Natural Language Processing for Under-resourced Languages: Developing a Welsh Natural Language Toolkit

Daniel Cunliffe[a], Andreas Vlachidis[b], Daniel Williams[a], Douglas Tudhope[a]

[a]School of Computing and Mathematical Sciences, University of South Wales, Trefforest, CF37 1DL, Wales, UK.

daniel.cunliffe@southwales.ac.uk (corresponding author), daniel.williams@southwales.ac.uk, douglas.tudhope@southwales.ac.uk

[b]Department of Information Studies, University College London, Gower Street, London, WC1E 6BT, UK.

a.vlachidis@ucl.ac.uk

Abstract

Language technology is becoming increasingly important across a variety of application domains which have become common place in large, well-resourced languages. However, there is a danger that small, under-resourced languages are being increasingly pushed to the technological margins. Under-resourced languages face significant challenges in delivering the underlying language resources necessary to support such applications.

This paper describes the development of a natural language processing toolkit for an under-resourced language, Cymraeg (Welsh). Rather than creating the Welsh Natural Language Toolkit (WNLT) from scratch, the approach involved adapting and enhancing the language processing functionality provided for other languages within an existing framework and making use of external language resources where available.

This paper begins by introducing the GATE NLP framework, which was used as the development platform for the WNLT. It then describes each of the core modules of the WNLT in turn, detailing the extensions and adaptations required for Welsh language processing. An evaluation of the WNLT is then reported. Following this, two demonstration applications are presented. The first is a simple text mining application that analyses wedding announcements. The second describes the development of a Twitter NLP application, which extends the core WNLT pipeline.

As a relatively small-scale project, the WNLT makes use of existing external language resources where possible, rather than creating new resources. This approach of adaptation and reuse can provide a practical and achievable route to developing language resources for under-resourced languages.

Key Words

Natural language processing, under-resourced languages, Welsh, Cymraeg, language technology

1 Introduction

Language technology is becoming increasingly important across a variety of application domains, including consumer electronics, e-learning, knowledge organisation systems, and big data analysis (Evas, 2013). New applications for text mining and information extraction have emerged, such as the analysis of social media, recommender systems and meta reviews of published papers and reports. Voice controlled systems, such as Apple's Siri, Amazon's Alexa and Google's Assistant are becoming widespread. As these applications become commonplace in large, well-resourced languages there is a danger that small, under-resourced languages are increasingly pushed to the technological margins. There is a widening technology gap between well-resourced and less-resourced languages, both within Europe (Evans, 2018) and beyond (Pretorius and Soria, 2017). Whilst under-resourced languages stand to gain most from language technologies (Pretorius and Soria, 2017; Rivera Pastor et al, 2017), they arguably also have much to lose if they are unable to take advantage of them. This is a significant concern, given that there are estimated to be almost 7,000 under-resourced languages in the world (Pretorius and Soria, 2017).

The advanced language-based applications expected by consumers depend upon an underlying infrastructure of language resources (Berger, et al., 2018). The challenge then is how to provide this underlying infrastructure when faced with a lack of resources. This under-resourcing may have many aspects, including lack of government support, lack of funding, lack of skills and knowledge, lack of linguistic resources, lack of visibility of existing resources and technologies, lack of interoperability between linguistic resources and between language technologies, and lack of commercial incentives. Pretorius and Soria (2017) suggest that this under-resourcing often results in the development of small resources with limited scope that are poorly integrated with other available resources.

It has been argued that the development of language resources for under-resourced languages should be community lead, rather than being driven by commercial interests (Hicks, et al., 2018). However, tools for identifying existing language resources and planning future language resources, such as BLARK (Krawer, 2003), may be less effective for under-resourced languages (Prys, 2006). Some more recent approaches, such as the Digital Language Vitality Scale (Ceberio, et al., 2018) and its associated Digital Language Survival Kit (Berger, et al., 2018), are more specifically oriented towards under-resourced languages. One section of the Digital Language Survival Kit, focussing on developing Digital Capacity, charts a path of recommended language resource development, from dictionary making, to speech synthesis and recognition.

1.1 The Welsh Context

This paper describes the development of a natural language processing toolkit for a particular under-resourced language, the Welsh language, Cymraeg. The Welsh language has had official status within Wales since 2011. Data from the 2011 census indicates that there were just over 560,000 Welsh speakers aged 3+ within Wales, approximately 19% of the population (StatsWales, n.d.). Approximately 56% of Welsh speakers in Wales use the language on a daily basis (Welsh Government, 2019). In terms of UNESCO's Language Vitality and Endangerment framework, the language is considered to be "vulnerable" (Moseley, 2010).

The beginnings of the use of the Welsh language online is described by Jones (2010, 2017) and a timeline of significant events is provided by ap Dyfrig (2013). The Language Technologies Unit[1] at Bangor University has been significant in developing language resources for the Welsh language,

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

including spelling and grammar checkers, electronic dictionaries and standardised terminologies (Prys, 2008). Recent work has included making available a Welsh Lemmatiser API (Jones, et al., 2015b), establishing a National Language Technology Portal (Prys and Jones, 2016) and the development of speech technology for Welsh (Prys and Jones, 2018). On the Digital Language Vitality Scale (Ceberio, et al., 2018), Welsh is "Developing", arguably tending towards "Vital" in some aspects.

The Welsh Government's Welsh language strategy, *Cymraeg 2050: A million Welsh speakers* (Welsh Government, 2017), has as one of its aims to "ensure that the Welsh language is at the heart of innovation in digital technology to enable the use of Welsh in all digital contexts". The Welsh Governments most recent *Welsh Language Technology Action Plan* emphases the need for language technologies, language technology skills and a culture of open innovation (Welsh Government, 2018).

This paper describes a project, funded by grants from the Welsh Government Welsh-language Technology and Digital Media programme, to develop a natural language processing toolkit for the Welsh language, oriented to developers not necessarily specialists in computational linguistics. Natural Language Processing (NLP) refers to the computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of human-like language computer processing for a range of tasks or applications (Liddy 2003). The overall aim of the project was to develop a suite of open source software modules that can facilitate and underpin Welsh Language NLP applications. The outcome is a Welsh Natural Language Toolkit (WNLT), comprising a set of core NLP tools, that supports the delivery of applications for the full range of NLP tasks in Welsh, both new applications and adaptation from English to Welsh of existing Information Extraction applications.

This paper begins by introducing the GATE NLP framework (Cunningham, 2002; Cunningham, et al., 2002) which was used as the development platform for the WNLT. It then describes each of the core modules of the WNLT in turn, detailing the extensions and adaptations required for Welsh language processing. An evaluation of the WNLT is then reported. Following this, two demonstration applications are presented. The first is a simple text mining application which analyses wedding announcements. The second describes the development of a Twitter NLP application which extends the core WNLT pipeline. The paper then discusses the WNLT as an open resource, before concluding with observations about the WNLT and developing language resources for under-resourced languages.

2 NLP background and choice of GATE framework

It was decided to employ the open-source GATE NLP framework (Cunningham, et al., 2002) as the development platform for the WNLT. This entailed adapting and enhancing the language processing functionality provided within the framework and making use of external language resources where appropriate. This approach of adaptation and reuse can provide a practical and achievable approach for under-resourced languages.

The decision to adopt GATE was based on two main considerations. Firstly, the platform is well-established and has been continuously maintained for over 20 years, ensuring sustainability, support, use, and dissemination of the toolkit across a strong community of researchers and users. Secondly, the availability of published examples discussing a rule-based adaptation and expansion of the GATE processing resources to a new language domain provided an important resource and a pathway to further development and contribution (e.g. Bontcheva, et al., 2003; Maynard, et al., 2003). In addition, our previous experience using the platform for the semantic annotation and

indexing of archaeological resources (Vlachidis and Tudhope, 2012) assisted the project and led to successful dissemination through the GATE Cloud platform[2].

The General Architecture for Text Engineering (GATE)[3] NLP framework provides an architecture and development environment for building and deploying natural language software components (Cunningham, et al., 2002). The architecture distinguishes two basic kinds of resources; *Language Resources* and *Processing Resources*. Language Resources can be text documents, including a wide range of formats, such as Plain text, MS word, PDF, HTML, XML, etc. Text documents can be loaded and processed individually or as a corpus, i.e. a collection of documents. Processing Resources, on the other hand, are the computational components (modules) of the architecture, such as tokenizer, part-of-speech tagger (POS), sentence splitter, gazetteers and specialised taggers, which are made available by a repository known as the Collection of Reusable Objects for Language Engineering (CREOLE). Processing resources are usually organised in a cascading processing order (pipeline), which performs a specific NLP task.

GATE includes a ready-to-run NLP pipeline, known as ANNIE (A Nearly-New Information Extraction System), which performs Named Entity Recognition (NER), a particular subtask of Information Extraction aimed at classifying units of information to predefined categories relevant to the intended application, such as the names of persons, locations, organisations, expressions of time (Nadeau and Sekine 2007). ANNIE utilises a set of JAPE rules in combination with language resources (e.g. gazetteers) and processing resources (e.g. the Sentence Splitter) in order to deliver the resulting named entities. At the core of the rule-matching mechanism is JAPE (Java Annotation Pattern Engine), a finite state transducer, which uses regular expressions for handling pattern-matching rules aimed at recognising textual snippets that conform to particular patterns (Cunningham, et al., 2000).

The multilingual features of GATE have gradually matured as a result of continuous development and participation in a range of research projects. Languages currently handled in GATE include English, Spanish, Chinese, Arabic, Bulgarian, French, German, Hindi, Italian, Cebuano, Romanian and Russian. The architecture enables software internationalisation via Unicode and offers extensive multilingual support to a wide range of languages. The Unicode features of the architecture were developed as part of the EMILLE project (Baker, et al., 2002), whereas adaption of the NER features, came as a result of the MUSE (Maynard, et al., 2002) and AMITIES (Hardy, et al. 2006) projects. The open architecture of GATE promotes and facilitates reuse and customisation of existing modules to new languages and allows porting of third-party resources (e.g. Stanford CoreNLP, Snowball Stemmer) in NER applications for less supported languages, such as Dutch and Swedish (Binding, et al., 2018). A Surprise Language Exercise (TIDES program) revealed the rapid adaptation qualities of the architecture to a new language (Cebuano), with no previous knowledge or training data available (Maynard, et al., 2003).

A key question when adapting a language processing tool to a new language, assuming that the tool is already developed, is how to limit the effort of development. Practical tips from various system developers for achieving multilinguality include the use of Unicode, modularity, simplicity of rules and the lexicon, and uniform input and output structures (Steinberger, 2010). Adapting to a new language is directly influenced by the modular and reusability qualities of an existing resource. The use of theory-neutral data types is an advantage for GATE because it facilitates reuse, whilst the component-based model of the language processing resources increases modularity. The NER features of ANNIE have been adapted to a range of languages beyond English, such as Arabic, Asian

[2] https://cloud.gate.ac.uk/shopfront#tagged=WNLT
[3] https://gate.ac.uk/

(e.g., Hindi, Bengali), Chinese, French, Germanic, Greek, Italian, Romanian, and Slavic (Bontcheva, et al., 2003). Support is also available for less popular languages, such as Cebuano (Maynard, et al., 2003). However, the multilingual features of GATE are not uniform across all languages, with some languages enjoying a greater availability of processing resources than others. For example, there is no POS tagger available in Bulgarian, whereas POS tagging in Danish relies on porting the Stanford CoreNLP POS tagger into GATE.

When adapting GATE to the Welsh language, it is of course necessary to incorporate a wide range of Welsh-based vocabulary resources as GATE gazetteers. However, the enhancement of GATE for Welsh must also take account of a range of linguistic phenomena that are particular to the Welsh language. The modular properties of GATE facilitated changes to grammars, and extension to English language modules. It was also possible to build on experience of using GATE in an investigation of semantic information extraction (entities and relations) from English language, archaeological grey literature (Vlachidis and Tudhope, 2016).

3 Welsh NLP development for WNLT

The core modules included within the WNLT are:

- Word Segmentation for separating text into words,
- Sentence Boundary Disambiguation for finding sentence boundaries
- Part of Speech Tagger for determining the part of speech of each word
- Morphological Analyser (Lemmatizer) for identifying the root form (lemma) of words. This takes account of prefix Welsh language mutation and incorporates grammatical rules and heuristics for delivering word stem results.
- CymrIE an adapted version for Welsh of the GATE - ANNIE Named Entity Recognition (NER) application for a range of entities such as Persons, Organisations, Locations, and date and time expressions.
- Tweet analysis module for Welsh language
- Standalone version of WNLT with API, command line interface and GUI

The enhancement of GATE for Welsh involved a series of tasks aimed at adapting, mirroring and expanding English language GATE resources into Welsh, including i) functionality expansion by adding new algorithmic arrangements, ii) enrichment of the knowledge base input by importing glossaries and gazetteers and iii) modification of rules and configuration files of the architecture. The Tokenizer, Sentence Splitter, POS tagger and Lemmatizer are core lower level NLP modules. In addition, the NER application (CymrIE) mirrors the English-language ANNIE application to Welsh, by combining the core modules with a range of named entity JAPE grammars and Welsh gazetteers. Eurfa (Donnelly, 2018), the largest Welsh dictionary under a free license containing verbal inflections and lemmas, provided a significant knowledge-based resource for the POS tagger and Lemmatizer. A specialist Welsh language tweet analysis module was developed. Effort was devoted to the development of a standalone version of the toolkit and an API so that other NLP developers are not required to work in the GATE environment.

The development of the WNLT took place over a period of two years. During the first year, the core pipeline was developed, namely the Tokenizer, Sentence Splitter, Part-of-Speech Tagger, Morphological Analyser and some core Named Entity Recognition. In total this took approximately 6 months of developer effort. The Morphological Analyser was the most complex, taking approximately six weeks to develop. During the second year, the text mining demonstrator, Twitter toolkit, Java API, Command Line Interface and standalone Graphical User Interface were developed, taking approximately 4 months of developer effort. The Twitter pipeline involved the most work and

took about 2 months to complete. The evaluation of the different elements of the project (see below) also took a significant amount of time during both phases of the project.

As a rough indication, the project created six Java files, containing over 1,200 lines of programming constructs to support the operation of the Welsh Tokenizer, Sentence Splitter, Part-of-Speech Tagger, and Morphological Analyser. A further 5,200 lines of named entity grammars (JAPE rules) have been created or adapted from English (GATE ANNIE) for the purposes of general Named Entity Recognition (CymrIE). In addition, a new Java-based framework has been created for enabling programmatic access (API) to WNLT language processing resources. The toolkit includes 210,000 words from the Eurfa Welsh Dictionary. Several vocabularies have been created or modified to provide the named entity grammars of CymrIE and TwitterCymrIE. The core named entity application (Person, Organisation, Place and Time), includes a vocabulary of 70,500 general-entity terms. The Twitter toolkit also enhanced the vocabulary with several hundred domain specific entries. Further details are provided in subsequent sections.

Extrapolating from the development effort involved in the WNLT to other projects is not straightforward as the actual time and effort required to develop a basic NLP toolkit for a given language will be influenced by a number of factors. This includes the characteristics of the language and its similarity to existing languages supported by GATE. As will be seen in the discussion below, some parts of the WNLT required more adaptation than others, due to the characteristics of the Welsh language, such as contact mutations.

Another factor is the range and quality of the language resources that are available for the language, and the amount of work required to repurpose them. If the language resources available are severely limited, of poor quality, or in an inappropriate format, it may be better to focus initial efforts on developing or improving these resources. These improved resources may also support other uses in the future. The development of the WNLT benefitted from the availability of a variety of appropriate language resources, most of which were already in digital format and required little effort to repurpose. The modular design of the GATE architecture also means that it is relatively straightforward to replace these resources as and when better resources become available.

The prior experience of the development team is also a factor, programming and NLP experience, but also language ability. Two developers were involved in developing the WNLT, in the first half of the project the developer had programming and NLP experience (including experience with GATE), while in the second half the developer had a wide range of applications development experience but limited NLP experience. Neither the developers, nor the project managers were fluent Welsh speakers, but most members had some degree of Welsh language ability (at Foundation and Intermediate level).

3.1 Tokenizer

The WNLT Tokenizer module extends the GATE Unicode Tokenizer to address the particular word tokenisation requirements of the Welsh language. The new module has modified the regular expression routines of the original Java classes and introduced an extensive post-processing transducer to address a range of linguistic behaviours relevant to the Welsh language. This expression of linguistic phenomena is based on Thorne's (1993) Welsh grammar, which has been used throughout the development of the WNLT.

Investigation revealed that the Welsh language makes use of punctuation elements, such as hyphen and apostrophe in a more elaborate way than the English language, which forms the basis of the default GATE tokenizer. While both languages use hyphenation to join loose compounds (e.g. semi-final) and the apostrophe for denoting the drop of a vowel (e.g. don't), Welsh extends the use of

such punctuation to a wider range of linguistic phenomena. In detail, the WNLT tokenizer module is designed to deliver unified word tokens for the following hyphenated cases: a) loose compounds (e.g. *cyd-aelod, cam-dro*), b) stressed compounds e.g. (*ail-law, di-flas*) c) elements of place names (e.g *Cil-y-cwm, Llys-faen*) and d) constituents such as dd-d, d-d, t-h (e.g *cybydd-dod, hwynt-hwy*). All other cases of hyphenation (e.g. *is-lywydd, lled-wyro*) are tokenised as separate words.

In terms of apostrophe, the WNLT tokenizer is designed to address cases of a) common contractions (i.e. “i, m, n, r, w, ch”, and “th”), b) initial vowel loss (e.g. *'Deryn*), c) medial vowel loss (e.g. *i'engoed*) and d) final consonant loss (e.g *cry' / cryf, hapusa' / hapusaf*), delivering a unified word token for the aforementioned cases. All other cases of apostrophe use, such as the adverbial use of “y” and “r” and verbal adjunct of “yn” are treated as separate tokens. In addition, the module is designed to deliver single word tokens for all ordinal cases, such as “1af, 2il, 3ydd, 31ain, 40fed, 80au, 90au”, and for the special cases *“Ar gyfer” , “Er mwyn” , “Yn erbyn”* and *“Oddi”* when the latter is followed by a preposition.

Grammatical forms beyond basic word tokenisation are addressed by respective modules, including the sentence splitter, part-of-speech tagger, and morphological analyser, as discussed in the following sections.

3.2 Sentence Splitter

The GATE sentence splitter module is described as a cascade of finite-state transducers which segment the text into sentences. The module uses a combination of regular expressions and lexical resources of common abbreviations to distinguish sentence-marking full stops from other kinds of full stops used when abbreviating. The WNLT sentence splitter is a Welsh version of the GATE module that primarily introduces an extended number of abbreviations relevant to the Welsh language. Similar to the original GATE module, the adapted version delivers annotations of type “Sentence” separated by breaks, such as a full stop, which are annotated as “Split” types.

The adapted version maintains an alternative ruleset, which considers newlines and carriage returns differently. As in the case of the original GATE sentence splitter, the alternative rule set can be used when a new line on the page indicates a new sentence. The module uses a large list of approximately 400 Welsh abbreviations divided into 5 distinct categories. The list contains abbreviations of a) linguistic type (e.g. *abs* [absolute], *cfst* [synonym]), b) narrative (e.g. *Brth* [British], *e.e* [for example]) c) scientific (e.g. *Seic* [Psychology], *Tiwt* [Teutonic] d) spatial (e.g. *Morg* [Glamorgan]) and e) temporal (e.g. *C.C* [B.C], *Mer* [Wednesday]). The list of abbreviations is derived from a review of dictionary resources available from the Welsh government portal Cymraeg[4] and particularly from the Welsh dictionary Geiriadur[5]

3.3 Part-of-Speech Tagger

The WNLT part-of-speech tagger is an adapted version to Welsh of the Hepple tagger (Hepple, 2000) which is the main part-of-speech tagger for English, distributed with GATE as part of its default information extraction application ANNIE. The original Hepple tagger uses two separate inputs for producing the part-of-speech tags; a default lexicon and a set of rules which have resulted from training on a large Wall Street Journal corpus. Work on adapting the tagger to the Ceubano Language suggests that using a comprehensive vocabulary and omitting the ruleset can deliver operational results (Maynard, et al., 2003). The adaptation of the Hepple tagger to the Welsh language follows a similar no-ruleset approach to the Ceubano case, using a lexicon for supporting

[4] https://cymraeg.llyw.cymru/
[5] http://www.welsh-dictionary.ac.uk/

the task of POS tagging. The Hepple part-of-speech tagger uses approximately 40 different tags for annotating words as nouns, pronouns, adjectives, verbs, etc. The list contains general tags and specialised tags, for example VB is used to annotate a verb whereas VBG is used for annotating the gerund form of a verb.

The new Welsh version of the Hepple tagger employs the Welsh dictionary Eurfa (Donnelly, 2016) which contains approximately 210,000 words, including mutated forms and inflected verbs of 10,393 lemmas. A mapping exercise was conducted for aligning the Eurfa part of speech labels with the Hepple tagger labels (Appendix A). Eurfa uses common dictionary labels for denoting the part of speech of its entries, for example *adv* for adverb. On the other hand, the Hepple tagger uses an internal set of tags which does not correspond to the schema of standard dictionary labels (e.g. adverb is tagged as *RB*). The alignment between the two tag schemas contained one-to-one relationships where a single Eurfa tag corresponds to a single Hepple tag (e.g. pre-determiner as *preq*[Eurfa] and *PDT*[Hepple]) but it also contained several one to many relationships with more than one Eurfa tags aligned to a single Hepple tag. The opposite direction of alignment never occurred as the EUFA schema is more elaborate than Hepple. Moreover, a set of new tags was introduced, following the labelling form of Hepple, for cases of tagging which are not available to English, such as distinction between masculine and feminine nouns, as well as for direct cases of tagging which are expressed periphrastically in English, for example imperfect, pluperfect and future tense of verbs.

The WNLT part-of-speech tagger also integrates a range of hard-coded regular expressions for detecting syntactic parameters and classifying words to various part of speech categories. Figure 1 presents in a flowchart the main stages of the pos-tagger algorithm. The first stage of the algorithm checks whether the token (word input) is known to the Eurfa dictionary and assigns the corresponding part-of-speech category from the dictionary. In case the input is unknown to the dictionary, the next step of the algorithm inspects for word suffix patterns of maximum size of 5 characters (Appendix B). Depending on the suffix the algorithm classifies the input as masculine noun, feminine noun, plural noun, verb, infinitive verb, or adjective. The regular expressions of the algorithm are consumed within nested conditional statements designed to strengthen accuracy by permitting explicit cases. For example, words ending in “a” in most cases denote a verb, however, words ending explicitly in “dra” and “fa” are tagged as masculine and feminine noun, respectively. The default category noun (NN) is assigned when all regular expressions and conditional statements are consumed, without achieving a match.

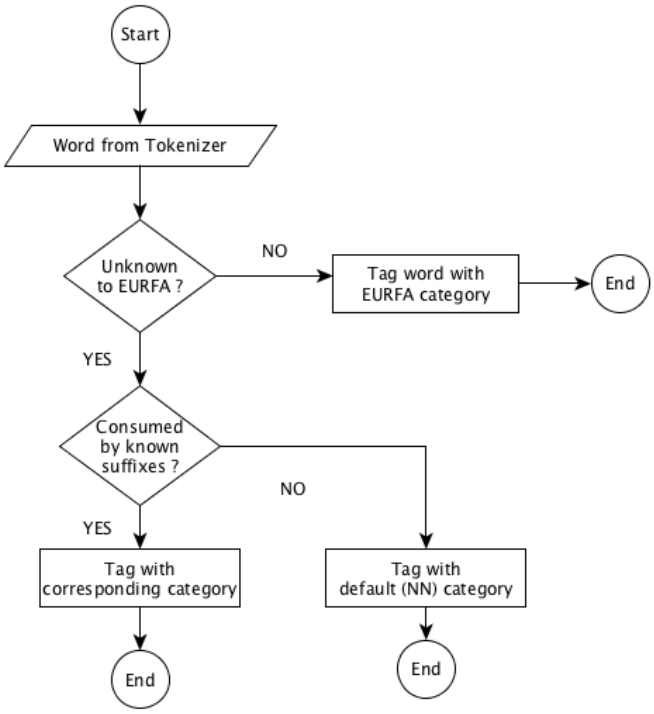

Figure 1 WNLT Part of Speech Tagger Flowchart

3.4 Morphological Analyser (Lemmatizer)

The WNLT Morphological Analyser significantly extends the original GATE Morphological Analyser, in order to address a particular Welsh grammatical feature, initial consonant mutation, which introduces changes to the beginning of words depending on preceding words or their role in a sentence. The WNLT module addresses contact mutation, which is the most common type of mutation in Welsh and manifested in three separate types: Soft, Nasal and Aspirate. Contact mutations are triggered by immediate or nearly immediate words. Other broader contextual types of mutation can exist in Welsh but are not addressed by the module (or WNLT).

The original GATE Morphological Analyser module takes as input a tokenized GATE document and considering a single token and its part of speech tag, one at a time, identifies its lemma and an affix. The module is based on regular expression rules, which contain a left-hand side (LHS) part responsible for matching a rule and a right-hand side (RHS) part, responsible for the function call after matching. The module allows modification and extension of the original ruleset to specific requirements. Different rules can be invoked depending on whether a word is identified as a noun or a verb. Furthermore, the ruleset allows definition of rules for irregular cases.

The Welsh version of the Morphological Analyser extends the original ruleset in order to address the plural grammatical form of Welsh nouns. A range of different lexemes (suffixes) can be used in the construction of plural grammatical forms in Welsh, which is significantly more challenging to address than in English. The modified version introduced 21 generic regular expression rules to address noun inflection of plural forms. The following rule for example,

```
<*>{A}+""esau"==>stem(2,"","au")
```

matches nouns, such as *“athrawesau”* (female teachers) and *“tywysogesau”* (princesses) and delivers the singular forms *“athrawes”* and *“tywysoges”,* respectively. Rule matching is invoked on the last four characters (i.e “esau”), but stemming is applied only on the last two characters by dropping the “au” ending. This is a useful operation that allows a generic based matching within a

certain level of restriction. In addition, 75 irregular cases of plural inflection were specified in the ruleset, addressing cases that cannot be matched by the generic rules, such as such as *nyrsys – nyrs* (i.e nurses-nurse). The module performs verb lemmatisation which is supported by the comprehensive input of the Eurfa dictionary that contains over 30,000 verbs and 4,000 lemmas of inflected verbs.

A dedicated transducer was developed to perform the task of lemmatisation on mutated noun forms. The transducer constitutes a significant extension to the original GATE module. The transducer as seen in Figure 2, resolves lemmas of mutated words through a cascading process that involves dictionary lookup, use of regular expression rules, and final validation of proposed lemmas (Figure 3). The proposed lemmas derive from a set of 12 separate JAPE rules, which are based on Welsh contact mutation grammars (Thorne, 1993). During the validation phase, the proposed word lemma is checked against the Eurfa dictionary containing 168,785 lemmas. Lemma propositions that fail to validate are dropped and the mutated form is assigned as the lemma attribute of the token.

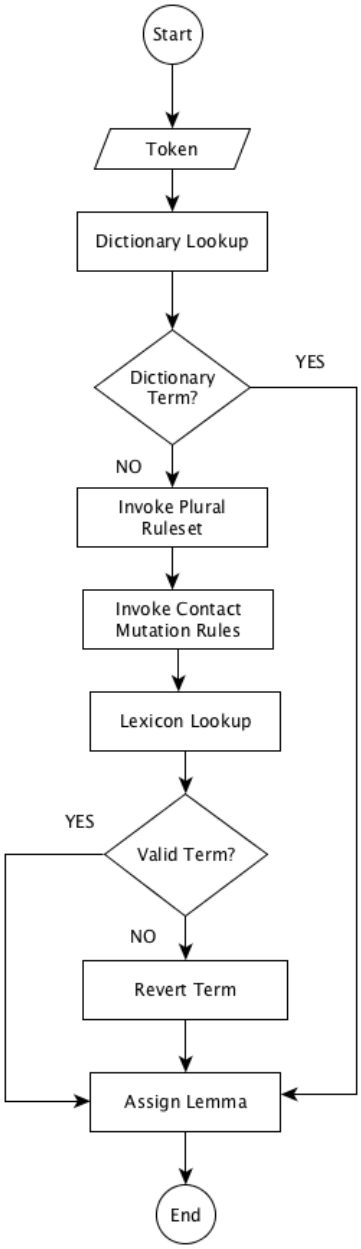

Figure 2. Flowchart of resolving mutated forms of Welsh noun and verbs using grammar rules and the Eurfa dictionary.

The construction of the matching rules is based on expressing abstract linguistic, contact-mutation rules as JAPE grammars. The common matching element to all rules is the presence of a triggering word which causes mutation to an adjunct word, noun and in some cases adjective. The rules have different levels of abstraction and can be grouped by their complexity. The rules of low complexity are highly abstract and are aimed for maximum recall. For example the following rule, matches any word token following a soft contact-mutation trigger.

```
(SOFTCONTACTMUTATION)({Token}):match .
```

A less abstract and more complex rule specifies the form and the context of the matched word. For example, the following rule, matches only feminine nouns following the trigger word *dwy* (two).

```
{Token.string ==~ "[Dd]wy"}({Token.category=="NNF"}):match
```

The most complex rules extend the specificity of the context by engaging further regular expression rules. For example, the following rule matches any word commencing with *Ngh* following the trigger word *yng* (in).

```
{Token.string =="yng"}({Token.string ==~ "(Ngh)(\\w)*"}):match
```

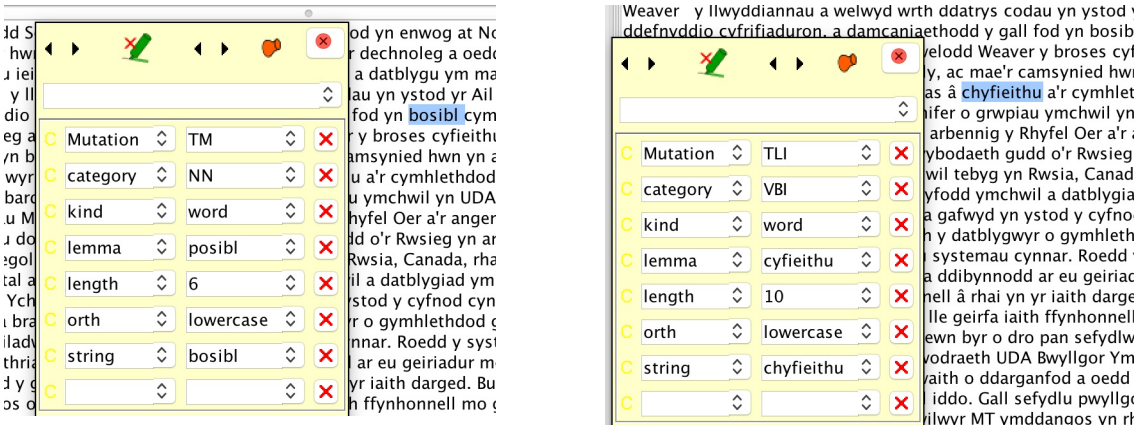

Figure 3. Two separate examples of mutated forms (bosibl, chyfiethu) resolved to original lemmas using contact mutation rules.

3.5 Named Entity Recognition

As part of the toolkit, a Welsh version (CymrIE) of the default GATE Named Entity Recognition (NER) module was developed. This required modification of the existing Named Entity transducer of ANNIE (default NER) and the introduction of a new set of rules for addressing linguistic constructs particular to the Welsh language. In addition, the Welsh version incorporates an extensive list of glossary resources (gazetteers), which supplement or replace the default English-based resources. Over 70,500 unique terms are included, grouped into 70 new lists, which contain Welsh person names, place names, abbreviations, time and number related terms, sport terms, etc. Whilst creating gazetteers of this size from scratch would require a significant effort, where existing resources are available (particularly digital resources), they can often be repurposed with relatively little effort. The lists used in CymrIE were populated with terms freely available on the web from Wikipedia[6], the

[6] https://cy.wikipedia.org/wiki/Gwledydd_y_byd

Welsh Assembly web pages[7], the NHS pages[8] and elsewhere. Additional terms were taken, with permission, from the directory of Welsh organisations and companies maintained by the publishing company Y Lolfa Cyf[9]. The system also retains 51 of the original ANNIE lists relating to person name, place name, company names, etc.

The relationship between the English gazetteers and the Welsh gazetteers is complex, at least in part due to the broader context of Welsh and English language use in Wales. In some cases there is a direct equivalence between an English and Welsh gazetteer because they refer to the same entities, such as the days of the week. In these cases the English gazetteer can be replaced with the Welsh gazetteer. In some cases there will not be a direct equivalence, for example person names or place names. The English person names gazetteer omits numerous Welsh names, so in this case a separate gazetteer of Welsh person names was created to be used alongside the English person names gazetteer. Similarly with place names, some, but not all places have a Welsh name, so a separate additional Welsh gazetteer was created. The retention of English gazetteers alongside Welsh gazetteers also caters for cases where there is a Welsh name for an entity, but that name is not commonly used, or the Welsh and English name are both commonly used. In a few cases there won't be an equivalent English gazetteer, either simply because one doesn't exist, or because the entities only have Welsh names, e.g. the Papurau Bro (Welsh language newsletters produced by the local community). These will typically refer to entities that are specific to Welsh speaking communities.

The rules of the ANNIE English transducer were modified and extended in order to address particular grammatical forms in Welsh. The original rules are to some extent language agnostic, being based on JAPE grammars and input from lookup gazetteers. Therefore, adaption of the rules to Welsh is achieved by deriving lookup matches from the Welsh lists and modifying, where necessary, the JAPE grammars to address particular constructs of the Welsh language. The modification addresses a) use of adjective in Welsh where (unlike English) the noun precedes the adjective, b) use of the definite article in organization names, c) cases of specific vocabulary in the rules, such as *ac*, (and), *Sant* (saint), etc. The resulting CymrIE annotation types and features see (e.g. Figure 4) are identical to ANNIE and are English language, for example Person (gender:male,female), Location (locType: region, airport, city, country, county), Organization (orgType: company, department, government, newspaper), etc.

[7] http://www.assembly.wales/en/memhome/Pages/memhome.aspx
[8] http://www.wales.nhs.uk/sites3/w-page.cfm?orgid=415&pid=33987
[9] https://www.ylolfa.com/directory

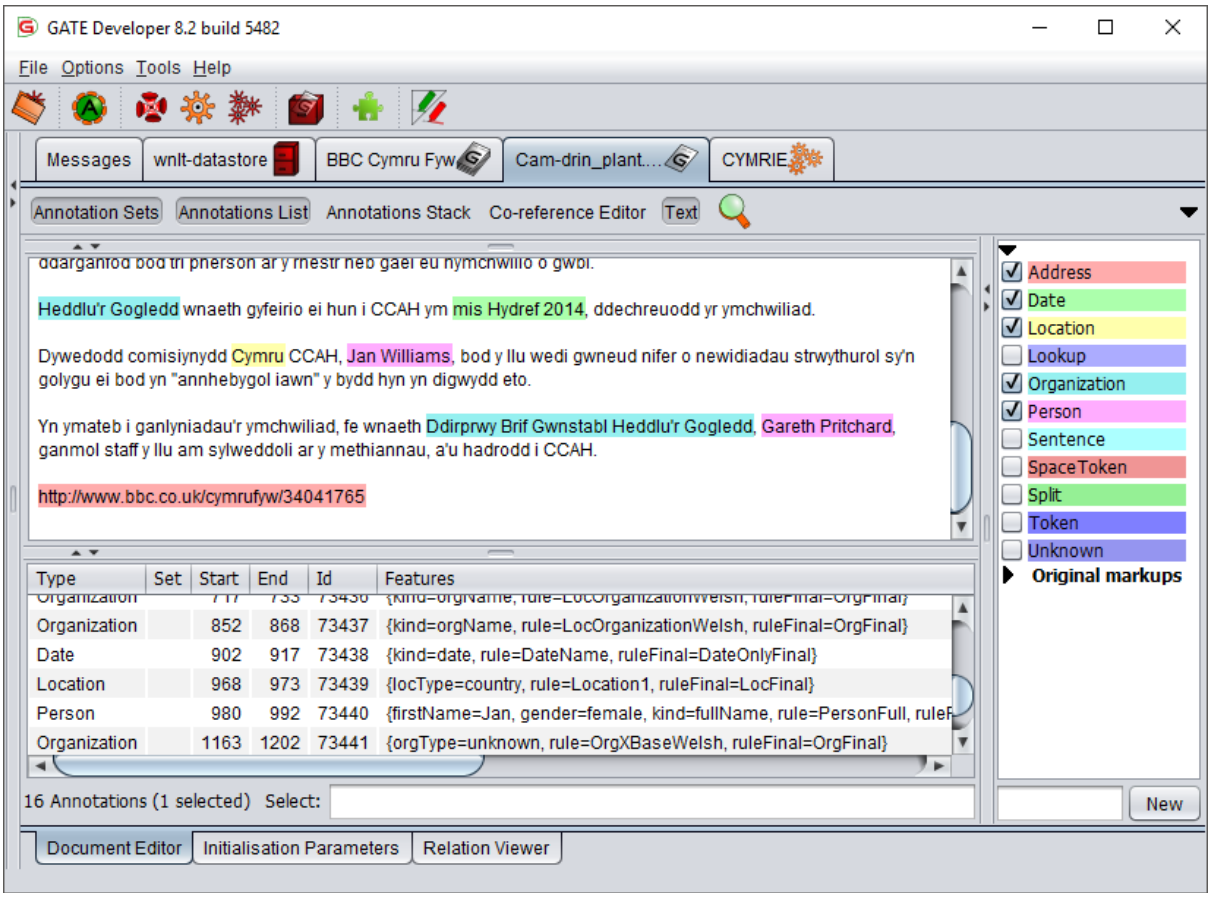

Figure 4 entity identification

4 Evaluation and discussion

The performance of the WNTL pipeline has been evaluated against a human annotated document of 2,200 words following the standard evaluation metrics of Recall, Precision and F-Measure**. Recall** is the fraction of the matches that are relevant to the query that are successfully retrieved. **Precision** is the fraction of retrieved matches that are relevant to the query. **F-Measure** is a harmonic mean that combines Precision and Recall. In most cases, efforts to increase Precision would cause a drop on Recall and vice-versa. Thus, F-measure is a useful metric for summarising the system's performance under a single metric.

The evaluation followed the 'Gold Standard' method where a human annotated input was specified for benchmarking the system performance against the desirable end result. The evaluation document comprised 10 news stories, including politics, life-style and community news originating from the BBC Cyrmru Fyw website[10]. The human annotations were delivered by a Welsh language expert, who received instructions about the task on how to produce word-based annotations including lemma attributes.

The evaluation conducted three separate system runs which assessed the performance of the Tokenizer, Part of Speech Tagger, and Morphological Analyser, respectively. The Recall, Precision and F-Measure scores of each run are presented and briefly discussed. The scores of Partially Correct, Missing, and False positives are also given. **Partially Correct** matches, as the name suggests, are the matches that have some agreement (overlap) with the human annotation span (these were

[10] https://www.bbc.co.uk/cymrufyw

treated as correct in the evaluation). **Missing** matches are the annotations that are not matched by the system. **False Positives** are matches that delivered by the system but not found in human input i.e. wrongly identified matches.

4.1 Tokenizer

The Tokenizer splits the text into simple tokens such as numbers, symbols and words of different types and distinguishes words in uppercase, lowercase, and between types of symbols. The module uses a slightly modified version of the original GATE Tokenizer rules file and an extended JAPE post-processing transducer adapting the generic output of the Tokenizer to the requirements of the Welsh part-of-speech tagger. The module delivered scores shown in Table 1.

Table 1 Results of the tokenizer

Recall	Precision	F-Measure	Correct	Partially correct	Missing	False positives
98.65%	97.86%	98.25%	2191	23	7	25

The task of Tokenization (i.e. identification of word segments) tends to be fairly mechanistic and less semantic. However, the Welsh language presents many challenges relating to use of hyphens and apostrophes. The Tokenizer addresses such cases well. The vast majority of the Partially Correct matches relate to the use of apostrophe (single quote) which in many cases is used instead of double quotes. The system treats such cases as an initial vowel loss, contraction or final consonant loss whereas the apostrophe has been used informally to wrap a phrase instead of a double quote. False Positives mainly relate to the use of certain less popular prefixes, such as *sgrin, aml* that are not captured by the system's rules.

The performance of the Tokenizer could be improved by restricting the system to recognize known contractions and removing the generic rules of initial vowel loss and final consonant loss.

4.2 Part-of-Speech Tagger

The WNLT part-of-speech tagger is based on a modified version of ANNIE's Hepple tagger. The tagger assigns a part-of-speech category (a tag) to each token previously produced by the tokenization process. Tokens can be single words, hyphenated words, numbers, symbols and punctuation. The complete list of part-of-speech categories delivered by the tagger is available in Appendix A. The part-of-speech module tagger performance in terms of Precision, Recall and F-Measure scores is shown in Table 2.

Table 2 Results from the part of speech tagger

Recall	Precision	F-Measure	Correct	Partially correct	Missing	False positives
81.36%	80.71%	81.03%	1807	13	401	419

Many of the false positive matches could have been described as partial matches but the evaluation tool was not able to recognise different levels of specificity. The Part of Speech Tagger uses finer grained (broader and more specific) category tags for nouns, verbs and adjectives which are not identified as being related by the GATE evaluation tool. For example, the Tagger uses NN as a generic label for nouns and NNF as a specific label for Feminine nouns. When a noun is annotated in the Gold Standard as a feminine noun NNF but the system tags it as NN (generic noun tag) this is regarded by the evaluation tool as a False Positive match but it could be considered a partial match.

Similarly, there are generic tags for verbs and specific tags for past or future tense verbs, generic tags for adjectives and specific tags for superlative and comparative adjectives.

Missing and False positive matches are also affected by the contextual use of words. For example, “a” can be used as a conjunction or as a pronoun. Similarly, “y” can be used as a determiner or as a pronoun. The current Tagger does not disambiguate such uses but it is possible to address such cases involving post-processing rules (as in the case of Morphological Analyser discussed below) or by developing generic rules via corpus training and machine learning. The tagger is equipped with an interface to port machine learning rules. In addition, the Eurfa dictionary which drives parts of the Tagger could be enhanced with additional terms to help improve the performance of the tagger.

The performance of the module is encouraging and can be considered a lower bound since many of the false positives (with more specific POS tags) could be considered matches. The WNLT provides the basis of an operational open-source, Part of Speech tagger that can be improved by future iterations.

4.3 Morphological Analyser (Lemmatizer)

The Morphological Analyser considers a single token and its part of speech tag, one at a time. It identifies its lemma, mutation form and in some cases an affix. The WNLT Morphological Analyser has significantly extended the original GATE Morphological Analyser to address the linguistic behaviour of Welsh with regards to inflection and mutation. The tool uses regular expression rules, a Lexicon of term-lemma pairs, a Gazetteer and a post-processing JAPE transducer for validating mutation propositions. The module delivered scores shown in Table 3.

Table 3 Results from the morphological analyser (lemmatizer)

Recall	Precision	F-Measure	Correct	Partially correct	Missing	False positives
80.01%	79.37%	79.69%	1777	2	442	460

The tool resolves word lemmas by a combination of dictionary look up, regular expression matching and contextual evidence post-processing. The contribution of the dictionary is critical not only because it delivers lemmas in the first instance but also because it is used to validate the lemmas proposed by the post-processing rules; extending the vocabulary will improve performance. The Lemmatizer would benefit from iterative evaluation techniques that introduce contextual patterns beyond the scope of the simple contact mutation patterns that are currently employed by the tool.

The task of lemmatization in Welsh is one of the most challenging tasks for Natural Language processing due to the use of three different type of mutation (soft, nasal, and aspirate). Performance is reaching operational levels and forms the basis of an open-source Lemmatizer that can be improved by future iterations.

4.4 Related work

The WNLT Welsh part-of-speech tagger was employed in work to develop an automatic Welsh language semantic annotation system, CySemTagger (Piao, et al., 2018), as part of the wide ranging CorCenCC project to create a national corpus of contemporary Welsh[11]. CySemTagger builds on part-of-speech tagging in order to produce broader annotations, such General/Abstract, Food/Farming,

[11] https://corcencc.org/

Emotion, Time[12]. Subsequently the CyTag part-of-speech tagger was developed as the Welsh part-of-speech tagger for the CorCenCC project (Neale, et al., 2018). In an evaluation of text coverage (percentage of words in the test corpus identified by the two taggers), CyTag achieved higher coverage than the WNLT part-of-speech tagger (92% vs 73% coverage respectively) mainly due to lemmatisation performance and is adopted by CySemTagger (Piao, et al., 2018). CyTag employs a two stage rule based constraint grammar method, developed previously for the multilingual Bangor Autoglosser[13] tagger (Donnelly and Deuchar 2011), which is likely to be a significant element in the higher text coverage by CyTag. The constraint grammar software is the result of a long running project involving some considerable effort in multiple languages. In related work, Ezeani, et al. (2019) report promising early results from a small scale experiment with a Welsh language neural network model for partof-speech and also semantic tagging. This was able to leverage the manually annotated CorCenCC project evaluation corpus and thus avoid the usual effort in constructing a training set for machine learning. In general constructing a representative training set from scratch is a major endeavour, although it is sometimes possible to repurpose existing resources or corpora.

5 Text mining demonstration application

In order to demonstrate the potential application of CymrIE in supporting text mining applications, and to validate the information extraction pipeline with a particular test case, a small text mining application was created in the domain of social history. The texts in question were wedding announcements taken from Llais Cwmtawe[14], a papur bro covering a number of communities in the Swansea valley in South Wales. Papurau bro are local, Welsh language newsletters produced by the community.

This type of application requires a more detailed and context dependent analysis than that required in general-purpose information extraction. For example general-purpose extraction might be satisfied in identifying a person, or a female person, whereas in a wedding announcement it is important to identify the female person who is the bride, as opposed to the female person who is the bride's mother, or the female person who officiated the ceremony.

Additional gazeteers[15] were added for the exercise. Heuristics for identifying entities were identified and JAPE rules were written for identifying and annotating specific person roles, such as Bride, Groom, BrideFather, BrideMother, GroomFather, GroomMother, and also ServiceLocation and ServiceDate. Figure 5 shows a fictional example. The results are shown in Table 4.

[12] http://ucrel.lancs.ac.uk/usas/
[13] http://bangortalk.org.uk/autoglosser.php
[14] https://sites.google.com/site/llaiscwmtawe/
[15] E.g. a partial gazetteer of Welsh chapels created specifically for the application, and a gazetteer of Welsh Choirs (from the Directory provided by Y Lolfa Cyf)

Fe briododd Lleucu Evans, Ystalafera, gyda Ewan Davies, Abertawe, yn Eglwys Abercraf ddydd Sadwrn 24 Medi. Merch Janet a Huw yw Lleucu ac wyres i'r diweddgar Meinwen a John Hughes, Ystalafera.

(*Lleucu Evans of Ystalafera married Ewan Davies of Swansea at Abercraf Church on Saturday 24 September. Lleucu is the daughter of Janet and Huw and the granddaughter of the late Meinwen and John Hughes of Ystalafera.*)

Figure 5 fictional wedding announcement (with translation)

Table 4 Results from wedding announcement text mining application using CymrIE

Annotation	Recall	Precision	F-Measure	Correct	Missing	False positives
Bride	67.7%	100%	80.74%	21	10	0
BrideFather	52.4%	91.7%	66.69%	11	10	1
BrideMother	54.2%	100%	70.30%	13	11	0
Groom	62.5%	95.2%	75.46%	20	12	1
GroomFather	57.9%	91.7%	70.98%	11	8	1
GroomMother	47.4%	90%	62.10%	9	10	1
ServiceLocation	58.3%	66.7%	62.22%	14	10	7
ServiceDate	89.7%	96.3%	92.88%	26	3	1

The system applied a set of 18 heuristics to the annotations produced by the general CymrIE resource. These heuristics embodied simple assumptions and observations about the typical format of the wedding announcements. For example, the first female person mentioned is likely to be the bride. If there is a date in the announcement, it is likely to be the date of the service. The set of heuristics is small and represents only a limited model of the actors and events that might be included in a wedding announcement. It is intended as an illustration of the type of text mining application that could be built on top of the base CymrIE system.

5.1 Evaluation

A sample of 32 wedding announcements were extracted and coded manually for the purpose of evaluation. As can be seen in Table 4, the results obtained for dates were good (a specific set of rules for Welsh language temporal expressions was developed) and those for person roles promising for an initial study. Location performed less well. In general, precise location results were hampered by insufficient glossary entries; with hindsight specific wedding venues (chapel and street names) were overly ambitious given the existing vocabulary resources.

It should be noted that the wedding announcements tended to be very similar in their content and organisation. They also tended to be very short. These factors make it possible to achieve relatively good results using only a small set of heuristics. The evaluation was also conducted on manually extracted wedding announcement texts. An important element of future work would be the automatic detection of potentially relevant text sections from the complete newsletter.

6 Twitter NLP tools demonstration application

In order to demonstrate how additional NLP tools could be developed on top of the CymrIE processing resources, an information extraction system was developed for Welsh tweets (see Figure 6). TwitterCymrIE reuses the Tokenizer, Sentence Splitter, Part of Speech Tagger, Morphological Analyser and Named Entity Recognition modules from CymrIE. It adds five processing resources found in the TwitIE Twitter information extraction system (Bontcheva, et al., 2013) available in GATE Developer. TwitIE itself reuses and adapts existing resources available within GATE, illustrating how the GATE framework can support multiple levels of reuse and adaptation to suit different purposes. Three of the TwitIE resources were used without adaptation, namely the Annotation Set Transfer, Language Identification and Emoticons Gazetteer. The remaining two, Hashtag Gazetteer and Tweet Normaliser, required some adaptation, as described below.

6.1 Annotation Set Transfer

This processing resource identifies the tweet and the metadata associated with the tweet. 'User mentions' and 'hashtags' are important metadata to identify. This processing resource generates the Tweet, Hashtag, UserID and URL annotations. This resource did not require adaptation.

6.2 Language Identification

TwitterCymrIE uses the default language models (Welsh, English, Spanish, German, Dutch and French) distributed with TextCat (Cavnar and Trenkle, 1994; Carter, et al., 2013) to classify the language of the tweet. The TextCat processing resource adds a "lang" feature to the "Tweet" annotation generated by the Annotation set transfer processing resource. This resource did not require adaptation.

6.3 Emoticons Gazetteer

This processing resource generates the Emoticon annotation. It uses a gazetteer to normalise emoticons such as :-) and :} to the emoticon :). This resource did not require adaptation.

6.4 Hashtag Gazetteer

A hashtag is a word or phrase that is used as a form of discovery metadata to categorise the tweet. The processing resource attempts to split the hashtag into multiple words. It employs JAPE rules to identify any camel casing and it was adapted to use the Eurfa dictionary and named entities used in CymrIE to break up the hashtag into separate words.

6.5 Tweet Normaliser

The Tweet Normaliser processing resource is of particular interest as it attempts to correct spelling mistakes, shortened words, slang etc. This resource was adapted to use the Eurfa dictionary and other gazetteers in CymrIE. Spelling mistakes are corrected by comparing the Levenshtein distance between each word, if a word is found in the gazetteers with a Levenshtein distance of 2 or less then it is corrected to that word. To normalise slang, a specialised gazetteer is used, this maps words such as 'l8r' to the full form, 'later'. A small additional gazetteer for Welsh slang was added, for instance mapping 'v' to the full form, 'fi'. TwitterCymrIE provides the ability to turn off normalisation, or to inspect the normalisations.

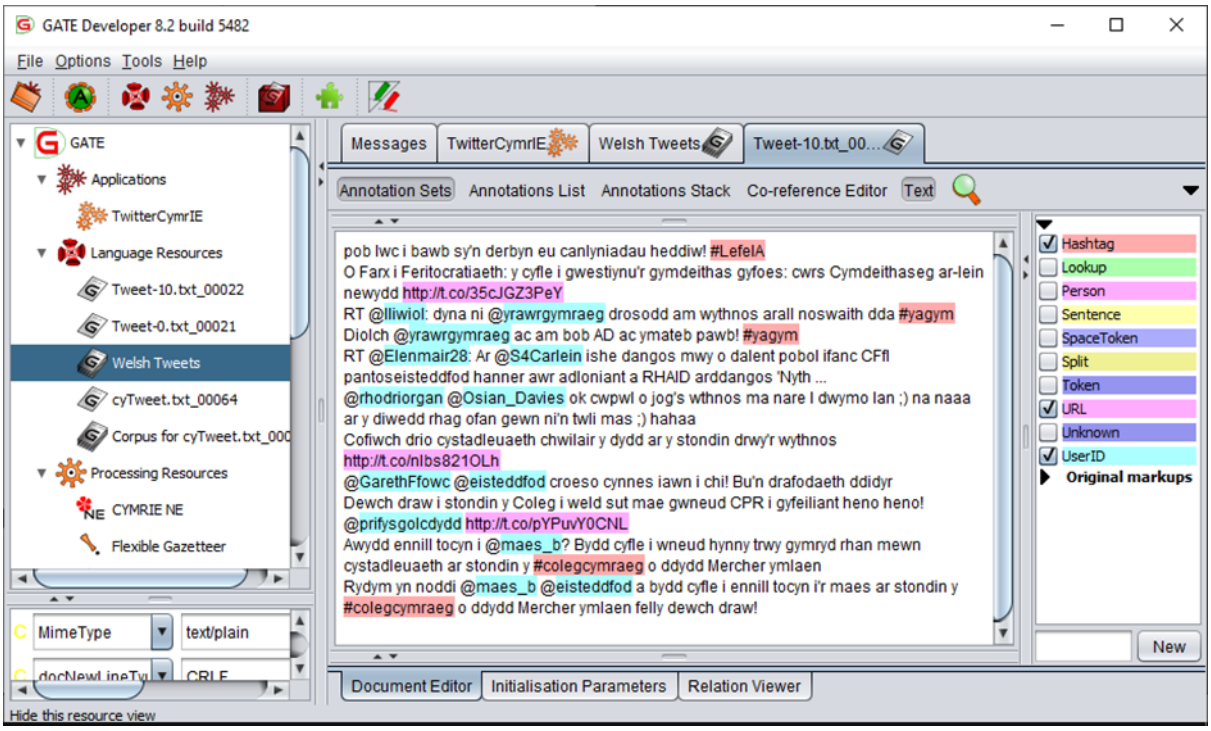

Figure 6 TwitterCymrIE showing tweet entity annotations for hashtags, URLs and account names

6.6 Evaluation

The performance of the TwitterCymrIE pipeline has been evaluated against a human annotated set of 120 Welsh language tweets selected from the Corpus of Welsh Language Tweets[16] (Jones, et al., 2015a). The evaluation followed the 'Gold Standard' method and the standard evaluation metrics of Recall, Precision and F-Measure were used**.** The human annotations were delivered by a Welsh language expert, who received instructions about the task and on how to produce word-based annotations including lemma attributes.

Analysis of the results revealed that TwitterCymrIE had annotated to more detailed part of speech categories than those employed in the Gold Standard. In order to compensate for this, the NLP output was normalised to the same categories used in the human annotation (*VBD, VBDP, VBDI, VDI, VBF -> VB, etc.)*.

Table 5 Results from TwitterCymrIE

Annotation	Recall	Precision	F-Measure	Correct	Overlap	Missing	False positives
Token	99.87%	87.66%	93.37%	2180	151	3	328
Token.category	81.32%	71.38%	76.03%	1770	128	436	761
Token.lemma	79.39%	69.69%	74.22%	1824	29	481	806

The evaluation assessed the performance of the pipeline on the output of three separate processes; tokenization, part-of-speech tagging and lemmatisation, as respectively shown by the three annotation types Token, Token.category, and Token.lemma, in Table 5. The Token annotation carries the boundaries of a meaningful textual snippet, in most cases a single word or a punctuation, symbol, etc. The *category* refers to the part of speech of a token assigned by the part-of-speech

[16] http://techiaith.cymru/data/corpora/twitter/?lang=en

tagger, whereas the *lemma* refers to the *root* form of a word that is delivered by the Welsh Morphological Analyser (lemmatiser). Figure 7 presents an example of the set of attributes for the word *gyfrwng* (*n.* medium) produced by the three separate processes, including the part of speech category noun (NN), the contact mutation form (TM) and the respective lemma (*cyfrwng*), together with a range of additional attributes about the size of the entry, its kind and letter case.

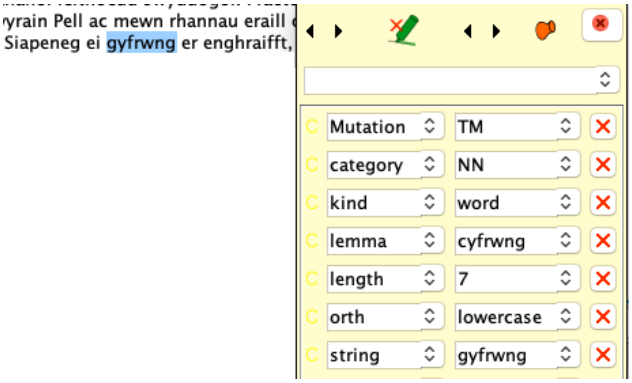

Figure 7 Example of Token, part-of-speech Category and Lemma attributes of (*gyfrwng*) - the mutated form of the word *cyfrwng*

The overall performance of the TwitterCymrIE pipeline compared with the CymrIE pipeline across all three annotation types is decreased by approximately 5%. As seen in Figure 8, the Recall rates between the two pipelines are comparable but the Precision is reduced by nearly 10%. This drop in performance can be explained by the inherent difficulties of microblogging content that relate to the use of colloquial language, extensive abbreviation and use of special characters. However, this is significantly less than the typical reduction in performance that affects English POS when used in microblogging context (Derczynski et. al, 2013). General purpose English POS taggers can deliver accuracy scores that reach as high as 97% but when applied to microblogging context their accuracy can reduce by up to 25%, delivering scores around 75%. Hence, it is necessary to retrain and introduce adaptations, including normalisation, gazetteer modification and regular expression-based tagging of Twitter-specific part-of-speech tags (Bontcheva et.al 2013). Our results demonstrate that the CymrIE modules can be used in the processing of microblogging content with some confidence, but further adaptions would benefit precision rates.

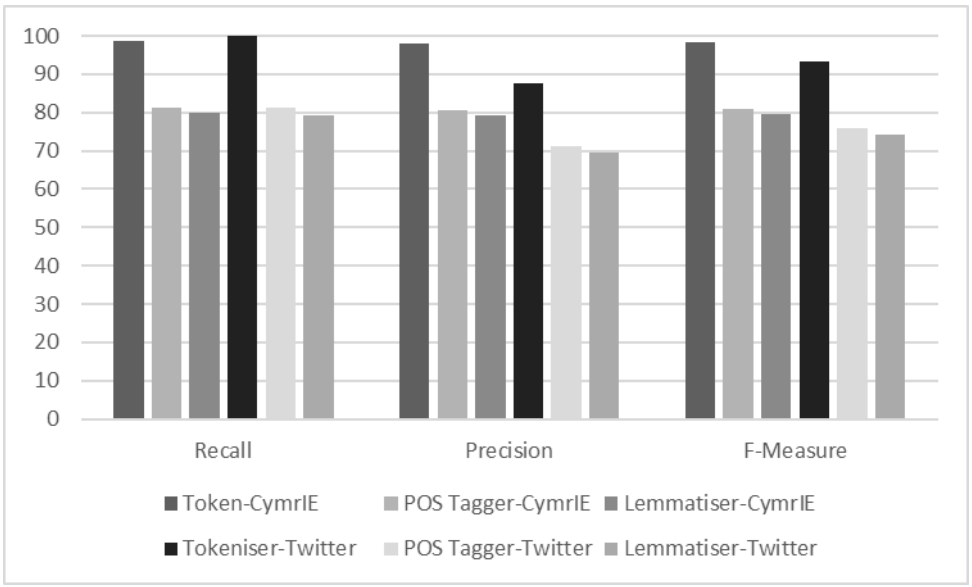

Figure 8 TwitterCymrIE pipeline compared with the CymrIE pipeline across the three annotation types

7 WNLT as an Open Resource

Given the lack of resources typically available to an under-resourced language, the openness, modularity, interoperability and reusability of the language resources that are available, is critical (Rivera Pastor, et al., 2017; Soria, et al., 2014; Witt, et al., 2009).

The WNLT is distributed on SourceForge[17] under a GPL open-source licence[18] as specified under the terms of the grant funding the project. This aligns with the Welsh Governments *Welsh Language Technology Action Plan* which emphases the need for a culture of open innovation (Welsh Government, 2018), though the terms of a GPL licence may restrict some commercial uptake.

Figure 9 Top level elements of the CymrIE API

The introduction of Welsh language functionality to the GATE framework exposes the Welsh language to the GATE developer community and the wide range of functionality that the GATE platform provides[19]. The WNLT GATE distribution requires developers to employ the GATE framework which is open source but inevitably will not suit every context and organisation. The WNLT toolkit was therefore made more accessible by creating a Java API, command line interface and a standalone graphical user interface for CymrIE. Future work could include the deployment of CymrIE and its low level processing resources via a REST API.

Software developers can use the toolkit's Java API to integrate the CymrIE pipeline and reuse required parts of the pipeline such as CymrIE's tokenizer and part of speech tagger. Thus application development can take place without requiring knowledge of the GATE platform. Software developers who wish to integrate CymrIE into their web, desktop or mobile applications can do so by including CymrIE's Java API and writing a few lines of code as detailed in the user manual. Figure 9 shows the top level elements of the CymrIE API.

[17] https://sourceforge.net/projects/wnlt-project/
[18] GNU Library or Lesser General Public License version 3.0 (LGPLv3)
[19] GATE sees itself as a 'One-Stop-Shop for Text Analytics and Semantics' https://gate.ac.uk/biz/usps.html

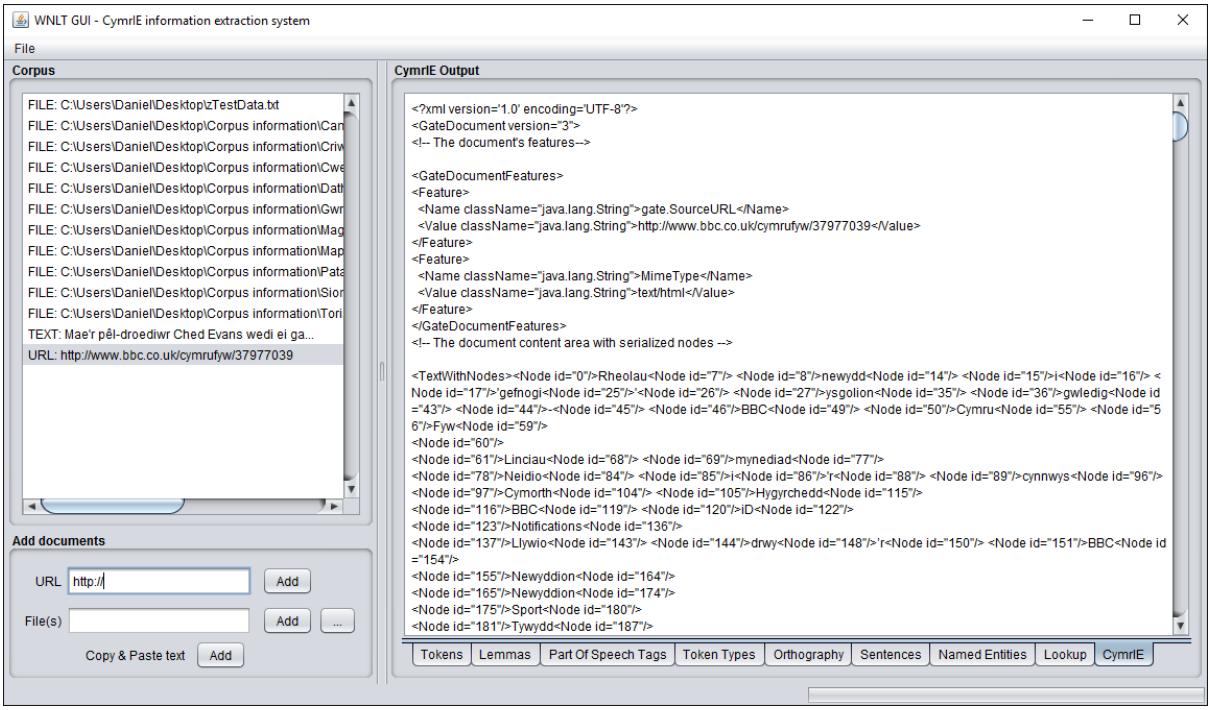

Figure 10 CymrIE standalone user interface

Increasing accessibility further, the core NLP components in the CymrIE pipeline can be utilised via a command line interface which is useful for scripts and other programming languages. The CymrIE pipeline also has its own graphical user interface so that non-technical users can obtain results from the CymrIE pipeline without needing to understand the GATE Developer's setup, configuration and graphical user interface. Figure 10 illustrates the CymrIE user interface. The TwitterCymrIE functionality is included in the standalone version and the CymrIE API.

8 Conclusions

The WNLT project has developed a core natural language processing toolkit for the Welsh language, *Cymraeg*. The project employed the open-source GATE NLP framework as its development platform. Existing language processing functionality provided within the framework was adapted and enhanced for use with Welsh, reducing the amount and complexity of the new code required. The WNLT is included in the official GATE plugins[20] and is thus available to the extensive GATE developer community. In addition, a Java API, command line interface and a standalone graphical user interface for CymrIE have been developed and are freely available.

As a relatively small scale project, the WNLT makes use of existing external language resources where possible, rather than creating new resources. This approach of adaptation and reuse can provide a practical and achievable route for under-resourced languages.

While the WNLT provides core natural language processing resources, it has several limitations. It is limited by the finite nature of its knowledge resources, such as dictionaries and gazetteers. In an under-resourced language these knowledge resources are likely to be limited in number and extent.

Another limitation is in the extent to which the WNLT fully reflects the features of the Welsh language. For example, only simple contact mutations are currently modelled. However, as the

[20] https://gate.ac.uk/gate/plugins/Lang_Welsh/

WNLT is open source, developers are able to extend the existing code to deal with more complex transitive mutations. The evaluation of the WNLT is encouraging but suggests areas for improvement. The modular GATE architecture makes it relatively easy for developers to implement changes or substitute knowledge resources and pipeline components.

The WNLT is not intended to be an end in itself, rather it provides essential language processing infrastructure upon which Welsh language applications can be built. The particular features of the Welsh language, such as mutation, make the use of NLP tools particularly important in search applications rather than relying on literal string matches. The text mining application illustrates the potential for constructing specialist information extraction tools with relatively little effort, post-processing the results obtained from using CymrIE as the base system. This case study demonstrates one way in which the CymrIE NER tools could be adapted for a user application. It also illustrates the potential for new applications of the Welsh NLP resource in providing tools for social/historical analysis of Welsh textual resources.

In addition to demonstrating how social media analysis tools can be developed by employing the CymrIE processing resources, TwitterCymrIE itself can provide a platform for applications analysing Welsh Twitter data, such as different forms of sentiment and opinion analysis, as well as study of language use in this medium (see for example, McMonagle, et al., 2018; Nic Giolla Mhichíl, et al., 2018).

In a wider context, the WNLT project could serve as an exemplar of a relatively low cost approach to the development toolkits that unlock the potential of language technologies in under-resourced languages.

Acknowledgements

Funding: The WNLT project was supported by grants from the Welsh Government, Welsh-language Technology and Digital Media programme.

Thanks are due to Y Lolfa Cyf for permission to use data from their Directory of Welsh organisations and companies.

References

ap Dyfrig, R. (2013). Hanes y We Gymraeg. http://www.tiki-toki.com/timeline/entry/84932/Hanes-y-We-Gymraeg/ Online publication.

Baker, P., Hardie, A., McEnery, T., Cunningham,H. and Gaizauskas R. (2002). EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation. In *Proceedings of 3rd Language Resources and Evaluation Conference* (LREC'2002), pp. 819-825

Berger, K.C., Hernaiz, A.G., Baroni, P., Hicks, D., Kruse, E., Quochi, V., Russo, I., Salonen, T. Sarhimaa, A. and Soria, C. (2018). *The DLDP Digital Language Survival Kit*. The Digital Language Diversity Project, www.dldp.eu

Binding C., Tudhope D. and Vlachidis A. (2018). A study of semantic integration across archaeological data and reports in different languages. *Journal of Information Science*, 45(3): 364-386. https://doi.org/10.1177/0165551518789874

Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M.A., Maynard, D. and Aswani, N., (2013). Twitie: An open-source information extraction pipeline for microblog text. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*,pp. 83-90.

Bontcheva, K., Maynard, D., Tablan, V. and Cunningham, H. (2003). GATE: A Unicode-based infrastructure supporting multilingual information extraction. In *Proceedings of Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages* (IESL'03) , Borovets, Bulgaria.

Carter, S., Weerkamp, W. and Tsagkias (2013). Microblogging language identification: Overcoming the limitations of short, unedited and idiomatic text. Language Resources and Evaluation, 47: 195-215.

Cavnar, W. and Trenkle, J. (1994). N-gram-based text categorization. In *Proceedings of the Annual Symposium on Document Analysis and Information Retrieval*, pp.161-175.

Ceberio, K., Gurrutxaga, A., Soria, C., Russo, I. and Quochi, V. (2018). *How to Use the Digital Language Vitality Scale*. The Digital Language Diversity Project, www.dldp.eu

Cunningham, H. (2002). GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36(2): 223-254.

Cunningham H, Maynard D, Bontcheva K. and Tablan V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications, *Proc. 40th Annual Meeting of Association for Computational Linguistics*; Philadelphia, pp 168-175.

Derczynski, L., Maynard, D., Aswani, N. and Bontcheva, K., (2013). Microblog-genre noise and impact on semantic annotation accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*,pp. 21-30.

Donnelly, K. (2018). Eurfa. http://eurfa.org.uk Online publication.

Donnelly, K., Deuchar, M. (2011). Using constraint grammar in the Bangor Autoglosser to disambiguatemultilingual spoken text. In *Constraint Grammar Applications: Proceedings of the NODALIDA 2011 Workshop*.
http://www.bangortalk.org.uk/publications/Donnelly2011_Constraint_Grammar.pdf

Evans (2018). European Parliament Committee on Culture and Education, *Draft Report on Language Inequality in the Digital Age, 2018/2028(INI)*. Brussels, European Parliament.

Evas, J. (2013). *Y Gymraeg yn yr Oes Ddigidol – The Welsh Language in the Digital Age*. META-NET White Paper Series. Available online at http://www.meta-net.eu/whitepapers

Ezeani, I. , Piao, S., Neale, S., Rayson, P., Knight, D. (2019) Leveraging pre-trained embeddings for Welsh taggers. In: Augenstein, I., Gella, S., Ruder, S., Kann, K., Can, B., Welbl, J., Conneau, A., Ren, X. and Rei, M., (eds.) *Proceedings of the 4th Workshop on Representation Learning for NLP* (RepL4NLP-2019). 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), 02 Aug 2019, Florence, Italy. Association for Computational Linguistics (ACL) , pp. 270-280.

Hardy, H., Biermann, A., Inouye, R.B., McKenzie, A., Strzalkowski, T., Ursu, C., Webb, N. and Wu, M., (2006). The Amitiés system: Data-driven techniques for automated dialogue. *Speech Communication*, 48(3-4):354-373.

Hepple, M. (2000). Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (ACL-2000), Hong Kong.

Hicks, D., Baroni, P., Berger, K.C., Hernaiz, A.G., Kruse, E., Quochi, V., Russo, I., Salonen, T., Sarhimaa, A. and Soria, C. (2018). *The DLDP Road Map*. The Digital Language Diversity Project, www.dldp.eu

Jones, R. (2010). Cilfachau electronig : geni'r Gymraeg ar-lein, 1989-1996. *Cyfrwng* 7: 21-36.

Jones, R. (2017). 'Porn shock for dons' (and other stories from Welsh pre-web history). In Gerard Goggin and Mark McLelland (Eds.), *Routledge Companion to Global Internet Histories*. New York: Routledge: pp.256-268.

Jones, D. B., Robertson, P. and Taborda, A. (2015a). Corpus of Welsh Language Tweets. http://techiaith.org/corpora/twitter/?lang=en Online publication.

Jones, D. B., Robertson, P. and Prys, G. (2015b) Welsh language Lemmatizer API Service. http://techiaith.cymru/api/lemmatizer/?lang=en Online publication.

Krauwer, S. (2003). The Basic Language Resourse Kit (BLARK) as the first milestone for the language resources roadmap. In: *Proceedings of the International Conference on Speech and Computer*, Moscow.

Liddy, E. (2003). Natural Language Processing. In Drake, M. (ed.), *Encyclopedia of Library and Information Science*, London: Taylor and Francis, pp. 2126–2136

Maynard, D., Tablan, V. and Cunningham, H. (2003). NE recognition without training data on a language you don't speak. In *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition*, vol 15, pp. 33-40.

Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., & Wilks, Y. (2002). Architectural elements of language engineering robustness. *Natural Language Engineering*, 8(2-3):257-274.

McMonagle, S., Cunliffe, D., Jongbloed-Faber, L., Jarvis, P. (2018). What can hashtags tell us about minority languages on Twitter?: A comparison of #cymraeg, #frysk, and #gaeilge. *Journal of Multilingual and Multicultural Development*, 40(1): 32-49. DOI: 10.1080/01434632.2018.1465429

Moseley, C. (ed.) (2010). Atlas of the World's Languages in Danger, 3rd edn. Paris, UNESCO Publishing. Online version. Retrieved 17 February 2011, from http://www.unesco.org/culture/en/endangeredlanguages/atlas

Nadeau D, Sekine S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*; 30(1): 3 – 26

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 Eleventh International Conference on Language Resources and Evaluation* (LREC-2018), Miyazaki, Japan, 3946-3954

Nic Giolla Mhichíl, M., Lynn, T. and Rosati, P. (2018). Twitter and the Irish language, #Gaeilge – agents and activities: exploring a data set with micro-implementers in social media. *Journal of Multilingual and Multicultural Development*, 39(10), 868-881, DOI: 10.1080/01434632.2018.1450414

Piao, S., Rayson, P., Knight, D. and Watkins G. (2018). Towards a Welsh Semantic Annotation System. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC-2018), Miyazaki, Japan, 980-985.

Pretorius, L. and Soria, C. (2017). Introduction to the special issue. *Language Resources and Evaluation*; 51(4), 891-895.

Prys, D. (2006). The BLARK matrix and its relation to the language resources situation for the Celtic languages. In *Proceedings of SALTMIL Workshop on Minority Languages*, organized in conjunction with LREC, Genoa, Italy, 31-32.

Prys, D (2008). The ultimate Welsh language survival kit: an overview of ten years of language technology work at Canolfan Bedwyr. *Mercator Media Forum*, 10(1): 4-10.

Prys D. and Jones D.B. (2016). National Language Technologies Portals for LRLs: A Case Study. In: Vetulani Z., Mariani J., Kubis M. (eds) Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2015. *Lecture Notes in Computer Science*, vol 10930. Springer, Cham.

Prys D. and Jones D.B. (2018). Gathering Data for Speech Technology in the Welsh Language: A Case Study. In: C. Soria, L. Besacier and L. Pretorius (Eds.), *Proceedings CCURL 2018 Workshop*, Miyazaki, Japan.

Rivera Pastor, R., Tarín Quirós, C., Villar García, J.P. Badia Cardús, T. and Melero Nogués, M. (2017). *Language Equality in the Digital Age: Towards a Human Language Project*. Scientific Foresight Unit (STOA), European Parliamentary Research Service, Brussels.

Soria, C., Calzolari, N., Monachini, M., Quochi, V., Bel, N., Choukri, K., Mariani, J., Odijk, J. and Piperidis, S. (2014). The language resource Strategic Agenda: the FLaReNet synthesis of community recommendations. *Language Resources and Evaluation*, 48:753-775.

StatsWales (n.d.). *Welsh speakers by local authority, gender and detailed age groups, 2011 Census*. https://statswales.gov.wales/Catalogue/Welsh-Language/WelshSpeakers-by-LocalAuthority-Gender-DetailedAgeGroups-2011Census Online publication.

Steinberger, R. (2010). Challenges and methods for multilingual text mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (LREC 2010), pp. 19-21.

Thorne, D.A. (1993). *A comprehensive Welsh grammar*. Wiley-Blackwell.

Vlachidis, A., and Tudhope, D. (2012) 'A pilot investigation of information extraction in the semantic annotation of archaeological reports'. *International Journal of Metadata, Semantics and Ontologies*, Vol. 7 No.3, pp.222-235.

Vlachidis A. and Tudhope D. (2016). A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain. *Journal of the Association for Information Science and Technology*; 67(5), 1138-1152.

Welsh Government (2017). Cymraeg 2050: Welsh Language Strategy. http://gov.wales/topics/welshlanguage/welsh-language-strategy-and-policies/cymraeg-2050-welsh-language-strategy/?lang=en

Welsh Government (2018). Welsh Language Technology Action Plan. https://gov.wales/topics/welshlanguage/welsh-language-strategy-and-policies/welsh-language-policies-upto-2017/wl-technology-and-digital-media/?lang=en

Welsh Government (2019). Welsh language results: Annual Population Survey, 2001- 2018 https://gov.wales/sites/default/files/statistics-and-research/2019-05/welsh-language-results-annual-population-survey-2001-to-2018.pdf

Witt, A., Heid, U., Sasaki, F. and Sérasset, G. (2009). Multilingual language resources and interoperability. *Languages Resources and Evaluation*, 43(1): 1-14.

Appendices

Appendix A: Mappings between Eurfa and Hepple POS tags

Hepple +	**Eurfa**
CC conjunction	conj
CD number	num
DT determiner	det.def, pron.rel
IN preposition	prep, prep+poss
INT interrogative	int
JJ adjective	adj, adj.eq, adj.pl, quan, quant, ord
JJR adjective comparative	adj.comp
JJS adjective superlative	adj.sup
NN noun	n(mf)-sg, n(n)sg, n(\N), m, h, \N
NNS noun plural	n(mf)-pl, n(n)-pl, m-pl
NNP proper noun singular	name
NNPS proper noun plural	name-pl
NNM noun masculine	n(m)-sg
NNF noun feminine	n(f)-sg
PDT pre-determiner	preq
PP pronoun	pron, pron.emph,, pron.rel.neg, pron.rel.neg.subj, adj.poss, adj.dem
RP particle	prt, prt.aff, prt.int, prt.neg, prt.neg+obj, prt.perf
RB adverb	adv

UH interjection	e, im
VB verb	v(/N), v(pres), v(pres.emph), v(pres.indef), v(pres.indef.neg), v(pres.neg), v(pres.rel), v(subj), v(cond), v(imper)
VBD verb past tens	v(past), v(subj.past), v(subj.past.hyp),
VBDP verb pluperfect	pluperf, pluperf.hyp
VBDI verb imperfect	v(imperf), v(imperf.neg),
VBI verb infinitive	v(infin)
VBF verb future	v(fut)
PN punctuation	’'[]()❴❵⟨⟩:,،、‒…-!.?’’“”"";\\/⁄
SC special characters	All other characters e.g. £$%* etc

Appendix B: POS Suffix Patterns

Pattern	**Tag**	**Description**
“id”, “od”, “awd”, “ydd”, “edd”, “yd”, “deb”, “ineb”, “fel”, “adur”, “dra”, “aint”, “cyn”, “der”, “iant”, “mon”, “wr”, “yr”	NNM	Masculine Noun
“oedd”, “eb”, “fa”, “wraig”, “ell”, “en”, “es”, “red”	NNF	Feminine Noun
“ydd”, “au”, “os”	NNS	Plural Noun
“lyd”, “ig”, “adwy”, ”aidd”, “ar”, “ed”, “gar”, “lon”, “us”	JJ	Adjective
“ad”, “aid”, “d”, “el”, “ur”, “a”, “fan”, “ain”, “o”	VB	Verb
“u”	VBI	Verb Infinitive

Vitae

Dr Daniel Cunliffe is an Associate Professor in the Faculty of Computing, Engineering and Science at the University of South Wales. He is a member of the Welsh Government's Welsh-language Technology Board. Since 2001, his research has applied insights and methods from fields such as computer-mediated communication, human-computer interaction and user experience design to the study of minority language use in information and communications technology.

Dr Andreas Vlachidis is a Lecturer/Assistant Professor in Information Science at the UCL Department of Information Studies. He has a long-term experience contributing to European and UK research projects focusing on cultural heritage data modelling and the multilingual application of Natural Language Processing in archaeological grey literature. His main research interests are in Information Extraction, Text Analytics, Knowledge-Based Systems and Ontologies. He is a certified text analyst, a fellow of the Higher Education Academy (FHEA) and a member of the British Computing Society (BCS). His research on the semantic indexing of cultural heritage resources has received several awards including the outstanding paper award from the Emerald Literati Network.

Dr Daniel Williams is leading AI projects at one of the biggest online schools in the UK, Wey Education. He is multidisciplinary with experience and knowledge that spans Project Management, Mathematics, Statistics, Computer Science, Artificial Intelligence and Software Development, and applies these invaluable problem-solving techniques to real world problems. Delivering solutions, insights and visualizations in the form of software, website and/or mobile applications.

Professor Douglas Tudhope leads the Hypermedia Research Group at University of South Wales. He was PI on the AHRC funded STAR, STELLAR and SENESCHAL projects (Semantic Tools for Archaeological Resources) and led the Linking Archaeological Data Work Package for the ARIADNE FP7 Infrastructures Project. Since 1977, he has been Editor of the journal, New Review of Hypermedia and Multimedia. He co-authored the JISC State of the Art Review on Terminology Services and Technology and the JISC Terminology Registry Scoping Study. He was a member ISO TC46/SC9/SC8 (and NISO) working group developing a new thesaurus standard (ISO 25964).
http://hypermedia.research.southwales.ac.uk/kos/