Open Welsh Language Resources for a Corpus Annotation Framework

Scott Piao[1], Steve Neale[2], Ignatius Ezeani[1], Paul Rayson[1], Dawn Knight[2], Kevin Donnelly[3]

[1]Lancaster University
[2]Cardiff University
[3]Independent Researcher

Given the increasing importance, wide utility and applications of corpora and corpus-based methods and tools, there is increasing understanding of the importance of developing corpora and corpus tools for under-resourced and minority languages. This paper focuses on the specific context of the Welsh language. It reports on the developments in the CorCenCC Project[1] (Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh), which is building the first large-scale community-driven corpus of Welsh representative of the language's use across communication types (circa 4 million spoken words, 4 million written, and 2 million e-language), genres, language varieties (regional and social) and contexts, with contributors representative of over half a million Welsh speakers in the UK.

As part of this project, we have been developing a series of corpus tools for the Welsh language, which enable users to automatically annotate a range of linguistic information covering morphological units, tokens, part-of-speech (POS) of words, multiword expressions (MWE), and semantic categories. Work on the CorCenCC project began in 2015, and over the past three years we have developed a software framework which enables users to mark up the aforementioned information in Welsh language data– the language resources included in this free/open-source framework will be the focus of the current presentation.

Our corpus tools are modelled on existing frameworks and techniques that have proven effective for carrying out automatic corpus annotation tasks, while incorporating new methods for dealing with specific features of the Welsh language. For example, the Welsh POS tagger employs a rule-based approach using constraint grammar (Karlsson et al., 1995) while the semantic tagger is based on the framework of the USAS corpus annotation system (Rayson et al., 2004; Piao et al., 2015). These tools employ a set of Welsh language resources compiled semi-automatically, which provide knowledge bases for the tools, such as POS classification rulesets, a gazetteer, semantic classification of lexical items etc. The construction of these Welsh language resources entailed significant effort, such as analysing large corpus resources and complex processing of the data. Indeed, much of our efforts have been dedicated to leveraging existing Welsh corpus resources including Eurfa (Donnelly, 2016), CEG (Cronfa Electroneg o Gymraeg) (Ellis et al., 2001), Kwici (Donnelly, 2014), the Corpus of Children's Literature in Welsh,[2] etc.

---

[1] For details of the CorCenCC Project, see website: http://www.corcencc.org/

[2] URL: http://www.egni.org

In fact, in addition to the software system, such language resources themselves provide valuable materials for the language research community. Currently this is the largest Welsh lexical resource of this type in existence. Table 1 lists the main Welsh languages resources that have been built and incorporated in the framework.

Language resource type	Information encoded	Supporting tool functionality	Size of resources
Welsh POS lexicon	Possible POS categories of words	Tokenisation, POS tagging	212,452 entries
Welsh grammar rules	Constraint grammar	POS tagging	243 rules
Gazetteer	Place names	POS tagging & Semantic tagging	1,760
Welsh Names	Welsh personal names, including surnames, male and female names	POS tagging & Semantic tagging	20,141
Welsh semantic Lexicon	Semantic categories of Welsh words	Semantic tagging	143,287 entries
Semantic Multiword Expression (MWE) templates	Possible semantic categories of MWE patterns	Semantic tagging of MWEs	Small sample entries

Table 1: Main Welsh languages resources included in the CorCenCC Welsh corpus annotation framework.

As shown in Table 1, there are two main groups of Welsh language resources used in our annotation framework: a) for part-of-speech tagging, and b) for semantic tagging.

With regards to resources for part-of-speech tagging, the Welsh POS lexicon provides a major knowledge base for identifying possible POS tags for each Welsh word. Next, the Welsh grammar rules allow the selection and disambiguation of POS tags in given contexts. In addition, the gazetteer and Welsh name collection, which contains 1,090 female names, 2,183 male names and 16,868 surnames, helps to identify named entities (this resource is also used for the semantic tagger).

The other group of resources are for tagging semantic categories of words and multiword expressions. In terms of the semantic classification of the lexical units, we apply a semantic annotation scheme developed in Lancaster University for the USAS (UCREL Semantic Analysis System)[3]. In addition to the gazetteer and person name collection, this group of resources contains two semantic lexicons, which contain information about semantic categories for each word or MWE in their usage contexts. For example, the word *ysgol* can denote a 'school' or 'ladder' in different contexts, and the lexicon encodes this word with these two candidate semantic categories, from which the tagging software selects the correct one according to the actual context. Below are two sample lexicon entries for a single word and an MWE respectively.

[3] For further details of the semantic annotation scheme, see website http://ucrel.lancs.ac.uk/usas.

Sample 1: amddifadedd noun A9-

Sample 2: cynllun*_NOUN {NOUN} bws_NOUN {ADJ/CONJ} am_PREP dim_NOUN M3/Q1.2

Sample 1 shows the structure of the Welsh semantic lexicon entry, in which the word *amddifadedd* (meaning 'deprivation') is mapped to USAS tag *A9-* (Getting and giving; possession). On the other hand, Sample 2 shows the template structure of the MWE semantic lexicon, which states that the word sequence *cynllun bws am ddim* is a stable phrase that has both senses of M3 (Vehicles and transport on land) and Q1.2 (Paper documents), because the phrase means 'free bus pass scheme.' The semantic lexicons provide semantic information for the software system, which will disambiguate the word and MWE senses based on their local context.

Although the main purpose of constructing these specialised Welsh language resources is to develop corpus annotation software tools, they themselves provide valuable materials for language studies and computational linguistics as well as a springboard towards other language tools. They will be made available along with the software towards the end of the CorCenCC project, under a free software (GPL version 3) licence. Please check our websites for updates on the launch of the corpus and the software[4].

The development of the Welsh language resources and corpus tools will continue, within and beyond the CorCenCC Project, with the aim, among other things, of introducing word vector based Welsh resources and machine learning techniques. Unsupervised learning of word embedding, popularised by the works of Mikolov et al (2013) and Pennington et al (2014), has proven to be efficient and effective for abstracting high-quality meaning representations from raw text, and has been successfully applied to many NLP tasks. Future research efforts will be geared towards training and applying Welsh embedding models to improve the performances of the existing Welsh POS and semantic tagging tools.

Acknowledgement

The research on which this presentation is based is funded by the UK Economic and Social Research Council (ESRC) and Arts and Humanities Research Council (AHRC) as part of the *Corpws Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh): A community driven approach to linguistic corpus construction* project (Grant Number ES/M011348/1).

References:

Donnelly, K. (2014). Kwici: a 4m-word corpus drawn from the Welsh Wikipedia. URL: http://cymraeg.org.uk/kwici.

Donnelly, K (2016). Eurfa, a GPLed dictionary of Welsh. Available at url: http://eurfa.org.uk.

Ellis, N. C., O'Dochartaigh, C., Hicks, W., Morgan, M., & Laporte, N. (2001). Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh. URL: http://corpws.cymru/ceg.

[4] URL: https://github.com/UCREL/Multilingual-USAS and https://github.com/CorCenCC

Karlsson, F., A. Voutilainen, J. Heikkilä, & A. Anttila (1995). Constraint Grammar: a Language-independent system for parsing unrestricted text. Berlin: Mouton de Gruyter.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *NIPS*, 1–9.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543. http://doi.org/10.3115/v1/D14-1162.

Piao, S., Bianchi, F., Dayrell, C., D'Egidio, A., & Rayson, P. (2015). Development of the multilingual semantic annotation system. *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015)*, Denver, Colorado, USA.

Rayson, Paul, Archer, D., Piao, S., & McEnery, T. (2004). The UCREL semantic analysis system. In *Proceedings of LREC-04 Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP Tasks*, pp. 7-12, Lisbon, Portugal.