Universal Dependencies for Manx Gaelic

Kevin Scannell

Saint Louis University &
Acadamh na hOllscolaíochta Gaeilge

Manx Gaelic

- **Manx Gaelic** is one of the three Q-Celtic (or Goidelic) languages, along with Irish and Scottish Gaelic.
- It is spoken primarily on the **Isle of Man**, located in the Irish Sea between Ireland and Scotland; see Figure 1.
- The language gradually fell out of widespread use during the 19th and 20th centuries, but **the number of speakers is now growing** thanks to language revitalization efforts, including a Manx-medium school on the island.
- Very little language technology exists for Manx; we believe the corpus presented here is **the first annotated corpus of any kind** for the language.

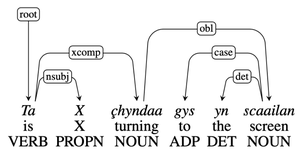
Fig. 1: The Isle of Man. Map by G. Bosanko (CC BY-SA 3.0)

Universal Dependencies

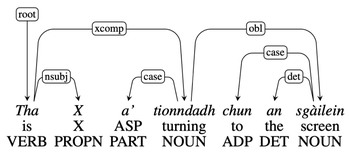
- Our main deliverable is a **new treebank for Manx**, annotated according to version 2 of the Universal Dependencies (UD) guidelines [5, 6]
- The treebank consists of **291 sentences (about 6000 tokens)** randomly sampled from a comprehensive web corpus of Manx containing more than 8 million words
- UD treebanks now exist for **five of the six Celtic languages**, with Manx joining Irish [3, 4], Breton [8], Scottish Gaelic [1], and Welsh [2] — only Cornish remains to be done.

The Q-Celtic Family

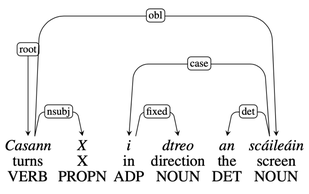
The grammar of Manx is very close to both Irish and Scottish Gaelic, sharing features such as VSO word order, initial consonant mutations, inflected prepositions, and extensive use of the verbal noun. Consider the following **trilingual example**, meaning "X turns toward the screen":



(a) Manx



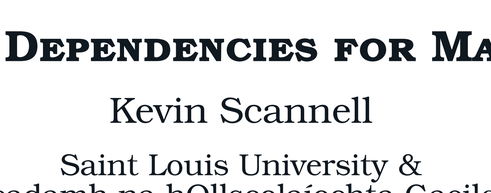
(b) Scottish Gaelic



(c) Irish

Annotation Details

- All of the Celtic languages have so-called **inflected prepositions**, e.g. Manx *lhiam* "with me" (Ir. *liom*, Sc.G *leam*). We diverge from the Irish and Scottish treebanks by decomposing into the constituent preposition and pronoun (as in the example tree below).
- **Indirect objects** do not occur in Irish or Scottish Gaelic, nor, to the best of our knowledge, in traditional Manx texts. There are examples of indirect objects in revived Manx, however, presumably under the influence of English.
- **Verbal nouns** play an important role in all of the Celtic languages but particularly so in Manx. We tag verbal nouns as `NOUN` and treat them syntactically as `xcomp` of the surrounding verb, as in the Irish and Scottish Gaelic treebanks.
- **Objects follow the verbal noun** much more frequently in Manx than in the other Gaelic languages.
- An unusual feature of Manx is its tendency to express past and future tense using **the verbal noun together with the verb *jean*** ("do", Ir. *déan*, Sc.G. *dèan*), where the other Gaelic languages would commonly use an inflected form of the verb itself. One sees this construction even with the verbal noun *jannoo* corresponding to *jean* itself.



S'treisht ***lhiam*** *dy vel shiu braew* "I hope that you are well"

Parsing Experiments

As an evaluation of the treebank, we performed several parsing experiments using UDPipe [7] with the default settings (hidden layer size of 200 and the projective parsing algorithm). The results of all experiments are reported in Table 1.

- We evaluated the UDPipe lemmatizer, POS tagger, and parser via **10-fold cross validation** on the Manx treebank itself. In the first experiment we parsed **plain text input**, which is to say we made no use of the gold standard tokens, lemmas, or POS tags.
- In the second set of experiments, we again evaluated the parser via **10-fold cross validation**, but this time we gave the tagger access to the gold-standard tokenization for making its predictions, and gave the parser access to the **gold tokens, lemmas, and POS tags** for making its predictions.
- Finally, we trained **cross-lingual delexicalized parsers** on the Irish and Scottish Gaelic treebanks and evaluated those models directly on the Manx treebank by providing the gold-standard Manx POS tags as the (only) input.

Results

Model	Lemma	POS	UAS	LAS
Manx 10-fold (plain text input)	87.43	89.06	72.83	65.20
Manx 10-fold (std dev)	1.55	1.13	2.74	2.96
Manx 10-fold (gold inputs)	90.40	92.19	82.61	76.29
Manx 10-fold (std dev)	1.22	1.11	1.82	2.30
Irish delexicalized	-	-	40.43	31.59
Scottish delexicalized	-	-	28.71	19.66

Table 1: F_1 scores for Manx lemmatization, POS tagging, and dependency parsing. Parser accuracy is reported as both unlabeled (UAS) and labeled (LAS) attachment scores.

Conclusions

- We created a new corpus for Manx Gaelic consisting of 291 sentences, annotated according to version 2 of the UD guidelines.
- We trained a dependency parser on the corpus and evaluated it using 10-fold cross-validation, obtaining encouraging results.
- We also experimented with delexicalized cross-lingual models using the Irish and Scottish Gaelic treebanks, with disappointing results.
- We believe this argues in favor of under-resourced language groups investing energy primarily into monolingual treebank development.

Acknowledgements

I created the treebank while visiting Acadamh na hOllscolaíochta Gaeilge in Carna, Co. na Gaillimhe as a Fulbright Scholar. All of the work was done during the COVID-19 lockdown in Ireland. Thanks to Teresa Lynn and Colin Batchelor for their invaluable work on the Irish and Scottish Gaelic treebanks. Thanks also to Fran Tyers and to the anonymous reviewers for many helpful comments and suggestions.

References

[1] Colin Batchelor. "Universal dependencies for Scottish Gaelic: syntax". In: *Proceedings of the Celtic Language Technology Workshop*. Dublin, 2019, pp. 7–15.

[2] Johannes Heinecke and Francis M. Tyers. "Development of a Universal Dependencies treebank for Welsh". In: *Proceedings of the Celtic Language Technology Workshop*. Dublin, 2019, pp. 21–31.

[3] Teresa Lynn and Jennifer Foster. "Universal Dependencies for Irish". In: *Proceedings of the 2nd Celtic Language Technology Workshop*. Paris, 2016.

[4] Teresa Lynn, Jennifer Foster, and Mark Dras. "Morphological features of the Irish Universal Dependency treebank". In: *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories*. Bloomington, Indiana, 2017.

[5] Joakim Nivre et al. "Universal Dependencies v1: A multilingual treebank collection". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 1659–1666.

[6] Joakim Nivre et al. "Universal Dependencies v2: An evergrowing multilingual treebank collection". In: *arXiv preprint arXiv:2004.10643* (2020).

[7] Milan Straka and Jana Straková. "Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe". In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. 2017, pp. 88–99.

[8] Francis M. Tyers and Vinit Ravishankar. "A prototype dependency treebank for Breton". In: *Actes de la conférence Traitement Automatique de la Langue-Naturelle, TALN*. Vol. 1. 2018, pp. 197–204.