CorCenCC Newsletter

Issue 2: May 2016

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Greetings from the PI

The third month of the CorCenCC project is now in full swing and it is great to report that work is going well. We are beginning to make some significant progress on individual work packages (WPs) on the project, and I am proud to say that we have already achieved a lot in this relatively short space of time. In this month's newsletter we will be providing brief updates on some of the on-going developments across specific WPs of the project and will be showcasing, in particular, recent work on the development of a Welsh semantic tagger that is being led by the team at Lancaster University (WP3). In addition to this, this month sees the inclusion of a new feature which will introduce you to individual members of the CorCenCC project team.

Happy reading, Dr Dawn Knight (Cardiff University)

Updates from CorCenCC Work Packages (WPs) 1, 2, 4, 5

Work on the CorCenCC project is distributed across 6 coordinated work packages, each with specific tasks, aims and objectives. WP0 involves on-going design, scoping and training activities, and involves all members of the project team. Brief updates on WPs 1, 2, 3, 4 are provided below (see the next section for updates on WP3).

Aim: Collect, transcribe and anonymise the data (lead: Steve Morris)

The WP1 team have been developing the sampling frame for data to be collected for CorCenCC. Following extensive research, a draft version of this has been drawn up and circulated to Welsh language and corpus experts for feedback. The team have also been working on devising transcription and anonymization conventions for the corpus.

Aim: Develop the part-of-speech tag-set/tagger (lead: Dawn Knight)

Progress continues on our bespoke part-of-speech tagset and tagging tools for Welsh. Additionally, we are putting plans in place for the production in the coming months of a gold-standard dataset for training/evaluating Welsh natural language processing tools!

Aim: Scope/construct the online pedagogic toolkit (leads: Enlli Thomas/Tess Fitzpatrick)

Work has begun on exploring the kinds of online tools already available to Welsh learners to avoid duplication of existing resources, and plans are underway to conduct a survey to investigate what resources learners and teachers already use and what they would ideally like to see developed as part of CorCenCC's pedagogical toolkit.

Aim: Construct infrastructure to host CorCenCC and build the corpus (lead: Irena Spasic)

Our project server is almost up and running and work is beginning on implementing our data crowdsourcing application, the first big steps in putting the CorCenCC infrastructure in place (keeping checking www.corcencc.org for developments).

Developing Semantic Annotation Tool for Welsh Language Analysis

In the CorCenCC Project, which aims to construct a large Welsh corpus and develop Welsh language tools, we are developing a suite of Welsh language processing software to assist the analysis and search for various information stored in the corpus data.

One of the major tools under development in this project at Lancaster University is the semantic annotation software, which aims to help to carry out the analysis of Welsh language data at the semantic level in a large scale. This tool will process Welsh language samples automatically and label each word or phrase with its coarse-grained meaning. Based on and extending the Lancaster USAS system (http://ucrel.lancs.ac.uk/usas/), it will be capable of assigning semantic categories, in the form of tags such as I1.3 (Arian: Prisiau/Money: Price), K5.1 (Chwaraeon/Sports), to Welsh words, idioms and other fixed phrases (also called multi-word expressions) based on a semantic classification scheme. This scheme consists of 232 semantic categories falling under 21 major categories, as shown below:

Tag	Definition	Tag	Definition
A	*TERMAU CYFFREDINOL A HANIAETHOL* *(*GENERAL AND ABSTRACT TERMS*)*	N	RHIFAU A MESUR (NUMBERS AND MEASUREMENT)
B	*B Y CORFF A'R UNIGOLYN* *(*THE BODY & THE INDIVIDUAL*)*	O	SYLWEDDAU, DEFNYDDIAU, GWRTHRYCHAU AC OFFER (SUBSTANCES, MATERIALS, OBJECTS AND EQUIPMENT)
C	*CELF A CHREFFT* *(*ARTS AND CRAFTS*)*	P	ADDYSG (EDUCATION)
E	*GWEITHREDIADAU, CYFLYRAU A PHROSESAU EMOSIYNOL* *(*EMOTIONAL ACTIONS, STATES & PROCESSES*)*	Q	GWEITHREDIADAU, CYFLYRAU A PHROSESAU IEITHYDDOL (LINGUISTIC ACTIONS, STATES AND PROCESSES)
F	*BWYD A FFERMIO* *(*FOOD & FARMING*)*	S	GWEITHREDIADAU, CYFLYRAU A PHROSESAU CYMDEITHASOL (SOCIAL ACTIONS, STATES AND PROCESSES)
G	*LLYWODRAETH A'R CYHOEDD* *(*GOVENMENT AND THE PUBLIC DOMAI*)*	T	AMSER (TIME)
H	*PENSAERNÏAETH, ADEILADAU, TAI A'R CARTREF* *(*ARCHITECTURE, BUILDINGS, HOUSES & THE HOME*)*	W	Y BYD A'N HAMGYLCHEDD (THE WORLD AND OUR ENVIRONMENT)
I	*ARIAN A MASNACH* *(*MONEY & COMMERCE*)*	X	GWEITHREDIADAU, CYFLYRAU A PHROSESAU SEICOLEGOL (PSYCHOLOGICAL ACTIONS, STATES AND PROCESSES)
K	*ADLONIANT, CHWARAEON A GEMAU* *(*ENTERTAINMENT, SPORTS AND GAMES*)*	Y	GWYDDONIAETH A THECHNOLEG (SCIENCE AND TECHNOLOGY)
L	*BYWYD A PHETHAU BYW* *(*LIFE AND LIVING THINGS *)*	Z	ENWAU A GEIRIAU GRAMADEGOL (NAMES AND GRAMMATICAL WORDS)
M	*SYMUD, LLEOLIAD, TEITHIO A CHLUDIANT* *(*MOVEMENT, LOCATION, TRAVEL AND TRANSPORT*)*		

It is a highly challenging task to develop an accurate semantic annotation tool. Our tool will combine a Welsh semantic lexical knowledge base (in essence a large machine readable dictionary) and a range of **word sense disambiguation** methods to achieve a high accuracy of the semantic annotation. Furthermore, we will explore crowdsourcing approaches, in which Welsh speaking communities will be invited to participate, to help to construct the semantic lexical knowledge base on a large scale and to improve the performance of the tool.

The Welsh semantic annotation tool can be useful for a variety of academic and practical purposes. Firstly, the entire Welsh corpus to be constructed in the CorCenCC Project will be tagged using this tool, which will facilitate various semantic analysis of the corpus data and a reliable and fast search of the corpus data by various semantic categories, as demonstrated by the Lancaster Wmatrix corpus research site (http://ucrel.lancs.ac.uk/wmatrix/). In addition, the semantic annotation tool can be used to improve Welsh language based ICT systems. For example, it can allow an ICT system to rapidly extract and analyse specific semantic information from online Welsh language sources on a large scale and help to link useful online sources for users.

Although the CorCenCC Project started only recently, the development of the semantic annotation tool is progressing well and we have released the translated set of categories as shown above.

Dr Paul Rayson (WP3 lead) and Dr Scott Piao (RA), Lancaster University

Meet the team

Every month we will be featuring a different member of the extended CorCenCC team in our newsletter. This will give everyone a chance to tell you a little about their background; what they want to see from CorCenCC and how they think it might contribute to their own work, or more broadly, to the work of others in Wales. This month we spotlight Dr Emyr Davies from CBAC-WJEC, who is part of the CorCenCC Project Advisory Group.

Profile: Dr Emyr Davies

At last! That was my reaction to the news that there would be a comprehensive corpus of the Welsh language. There have been small corpora developed in the past, but nothing on this scale with so many potential uses. I've been interested in language and linguistics for a long time. After graduating in Welsh from the University of Wales Aberystwyth and gaining my PhD, I spent 11 years teaching Welsh to adults in Trinity College, Carmarthen. I spent a lot of time developing new resources, but this was before the digital revolution; ideas such as electronic corpora simply didn't exist. I took a year off in 1996 to study for a Higher Diploma in General Linguistics at UCD Dublin, which I enjoyed very much. It's a useful grounding for any one working in languages, even if they then go on to work in some aspect of applied linguistics, which is exactly what I did.

I left Carmarthen to join CBAC-WJEC, the exam board based in Cardiff, in 2001 and have been working primarily in assessment and developing resources for adult learners of Welsh since then. The main part of my job has been to establish a suite of language exams for the sector, which are now in place. Another part of my job has been to represent the Welsh language exams in ALTE, the Association of Language Testers in Europe. This has been a great experience for me and given me opportunities to attend workshops and learn from some of the leading experts in the field. Not many people work purely on assessment, certainly in Welsh, and ALTE (for one thing) provides a professional community of organisations focused on language testing. ALTE also requires its members to be audited, or to provide evidence that we reach a high quality profile. The same robust standards apply to all languages and organisations who wish to be full members, including all major European languages and many less widely used languages, including Basque, Irish and Welsh.

ALTE's approach is based on an evidence-based approach. In other words, can we provide evidence to the support the decisions we make? A comprehensive corpus could be useful in developing tests, as well as resources for learning, and would provide us with evidence to make a thousand and more decisions. Many questions spring to mind: *How do we choose aspects of language to teach (and test) first? Do frequency tables for Welsh vocabulary correspond to English? Are there collocations which we should include in our teaching materials? How can we include 'redundant' language, e.g. Well... uhm... in an authentic way? Which aspects of dialect are most useful for learners? Is this or that phrase actually useful, or is it simply a quaint expression that nobody uses any more except learners?* A corpus could guide these and other decisions, rather than our intuitions.

I consider myself to be a first language Welsh speaker, having been raised in Welsh in a Welsh speaking area of south west Wales. However, our notions of who or what is a 'first' language and 'second' language user are changing fast. Having a comprehensive picture of how the language is used, especially spoken language, will be useful in redefining these notions, and how we approach learning, teaching and assessment in Welsh. We need all the resources and researchers we can find in order to explore the possibilities.

Dr Emyr Davies, emyr.davies@cbac.co.uk

CorCenCC online

You can keep up to date with developments on the project via Facebook www.facebook.com/CorCenCC/; Twitter https://twitter.com/corcencc (Tweet us @CorCenCC). You can also contact us on the project email address: corcencc@cardifff.ac.uk or visit our holding website at: http://sites.cardiff.ac.uk/corcencc/

CorCenCC @ conferences and events (May)

- **Piao, S., Rayson, P.,** Archer, D., Bianchi, F., Dayrell, C. El-Haj, M., Jiménez, R-M., **Knight, D.,** Michal Křen, M., Löfberg, L., Nawab, R., Shafi, J., The, P-L. and Mudraya, O. (2016). Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. Poster presentation to be delivered at the *LREC (Language Resources Evaluation) 2016 Conference,* May 2016, Slovenia.

CorCenCC is an ESRC/AHRC funded research project (Grant Number ES/M011348/1). The CorCenCC team includes **PI** - Dawn Knight; **CIs -** Tess Fitzpatrick, Steve Morris, Irena Spasic, Paul Rayson, Enlli Thomas, Mark Stonelake and Jeremy Evas; **RAs** - Steven Neale, Jennifer Needs, Mair Rees, Scott Piao and Gareth Watkins; **Consultants** - Kevin Donnelly, Kevin Scannell, Laurence Anthony, Tom Cobb, Michael McCarthy and Margaret Deuchar; **Project Advisory Group** - Karen Corrigan, Llion Jones, Maggie Tallerman, Mair Parry-Jones and Gwen Awbery, Emyr Davies (CBAC-WJEC), Gareth Morlais (Welsh Government), Owain Roberts (National Library of Wales), Aran Jones (Saysomethingin.com) and Andrew Hawke (University of Wales Dictionary of the Welsh Language).

If you have any comments or questions about the content of this newsletter please contact Dr Dawn Knight: KnightD5@cardiff.ac.uk