjað var á nótum í miðaldahandritum og síðan pappírshandritum fram á 19. öld. Um aldamótin bættust við elstu hljóðrit af vaxhólkum og þjóðfræðihljóðritanir Árnastofnunar sem sumar innihalda tónlistarefni. Orgel og tónmenning í kirkjum landsins var skrásett og þannig áfram. Tónlistarsafn Íslands varð til 2009 og tók við þessu starfi. Nú er svo komið að Ísmús er orðið einstakt tæki sem opnar samþætta gátt að tónmenningu Íslands fyrr og nú. Markmið til framtíðar er að auka og bæta upplýsinga- og fræðslugildi verkefnisins eins og frekast er kostur.
Kirkjur
Hljóðfæri
Tónlistarfólk
Hljómsveitir
Viðburðir
Meira
Þjóðfræði
Þjóðfræðisafn Stofnunar Árna Magnússonar í íslenskum fræðum inniheldur hljóðrit sem safnað hefur verið af starfsmönnum stofnunarinnar. Söfnunin fór fram á seinni hluta 20. aldar um allt land og í Íslendingabyggðum vestan hafs. Þar er einnig að finna afrit af elstu íslensku hljóðritum sem varðveist hafa frá upphafi 20. aldar og ýmis minni söfn sem stofnuninni hafa verið afhent til varðveislu. Í Sagnagrunni og Ævintýragrunni eru skráðar sagnir og ævintýri sem hafa verið prentuð í ýmsum þjóðsagnasöfnum.
Hljóðrit
Sagnagrunnur
Ævintýragrunnur
Þjóðfræðileit
Meira
Um Ísmús
Ísmús – íslenskur músík- og menningararfur – er gagnagrunnur sem geymir og birtir á vefnum gögn er varða íslenska menningu fyrr og nú: hljóðrit, þjóðsögur, ljósmyndir, kvikmyndir, handrit og texta. Verkefnið er í umsjá Landsbókasafns Íslands – Háskólabókasafns og Stofnunar Árna Magnússonar í íslenskum fræðum.
Nánar
Fólk
Nöfn fólks sem skráð er í gagnagrunninum. Þetta getur verið heimildarmaður eða flytjandi (sá sem talað er við, segir frá, kveður, syngur eða flytur efni á annan hátt); spyrill eða upptökumaður. Á listanum er einnig tónlistarfólk og höfundar kvæða, vísna, rímna og sálma og fólk sem tengist kirkjum, svo sem prestar, organistar og forsöngvarar.
Nánar
Staðir
Skráðir eru allir staðir sem koma fyrir sem upptökustaðir og sem fæðingarstaðir og heimili einstaklinga í gagnagrunninum. Á listanum eru bæjarnöfn, heimilisföng og hús, en einnig heiti heilla sveita og byggðarlaga, sýslna, borga og landa. Hér eru einnig skráðar kirkjur, prestaköll og prófastsdæmi.
Nánar
Kort

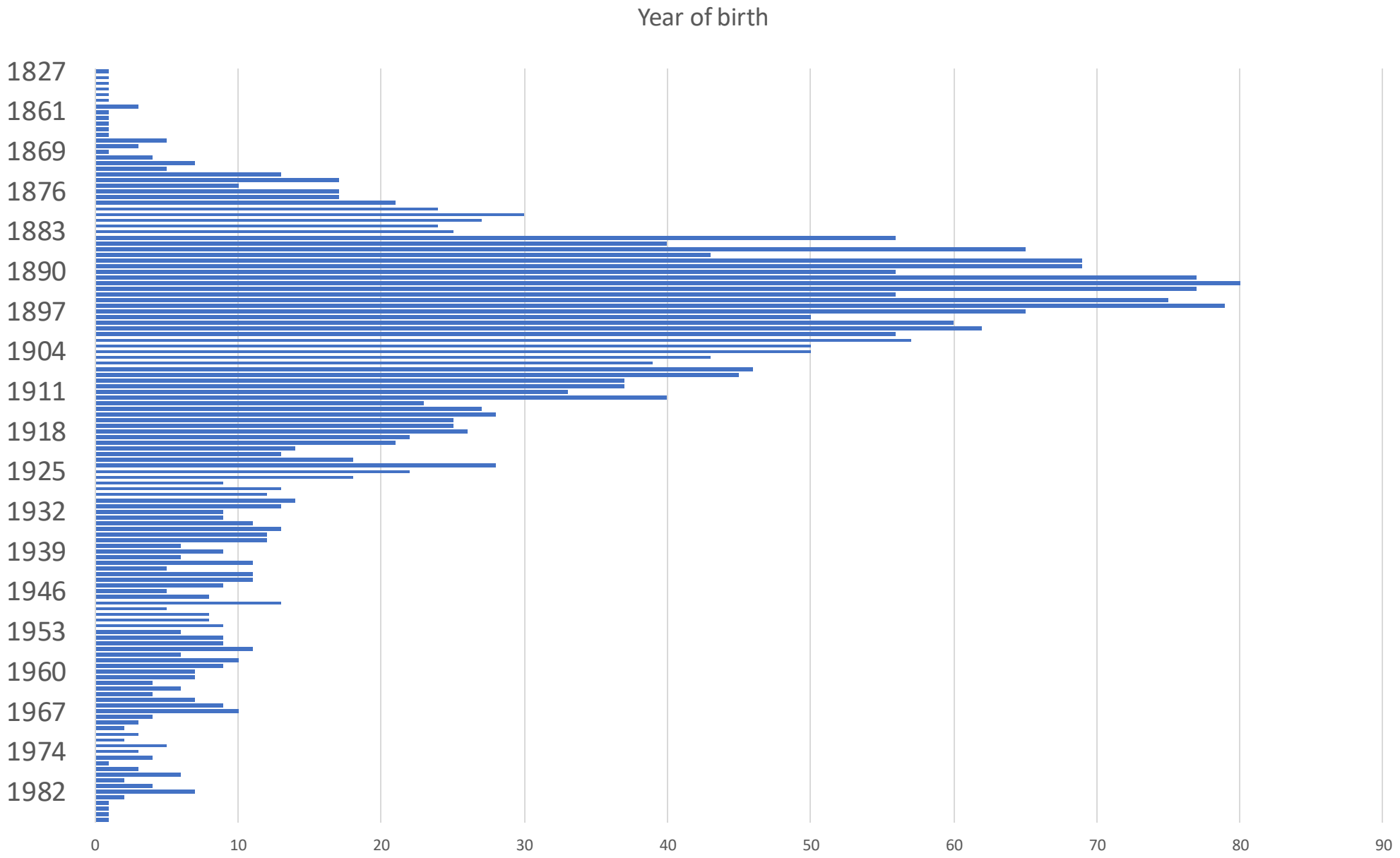
Digitized - post-digitization - what's next?

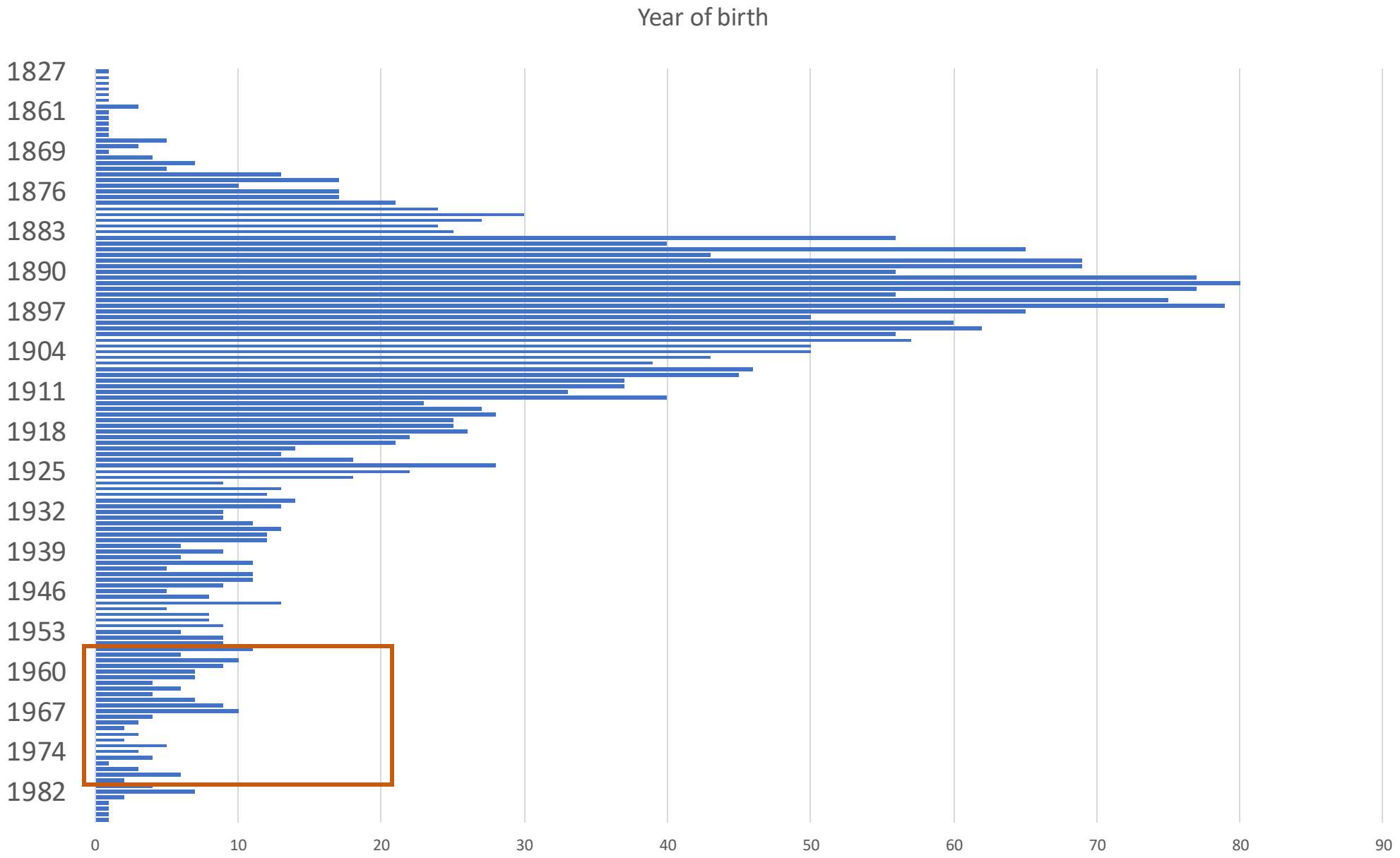
- Current limitations
  - Limited metadata
  - Keywords
  - Short description

Automatic Transcriptions

- The projects aim:
  - Train ASR system to automatically transcribe audio recordings from the Ethnology collection of Árnastofnun
  - Create a text corpus of speech data from the collection
- Funding from The Icelandic Center for Research (RANNÍS) through Centre for Digital Humanities and Arts
- Collaboration with technology company Tíró
- Participants:
  - Rósa Þorsteinsdóttir (The Árni Magnússon Institute for Icelandic Studies)
  - Trausti Dagsson (The Árni Magnússon Institute for Icelandic Studies)
  - Finnur Ingimundarson (The Árni Magnússon Institute for Icelandic Studies)
  - Luke O'Brien (Tíró)

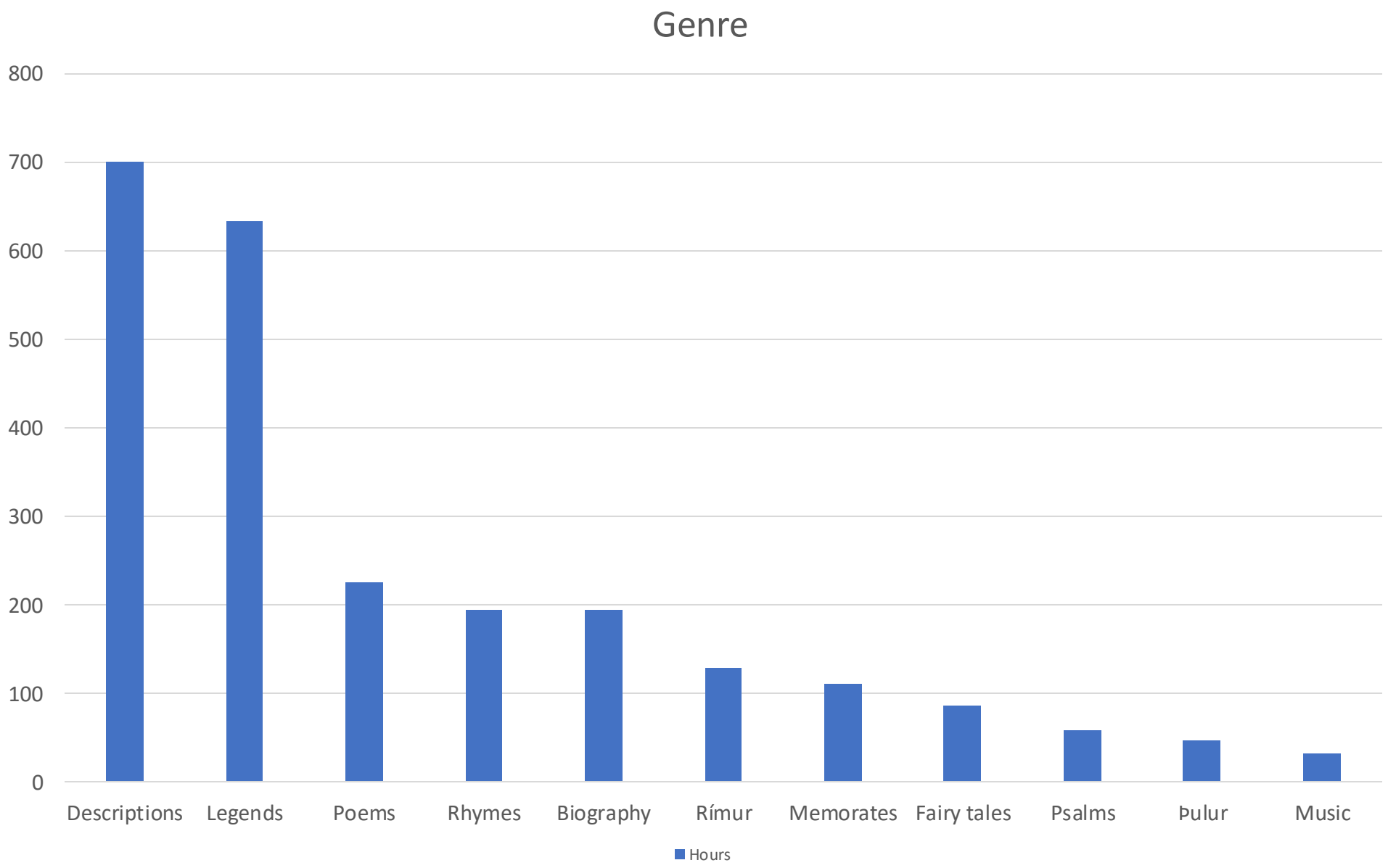

Years of Recording
25000
20000
15000
10000
5000
0
1900-1909
1910-1919
1920-1929
1930-1939
1940-1949
1950-1959
1960-1960
1970-1979
1980-1989
1990-1999
2000-2009


Year of birth
1827
1861
1869
1876
1883
1890
1897
1904
1911
1918
1925
1932
1939
1946
1953
1960
1967
1974
1982
0
10
20
30
40
50
60
70
80
90


Year of birth
1827
1861
1869
1876
1883
1890
1897
1904
1911
1918
1925
1932
1939
1946
1953
1960
1967
1974
1982
0
10
20
30
40
50
60
70
80
90

Gender of speakers

- 57% male
  - 1504 hours – 65%
- 42 % female
  - 821 hours – 35%


Genre
800
700
600
500
400
300
200
100
0
Descriptions
Legends
Poems
Rhymes
Biography
Rímur
Memorates
Fairy tales
Psalms
Þulur
Music
Hours

Preparation and Training

- Correction and time align of existing text data
- Training the model using neural network
- Testing and Correcting

```
1
00:00:00,000 --> 00:00:02,200
Jón M.:
2
00:00:02,200 --> 00:00:06,879
Nú en annars líkaði mér svona heldur vel í skóla.
3
00:00:07,089 --> 00:00:11,349
En ég var enginn námsmaður og hef aldrei verið
```

Models

- Acoustic Model
  - Relationship between audio and phonemes
  - TDNN (time-delayed neural network) chain model trained in Kaldi
  - Trained on:
    - Althingi's Parliamentary Speeches – corpus of 514 hours
    - 114 hours of speech from the first Samrómur release
    - 173 hours of unverified Samrómur data
    - 228 hours of dataset from RÚV (National Broadcast Company)

Models

- Language Model
  - N-gram language model
  - Trained on:
    - Corrected OCR data
    - The Icelandic Gigaword Corpus
    - Ethnographic data from the National Museum of Iceland
    - Audio file descriptions from Ísmús for their content
    - Place name data from the Icelandic Place Name Collection
  - Vocabulary data
    - The Database of Icelandic Morphology, etc.

The Results

- ~12.000.000 words
- Surprisingly good in most cases but sometimes surprisingly bad
- Still being evaluated

Challenges

- Poor audio quality
- Multiple speakers
- Not split into sentences
- "Já já, já já já" "ha?"
- Background noise
  - Children, spouses, clocks, telephones
- Uncertainty...

Challenges

- Poor audio quality
- Multiple speakers
- Not split into sentences
- "Já já, já já já" "ha?"
- Background noise
  - Children, spouses, clocks, telephones
- Uncertainty of the mindless robot

Usage

- Search functionality for Ísmús
  - Folklore search
  - Word distribution-based text linking
- Language research
  - Speech Text Corpus
- Usage of model for further transcriptions
  - E.g., RÚV (National Broadcast Company), National Library, etc.

Examples

- Narrative: Elves (huldufólk), Jenný Jónasdóttir (1904-1991)
- Fairy tale: Steinunn Þorsteinsdóttir, (1887-1973)

Examples

- Singing: Sjö sinnum það sagt er mér, Jón Stefánsson (1880-1971)
- Singing: Kom ég upp í Kvíslarskarð, Tryggvi Sigtryggsson (1894-1986)