THE UNIVERSITY of EDINBURGH

‘Progress Towards Accurate Automatic Speech Recognition for Scottish Gaelic’

Soillse Seminar Series

Will Lamb

27 April 2022

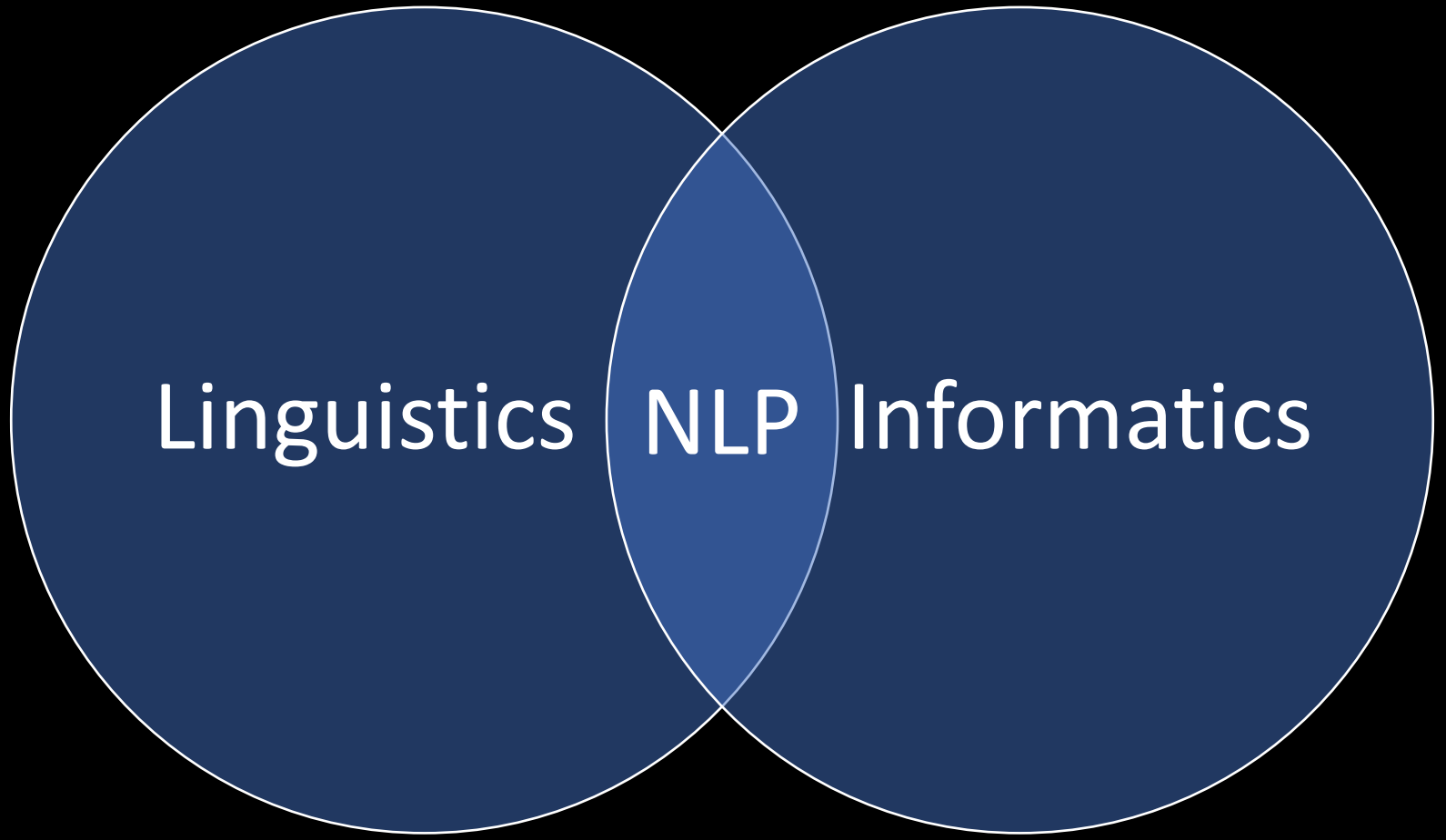
Questions for today

What resources are available for developing Gaelic NLP and ASR, specifically?

What Gaelic NLP applications currently exist?

What is needed to develop a robust, accurate ASR system for Gaelic?

What is Natural Language Processing (NLP)?



Why is NLP important for Gaelic?

Start learning

Scottish Gaelic!

Which of these is "dog"?

balach

cù

cat

muc

Innovation Award, Nov 2021

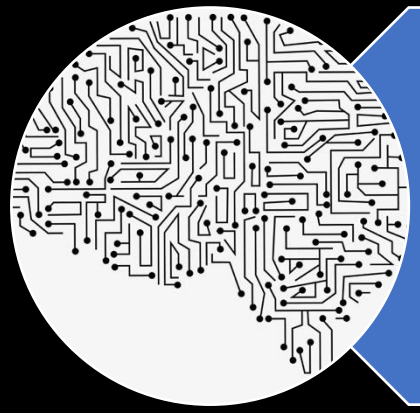
Types of NLP

Text-based

Audio-based

Types of NLP

Rule based



Statistical

Part-of-speech tagger

Given a text input, returns tokens that are morphologically classified

Input: ‘tha na coin mhòra ann’

Token	Lemma	Tag	Gloss
tha	bi	Vp	'Verb: present tense'
na	na	Td	'Article: definite'
coin	cù	Nc	'Noun: common'
mhòra	mòr	Aq	'Adjective: attributive'
ann	e	Pr	'Prepositional pronoun'

Audio based NLP

Speech synthesis

Text to speech (TTS)

Speech recognition

Speech to text (STT)

NLP Resources for Gaelic

Digital lexicon

- Am Faclair Beag

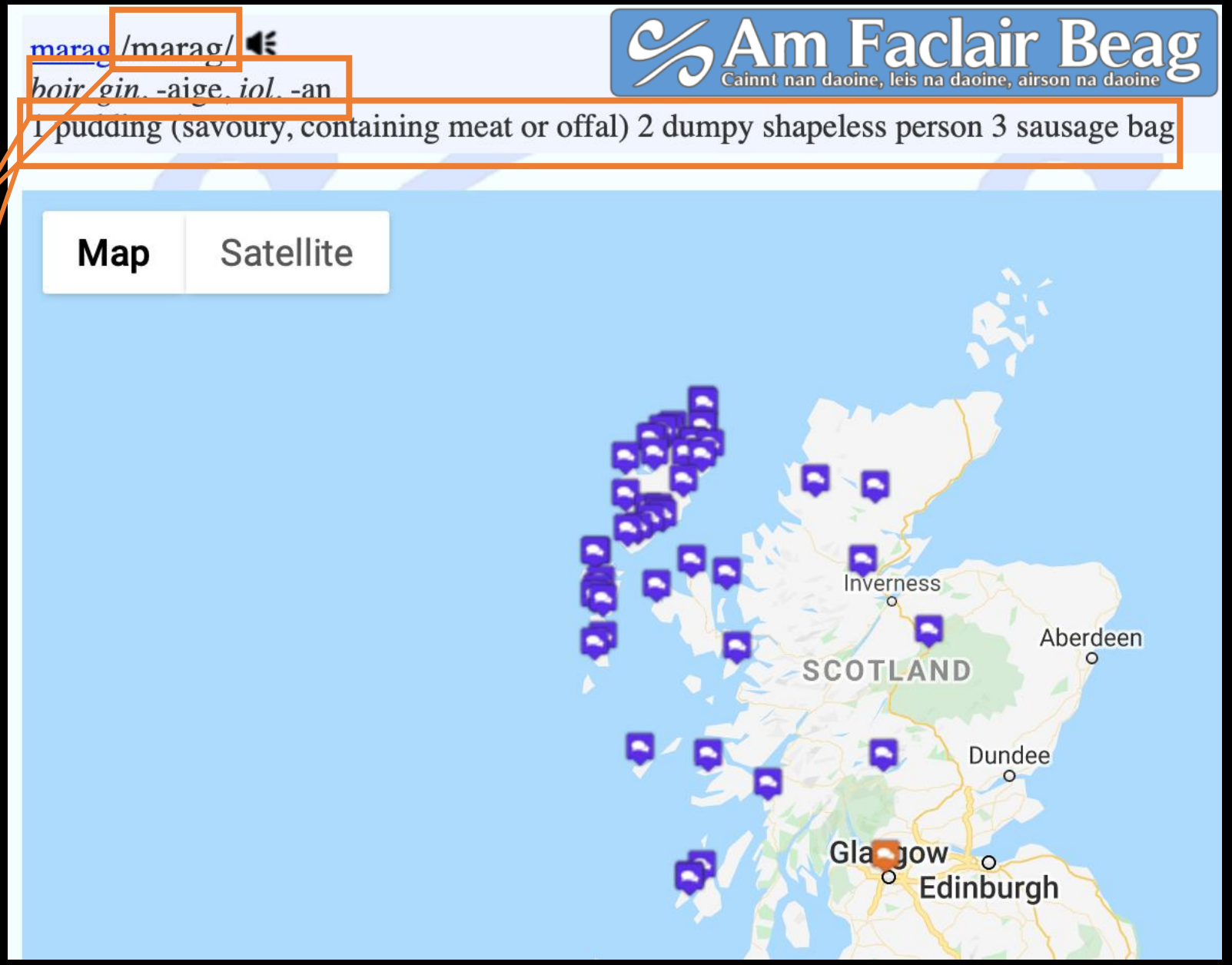
Digital corpora

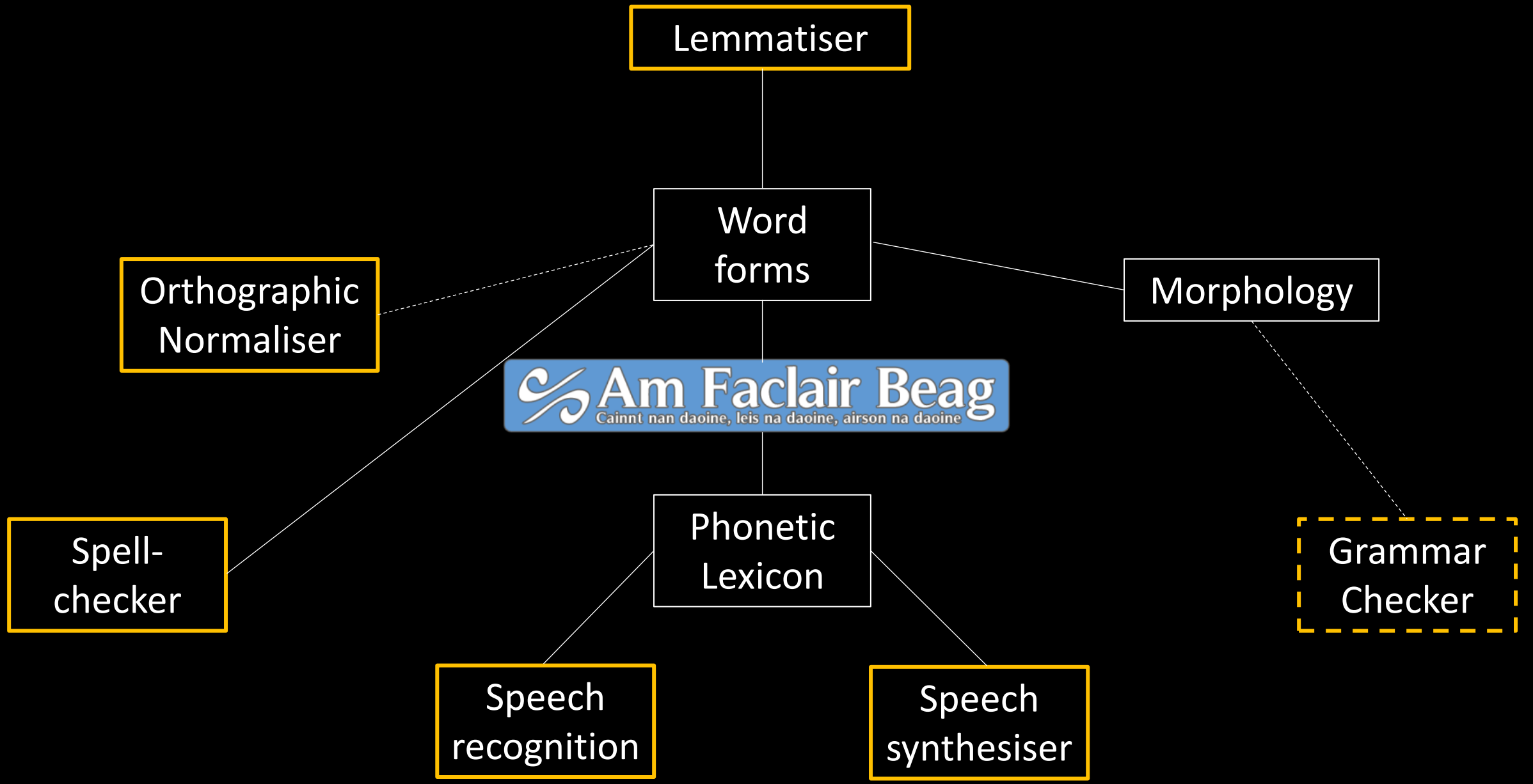
- *Annotated*
  - ARCOSG
- *Non-annotated*
  - Corpas na Gàidhlig (DASG, U of Glasgow)
  - Island Voices
  - The GD corpus (from An Crúbadán)

Phonology

Morphology

Semantics




Lemmatiser
Word forms
Orthographic Normaliser
Morphology
Am Faclair Beag
Cainnt nan daoine, leis na daoine, airson na daoine
Spell-checker
Phonetic Lexicon
Grammar Checker
Speech recognition
Speech synthesiser

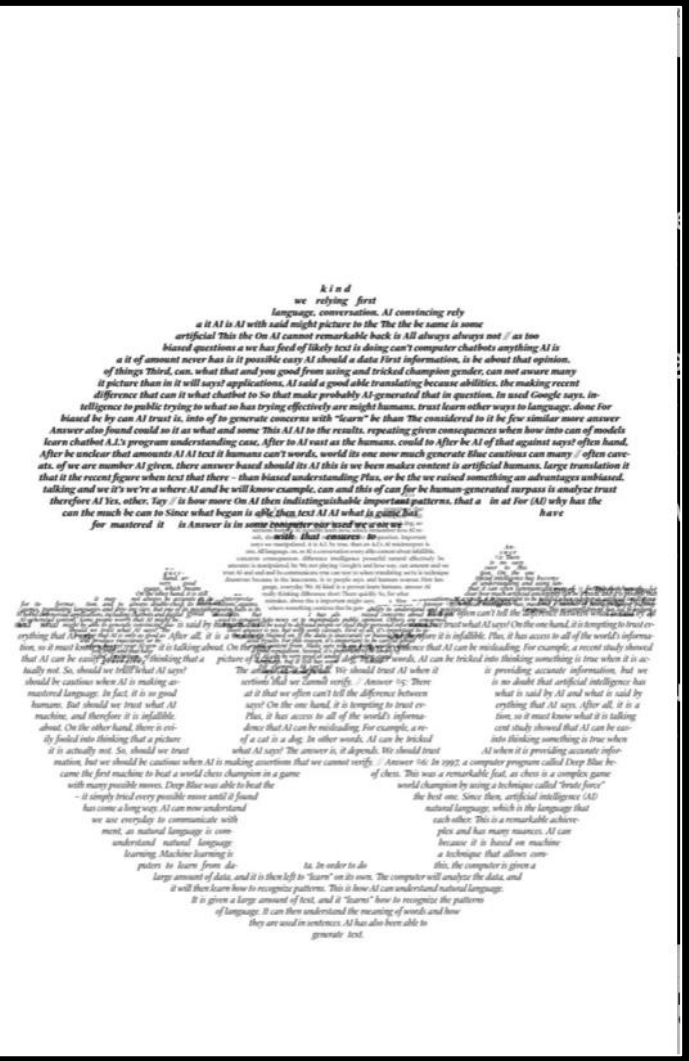
Non-annotated corpora / texts: Examples

- The GD Crúbadán corpus
  - >5.7 million words of web-crawled text hosted by Kevin Scannell
- Island Voices project
  - ~ 350k words
  - Transcribed interviews with Gaelic speakers
- Raw texts from Corpas na Gàidhlig
  - Possible to download individual texts that are out of copyright
- By 2024: Transcriptions from the School of Scottish Studies Archives
  - Approx 7 million words

The New York Times Magazine

A.I. Is Mastering Language. Should We Trust What It Says?

OpenAI's GPT-3 and other neural nets can now write original prose with mind-boggling fluency — a development that could have profound implications for the future.



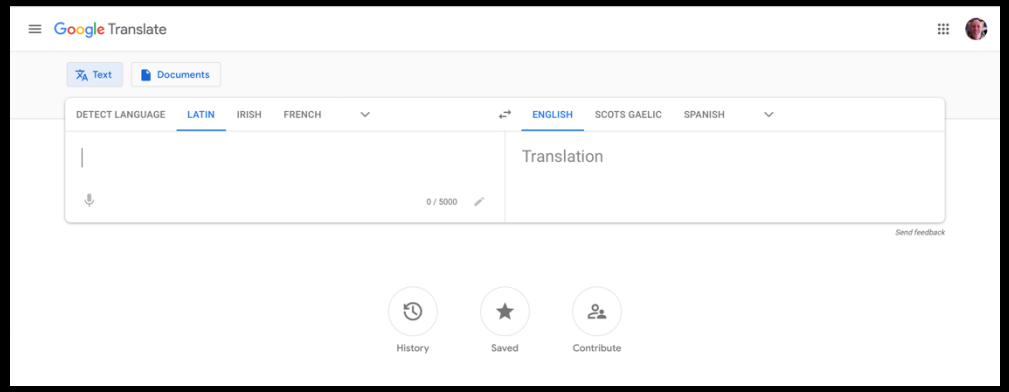
https://www.nytimes.com/2022/04/15/magazine/ai-language.html?unlocked_article_code=AAAAAAAAAAAAAAAACEIPuomT1JKd6J17Vw1cRCfTTMQmqxCdw_PIxftm3iWka3DJDm8fiOEcDIGS-kHAIqNiIt862TuGTdBMdr8zQfg4hsluA3tQcSj66J2VhMZCZCwvtYO4Wm5x1M-QB-1m-2GvMja3I6IizrHi-kLZbmL1WKKM0SciJV5kqMdhZkjqjSJTvtrNG-J029d_3fszVNstFXpbOn7877S_AA5-Od6Gchja9gI2PulYUjzVltaagKkSJEQQURmVCSMivhtvrY9UK9gVP63gLhI_e8GYgb8ZD2FgLIrBFIR5Ch6YSpPo7suxTdZ3OS5A&smid=url-share

Currently available Gaelic NLP tools

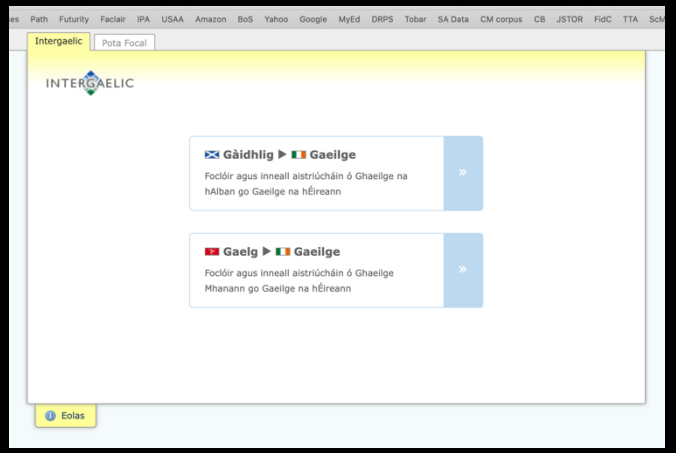
- Machine translation
- Speech synthesis (text to speech)
- Lemmatisation / POS tagging / parsing
- Handwriting recognition

Available NLP apps

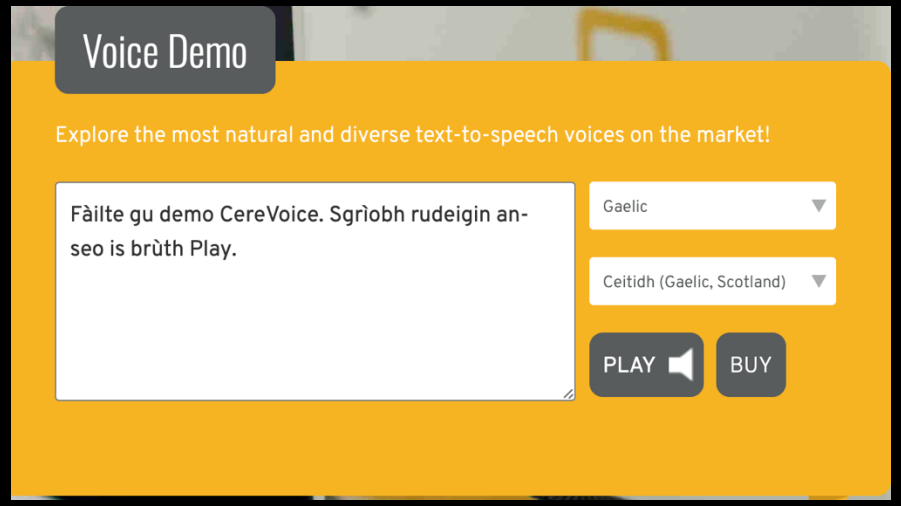
Google Translate



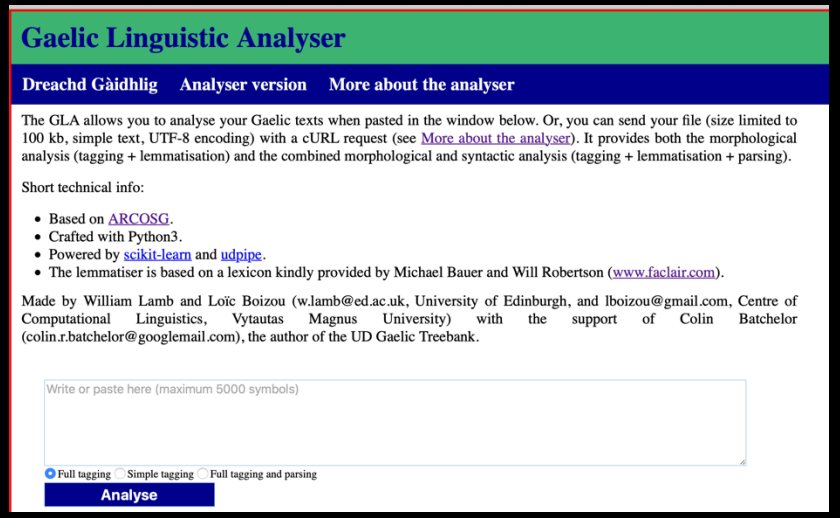
Intergaelic (ScG → Ir)



Cereproc Voice Synthesiser



Gaelic Linguistic Analyser



See: https://guthan.wordpress.com/2021/03/20/google-learns-gaelic/

Gaelic Linguistic Analyser

Dreachd Gàidhlig **Analyser version** **More about the analyser**

The GLA allows you to analyse your Gaelic texts when pasted in the window below. Or, you can send your file (size limited to 100 kb, simple text, UTF-8 encoding) with a cURL request (see More about the analyser). It provides both the morphological analysis (tagging + lemmatisation) and the combined morphological and syntactic analysis (tagging + lemmatisation + parsing).

Short technical info:

- Based on ARCOSG.
- Crafted with Python3.
- Powered by scikit-learn and udpipe.
- The lemmatiser is based on a lexicon kindly provided by Michael Bauer and Will Robertson (www.faclair.com).

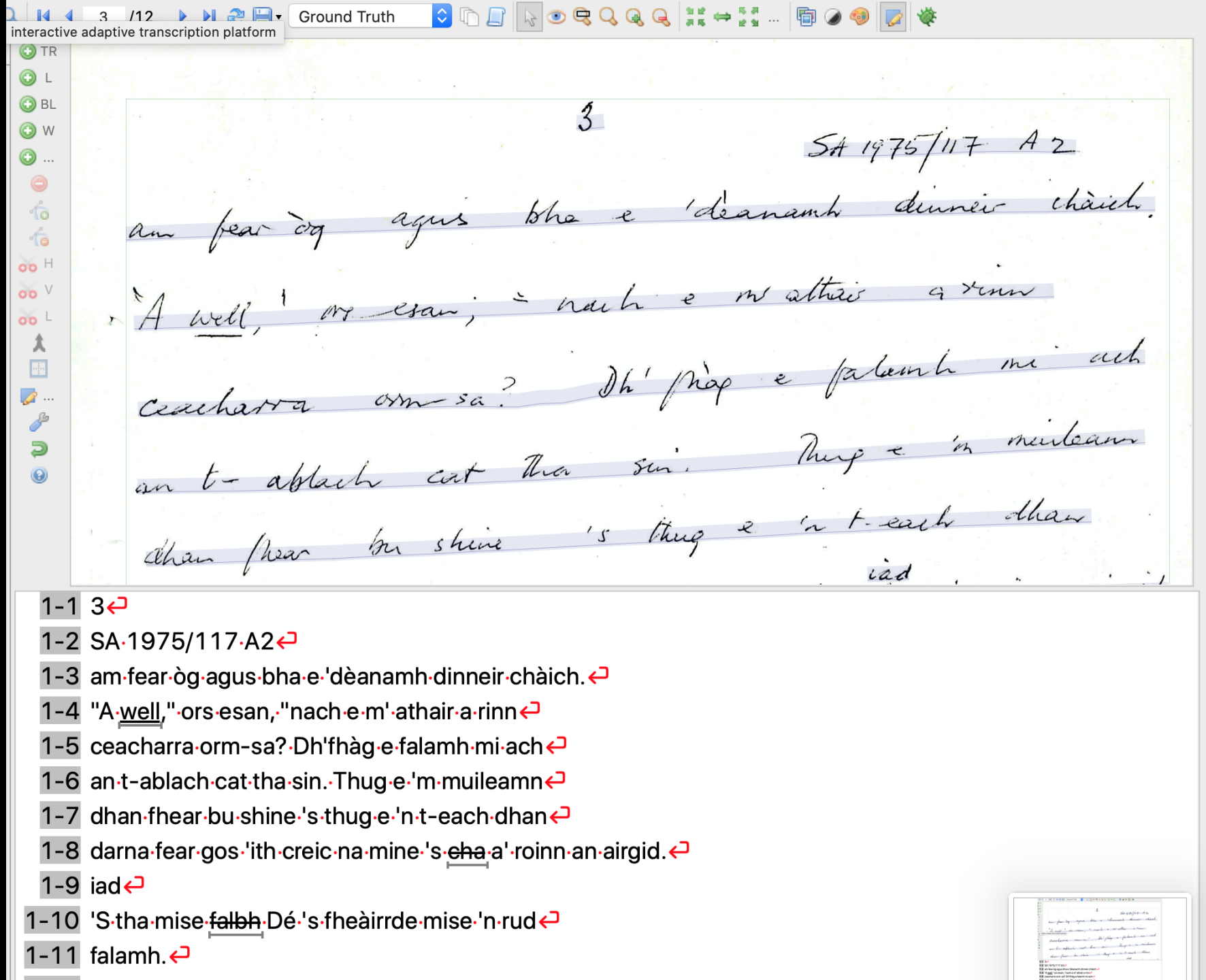
Made by William Lamb and Loïc Boizou (w.lamb@ed.ac.uk, University of Edinburgh, and lboizou@gmail.com, Centre of Computational Linguistics, Vytautas Magnus University) with the support of Colin Batchelor (colin.r.batchelor@googlemail.com), the author of the UD Gaelic Treebank.

Write or paste here (maximum 5000 symbols)

(•) Full tagging ( ) Simple tagging ( ) Full tagging and parsing

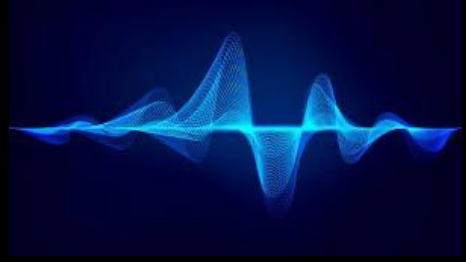
Analyse

Handwriting Recognition: Transkribus



Gaelic Speech Recognition Project

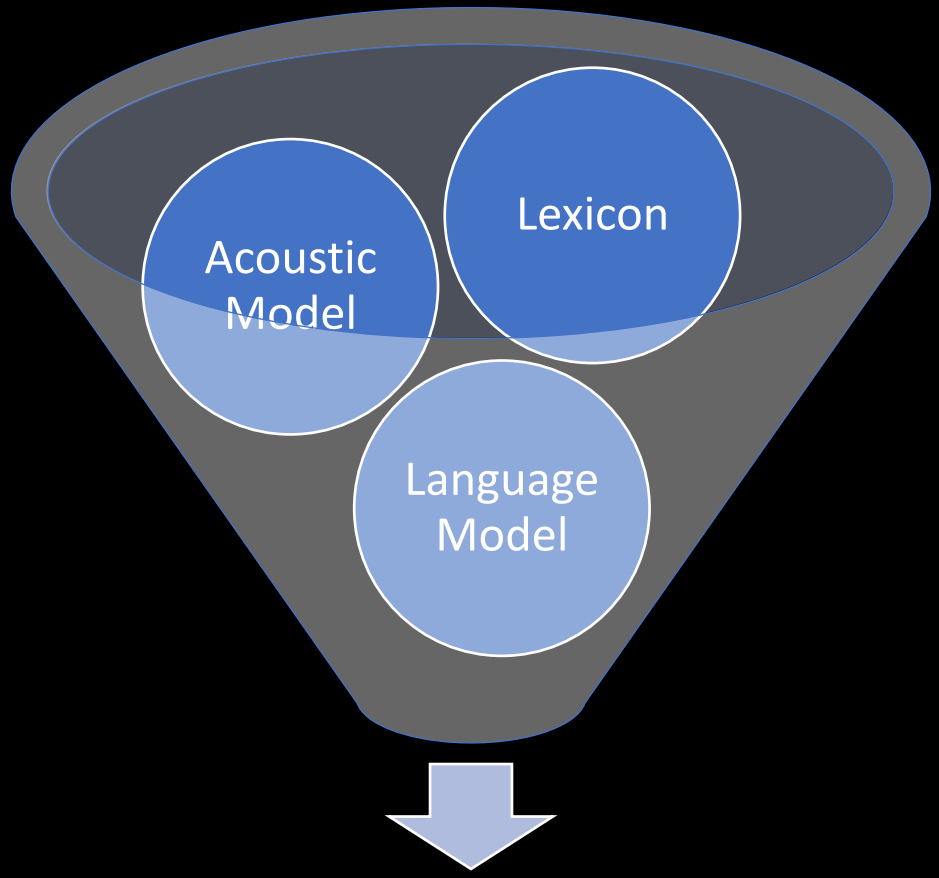
Sept 2020 to July 2021



Goals

1. Automatically transcribe Gaelic narrative audio
   - School of Scottish Studies Archives
   - National Library of Scotland
2. Produce an orthographic normalisation tool
   - To enable rapid inclusion of digitised texts to training pipeline
3. Aid Gaelic revitalisation

Components



Data Preparation: Text Normalisation

Original Text	**Goal Text**
A’ cur uèirichean ri pluga.	a cur uèirichean ri pluga
Bha, bha e ann am Poll a’ Charra ann an 1860.	bha bha e ann am poll a charra ann an ochd ceud deug trì fichead
EC—00:05: Dè bha ceàrr air, air obair a’ bhanca?	dè bha ceàrr air air obair a bhanca

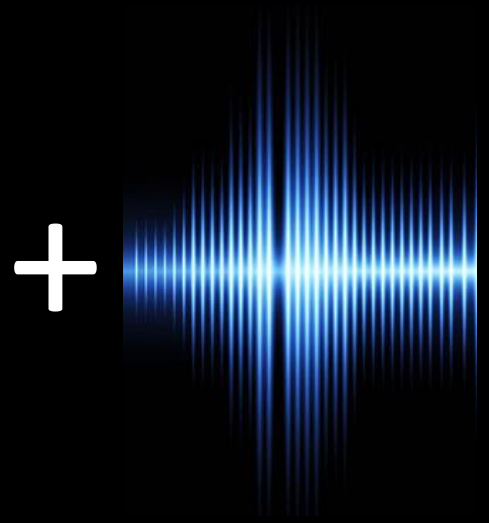
Forced Alignment

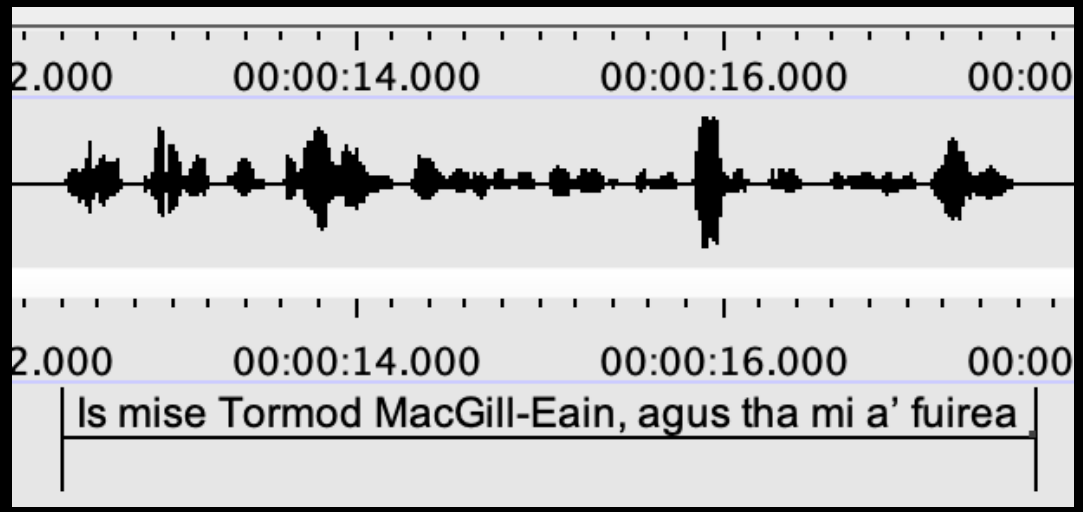
Recorded from Angus MacLellan, Sou
by Morag MacLeod, translated by Ca

Angus MacLellan: 'Nuair a dh'fhàs Rob Ruadh Mac Griogar, dh'fhàs e sean, phòs e agus sguir e dhan robaigeadh, agus bha e fuireach, e fhéin 's a bhean ann an taigh 's cha robh ann ach 'ad fhéin.

Calum Maclean: Seadh dìreach.

AM: Agus 'nuair a chunnaic Diùc Earra-Ghàidheal gu robh Rob air a dhol bhuaithe, bha e air son diola-fiach thoir dheth airson na rinn e robaigeadh air, agus 's ann a chuir





Phoneset mappping

Word	Original (Gaelic IPA)	Standard IPA	ARPABET	Quorate system
uisge	ɯ ʃ gʲ ə	ɯ ʃ kʲ ə	UX SH K AX	uh sh k ax
gorm	g ɔ r ɔ m	k ɔ r ɔ m	K AO DX AO M	k ao r ao m

G2P model

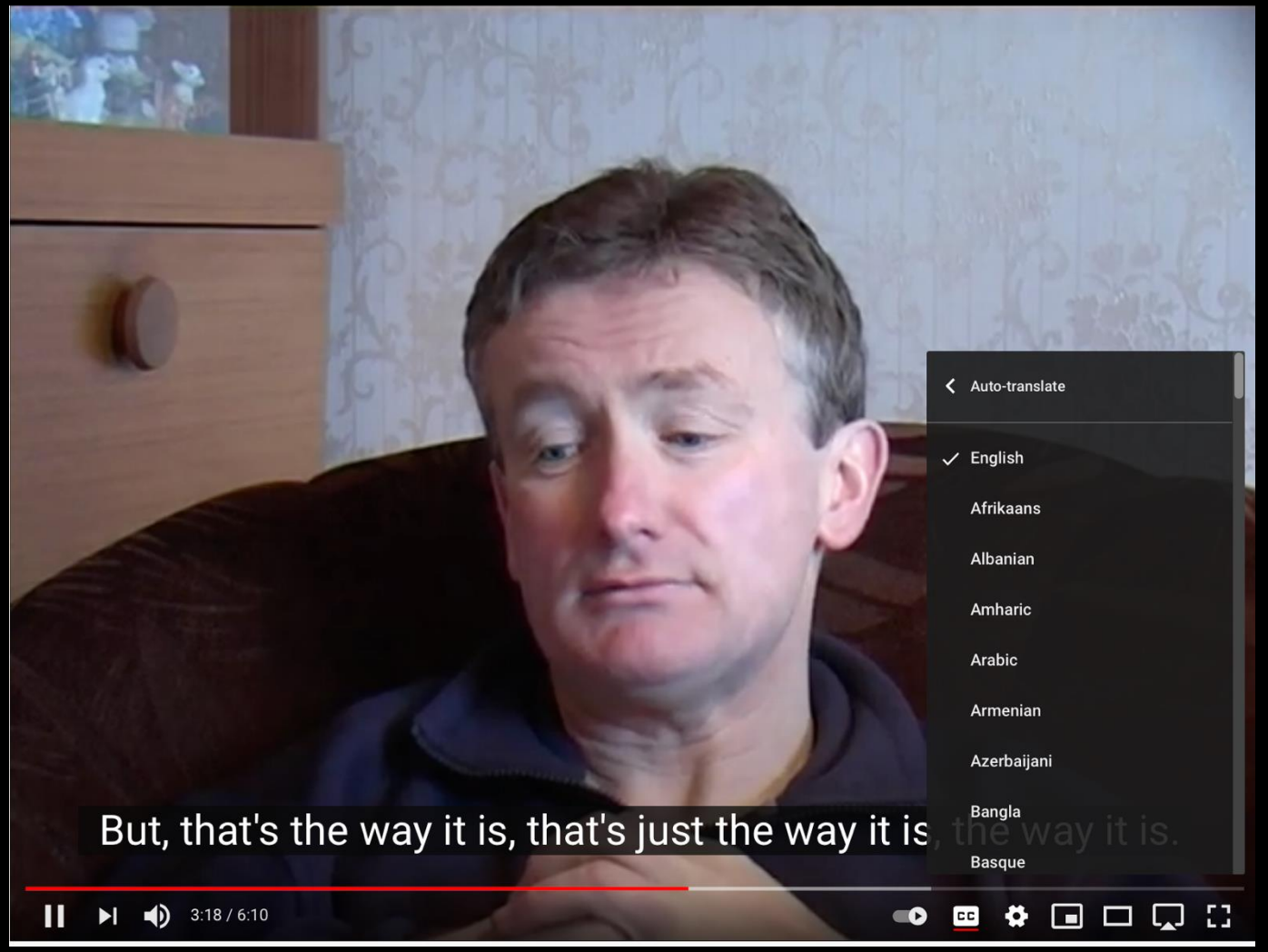
- Converted pronunciation lexicon (30k words)
- Trained G2P model using sequitur-g2p Python toolkit
- Achieved symbol error rate of 3.82 (96.18% accuracy)

```
h-uisgeanan → hh uh sh k ih n aa n
fuaimeannan → f uw ax iy m aa en aa n
galachan → k aa el ax k aa n
```

NB: not IPA /x/

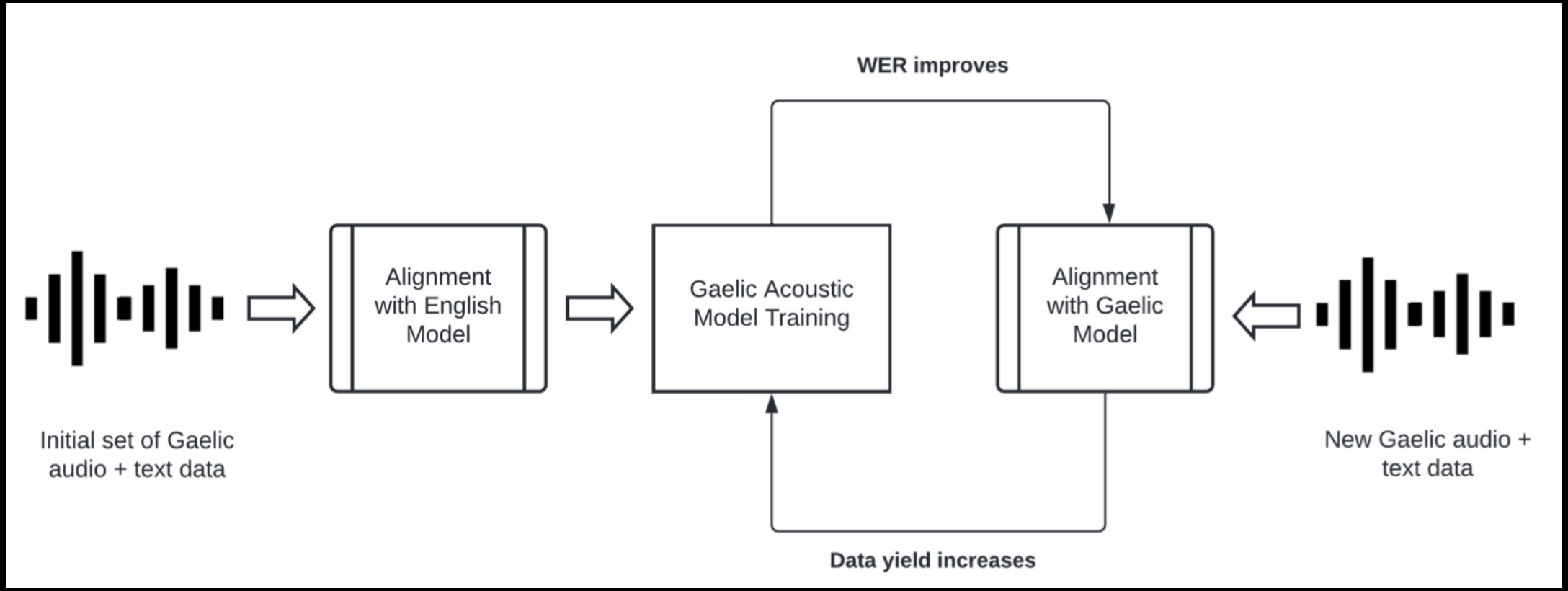
Automatic subtitles

Automatic translation



Gaelic ASR system: Demo

Training process

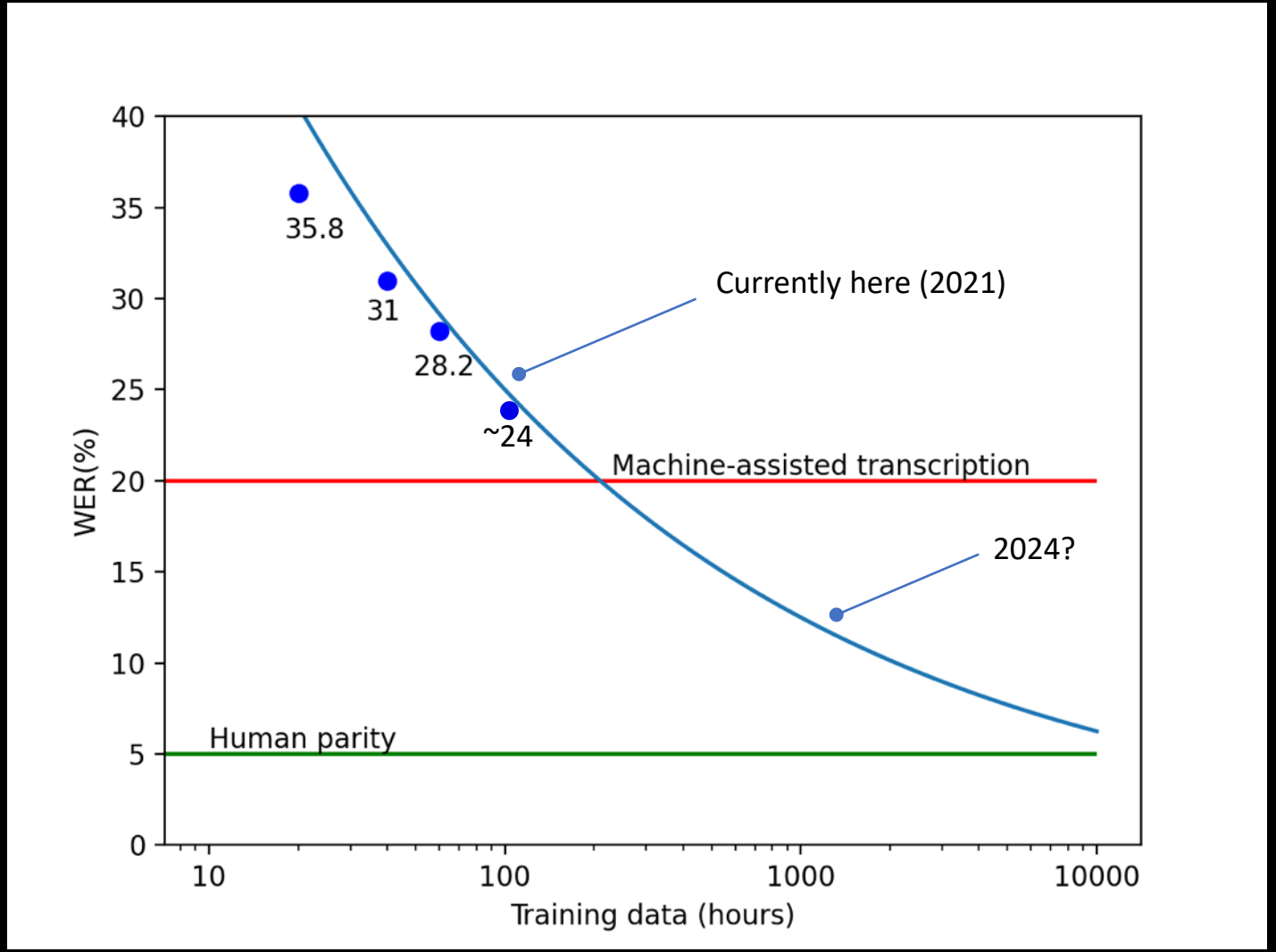


Training data / current word error rate

- Audio: 103.5 hours
- Text: 8,593,567 words (perplexity = 81.27)
- WER: 26.3% (accuracy: 73.7%)

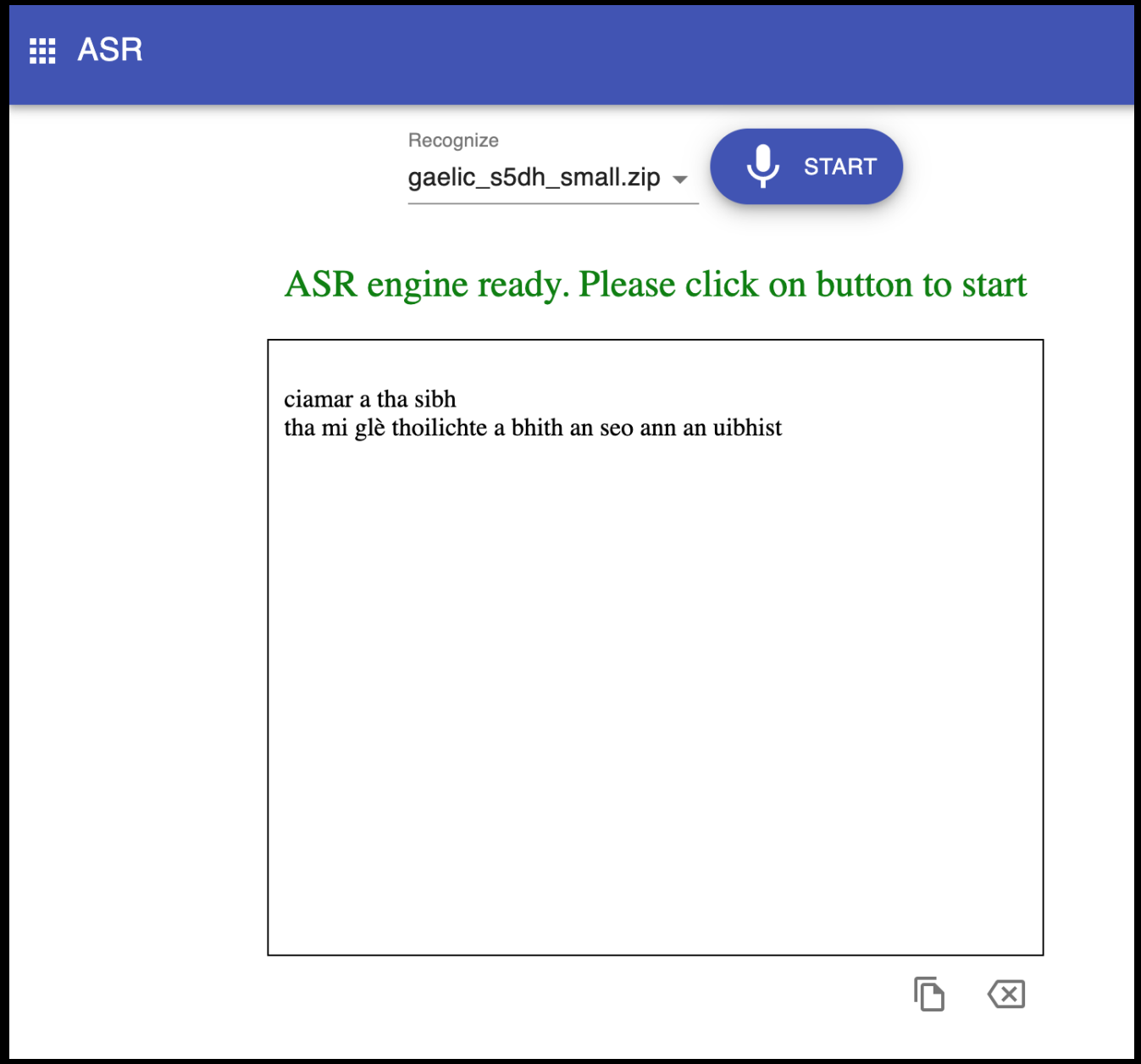
Model	AM data (hrs)	LM	WER(%)
s5	21.2	H	35.80
s5b	39.1	H	31.00
s5c	63.3	H	28.20
s5d	103.5	I	27.40
s5d-small	103.5	I	26.30

Errors per 100 words



The amount of training data (hours)

Prototype web service



An t-àm ri teachd
The future

Future goals (5 years)

- Experiments using multilingual (mainly Irish language) data
  - Irish is better resourced than ScG, so this may prove to be helpful
- Crowdsourcing data acquisition / correction
  - Cf. Meitheal Dúchais - https://www.duchas.ie/en/meitheal/
- Language technology applications of ASR
  - Providing live Gaelic subtitles (e.g. on radio and TV)
  - Automatic searching / indexing of audio content on line (e.g. on Tobar an Dualchais)
  - Enabling coaching of Gaelic phonology (e.g. on Duolingo)

Resources for ASR: desiderata

- More speech data: audio
- More annotated data: audio + text (i.e. transcriptions or scripts)

Examples of data sources

- BBC: Coinneach Macl̀omhair’s programme
  - ~3k-4k hrs of audio data
- Radio nan Gàidheal: Naidheachdan
  - ~100 hours of audio and textual data
- School of Scottish Studies: Tale Archive
  - ~500 to 600 hours of audio and textual data

More information

- Blog for the Gaelic Algorithmic Research Group (GARG)
- Twitter feeds
  - Will Lamb @UilleamUan
  - Kevin Scannell @kscanne
  - Michael Bauer @akerbeltzalba
- Proceedings of the Celtic Language Technology Workshop (1st, 2nd, 3rd)

Mòran taing dhar buidhnean-maoineachaidh

Many thanks to our funders and external partners


ASR
Recognize
gaelic_s5dh_small.zip
STOP
Running
Mòran Taing!

Gaelic Algorithmic Research Group

Rannsachadh digiteach air a' Ghàidhlig ~ Goireasan digiteach airson nan Gàidheal

Digital Research on Gaelic – Digital Resources for Gaelic Speakers