

Universal dependencies for Scottish Gaelic: syntax

Colin Batchelor

Royal Society of Chemistry, Thomas Graham House, Cambridge, UK CB4 0WF

colin.r.batchelor@gmail.com

Abstract

We present universal dependencies for Scottish Gaelic and a treebank of 1021 sentences (20 021 tokens) drawn from the Annotated Reference Corpus Of Scottish Gaelic (ARCOSG). The tokens are annotated for coarse part-of-speech, fine-grained part-of-speech, syntactic features and dependency relations. We discuss how the annotations differ from the treebanks developed for two other Celtic languages, Irish and Breton, and in preliminary dependency parsing experiments we obtain a mean labelled attachment score of 0.792. We also discuss some difficult cases for future investigation, including cosubordination. The treebank is available, along with documentation, from <https://universaldependencies.org/>.

1 Introduction

Scottish Gaelic is an low-resourced language which has hitherto lacked a robust parser, despite the recent development of the Annotated Reference Corpus of Scottish Gaelic (ARCOSG) (Lamb et al., 2016). Previous work (Batchelor, 2016) has leveraged ARCOSG to produce a medium-coverage categorial grammar but not a gold standard corpus that would enable the grammar to be properly evaluated. In this work we fill this gap by creating a dependency treebank for Scottish Gaelic of similar size to the existing treebanks in Irish (Lynn, 2016) and (Lynn and Foster, 2016) and Breton (Tyers and Ravishankar, 2018). An

important advantage of the dependency grammar approach is that the tools are better developed and less closely tied to English than for combinatory categorial grammar (CCG). Indeed, the universal dependencies (UD) framework has been developed to cover as wide a range of languages as possible, and recent CoNLL shared tasks have been explicitly multilingual, for example the 2018 task which focussed on extracting universal dependencies for 82 treebanks in 57 languages (Zeman et al., 2018). Given a corpus in CoNLL format, the `udpipe` package (Straka and Straková, 2017) can be used to train a tokeniser, a POS tagger and a dependency parser. In this work we will concentrate on the last of these and present preliminary results.

2 Scottish Gaelic

Scottish Gaelic, hereafter Gaelic, is a Celtic language of the Goidelic family closely related to Irish and Manx. It is spoken mainly in the Highlands and islands of Scotland, in the cities of the Central Belt and in Cape Breton in Canada. Its usage has been declining since the Middle Ages, when placename evidence attests its presence as far southeast as Fife and even East Lothian (Gullane, Innerwick and Ballencrieff all have Gaelic etymologies) and according to the UNESCO Atlas of the World’s Languages in Danger it is “definitely endangered” (Moseley, 2010). Lamb (2003) has published an accessible grammar of the language, but for a fuller account see Cox (2017) and for a short practical account focussing on contemporary usage see Ross et al. (2019).

The main electronic corpora for the language are ARCOSG and *Corpas na Gàidhlig* ‘Corpus of Gaelic’, part of the Digital Archive of Scottish Gaelic (DASG) (University of Glasgow, 2019).

The usual word order is VSO, but periphrastic

© 2019 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

constructions are very common and in contrast to Irish the usual way of expressing that something is taking place in the present is to use the verb *bi* and the verbal noun, for example *Tha mi a' dol* 'I am at going' for 'I am going'. We will discuss other differences from Irish, mainly that certain constructions are much more common in one language than the other, in more depth later on. There are two genders (masculine and feminine), three numbers (singular, plural and a separate dual form for nouns) and four cases (nominative, vocative, genitive and dative). Like the other Celtic languages Gaelic has conjugated prepositions such as *agam* 'at me' and *oirnn* 'on us'. These were already single words by the time of Old Irish (the 8th and 9th centuries CE) (Stifter, 2006). Both Irish and Scottish Gaelic exhibit cosubordination (Lamb, 2003). This is where the coordinator *agus* is followed by a nominal subject and an adjectival predicate or a small clause. Usually in Scottish Gaelic the cosubordinated clause follows the main clause, but it is not unusual in Irish to see it fronted. We will discuss options for handling this later on. Lastly, it is very common to express psychological states by a combination of the verb *bi*, a noun and two prepositional phrases, for example 'I love her' is *Tha gaol agam oirre* 'There is love at me on her'.

3 Related work

The first work on a dependency treebank for Gaelic was by Batchelor (2014) which predated the release of ARCOSG and was built from a tiny collection (82 sentences) of hand-picked sentences. At the same time Lamb and Danso presented a part-of-speech tagger for Gaelic based on ARCOSG (Lamb and Danso, 2014). Subsequently Tyers and Ravishankar (2018) have presented a Universal Dependencies treebank for Breton.

4 Corpus

ARCOSG is a corpus of 76 texts from a variety of genres, including conversations, sports commentary, fiction and news. Part of the context for the development of ARCOSG is given in (Lamb, 1999) where Lamb describes the development of the news reports and the language used on Radio nan Gàidheal. The texts have been part-of-speech tagged by hand according to a tagging scheme described in (2014) and based on the PAROLE tagset used by Uí Dhonnchadha (2009).

ARCOSG is made available in Brown Corpus format. This is converted into CoNLL-U format with a short Python script and the fine-grained part-of-speech tags mapped to the coarse-grained UD v2 tagset. A sample of the mapping is shown in Table 1. The LEMMA field in the treebank is populated using an extended version of the lemmatiser described in Batchelor (2016). The FEATS field is populated in the conversion process, largely following the feature set for Irish (Lynn et al., 2017). The corpus is broken into sentences on full stops and sentence boundaries corrected in manual postprocessing.

We use the texts in the following subcorpora: narrative, news, fiction, formal prose and popular writing. We have initially excluded the conversation and interview subcorpora because of the large fraction of single-token utterances. We also exclude the sports subcorpus for the moment as it largely consists of highly paratactic football commentary. Table 2 gives an overview. 752 sentences (25 593 tokens) remain to be added to the corpus and 30 sentences (1043 tokens) are awaiting a better treatment of cosubordination. Lastly there are five sentences in the narrative subcorpus which have a total of 5092 tokens between them.

5 Annotation

In this section we describe the process and look at some special cases in the light of the Irish and Breton treebanks.

5.1 Guidelines

We use the generic Universal Dependencies guidelines (Universal Dependencies contributors, 2016) and refer to the Irish UD corpus (Lynn and Foster, 2016) and our own list of special cases¹ in case of doubt. There is a single human annotator, an experienced adult learner of Gaelic, but the tagset used in ARCOSG is extremely rich, with well over 200 POS tags and marks for tense, case, number and gender and hence does a great deal of disambiguation. We follow Lynn et al. (2013) in marking up a small portion of the corpus (around 1000 tokens in this case) by hand and then iteratively training MaltParser (Nivre et al., 2006) with the standard settings on an annotated/checked part of the corpus and using it to parse the unchecked part. We

¹See `guidelines.md` and the documentation for the python conversion scripts in <https://github.com/colinbatchelor/gdbank/releases/tag/v0.2-alpha>

ARCOSG	UD	Comments	Examples
A*	ADJ	adjectives	
Dd	DET	determiners	<i>seo</i> ‘this’, <i>ud</i> ‘yon’
Dp*	DET	possessive pronouns	<i>mo</i> ‘my’
Dq	DET	determiners	<i>gach</i> ‘every’, <i>a h-uile</i> ‘every’
Cc	CCONJ	coordinators	<i>agus</i> ‘and’, <i>ach</i> ‘but’, <i>oir</i> ‘for’
Cs	SCONJ	subordinators	<i>ged</i> ‘although’, <i>nuair</i> ‘when’
F*	PUNCT	punctuation	
I	INTJ	interjections	<i>O</i> ‘oh’, <i>Seadh</i> ‘aye’, <i>Uill</i> ‘well’
M*	NUM	numbers (non-human series)	
Nc*	NOUN	common nouns	
Nf	ADP	fossilized nouns	<i>airson</i> ‘for’, <i>air feadh</i> ‘throughout’
Nn*	PROPN	proper nouns	
Nt	PROPN	toponyms	
Nv	NOUN	verbal nouns	<i>lorg</i> ‘going’
Pp*	PRON	personal pronouns	<i>mi</i> ‘I’
Pr*	ADP	conjugated prepositions	<i>orm</i> ‘on me’, <i>aige-san</i> ‘at him’
Q*	PART	particles	<i>a</i> (relativiser), <i>cha</i> (negative particle)
R*	ADV	adverbs	<i>a-mach</i> ‘out’, <i>cuideachd</i> ‘also’
Sa	PART	aspectual markers	<i>a’</i> , <i>ag</i> , <i>ri</i>
Sp	ADP	prepositions	<i>aig</i> ‘at’
Td*	DET	articles	<i>an</i> , <i>na</i> , <i>nan</i> ‘the’, ‘of the’
U*	PART	particles	<i>a</i> (adverbialiser), <i>a</i> (vocative)
Uf	NOUN	fossilized nouns	<i>’S urrainn</i> ‘can’, <i>’S dòcha</i> ‘maybe’
Um	ADP	‘than’	<i>na</i>
Uo	PART	numerical prefix	<i>a naoi</i> ‘nine’ (but see below for <i>h-</i> , <i>n-</i> , <i>t-</i>)
Up	PROPN	part of proper name	<i>Mac</i>
Uq	—	interrogatives	<i>Dè</i> ‘what’ (PRON), <i>Ciamar</i> ‘how’ (ADV)
V*	VERB	verb	
W*	AUX	copula	<i>B’</i> ‘was’, <i>’sann</i> (see below)
Xfe	—	foreign word	X for running text, NOUN where foreign noun
Y	NOUN	abbreviation	<i>Mgr</i> ‘Mr’, <i>a BhBC</i> ‘of the BBC’

Table 1: Mapping of the most important part-of-speech classes from ARCOSG to the UD coarse-grained tagset. The asterisk in A*, for example, indicates that all of the tags beginning with A map on to ADJ.

proofread the trees, add them to the training data, and iteratively improve the unchecked portion of the corpus. We keep trees which feature cosubordination in a separate file for future work.

5.2 Tokenisation

By and large we follow the tokenisation in ARCOSG but we do have to make some adjustments to match the UD scheme. Firstly, ARCOSG treats a number of multiword expressions such as place-names and the prepositions *ann an*, *anns an*, ‘in’, ‘in the’ and variants as a single token. We re-tokenise these on whitespace and assign, for the moment, the same part-of-speech tag to all of them. Secondly the prefixes *h-*, *n-* and *t-*, which are in-

separable parts of the word written without a hyphen in Irish, are treated as independent tokens with type Uo. These we unite with the tokens that follow them into a single token. There is a small number of multitoken compound prepositions that according to ARCOSG are three tokens, including a punctuation mark, for example *a-rèir* ‘according to’, which we collapse into a single token.

In addition we have to make some assumptions about reconstructing the original text, which is absent from the corpus. To this end we use the Gaelic Orthographic Conventions (GOC) ((SQA), 2009) for consistency in reconstructing spacing, but don’t apply any other corrections. We retain spaces after *a’*, *b’*, *d’*, *m’*, *th’* and *bh’*. If an elided *a’* or

Subcorpus	# sentences	Mean # tokens	Longest	Shortest
Fiction	397	17.0	61	2
Formal prose	120	24.9	113	5
News	167	22.2	52	7
Narrative	132	18.2	112	3
Popular writing	205	20.4	59	4

Table 2: Overview of the subset of ARCOSG in the treebank.

ag before a verbal noun is indicated by ', this is combined with the following token.

We make limited use of UD’s word–token distinction at present. The Irish and Breton treebanks differ on how to treat conjugated prepositions, with Breton dividing the single token *ganto* ‘with them’ into two words, *gant* and *o*, and Irish keeping *orm* ‘on me’ as a single word rather than dividing it into *ar* and *mé*. We follow the Irish example, but do divide tokens that have been tagged as fused tokens by ARCOSG. One example of this is *cúimhneam* ‘memory at me’, which has the POS `Ncsfn+Pr1s` (singular feminine common noun and first person singular conjugated preposition). In this case we divide it into the two words *cúimhne* and *agam*. There are currently ten examples of this in the treebank.

5.3 Personal names

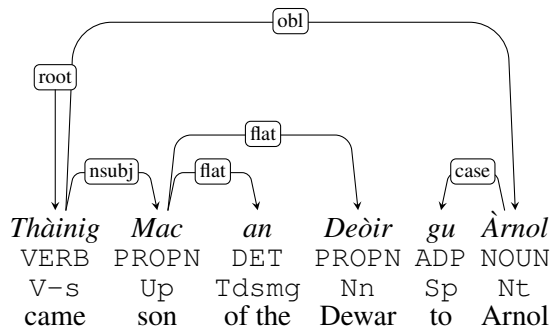


Figure 1: ‘Dewar came to Arnol’ (part of tree pw05_017), showing how compound proper names are handled.

We treat personal names, following the UD v2 guidelines, as a flat structure even in the case of surnames such as *Mac an Deòir* ‘Dewar’ that have internal grammatical structure (Fig. 1).

5.4 Copular constructions

An important use of the `fixed` relation is for the ‘dummy’ pronouns *e* and *ann* in the copular constructions ‘*s e*, *b’ e*, ‘*s ann* and *b’ ann*, which are sometimes written as a single word. Two examples

are given in Fig. 2. ‘*S e* (or in this case ‘*S i*) introduces an NP, and ‘*S ann* (or ‘*Sann*) is used for other sorts of constituent, here *an Ìle*, ‘in Islay’. The expression that follows the dummy pronoun as being the root of the clause, and the expression that follows that as the subject. It may be a definite NP, in which case we use the relation `nsubj`, or clausal, where we use `csubj:cop`. This is broadly similar to the Irish treebank except that Irish does not usually have the dummy pronouns.

5.5 The verbal noun and inversion structures

The verbal noun is intermediate between an archetypal verb and an archetypal noun. In some ways it behaves like a noun. It can be qualified by an adjective preceding it, and in progressive tenses, the NPs governed by a verbal noun were historically always in the genitive, however Ross et al. (2019) say that the genitive ‘is no longer required’ for indefinite singular nouns in these cases. Equally, it is part of a clause. In the Irish treebank verbal nouns are tagged as NOUN and treated as `xcomp` (externally-controlled clausal complements) of the controlling verb. Conversely in the Breton treebank they are tagged as VERB and treated as the root, with what would be the controlling verb in Irish treated as `aux` (an auxiliary). The approach where the verbal noun is tagged as a NOUN and the controlling verb as AUX is disallowed by the validation script. We have chosen to follow the Irish scheme, though these contrasting approaches, shown in Figure 3, both have their merits.

The two main ways of indicating the passive in Scottish Gaelic are synthetic: *Rugadh is thogadh mi* ‘I was born and raised’ (Fig. 4) and analytic: *Chaidh a’ nighean a lorg mu ochd uairean a-raoir*. ‘The girl was found about eight o’clock last night’. In the latter case we treat the subject *nighean* as a dependent of the verbal noun *lorg*. The root of the clause is the verb *Chaidh* ‘went’. This is shown in Fig. 5. This is different from the approach in

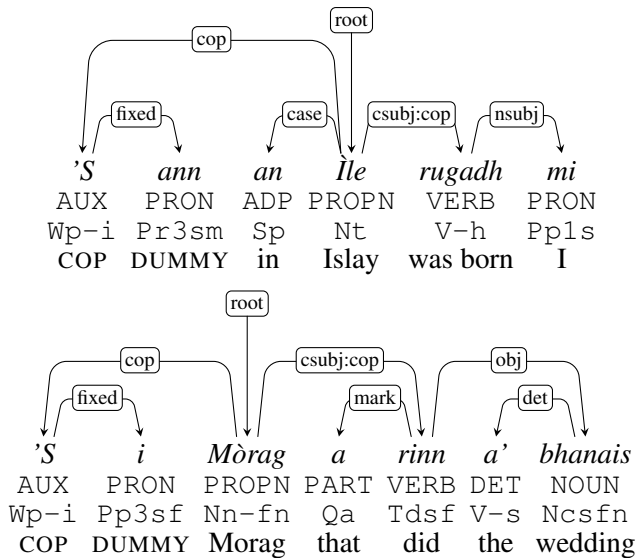


Figure 2: Dependency trees for ‘It was in Islay that I was born’ and ‘It was Morag who had the wedding’.

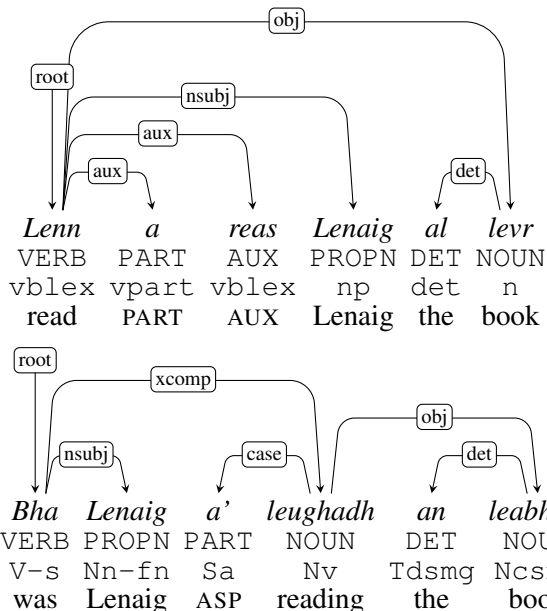


Figure 3: Dependency trees for ‘Lenaig read the book’ in the Breton (sentence ID grammar.vislsg.txt:28:654) and ‘Lenaig was reading the book’ in the Gaelic approach.

English where the word ‘was’ is treated as an auxiliary and the verb ‘found’ treated as the root, and of course from the Breton example above, but is identical to the other ‘inversion’ structures in Scottish Gaelic, where the object goes before the verbal noun (Ross et al., 2019), for example obligation: *Feumaidh mi cofaidh òl* ‘I must drink coffee’, or possibility: *Is urrainn dhuinn an duine a chuideachadh* ‘We can help the person’.

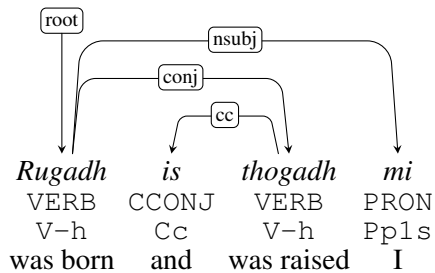


Figure 4: Dependency tree for ‘I was born and raised’ (analytic passive).

5.6 Numbers

As in Irish, there are two sets of cardinal numbers, one for people and the other for everything else. In ARCSOG *triùir* ‘three people’ is tagged as a noun (Ncsfn), and *triùir mhac* ‘three sons’ as Ncsfn Ncpmg. For this reason we treat *triùir* as the content word and *mhac* as a modifier. This is one of the many non-possessive constructions in Gaelic where a noun in the genitive modifies another noun, so we do not follow Breton in using the `nmod:poss` relation in these cases. Cardinal numbers on the other hand are tagged as Mc and we treat them as modifying their nominal head (`nummod`) unless they are the subject, object or oblique argument. See Figure 7 for examples.

5.7 Cosubordination

Cosubordination is an important grammatical phenomenon in Gaelic, found in all registers, and it is not clear how to cover it from the UD guidelines. Here is an example, a simplified version of sen-

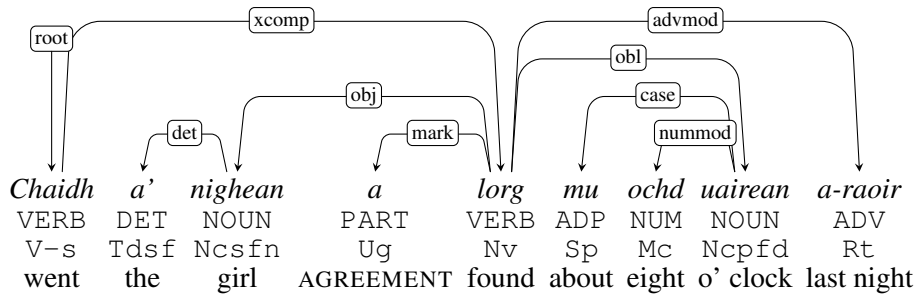


Figure 5: Dependency tree for ‘The girl was found about eight o’ clock last night’ (analytic passive, simplified version of ns10_021 in the as-yet unchecked part of the corpus).

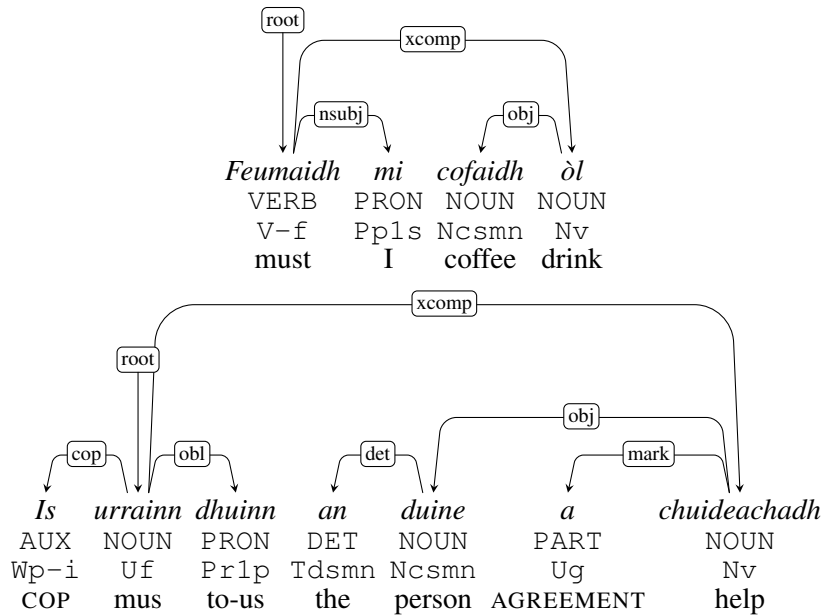


Figure 6: Dependency trees for inversion structures (Ross et al., 2019): ‘I must drink coffee’ and ‘We can help the person’.

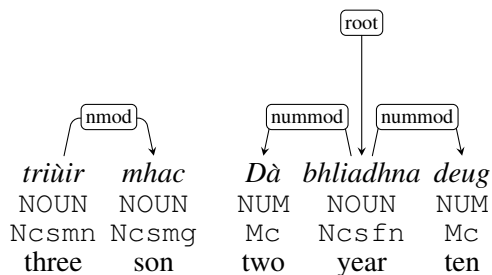


Figure 7: Contrasting treatments of personal numbers (left, from n05_009) and impersonal numbers (right, f08_036).

tence n05_004: *Chaidh e air chall ann an ceò, 's e 'g iasgach*. ‘He went missing in the fog when he was fishing’, or more literally ‘He went missing in the fog, and he fishing’.

We illustrate two approaches in Figure 8. The first one is to use the adnominal clause modifier (acl) relation as in the depictive example in the UD guidelines ‘She entered the room sad’. This

introduces a non-projective arc between the words *e* ‘he’ in each clause. The analogous sentence in English is projective because of SVO word order instead of VSO. It would perhaps be clearer to subclass the pertinent relations as *conj:cosub* and *acl:cosub*. The second approach is to assume that the word *bha* has been elided. This has two advantages: firstly clear UD guidelines, indicating that the correct thing to do is to treat *e* as a conjugate of the root verb *Chaidh* ‘went’ and connect it to the verbal noun with the *orphan* relation, and secondly maintaining a projective structure. And yet it is not obvious that cosubordination actually is ellipsis. One argument that it isn’t is that while in Gaelic the cosubordinate clause usually appears after the main clause, it is not uncommon in Irish to see preposed cosubordinate clauses, for example the phrase *Agus mé óg* ‘When I was young’.

The chief drawback of the depictive approach as opposed to the elliptical is non-projectivity, but we

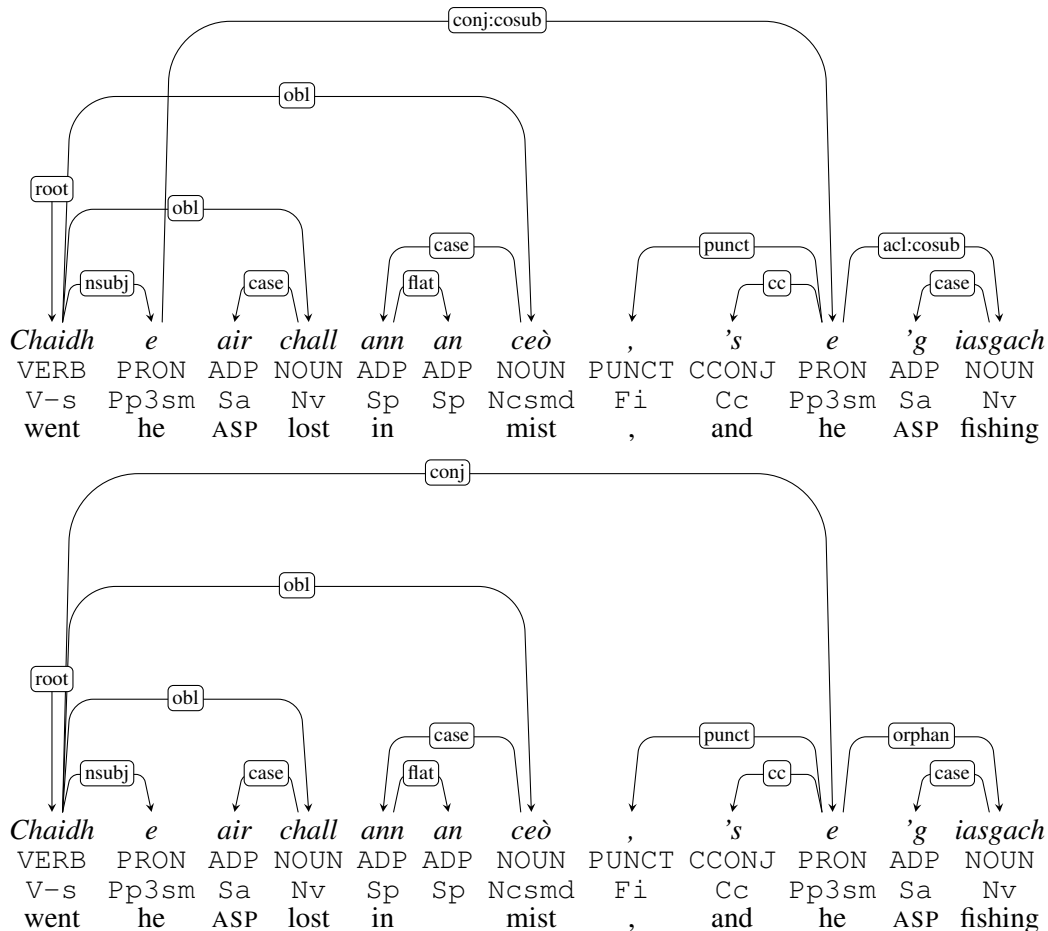


Figure 8: Two approaches to cosubordination. (above) Depictive (below) elliptical.

feel that it is better in terms both of representing the language as it is found in all registers, and in the specificity of the relation used: `acl` as opposed to the more generic `orphan`. A full investigation of how common non-projectivity is in the treebank as a whole should help decide how much weight to place on projectivity as a criterion for annotation choices.

6 Experiments

Transition system	LAS
projective	0.796 [0.752, 0.839]
swap (non-projective)	0.789 [0.747, 0.825]
link2 (non-projective)	0.792 [0.750, 0.835]

Table 3: Labelled-attachment scores (LAS) for parsing the treebank with different transition systems for `udpipe`'s `parsito` parser. The LAS are the mean values from ten-fold cross-validation. There were 1021 sentences and 20031 words.

In this section we examine briefly how learnable the annotation scheme is by a dependency parser and the effects of different parsing algorithms. A

fuller account awaits a larger treebank, which on average will have longer sentences. Because of the small size of the corpus we use ten-fold cross validation. We use `udpipe`'s `parsito` parser and `MaltEval` (Nilsson and Nivre, 2008), both with the default settings. Table 3 gives the ten-fold cross-validated labelled-attachment scores (LAS) for the three transition systems, one projective and two non-projective. We find that the default, projective transition system performs best, with a mean LAS of 0.796, but the three are very similar, and the scores may be flattered by the relatively short mean sentence length (19.6 words). These scores are comparable to the scores for much larger treebanks given by Nivre and Fang (2017).

7 Conclusions and future work

This is the first reasonably-sized dependency treebank of Scottish Gaelic and the first demonstration of a fast parser for the language. The treebank can also be used for tokenisation and part-of-speech tagging using `udpipe`, though as mentioned before

the tokenisation scheme is different from that used in ARCOSG.

The treebank presented here is not quite ready for an official release, despite its passing the validation script. Nonetheless a significant part of the treebanking process, annotating a part-of-speech corpus with lemmas, universal part-of-speech tags, syntactic features and dependency relations, has been achieved. A substantial part of ARCOSG has not been covered in this work—the conversation, sport and interview subcorpora, and five of the texts in the narrative subcorpus. There are also over 750 trees that are not yet in the treebank but they will be processed in due course. Lastly, given the close relation between the two languages and that the annotation scheme presented here is as close as possible to the Irish one, it would be very interesting to repeat the parsing experiments on a combined Irish and Scottish Gaelic corpus.

References

- Batchelor, Colin. 2014. gdbank: The beginnings of a corpus of dependency structures and type-logical grammar in Scottish Gaelic. In *Proceedings of the First Celtic Language Technology Workshop*, pages 60–65, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.
- Batchelor, Colin. 2016. Automatic derivation of categorical grammar from a part-of-speech-tagged corpus in Scottish Gaelic. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 6 : CLTW*, Paris, France, July.
- Cox, Richard A. V. 2017. *Geàrr-Ghràmar na Gàidhlig*. Clann Tuirc, Tigh a' Mhaide, Ceann Drochaid, Alba.
- Lamb, William and Samuel Danso. 2014. Developing an Automatic Part-of-Speech Tagger for Scottish Gaelic. In *Proceedings of the First Celtic Language Technology Workshop*, pages 1–5, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.
- Lamb, William, Sharon Arbuthnot, Susanna Naismith, and Samuel Danso. 2016. Annotated Reference Corpus of Scottish Gaelic (ARCOSG), 1997–2016 [dataset]. Technical report, University of Edinburgh; School of Literatures, Languages and Cultures; Celtic and Scottish Studies. <http://dx.doi.org/10.7488/ds/1411>.
- Lamb, William. 1999. A diachronic account of Gaelic news-speak: The development and expansion of a register. *Scottish Gaelic Studies*, XIX:141–171.
- Lamb, William. 2003. *Scottish Gaelic, 2nd edn*. Lincom Europa, Munich, Germany.
- Lynn, Teresa and Jennifer Foster. 2016. Universal Dependencies for Irish. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 6 : CLTW*, Paris, France, July.
- Lynn, Teresa, Jennifer Foster, and Mark Dras. 2013. Working with a small dataset - semi-supervised dependency parsing for Irish. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–11, Seattle, Washington, USA, October. Association for Computational Linguistics.
- Lynn, Teresa, Jennifer Foster, and Mark Dras. 2017. Morphological Features of the Irish Universal Dependency Treebank. In *TLT 2017: Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories*, Aachen, Germany.
- Lynn, Teresa. 2016. *Irish Dependency Treebanking and Parsing*. Ph.D. thesis, Dublin City University and Macquarie University.
- Moseley, Christopher, editor. 2010. *Atlas of the World's Languages in Danger*. UNESCO Publishing, Paris, 3rd edition.
- Naismith, Susanna and William Lamb. 2014. Scottish Gaelic Part-of-Speech Annotation Guidelines. Celtic and Scottish Studies, University of Edinburgh.
- Nilsson, Jens and Joakim Nivre. 2008. MaltEval: an evaluation and visualization tool for dependency parsing. In *LREC 2008*.
- Nivre, Joakim and Chiao-Ting Fang. 2017. Universal dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden, May. Association for Computational Linguistics.
- Nivre, Joakim, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).
- Ross, Susan, Mark McConville, Wilson McLeod, and Richard Cox. 2019. Stiùireadh Gràmair. <https://dasg.ac.uk/grammar/grammar.pdf>, accessed 23 May 2019.
- (SQA), Scottish Qualifications Authority. 2009. *Gaelic Orthographic Conventions*. Scottish Qualifications Authority, Dalkeith.
- Stifter, David. 2006. *Sengoidelc: Old Irish for beginners*. Syracuse University Press, Syracuse, New York.

- Straka, Milan and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipes. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.
- Tyers, Francis M. and Vinit Ravishankar. 2018. A prototype dependency treebank for Breton. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018 : Volume 1 : Articles longs, articles courts de TALN*, pages 197–204, Rennes, France, May.
- Uí Dhonnchadha, Elaine. 2009. *Part-of-speech tagging and partial parsing for Irish using finite-state transducers and constraint grammar*. Ph.D. thesis, Dublin City University.
- Universal Dependencies contributors. 2016. UD Guidelines, <https://universaldependencies.org/guidelines.html>. Accessed 9 July 2019.
- University of Glasgow. 2019. Corpas na Gàidhlig, Digital Archive of Scottish Gaelic (DASG), <https://dasg.ac.uk/corpus/>. Accessed 9 July 2019.
- 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 *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October. Association for Computational Linguistics.