Towards a Welsh Semantic Tagger: Creating Lexicons for A Resource Poor Language

Scott Piao (Lancaster University, UK), Paul Rayson (Lancaster University, UK), Gareth Watkins (Cardiff University, UK), Dawn Knight (Cardiff University, UK) and Kevin Donnelly (Independent Researcher, UK)

Semantic annotation and analysis is an important part of corpus linguistics and other research areas (Gacitua et al., 2008; Potts and Baker, 2013; Demmen et al., 2015), and semantic tagging tools have facilitated this type of research to be conducted on a large scale. A major tool is the USAS semantic tagger developed at Lancaster University (Rayson et al., 2004), originally designed for English but has been extended to cover more languages, including Italian, Chinese, Spanish, Portuguese etc. (Löfberg et al., 2005; Mudraya et al., 2006; Piao et al., 2015; Piao et al., 2016). The USAS framework employs a lexical semantic classification scheme containing 21 major semantic categories that are further sub-divided into 232 categories. The multilingual capability of the tagger enables multilingual/cross-lingual semantic analysis under this unified scheme. In the CorCenCC Project (Corpws Cenedlaethol Cymraeg Cyfoes: The National Corpus of Contemporary Welsh)[1], we are extending the USAS to automatically annotate Welsh language data, particularly the CorCenCC corpus, with the semantic information.

A critical part of the USAS extension is the development of a Welsh semantic lexicon which provides a wide coverage of Welsh vocabulary and a high quality of semantic classification. Various Welsh lexical resources have been drawn on in building the Welsh semantic lexicon. A major such resource is the Eurfa Welsh/English bilingual lexicon developed by Donnelly (2016), and made available under an open license (GPL). This resource contains valuable lexical information about a large number of Welsh words, including lemma forms, part-of-speech (POS) categories, many multi-word expressions (MWEs), and English translations. From this resource, we extracted 136,468 single Welsh words (inflected forms) and classified them into USAS semantic categories via their English translation equivalents and English semantic lexicons to form the basis of the new Welsh semantic lexicon. In addition, the words are mapped to their lemma forms in order to improve text coverage of the lexicon by allowing each lexicon entry to match with all inflectional variants of the same lemma in the text.

In addition, we expanded the Welsh semantic lexicon by manually compiling closed-class word lists, such as prepositions, conjunctions etc., in order to cover the highly frequent closed-class words in the running text. Another important source for the lexicon are Welsh names, including person names and place names, which we have collected from a number of resources, including the Language Technologies Unit of Bangor University, UK and websites including "Behind The Name", "Think Baby Names", and "Wales"[2]. Through these approaches, the Welsh semantic lexicon was expanded to contain 143,290 Welsh words.

---

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

[2] Permissions to use their name resources were obtained from the following organisations:

a) Language Technologies Unit of Bangor University (https://www.bangor.ac.uk/canolfanbedwyr/technolegau_iaith.php.en),
b) Behind The Name (http://www.behindthename.com/names/usage/welsh),
c) Think Baby Names (http://www.thinkbabynames.com/names/1/welsh), and
d) Wales UK (http://www.walesuk.info/wales.html).

We carried out an initial lexical coverage evaluation of the Welsh semantic lexicon using a gold standard corpus, which was constructed for evaluating corpus tools in the CorCenCC project. The gold corpus consists of around 15,000 words and contains selected materials from four existing corpora: *Kwici* (Welsh Wikipedia)[3], *Kynulliad3* (Welsh Assembly Proceedings)[4], *Meddalwedd* (software translations)[5], and *LER-BIML* (a small corpus of 10 multi-domain texts)[6]. The first three corpora were stored in databases, so the selection was made for the first two by selecting the first 100 items where the length was between 20 and 40 words, and for the third by selecting the first 100 longest items. For the fourth, two of the texts were chosen. The aim was to give a reasonable selection (between 2,000 and 4,000 words) of text from different domains, different sources, and differing lengths in order to create a balanced and representative test corpus. Once the text was gathered, it was cleaned (for example, HTML tags were removed, as were items that contained little Welsh), and then typos and punctuation errors were corrected. Items which contained English words were retained: since modern, less formal Welsh usually contains some English code switches, it is considered desirable that our part-of-speech (POS) tagging and semantic tagging systems have the capability to handle such noise.

	Content words	**Function words**	**Person/Place names**
Number of entries	136,468	264	6,558
Sample lexicon entries	abacws Eg N3.1 bri Eg X9.2+ chwerthin B E4.1+/X3.2 defnyddio B A1.5.1 S7.1+ llwybro B M1 L2 X9.2+/A12+ llwydda B X9.2+ N4 plesiwn B E4.2+ E2+ X7+ plicio B A9+ A1.1.1 F1 tripio B M1 M2 S8- walio B H2 warws Eg A9+/H1	â Cy Z5 ag Cy Z5 amdanat A Z5 atoch A Z5 chithau Cy Z5 cyn Cy Z5 erbyn Cy Z5 fel Cy Z5 gyda Cy Z5 hebddi A Z5 imi A Z5 lle Cy Z5 mai Cy Z5 na Cy Z5 oddiar A Z5 wrthych A Z5	#Person names Anwen Ep Z1 Arwel Ep Z1 Bedwyr Ep Z1 Bethan Ep Z1 Bleddyn Ep Z1 Blodeuyn Ep Z1 #Place names Aberhafesb Ep Z2 Barry Ep Z2 Carreghwfa Ep Z2 Geufron Ep Z2 Gifftan Ep Z2 Penygroes Ep Z2 Wigau Ep Z2

Table 1: Statistics of Welsh semantic lexicon and sample entries.

In the evaluation, our prototypical Welsh semantic tagger based on the current version of the semantic lexicon covered 72.42% of the words in the gold corpus. If the noise in the text mentioned above is excluded, a higher lexical coverage can be expected. Table 1 shows the sizes of three main types of the Welsh semantic lexicon entries and some sample entries from each type. As shown in Table 1, the first column in each lexicon entry contains a word/lemma, the second column contains a part-of-speech tag, and the last column contains possible USAS semantic tag/s (for definitions of the USAS tags, see website: *http://ucrel.lancs.ac.uk/usas*). The tags contained in the sample entries are from a new Welsh POS tagset developed in the CorCenCC Project, which are defined as follows:

[3] See website http://cymraeg.org.uk/kwici.
[4] See website: http://cymraeg.org.uk/kynulliad3
[5] See website http://techiaith.cymru/corpws/Moses/Meddalwedd
[6] See website http://www.lancaster.ac.uk/fass/projects/biml

Eg: Noun
B: Verb
Cy: Conjunction
A: Preposition
Ep: Proper noun

We are further expanding the Welsh semantic lexicon in order to achieve a higher lexical coverage and better quality in the semantic classification of the words. For example, we have extracted additional large Welsh word lists from a number of existing Welsh corpora, including *CEG Cronfa Electroneg o Gymraeg* (Ellis et al., 2001), *Kwici* (Corpus of Welsh Wikipedia *http://cy.wikipedia.org*) and C*orpus of Children's Literature in Welsh* (*http://www.egni.org*). When we repeated the lexical coverage evaluation including the additional word list, our Welsh lexicons covered over 97% of the text of the gold test corpus. Note that the additional raw Welsh word collection has not yet been built into the semantic lexicon, but such a high lexical coverage figure shows the potential of our semantic tagger that can be achieved when a major part, if not all, of the words already extracted are integrated into the formal semantic lexicon. A prototype of the semantic tagger tool has been built based on the existing Welsh semantic lexicon for testing, which is available at website *http://phlox.lancs.ac.uk/ucrel/semtagger/welsh* and the Welsh lexicon is available to download under a Creative Commons licence at *https://github.com/UCREL/Multilingual-USAS*. The semantic tagger will be continuously improved during the project, and we will provide a demonstration of the current version tool.

The current Welsh semantic tagger is at an early stage of development. As the CorCenCC project progresses, we will continue to expand and refine the Welsh semantic lexicons and improve the semantic tagger for annotating Welsh corpus with a high accuracy. In future work, we will: 1) further expand the size and improve quality of the semantic classification of the Welsh lexicon entries, 2) build Welsh multiword expression semantic lexicon, and 3) develop an efficient Welsh semantic annotation tool by combining these semantic lexical resources and word sense disambiguation methods with the Welsh part-of-speech tagger and lemmatiser being created in the CorCenCC project.

Acknowledgement

The research on which this article 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)* (Grant Number ES/M011348/1).

References

Demmen, J.E., E. Semino, Z. Demjen, V. Koller, A. Hardie, P. Rayson and S. Payne (2015). A computer-assisted study of the use of violence metaphors for cancer and end of life by patients, family carers and health professionals. *International Journal of Corpus Linguistics*. 20 (2), 205-231.

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

Ellis, N. C., C. O'Dochartaigh, W. Hicks, M. Morgan and N. Laporte (2001). Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh. URL: www.bangor.ac.uk/canolfanbedwyr/ceg.php.en

Gacitua, R., P. Sawyer and P. Rayson (2008). A flexible framework to experiment with ontology learning techniques. *Knowledge-Based Systems*, 21(3), 192-199.

Mudraya, O.V., B.V. Babych, S. Piao, P. Rayson and A. Wilson (2006). Developing a Russian semantic tagger for automatic semantic annotation. In *Proceedings of the International Conference "Corpus Linguistics - 2006"*, St.-Petersburg, Russia, 290-297.

Löfberg, L., S. Piao, A. Nykanen, K. Varantola, P. Rayson and J. P. Juntunen (2005). A semantic tagger for the Finnish language. In *Proceedings of the Corpus Linguistics Conference 2005*, Birmingham, UK.

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

Piao, S., P. Rayson, D. Archer, F. Bianchi, C. Dayrell, M. El-Haj, R. Jiménez, D. Knight, M. Křen, L. Löfberg, R. M. Adeel Nawab, J. Shafi, P. L. The and O. Mudraya (2016). Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference (LREC2016)*, Portorož, Slovenia.

Potts, A. and P. Baker (2013). Does semantic tagging identify cultural change in British and American English?. *International Journal of Corpus Linguistics*, 17(3), 295-324.

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