A Dependency Treebank for Welsh

Johannes Heinecke

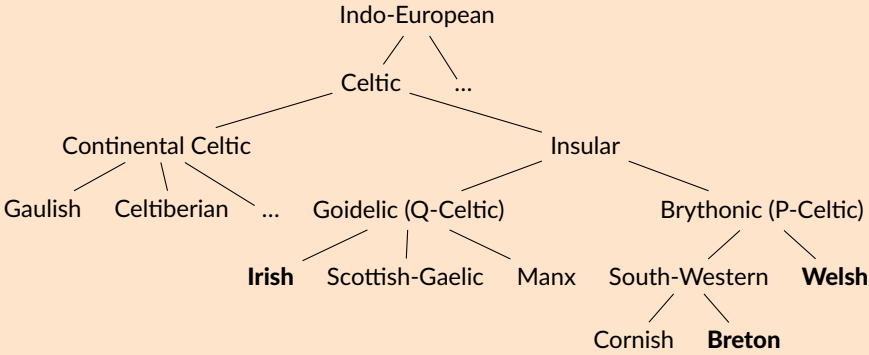
Orange / TGI / Data & IA / DESKIÑ

Universal Dependencies Workshop 2019

Table of Contents

Welsh I

Affiliation



- 500 000 speakers (Wales and Chubut, Argentina)
- all bilingual (English, or Spanish)

Welsh II

Like Breton or Irish

- initial consonant mutation
- inflected prepositions *ar* “on” *arnaf fi* “on me”
- genitive construction with single determiner (cf. Arabic):
  - *tŷ'r brenin* “the house of the king”
  - vs. *tŷ brenin* “a king’s house”
- VSO basic word order
- composed numbers (nouns are always in singular after a numeral)
  - *tri phlentyn ar ddeugain* “three child on two twenty” → “43 children”
  - *unfed ganrif ar hugain* “first century on twenty” → “21st century”
- impersonal forms (including intransitive verbs) *gwelwyd*, *cysgir* “one saw, one will sleep”
- possession expressed with a prepositional phrase: *Mae arian gen i* “there is money with me”
- singulatives (add suffix (plus Umlaut) to have singular from basic plural):
  - *plant* “children”, *adar* “birds”, *llygod* “mice”, *caws* “cheese”
  - *plentyn* “child”, *aderyn* “bird”, *llygoden* “mouse”, *cosyn* “piece of cheese”

Welsh III

Like Irish but unlike Breton

- no auxiliary corresponding to “to have”, thus no composed tenses
- no participles (but verbal adjectives)
- periphrastic constructions with *bod* “to be” and tense-aspect-markers for most tenses and aspects
- pronouns not subject/object (NOM/ACC) but independent (subject position)/dependent (possessives and object position)
- no infinitives but verbnouns (direct object marked differently on verbnouns than on verbs)
  - patient marked in the same way as possessives using dependent pronouns (cf. *dy gwrw dŷ* [*ti*])
  - different to inflected verb forms, were nominal direct objects undergo soft mutation, and independent pronouns are used for pronominal direct objects.
  - cf. German:

inflected:	*ich sehe den Hund*$_{ACC}$	“I see the dog”
infinitive:	*den Hund sehen*$_{ACC}$	“to see the dog”
verbnoun:	*das Sehen des Hundes*$_{GEN}$	“the sight/seeing of the dog”

  - Welsh

inflected:	*Mi welodd o ti*$_{indep}$	“He saw you” (lit. “(aff) saw he you”)
verbnoun:	*Roedd o'n dy*$_{dep}$ *weld*	“He was seeing you” (lit. “Was he (impf) your seeing”)

NLP resources I

Research

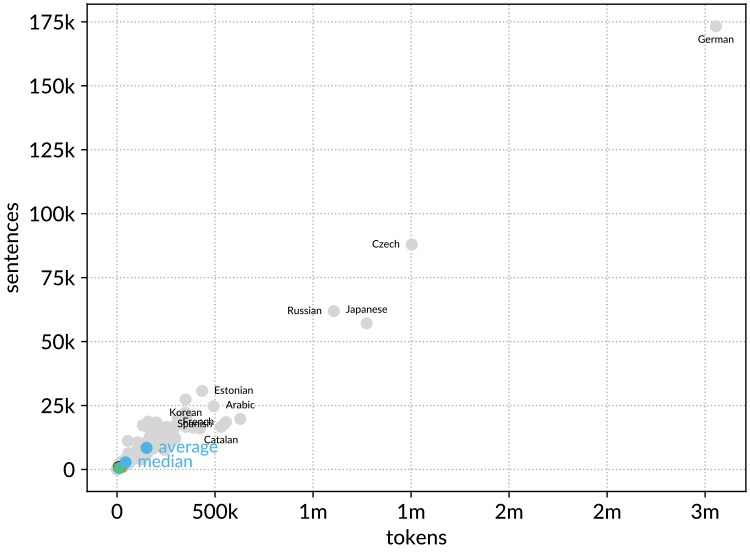
- important work on syntax using frameworks like LFG, GB, HPSG (Awbery 1976, Rouveret 1990, Sadler 1998, Sadler 1999, Roberts 2004, Borsley, T. Tallerman, et al. 2007, M. Tallerman 2009, Borsley 2010)
- orthographic correction
- automatic speech recognition (ASR) and speech synthesis (TTS) (Williams 1999, Williams, Jones, and Uemlianin 2006, Williams and Jones 2008)
- natural language understanding (NLU)
- cf. Canolfan Bedwyr, Bangor University (https://www.bangor.ac.uk/canolfanbedwyr)

NLP resources II

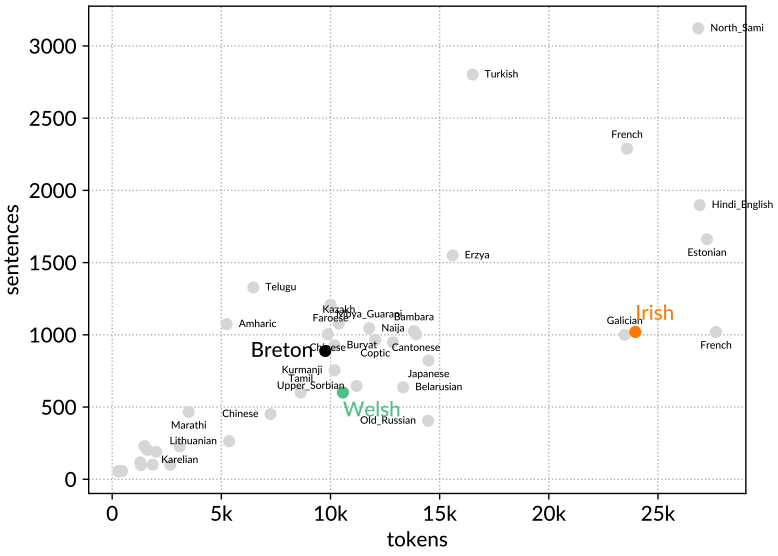
(Publicly available) data

- *Cronfa Electroneg o Gymraeg* (CEG, Ellis, O'Dochartaigh, et al. 2001): 1 000 000 tokens annoted with lemmas and POS
- National Corpus of Contemporary Welsh (http://www.corcencc.org/)
- Wikipedia (+100 000 pages), word embeddings: fastText (Bojanowski, Grave, et al. 2017), BERT (Lample and Conneau 2019)
- Wictionary (and Unimorph) very little data (http://www.unimorph.org/)
- *Eurfa* full form dictionary (210 000 forms, 10 000 lemmas and English glosses, http://eurfa.org.uk/)
- parallel Welsh/English corpus from Welsh Assembly (http://cymraeg.org.uk/kynulliad3/)

The Welsh language in UD



The Welsh language in UD



Corpus of the Treebank

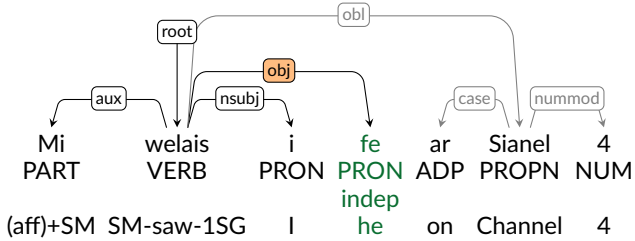
10 756 tokens, 601 sentences

- shortest sentence: 4 words, longest: 59, average length: 17.9, median length: 16
- Wikipedia (pages on items of Wales)
- Welsh assembly corpus
- Media (BBC Cymru, Y Golwg)
- Web sites of Welsh universities and organisations (*Cymdeithas yr Iaith*, *Urdd Gobaith Cymru*, county councils)
- Welsh language blogs
- Welsh Grammars and Textbooks

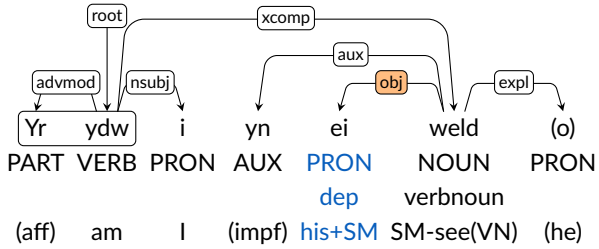
(pre)processing

- UD-isation of CEG and training UDpipe with CEG and external dictionary (*Eurfa*)
- POS annotation/lemmatisation of treebank sentences (using model trained on CEG)
- manual validation of lemmas, XPOS, UPOS and annotation of dependency relations
- validation scripts (checking (some) features, adding mutation type etc.)

Periphrastic verbal constructions

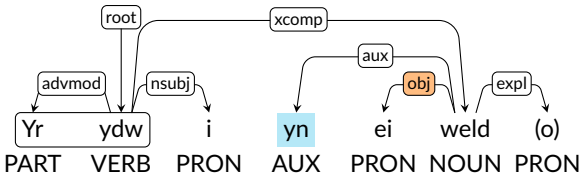


"I saw him on Channel 4"



"I am seeing him"

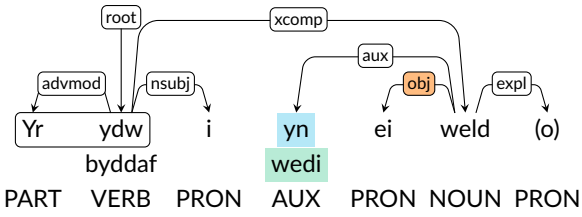
Periphrastic verbal constructions: Tense-aspect-markers



“I’m seeing him”

(Heinecke 1999)

Periphrastic verbal constructions: Tense-aspect-markers

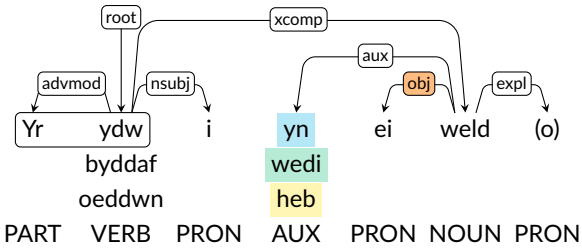


"I'm seeing him"

"I will have seen him" ("I'm after seeing him")

(Heinecke 1999)

Periphrastic verbal constructions: Tense-aspect-markers



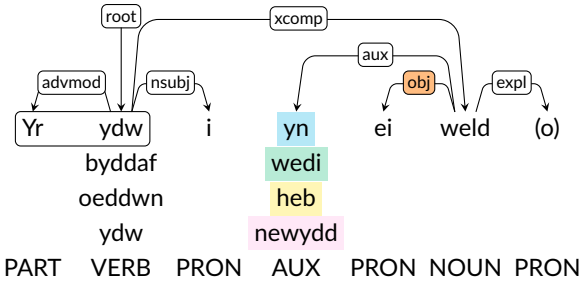
"I'm seeing him"

"I will have seen him" ("I'm after seeing him")

"I had not seen him" ("I'm without seeing him")

(Heinecke 1999)

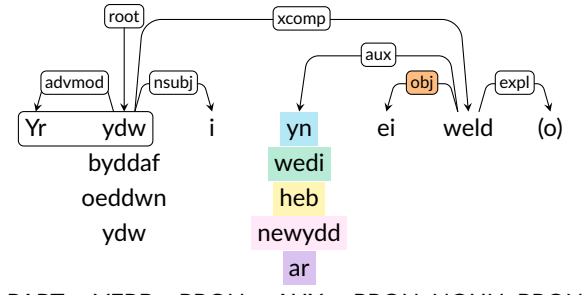
Periphrastic verbal constructions: Tense-aspect-markers



"I'm seeing him"
"I will have seen him" ("I'm after seeing him")
"I had not seen him" ("I'm without seeing him")
"I have just seen him" ("I'm new seeing him")

(Heinecke 1999)

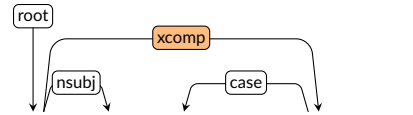
Periphrastic verbal constructions: Tense-aspect-markers



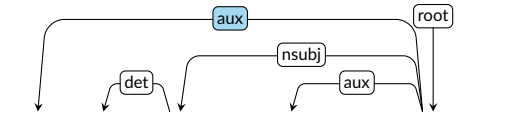
"I'm seeing him"
"I will have seen him" ("I'm after seeing him")
"I had not seen him" ("I'm without seeing him")
"I have just seen him" ("I'm new seeing him")
"I'm about to see him" ("I'm on seeing him")

(Heinecke 1999)

Periphrastic verbal constructions in Irish and Breton



"He is fishing"
(Lynn and Foster 2016)

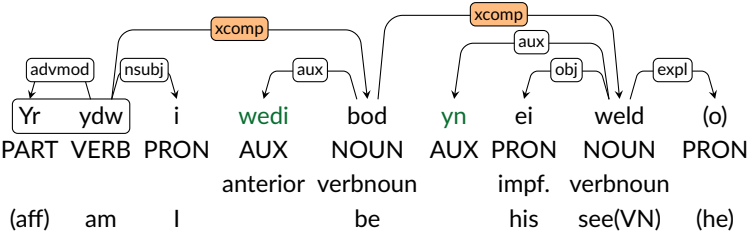


"The children are going down."
(Tyers and Ravishankar 2018)

- The Welsh treebank follows the Irish example in making "to be" head and attach the verb as *xcomp*
- Difference: Irish attaches TAM (ADP) as *case*, Welsh attaches TAM (AUX) as *aux*: Harmonisation needed

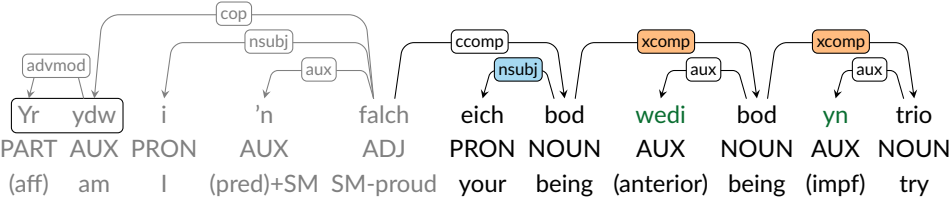
Nested periphrastic verbal constructions

- Anterior Present, Imperfective



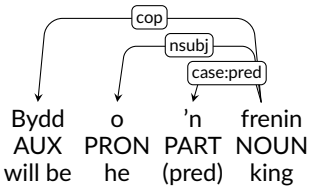
"I have been seeing him"

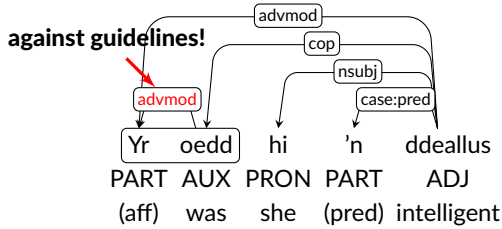
- Subordinate (no finite Tense)

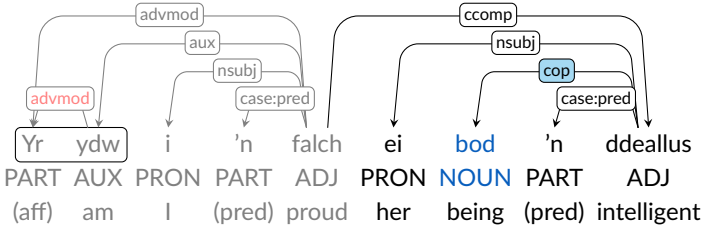


"I'm proud that you have been trying"

Nonverbal predicates







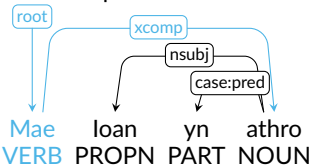
"I'm proud that she is intelligent"

- N.B. *yn*: 3 syntactically distinct homographs:
  - imperfective TAM (before verbnouns, no mutation triggered)
  - predicative marker (before nouns and adjectives, triggers soft mutation)
  - preposition "in" (triggers nasal mutation)

Comment:
Copula can be dropped:
Liz yn frenhines ``Liz (is) queen''

Constructions without *bod*

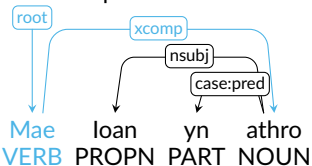
- Nonverbal predicate



“Ioan (is) teacher”

Constructions without *bod*

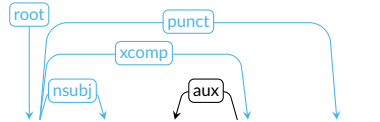
- Nonverbal predicate



“Ioan (is) teacher”

Constructions without *bod*

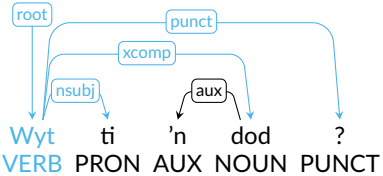
- Verbnoun: Head of subject (*bod*) dropped (create empty word or new annotation)?



"(Do) you come?"

Constructions without *bod*

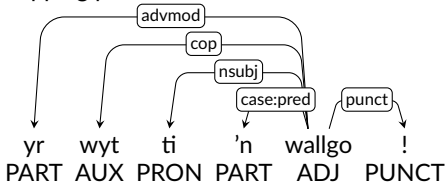
- Verbnoun: Head of subject (*bod*) dropped (create empty word or new annotation)?



"(Do) you come?"

Ellipsis

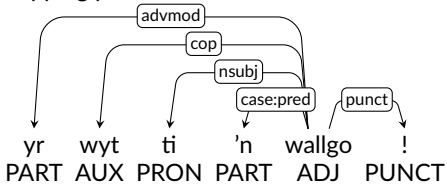
- Dropping predicate:



"you are stupid!"

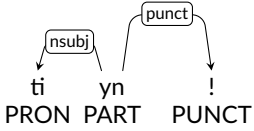
Ellipsis

- Dropping predicate:



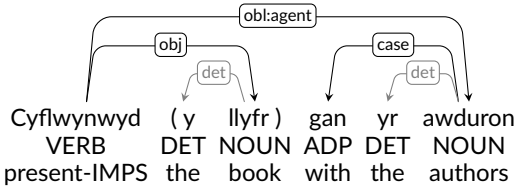
"you are stupid!"

- Which syntactic structure?

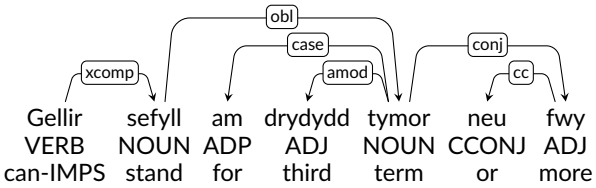


("Oh, I'm stupid!") "you are!"

Impersonal Forms ≠ Passive, *cael*-"Passive"

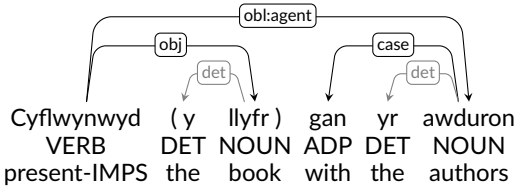


"(The book) was presented by the authors" (lit. "One presented the book by the authors")
→ direct object faculative

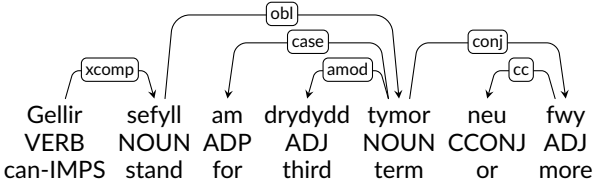


"One can stand for a third term or more"
→ impersonal modal verb

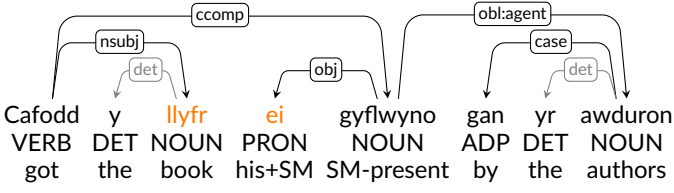
Impersonal Forms ≠ Passive, *cael*-"Passive"



"(The book) was presented by the authors" (lit. "One presented the book by the authors")
→ direct object faculative

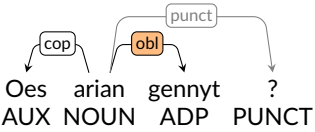


"One can stand for a third term or more"
→ impersonal modal verb

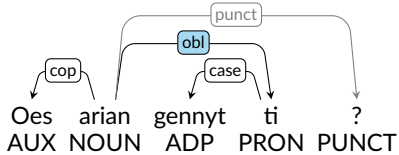


"The book was presented by the authors"
(lit. "got the book his presenting by the authors")

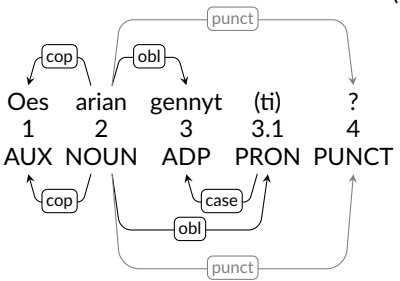
Inflected prepositions



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

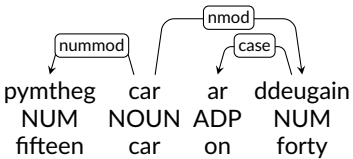


“Do you have money?”
(lit. “Is money with-2SG you-SG?”)



Compound numbers

- vigesimal system
  - 20 = *ugain*
  - 30 = *deg ar hugain* “ten on twenty”
  - 40 = *deugain* “two twenties”
  - 60 = *trigain* “three twenties”



“55 cars” (lit. “15 cars on 2*20”)

Statistics (POS)

- 10756 tokens, 601 sentences

UPOS	%	*XPOS*	%	*XPOS*	%	*XPOS*	%
NOUN	30.1	noun	21.5	aff	1.3	ord	0.1
ADP	12.9	prep	12.5	person	1.3	refl	0.1
PUNCT	9.7	punct	9.7	num	1.2	intr	0.1
ADJ	6.9	**verbnoun**	9.1	ante	1.1	sym	0.1
DET	6.5	art	6.5	neg	0.9	int	0.0
PRON	6.3	**verb**	6.3	dem	0.6	work	< 0.0
VERB	6.3	pos	6.0	cprep	0.6	perf	< 0.0
AUX	5.9	cconj	2.9	sup	0.5	contr	< 0.0
PART	4.4	dep	2.7	sconj	0.5	card	< 0.0
PROPN	3.7	impf	2.5	rel	0.4	propn	< 0.0
CCONJ	2.9	indep	2.4	cmp	0.3		
ADV	1.9	place	2.3	pron	0.2		
NUM	1.3	pred	2.1	org	0.1		
SCONJ	0.5	adv	2.0	post	0.1		
SYM	0.1	aux	1.9	eq	0.1		

Statistics (dependency relations)

- 10756 tokens, 601 sentences

deprel	%	*deprel*	%	*deprel*	%
case	10.5	mark	2.1	csubj	0.1
punct	9.7	cop	2.0	nmod:agent	0.1
nmod	8.4	ccomp	1.7	compound	< 0.0
det	6.9	nmod:poss	1.7	iobj	< 0.0
obl	6.0	acl	1.6		
nsubj	5.7	advcl	1.4		
root	5.6	acl:relcl	1.2		
obj	5.2	flat:name	0.8		
advmod	5.2	flat	0.6		
amod	4.8	nummod	0.6		
xcomp	4.3	fixed	0.5		
aux	3.7	appos	0.5		
conj	3.2	expl	0.2		
cc	2.9	obl:agent	0.2		
case:pred	2.1	parataxis	0.2		

Evaluation (UDPipe)

- train: 80%, dev: 10%, test: 10%
- 10-fold cross-evaluation
- tagging and lemmatisation

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

- dependency parsing

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

- No increase in UAS, LAS, CLAS with word embeddings (fastText)

Errors: UPOS

UPOS	errors	all	%	wrong UPOS
PUNCT	0	1039	0.0	
DET	7	698	1.0	PART:6 NOUN:1
ADP	91	1389	6.6	NOUN:25 PART:24 CCONJ:12 AUX:11 ADV:5 PRON:5 ADJ:4 SCONJ:2 VERB:2 PROPN:1
CCONJ	25	312	8.0	ADP:15 PRON:4 NOUN:2 SCONJ:2 PROPN:1 ADJ:1
NOUN	290	3242	8.9	ADJ:81 PROPN:73 ADP:34 ADV:30 VERB:19 AUX:17 PRON:16 NUM:9 CCONJ:7 PART:2 SCONJ:1 DET:1
NUM	13	134	9.7	NOUN:8 PRON:3 ADJ:1 PROPN:1
PRON	68	676	10.1	ADP:23 CCONJ:23 NOUN:10 PART:5 ADJ:4 AUX:2 VERB:1
PART	68	473	14.4	AUX:30 ADP:16 DET:16 PRON:4 NOUN:1 CCONJ:1
SCONJ	7	48	14.6	CCONJ:3 ADP:3 NOUN:1
VERB	111	674	16.5	AUX:50 NOUN:40 ADP:10 ADJ:7 PRON:2 PROPN:1 CCONJ:1
ADJ	145	738	19.6	NOUN:98 PROPN:11 ADP:9 VERB:9 ADV:8 PRON:4 AUX:3 SCONJ:1 NUM:1 CCONJ:1
AUX	126	631	20.0	VERB:80 ADP:17 NOUN:15 PART:12 ADJ:2
ADV	52	206	25.2	NOUN:27 ADJ:8 VERB:6 ADP:3 PROPN:2 PART:2 PRON:2 CCONJ:2
SYM	2	6	33.3	PUNCT:1 NOUN:1
PROPN	148	396	37.4	NOUN:109 ADJ:21 ADP:5 ADV:4 PART:2 PRON:2 NUM:2 AUX:2 CCONJ:1
total:	1153	10756	10.7	

Errors: dependency relations I

- Parsing on gold UPOS

deprel	errors	all	%	wrong deprel
case:pred	0	227	0.0	
punct	0	1039	0.0	
aux	5	396	1.3	cop:2 case:1 advmod:1 case:pred:1
det	11	736	1.5	nmod:8 obl:1 advmod:1 nsubj:1
cc	6	311	1.9	mark:6
case	40	1121	3.6	mark:31 advmod:4 fixed:2 obl:1 obj:1 nsubj:1
amod	23	511	4.5	advmod:10 obl:3 det:2 nmod:2 root:2 obj:1 nsubj:1 conj:1 flat:name:1
nmod:poss	11	179	6.1	obj:4 nmod:4 obl:1 expl:1 nsubj:1
cop	19	208	9.1	root:8 acl:relcl:3 advmod:2 advcl:2 aux:2 acl:1 punct:1
advmod	52	554	9.4	amod:26 case:10 root:4 ccomp:3 obj:2 nsubj:2 obl:1 cc:1 appos:1 advcl:1 acl:1
obj	70	556	12.6	obl:23 nsubj:15 nmod:14 ccomp:4 xcomp:3 flat:name:3 root:2 case:1 cc:1 nummod:1 acl:relcl:1 conj:1 nmod:poss:1
nmod	122	899	13.6	obl:56 acl:15 flat:13 nsubj:7 conj:6 obj:3 flat:name:3 appos:3 nmod:poss:3 mark:2 acl:relcl:2 nummod:2 xcomp:1 root:1 advmod:1 det:1 case:1 amod:1 obl:agent:1
xcomp	68	461	14.8	acl:29 ccomp:12 advcl:9 nmod:7 obl:4 obj:3 root:2 amod:1 nsubj:1
mark	39	226	17.3	case:30 cc:8 advmod:1
root	109	601	18.1	acl:19 nsubj:15 nmod:12 conj:12 acl:relcl:9 cop:8 ccomp:8 advcl:7 amod:4 advmod:4 obj:3 appos:3 xcomp:2 obl:1 cc:1 flat:name:1

Errors: dependency relations II

nsubj	150	609	24.6	obj:60 nmod:23 root:20 xcomp:9 obl:8 mark:6 conj:5 amod:5 nmod:poss:3 appos:2 advmod:2 ccomp:2 case:1 det:1 flat:name:1 expl:1 nummod:1
conj	85	345	24.6	nmod:35 appos:8 acl:8 advmod:5 ccomp:5 acl:relcl:5 amod:4 obj:3 nsubj:3 xcomp:2 obl:2 cop:1 root:1 advcl:1 nmod:poss:1 nummod:1
nummod	16	64	25.0	nmod:8 obl:3 obj:2 case:1 mark:1 conj:1
acl:relcl	35	132	26.5	conj:8 acl:8 root:5 nmod:4 xcomp:3 obl:2 advcl:2 cop:1 parataxis:1 amod:1
flat:name	23	86	26.7	nmod:15 flat:2 appos:2 nummod:2 obj:1 punct:1
ccomp	77	181	42.5	xcomp:17 advcl:15 nmod:8 obl:7 obj:7 advmod:5 acl:4 amod:3 conj:3 root:3 cop:2 acl:relcl:2 nsubj:1
obl	277	639	43.3	nmod:149 obj:39 root:31 case:10 conj:10 xcomp:8 nummod:7 nsubj:5 acl:5 advcl:4 appos:2 mark:2 ccomp:2 flat:name:2 advmod:1
acl	78	167	46.7	nmod:26 xcomp:14 advcl:10 ccomp:8 acl:relcl:6 conj:4 obl:3 root:3 amod:2 obj:1 nsubj:1
obl:agent	15	22	68.2	obl:7 nmod:5 acl:1 root:1 nsubj:1
appos	39	57	68.4	nmod:12 conj:12 obl:4 nsubj:3 acl:2 flat:name:2 nummod:1 advmod:1 advcl:1 mark:1
fixed	41	56	73.2	case:24 obl:6 advmod:3 nmod:3 cc:2 conj:2 acl:1
flat	52	66	78.8	nmod:50 flat:name:2
advcl	120	150	80.0	acl:28 root:22 obl:15 xcomp:11 nmod:11 advmod:10 ccomp:9 conj:8 amod:4 obj:1 acl:relcl:1
expl	20	24	83.3	nsubj:9 nmod:6 obj:4 obl:1
parataxis	16	17	94.1	nmod:5 acl:3 ccomp:2 xcomp:1 obl:1 nsubj:1 acl:relcl:1 advcl:1 conj:1
iobj	1	1	100.0	nmod:1
compound	5	5	100.0	nmod:3 advmod:1 amod:1
nmod:agent	6	6	100.0	obl:2 nmod:2 obl:agent:1 appos:1
csubj	10	10	100.0	root:4 acl:3 xcomp:1 conj:1 nmod:1
total:	1641	10756	15.3	

Conclusion

So far ...

- First Welsh treebank, third Celtic language treebank in UD
- 10 000 tokens, allows a relatively robust POS tagging (89.2%), and parsing of sentences

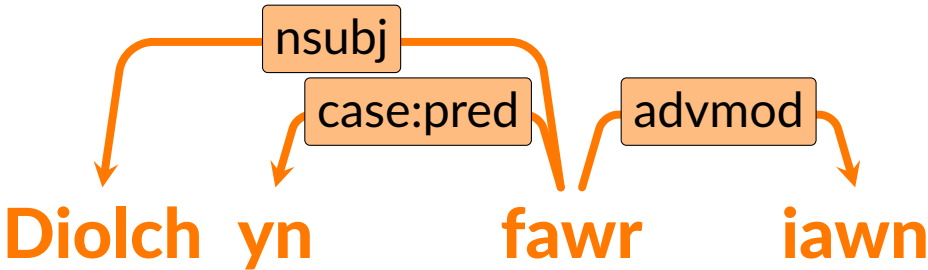
Conclusion

So far ...

- First Welsh treebank, third Celtic language treebank in UD
- 10 000 tokens, allows a relatively robust POS tagging (89.2%), and parsing of sentences

Next steps

- needs to be expanded
- more annotators needed
- adding enhanced dependencies
- adding translations of sentences and glosses to words
- harmonisation with similar constructions in other languages?



https://github.com/UniversalDependencies/UD_Welsh-CCG/
johannes.heinecke@orange.com

References I

- Awbery, Gwenllian M. (1976). *The Syntax of Welsh. A transformational Study of The Passive*. Cambridge: Cambridge University Press. ISBN: 0-521-21341-X.
- Bojanowski, Piotr, Edouard Grave, et al. (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.
- Borsley, Robert D. (2010). "An HPSG Approach to Welsh Unbounded Dependencies". In: *Proceedings of the 17th International Conference on Head-Driven Phrase Structure Grammar*. Ed. by Stefan Müller. Stanford: CSLI Publications, pp. 80–100.
- Borsley, Robert D., Taggie Tallerman, et al. (2007). *The Syntax of Welsh*. Cambridge: Cambridge University Press. ISBN: 978-051148622-7.
- Ellis, Nick C., Cathair O'Dochartaigh, et al. (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. Heidelberg: Winter. ISBN: 3-8253-0922-3.
- 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.

References II

- Roberts, Ian G. (2004). *Principles and Parameters in a VSO Language. A Case Study in Welsh*. Oxford Studies in Comparative Syntax. Oxford: Oxford University Press. ISBN: 0-19-516821-6.
- Rouveret, Alain (1990). "X-Bar Theory, Minimality, and Barrierhood in Welsh". In: *The Syntax of the Modern Celtic Languages*. Ed. by Hendryck Randall. Syntax and Semantics 23. New York: Academic Press, pp. 27–77. ISBN: 0-12-613523-1, -606104-1.
- Sadler, Louisa (1998). "Welsh NPs without Head Movement". In: *Proceedings of the LFG98 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.
- — (1999). "Non-Distributive Features in Welsh Coordination". In: *Proceedings of the LFG99 Conference*. Ed. by Miriam Butt and Tracy Holloway King. Stanford: CSLI Publications.
- Tallerman, Maggie (2009). "Phrase structure vs. dependency: The analysis of Welsh syntactic soft mutation". In: *Journal of Linguistics* 45(1), pp. 167–201.
- Tyers, Francis M. and Vinit Ravishankar (2018). "A prototype dependency treebank for Breton". In: *Traitement Automatique des Langues Naturelles (TALN)*. Rennes, pp. 197–204.
- 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.

References III

- 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.