Development of a Universal Dependencies treebank for Welsh

Johannes Heinecke
Orange Labs
2 rue Pierre Marzin
F - 22307 Lannion cedex
johannes.heinecke@orange.com

Francis M. Tyers
Department of Linguistics
Indiana University
Bloomington, IN
ftyers@iu.edu

Abstract

This paper describes the development of the first syntactically-annotated corpus of Welsh within the Universal Dependencies (UD) project. We explain how the corpus was prepared, and some Welsh-specific constructions that require attention. The treebank currently contains 10 756 tokens. An 10-fold cross evaluation shows that results of both, tagging and dependency parsing, are similar to other treebanks of comparable size, notably the other Celtic language treebanks within the UD project.

1 Introduction

The Welsh Treebank is the third Celtic language within the Universal Dependencies project (Nivre et al., 2016), after Irish (Lynn and Foster, 2016) and Breton (Tyers and Ravishankar, 2018). The main goal of the Universal Dependencies treebanks is to have many different languages annotated with identical guidelines and universally defined set of universal POS tags and dependency relations. These cross-linguistically consistent grammatical annotations can be used, for instance, for typological language comparison or developping and evaluating cross-linguistic tagging and parsing tools.

The motivation was twofold: To have a Welsh treebank annotated using the same guidelines as many existing treebanks which permits language comparison and to have a (start for a) treebank which can be used to train dependency parsers. Since the UD project already contains 146 treebanks for 83 languages and provides annotation principles which have been used in typological very different languages, we chose to develop the Welsh treebank within the UD project. At the time of writing 601 sentences with 10 756 tokens in total have been annotated and released with UD version 2.4.

The paper is laid out as follows: in Section 2 we give a short typological overview of Welsh. Section 3 describes briefly prior work for Welsh in computational linguistics and syntax as well as existing resources. In sections 4 and 5 we describe the annotated corpus, the preprocessing steps and present some particularities of Welsh and how we annotate them. Section 6 explains the validation process. We conclude with a short evaluation in section 7 and some remarks on future work (section 8).

2 Welsh

Welsh is a Celtic language of the Brythonic branch[1] of the Insular Celtic languages. There are about 500 000 (Jones, 2012) native speakers in Wales (United Kingdom). Apart from very young children, all speakers are bilingual with English. There are also a few thousand Welsh speakers in the province of Chubut, in Argentina, who are the descendants of Welsh emigrants of the 1850s, who are now all bilingual with Spanish. A short overview on the Welsh Grammar is given in Thomas (1992) and Thorne (1992), more detailed information can be found in Thorne (1993) and King (2003).

Even though Welsh is a close cognate to Breton (and Cornish), it is different from a typological point of view. Like Breton (and the Celtic languages of the Goidelic branch), it has initial conso-

[1] Together with Breton and Cornish

nant mutations, inflected prepositions (*ar* "on", *arnaf i* "on me"), genitive constructions with a single determiner (*tŷ'r frenhines*, lit. "house the queen": "the house of the queen") and impersonal verb forms. However, unlike Breton, Welsh has a predominantly verb-subject-object (VSO) word order, does not have composed tenses with an auxiliary corresponding to "have" and uses extensively periphrastic verbal clauses to convey tense and aspect (Heinecke, 1999). It has only verbnouns instead of infinitives (direct objects become genitive modifiers or possessives). Like Irish but unlike Breton Welsh does not have a verb "to have" to convery possession. It uses a preposition "with" instead: *Mae dau fachgen gen i* "I have two sons", lit. "There is two son(SG) with me". Another feature of Welsh is the vigesimal number system (at least in the formal registers of the language) and non-contiguous numerals (*tri phlentyn ar hugain* "23 children", lit. "three child(SG) on twenty").

3 Related Work

Welsh has been the object of research in computational linguistics, notably for speech recognition and speech synthesis (Williams, 1999; Williams et al., 2006; Williams and Jones, 2008), as well as spell checking and machine translation. An overview can be found in Heinecke (2005), more detailed information about existing language technology for Welsh is accessible at `http://techiaith.cymru`. Research on Welsh syntax within various frameworks is very rich: Awbery (1976), Rouveret (1990), Sadler (1998), Sadler (1999), Roberts (2004), Borsley et al. (2007), Tallerman (2009), Borsley (2010) to cite a few.

The only available annotated corpus to our knowledge is *Cronfa Electroneg o Gymraeg* (CEG) (Ellis et al., 2001), which contains about one million tokens, annotated with POS and lemmas. The CEG corpus contains texts from novels and short stories, religious texts and non-fictional texts in the fields of education, science, business or leisure activities. It also contains texts from newspapers and magazines and some transcribed spoken Welsh.

Currently work is under way for the National Corpus of Contemporary Welsh[2]. It is a very large corpus which contains spoken and written Welsh. currently the existing data is not annotated in syntax (dependency or other). Members of the CorCenCC project also work on WordNet Cymraeg, a Welsh version of WordNet. Further corpora (including CEG) are available at University of Bangor's Canolfan Bedwyr [3].

Other important resources are the proceedings of the Third Welsh Assembly[4] and *Eurfa*, a full form dictionary[5] with about 10 000 lemmas (210 578 forms). There is also the full form list for Welsh of the Unimorph project[6]. Currently, however, this list contains only 183 lemmas (10 641 forms).

The Welsh treebank is comparable in size to the Breton treebank (10 348 tokens, 888 sentences). The Irish treebank is twice as big with 23 964 tokens (1 020 sentences). UD treebanks vary very much in size. Currently the largest UD treebank is the German-HDT with 3 055 010 tokens. The smallest is Tagalog with just 292 words. Average size for all treebanks is 150 827 tokens, median size is 43 754 tokens.

4 Corpus

Like every language, Welsh has formal and informal registers. All of those are written, which makes it difficult to constitute a homogeneous corpus. The differences are not only a question of style, but are also of morphological and sometimes of syntactic nature. Usually for the written language distinction is made between Literary Welsh (cf. grammars by Williams (1980) and Thomas (1996)) and Colloquial Welsh (Uned Iaith Genedlaethol (1978)) including an attempted new standard, *Cymraeg Byw* "Living Welsh" (Education Department, 1964; Davies, 1988)). Cymraeg Byw, however, has fallen out of fashion since. For the UD Welsh treebank, we chose sentences of Colloquial written Welsh.

The sentences of the initial version of the treebank have been taken from varying sources, like the Welsh language Wikipedia, mainly from pages on items about Wales, like on the *Urdd Gobaith Cymru*, the *Eisteddfodau* or Welsh places, since it is much more probable that native Welsh speakers contributed to these pages. Other sources for individual sentences were the Welsh Assembly corpus mentioned above, Welsh Grammars (in order to cover syntactic structures less frequently seen) or

[2] `http://www.corcencc.org/`

[3] cf. Canolfan Bedwyr, University of Bangor, `https://www.bangor.ac.uk/canolfanbedwyr/ymchwil_TI.php.en` and `http://corpws.cymru`

[4] `http://cymraeg.org.uk/kynulliad3/`

[5] `http://eurfa.org.uk/`

[6] `http://www.unimorph.org/`

web sites from Welsh institutions (Welsh Universities, *Cymdeithas yr Iaith*). Finally some sentences origin from Welsh language media (*Y Golwg*, *BBC Cymru*) and blogs. Even though a few of the sentences from Wikipedia may look awkward or incorrect to native speakers, these sentences are the reality of written Welsh and are therefore included in the treebank.

The different registers of Welsh mean, that theoretically "identical" forms may appear in diverging surface representations: so the very formal *yr ydwyf i* "I am" (lit. "(affirmative) am I") can take the following (more or less contracted) forms in written Welsh: *rydwyf (i)*, *rydwi*, *rydw (i)*, *dwi*. In the treebank, we annotate these forms as multi-token words. For layout reasons, we do not split these forms in all examples in this paper. Where it is the case, we mark multi-token words with a box around the corresponding words. The same applies for the negation particle *ni(d)* which is often contracted with the following form of *bod*, if the latter has an initial vowel: *nid oedd* "(he) was not" > *doedd*. Sometimes dialectal variants appear in the Written language: *oeddan* vs *oedden* "(we) were". The corpus of the Welsh treebank retains the original forms. However, we use standard lemmas (Thomas and Bevan, 1950 2002).

4.1 Preprocessing

In order to initiate and to speed up the annotation process, we transformed the CEG corpus (forms, lemmas and POS) into UD's CoNLL-U format (cf. figure 1) and replaced CEG's part-of-speech tags into UD UPOS. During this step we also corrected annotation errors (notably non-ambiguous cases) and added information about which consonant mutation is present, if any. We then used the *Eurfa* full-form dictionary to enrich the CoNLL-U format with morpho-syntactic features. On this UD compatible Welsh corpus, we trained the UDpipe tagger and lemmatizer (Straka and Straková, 2017) using word embeddings for Welsh trained on the Welsh Wikipedia with FastText and provided by Bojanowski et al. (2017). With this model we POS-tagged our corpus. A second script pre-annotated some basic dependency relations (`case`- and `det` relations).

In addition to the UD standard, we defined language specific XPOS (table 1), a morphological feature `Mutation` with three values to indicate the consonant mutation, since they carry syntactically relevant information, and some subtypes for dependency relations, also frequently used in other languages: `acl:relcl` (relative clauses) and `flat:name` (flat structures for multi-word named entities).

UPOS	*Welsh specific XPOS*
ADJ	pos, cmp, eq, ord, sup
ADP	prep, cprep
AUX	aux, impf, ante, post, verbnoun
NOUN	noun, verbnoun
PRON	contr, dep, indp, intr, pron, refl, rel
PROPN	org, person, place, propn, work

Table 1: Welsh specific XPOS

5 Dependencies

The POS-tagged corpus was then manually annotated[7] and all layers were validated: lemmas, UPOS, XPOS (see section 5), and dependency relations using the annotation guidelines of Universal Dependencies. The annotation were made by a single annotator.

The following subsections discuss some of the particularities of the Welsh language, and how these were annotated.

5.1 Nominal genitive construction

Similar to the other Celtic languages, but also to genetically very different languages like Arabic, nominal genitive constructions are juxtaposed nounphrases. Only the last nounphrase can have a determiner (article, possessive), which determines the whole construction (fig. 2).

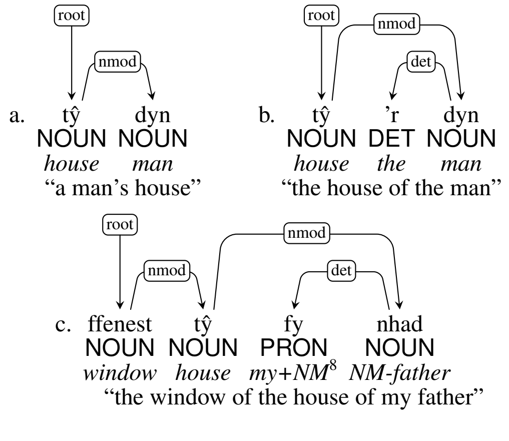


Figure 2: noun phrases

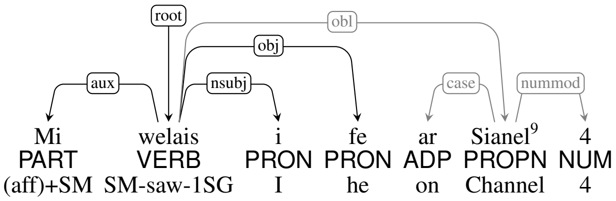
[7] using https://github.com/Orange-OpenSource/conllueditor/

```
#id form   lemma  UPOS XPOS features                  head deprel enh.deps misc
1   tai    tŷ     NOUN noun Gender=Masc|Number=Plur 0    root   _        SpaceAfter=No
2   'r     y      DET  art  _                       3    det    _        _
3   brenin brenin NOUN noun Gender=Masc|Number=Sing 1    nmod   _        _
```

Figure 1: CoNLL-U format: Every token is a line of 10 tab-separated columns UPOS are universal POS tags, XPOS are language specific. The enhanced dependencies column adds a second layer of annotation (not used yet in Welsh). The misc column provides information about inter-token spaces, glosses, transcription etc. For details, see `https://universaldependencies.org/format.html`

5.2 Periphrastic verbal construction

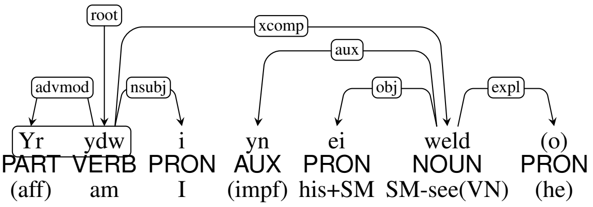
The verb can be seen as the central word in dependency syntax. Since the Welsh verb has only two tenses in the indicative mood (Future and Past, both denoting perfective aspect), all other tense and aspect forms are built using periphrastic constructions using one or more forms of the verbnoun *bod* "to be". Whereas inflected verbs (fig. 3) are annotated in a straight forward way, the periphrastic forms need some attention.



"I saw him on Channel 4"

Figure 3: Inflected verb

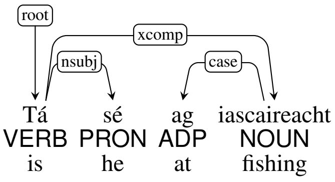
As said before, Welsh has no infinitives, but verbnouns, which mark objects differently compared to inflected verbs. Whereas in English a direct object is the same (*I saw **him***; *to see **him***), in Welsh the inflected form uses a different pronoun series (called independent pronouns in Welsh grammar tradition) than the verbnoun. Note that Welsh does not distinguish between subject and object pronouns, but between independent and dependent pronouns. The former are used in subject and object position of inflected verbs, the latter for possessives and "object" relations on verbnouns, e.g. *ei gar* "his car" or mark the direct object for verbnouns *ei gweld* literally "his seeing", e.g. "to see him". For this reason, verbnouns have a different language specific XPOS (verbnoun vs. noun) but the same UPOS (NOUN) as nouns. Other treebanks in the UD project with verbnouns do the same (notably Irish and, in some well defined, cases Polish). The periphrastic construction (fig. 4) employs at least one (inflected) form of *bod* (here *ydw*) and a time or aspect marker (TAM) like *yn* or *wedi* etc. The (independent) pronoun after the verbnoun (VN) is facultative and repeats the (dependent) pronoun before the verbnoun *gweld* (here undergone soft mutated to *weld*).



"I am seeing him"
(lit. "I am (impf) his seeing he")

Figure 4: Periphrastic construction

The same annotation can be found in the Irish treebank (fig. 5). In the Breton treebank, however, the infinitive is the phrasal head (6) to which the auxiliary verb is attached as an `aux`.

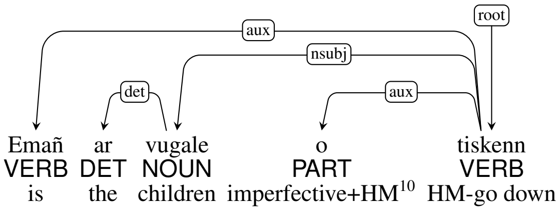


"He is fishing"

Figure 5: Irish periphrastic construction (from UD Irish-IDT:948

[8]+NM means that this word triggers soft mutation on the following word, NM- means that this word undergoes nasal mutation. Similarly we use SM and AM for soft and aspirated mutation, respectively. For more details on mutations, see King (2003, pp. 14ff). Some mutations are triggered by syntactic functions and not by a preceding word, e.g. temporal and spatial adverbials or undefinite direct objects.

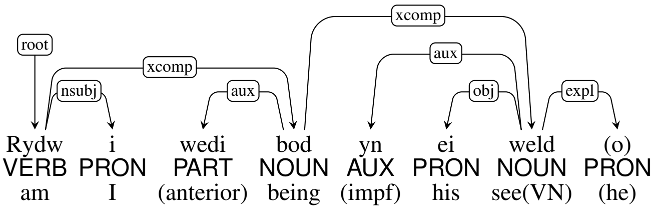
[9]Dependency relations in gray are irrelevant for the point made in the example.



"The children are going down."

Figure 6: Breton periphrastic construction (from UD Breton-KEB:grammar.vislcg.txt:54:1065

Periphrastic constructions can be nested, so to have the imperfective version of figure 4 we get the Sentence shown in figure 7.



"I have been seeing him"
(lit. "I am after being his seeing he")

Figure 7: Nested periphrastic construction

If a periphrastic phrase is used as a subordinate, even the head (*bod*) is a verbnoun and the subject is marked using the dependent (possessive) pronoun (fig. 8, next page).

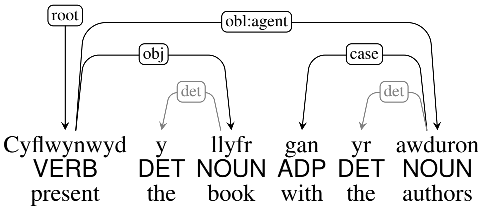
In Welsh the main (semantic) time (past, present, future) is nearly always expressed as a form of *bod*. Relative time positions (before or after) are marked using other TAM markers (*ar*, *am* (posteriority, "about to"), *wedi* (anteriority, cf. English Present/Past Perfect), *hen* (distant anteriority), *newydd* (recent anteriority, cf. Heinecke (1999, p. 271)). We have decided to use the inflected form of *bod* as the syntactic head and link the subsequent verbnoun *bod* and finally the verbnoun carrying the meaning with as `xcomp` to avoid completely flat trees which do not show the inherent structure of these phrases.

5.3 Subjects in subordinate phrases

Subordinate phrases very often do not have inflected verbs, but use TAM or prepositions to establish a relative time with respect to the main phrase. In these cases the subject has a `case` dependant (preposition *i* "to", cf. fig. 9 next page, where the preposition is inflected and appears as *iddo*).

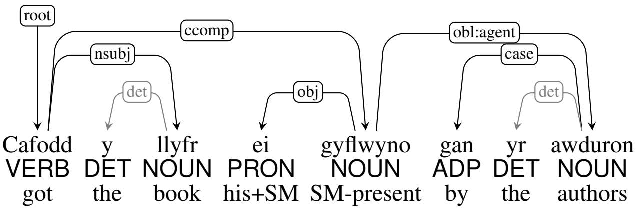
5.4 Impersonal and *cael*-periphrastics

Like the other Celtic languages Welsh has impersonal forms (which are often translated using passives). In this construction the demoted agent can be expressed using the preposition *gan* "with". As in the Irish and Breton treebanks, the core argument of an Impersonal is marked `obj` (fig. 10). A periphrastic construction is possible using the verb *cael* "get" (fig. 11).



"The book was presented by the authors"
(lit. "One presented the book by the authors")

Figure 10: Impersonal verb form



"The book was presented by the authors" (lit. "got the book his presenting by the authors")

Figure 11: Periphrastic construction with*cael*

5.5 Nonverbal predicates

Like in most other languages, adjectives and nouns can be head, if they are the predicate. In Welsh, however, such adjectives and nouns need a special predication marker *yn*[11] (fig. 12 and 13).

[10] HM: Breton hard mutation

[11] There are three forms *yn* in Welsh with tree different functions: 1) predicative marker preceding nominal and adjectival predicates, 2) imperfective marker (Isaac, 1994), which precedes a verbnoun (cf. fig. 4), and 3) preposition "in" which triggers nasal mutation). The first two are shortened to *'n* if the preceding word terminates with a vowel.

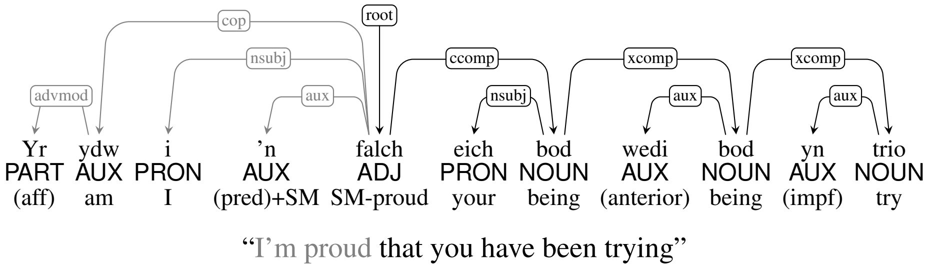

root
cop
nsubj
aux
advmod
ccomp
nsubj
xcomp
aux
xcomp
aux
Yr ydw i 'n falch eich bod wedi bod yn trio
PART AUX PRON AUX ADJ PRON NOUN AUX NOUN AUX NOUN
(aff) am I (pred)+SM SM-proud your being (anterior) being (impf) try
"I'm proud that you have been trying"

Figure 8: Subordinate phrase

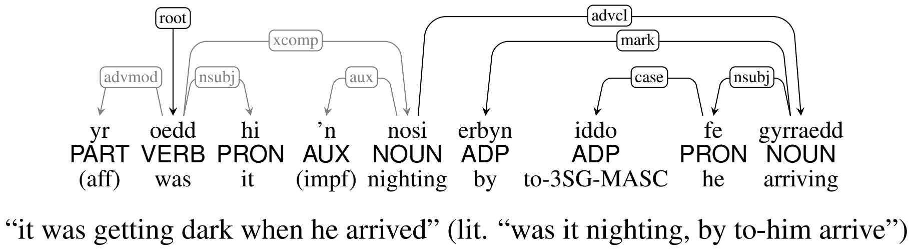

root
advcl
mark
xcomp
advmod
nsubj
aux
case
nsubj
yr oedd hi 'n nosi erbyn iddo fe gyrraedd
PART VERB PRON AUX NOUN ADP ADP PRON NOUN
(aff) was it (impf) nighting by to-3SG-MASC he arriving
"it was getting dark when he arrived" (lit. "was it nighting, by to-him arrive")

Figure 9: subject in a subordinate phrase

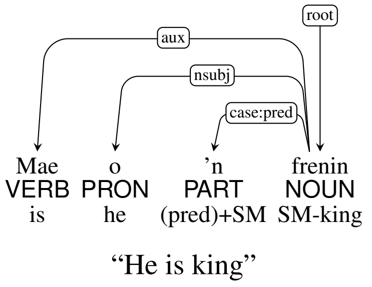


Figure 12: Nonverbal predicates (noun)

Since the predicative *yn* is not a preposition[12] but in the same position as a preposition we decided to use the relation `case:pred`.

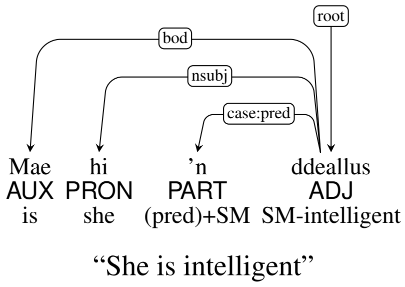


Figure 13: Nonverbal predicates (adjective)

The predication marker *yn* + adjective is also used to have adverbs on verbnouns (fig. 14). In subordinates, the copula *bod* becomes a verbnoun, the subject is attached as possessive using a dependent pronoun (fig. 15, next page).

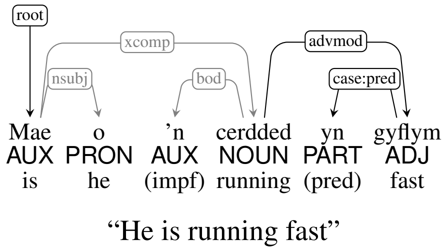

root
xcomp
nsubj
bod
advmod
case:pred
Mae o 'n cerdded yn gyflym
AUX PRON AUX NOUN PART ADJ
is he (impf) running (pred) fast
"He is running fast"

Figure 14: adverbial

5.6 Inflected prepositions

All Celtic languages have contracted forms of prepositions and following pronouns. In Welsh, the pronoun can follow the contracted preposition, so it is more adequate to speak of inflected prepositions instead (Morris-Jones (1913, p. 397), King (2003, p. 268)). This requires a different annotation, since the inflected prepositions (like *gennyt* "with you" in fig. 16) incorporates the `obl` argument. The pronoun "you" is dropped.

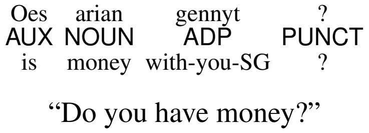
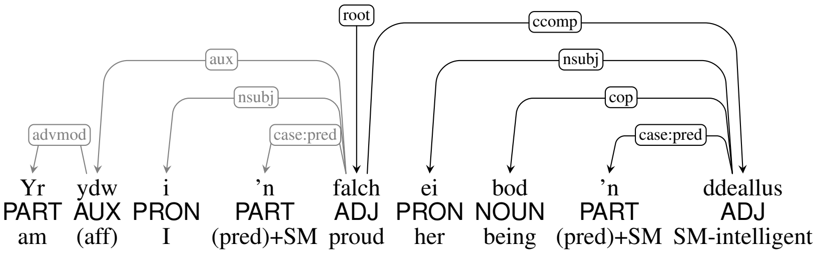


Figure 16: Inflected prepositions (pronoun dropped)

In fig. 17, where the pronoun *ti* is present (`obl`),

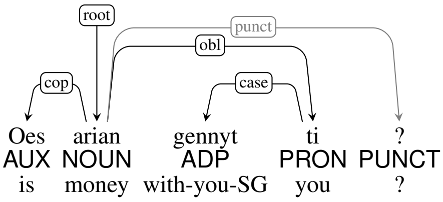
[12] In Welsh dictionaries the predicative *yn* is tagged as an adverb.


root
ccomp
aux
nsubj
nsubj
cop
advmod
case:pred
case:pred
Yr ydw i ’n falch ei bod ’n ddeallus
PART AUX PRON PART ADJ PRON NOUN PART ADJ
am (aff) I (pred)+SM proud her being (pred)+SM SM-intelligent

“I’m proud that she is intelligent”

Figure 15: subordinate nonverbal predicate

gennyt is attached as a simple `case` to the pronoun.



“Do you have money?”

Figure 17: Inflected prepositions (pronoun present)

Using empty nodes and enhanced dependencies, the annotation of an inflected preposition without pronoun becomes more similar to the case with pronoun (fig. 18). The current version of the Welsh treebank is not yet annotated this way.

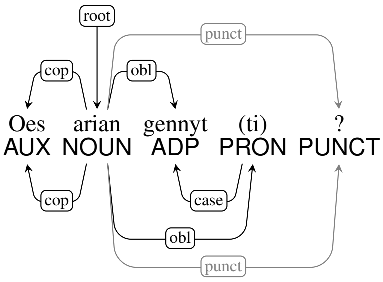
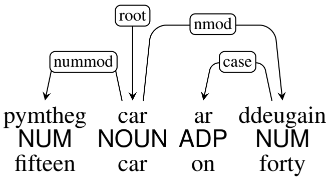


Figure 18: Inflected prepositions with empty words and enhanced dependencies

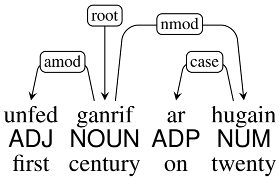
5.7 Compound numbers

The traditional Welsh numbering is based on a vigesimal number system (20 = *ugain*, 30 = *deg ar hugain* “ten on twenty”, 40 = *deugain* “two twenties”, 60 = *trigain* “three twenties”). Notably for compound numbers, the counted item comes between the units and the tens of the number, in both cardinals (fig. 19, nouns are always in singular after a numeral) and ordinals (fig. 20).



“55 cars” (lit. “15 cars on 2*20”)

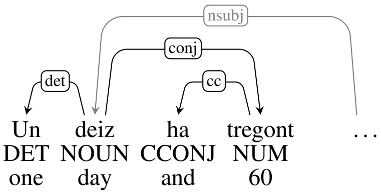
Figure 19: Compound numbers (cardinals)



“twenty first century”
(lit. “first century on twenty”)

Figure 20: Compound numbers (ordinals)

Breton and Irish use(d) a similar system. There is one example in the Breton UD treebank (fig. 21), which is annotated in a similar way apart from the fact, that Breton uses coordination to join numbers instead of a preposition, as Welsh does). Irish dialects traditionally have a similar structure, even though none is attested in the UD treebank: *dhá bhád is ceithre fichid* “82 boats” (lit. “two boat(sg) and four twenty”).



“61 days” (lit. “one day and sixty”)

Figure 21: Compound numbers in Breton (from UD Breton-KEB, wikipedia.vislcg.txt:112:3736)

6 Validation

After having all sentences annotated, a post-validation script checked some semantic aspects like the XPOS of inflected prepositions, TAM markers, preverbal particles and consonant mutations (adding the corresponding feature) and looked for potential errors (e.g. a `det` with a VERB head). This script also checked all forms of the verb *bod* and completed morphological features. For nouns and adjectives the script gives an alert if it cannot determine number (on the base of regular suffixes etc.). The final step is the validation script provided by the UD project which finds formal errors (e.g. dependants on words with a `case` or `aux` relation).

Currently the Welsh treebank contains 601 sentences with 10 756 tokens (including punctuation). The average sentence length is 17,9 tokens (shortest sentence: 4 tokens, longest sentence: 59 tokens, median length: 16 tokens). Since verbnouns have the UPOS NOUN, 30.1% of all UPOS are NOUN (table 2).

UPOS	%	*XPOS*	%
NOUN	30.1	noun	21.3
ADP	12.9	prep	12.4
PUNCT	9.7	punct	9.7
ADJ	6.9	verbnoun	9.1
DET	6.5	art	6.5
PRON	6.3	verb	6.3
VERB	6.3	adj	6.0
AUX	5.9	cconj	2.9
PART	4.4	dep	2.7
PROPN	3.7	impf	2.5
CCONJ	2.9	indep	2.3
ADV	1.9	pred	2.1
NUM	1.3	aux	1.9
		adv	1.9

Table 2: relative frequency of (some) UPOS, XPOS

The relatively small number of tokens with UPOS VERB is due to the fact that verbnouns have the UPOS NOUN. This is relativized if we regard the distribution of the XPOS: noun 21.3%, verbnoun 9.1% + verb 6.3% = 15.4% "verbs".

Table 3 shows all 34 dependency relations used in the Welsh treebank including their frequency. 6 out of the 37 dependency relations proposed[13] by the UD guidelines are not used. These are `clf` used for classifiers (absent in Welsh), `orphan` (used to annotate ellipses), `discourse` (interjections) `goeswith` and `reparandum` used to correct errors in spelling or tokenization (currently all sentences in the treebank are correctly tokenised) and `dep`, the default label, if no more specific relation can be chosen).

deprel	%	*deprel*	%
case	10.5	ccomp	1.7
punct	9.7	nmod:poss	1.7
nmod	8.4	acl	1.6
det	6.9	advcl	1.4
obl	6.0	acl:relcl	1.2
nsubj	5.7	flat:name	0.8
root	5.6	flat	0.6
obj	5.2	nummod	0.6
advmod	5.2	fixed	0.5
amod	4.8	appos	0.5
xcomp	4.3	expl	0.2
aux	3.7	obl:agent	0.2
conj	3.2	parataxis	0.2
cc	2.9	csubj	0.1
case:pred	2.1	nmod:agent	0.1
mark	2.1	compound	< 0.1
cop	2.0	iobj	< 0.1

Table 3: relative frequency of all 34 used deprels

7 Evaluation

Even though the treebank currently contains only 601 sentences, tests for tagging and dependency parsing (table 6) show results comparable with similar sized treebanks (Tyers and Ravishankar (2018) report a LAS between 64.14% and 74.29% for the Breton treebank, Zeman et al. (2018) mention a LAS of 70.88 for Irish). We used Udpipe (v2, single model for tagging and parsing). Tests with Wikipedia embeddings (500 dimensions) trained with fastText (Bojanowski et al., 2017) did not improve the parsing. This might be caused by the relatively small corpus on which the embeddings have been trained (the Welsh Wikipedia (April 2019) contains only 62MB of compressed raw data (104 000 pages).

For the evaluation we split the 601 sentences into training (80%), dev (10%) and test (10%) corpora and performed a 10-fold cross evaluation. We used the official CoNLL-2018 evaluation script[14]

[13] `https://universaldependencies.org/u/dep/all.html`

[14] `https://universaldependencies.org/`

to calculate all scores. Table 4 shows the results of POS tagging and lemmatisation without and with the Eurfa dictionary, and table 5 the results per UPOS.

	UPOS	XPOS	Lemma
baseline	89.2	87.3	86.7
+*Eurfa*	87.9	87.5	93.5

Table 4: results of POS tagging and lemmatisation (F-measure

Nearly half of the word forms in the test corpus are out-of-vocabulary (OOV) with respect to the training corpus. The dictionary provided roughly half of the missing words. Thus a quarter of the words in the test corpus remains OOV, which may explain the unexpected low performance (UDpipe switches off its guesser, if a dictionary is provided).

The results of dependency analysis are presented in table 6 using 3 of the standard metrics for dependency parsing (Nivre and Fang, 2017), Labelled Attachment Score (LAS, evaluates heads and dependency labels) or Content Word LAS (CLAS, as LAS, but only for dependency relations of content words (excluding `aux`, `cop`, `mark`, `det`, `clf`, `case`, `cc`).

We run four tests, a model trained solely on the treebank, with dependencies parsed on the results of the tagger, and dependencies parsed using gold tags. The other two tests use the Eurfa dictionary again. The better results of tagging with the full form lexicon, also improves the dependency parsing, if the parsing is done on predicted POS tags. All three metrics increase accordingly.

	POS tags	UAS	LAS	CLAS
baseline	predicted	74.3	63.9	54.8
	gold	82.2	76.2	69.6
+*Eurfa*	predicted	75.5	64.3	55.4
	gold	81.9	75.9	69.3

Table 6: results of dependency parsing

8 Future Work

The most obvious work is to increase the number of sentences annotated. The current 601 sentences may be a start, but do not cover enough examples to train a robust dependency parser. Another important problem is the absence of very formal Welsh (as in the Bible (in its 1588 translation) and some literary works) and of very informal written Welsh (as is used by some Welsh bloggers). Since Welsh is not one of the most widely learned languages, we plan adding glosses and translations to the existing sentences.

With word embeddings becoming more important, work on Welsh word embeddings is needed too. We need to dig into cross-lingual approaches too (e.g. with BERT, (Lample and Conneau, 2019) or UDify (Kondratyuk, 2019)) and/or provide much larger Welsh text corpora than Wikipedia to train word embeddings.

References

Awbery, Gwenllian M. 1976. *The Syntax of Welsh. A transformational Study of The Passive*. Cambridge University Press, Cambridge.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Borsley, Robert D., Taggie Tallerman, and David Willis. 2007. *The Syntax of Welsh*. Cambridge University Press, Cambridge.

Borsley, Robert D. 2010. An HPSG Approach to Welsh Unbounded Dependencies. In Müller, Stefan, editor, *Proceedings of the 17th International Conference on Head-Driven Phrase Structure Grammar*, pages 80–100, Stanford. CSLI Publications.

Davies, Cennard. 1988. Cymraeg Byw. In Ball, Martin J., editor, *The Use of the Welsh*, pages 200–210. Multilingual Matters, Clevedon.

Education Department, University College of Swansea. 1964. *Cymraeg Byw I*. Llyfrau'r Dryw.

Ellis, Nick C., Cathair O'Dochartaigh, William Hicks, Menna Morgan, and Nadine Laporte. 2001. Cronfa Electroneg o Gymraeg (CEG). A 1 million word lexical database and frequency count for Welsh. `http://www.bangor.ac.uk/ar/cb/ceg.php.en`.

Heinecke, Johannes. 1999. *Temporal Deixis in Welsh and Breton*. Anglistische Forschungen 272. Winter, Heidelberg.

Heinecke, Johannes. 2005. Aspects du traitement automatique du gallois. In *Actes de TALN 2005. Atelier "langues peu dotées"*. ATALA.

Isaac, Graham R. 1994. The Progressive Aspect Marker: W "yn" / OIr. "oc". *Journal of Celtic Linguistics*, 3:33–39.

`conll18/conll18_ud_eval.py`

	baseline			+*Eurfa*		
	precision	recall	f-measure	precision	recall	f-measure
ADJ	82.0	80.5	81.1	92.2	55.9	69.4
ADP	90.6	93.4	91.9	91.2	93.2	92.2
ADV	76.3	76.4	76.0	94.5	74.1	82.9
AUX	81.5	80.3	80.8	76.6	70.1	73.0
CCONJ	84.8	91.8	88.1	88.4	92.3	90.2
DET	97.6	99.0	98.3	98.8	98.0	98.4
NOUN	89.8	91.0	90.4	82.6	97.1	89.2
NUM	89.5	90.3	89.5	96.8	88.0	91.6
PART	88.2	85.3	86.6	79.4	82.5	80.8
PRON	93.8	89.8	91.7	98.5	90.9	94.5
PROPN	75.1	63.0	68.0	92.9	48.4	62.7
PUNCT	99.9	100.0	100.0	99.9	100.0	100.0
SCONJ	89.6	87.0	87.1	97.5	75.1	84.4
SYM	66.7	66.7	66.7	33.3	33.3	33.3
VERB	82.9	83.6	83.1	80.2	83.9	81.8

Table 5: results of POS tagging and lemmatisation per UPOS

Jones, Hywel M. 2012. *A statistical overview of the Welsh language*. Bwrdd yr Iaith Gymraeg/Welsh Language Board, Cardiff.

King, Gareth. 2003. *Modern Welsh. A comprehensive grammar*. Routledge, London, New York, 2 edition.

Kondratyuk, Daniel. 2019. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. `http://arxiv.org/abs/1904.02099`.

Lample, Guillaume and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. `http://arxiv.org/abs/1901.07291`.

Lynn, Teresa and Jennifer Foster. 2016. Universal Dependencies for Irish. In *CLTW*, Paris.

Morris-Jones, John. 1913. *A Welsh Grammar. Historical and Comparative*. Clarendon Press, Oxford.

Nivre, Joakim and Chiao-Ting Fang. 2017. Universal Dependency Evaluation. In Marneffe, Marie-Catherine de, Joakim Nivre, and Sebastian Schuster, editors, *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies*, pages 86–95, Göteborg.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Yoav Goldberg, Jan Hajič, Manning Christopher D., Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *the tenth international conference on Language Resources and Evaluation*, pages 23–38, Portorož, Slovenia. European Language Resources Association.

Roberts, Ian G. 2004. *Principles and Parameters in a VSO Language. A Case Study in Welsh*. Oxford Studies in Comparative Syntax. Oxford University Press, Oxford.

Rouveret, Alain. 1990. X-Bar Theory, Minimality, and Barrierhood in Welsh. In Randall, Hendryck, editor, *The Syntax of the Modern Celtic Languages*, Syntax and Semantics 23, pages 27–77. Academic Press, New York.

Sadler, Louisa. 1998. Welsh NPs without Head Movement. In Butt, Miriam and Tracy Holloway King, editors, *Proceedings of the LFG98 Conference*, Stanford. CSLI Publications.

Sadler, Louisa. 1999. Non-Distributive Features in Welsh Coordination. In Butt, Miriam and Tracy Holloway King, editors, *Proceedings of the LFG99 Conference*. CSLI Publications, Stanford.

Straka, Milan and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *ACL 2017*, pages 88–99, Vancouver.

Tallerman, Maggie. 2009. Phrase structure vs. dependency: The analysis of Welsh syntactic soft mutation. *Journal of Linguistics*, 45(1):167–201.

Thomas, R. J. and Gareth A. Bevan. 1950-2002. *Geiriadur Prifysgol Cymru. A Dictionary of the Welsh Language*. Gwasg Prifysgol Cymru, Caerdydd.

Thomas, Alan R. 1992. The Welsh Language. In MacAulay, Donald, editor, *The Celtic languages*, Cambridge language surveys. Cambridge University Press, Cambridge.

Thomas, Peter Wynn. 1996. *Gramadeg y Gymraeg*. Gwasg Prifysgol Cymru, Caerdydd.

Thorne, David. 1992. The Welsh Language, its History and Structure. In Price, Glanville, editor, *The Celtic Connection*, pages 171–205. Colin Smythe, Gerrards Cross.

Thorne, David. 1993. *A Comprehensive Welsh Grammar*. Blackwell, Oxford.

Tyers, Francis M. and Vinit Ravishankar. 2018. A prototype dependency treebank for Breton. In *Traitement Automatique des Langues Naturelles (TALN)*, pages 197–204, Rennes.

Uned Iaith Genedlaethol. 1978. *Gramadeg Cymraeg Cyfoes. Contemporary Welsh Grammar*. Brown a'i Feibion, Y Bontfaen.

Williams, Briony and Rhys James Jones. 2008. Acquiring Pronunciation Data for a Placenames Lexicon in a Less-Resourced Language. In *The sixth international conference on Language Resources and Evaluation*, Marrakech, Maroc. European Language Resources Association.

Williams, Briony, Rhys James Jones, and Ivan Uemlianin. 2006. Tools and resources for speech synthesis arising from a Welsh TTS project. In *The fifth international conference on Language Resources and Evaluation*, Genoa, Italy.

Williams, Stephen Joseph. 1980. *Elements of a Welsh Grammar*. University of Wales Press, Cardiff.

Williams, Briony. 1999. A Welsh speech database. Preliminary result. In *EuroSpeech 1999. Proceedings of the Sixth European Conference on Speech Communication and Technology, Budapest, Hungary, September 5-9, 1999*, Budapest.

Zeman, Daniel, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In Zeman, Daniel and Jan Hajič, editors, *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels. Association for Computational Linguistics.