Mobilising the Oxford Text Archive in the Development of Computational Linguistic Tools

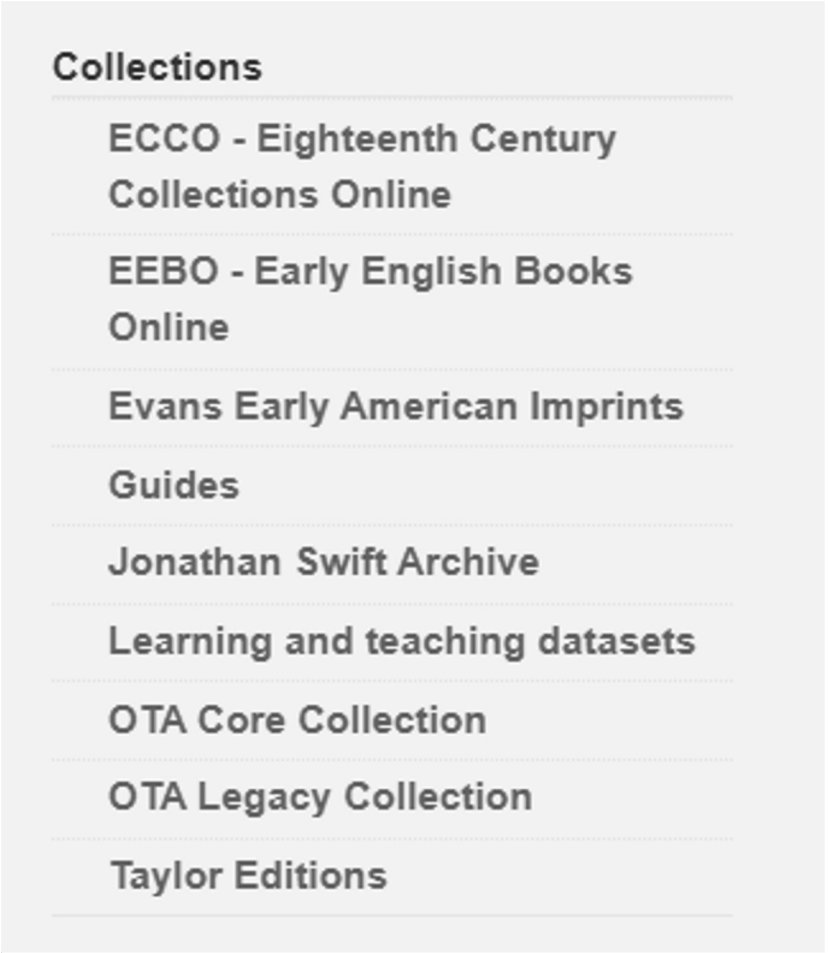
OXFORD
TEXT ARCHIVE

<ota since="1976"/>

Fiesole Retreat 2026

Oxford Text Archive (OTA)

- Archive of electronic texts created for linguistic and literary research.
- Founded in 1976 by Lou Burnard and Susan Hockey of the Oxford University Computing Services.
- Contains over 70,000 texts from a range of time periods and languages.
- Contains digital editions, electronic corpora, and analytical tools.
- A national repository for literary and linguistic resources.
- Currently funded by the AHRC iDAH programme.



Text Creation Partnership

- Partnership between University of Michigan Library, Bodleian Libraries at the University of Oxford, ProQuest, and the Council on Library and Information Resources.
- Created standardized, accurate XML/SGML encoded electronic text editions of early print books.
- TEI transcribed files of texts in:
  - Early English Books Online
  - Gale Cengage's Eighteenth Century Collections Online (2,231 texts)
  - Readex's Evans Early American Imprints

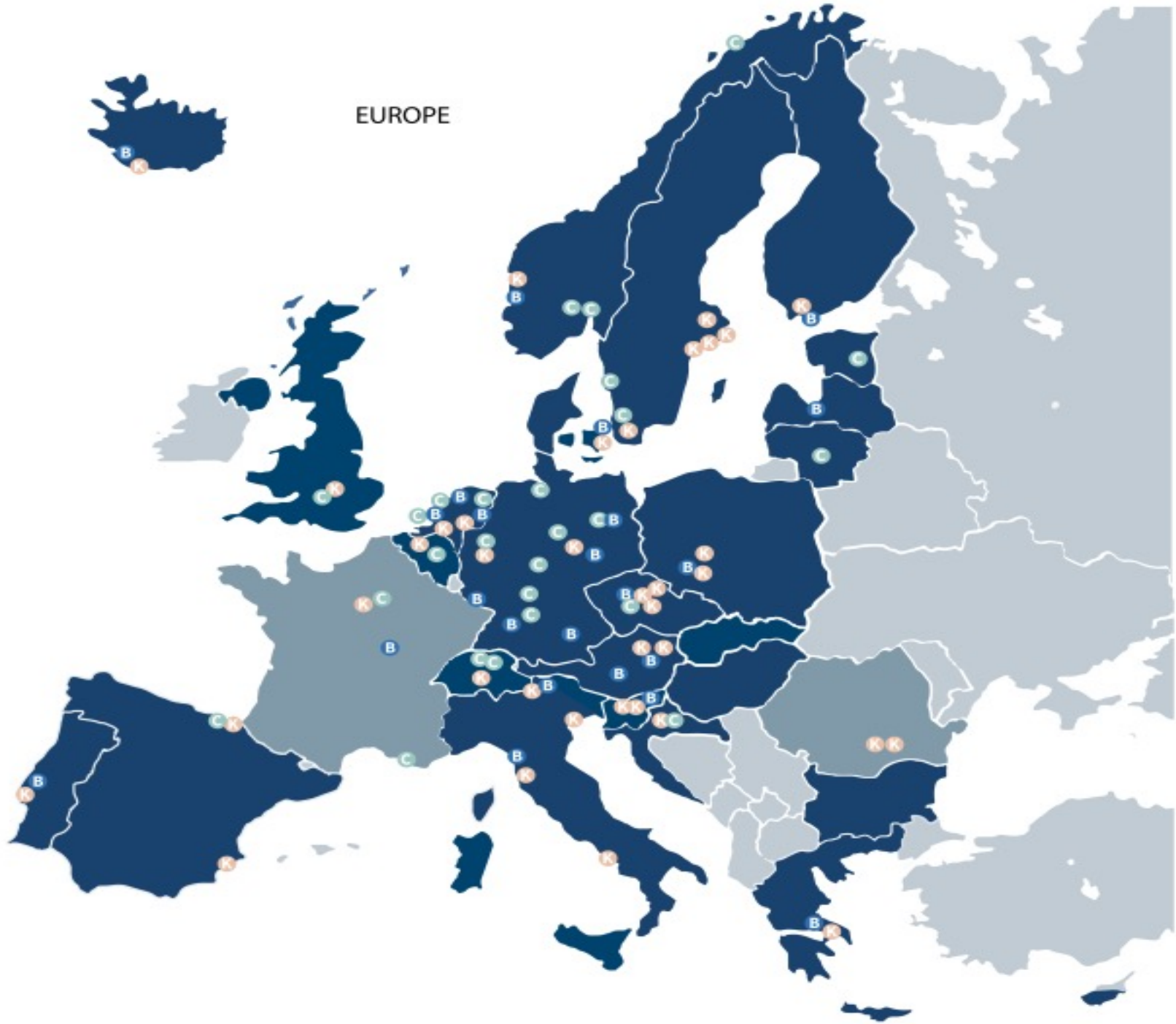
Strengths and challenges

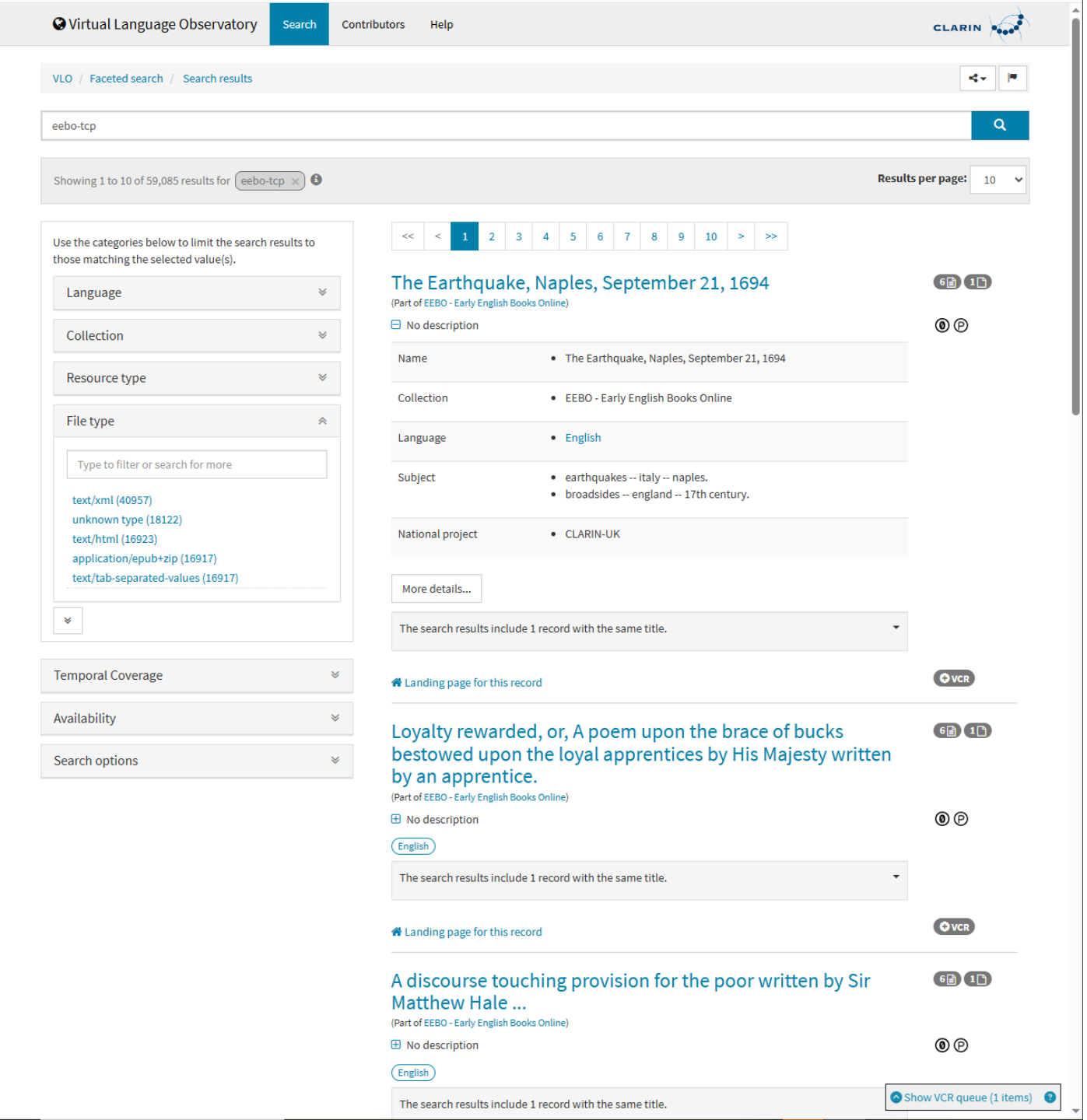
- A large, varied collection of textual data.
- Academically sourced texts hopefully ensuring a level of quality and ethics.
- Carefully curated metadata with copyright information.
- Persistent identifiers.
- A collection built over 50 years!

- Not balanced; no conscious sampling or building strategy.
- 50 years of metadata standards.
- 50 years of encoding standards.
- Lack of interoperability between deposits.
- Multiple migrations and unstable institutional identity.
- Ever-evolving …

CLARIN

- ERIC members
- Observers
- Countries with participating centres
- B Centre Providing Data
- C Centre Providing Metadata
- K Knowledge Centre





Enhancement through CLARIN

- Searching in VLO via shared metadata.
- Interoperability – Language Resource Switchboard.
- Improving metadata and updating it in response to AI.

CLARIN-UK

Infrastructure for Digital Language Resources and Tools

Home | Centres | Resources | Knowledge Centre | Deposit | People | News | Events | Blog | About | CLARIN in other countries

- British Library
- Centre for Corpus Research (Birmingham)
- Centre for Translation Studies (Leeds)
- CorCenCC (Cardiff)
- CTTR University of Wolverhampton
- Faculty of Linguistics, Philology and Phonetics (Oxford)
- King's College London
- Lancaster University
- Language Technology Group (Edinburgh)
- Natural Language Processing Group (Sheffield)
- School of Critical Studies (Glasgow)
- School of Humanities (Coventry)

https://www.clarin.ac.uk/

Centres

CLARIN-UK is a consortium of centres of expertise involved in research and resource creation involving digital language data and tools. The consortium includes the national library, and academic departments and university centres in linguistics, languages, literature and computer science. The members are listed below. More are very welcome to join - please get in touch!

Current members of the consortium

British Library

Centre for Corpus Research

University of Birmingham

Centre for Translation Studies

University of Leeds

CorCenCC

The Welsh National Corpus team at Cardiff University

CTTR University of Wolverhampton

Wolverhampton University

Faculty of Linguistics, Philology and Phonetics

University of Oxford

King's College London

King's Digital Lab and Department of Digital Humanities

Lancaster University

CASS and UCREL

Language Technology Group

University of Edinburgh

NLPG

Natural Language Processing Group, University of Sheffield

School of Critical Studies

University of Glasgow

School of Humanities

Coventry University

DR-LIB K-Centre

- Digital Resources for the Languages in Ireland and Britain.
- Offer advice concerning digital resources for the native languages of Britain and Ireland with a special focus on lesser-resourced languages.
- Build a network of experts on these languages.
- Develop our resources in these languages.
- Partners: Cardiff University, Dublin City University, Lancaster University, University of Edinburgh, University of Oxford.

Language	Name	Description
Breton	An Drouizig	Tools for translation, spellcheckers, Breton keyboard, Breton fonts, Breton dictionaries.
Breton	Porched niverel ar brezhoneg	Breton language technology portal, promoting various digital tools and resources.
Cornish	BBC news in Cornish	
Cornish	Gerlyver Kernewek	Cornish dictionary.
Cornish	Korpus kernewek	Cornish corpus.
English	DANTE lexical database	Corpus-based description of the core vocabulary of English.
English Welsh, etc.	PymUSAS	Python Multilingual Ucrel Semantic Analysis System.
English Irish Welsh	Seamless Communication	Translation and S2T Models.

https://www.clarin.ac.uk/article/digital-resources-languages-ireland-and-britain

Unlocking AI for the Languages in Britain and Ireland

- Gather existing generic AI skills training for the humanities, make them more accessible, and promote them.
- Unlock good quality training data for training specialized language models.
- Improve language models for lesser-resourced languages.
- Run experiments on LLM creation and fine-tuning using data deposited in the OTA.

Training materials

https://www.clarin.ac.uk/ai-skills-humanists

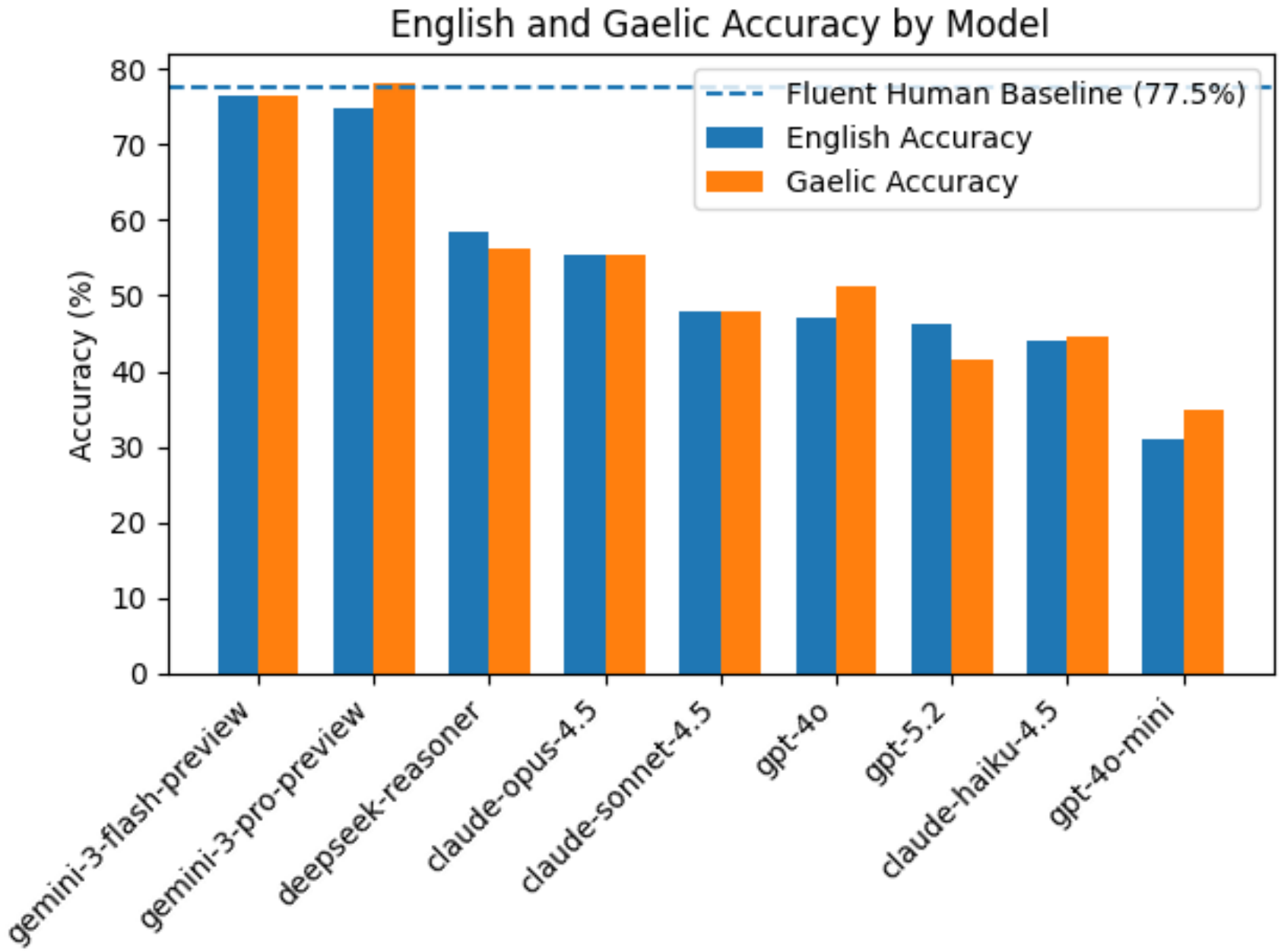
- RANLP2025 conference proceedings
  - https://ranlp.org/ranlp2025/index.php/tutorials/ Includes tutorials on "Legal NLP in the LLM era" and "Building Affordable Language Models with Limited Resources", plus a summer school on deep learning and LLMs (https://ranlp2025-summer-school.github.io/)
- CLARIN Learning Hub https://www.clarin.eu/content/learning-hub
  - A catalogue of 35 learning resources, example related courses are:
  - https://www.clarin.eu/content/ai-understand-and-connect-people (introductory AI concepts)
  - https://www.clarin.eu/content/processing-texts-and-corpora (Processing Texts and Corpora)
  - https://www.clarin.eu/content/natural-language-processing-methods-0 (Natural Language Processing Methods)
- Lancaster Corpus MOOC: https://www.edx.org/learn/social-sciences/lancaster-university-corpus-linguistics-and-new-technologies-data-language-society
- Programming Historian https://programminghistorian.org/
  - Online tutorials in English, Spanish, French and Portuguese
  - Includes AI, machine learning tutorials in facial recognition, clustering documents, and image classification
- UCREL NLP summer schools 2024 GitHub repositories: https://github.com/UCREL/USS2024
  - These are summer schools which ran in 2024 aimed at computer scientists and early career NLP researchers, and includes several practical sessions which are LLM related

Benchmarking multilingual LLMs

- Evaluating multilingual LLMs ‘shadow’ capabilities in under-resourced languages they do not officially support.
- Can multilingual LLMs answer questions about under-resourced languages without resulting to English translation?
- Need a model that can account for linguistic and cultural competency.
- Ongoing work in Gaelic, Irish, and Welsh.

GaelEval

- Multidimensional benchmark for Gaelic.
- Proprietary models perform better than open-weight models.
- Consistent advantage gained by in-language prompting.



See full paper at https://arxiv.org/pdf/2604.02135

Future ambitions

- Further benchmarking exercises.
- Recommendations and general principles for benchmarking exercises in other under-resourced languages.
- Finetuning and RAG with the OTA.
- AI-enhanced search; AI-enhanced standards.
- Linking our data with other data services in the UK.

Lessons learned

Enduring value of authentic linguistic data – especially historical data that cannot easily be replaced by synthetic data.

Importance of updating metadata and rebranding.

Not everyone needs to build an LLM.

Need to create pipelines for accessing and processing legacy data.

Thanks – and please do get in touch!

- The Oxford Text Archive at LLDS:
  - https://llds.ling-phil.ox.ac.uk/
- CLARIN-UK website and blog:
  - https://www.clarin.ac.uk/
- CLARIN-UK mailing list:
  - https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=CLARIN-UK
- OTA mailing list:
  - oxford-text-archive-subscribe@maillist.ox.ac.uk
- DR-LIB help desk:
  - contact-dr-lib@forum.clarin.eu

martin.wynne@ling-phil.ox.ac.uk

megan.bushnell@ling-phil.ox.ac.uk