Proceedings of the CLTW 4 @ LREC2022 , pages 30–39
Marseille, 20-25 June 2022

Evaluation of Three Welsh Language POS Taggers

Gruffudd Prys, Gareth Watkins
Bangor University,
Bangor, Wales,
{g.prys, g.watkins}@bangor.ac.uk

Abstract

In this paper we describe our quantitative and qualitative evaluation of three Welsh language Part of Speech (POS) taggers. Following an introductory section, we explore some of the issues which face POS taggers, discuss the state of the art in English language tagging, and describe the three Welsh language POS taggers that will be evaluated in this paper, namely WNLT2, CyTag and TagTeg. We then describe the challenges involved in evaluating POS taggers which make use of different tagsets, and introduce our mapping of the taggers' individual tagsets to an Intermediate Tagset used to facilitate their comparative evaluation. We introduce our benchmarking corpus as an important component of our methodology, before describing how the inconsistencies in text tokenization between the different taggers present an issue when undertaking such evaluations, and discuss the method used to overcome this complication. We proceed to illustrate how we annotated the benchmark corpus, then describe the scoring method used. We provide an in-depth analysis of the results followed by a summary of the work.

Keywords: POS Tagger, Welsh, Evaluation, Machine Learning

1. Introduction

POS tagging remains an important tool for modern methods of extracting information from data, and it is often used alongside the artificial intelligence techniques that currently claim the headlines. Due to the growing use of these methods, it is important to ensure that the POS taggers available to the Welsh language are of a high standard, that their strengths and weaknesses are known, and that they are proven to be fit for purpose.

However, creating a fair quantitative comparison between existing taggers is not a straightforward task due to their use of different tagsets and their reliance on differing methodologies (namely rule-based and statistical approaches) where the methods used to develop and evaluate the taggers are not directly comparable.

1.1. Impartiality

This paper summarizes an unpublished report on Welsh part-of-speech taggers that we were commissioned to write for the Welsh Government as one of the outputs of the Text, Speech and Translation Technologies for the Welsh Language project - the same project which also funded our work on the the TagTeg tagger. We therefore find ourselves evaluating taggers that include our own, and wish to make our interests clear. Whilst we have strived to be open and impartial in our evaluation, it is inevitable that TagTeg will fit closely with our ideal of how a tagger should behave as we could influence its design. To ensure a fair test of each of the three taggers, we have opted for a simple evaluation that treats different linguistic theoretical perspectives as equally valid, accepting tagging that is different from our preferred interpretation providing it has linguistic justification and is not clearly a mechanical error on the part of the tagger. We have also sought to justify our criticisms (especially in the more qualitative aspects of the evaluation), in an open and transparent way that allows the reader to draw their own conclusions.

2. Taggers

Accurate automatic tagging is not a simple task. As Hagerman (2012) notes in reference to English "Many of the most common used words have more than one possible usage, making their part-of-speech ambiguous". Automatically tagging Celtic languages such as Welsh is further challenged by complex morphological processes such as initial letter mutations which can lead to what Lamb and Danso (2014) call 'data sparsity', as well as an increase in ambiguous forms.

2.1. Accuracy of English Language Taggers

Over a decade ago, Manning (2011) reported that state-of-the-art English language taggers could achieve an accuracy of 97.3% at word level, and that such accuracy was comparable or even better than that of a human annotator. Figures reported by the Association for Computational Linguistics (2019) show that systems have not improved significantly since 2011 in terms of accuracy. It appears that rules based methods are currently less used than statistical methods, which 'have become the mainstream ones obtaining state-of-the-art performance' (Nguyen et al., 2016). Sadredini et al. (2018) appear to agree, in part, noting that 'Generally in NLP, and specifically in POS tagging, statistical and neural network (NN)-based approaches have been favored over rule-based approaches, because they have shown higher accuracy and the training is straightforward to automate'.

2.2. The Taggers Selected for Evaluation

Due to time constraints, our funder's interests, the complexities involved in evaluating taggers which are fundamentally different,[1] and the need to map multiple tagsets to a common interset (discussed in section 5 below), we limited our evaluation to a cross section of taggers recently developed within Welsh universities with public funding. Thus we describe the evaluation of the University of South Wales' Welsh Government funded WNLT2 tagger (Cunliffe et al., 2022), the CyTag tagger (Neale et al., 2018), produced as part of the Cardiff University-led AHRC and ESRC funded CorCenCC project (Knight et al., 2020), and Bangor University's TagTeg tagger (Prys et al., 2020), also funded by the Welsh Government. In the future we also hope to evaluate other taggers, such as the Cyslib tagger (Hicks, 2004; Jones et al., 2015) (part of the Welsh spell/grammar checker Cysill) and the Autoglosser 2 tagger (Donnelly, 2018).

2.3. WNLT2

WNLT (Welsh Natural Language Toolkit) predates the other taggers. The first version was developed by the University of South Wales Hypermedia Research Group between 2015 and 2016. A second version, WNLT2,[2] was developed in a follow-up project between 2016 and 2017. It uses the GATE (General Architecture for Text Engineering)[3] architecture originally developed by the University of Sheffield in 1995. WNLT2's tagging component is based on the Hepple tagger (Hepple, 2000), but with major modifications designed to enable it to categorise Welsh language input (Cunliffe et al., 2017).

A rules-based tagger, WNLT2's lexicon is based on a version of Eurfa (Donnelly, 2013) modified to use the Hepple tagset. However, WNLT2 only uses rules when trying to tag words not found in the lexicon. It does so based on their endings (e.g. by specifying that an unfamiliar word ending with ending with 'fa' is a feminine noun). Ambiguous wordforms appear to be given the same default POS in all contexts. For instance, in the lexicon 'mae' (English:it is) is listed as a verb. This is correct however 'mae' can also be a mutated form of the noun 'bae' (English:bay). The implication of this is that 'mae' will never be correctly tagged as a noun when it acts as such. The basis on which one possible tag is prioritized over another in the WNLT2 data is not clear, but the logical choice would have been to choose based on frequency. (Jurafsky and Martin, 2021) note that accuracy of up to 92% could be achieved with a similar approach in the case of English. When the tagger is unable to find a wordform in the lexicon, and when its rules are unable to determine the POS of the wordform, WNLT2 assigns its noun tag (NN) as default (a common tactic to improve the score a tagger is likely to get).

2.3.1. Ease of Use

The WNLT2 team developed a simple user interface for their tool, one which benefits novice users as it does not require them to learn how to use the more complex GATE architecture. However, tagging 1500 sentences using this simple interface proved very slow, even on powerful machines (CPU i7, 32Gb RAM). It was also necessary to turn to Mac computers for the purpose of the evaluation. We failed to get the program to work on Linux machines, and although it worked on Windows machines, it was restricted to using the Windows default encoding,[4] thus on Windows machines UTF-8 characters such as 'ŷ' and 'ŵ' were corrupted. We chose to ignore these UTF-8 problems for the evaluation, but the need for a Mac computer to make real use of the software is potentially problematic.

2.3.2. Reported Accuracy

WNLT2 authors reported an accuracy of 81% from the first version of WNLT on a gold corpus of 2221 tokens (Williams, 2017).

2.4. CyTag

CyTag[5] is another rule-based Welsh POS tagger. Neale et al. (2018) note that their "motivations for developing a bespoke solution for Welsh POS tagging are based on the requirements, aims and scope of the CorCenCC ... project", that is, CyTag was created to tag the corpus of contemporary Welsh that would form the main outcome of that project.

In common with WNLT2, CyTag uses a version of Eurfa for its core lexicon. CyTag is based on the VISL-CG3 library,[6] a software library for implementing constraint grammar (Karlsson, 1990), a technique used for tag disambiguation. It works by implementing rules handwritten by linguists to identify the syntactic context of a token and limit the number of possible interpretations for the token's tag accordingly. Thus, unlike WNLT, CyTag can select the appropriate tag for a POS-ambiguous wordform according to its syntactic context.However many rules are required to enable accurate disambiguation, and although rule-based taggers have historically produced good results, one of their disadvantages is that developing and maintaining these rules while avoiding conflict between them is specialized and often difficult work. In the case of CyTag, it appears that the rules do not always resolve some common cases where more than one tag corresponds to a single wordform. In the case of the wordform 'ceir', for example, the lexicon indicates that it can represent a verbal form of 'cael' (English:to have) or a plural

[1] I.E. rule based v statistical.

[2] Available free of charge under the LGPL3 license from https://sourceforge.net/projects/wnlt-project/

[3] See https://gate.ac.uk/

[4] Windows-1252

[5] Available for download under the GPL-3.0 license from https://github.com/CorCenCC/CyTag

[6] Available for download under the GPL-3.0 license from https://visl.sdu.dk/constraint_grammar.html

noun (English:cars), but the tagger does not successfully disambiguate between them, and at times suggests a preposition. In addition, there appears to be no provision for coping with common words missing from the program's lexicon. As a result, words that are not in its lexicon are tagged with the unk (unknown) tag.

2.4.1. Ease of Use

We were able to follow the instructions, download CyTag and install VISL-CG3 without issue. Users will need to be comfortable using the command line and Python to do so. Python is often used for doing NLP work and has a reputation for being relatively easy to learn. However, the documentation for VISL-CG3 starts with a prominent *Caveat Emptor* section, and users are instructed to download the latest nightly version rather than a proven release. The coding conventions and structure of CyTag seemed streamlined, but the lack of version information on the GitHub meant we could not ensure that the CyTag we tested was the same as that described by Neale et al. (2018) in 2018.

2.4.2. Reported Accuracy

An early version of CyTag was reported to have reached 93% accuracy when tagging with basic tags on a gold standard corpus of 611 tokens (Neale et al., 2018).[7]

2.5. TagTeg

TagTeg[8] is our statistical Welsh-language POS tagger, based on the tagger found in spaCy's[9] NLP library. spaCy offers several advantages. It provides clear and comprehensive documentation. It is a free and open source library that is actively developed and updated. The impressive results reported by the developers of spaCy (2022) are supported by academic and peer-reviewed experiments and comparisons such as Jiang et al. (2016) and Schmitt et al. (2019) which have shown that spaCy compares well with similar technology, being both fast (Choi et al., 2015; Schmitt et al., 2019) and accurate (Partalidou et al., 2019).

The way spaCy's tagger works is not based on rules set by the developer. Rather, it must be trained with a corpus of human annotated sentences. To this end, a corpus of Welsh language sentences was collected and annotated. Prodigy[10] and spaCy were used to facilitate the annotation. In order to further improve results, the tagger was also trained with a list of 76,000 individual words where each had only one possible interpretation in terms of their POS. These words were sourced from Bangor University's comprehensive lexicon (Prys et al., 2021), however their inclusion as single word training sentences should not be seen as adding a lexicon to the model but rather as a means of influencing the probabilities contained within the model.

The Universal Dependencies (UD) tagset was used to tag the sentences' tokens. This tagset is based on "an evolution of (universal) Stanford dependencies (de Marneffe et al., 2006, 2008, 2014), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008)." (Universal Dependencies, 2021a).

2.5.1. Ease of Use

As with CyTag, installing spaCy is a simple matter for anyone who is familiar with Python. Also as with CyTag, a non-technical user-friendly interface, such as that provided by WNLT2, is not currently available.

2.5.2. Reported Accuracy

An early model resulted in 91% accuracy when testing using a test corpus that was not part of the training data.[11]

3. Tagsets

One of the major challenges identified during evaluation was the taggers' use of different tagsets. WNLT2 uses a tagset of 27 tags based on the Hepple tagset, which itself closely matches the well-known Penn Treebank tagset (Gorrell et al., 2010). CyTag uses two tagsets, one 'basic', one 'enriched'. The basic tagset consists of 13 tags (14 if we count the unk tag), but maps to an enriched tagset of 145 categories for a more detailed description of the Welsh language's morphological features. The tagset follows Expert Advisory Group on Language Engineering Standards (EAGLES) to enable mapping to other languages through 'Intermediate Tags' (Leech and Wilson, 1996). As mentioned, TagTeg uses the UD tagset. This tagset is a relatively simple tagset containing 17 POS tags, and was designed to facilitate crosslingual tagging. UD also provides for a comprehensive list of morphological features as additional features.

3.1. Rules-Based Tagging and Statistical Tagging: Implications

In the case of rule-based taggers, the tagger is closely tied to the tagset used within the tagger's tag rules. Changing this tagset is no trivial task, and as will be discussed in the following paragraph, neither is mapping between tagsets. However, statistical taggers can be trained to use any tagset by feeding it a corpus of materials annotated using that tagset. In addition, by annotating corpora rather than developing grammar rules, alternative statistical taggers can be trained on

[7] Neale et al. (2018) also describe the content of this corpus. We believe this is the test set used: `https://github.com/CorCenCC/welsh_pos_sem_tagger/blob/master/data/cy_both_tagged.data`

[8] Available to download under the MIT licence from https://github.com/techiaith/model-tagiwr-spacy-cy

[9] See https://spacy.io/

[10] See https://prodi.gy/

[11] Available from https://github.com/techiaith/brawddegau-tagiedig

the same data.[12] This makes it possible to maintain lists such as those of the Association for Computational Linguistics (2019) which directly compare the accuracy of the different taggers trained on the same corpora and using the same tagsets.

However, as this evaluation consisted of two rule-based (non-trainable) taggers, and a total of three different tagsets that did not map directly to each other, evaluation based on a pattern similar to that used in ACL's 'State of the Art' list was not possible. Therefore, to enable a fair and valid comparison between taggers, it was necessary to develop an Intermediate Tagset as illustrated in Table 1. Direct mapping between the three tagsets was impossible. As can be seen in Table 1, Tagteg's ADJ tag can be represented by several tags in WNLT2: namely JJ, JR, JJS and PDT. However WNLT2's PDT tag can also be represented by TagTeg's DET tag. The inability to map one tagset to another directly is compounded when we attempt to map additional tagsets together. In order to overcome these difficulties, we adopted an approach of generalisation. We also decided to allow for multiple 'correct' answers within the gold corpus, so that individual taggers were not penalized incorrectly if the tag used was justified under the schema used by the tagger. This point is exemplified in section 6.1 below. The process resulted in a simplified tagset featuring basic tags to which the the three taggers' complex tags were mapped. By functioning as a bridge between the different tagsets, the Intermediate Tagset enables a comparison of the respective taggers' output. Such a technique has been used by others to facilitate the comparison of different NLP systems (see, for instance Jiang et al. (2016) and Schmitt et al. (2019)).

Intermediate tag	WNLT2 tag	CyTag tag	TagTeg tag
ADF	RB	Adf	ADV
ANS	JJ, JR, JJS, PDT	Ans	ADJ
ARDD	IN	Ar	ADP
ATALN	PN	Atd	PUNCT
BAN	DT, PDT	Ban, YFB	DET
BERF	VB, VBD VBDI, VBDP, VBF, VBI	B	AUX, VERB
CYS	CC	Cys	CCONJ, CONJ, SCONJ
EBYCH	INTJ	Ebych	INTJ
ENW	NN, NF, NNM, NNS	E	NOUN
GEIR	RP	U	PART
MISC	SC	Gw	SYM, X
PRIOD	NNP, NNPS	Ep	PROPN
RHAG	INT, PP	Rha	PRON
RHIF	CD	Rhi	NUM
?		unk	

Table 1: Mapping between tagsets.

4. The Benchmarking Corpus

In order to compare the accuracy of different taggers, we curated a corpus of 1,500 Welsh sentences drawn from a variety of different sources. This Benchmarking Corpus was specifically designed to include a broad representation of contemporary Welsh. The corpus contains a variety of registers and styles to reward taggers that are able to generalize and recognize less-standard forms and orthography in addition to the more literary and formal forms. The Benchmarking Corpus contains examples of transcribed Welsh from recordings of spoken Welsh that use standard informal written apostrophed forms e.g. 'cer'ed' (= cerdded, English:walk). The corpus also includes natural informal written Welsh from text messages and emails where there is less use of the apostrophes than found in 'standard' informal Welsh. Efforts were made to ensure that the sentences also included a variety of dialects and subject matter, and reference was made to the sample frameworks used by CEG (Ellis et al., 2001) and CorCenCC in doing so. The sentences featured a variety in terms of person and tense, and were of varied lengths. The corpus contains sentences from important sources such as Wikipedia, Coleg Cymraeg Cenedlaethol Cymru[13] materials and the CorCenCC and Siarad (Deuchar et al., 2009) corpora. The corpus is available for distribution under the CC-BY-SA license, as used and required by some of these constituent resources.

5. Tokenization

One of the other considerations that complicates the comparison of different taggers is that each tagger may tokenize differently Paroubek (2007), and the three taggers described in this paper each tokenize some texts differently. For instance URLs are tagged differently by all three taggers. Moreover WNLT 2 tokenize 'ar gyfer' (for) and 'er mwyn' (for the sake of) and other commonly used multi word prepositions as single tokens. However, this decision isolates 'ar gyfer' (for) from related forms such as 'ar ei gyfer' (for it/him), and is not followed by CyTag and TagTeg which choose to tokenize multi word prepositions as individual tokens. As a result, comparison between a sentence tagged by a tagger and the 'gold standard' sentence is not straightforward. To facilitate comparison, it was decided to limit the evaluation to identically tokenized sentences.

[12] For example, we understand that Dr Johannes Heinecke has already used the data annotated by us to train a UDPipe tagger.

[13] See https://www.colegcymraeg.ac.uk/

This gave us just over 500 sentences, which we deemed sufficient to give the taggers a fair and useful evaluation.

6. Gold Tagged Benchmarking Corpus

6.1. Annotation Method

The 500 sentence gold corpus was annotated or hand-tagged by an experienced researcher and verified by a senior researcher. In most cases, each token was annotated with one POS tag from the Intermediate set, as can be seen in Table 2. Where it was not possible to map the expected tagger-assigned tag to one specific Intermediate tag, more than one acceptable tag was considered permissible, as exemplified in Table 3. These were separated by a comma. Where it was not possible to map the expected tagger-assigned tag to one specific Intermediate tag, more than one acceptable tag was considered permissible, as exemplified in Table 3. These were separated by a comma.

Mae	Huw	yn	siarad	Cymraeg
BERF	PRIOD	GEIR	BERF	PRIOD

Table 2: Example tagged sentence (sentence literal translation: 'Be Huw is speak Welsh').

Beth	sydd	angen	ei	wneud
RHAG	BERF	ENW	BAN, RHAG	BERF

Table 3: Tagging with more than one tag (sentence literal translation: 'what is need it doing').

6.2. Dealing with Taggers that Offer More than One Possible Tag

While TagTeg and WNLT2 assign one tag per token, CyTag often offers a number of possible tags where the tagger failed to reach a specific conclusion. This is problematic when trying to determine the appropriate method for evaluating these taggers alongside each other. To a degree, the desired behaviour of a tagger depends upon its intended use. In some circumstances, it is arguably better to offer a choice of possible tags rather than risk suggesting the wrong tag. This is the case with the tagger used by Welsh spell/grammar checker Cysill, for example, where it is essential that the checker does not misinterpret the grammar of a text as this could lead it to recommend that the user amends a correct text. Nevertheless, most typical applications expect taggers to output an explicit and unambiguous output, and the inability to select one tag from amongst the number of possible tags should arguably be considered a shortcoming of the tagger. However, as the most appropriate behaviour is task-dependent, we decided to evaluate CyTag's output twice; once in a 'strict' manner, penalizing any ambiguous tagging as if it was an incorrect tag, and again in a 'generous' way by marking ambiguous tagging as correct (where the correct tag was included). By reporting both scores, we let the reader decide on the appropriate interpretation.

7. Scoring Method

We started by using the latest available versions of the three taggers to tokenize and tag the 1500 sentences found in the Benchmarking Corpus. From those 1500 sentences we selected 500 sentences where the tokenization was consistent between each tagger (see Section 5). Those 500 sentences were then manually annotated using the Intermediate Tagset to create the gold standard evaluation corpus. The tags assigned by each tagger were then mapped and converted to the corresponding tags in the Intermediate set. For example, each WNLT2 tag which corresponded to a noun, namely NN, NNF, NNM and NNS, was mapped to the ENW intermediate tag. In doing so, each sentence, along with its corresponding tags, was converted to a common structure in order to compare each of them in turn with the corresponding gold sentences and their associated tags, as can be seen in Table 4.

To facilitate the scoring, we created a benchmarking script that reads the output of each tagger in turn, and works its way through the sentences using these structures to compare the tagger's tags with the corresponding gold tags. The script records the correctly and incorrectly assigned tags and records which combination of token and tag was problematic for the tagger. This provided a score in the form of a percentage of the correct tags in a sentence, and allowed us to calculate a total for all text in the 500 sentence selection from the benchmarking corpus. It also provided an overview of the number of tagging errors and a list of all the tokens incorrectly tagged. To concentrate on a simple, clean cut comparison we avoided mention of precision, recall and F scores in our report.

In addition, we were able to create a complete report for each sentence, which shows every token in the sentence and displays the tag assigned originally by the tagger (following conversion to the relevant Intermediate tag), whether that tag was correct or not, and, where the assigned tag was incorrect, the correct or expected tag. We used that feature to ensure that we were not penalizing taggers whose interpretation was correct. Figure 1 shows a Scoring Report for one specific sentence.

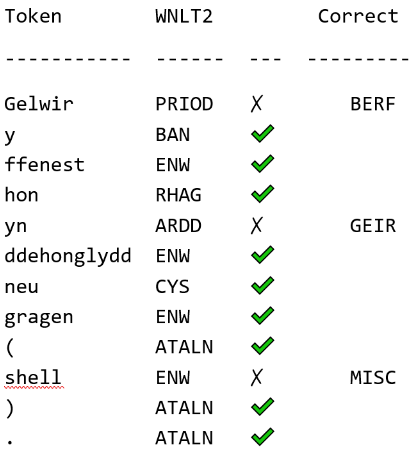
8. Results

8.1. Accuracy of Tokens

The 500 sentences contained a total of 7,675 tokens. We believe that this total is sufficient to prevent the percentages we report being unduly affected by any minor evaluation errors. Table 5 provides an overview of the main results of the evaluation, displaying the number of tokens correctly tagged by each tagger and the percentage of the total that that number represents.

WNLT2	CyTag	TagTeg	Gold
[(‘Ynddi’,‘ARDD’), (‘mae’,‘BERF’), (‘20’,‘RHIF’), (‘o’,‘ARDD’), (‘ganeuon’,‘ENW’)]	[(‘Ynddi’,‘ARDD’), (‘mae’,‘BERF’), (‘20’,‘RHIF’), (‘o’,‘ARDD’), (‘ganeuon’,‘ENW’)]	[(‘Ynddi’,‘ARDD’), (‘mae’,‘BERF’), (‘20’,‘RHIF’), (‘o’,‘ARDD’), (‘ganeuon’,‘ENW’)]	[(‘Ynddi’,‘ARDD’), (‘mae’,‘BERF’), (‘20’,‘RHIF’), (‘o’,‘ARDD’), (‘ganeuon’,‘ENW’)]

Table 4: Common structure for evaluation (sentence literal translation: ‘In it it is 20 of songs’).



```
Token        WNLT2         Correct
-----------  ------  ---  ---------
Gelwir       PRIOD   ✗       BERF
y            BAN     ✓
ffenest      ENW     ✓
hon          RHAG    ✓
yn           ARDD    ✗       GEIR
ddehonglydd  ENW     ✓
neu          CYS     ✓
gragen       ENW     ✓
(            ATALN   ✓
shell        ENW     ✗       MISC
)            ATALN   ✓
.            ATALN   ✓
```

Figure 1: Scoring Report.

Tagger	Number of Correct Tags	Token Accuracy (%)
WNLT2	5992/7675	78%
CyTag	6304/7675	82%
TagTeg	7029/7675	92%

Table 5: Main results of evaluation.

By running the evaluation twice, we calculated that CyTag’s score of 82% would be 84% if we were to allow multiple tags. We feel that disallowing multiple tags is appropriate as neither WNLT2 nor TagTeg offer ambiguous results. However, as noted, we include the more generous figure here so that the reader can come to their own conclusions. TagTeg has benefited somewhat because the predicative ‘yn’ and the preverbal ‘yn’ have both been treated as particles within this evaluation, rather than being divided into two distinct categories. On the other hand, WNLT2 and CyTag have benefited from us allowing a verbnoun (such as ‘canu/to sing’) to be tagged as either a noun OR a verb. TagTeg however attempts to distinguish between the two uses and is penalised when it gets this wrong.

8.2. Sentence Level Accuracy

Manning (2011) questions measuring accuracy at the Token level when taggers routinely score in the high 90s, and suggests using sentence accuracy as an alternative benchmark. In table 6 we therefore provide an overview of sentence accuracy for each tagger. As Manning suggests, the results for the sentences give a better impression of the ability of taggers to correctly tag entire sentences or texts. This is important if the ultimate goal is for computers to correctly understand the information contained in textual data.

Tagger	Number of Sentences 100% Accurate	Sentence Accuracy (%)
WNLT2	41/500	8%
CyTag	48/500	10%
TagTeg	168/500	34%

Table 6: Sentence accuracy.

8.3. Analysis of Results

These results show that TagTeg is significantly more accurate than CyTag and WNLT2, the two rules-based taggers. This is despite the fact that TagTeg is currently trained on a relatively small collection of complete sentences. Although some of the differences between those scores are due to problems specific to CyTag and WNLT2, we believe this generally shows, contrary to Neale et al.’s suggestion (Neale et al., 2018), that statistical methods can be effective with a relatively small amount of data.

The difference between the method used to train TagTeg and that used, for example, by Lamb and Danso (2014), was that, as noted in section 2.5, the training sentences were ‘reinforced’ with one word tagged ‘sentences’ in the form of 76,000 inflected wordforms. The success of this method is welcome news for less resourced languages (which often have dictionary style resources that can be adapted to be used in a similar way). It suggests that collecting and tagging training sentences is an easier and less specialized task than the formulation of grammatical rules, especially when those rules begin to increase in complexity and start to conflict with other rules.

Despite these relatively high token-level scores, at the sentence level the results are significantly poorer. Table 5, where the highest score is 34%, shows that there is still much work to be done to improve taggers to a point where they can be considered completely reliable.

8.4. General Findings

Analyzing the data from a general perspective, we summarize our overall findings on the performance of the three taggers.

8.4.1. The Importance of recognizing English

English words occur frequently within contemporary Welsh texts, whether in the names of companies or organizations, in quotations, or when code switching occurs between Welsh and English. As a result, a useful modern tagger should be able to cope with English words. CyTag is able to identify English words if those are found within its lexicon of English words. WNLT2 lacks this ability completely. TagTeg is able to specify English forms as X (the UD tag for foreign words, among other things) but its ability to specifically label these forms as English could be further improved, as will be discussed in section 8.7.

8.4.2. Informal Welsh

Informal and dialectal Welsh is common on social media, as are misspellings. It is therefore important that a Welsh tagger can cope with the variety of non-standard language contained within such discourse. CyTag can correctly tag many of the most commonly used 'standard' forms of informal Welsh words, but informal vocabulary seems to be a problem for WNLT2. TagTeg now has a normalization component to deal with less standard language, but would also benefit from the inclusion of more spoken sentences in the training data so that there would be no need to include a normalization component in the pipeline to deal with informal forms appropriately.

8.4.3. Destructive Tokenization

One issue that can affect taggers is that of 'destructive tokenization'. This refers to the loss of information detailing the location of spaces and tabs etc. in the tagged output, which can make reproducing the original texts impossible if discarded or lost. Whilst WNLT and TagTeg keep a note of where the spaces were found within a sentence so that the original raw sentences can be reproduced after tokenization, this is not true of CyTag. This is also a problem with the version of the tagged CorCenCC corpus that was shared with us, and may be a significant problem for the future if plain text copies of the original corpus data were not retained.

8.5. Discussion of WNLT2 Results

Thanks to its use of the Eurfa lexicon and of rules to tag unknown words, WNLT2 can provide a tag for most words found in a text. However, its inability to disambiguate wordforms which may correspond to multiple POS tags is problematic. This means that it cannot attribute the correct tag to words when they occur in their alternative function, and users are not alerted to this when using the program. For example, it can assign only one tag to ambiguous words such as 'y' (English:the/that/which), 'yn' (English:is/in), 'i' (English:for/to/me) and 'a' (English:and/that/which). These wordforms make up circa 15% of the words in Welsh texts. As this issue affects such a large proportion of Welsh words, it has a significant impact on the accuracy of the tagger. It's worth noting, however, that Cunliffe et al. (2022) recognise the lack of disambiguation as an issue, noting “The current Tagger does not disambiguate such uses but it is possible to address such cases involving post-processing rules [..] or by developing generic rules via corpus training and machine learning.” Moreover, “The WNLT provides the basis of an operational open-source, Part of Speech tagger that can be improved by future iterations.” Thus, this open source tool is a starting point, ripe for further development.

8.6. Discussion of CyTag Results

As CyTag, like WNLT2, uses the Eurfa lexicon, it succeeds in tagging most of the Welsh words it encounters, but is less effective at identifying unknown words, tagging a number of words which would be assigned meaningful tags by WNLT2 and TagTeg, with the unk tag.

Importantly, CyTag is more sophisticated than WNLT2 in its ability to appropriately tag wordforms which may correspond to more than one tag. However, it does not always succeed in disambiguating between multiple possible tags as there are instances where CyTag outputs multiple tags for a token whereas WNLT2 and TagTeg consistently specify a single tag only. Furthermore, some obvious words are simply tagged incorrectly. For instance, it is difficult to understand why the verb 'ceir' (English:to have) is sometimes tagged as a preposition.

CyTag's main weakness is that the tagging rules of the version tested for this paper appear inconsistent in places. The most obvious example of this is that 'yn' and its shortened enclitic form ''n' are treated differently without obvious justification. Some pronouns are classified as both pronouns and determiners, whilst other similar pronouns are treated as pronouns only. These factors mean that the tagger has scored lower than it could. It should also be noted that CyTag was developed stage by stage using their test set. That is, the test set was also used to develop the tagger's grammar rules (Neale, 2022). Thus, the discrepancy can be attributed to the reported figure representing CyTag's performance on the test set rather than its typical performance on completely unseen texts.

8.7. Discussion of TagTeg Results

As a statistical tagger, TagTeg is not dependant on rules and a lexicon, but rather on annotated sentences. One of its main advantages is its ability to generalize from the training data and learn to tag unfamiliar words appropriately based on similar patterns of sentence placement and prefixes and endings. We believe this partly explains why TagTeg's accuracy is at least 10% higher than the other taggers evaluated here.

One of the issues with TagTeg is that it is difficult to identify a pattern to its errors. It will occasionally fail to appropriately tag a wordform that is otherwise routinely tagged correctly. For example, occasionally TagTeg will incorrectly tag proper nouns such as ‘Sioned’ even though it usually tags them correctly. Without many examples of ‘Sioned’ in the training data, it may be that the tagger’s probabilistic model is influenced by the fact that -ed is a common verbal suffix.

We believe the addition of further training sentences including ‘Sioned’ and other proper nouns will improve this situation. Further examples should also solve TagTeg’s issue where it should tag title tokens such as ‘Hybu Cig Cymru’ (English:Meat Promotion Wales) with the POS of the common word (eg VERB for ‘Hybu’ (English:Promotion)) as Universal Dependencies guidelines dictate (Universal Dependencies, 2021b), rather than PROPN.

Another current shortcoming is that it does not consistently identify some common dialectal forms such as ‘chdi’ (English:you) and ‘isho’ (English:want) as there are currently no examples of such forms in the training data. We intend to add additional dialectal sentences to the training data to address this.

Another issue that arises from analysing the TagTeg results is the manner in which it tags English words such as ‘slow’ when found within a Welsh sentence such as ‘mynd yn slow iawn’ (English:going very slowly) with Welsh POS tags, instead of the expected X tag. To err on the side of caution, we have penalized TagTeg here, but its interpretation is arguably correct under certain theoretical approaches, especially those favouring more descriptive analysis over prescriptivism and linguistic purism. Interestingly, however, TagTeg is very good at identifying chains of English words which combine to form a title, such as ‘The Phantom of the Opera’. We believe that TagTeg has the potential to improve its ability to tag individual English words given additional training with appropriate data.

Overall, an accuracy of 92% meant that many of the TagTeg tagged sentences contained few, if any, mistakes. As a result, we believe that TagTeg represents a successful tagger with plenty of scope for improvement. Unlike the case with rule-based taggers, we believe that this improvement can be achieved relatively easily by identifying and annotating additional training sentences that target the current areas of weakness.

8.8. Further Work

As mentioned, this evaluation is not an exhaustive evaluation of all Welsh-Language taggers. In the future, we hope to expand our evaluation to include taggers such as UDPipe (based on Welsh Syntax Corpus data forthcoming by Dr Johannes Heinecke), the Cyslib tagger (historically used in Cysill), and Autoglosser 2, a rule based tagger which may improve on the results given by CyTag or WNLT.

9. Conclusions

In this work we have described three Welsh language POS taggers and introduced our tagger evaluation methodology. In order to be able to compare the performance of the three different Welsh POS taggers, their output was converted to a consistent general format so that they all display the same tag for nouns, verbs, adjectives and so on. Accuracy was scored by comparing the output of the three taggers when used on the same set of 500 sentences with corresponding annotations made by experienced linguists.

The results show a significant difference in accuracy between the TagTeg statistical tagger and the two rules-based taggers, with a 10% difference between TagTeg and the nearest tagger. This difference can be attributed to three factors, the first being the superiority of the statistical method over the rules-based method. Internationally, statistical methods have proven to be dominant over rule-based ones for many years. Cole et al. (1997) noted that statistical methods ‘have been dominant since the early 1980s’. Brants (2000) too notes that statistical approaches ‘yield better results’. More recently, it is telling that all of the taggers listed in ACL’s regularly updated POS Tagging (State of the art) list are all statistical taggers. Some of the benefits of the statistical method include their ability to generalize and assign appropriate POS tags to unfamiliar words based on features such as their sentence placement, capitalization, prefixes and suffixes. This also means that they can better cope with the misspellings, dialectal forms and unfamiliar proper nouns that characterize real-life data. Moreover, it is easier to maintain and develop a statistical tagger than a rule-based tagger as writing and tagging training sentences is easier than trying to write rules that build on one another whilst also ensuring that the rules do not conflict with each other. The second reason is that CyTag has no method for guessing unfamiliar words, so words that aren’t already in the tagger’s vocabulary are tagged as unk. CyTag also tags some frequently occurring words incorrectly or inconsistently. Thirdly, WNLT2 does not attempt to disambiguate wordforms that have a different POS in different contexts.

In addition to achieving better results, the statistical Machine Learning approach also allows statistical taggers other than the one used by TagTeg to be trained on the same data. This ensures that the Welsh language is not tied to one specific piece of software, such as spaCy, in perpetuity. That being said, we believe that spaCy is a good choice to form the basis of a broader NLP framework for the Welsh language, as it is a modern, well-documented, accessible library that is available free of charge under a permissive open license. With this in mind we are investing further in building a modern NLP pipeline. This will include creating additional tools, such as a Welsh language dependency parser and an NER component, so that the current and future technical needs of the Welsh language are answered.

10. Acknowledgements

We are grateful to the Welsh Government for funding this work as part of the Text, Speech and Translation Technologies for the Welsh Language project.

11. Bibliographical References

Association for Computational Linguistics. (2019). Pos tagging (state of the art).

Brants, T. (2000). Tnt - a statistical part-of-speech tagger.

Choi, J. D., Tetreault, J., and Stent, A. (2015). It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 387–396. Association for Computational Linguistics.

Cole, R. A., Chief, I., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V., Varile, G., Zampolli, A., Cole, R., Zue, V., Zue, V., and Cole, R. (1997). Survey of the state of the art in human language technology. In *Studies In Natural Language Processing, XIIXIII.*

Cunliffe, D., Tudhope, D., Vlachidis, A., and Williams, D. (2017). Pecyn Cymorth Iaith Naturiol Cymru Fersiwn 2.2.

Cunliffe, D., Vlachidis, A., Williams, D., and Tudhope, D. (2022). Natural language processing for under-resourced languages: Developing a Welsh natural language toolkit. *Computer Speech Language*, 72:101311.

Hagerman, C. (2012). Evaluating the performance of automated part-of-speech taggers on an l2 corpus.

Jiang, R., Banchs, R. E., and Li, H. (2016). Evaluating and combining name entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop*, pages 21–27. Association for Computational Linguistics.

Jurafsky, D. and Martin, J. H., (2021). *Sequence Labeling for Parts of Speech and Named Entities*. Stanford University, online.

Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., Thomas, E.-M., Lovell, A., Morris, J., Evas, J., Stonelake, M., Arman, L., Davies, J., Ezeani, I., Neale, S., Needs, J., Piao, S., Rees, M., Watkins, G., Williams, L., Muralidaran, V., Tovey-Walsh, B., Anthony, L., Cobb, T., Deuchar, M., Donnelly, K., McCarthy, M., and Scannell, K. (2020). CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh.

Lamb, W. and Danso, S. (2014). Developing an automatic part-of-speech tagger for scottish gaelic. Proceedings of the First Celtic Language Technology Workshop, pages 1–5. Association for Computational Linguistics and Dublin City University.

Leech, G. and Wilson, A. (1996). Recommendations for the morphosyntactic annotation of corpora. Report.

Manning, C. D. (2011). Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? Computational Linguistics and Intelligent Text Processing, pages 171–189. Springer Berlin Heidelberg.

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)*. European Language Resources Association (ELRA).

Neale, S. (2022). Private Communication.

Nguyen, D. Q., Nguyen, D. Q., Pham, D. D., and Pham, S. B. (2016). A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. *AI Communications*, 29(3):409–422, apr.

Paroubek, P., (2007). *Evaluating Part-of-Speech Tagging and Parsing*, pages 99–124. Springer Netherlands, Dordrecht.

Partalidou, E., Spyromitros-Xioufis, E., Doropoulos, S., Vologiannidis, S., and Diamantaras, K. (2019). Design and implementation of an open source Greek pos tagger and entity recognizer using spacy.

Sadredini, E., Guo, D., Bo, C., Rahimi, R., Skadron, K., and Wang, H. (2018). A scalable solution for rule-based part-of-speech tagging on novel hardware accelerators. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '18, page 665–674, New York, NY, USA. Association for Computing Machinery.

Schmitt, X., Kubler, S., Robert, J., Papadakis, M., and Le Traon, Y. (2019). A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate, 2019-10-22. ISET Engineering Sciences [physics]/AutomaticConference papers.

spaCy. (2022). Facts & figures.

Universal Dependencies. (2021a). Introduction.

Universal Dependencies. (2021b). Propn: proper noun.

Williams, D. (2017). Welsh Natural Language Toolkit.

12. Language Resource References

Cunliffe, D., Vlachidis, A., Williams, D., and Tudhope, D. (2022). Natural language processing for under-resourced languages: Developing a Welsh natural language toolkit. *Computer Speech Language*, 72:101311.

Deuchar, M., Carter, D., Davies, P., Donnelly, K., Herring, J., del, M., Stammers, J., Aveledo, F., Fusser, M., Jones, L., Lloyd-Williams, S., Prys, M., and Robert, E. (2009). Siarad corpus.

Donnelly, K. (2013). Eurfa.

Donnelly, K. (2018). Autoglosser2.

Ellis, N. C., O'Dochartaigh, C., Hicks, W., Morgan, M., and Laporte, N. (2001). Cronfa Electroneg o Gymraeg (CEG).

Gorrell, G., Maynard, D., and Roberts, A. (2010). Module 2: Introduction to IE and ANNIE.

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*, pages 278–277.

Hicks, W. J. (2004). Welsh proofing tools: Making a little nlp go a long way. In *the 1st Workshop on International Proofing Tools and Language Technologies*.

Jones, D. B., Robertson, P., and Prys, G. (2015). Gwasanaeth API Tagiwr Rhannau Ymadrodd Cymraeg.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 3*, COLING '90, page 168–173, USA. Association for Computational Linguistics.

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)*. European Language Resources Association (ELRA).

Prys, D., Prys, G., and Watkins, G. L. (2020). Model Tagio Rhannau Ymadrodd Cymraeg/Welsh Language Part of Speech Tagging Model.

Prys, D., Jones, D. B., Prys, G., and Watkins, G. L. (2021). Lecsicon Cymraeg Bangor Welsh Lexicon.