# Towards A Welsh Semantic Annotation System

**Scott Piao[1], Paul Rayson[1], Dawn Knight[2] and Gareth Watkins[2]**
[1]School of Computing and Communications, Lancaster University, UK
[2]School of English, Communication and Philosophy, Cardiff University, UK
{s.piao;p.rayson}@lancaster.ac.uk, knightd5@cardiff.ac.uk, gllwatkins@gmail.com

**Abstract**

Automatic semantic annotation of natural language data is an important task in Natural Language Processing, and a variety of semantic taggers have been developed for this task, particularly for English. However, for many languages, particularly for low-resource languages, such tools are yet to be developed. In this paper, we report on the development of an automatic Welsh semantic annotation tool (named CySemTagger) in the CorCenCC Project, which will facilitate semantic-level analysis of Welsh language data on a large scale. Based on Lancaster's USAS semantic tagger framework, this tool tags words in Welsh texts with semantic tags from a semantic classification scheme, and is designed to be compatible with multiple Welsh POS taggers and POS tagsets by mapping different tagsets into a core shared POS tagset that is used internally by CySemTagger. Our initial evaluation shows that the tagger can cover up to 91.78% of words in Welsh text. This tagger is under continuous development, and will provide a critical tool for Welsh language corpus and information processing at semantic level.

**Keywords:** Welsh semantic tagger, corpus annotation, Welsh language, Welsh corpus, CorCenCC

## 1. Introduction

Automatic semantic annotation and analysis is an important task for Natural Language Processing (NLP), and semantic taggers have been developed and used for carrying out semantic analysis of language data on a large scale. A major tool built for such a purpose is USAS (UCREL Semantic Analysis System)[1] (Rayson et al., 2004; Piao et al., 2017), which is designed to annotate each word or phrase in text with lexical semantic categories derived from Tom McArthur's Longman Lexicon of Contemporary English (McArthur, 1981), such as "Food" or "Time". The USAS semantic scheme contains 21 major semantic fields that are further divided into 232 sub-categories. Initially developed for processing English text, it has been extended to tag texts of a number of languages, including Italian, Chinese, Spanish etc. (Piao et al., 2015).

In the CorCenCC Project[2], we have been developing a Welsh semantic tagger, named CySemTagger, modelled on the USAS framework, which employs a translated USAS semantic classification scheme and tagset for Welsh language. During the course of development, we have first constructed large Welsh semantic lexicons containing approximately 136,468 Welsh word entries, which provide a lexical knowledge base for the semantic tagging system[3]. Based on the lexicons, we have developed an initial version of the tagger software system that is designed to accommodate different Welsh POS taggers and tagsets that exist or will be developed. In the CorCenCC project, CySemTagger will be mainly based on a new Welsh POS tagger, named CyTag, which has been developed in this Project (Neale et al., 2018). In this paper, we describe the CySemTagger system and report on its initial evaluation.

The remaining part of this paper is organised as follows. Section 2 will discuss related work, Section 3 will describe the architecture of the CySemTagger, Section 4 will discuss the construction of Welsh semantic lexicon, Section 5 will explain about detailed working mechanism of CySemTagger's main components, Section 6 will discuss an evaluation of the current version of CySemTagger, and Section 7 will conclude our work and discuss future work.

## 2. Related Work

Over recent years, various semantic annotation tools have been developed in the NLP community. These tools are used to automatically recognise and annotate various semantic categories and concepts at different syntactic levels, such as word level, phrase level, sentence level etc.

Among the major existing semantic taggers developed in NLP communities is USAS (Rayson et al., 2004; Piao et al., 2017), GATE[4] (Cunningham et al., 2011), Freeling (Padro and Stanilovsky, 2012), NLTK[5] (Bird et al., 2009) etc., which provide functionalities of semantic annotation of various types, such as WordNet's Word sense IDs or Named Entity types etc. For example, GATE and KIM (Popov et al., 2003), combined together, provide multilingual semantic tagging function based on ontologies. Freeling is capable of detecting and tagging multilingual texts with named entity types and WordNet senses. Zhang and Rettinger (2014) developed a toolkit that carries out Wikipedia-based annotation. NLTK (Bird et al., 2009) provides a function for analysing the meaning of sentences.

What is directly related to our work is the past development of multilingual functionality of the USAS framework. As mentioned earlier, initially developed for English, it has been extended and modified to cover an increasing number of languages. Currently USAS is capable of carrying out semantic annotation on 12 languages, including Italian, Finish, Russian, Chinese, Spanish, Portuguese, Swedish,

[1]For further details of USAS, see website http://ucrel.lancs.ac.uk/usas/

[2]For information about this project, see website http://www.corcencc.org/

[3]These lexicons are made available at https://github.com/UCREL/Multilingual-USAS

[4]See website https://gate.ac.uk/

[5]http://www.nltk.org/

Dutch etc. (Lofberg et al., 2005; Mudraya et al., 2006; Piao et al., 2015).

Our research draws upon the experiences of the previous work, while extending the capability of automatic semantic annotation of existing tools to the Welsh language, for which no existing NLP tools can provide such a comprehensive semantic annotation as CySemTagger aims to perform. Our new tagger will facilitate new semantic based research on Welsh language data, and via the CorCenCC framework will assist with improving the understanding of real-life communication for Welsh speakers, teachers and learners.

## 3. Outline of CySemTagger's Architecture

The CySemTagger is a software system which has an architecture consisting of a set of lexical knowledge resources and disambiguation rules and models wrapped in software components that interact with each other. Figure 1 illustrates the outline of the architecture and main workflows of the system.

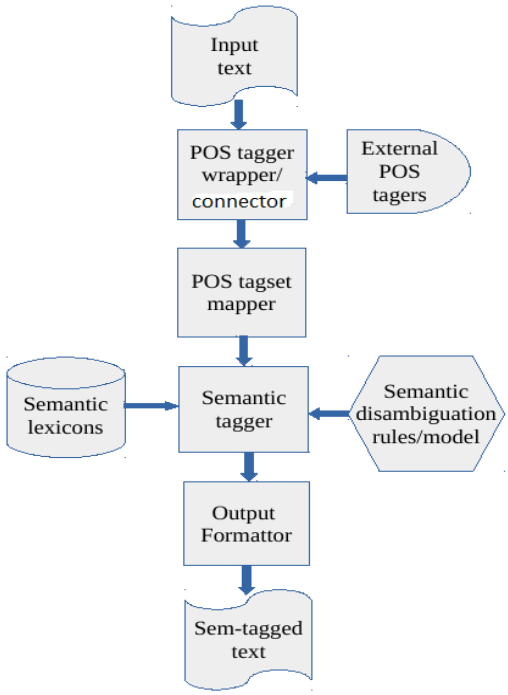


Figure 1: CySemTagger Software Architecture

As shown in the figure, a set of software modules provide main functionalities, such as receiving input data stream, calling external part-of-speech taggers to carry out morpho-syntactic analysis, produce tagged output in required formats etc. Together, they form a pipeline to interact with each other to complete the whole task of semantic tagging. The main challenge for building such a system for a new language, Welsh in our case, is the development of semantic lexicons and word sense disambiguation rules and models for the new language. Because there are very few Welsh semantic lexical resources available, we were faced with a tough challenge for building the Welsh semantic tagger.

## 4. Semantic Lexicon Construction

As a system based on linguistic knowledge, a core part of the semantic tagger system is a set of semantic lexicons which provide candidate semantic categories for each word, hence Welsh semantic lexicon construction is the first main step for the development of Welsh semantic tagger. For this purpose, we exploited various lexical and corpus resources. The main lexical resource used for this work is the Eurfa Welsh/English bilingual lexicon developed by Donnelly (2017). This bilingual lexicon contains a large number of Welsh words and their English translations along with useful information such as lemma forms and part-of-speech (POS) labels. It also contains many Welsh multi-word expressions (MWEs), which is a valuable resource for creating semantic MWE lexicons for the semantic tagger in later stages. Because it is time-consuming work to manually compile new semantic lexicons from scratch, we applied automatic methods by mapping and porting semantic categories and tags for Welsh words via their English translations through the existing English semantic lexicons. This method has been proven effective in our previous research on other languages (Piao et al., 2016). The high quality of the Eurfa bilingual lexicon helped us to achieve a good initial result for the automatically generated Welsh semantic lexicon. Obviously, the automatically generated lexicon will need be pruned manually or by other methods to guarantee the accuracy of the semantic annotation. Through the automatic process, we extracted a lexicon containing 136,468 Welsh words (including many inflected forms) mapped to semantic category/ies. It provides a solid basis for developing a system of Welsh semantic tagger.

In addition to the automatic lexicon generation, we also collected 264 Welsh closed class words and integrated them into the lexicon, mainly including function words such as prepositions, conjunctions etc. Although there are a limited number of such words, they are critical for correctly understanding the meaning of the text, and typically are high frequency items. Another important lexical source is the Welsh names. Our initial observation showed that a significant proportion of corpus data consists of various names. Therefore, we searched and collected Welsh names, including person names and place names, from various sources[6], including the Language Technologies Unit of Bangor University[7], UK and the websites of "Behind The Name"[8], "Think Baby Names"[9], and "Wales UK"[10]. As a result, we collected 6,553 Welsh names to expand the semantic lexicon. Combining automatic and manual processes, currently we have constructed a Welsh single word semantic lexicon of 143,287 entries. Table 1 shows a sample of the single word semantic lexicon, where the semantic tags are from the USAS semantic tagset[11].

[6]The creators/owners of these name sources gave us permissions to use their Welsh name resources for the purpose of developing the Welsh semantic tagger

[7]https://www.bangor.ac.uk/canolfanbedwyr/technolegau_iaith.php.en

[8]http://www.behindthename.com/names/usage/welsh

[9]http://www.thinkbabynames.com/names/1/welsh

[10]http://www.walesuk.info/wales.html

[11]For definitions of the semantic tags, see http://ucrel.lancs.ac.uk/usas/usas_guide.pdf

Table 1: Sample of Single Word Semantic Lexicon

| Word | POS-tag | Sem-tag |
|---|---|---|
| abacws | noun | N3.1 |
| corfforaethol | adj | I2.1/S5 |
| galluogai | verb | S8+ |
| gambled | verb | A15- I1/A1.4 K5.2/I1 |
| sobr | adv | A13.3 |

Our semantic tagger also needs a Welsh MWE semantic lexicon, which provides semantic categories for MWE terms and non-compositional idiomatic expressions, e.g. "take care of" and "kick the bucket" (for English) etc. Currently we use a sample MWE semantic lexicon for testing software, but work is under way to construct a large MWE semantic lexicon as the project progresses. Table 2 shows samples of MWE semantic lexicon entries, including template codes.

Table 2: Sample of MWE Semantic Lexicon

| MWE/Templates | Sem-tag |
|---|---|
| Adran_NOUN Iechyd_NOUN | G1.1 |
| gofyn_VERB {NOUN/VERB} i_PREP | Q2.1 |
| parhad_NOUN busnes_NOUN | I2.1/T2+ |
| pwyllgor*_NOUN addysg_NOUN | P1/G1.1 G1.2 |

Besides the existing English/Welsh bilingual lexical resources, we are also considering Welsh corpus resources, particularly Welsh/English parallel corpora, for further expanding and improving the Welsh semantic lexicons. We carried out an initial experiment by collecting words from the existing Welsh corpora, including CEG Cronfa Electroneg o Gymraeg (Ellis et al., 2001), Kwici (Corpus of Welsh Wikipedia)[12] and Corpus of Children's Literature in Welsh[13], and estimated the proportion of text that can be covered by the word list, together with the formal semantic lexicon. Our experiment shows that, if all the newly collected words can be integrated into the Welsh semantic lexicons, our semantic tagger could achieve over 97% of text coverage. Therefore we aim to semantically classify as many words in the word collection as possible and integrate them into the semantic lexicon in order to achieve a high lexical/text coverage.

Some unique features of Welsh language present a tough challenge for the Welsh semantic lexicon building. As a Celtic language, the Welsh language's linguistic features are widely different to those of the English language (a Germanic language). For instance, while the English alphabet consists of 26 letters, the Welsh alphabet consists of 29 letters, including some letters which are made up of two characters (namely "Ch", "Dd", "Ff" "Ng", "Ll", "Ph", "Rh" and "Th"). The Welsh language is different from English in terms of grammar, there is no indefinite article in the Welsh language, for instance. Furthermore, the Welsh language is unique in that it employs a system of mutation, that is, under certain circumstances the first letter of a word is substituted for another. For example, Welsh feminine nouns which follow the Welsh definite article "y" mutate as shown in 3 (see the highlighted letters).

Table 3: Example of mutation following the definite article in Welsh

| Welsh | English | Mutation |
|---|---|---|
| **P**ioden | Magpie | Y **B**ioden |
| **C**raith | Scar | Y **G**raith |
| **T**eml | Temple | Y **D**eml |

In respect of English equivalents to Welsh words, a one-to-one relationship does not always exist, as shown in 4. For instance, the English word "together" would normally be translated as the MWE "gyda'i gilydd" in Welsh. Conversely, the Welsh word "haprif" would be equivalent to "random number" in English. These factors present a tough challenge for the automatic or semi-automatic creation of Welsh semantic lexicons based on existing English ones, and requires more manual efforts in this process, particularly for MWE lexicon construction.

Table 4: Examples of Equivalence Between English and Welsh Single Words and MWEs

| English | Welsh |
|---|---|
| beyond repair | anadferadwy |
| random number | haprif |
| take pride in | ymfalchïo |
| lifeboat | bad achub |
| yorker (a cricket term) | pelen lawn |
| Iceland | Gwlad yr Iâ |
| reserve (in sport) | chwaraewr wrth gefn |
| desktop | bwrdd gwaith |
| toadstool | caws llyffant |

## 5. Main Components of CySemTagger

The current version of the CySemTagger is based on the Welsh semantic lexicons constructed so far. With regards to major functionality, the system mainly consists of four modules:

1) Lexicon look-up (both for single words and MWEs),
2) Part-of-speech tagging,
3) Semantic category disambiguation,
4) Output formatting.

The first main module is for loading the lexicons into the system and looking up candidate categories for each word. For single words this is a straightforward process, whereas a complex algorithm is needed for the MWE lookup. In the USAS framework, the MWE entries can contain specified template codes and format which are used to represent similar MWEs with a single lexicon entry. For example, the entry below:

```
spe*d_* {R*} off_RP
```

represents MWEs "sped off", "speed off", and "sped quickly off" etc. Internally in the software, the MWE entries are transformed into regular-expression based matchers when loaded into the system. This technique allows the

[12] http://cy.wikipedia.org
[13] http://www.egni.org

semantic tagger to identify and annotate a large amount of MWE variants using a moderate-sized MWE lexicon.

Next, the part-of-speech module is a software wrapper for external POS taggers. It is a program component that links and wraps external Welsh POS taggers built independently into the CySemTagger system. We started with an existing Welsh POS tagger, WNLT (Welsh Natural Language Toolkit)[14], which was publicly available when our project began. With the development of the new Welsh POS tagger, CyTag, in the CorCenCC Project (Neale et al., 2018), it has also been integrated into the semantic tagger. Considering the existence of multiple POS taggers, and in order to make CySemTagger compatible with different POS tagsets, we designed a core POS tagset that provides sufficient information for the semantic annotation purpose and to which other Welsh POS tagsets can easily be mapped. Such a design makes CySemTagger flexible to potentially accommodate existing and future POS taggers and provides a wider choice for users. Table 5 lists the core POS tagset.

Table 5: Core POS Tags

| POS Tag | Definition |
|---|---|
| noun | Noun |
| verb | Verb |
| adj | Adjective |
| adv | Adverb |
| num | Numerals |
| pnoun | Proper noun |
| intj | Interjection |
| art | Article |
| part | Particle |
| prep | Preposition |
| conj | Conjunction |
| pron | Pronoun |
| code | Special code, e.g. Maths symbol |
| punc | Punctuation |
| fw | Foreign word |
| abbrev | Abbreviation |
| lett | Letter |
| xx | Unregnized token |

Table 6 lists the detailed mapping from the CyTag POS tagset[15] and WNLT POS tagset[16] to the core tagset. As shown in the table, different POS tagsets can have widely different levels of granularity. For instance, the CyTag POS tagset has 59 fine-grained sub-categories for Welsh verbs. As a result, the mapping of the tagsets is not straightforward for some POS categories such as *abbreviation* and *foreign words* etc. The core POS tagset is designed to accommodate all of the POS sub-categories included in the Cytag and WNLT tagsets, with four categories (*fw*, *abbrev*, *lett* and *XX*) only mapping with CyTag tags without corresponding WNLT tags. This may cause slight loss of POS information if the WNLT is used, but we expect that such a design provides the optimal practical solution for our system.

The disambiguation module for CySemTagger is in early stages of development, and will be reported in future papers. Various context aware algorithms will be tested and integrated to improve the accuracy of the tagger. The tagged output can be presented to users in various formats as necessary, such as CTV and XML.

The current version of CySemTagger has been built as a Web service for the convenience of integrating software written in different programming languages, including Java and python programs and the VISL's Constraint Grammar v3 (CG-3) package, and it can be accessed via a demo web site[17] and a desktop client GUI application[18].

## 6. Evaluation

In order to assess the performance of the current version of CySemTagger, we carried out a test based on a Welsh test corpus, a gold corpus, which is specifically compiled for the tool evaluation task in the CorCoeCC Project. The test corpus consists of text segments selected from four existing corpora, Kwici (Welsh Wikipedia)[19], Kynulliad[20] (Welsh Assembly Proceedings), Meddalwedd (software translations)[21], and LER-BIML (a small corpus of 10 multi-domain texts)[22], and contains around 15,000 words. In this experiment, we focused on examining the average text coverage of the tool, i.e. what percentage of the words in the test corpus can be identified by CySemTagger.

In detail, we examined the text coverage of the CySemTagger when it is linked to the two POS taggers CyTag (internal prototype version) and WNLT respectively. Because CyTag and WNLT use different tokenisation rules and algorithms, they produced different number of tokens. CyTag produced 13,220 words (excluding punctuations) of which 3,716 (28.11%) are function words; WNLT produced 14,435 words (excluding punctuations) of which 5,314 words (36.81%) are function words. Therefore, the text coverages are based on different word numbers, but they are comparable in terms of tool performance. Table 7 shows the text coverage statistics in terms of content words, function words and whole text respectively.

As shown by our experiment, CySemTagger is capable of covering about 91.78% of Welsh running text when it uses CyTag POS tagger, which is under continuous development in the CorCenCC Project. It still covers 72.92% of text when using the WNLT. Our initial analysis reveals that an important factor for the difference of the text coverages is the performance of lemmatisation of Welsh words, for which CyTag has a superior accuracy compared to WNLT.

Due to the lack of manually annotated Welsh test data, it was not possible to carry out an evaluation on the quality of the semantic annotation. Currently a test corpus for

[14] See https://sourceforge.net/projects/wnlt/

[15] For details of the CyTag POS tagset, see website: http://cytag.corcencc.org/tagset

[16] For details of the WNLT POS tagset, see website: https://sourceforge.net/projects/wnlt/files/user-guide.pdf

[17] See demo web site http://phlox.lancs.ac.uk/ucrel/semtagger/welsh

[18] A Java desktop application downloadable at http://ucrel.lancs.ac.uk/usas/gui/

[19] http://cymraeg.org.uk/kwici

[20] http://cymraeg.org.uk/kynulliad3

[21] http://techiaith.cymru/corpws/Moses/Meddalwedd

[22] http://www.lancaster.ac.uk/fass/projects/biml

Table 6: Map of Welsh POS Tagsets

| Core Tags | CyTag Tags | WNLT Tags |
|---|---|---|
| noun | Egu Ebu Egll Ebll Egbu Egbll | NN NNS NNM NNF |
| verb | Be Bpres1u Bpres2u Bpres3u Bpres1ll Bpres2ll Bpres3ll Bpresamhers Bpres3perth Bpres3amhen Bdyf1u Bdyf2u Bdyf3u Bdyf1ll Bdyf2ll Bdyf3ll Bdyfamhers Bgorb1u Bgorb2u Bgorb3u Bgorb1ll Bgorb2ll Bgorb3ll Bgorbamhers Bamherff1u Bamherff2u Bamherff3u Bamherff1ll Bamherff2ll Bamherff3ll Bamherffamhers Bgorff1u Bgorff2u Bgorff3u Bgorff1ll Bgorff2ll Bgorff3ll Bgorffamhers Bgorffsef Bgorch2u Bgorch3u Bgorch1ll Bgorch2ll Bgorch3ll Bgorchamhers Bdibdyf1u Bdibdyf2u Bdibdyf3u Bdibdyf1ll Bdibdyf2ll Bdibdyf3ll Bdibdyfamhers Bamod1u Bamod2u Bamod3u Bamod1ll Bamod2ll Bamod3ll Bamodamhers | VB VBD VBDP VBDI VBI VBF |
| adj | Anscadu Anscadbu Anscadll Anscyf Anscym Anseith | JJ JJR JJS PDT |
| adv | Adf | RB |
| num | Rhifol Rhifold Rhifolt Rhitref Rhitrefd Rhitreft Gwdig Gwrhuf | CD |
| pnoun | Epg Epb | NNP NNPS |
| intj | Ebych | UH |
| art | YFB | DT |
| part | Uneg Ucad Ugof Utra Uberf | RP |
| prep | Arsym Ar1u Ar2u Ar3gu Ar3bu Ar1ll Ar2ll Ar3ll | IN |
| conj | Cyscyd Cysis | CC |
| pron | Rhapers1u Rhapers2u Rhapers3gu Rhapers3bu Rhapers1ll Rhapers2ll Rhapers3ll Rhadib1u Rhadib2u Rhadib3gu Rhadib3bu Rhadib1ll Rhadib2ll Rhadib3ll Rhamedd1u Rhamedd2u Rhamedd3gu Rhamedd3bu Rhamedd1ll Rhamedd2ll Rhamedd3ll Rhacys1u Rhacys2u Rhacys3gu Rhacys3bu Rhacys1ll Rhacys2ll Rhacys3ll Rhagof Rhadangg Rhadangb Rhadangd Rhaperth Rhaatb Rhacil | PP INT |
| code | Gwfform Gwsym | SC |
| punc | Atdt Atdcan Atdchw Atdde Atdcys Atddyf | PN |
| fw | Gwest | |
| abbrev | Gwacr Gwtalf | |
| lett | Gwllyth | |
| xx | Gwann | |

Table 7: Text Coverage of CySemTagger

| Word-type | CyTag | WNLT |
|---|---|---|
| content words | 88.69% | 65.42% |
| function words | 99.70% | 85.79% |
| total | 91.78% | 72.92% |

the semantic tagger is under compilation by manually annotating the gold corpus by language experts in the CorCenCC project. When this test data becomes available, we will carry out a full scale evaluation of the semantic tagger, including evaluation of the contextual disambiguation accuracy.

## 7. Conclusion

In this paper, we have reported on our development of a Welsh Semantic tagger carried out in the CorCenCC Project. We applied various approaches for rapidly constructing Welsh semantic lexicons and extended and modified existing USAS software framework to develop the CySemTagger system, aiming to provide a tool for automatic annotation of Welsh language corpus data in large scales. In our evaluation, the prototype Welsh semantic tagger demonstrated an encouraging performance and, as it stands, already provides a useful tool for the semantic analysis of Welsh language data. This system is under continuous development, and we will investigate using cross-lingual word embeddings and other techniques, and integrate more efficient algorithms into the system to develop a wide-coverage and accurate Welsh semantic tagger, which will support a range of new research on Welsh Language data in a large scale.

## 8. Acknowledgement

This research work is funded by the UK Economic and Social Research Council (ESRC) and Arts and Humanities Research Council (AHRC) as part of the CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes; The National Corpus of Contemporary Welsh) Project (Grant Number ES/M011348/1).

## References

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Cunningham, H., Maynard, D., and Bontcheva, K. (2011). *Text Processing with GATE*. Gateway Press CA.

Donnelly, K. (2017). Eurfa, a gpled dictionary of welsh. `http://eurfa.org.uk`. Online publication.

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

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

McArthur, T. (1981). *Longman Lexicon of Contemporary English*. Longman London.

Mudraya, O., Babych, B., Piao, S., Rayson, P., and Wilson, A. (2006). Developing a Russian semantic tagger for automatic semantic annotation. In *Proceedings of International Corpus Linguistics Conference 2006*, pages 290–297.

Neale, S., Donnelly, K., Watkins, G., and Knight, D. (2018). Leveraging lexical resources and constraint grammar for rule-based part-of-speech tagging in Welsh. In *Proceedings of the LREC 2018 Conference*.

Padro, L. and Stanilovsky, E. (2012). Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*.

Piao, S., Bianchi, F., Dayrell, C., D'Egidio, A., and Rayson, P. (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)*. Association for Computational Linguistics.

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jimenez, R.-M., Knight, D., Kren, M., Lofberg, L., Nawab, R., Shafi, J., Teh, P., and Mudraya, O., (2016). *Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages*, pages 2614–2619. European Language Resources Association (ELRA), 5. The LREC 2016 Proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Piao, S., Dallachy, F., Baron, A., Demmen, J., Wattam, S., Durkin, P., McCracken, J., Rayson, P., and Alexander, M. (2017). Time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation. *Computer Speech and Language*, 46:113–135.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., and Goranov, M. (2003). Kim - semantic annotation platform. In *Proceedings of 2nd International Semantic Web Conference (ISWC2003)*, pages 834–849.

Rayson, P., Archer, D., Piao, S., and McEnery, A. (2004). The UCREL semantic analysis system. In *Proceedings of the Beyond Named Entity Recognition Semantic Labelling for NLP Tasks Workshop*, pages 7–12.

Zhang, L. and Rettinger, A. (2014). Semantic annotation, analysis and comparison: A multilingual and cross-lingual text analytics toolkit. In *Proceedings of the Demonstrations at the EACL 2014*, pages 13–16.