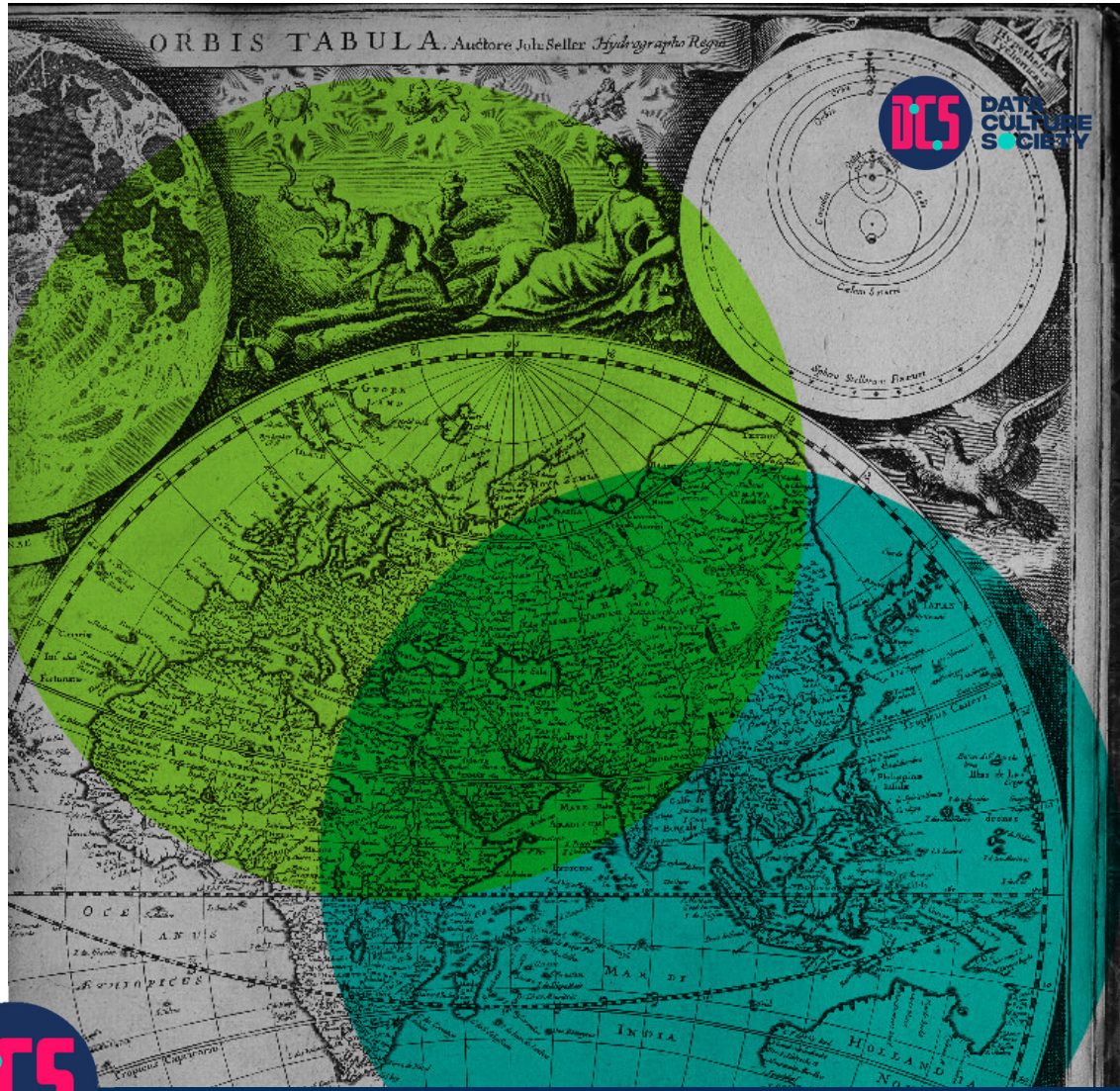
THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

Coding my way out of a box: Achieving emancipation and collaboration through the Digital Humanities



Will Lamb,
Professor in Gaelic Ethnology and Linguistics, University of Edinburgh

Coding my way out of a box:

Achieving emancipation and collaboration through DH

Will Lamb

2 July 2024

DHRSE Summer School

THE UNIVERSITY *of* EDINBURGH

Open to the world

Outline

What are the Digital Humanities?

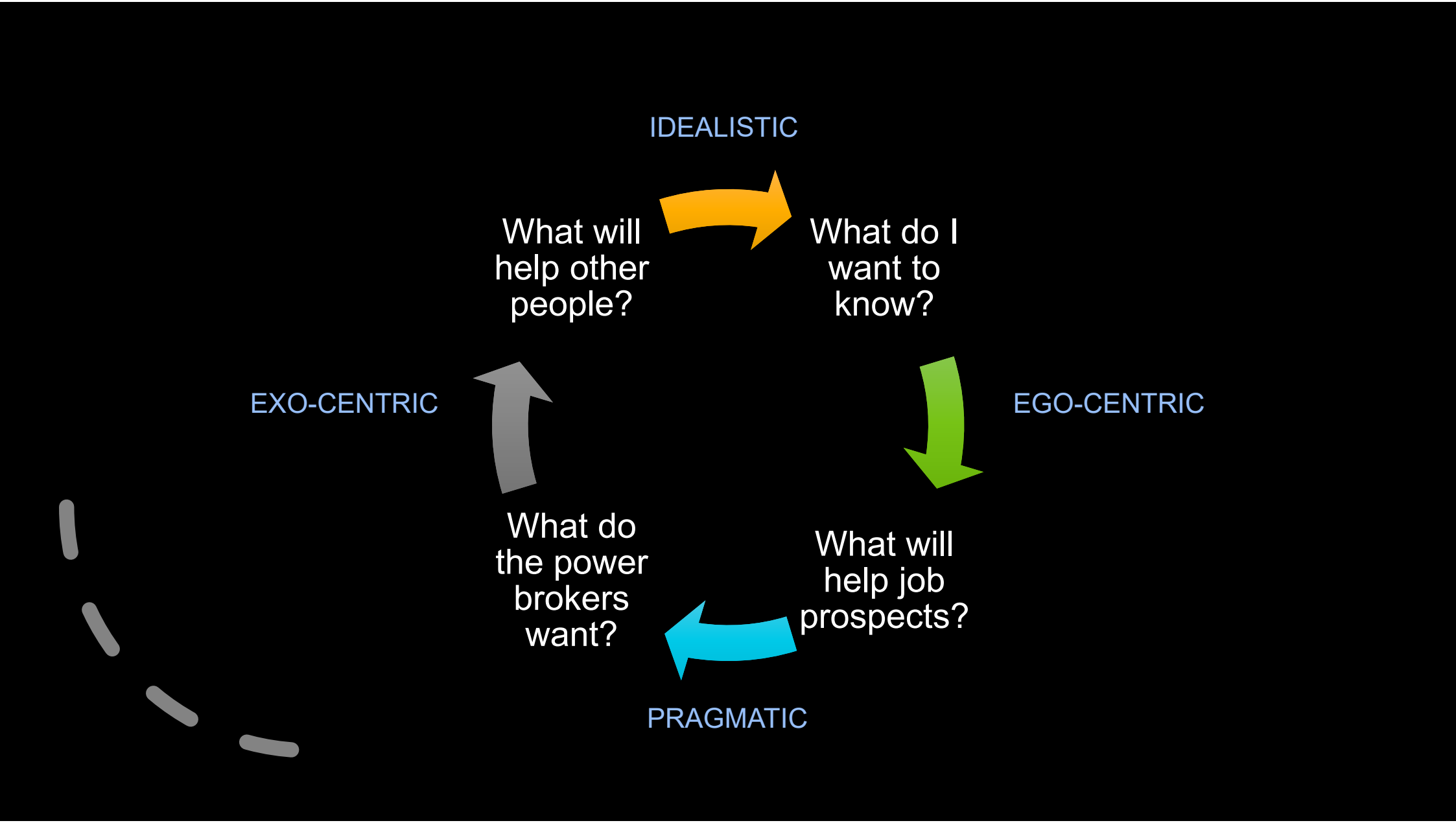
How do I decide what to work on?

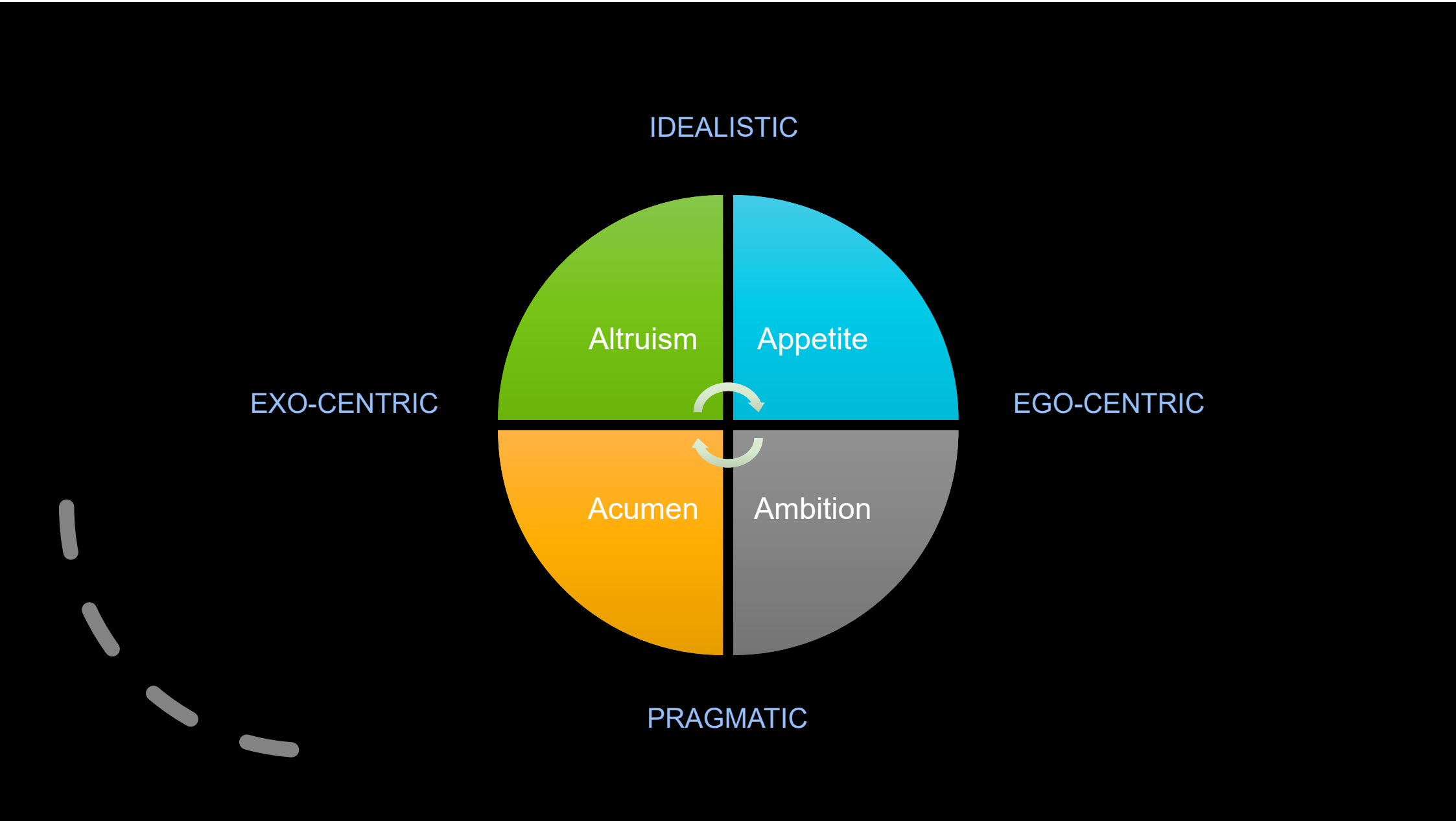
What imperatives drive my own research?

What has my transition from Gaelic linguistics to NLP been like?


Defining the Digital Humanities

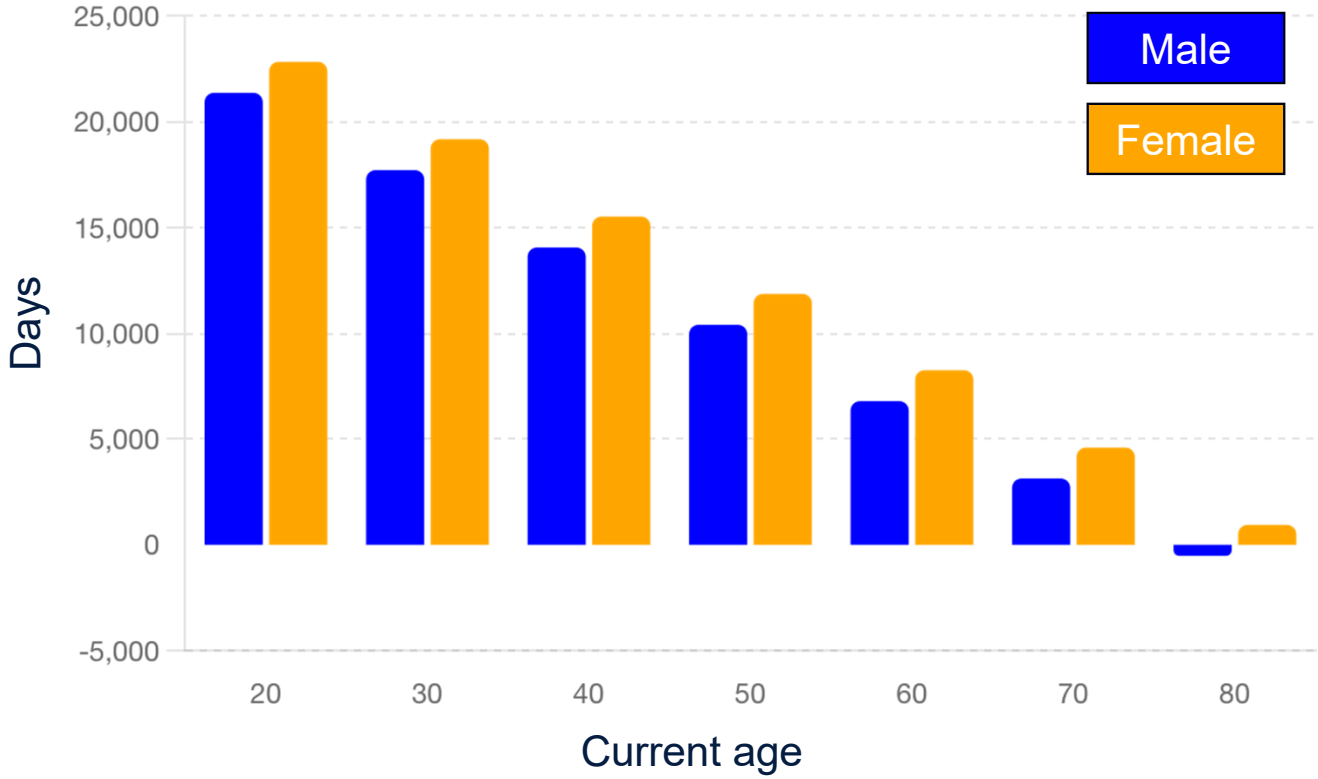

digital
humanities
research
data
can
public
technology
culture
understanding
facebook
archives
humanists
methods
computation
technologies
tools
amazon
also
netflix
projects
insights
network
combine
diffusion
available
ensembles
transforming
work
microsoft
analysis
use
creation
web
age
according
brought
idea
question
large
spotify
personal
like
disciplinary
computer
users
early
social
across
ability
techniques
information
computer-based
score
databases
media
humanistic
artworks
collections
internet
materials
seeing
used
billion
app
shin
academy
ways
greater
areas
possible
science
developed
giants
theorised
part
now
interventions
frameworks
storage
photo
project
provide
help
knowledge
crucial
associated
difficult
artefacts
reproductions
initial
intelligence
produce
communications
making
understand
tradition
expertise
century
laid
concern
however
complex
funded
powered
accessed
reject
worth
adopters
capacity
machine-learning
might
berry
disciplines
world
tremendously
larger
boundaries
giving
policy
openly
modelling
handling
interdisciplinary
enabled
called
centrality
transferring
different
fields
study
publicly
wider
ground
searching
print
concerns
expect
university
vastly
exemplary
music
field
news
focused
easily
commonly
expanding
literature
books
companies
years
david
sussex
google
focus
move
ends
reflect
group
communicate
files
lives
well
search
image
grown
arts
private
digitised
tremendous
one
software
fake
almost
topic
professor
open
ease
contain
ever
times
need
often
imaging
originally
golden
transform
ask
developing
bias
advances
sets
cultures
much
feb
increasingly
applying
values
sound
text
retrieval
recent
important
mere
present
apple
edge
initially
literary
continue
rapid
life
discrimination
sophisticated
leading
problems
artificial
together
equally
images
networks
apply
ideas
york
many
increase
synchonia
intrusive
created
material
engineering
comparison


IDEALISTIC
What do I want to know?
EGO-CENTRIC
What will help job prospects?
PRAGMATIC
What do the power brokers want?
EXO-CENTRIC
What will help other people?

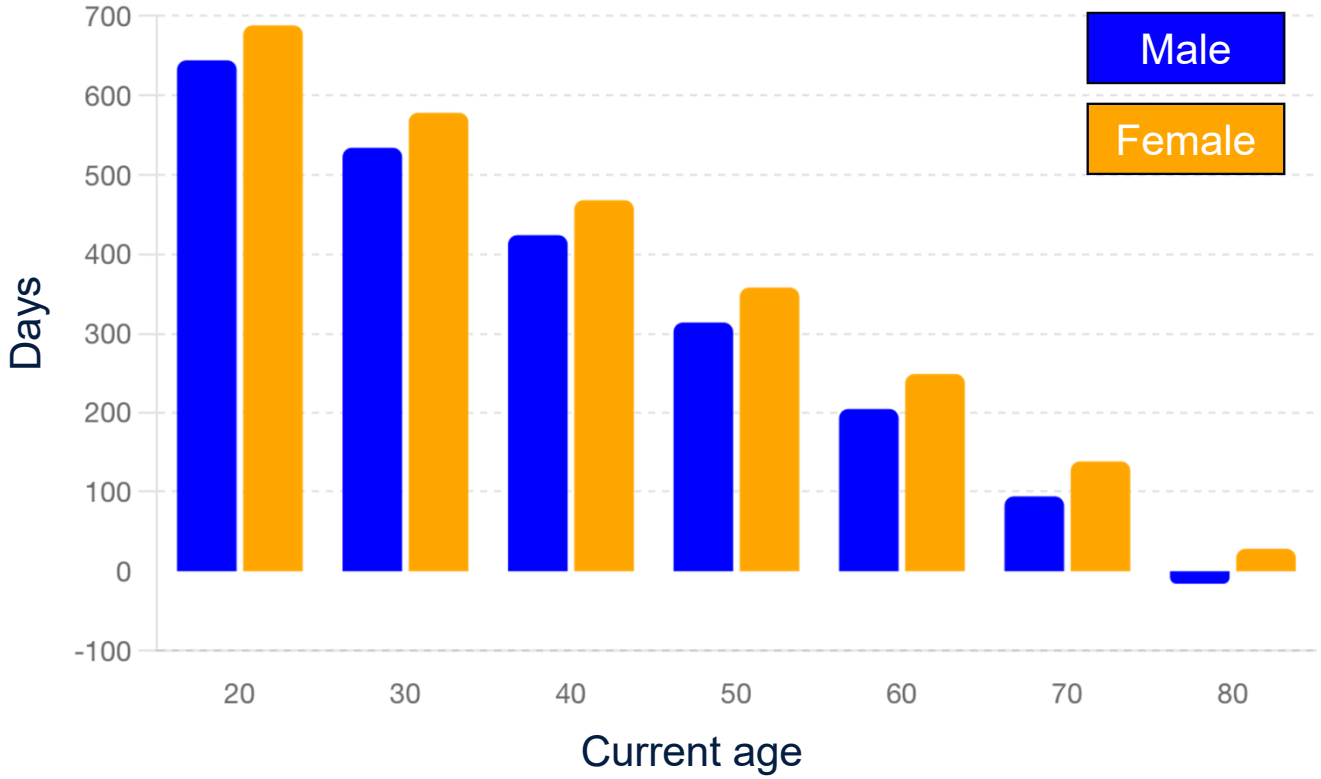

IDEALISTIC
Altruism
Appetite
EXO-CENTRIC
EGO-CENTRIC
Acumen
Ambition
PRAGMATIC


What drives you?

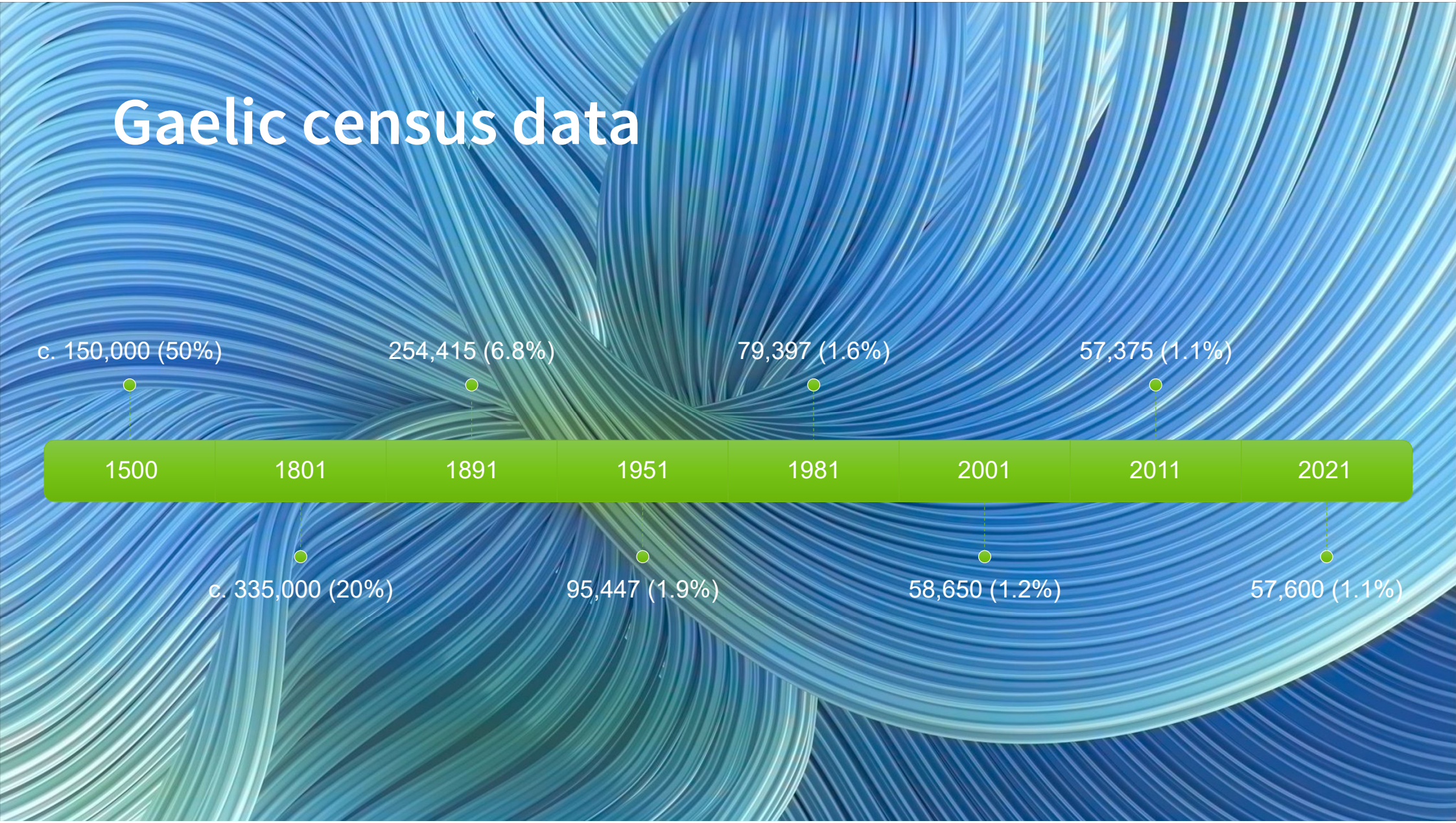
Average days left according to age (UK)



Average number of >20c days left (Edinburgh)



Gaelic census data

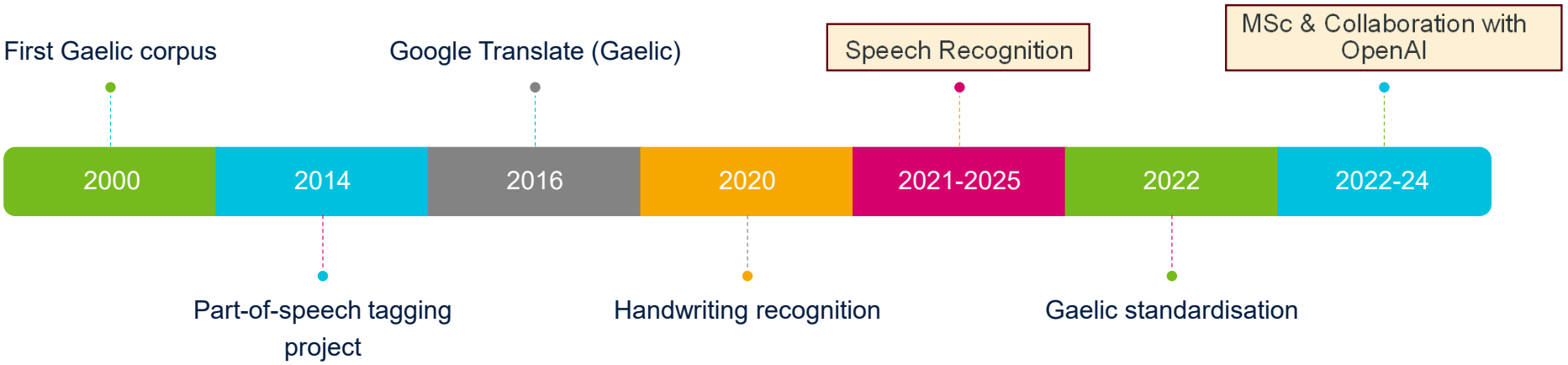


duolingo

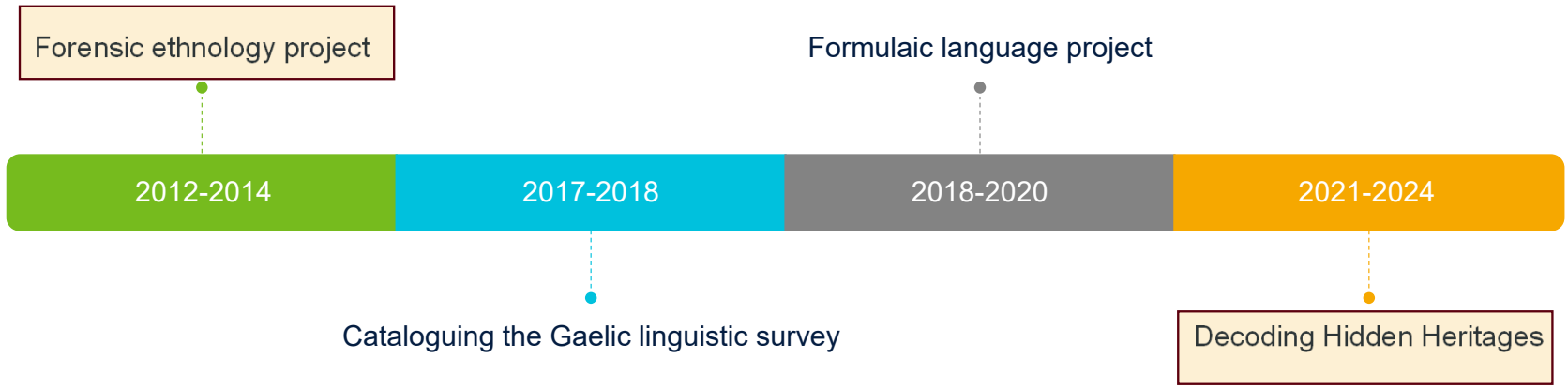
?

PhD: Gaelic register variation (2002)

Language Processing



Digital Humanities



Forensic Ethnology (2012-2014)

- Donald John MacDonald collection (School of Scottish Studies Archives)
  - 1953 through 1958
  - 26 bound volumes (Upper Library)
  - > 1500 pages attributed to Duncan and Neil
  - Largest manuscript collection of SSSA

Dice similarity scores

The highest similarity is between D-1944 and N-1955

D-1944 = a published text transcribed from Duncan by K. C. Craig

N-1955 = an unpublished ‘transcription’ in DJM’s mss from Duncan’s brother Neill

Shows indisputable evidence of copying

Text 1	Text 2	Relation
D-1944	**N-1955**	**0.87**
D-1944	D-1953	0.82
D-1953	D-1950	0.82
D-1944	D-1950	0.81
D-1947	D-1944	0.81
D-1947	D-1953	0.80
D-1947	D-1950	0.79
N-1955	D-1953	0.78
D-1944	D-1936	0.77
N-1955	D-1950	0.77
D-1947	N-1955	0.77
D-1947	D-1936	0.76
N-1955	D-1936	0.76
D-1953	D-1936	0.75
D-1950	D-1936	0.75

Visual evidence 1: ‘banais’

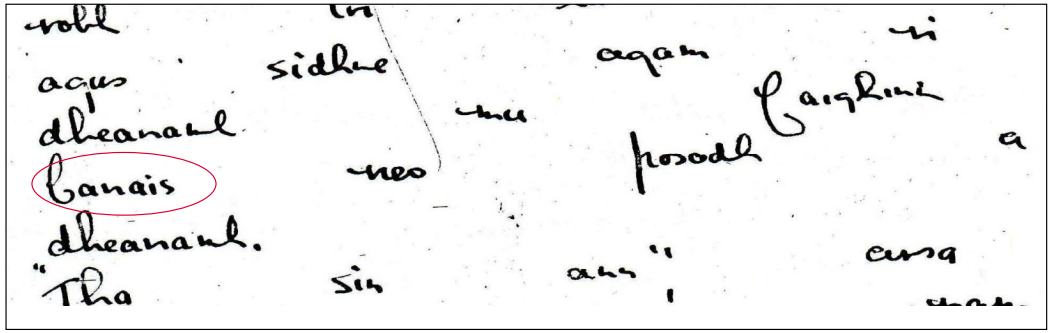
D-1944 pg 24

Chaidh mo chur a chadal ann an sobhal fada fàs an oidhche sin a rithist. Agus thàinig guth chon na h-uinneig, agus dh’ éibh e gu robh dà latha seilgeadh agus sìdhneadh fhathast agam ri dhianamh mum faighinn banais no pòsadh.

‘Tha sin ann,’ orsa mi fhìn, ‘agus nam biodh an còrr ann, cha rachadh tus’ a dh’ inns’ an ath sgeul.’

24

DJM MSS pg 3557



Visual evidence 2: ‘bainis’

D-1944 pg 26

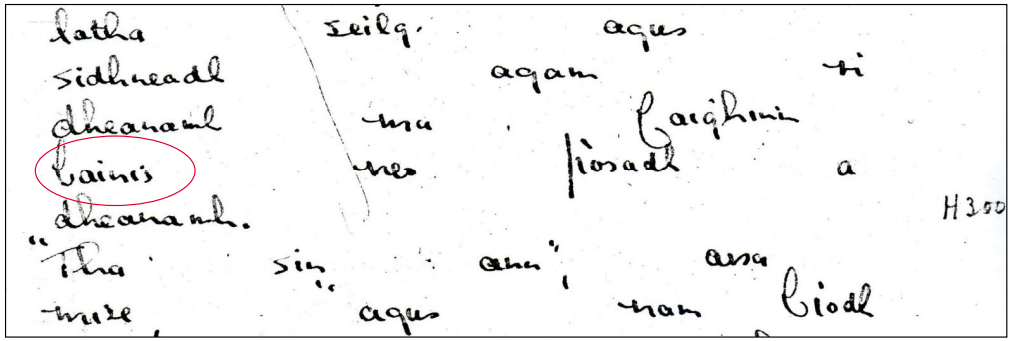
dhomh gu robh latha seilgeadh agus sìdhneadh agam ri dhianamh mum faighinn bainis no pòsadh.’

‘Tha sin ann,’ orsa mi fhìn, ‘agus nam biodh an còrr ann, cha rachadh tus’ a dh’ inns’ an ath sgeul.’

‘Cha bu lughaide do chuid-sa a gheasachd an eilein sin,’ ors am fear eile, ‘mura cuirinn-sa na geasaibh ud ort, chuireadh fear eil’ ort iad.’

26

DJM MSS pg 3557



What I learnt from this project

1. A Gaelic handwriting recogniser would be useful
2. A Gaelic speech recogniser would be *very* useful
3. There is a huge amount of data in the School of Scottish Studies Archives

The School of Scottish Studies Archives

- Research institution inaugurated in Jan 1951
- Tasked with collecting and interpreting Scotland's tangible and intangible cultural heritage

Decoding Hidden Heritages

UKRI Arts and Humanities Research Council

IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn

THE UNIVERSITY of EDINBURGH

Open to the world

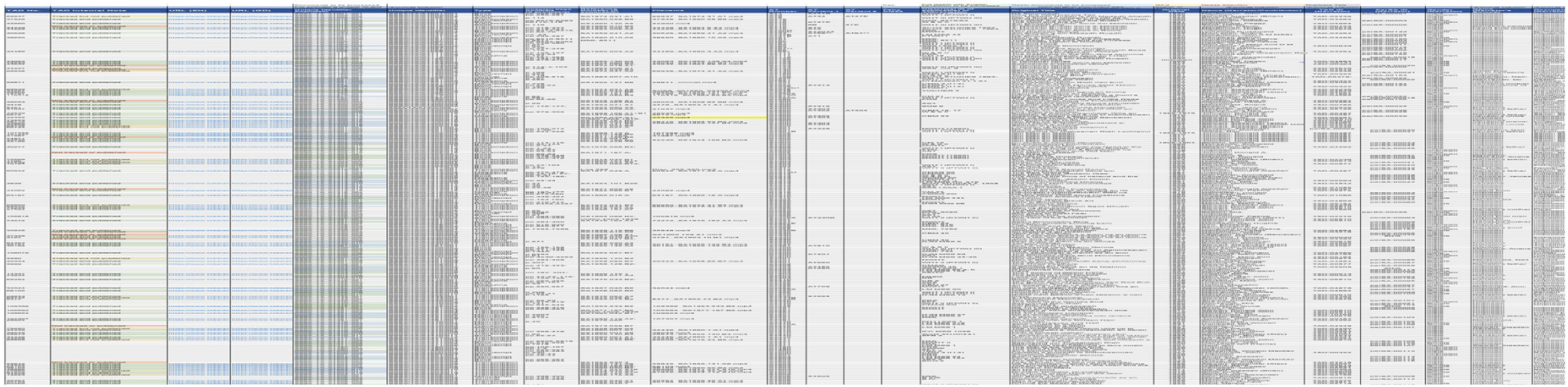
DHH Research Questions

- How can AI help to digitise and automatically recognise transcriptions of Gaelic and Irish folklore?
- What can these texts tell us about Scotland and Ireland's shared history and culture?
- How can traditional folkloristic and AI-based methods help to bring this to light?

DHH Principal Activities

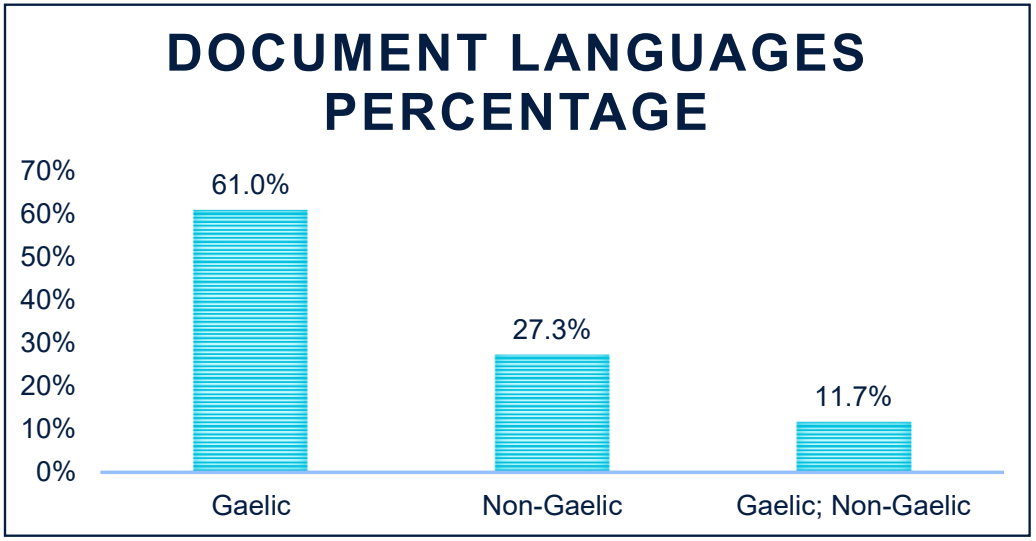
- Digitising and generating metadata for the Tale Archive at the School of Scottish Studies Archives (University of Edinburgh)
- Recognising printed and handwritten materials semi-automatically
- Clearing material for dissemination (e.g. copyright permission)
- Analysing the material with traditional methods and AI-driven techniques

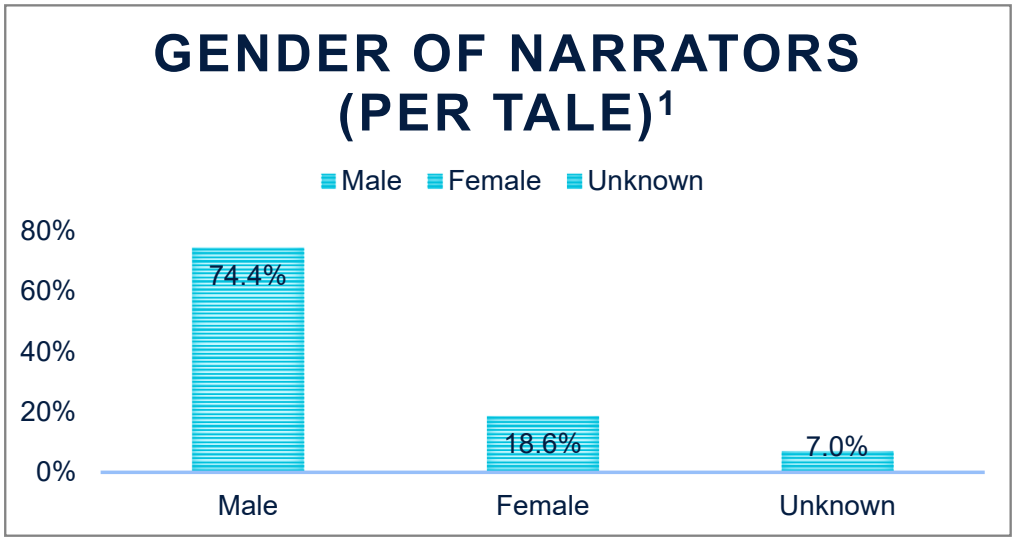
SSSA Tale Archive: Basic Stats



Scanned images	Metadata
Document images scanned: 26,201 Index cards scanned: 12,419 Total scanned images: 38,620	Document metadata rows completed: 3,860 Index cards metadata completed: 5,384 Total metadata entries: 9244

SSSA Tale Archive: Basic Stats (cont.)

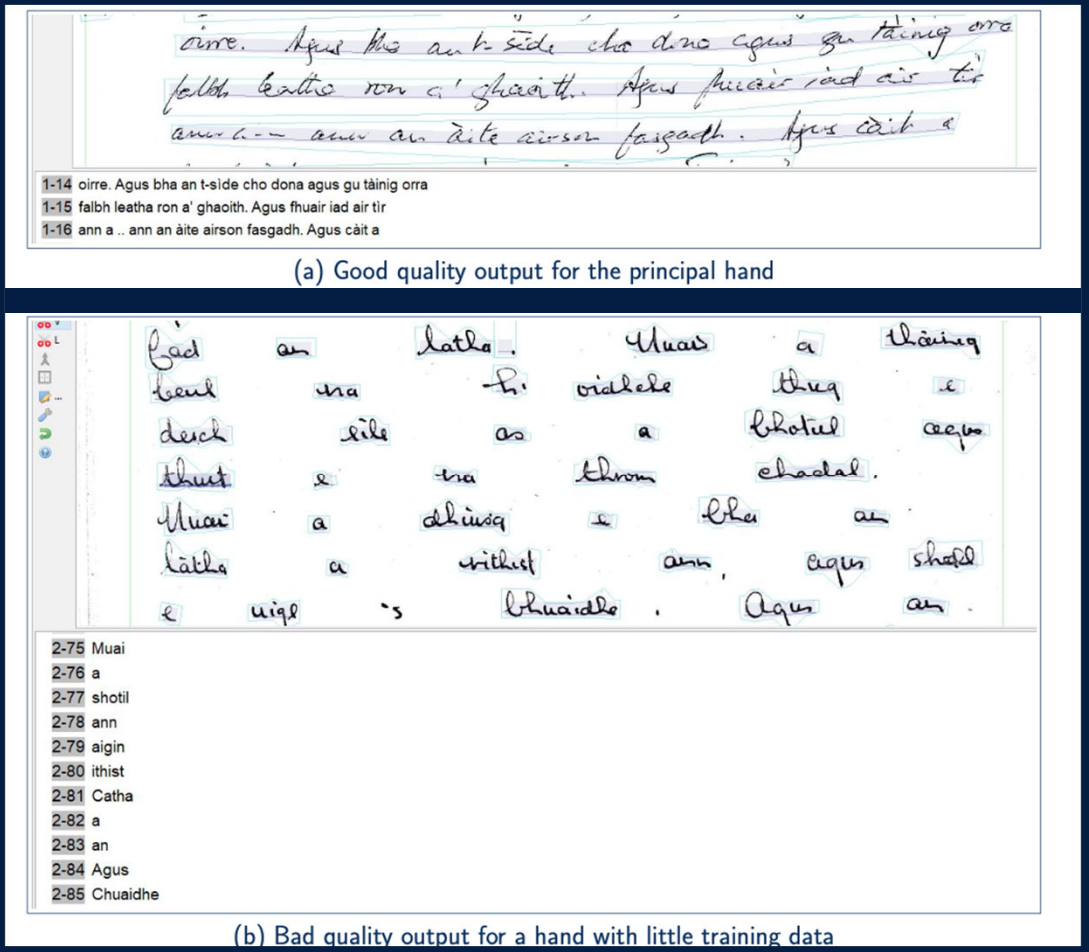




[1] Decoding Hidden Women: Feminist digitisation practices in the Tale Archive

Handwriting Recognition

- Uses Transkribus platform
- Pages recognised: 8489
- Words recognised: 1.9M
- Accuracy levels
  - Char level accuracy: 97.5%
  - Word level accuracy: 95%



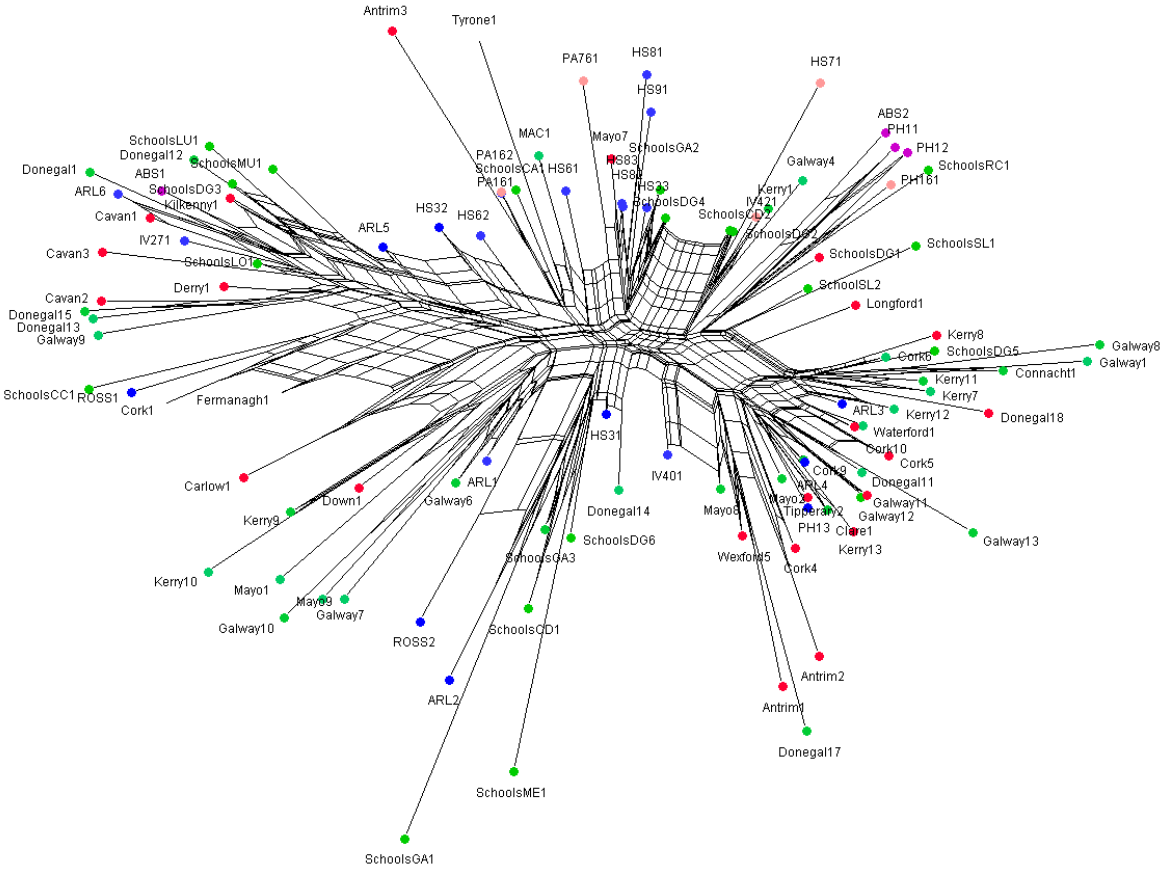
(a) Good quality output for the principal hand

(b) Bad quality output for a hand with little training data

Good vs bad quality Transkribus output

DHH text mining workstreams

1. Phylogenetic analysis
2. Topic modelling / automatic clustering
3. Automatic identification of formulaic language

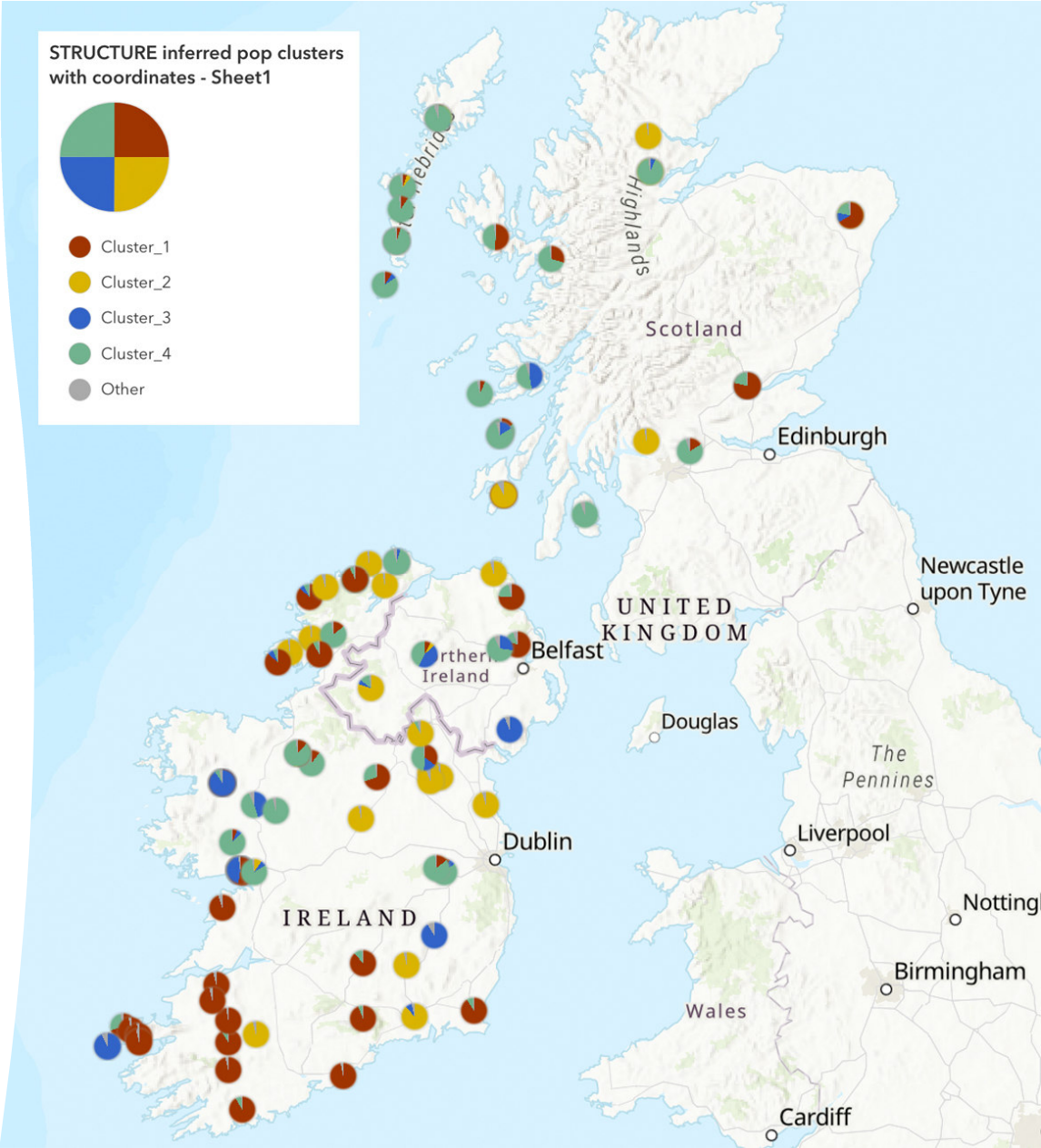


Phylogenetic network analysis: ATU 503

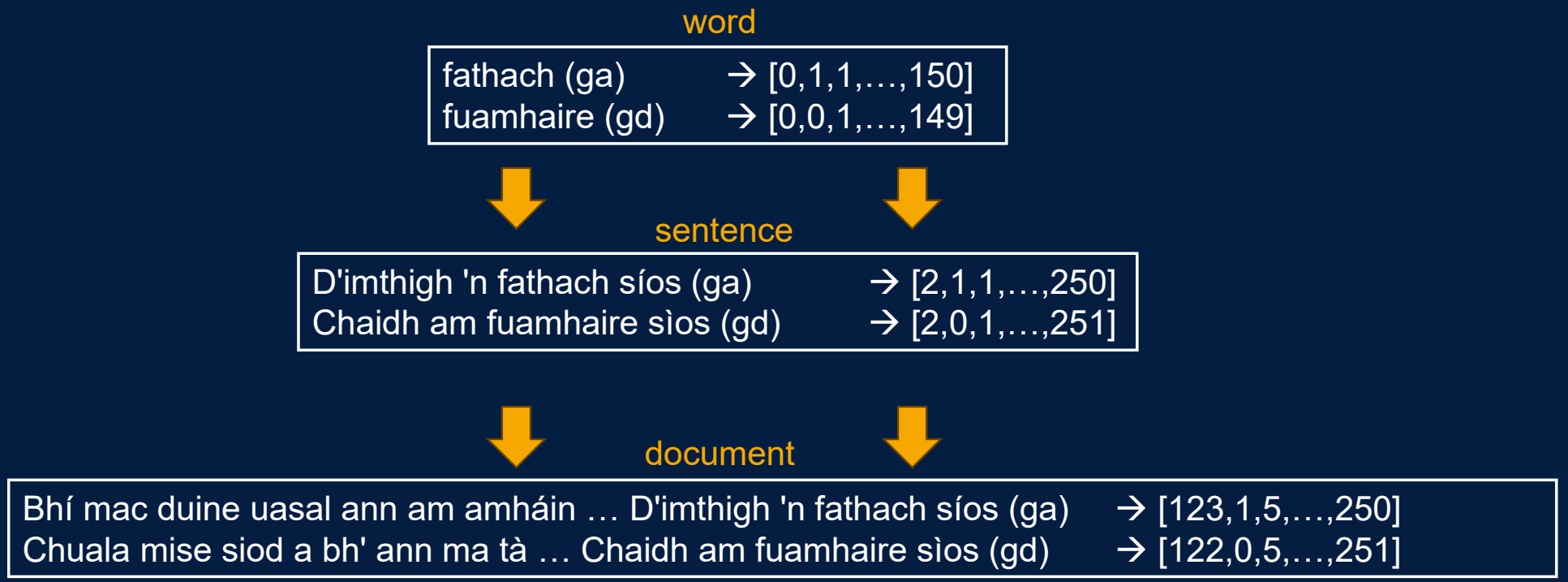
- Network analysis has a highly reticulated structure
- Suggests complicated patterns of transmission between communities and linguistic traditions
- Bottom line: lots of borrowing and mixing of motifs

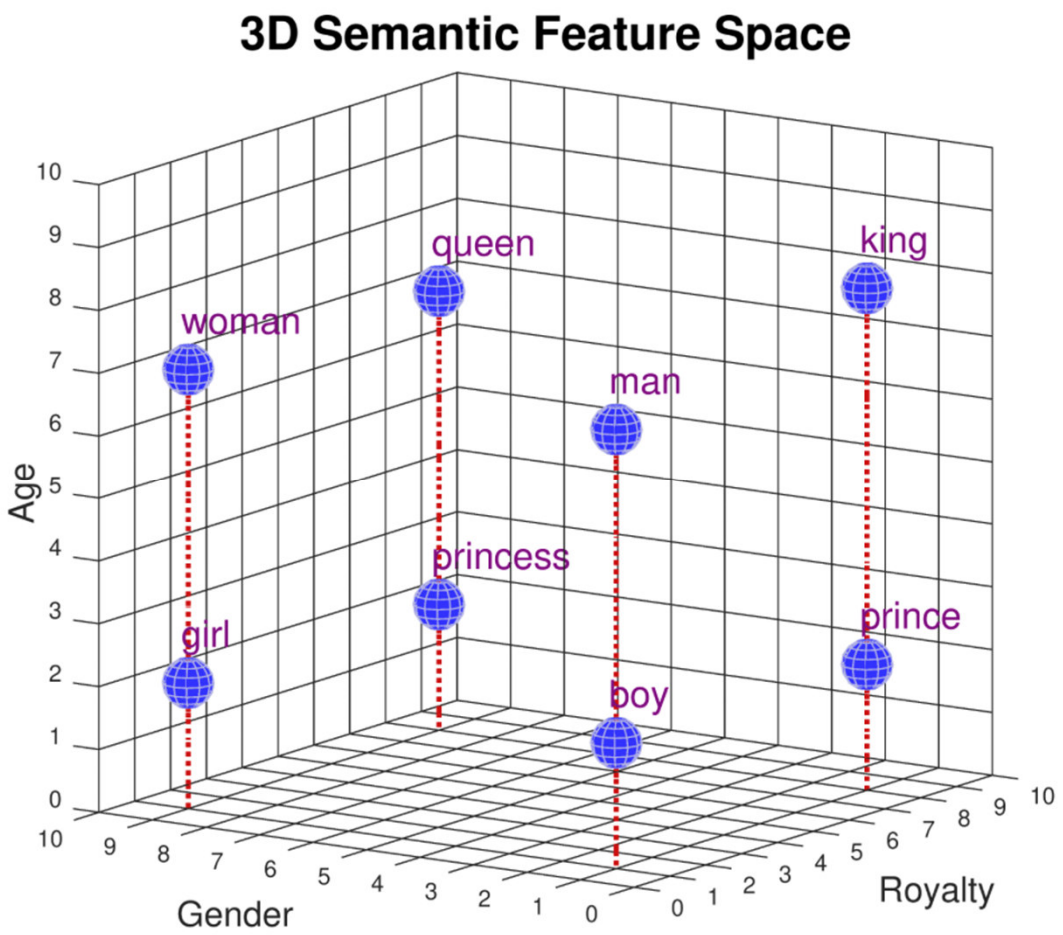
Diatopic cluster distribution: ATU 503

- Shows some effect of geographical proximity (<50km)
- The effect of linguistic similarity, however, is stronger
  - Linguistic groups (e.g. Gaelic speakers in Scotland) have more in common with themselves than with other, geographically proximate groups



Word to Sentence to Topic Embeddings



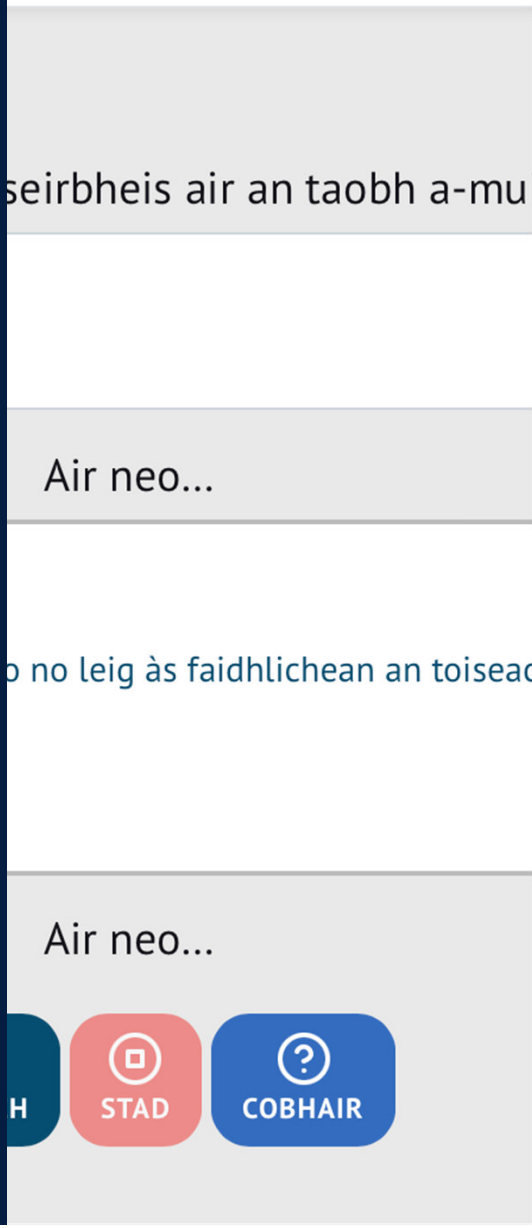

3D Semantic Feature Space
queen
king
woman
man
princess
prince
girl
boy
Age
Gender
Royalty
0
1
2
3
4
5
6
7
8
9
10

Decoding Hidden Heritages: Next steps

- Run repertoire analyses using the coordinates and metadata
  - Particularly interested in gendered distinctions
- Wrap up the phylogenetics and formulaic language research
- Launch the project website (Q4, 2024)
  - Will provide searchable texts and PDFs
- Prepare *Computational Folkloristics,* an edited collection of papers from the project

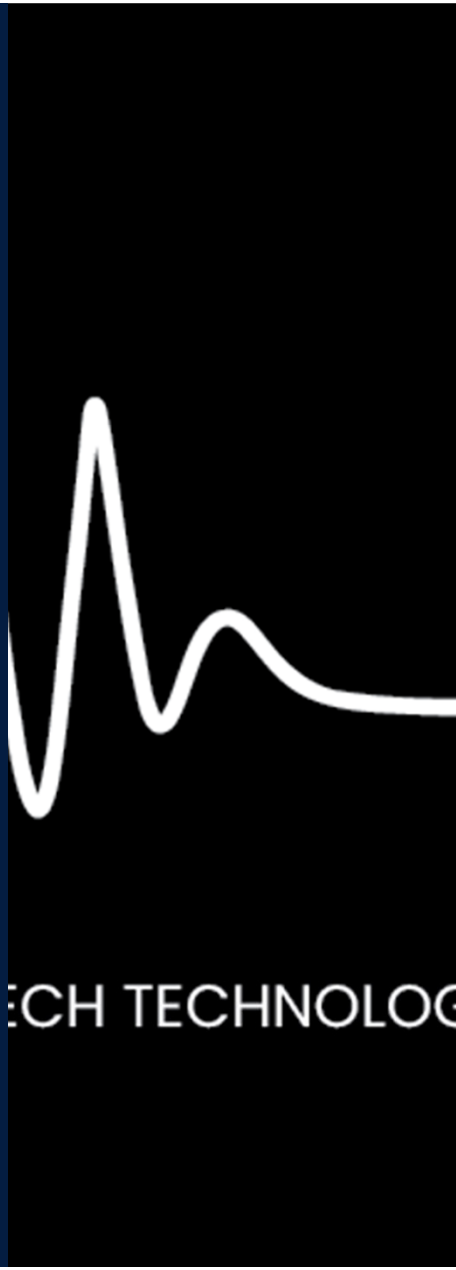
Gaelic speech recognition

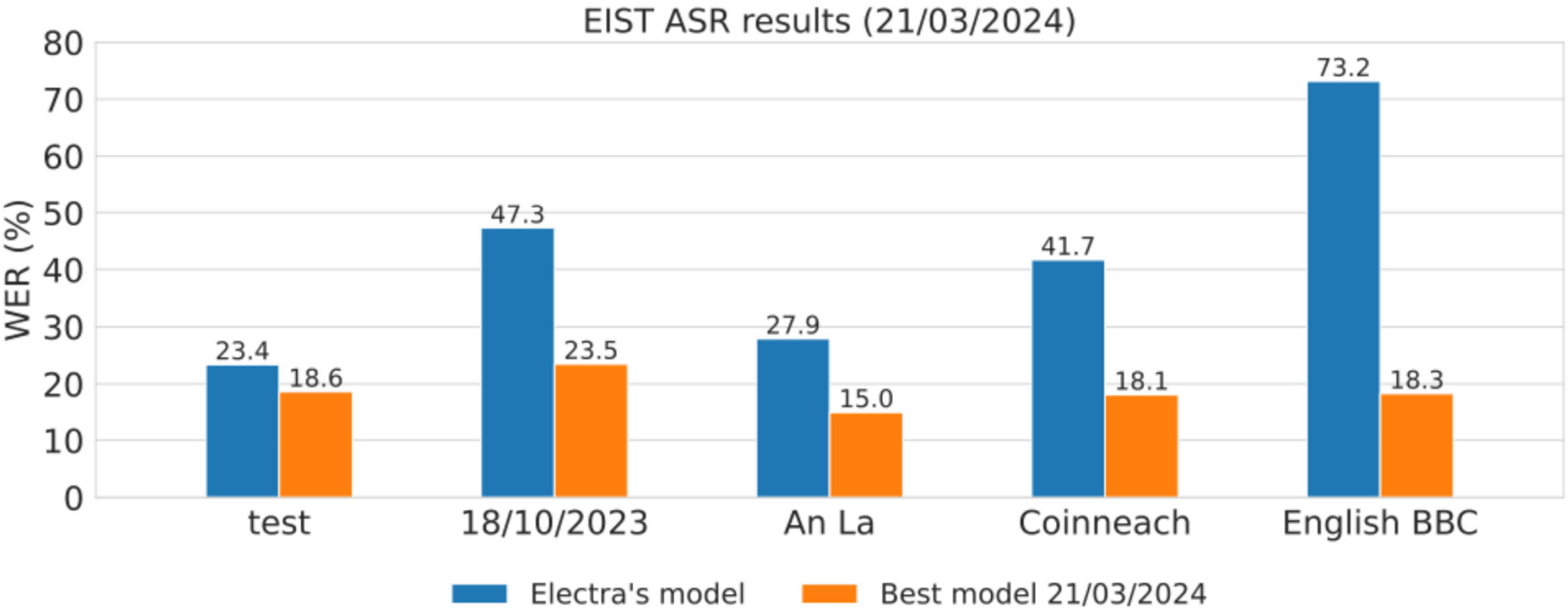
Scottish Government
Riaghaltas na h-Alba



ÈIST project: Current aims

1. Produce a subtitling system for the BBC (radio and television)
2. Produce a transcription app for Gaelic school children
3. Expedite transcriptions of ethnographic material for Tobar an Dualchais / Kist o Riches




EIST ASR results (21/03/2024)
WER (%)
80
70
60
50
40
30
20
10
0
23.4
18.6
47.3
23.5
27.9
15.0
41.7
18.1
73.2
18.3
test
18/10/2023
An La
Coinneach
English BBC
Electra's model
Best model 21/03/2024

```
N_AGENT = 30 # number of agent generate sentences

SYSTEM_MSG_SUMMARY = "Your role is to summarise the given Scottish Gaelic in 3 to 4 sentences in
SYSTEM_MSG = "Your role is to generate news story in Scottish Gaelic given the English summary."

ROOT_PATH = os.getcwd()
DATA_PATH = os.path.join(ROOT_PATH, "data")
INPUT_FILE_NAME = "bbc_dongge_with_timestamps.jl"
OUTPUT_PATH = os.path.join(ROOT_PATH, "output")
```

```
with open(os.path.join(ROOT_PATH, DATA_PATH, INPUT_FILE_NAME), 'r', encoding='utf-8') as file:
    content = file.read()
df_base = pd.read_json(content, lines=True)
df_base
```

```
/var/folders/mt/z9020bs533j19ypwxz5g2kdm0000gq/T/ipykernel_72101/2124592222.py:3: FutureWarning
version. To read from a literal string, wrap it in a 'StringIO' object.
  df_base = pd.read_json(content, lines=True)
```

MSc SLP + OpenAI project

Many thanks to our funders and collaborators
Taing mhòr dhan luchd-taic againn

Am Faclair Beag, Ceòlas Uibhist Ltd, Digital Archive of Scottish Gaelic (DASG: U of Glasgow), European Ethnological Research Centre, Faclair na Gàidhlig, Grace Note Publications, Guthan nan Eilean / Island Voices, LearnGaelic (MG Alba), National Folklore Collection (University College Dublin), OpenAI, Ruairidh MacIlleathain, Sabhal Mòr Ostaig, The National Library of Scotland, The School of Scottish Studies Archives, Tobar an Dualchais / Kist o Riches, University of Bangor, University of the Highlands and Islands, University of Glasgow and Riaghaltas na h-Alba (The Scottish Government)

UKRI Arts and Humanities Research Council

SOILLSE

Scottish Government Riaghaltas na h-Alba

DDI Data-Driven Innovation

THE UNIVERSITY of EDINBURGH

Open to the world

THE UNIVERSITY *of* EDINBURGH

Ceud Mìle Taing
Thank you

Open to the world