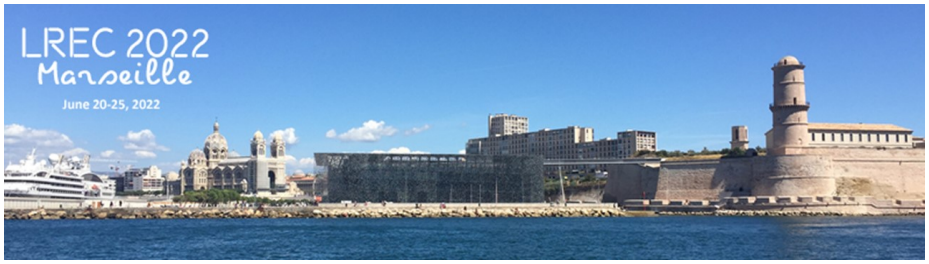

LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

4th Celtic Language Technology Workshop
(CLTW 4)

CLTW

PROCEEDINGS

Editors:
Theodorus Fransen, William Lamb and Delyth Prys

Proceedings of the 4th Celtic Language Technology Workshop at LREC 2022 (CLTW 4)

Edited by:
Theodorus Fransen, William Lamb and Delyth Prys

ISBN: 979-10-95546-73-3
EAN: 9791095546733

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

© European Language Resources Association (ELRA)

These workshop proceedings are licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License

Preface

These proceedings include the programme and papers presented at the 4th Celtic Language Technology Workshop (CLTW 4), co-located with the Language Resources and Evaluation Conference (LREC) in Marseille, June 2022.

In Classical times, Celtic languages were found across a wide swathe of modern Eurasia. Today, they are spoken in regions of the UK, France and Ireland, as well as in emigrant communities in Argentina and Canada. The modern languages are: Breton, Cornish, Irish, Manx, Scottish Gaelic and Welsh. Although the hereditary communities of these languages are small compared to those of most other European languages, they continue to have a vibrant presence in their traditional areas as well as in urban centres. While Irish is the only Celtic language that has official EU language status (since 2007), Welsh, Gaelic and Manx have co-official status. Breton and Cornish also have some limited status in their home regions. That said, all Celtic languages face the same issue in lacking natural language processing (NLP) resources to ensure continued technology support in the digital era.

Until recently, the Celtic languages lagged behind in the areas of NLP and applied language technology. Consequently, research and resource provision for this language group was poor. In recent years, however, some Celtic languages have been able to benefit from improved provisions for under-resourced languages in academia and the tech industry. Some now also have dedicated research teams working on language and speech processing technologies and related resources. The CLTW community and workshop, inaugurated at COLING (Dublin) in 2014, provides a forum to help connect these researchers and their associates to one another, to disseminate cutting-edge work and to raise the profile of Celtic language technology, more generally.

The accepted papers cover an extremely wide range of topics, including: computer-assisted language learning (CLL); automatic speech recognition (ASR), handwriting recognition; speech synthesis; syntactic parsing; part-of-speech tagging; NLP with mediaeval languages and coreference resolution.

We thank our invited speaker, Prof Kevin Scannell of Saint Louis University. We also thank our authors and presenters for their hard work, and workshop attendees for their participation. We are also very grateful to our programme committee for reviewing and providing invaluable feedback on the work published.

The CLTW 4 Workshop Organisers

Dr Theodorus Fransen, National University of Ireland, Galway
Prof William Lamb, The University of Edinburgh
Prof Delyth Prys, Bangor University

Organisers

Theodorus Fransen, National University of Ireland, Galway
William Lamb, The University of Edinburgh
Delyth Prys, Bangor University

Programme Committee

Beatrice Alex, The University of Edinburgh
Colin Batchelor, Royal Society of Chemistry
Ann Foret, Université Rennes 1
John Judge, ADAPT, Dublin City University
Teresa Lynn, Dublin City University
Mark McConville, University of Glasgow
John P. McCrae, National University of Ireland, Galway
Marieke Meelen, University of Cambridge
Ailbhe Ní Chasaide, Trinity College Dublin
Neasa Ní Chiaráin, Trinity College Dublin
Brian Ó Raghallaigh, Fiontar, Dublin City University
Thierry Poibeau, Laboratoire Lattice, CNRS, École Normale Supérieure, Sorbonne Nouvelle
Kevin Scannell, Saint Louis University
Elaine Uí Dhonnchadha, Trinity College Dublin
Monica Ward, Dublin City University
Pauline Welby, Laboratoire Parole et Langage, CNRS, Aix Marseille Université
David Willis, University of Oxford

Additional Reviewers

Adrian Doyle, National University of Ireland, Galway

Invited Speaker

Kevin Scannell, Saint Louis University

Table of Contents

Multilingual Abstract Meaning Representation for Celtic Languages
Johannes Heinecke and Anastasia Shimorina 1

Diachronic Parsing of Pre-Standard Irish
Kevin Scannell 7

Creation of an Evaluation Corpus and Baseline Evaluation Scores for Welsh Text Summarisation
Mahmoud El-Haj, Ignatius Ezeani, Jonathan Morris and Dawn Knight 14

CLILSTORE.EU - A Multilingual online CLIL platform
Caoimhín Ó Dónaill 22

Evaluation of Three Welsh Language POS Taggers
Gruffudd Prys and Gareth Watkins 30

Iterated Dependencies in a Breton treebank and implications for a Categorial Dependency Grammar
Annie Foret, Denis Béchet and Valérie Bellynck 40

Automatic Speech Recognition for Irish: the ABAIR-ÉIST System
Liam Lonergan, Mengjie Qian, Harald Berthelsen, Andy Murphy, Christoph Wendler, Neasa Ní Chiaráin, Christer Gobl and Ailbhe Ní Chasaide 47

Development and Evaluation of Speech Recognition for the Welsh Language
Dewi Jones 52

Handwriting recognition for Scottish Gaelic
William Lamb, Beatrice Alex and Mark Sinclair 60

Celtic CALL: strengthening the vital role of education for language transmission
Neasa Ní Chiaráin, Madeleine Comtois, Oisín Nolan, Neimhin Robinson-Gunning, John Sloan, Harald Berthelsen and Ailbhe Ní Chasaide 71

Cipher – Faoi Gheasa: A Game-with-a-Purpose for Irish
Elaine Uí Dhonnchadha, Monica Ward and Liang Xu 77

Towards Coreference Resolution for Early Irish
Mark Darling, Marieke Meelen and David Willis 85

Use of Transformer-Based Models for Word-Level Transliteration of the Book of the Dean of Lismore
Edward Gow-Smith, Mark McConville, William Gillies, Jade Scott and Roibeard Ó Maolalaigh 94

Introducing the National Corpus of Irish Project
Mícheál Ó Meachair, Úna Bhreathnach and Gearóid Ó Cleircín 99

BU-TTS: An Open-Source, Bilingual Welsh-English, Text-to-Speech Corpus
Stephen Russell, Dewi Jones and Delyth Prys 104

Developing Automatic Speech Recognition for Scottish Gaelic
Lucy Evans, William Lamb, Mark Sinclair and Beatrice Alex 110

Handwritten Text Recognition (HTR) for Irish-Language Folklore
Brian Ó Raghallaigh, Andrea Palandri and Críostóir Mac Cárthaigh 121

AAC don Ghaeilge: the Prototype Development of Speech-Generating Assistive Technology for Irish
Emily Barnes, Oisín Morrin, Ailbhe Ní Chasaide, Julia Cummins, Harald Berthelsen, Andy Murphy, Muireann Nic Corcráin, Claire O'Neill, Christer Gobl and Neasa Ní Chiaráin 127

Conference Programme

Monday, June 20, 2022

09:00–09:10 *Welcome*
CLTW organisers

09:10–09:50 *Keynote*
Prof Kevin Scannell, Saint Louis University

09:50–10:15 Oral Session 1

09:50–10:15 *Multilingual Abstract Meaning Representation for Celtic Languages*
Johannes Heinecke and Anastasia Shimorina

10:15–11:00 Coffee Break / Poster Session

10:15–11:00 *Diachronic Parsing of Pre-Standard Irish*
Kevin Scannell

10:15–11:00 *Creation of an Evaluation Corpus and Baseline Evaluation Scores for Welsh Text Summarisation*
Mahmoud El-Haj, Ignatius Ezeani, Jonathan Morris and Dawn Knight

10:15–11:00 *CLILSTORE.EU - A Multilingual online CLIL platform*
Caoimhín Ó Dónaill

10:15–11:00 *Evaluation of Three Welsh Language POS Taggers*
Gruffudd Prys and Gareth Watkins

10:15–11:00 *Iterated Dependencies in a Breton treebank and implications for a Categorial Dependency Grammar*
Annie Foret, Denis Béchet and Valérie Bellynck

10:15–11:00 *Automatic Speech Recognition for Irish: the ABAIR-ÉIST System*
Liam Lonergan, Mengjie Qian, Harald Berthelsen, Andy Murphy, Christoph Wendler, Neasa Ní Chiaráin, Christer Gobl and Ailbhe Ní Chasaide

10:15–11:00 *Development and Evaluation of Speech Recognition for the Welsh Language*
Dewi Jones

Monday, June 20, 2022 (continued)

10:15–11:00 *Handwriting recognition for Scottish Gaelic*
William Lamb, Beatrice Alex and Mark Sinclair

10:15–11:00 *Celtic CALL: strengthening the vital role of education for language transmission*
Neasa Ní Chiaráin, Madeleine Comtois, Oisín Nolan, Neimhin Robinson-Gunning, John Sloan, Harald Berthelsen and Ailbhe Ní Chasaide

10:15–11:00 *Cipher – Faoi Gheasa: A Game-with-a-Purpose for Irish*
Elaine Uí Dhonnchadha, Monica Ward and Liang Xu

10:15–11:00 *Towards Coreference Resolution for Early Irish*
Mark Darling, Marieke Meelen and David Willis

10:15–11:00 *Use of Transformer-Based Models for Word-Level Transliteration of the Book of the Dean of Lismore*
Edward Gow-Smith, Mark McConville, William Gillies, Jade Scott and Roibeard Ó Maolalaigh

10:15–11:00 *Introducing the National Corpus of Irish Project*
Mícheál Ó Meachair, Úna Bhreathnach and Gearóid Ó Cleircín

11:05–11:55 Oral Session 2

11:05–11:30 *BU-TTS: An Open-Source, Bilingual Welsh-English, Text-to-Speech Corpus*
Stephen Russell, Dewi Jones and Delyth Prys

11:30–11:55 *Developing Automatic Speech Recognition for Scottish Gaelic*
Lucy Evans, William Lamb, Mark Sinclair and Beatrice Alex

Monday, June 20, 2022 (continued)

11:55–13:00 Oral Session 3

11:55–12:20 *Handwritten Text Recognition (HTR) for Irish-Language Folklore*
Brian Ó Raghallaigh, Andrea Palandri and Críostóir Mac Cárthaigh

12:20–12:45 *AAC don Ghaeilge: the Prototype Development of Speech-Generating Assistive Technology for Irish*
Emily Barnes, Oisín Morrin, Ailbhe Ní Chasaide, Julia Cummins, Harald Berthelsen, Andy Murphy, Muireann Nic Corcráin, Claire O'Neill, Christer Gobl and Neasa Ní Chiaráin

12:45–13:00 *Valedictory Session*
CLTW organisers