RESEARCH PROJECT FUND

FINAL REPORT: 'Developing Automatic Speech Recognition for Scottish Gaelic'

TEAM

Dr William Lamb (Principal Investigator): w.lamb@ed.ac.uk
Prof Conchúr Ó Giollagáin (Co-investigator): cog.smo@uhi.ac.uk
Dr Mark Sinclair (Consultant): mark.sinclair@quoratetechnology.com
Mr Gordon Wells (Co-investigator): gw.smo@uhi.ac.uk
Ms Lucy Evans (Research Assistant)

BACKGROUND

This was a collaborative, 6-month project that aimed to develop a general **Language Model (LM)** and working **Automatic Speech Recognition (ASR) system** for Scottish Gaelic. Given an input of text, an LM is used to predict the next likely word. ASR systems, also called Speech-To-Text (STT) systems, convert spoken audio to text. The applications for these cornerstones of modern language technology are vast, and include: voice assistants (Alexa, Siri); predictive texting; video subtitling; and automatic transcription. In the long term, the team intends to enable the automatic generation of transcripts and/or subtitles for pre-existing Gaelic recordings and videos. This will add value to these resources by rendering them searchable by word or topic.

The project brought together researchers from the University of Edinburgh, the University of the Highlands and Islands and Quorate Technology Ltd. The team also included a Research Assistant who had recently completed the MSc in Speech and Language Processing at the University of Edinburgh.

The project articulated with two prior funded projects at the University of Edinburgh: the Gaelic Part-of-Speech Tagging Project[1] and the Gaelic Handwriting Recognition Project.[2] It also incorporated two recent spoken language ethnographic recording projects conducted under the auspices of UHI/Soillse (*Saoghal Thormoid* and *Stòras Beò nan Gàidheal*). Its orientation was towards enhancing computer-assisted language learning (CALL), broadening communicative domains in Gaelic (e.g. via interaction with smart devices) and facilitating automatic transcription and translation.

OBJECTIVES

The project successfully completed its main research objectives, which were:

- Collate as much speech and text data as possible for Scottish Gaelic
- Process it into a common format to establish a formalised training corpus
- Use the training corpus to develop an LM
- Develop and evaluate a baseline, standardised Scottish Gaelic recipe for the open-sourced Kaldi ASR toolkit

It also accomplished several adjunct objectives, which are discussed in further detail in 'Outputs and Future Steps':

- Conference talk
- Links made to industry and other research teams
- Major funding application

---

[1] https://www.aclweb.org/anthology/W14-4601.pdf

[2] https://blogs.ed.ac.uk/garg/2020/02/04/cif_project_feb2020/

The only objective remaining to be fulfilled is writing a peer-reviewed research article. The team intends to complete this late in 2021 or early in 2022.

OUTPUTS AND FUTURE STEPS

The most substantial output to date is that a forced alignment tool – an intermediary step in the ASR development – was used to automatically subtitle 75 videos from Island Voices – Guthan nan Eilean, as well as 3 from Stòras Beò and 2 from Guth nan Siarach. A beneficial side-effect is that the subtitles can be viewed in any language covered by Google Translate. This has greatly widened their accessibility and enhanced them as language learning tools. These developments were detailed in several blogs on the GARG and Island Voices sites. A demonstration of the alignment tool on Twitter was viewed over 6000 times. On the back of this publicity, William Lamb and Gordon Wells gave several interviews to BBC Radio nan Gàidheal and BBC Alba for the programmes 'Aithris na Maidne', 'Coinneach Macìomhair' and 'An Là'. Dr Lamb also gave an invited lecture on the project in March 2021, as part of the Formal Approaches to Celtic Linguistics series (University of Arizona).

In Dec 2020, Dr Lamb secured additional funding (£49,991) to expand the project and extend it to 31 July 2021. Two new RAs were hired for this phase. The funding came from a DDI/SFC 'Building Back Better' Open Call grant. In March 2021, the team developed a working ASR system for the language, which is now being trialled for core transcription tasks (e.g. transcribing Gaelic narrative audio from Tobar an Dualchais). The team also submitted a £400k funding application to the AHRC for a Digital Humanities project, involving academic and third-sector partners in Ireland and Durham. If the bid is successful, this project will utilise some of the technologies developed as part of the Soillse project (e.g. the LM). The team is currently seeking additional funding to continue the work on ASR beyond 31 July 2021 and is in conversations with CALL Scotland about approaching the Scottish Government and the Scottish Qualifications Authority. Regardless of whether additional bids are successful, the team intends to release the tools developed so far in August 2021.

16 April 2021

Timeline

1 Jun 2020: Initial conversations with School of Scottish Studies towards release of Gaelic narrative audio for project
July 2020: Interviews for RA post
1 Aug 2020: Project commences, along with initial preparation and collation of corpus data; contacts made with Abair (Ireland), Duolingo and Cereproc
1 Sept 2020: RA begins, based at Quorate:[3] induction and discussion of project aims and timeline; initial publicity – institutional press release, tweets and blogs
5 Sept 2020: Phase 1 – Data collation and initial model development
1 Dec 2020: Phase 2 – Develop Kaldi recipe for ScG and LM
5 Jan 2021: Phrase 3 – Testing other approaches
29 Jan 2021: RA finishes – submits report to research team
1 Feb 2021: Formal evaluation of models and dissemination of early results (blogs, tweets)
1 June 2021: Project end - final report delivered; public-facing LM and ASR models released on GitHub
Summer 2021: Delivery of conference paper / peer-reviewed publication; additional press release and publicity timed to publication of results

Personnel

Dr William Lamb (University of Edinburgh) is a Senior Lecturer in Celtic and Scottish Studies, in the School of Literatures, Language and Cultures. He has over 20 years of experience carrying out data-intensive work on the Gaelic language. As PI, he has successfully led research projects into natural language programming and archival cataloguing, including 'The Gaelic Part-of-Speech Tagging Project' (2013-15: Carnegie Trust Large Grant £40k; Bòrd na Gàidhlig £19k), 'Cataloguing the Gaelic Materials of the Linguistic Survey of Scotland' (2018-19: John Lorne Campbell Trust £23,814) and 'Automatic Handwriting Recognition for Scottish Gaelic' (2019: UoE Challenge Investment Fund £12k). Recently, he was also a collaborator on the £130k inter-university consultation, 'Gaelic Corpus Development' (Bòrd na Gàidhlig). With Dr Mark Sinclair, he piloted the first neural network for Scottish Gaelic (2016: 'Developing embedding models for Scottish Gaelic') and has authored numerous research publications in the areas of Gaelic ethnology, linguistics and language technology.

---

[3] *Pending relaxation of social distancing rules relating to the COVID19 pandemic. Otherwise, we will adopt home-based models, which are already in place and working well for Quorate*

Dr Mark Sinclair (Quorate Technology Limited) has more than 10 years of experience working in the field of speech and language technology both academically and commercially. He received his PhD from the Centre for Speech Technology Research (CSTR) at the University of Edinburgh in 2016 where he continues to contribute research, teaching and MSc project supervision. He has most recently worked on the SCRIPT project, which aimed to develop 'Speech Synthesis for Spoken Content Production' with a particular focus on low data resource languages such a Swahili, Hausa and Bengali. The majority of his time is spent working with Quorate Technology Ltd. – a spin-out from CSTR which commercialised Automatic Speech Recognition (ASR) research from the Augmented Multi-party Interaction (AMI) project – providing products and services to a broad range of clients in both the public and private sectors.

Prof Conchúr Ó Giollagáin (University of the Highlands and Islands) is the Gaelic Research Professor in the University of the Highlands and Islands, Scotland and academic director of the Soillse sociolinguistic research partnership. He recently established the UHI Language Sciences Institute based in Inverness. In 2015 he was appointed as an Adjunct Professor in the School of Political Science and Sociology, National University of Ireland, Galway. He co-authored the government-commissioned Gaeltacht survey *Comprehensive Linguistic Study of the Use of Irish in the Gaeltacht* (2007). The *Update of the Comprehensive Linguistic Study of the Use of Irish in the Gaeltacht: 2006-2011* was published in 2015. He co-authored the first major study of bilingual acquisition in Ireland, *Assessment of Bilingual Competence: Language acquisition among people in the Gaeltacht*. He co-edited the ground-breaking *Beartas Úr na nGael: Dálaí na Gaeilge san Iar-Nua-Aoiseachas* [A New Deal for the Gaels: Irish in Postmodernity] (2016). He is the lead author of the in-depth Islands Gaelic Research Project, a social survey of Gaelic vernacular communities publishing in June 2020.

Mr Gordon Wells (University of the Highlands and Islands) is Project Manager for the Soillse inter-university Gaelic research partnership, based at Gaelic College, Sabhal Mòr Ostaig on the Isle of Skye. He is also a key member of the UHI Language Sciences Institute, through which he has been closely involved in making video recordings for the "Stòras Beò nan Gàidheal" project, overseeing their transcription, and placing these online using the SMO-developed Clilstore platform. This project follows on from the preceding pilot project "Saoghal Thormoid" for which he acted as both interviewer and recording technician. He also plays a leading role in the LSI's "Mediating Multilingualism" project which explores links with similar concerns amongst the network of Centres for Endangered Languages in India, and continues to co-ordinate the Island Voices/Guthan nan Eilean online language capture and curation project based in the Western Isles.