Published in: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/stevenson, Mark/Sick, Theresa (eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. - Mannheim: Institut für Deutsche Sprache, 2017. Pp. 13-14

Creating CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes - The National Corpus of Contemporary Welsh)

Dawn Knight[1], **Tess Fitzpatrick**[2], **Steve Morris**[2], **Jeremy Evas**[1], **Paul Rayson**[3], **Irena Spasić**[1], **Mark Stonelake**[2], **Enlli Môn Thomas**[4], **Steven Neale**[1], **Jennifer Needs**[2], **Scott Piao**[3], **Mair Rees**[2], **Gareth Watkins**[1], **Laurence Anthony**[4], **Thomas Michael Cobb**[5], **Margaret Deuchar**[6], **Kevin Donnelly**[7], **Michael McCarthy**[8], **Kevin Scannell**[9]

[1]Cardiff University, [2]Swansea University, [3]Lancaster University, [4]Bangor University, [4]Waseda University, [5]University of Quebec at Montreal, [6]University of Cambridge, [7]Freelance, [8]University of Nottingham, [9]Saint Louis University

CorCenCC is an inter-disciplinary and multi-institutional project that is creating a large-scale, open-source corpus of contemporary Welsh. CorCenCC will be the first ever large-scale corpus to represent spoken, written and electronically-mediated Welsh (compiling an initial data set of 10 million Welsh words), with a functional design informed, from the outset, by representatives of all anticipated academic and community user groups.

The CorCenCC project is led by Cardiff University with academic partners at Swansea, Lancaster and Bangor Universities. It has received major funding of £1.8M from two UK research councils (ESRC and AHRC) and attracted contributions and support from stakeholders including the Welsh Government, National Assembly for Wales, BBC, S4C, WJEC, Welsh for Adults, Gwasg y Lolfa, and University of Wales Dictionary of the Welsh Language. Nia Parry (TV presenter, producer and researcher; Welsh tutor, Welsh in a week (S4C)); Nigel Owens (international rugby referee; TV presenter), Cerys Matthews (Musician author; radio and TV presenter) and Damian Walford Davies (Prof. of English Literature; poet Chair of Literature Wales) are the official ambassadors of the CorCenCC project which started in March 2016, and lasts for 3.5 years.

The corpus will enable, for example, community users to investigate dialect variation or idiosyncrasies of their own language use; professional users to profile texts for readability or develop digital language tools; to learn from real life models of Welsh; and researchers to investigate patterns of language use and change. Corpus design and construction in a minority language context such as that of Welsh poses interesting challenges, but also presents opportunities perhaps not open to developers of corpora for larger languages.

In our presentation, we provide an overview of the whole project highlighting key elements such as:

- Collection, transcription and anonymisation of the data: so far, we have extended our initial plans and developed a sampling frame for the corpus

- Development of the part-of-speech tagset and tagger: including ongoing work to create a gold-standard data for training and evaluating the Welsh natural language processing tools

- Development of a semantic annotation tool: the project has adapted the UCREL Semantic Analysis System (USAS) taxonomy for Welsh and a prototype semantic tagger has been created

- Scoping and construction of an online pedagogic toolkit: to date we have undertaken surveys with stakeholders, national and international advisors in order to collect requirements for this tool

- Infrastructure to collect and host the resulting corpus: this involves designing and building a crowdsourcing app (currently available for iOS with Android under development) for the general population to donate their conversational data, alongside the design of storage and retrieval software

Four presentations at the main CL2017 conference (Rees et al., 2017; Piao et al., 2017; Needs et al., 2017; Neale et al., 2017) provide more detail on these aspects. Further details of the project are available from the website: http://www.corcencc.org/

Acknowledgments

CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes The National Corpus of Contemporary Welsh): A community driven approach to linguistic corpus construction is an interdisciplinary, collaborative project led by the School of English, Communication and Philosophy at Cardiff University. The project commenced on 1st March 2016 and is funded by the Economic and Social Research Council (ESRC) and the Arts and Humanities Research Council (AHRC) (ref. ES/M011348/1).

References

Rees, M., Watkins, G., Needs, J., Morris, S. and Knight, D. 2017. Creating a Bespoke Corpus Sampling Frame for a Minoritised Language: CorCenCC, the National Corpus of Contemporary Welsh. *Paper presented at the CL2017 conference*, University of Birmingham, Birmingham, 24-28 July 2017.

Piao, S., Rayson, P., Knight, D., Watkins, G. and Donnelly, K. 2017. Towards a Welsh Semantic Tagger: Creating Lexicons for A Resource Poor Language. *Paper presented at the CL2017 conference*, University of Birmingham, Birmingham, 24-28 July 2017.

Needs, J., Knight, D., Morris, S., Fitzpatrick, T., Thomas, E. and Neale, S. 2017. ”How will you make sure the material is suitable for children?”: User-informed design of Welsh corpus-based learning/teaching tools. *Paper presented at the CL2017 conference*, University of Birmingham, Birmingham, 24-28 July 2017.

Neale, S., Spasi, I., Needs, J., Watkins, G., Morris, S., Fitzpatrick, T., Marshall, L. and Knight, D. 2017. The CorCenCC Crowdsourcing App: A Bespoke Tool for the User-Driven Creation of the National Corpus of Contemporary Welsh. *Paper presented at the CL2017 conference*, University of Birmingham, Birmingham, 24-28 July 2017.