A Dependency Treebank for Welsh

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

Abstract

In this paper we describe the development of the first syntactically-annotated corpus of Welsh within the Universal Dependencies (UD) project. We explain how the corpus was prepared, and we discuss some Welsh-specific syntactic 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). At the time of writing 601 sentences with 10 756 tokens in total have been annotated and released in version 2.4, with another 200 already annotated. The motivation for this work 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 for Welsh. 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. Another argument for the UD annotation is the possibility of improving a Welsh parser with data from other languages using cross-lingual approaches (Kondratyuk, 2019) or transfer learning.

The paper is laid out as follows: in Section 2 we give a short typological overview of Welsh. Section 3 describes briefly prior work on Welsh in computational linguistics and syntax and 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 The Welsh Language

Welsh is a Celtic language of the Brythonic branch 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 and who are 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 other Insular Celtic languages), it has initial consonant mutations, inflected prepositions (*ar* "on", *arnaf i* "on me"), genitive constructions with a single determiner (*tŷ'r brenin*, lit. "house the king": "the house of the king") and impersonal verbforms. However, unlike Breton, Welsh has predominantly VSO word order, does not have composed tenses with an auxiliary corresponding to "have", but uses extensively periphrastic verbal clauses to convey tense and aspect (Heinecke, 1999). Welsh has only verbnouns instead of infinitives (direct objects become genitive modifiers or possessives). Another feature of Welsh is the vigesimal number system (at least in the formal registers of the language)

and non-contiguous numerals for cardinals (*tri phlentyn ar hugain* "23 children" (lit. "three child(SG) on twenty") and ordinals (*unfed ganrif ar hugain* "21st century" (lit. "first century 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[1]. An overview can be found in Heinecke (2005). Research on Welsh syntax applying a wide range of formalisms 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 and leisure activities. It also contains texts from newspapers and magazines and some transcribed spoken Welsh.

Other important resources are the (non annotated) National Corpus of Contemporary Welsh[2], the parallel Welsh-English corpus of the Welsh Assembly[3] and *Eurfa*, a full form dictionary[4]. The Unimorph project[5] also provides Welsh data, however, currently this list contains only 10 641 forms (183 lemmas).

4 Corpus of the Treebank

As every language, Welsh has several formal and informal registers. All of those are written, which makes it difficult to constitute a homogeneous corpus. The differences are often in morphology and sometimes of syntactic nature. Frequently a 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.

Some sentences of the treebank have been taken from the Welsh language Wikipedia, mainly from pages on items about Wales, like 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 where the Welsh Assembly corpus, Welsh Grammars (in order to cover syntactic structures less frequently seen), web sites from Welsh institutions (Welsh Universities, *Cymdeithas yr Iaith*), Welsh language media (e.g. *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: *rdwyf (i)*, *rydwi*, *rydw (i)*, *dwi*. The same applies for the negation particle *ni(d)* which is contracted with the following form of *bod*, if the latter has an initial vowel: e.g. *nid oedd* "(he) was not" > *doedd*. Sometimes dialectal variants appear in the written language: *oeddan ni* vs *oedden ni* "we were". The corpus of the Welsh treebank retains the original forms. However, we use standardized lemmas (Thomas and Bevan, 1950 2002).

4.1 Preprocessing

In order to initiate the annotation, we transformed the CEG corpus into UD's CoNLL-U format and changed CEG's part-of-speech tags to UD UPOS. During this step we also corrected annotation errors (non-ambiguous cases) and added information about which consonant mutation is present, if any. We

[1] cf. Canolfan Bedwyr, University of Bangor, https://www.bangor.ac.uk/canolfanbedwyr/ymchwil_TI.php.en
[2] http://www.corcencc.org/
[3] http://cymraeg.org.uk/kynulliad3/
[4] http://eurfa.org.uk/, 210 000 forms, 10 000 lemmas and English glosses
[5] http://www.unimorph.org/

then used the *Eurfa* full-form dictionary to enrich the CoNLL-U format with morpho-syntactic features. With 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 (provided by Bojanowski et al. (2017)). We then POS-tagged and lemmatized our corpus with UDpipe. A second script pre-annotated some basic dependency relations (such as articles and prepositions). In addition to the UD standard, we defined language specific XPOS (see table 1), a morphological feature `Mutation` with three values to indicate the consonant mutation, since they carry syntactically relevant information, and subtypes for dependency relations `case:pred` (including those frequently used in other UD treebanks: `acl:relcl` and `flat:name`). The motivation to add XPOS was to be able to distinguish UPOS categories where there is a different syntactic context (e.g. independent vs. dependent pronouns, where the latter are used for direct objects of verbnouns, the former for direct objects of inflected verbs) without recurring to morphological features.

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

Table 1: Welsh specific XPOS

5 Dependencies

In the following step we annotated manually[6] and validated all layers: lemmas, UPOS, XPOS, and dependency relations using the annotation guidelines of Universal Dependencies. The annotation for this initial version were made by a single annotator.

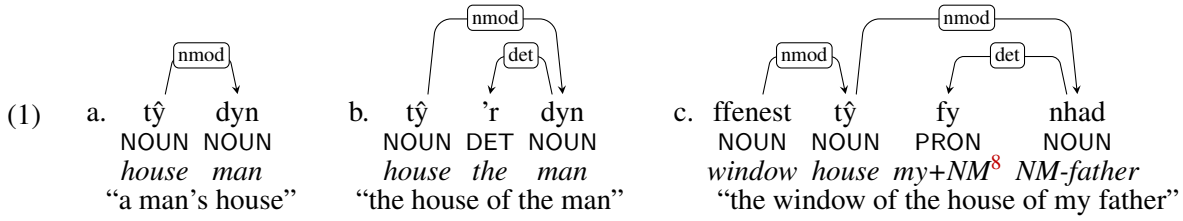
Since Irish is, amongst the languages present in the UD project, the language closest to Welsh, from typological point of view, we consulted the Irish data in order to annotate similar structures in a similar way (notably constructions with the auxiliary verb "to be" + time or aspect marker (TAM) + verbnoun (cf. section 5.2).

In the following subsections we discuss some of the particularities of the Welsh language, and how these are annotated within the UD v2 guidelines[7].

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 (example 1). This annotation is identical to Irish, Breton and Arabic treebanks:

(1)



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

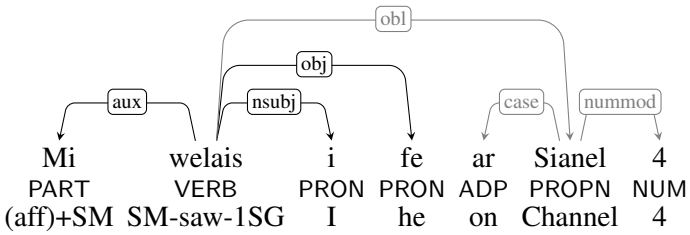
[7] https://universaldependencies.org/guidelines.html

[8] +NM means 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 be a preceding word, e.g. temporal and spatial adverbials or undefinite direct objects.

5.2 Periphrastic verbal construction

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 expressed by periphrastic constructions using forms of the verbnoun *bod* "to be". Whereas the inflected verbs (2)[9] are annotated in a straight forward way, the periphrastic forms need some attention (2). 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:

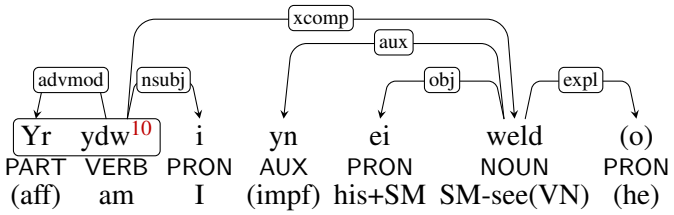
(2)



"I saw him on Channel 4"

The periphrastic construction (3) employs at least one (inflected) form of *bod* and a time or aspect marker (TAM). The (independent) pronoun after the verbnoun (VN) is facultative and repeats the (dependent) pronoun before the verbnoun *gweld*.

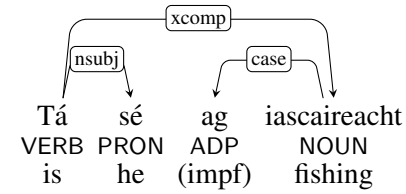
(3)



"I am seeing him"

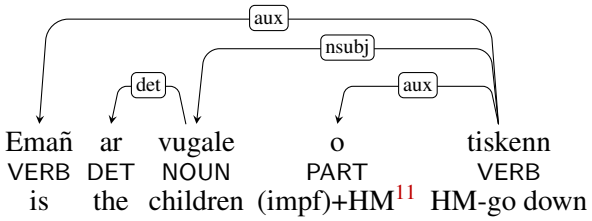
The same annotation can be found in the Irish treebank (4a). In the Breton treebank, however, the infinitive is the phrasal head (4b).

(4) a.



"He is fishing"

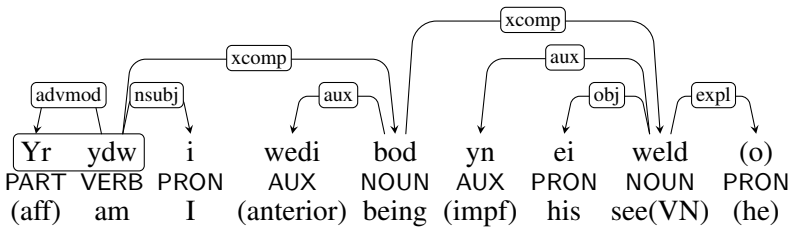
b.



"The children are going down."

There are at least two reasons which are at the origin of this annotation: (1) In the Irish treebank these constructions are annotated similary, and (2) this avoids totally flat trees in case of nested periphrastic constructions, if the verb *bod* and the TAM marker were annotated both as an `aux`. For instance a nested periphrastic constructions exists for the imperfective version of (3), shown in (5)

(5)

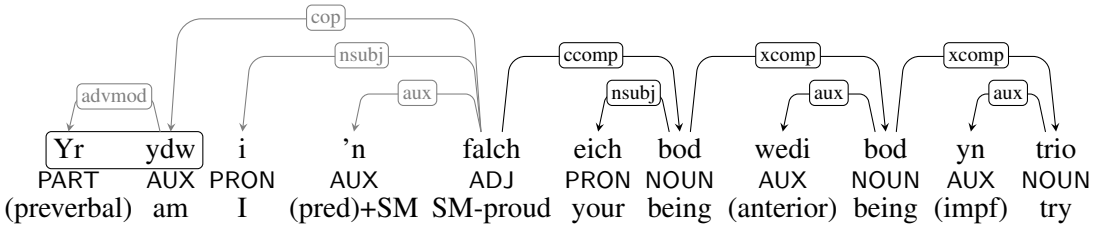


"I have been seeing him"

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:

[9] Dependency relations in gray are irrelevant for the point made in the example.
[10] Multitoken word *Rydw* in the original sentence.
[11] HM: Breton hard mutation

(6)



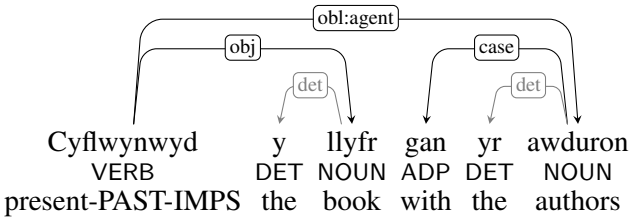
"I'm proud that you have been trying"

Other combinations are possible too, also using other TAM (*ar*, *am* (posteriority, "about to", *wedi* (anteriority, cf. English Present/Past Perfect), *hen* (distant anteriority), *newydd* (recent anteriority), cf. Heinecke (1999, p. 271). We decided to use the inflected form of *bod* as the syntactic head and link the subsequent verbnoun *bod* and finally the verbnoun which carries the meaning with as `xcomp` to avoid completely flat (sub)trees which do not show the internal structure of these phrases.

5.3 Impersonal and *cael*-periphrastics

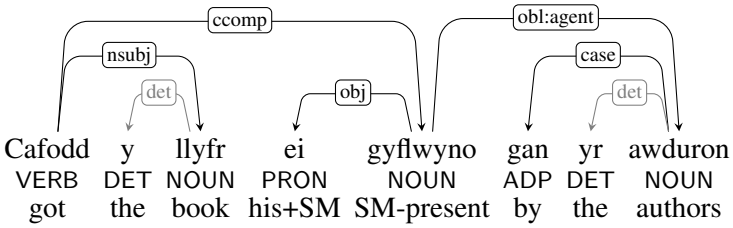
Like the other Celtic languages Welsh verbs (including intransitive verbs) have impersonal forms in all tenses. In this construction the demoted agent can be expressed using the preposition *gan* "with". As in the Irish and Breton treebanks, its core argument is marked `obj` (7). Even though impersonal are often translated into English using passive voice, they cannot be considered as passive, since the direct object is not promoted to subject position and impersonals exist for intransitive verbs. A periphrastic construction exists using the verb *cael* "get" (8).

(7)



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

(8)

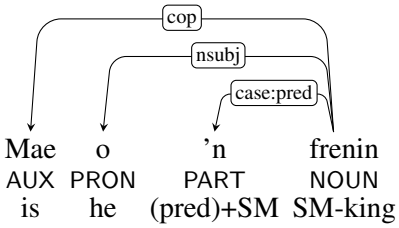


"The book was presented by the authors"
(lit. "got the book his presenting by the authors")

5.4 Nonverbal predicates

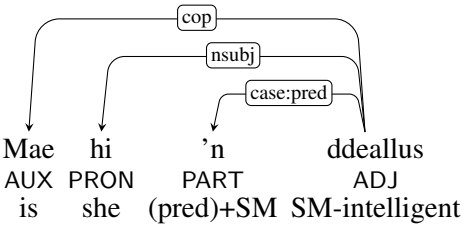
Like in most other languages, adjectives and nouns can be head, if they are the predicate. In Welsh, however, such adjectives and nouns need a special predicate marker *yn* attached by `case:pred` (ex. 9).[12]

(9) a.



"He is king"

b.

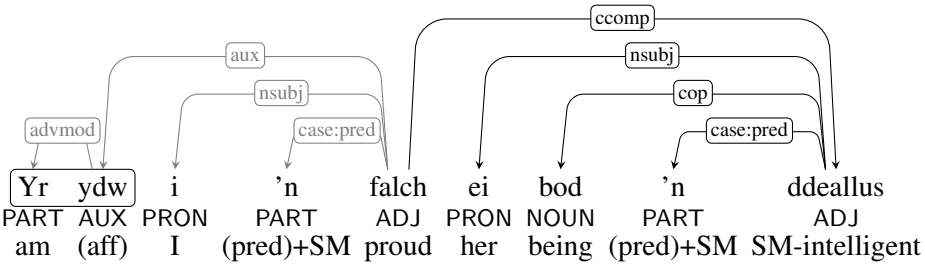


"She is intelligent"

[12]There are three forms *yn* in Welsh with tree different functions: 1) predicative marker which triggers soft mutation, 2) imperfective marker, which precedes a verbnoun as seen above (Isaac, 1994), 3) preposition "in" which triggers nasal mutation). All three are shortened to *'n* if the preceding word terminates with a vowel.

In subordinates, the inflected form of *bod* becomes a verbnoun, so the subject is attached as possessive (dependent) pronoun.

(10)

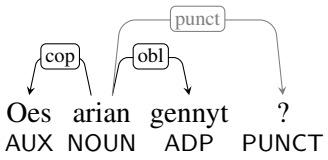


"I'm proud that she is intelligent"

5.5 Inflected prepositions

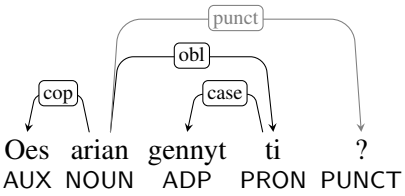
All Celtic languages have contracted forms of prepositions and following pronouns. In Welsh, the pronoun can follow the contracted preposition, so it is more adequate to speak of inflected prepositions (Morris-Jones, 1913; King, 2003). This requires a different annotation, since in (11a) the inflected preposition incorporates the `obl` argument. In (11b) *gennyt* is attached as a simple `case` to the pronoun.

(11) a.



"Do you have money?"
(lit. "Is money with-2SG?")

b.



"Do you have money?"
(lit. "Is money with-2SG you-SG?")

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 warned on potential errors (e.g. a `det` with a VERB head). This script also checked all forms of the verb *bod* and completes morphological features. For nouns and adjectives the script warns 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 (average sentence length: 17,9 tokens, shortest: 4 tokens, longest: 59 tokens, median length: 16). Since verbnouns have the UPOS NOUN, 30% of all UPOS are NOUN, (ADP 12.9%, ADJ 6.9% DET 6.5%, VERB and PRON 6.3%). The relatively small number of VERB is relativized if we regard the distribution of the XPOS: noun 21.3%, verbnoun 9.1% + verb 6.3% = 15.4% "verbs". Amongst the core dependency relations, the most frequent are `nsubj` 5.7%, `obj` 5.2%, `xcomp` 4.3% and `ccomp` 1.7%. Overall, the most frequent relations are `case` 10.4%, `nmod` 8.3%, `det` 6.8% and `obl` 5.9%. All but 6 dependency relations defined in the UD guidelines are used, the absent relations 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).

7 Evaluation

Even though the treebank is still small, tests for tagging and dependency parsing show results comparable with similar sized treebanks (Tyers and Ravishankar (2018) report a LAS between 64.14% and 74.29%, Zeman et al. (2018) mention a LAS of 70.88 for Irish). We used Udpipe (v2, single model for tagging and parsing). Test with Wikipedia embeddings (500 dimensions) trained with fastText (Bojanowski et

al., 2017) did not improve the parsing[13]. Therefore, the numbers in table 2b are the results without word embeddings.

For the evaluation, we shuffled and 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] to calculate all scores. Table 2a shows the results of POS tagging and lemmatisation without and with the *Eurfa* dictionary.

The UPOS tag with the biggest error rate is PROPN (37.4% errors (148 tokens)). In 109 cases, the parser chose NOUN, which is, after all, not the worst guess. Out of 674 tokens with a gold VERB UPOS tag, 111 tokens were missed (the wrong UPOS were mainly AUX (50 tokens) and NOUN (40). Out of 3242 NOUN tags, 290 were wrong (73 PROPN).

For the parsing (on gold tags) dependency relations like `case` or `det` are relatively well predicted (40 wrong out of 1121 for the former (3.6%), 11 out of 736 (1.5%) for the latter). For more central relations the error rate is higher `nsubj` (150/609, 24.6%), `obj` (70/556, 12.6%), `obl` (277/639, 43.4%). However, subordinations pose problems: `acl` (78/167, 46.7%) `advcl` (120/150, 80% `ccomp` (77/181, 42.5%) (however `xcomp` has a lower score, due to the fact it is frequent in periphrastic constructions with forms of the verb *bod*, 68 wrong out of 461 (14.8%). A more detailed analysis, notably whether these error are due to a certain context, is yet to be done.

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

		POS tags	UAS	LAS	CLAS
b)	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 2: POS tagging and lemmatisation (left) dependency parsing (right). Best results in bold.

Nearly half of the word forms in the test corpus are out-of-vocabulary (OOV) with respect to the training corpus. The *Eurfa* 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 with an additional dictionary (UDpipe switches off its guesser, if a dictionary is provided). Results of the parsing are presented in table 2b. We run four tests, a model trained solely on the treebank, with dependencies predicted on the results of the tagger, and dependencies predicted on gold tags, both tests with and without the *Eurfa* dictionary. Using the dictionary only slightly increases UAS, LAS or CLAS.

8 Conclusion and Future Work

The most obvious work is to increase the number of sentences annotated in order to see whether this leads to better tagging and parsing results. Another important problem is the absence of very formal Welsh (as in the Bible and some literary works) and of very informal written Welsh (as is used by some Welsh bloggers).

In order to increase the tagging and parsing results, we envisage tests with transfer learning might be a solution. From a typological point of view, from all existing UD treebanks, the Irish data is the most promising, since Irish shares the compound tenses with TAM markers and the rather strict VSO word order.

With word embeddings becoming more important, work on Welsh word embeddings is needed too. We need to dig into cross-lingual approaches (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.

[13] This might be due to the relatively small corpus used to train the embeddings: the Welsh Wikipedia (april 2019) contains 62MB of compressed raw data (104 000 pages.

[14] https://universaldependencies.org/conll18/conll18_ud_eval.py

References

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

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

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

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

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

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

Nick C. Ellis, 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.

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

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

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

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

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

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

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

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

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

Joakim Nivre, 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.

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

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

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

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

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

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

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

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

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

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

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

Francis M. Tyers 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.

Briony Williams 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.

Briony Williams, 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.

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

Briony Williams. 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.

Daniel Zeman, 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 Daniel Zeman 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.