COLING 2025

Proceedings of the 5th Celtic Language Technology Workshop (CLTW 5)

January 20, 2025
Abu Dhabi, UAE (Online)

ISBN 979-8-89176-212-1

Preface

These proceedings include the programme and papers presented at the 5th Celtic Language Technology Workshop, co-located with COLING, Abu Dhabi, January 19–24, 2025. The fifth edition has been organised as a virtual event to allow for higher attendance to the workshop.

In classical antiquity, Celtic languages were spoken across much of present-day Eurasia. In modern times, Celtic languages survive primarily in select regions of the United Kingdom, France, and Ireland, while also finding homes in diaspora communities in Argentina and Canada. The surviving Celtic languages comprise Welsh, Irish, Scottish Gaelic, Manx, Breton, and Cornish.

While these languages have relatively small speaker populations compared to major European languages, they maintain strong cultural and social significance in their traditional territories and urban areas. Among them, Irish holds a distinctive position as the only Celtic language with full European Union official status, achieved in 2007. Welsh, Gaelic, and Manx enjoy co-official recognition in their respective regions, while Breton and Cornish receive limited official acknowledgment in their historical territories.

A significant challenge facing all Celtic languages is their historical lack of resources and natural language processing (NLP) applications, which are crucial for maintaining relevance in our digital age. However, the landscape has begun to shift positively in recent years. These languages are increasingly benefiting from new academic and technological initiatives designed to support under-resourced languages. Dedicated research teams now focus on developing language and speech processing technologies for Celtic languages.

A significant milestone in this development was the establishment of CLTW. With the fifth edition, CLTW celebrates its tenth anniversary—the first workshop was held just over ten years ago, also at COLING (Dublin 2014). Over the last ten years, this forum has served as a vital platform for researchers to collaborate, share innovative work, and elevate the profile of Celtic language technology in the global linguistic community.

Despite a lower submission rate than previous years and an exclusive focus on Scottish Gaelic and Irish this time, the accepted papers in this edition represent an interesting mix of work; they cover automatic speech recognition (ASR), game-based learning, the use of LLMs for text expansion and translation, and tokenisation for Old Irish.

We thank our invited speakers for their valuable contributions: Linda Heimisdóttir (Miðeind, Reykjavík, Iceland) and Dr. Alham Fikri Aji (Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE). We also extend our gratitude to all presenters for their hard work and to the workshop attendees for their active participation. Finally, we are deeply grateful to our program committee members for their thorough reviews and invaluable feedback on the published work.

The CLTW 5 Organisers,
Brian Davis, Theodorus Fransen, Elaine Uí Dhonnchadha, and Abigail Walsh

Program Committee

Colin Batchelor, Royal Society of Chemistry, UK

Inge Birnie, University of Strathclyde, UK

Alan Cowap, Dublin City University, Ireland

Adrian Doyle, Insight SFI Centre for Data Analytics, Data Science Institute, University of Galway, Ireland

Johannes Heinecke, Orange Innovation, France

Mélanie Jouitteau, CNRS, France

John Judge, ADAPT Centre, Dublin City University, Ireland

Dawn Knight, Cardiff University, UK

William Lamb, The University of Edinburgh, UK

John P. McCrae, Insight SFI Centre for Data Analytics, Data Science Institute, University of Galway, Ireland

Simon Mille, ADAPT Centre, Dublin City University, Ireland

Caoimhín Ó Donnaíle, Sabhal Mòr Ostaig UHI, UK

Paul Rayson, Lancaster University, UK

Kevin Scannell, Cadhan Aonair, LLC, St. Louis, Missouri, USA

Monica Ward, Dublin City University, Ireland

David Willis, University of Oxford, UK

Organising Committee

Brian Davis, ADAPT Centre, Dublin City University, Ireland

Theodorus Fransen, Università Cattolica del Sacro Cuore, Milan, Italy

Elaine Uí Dhonnchadha, Trinity College Dublin, Ireland

Abigail Walsh, ADAPT Centre, Dublin City University, Ireland

Invited Speakers

Linda Heimisdóttir, Miðeind, Reykjavík, Iceland

Alham Fikri Aji, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

Sponsors

The workshop has been funded in part by ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at Dublin City University, under Grant Agreement No. 13/RC/2106_P2. It has also received funding from the eSTÓR project, which is funded by the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media, Government of Ireland. Additionally, support comes from the Common European Language Data Space, which is funded by the European Union under contract LC-01936389.

Table of Contents

Conference Program

Please note all times are UTC+0

09:00–09:10 ***Welcome***

09:10–09:50 *Keynote Speech*
Dr. Alham Fikri Aji, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

09:50–10:15 *An Assessment of Word Separation Practices in Old Irish Text Resources and a Universal Method for Tokenising Old Irish Text*
Adrian Doyle and John P. McCrae

10:15–10:35 ***Break***

10:35–11:00 *Synthesising a Corpus of Gaelic Traditional Narrative with Cross-Lingual Text Expansion*
William Lamb, Dongge Han, Ondrej Klejch, Beatrice Alex and Peter Bell

11:00–11:25 *A Pragmatic Approach to Using Artificial Intelligence and Virtual Reality in Digital Game-Based Language Learning*
Monica Ward, Liang Xu and Elaine Uí Dhonnchadha

11:25–11:45 ***Break***

11:45–12:25 *Keynote Speech*
Linda Heimisdóttir, CEO of Miðeind, a leader in language technology and artificial intelligence for Icelandic

12:25–12:50 *Fotheidil: an Automatic Transcription System for the Irish Language*
Liam Lonergan, Ibon Saratxaga, John Sloan, Oscar Maharg Bravo, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl and Ailbhe Ní Chasaide

12:50–13:15 *Gaeilge Bhriste ó Shamhlacha Cliste: How Clever Are LLMs When Translating Irish Text?*
Teresa Clifford, Abigail Walsh, Brian Davis and Mícheál J. Ó Meachair

13:15–13:25 *Concluding Remarks*
CLTW Committee