Universal Dependencies for Manx Gaelic

Kevin P. Scannell
Department of Computer Science
Saint Louis University
St. Louis, Missouri, USA
kscanne@gmail.com

Abstract

Manx Gaelic is one of the three Q-Celtic languages, along with Irish and Scottish Gaelic. We present a new dependency treebank for Manx consisting of 291 sentences and about 6000 tokens, annotated according to the Universal Dependency (UD) guidelines. To the best of our knowledge, this is the first annotated corpus of any kind for Manx. Our annotations generally follow the conventions established by the existing UD treebanks for Irish and Scottish Gaelic, although we highlight some areas where the grammar of Manx diverges, requiring new analyses. We use 10-fold cross validation to evaluate the accuracy of dependency parsers trained on the corpus, and compare these results with delexicalized models transferred from Irish and Scottish Gaelic.

1 Introduction

Manx Gaelic, spoken primarily on the Isle of Man, is one of the three Q-Celtic (or Goidelic) languages, along with Irish and Scottish Gaelic. Although the language fell out of widespread use during the 19th and 20th centuries, it has seen a vibrant revitalization movement in recent years. The number of speakers is now growing, thanks in part to a Manx-language primary school on the island. The language also has a strong online presence relative to the size of the community.

Several Manx dictionaries were published in the 19th and 20th centuries, and a number of these have been digitized in recent years and published online.[1] The full text of the Bible, originally published in the 18th century, is also available and provides an easily-accessible source of parallel text.[2] In terms of spoken language, the Irish Folklore Commission recorded fluent speakers on the Isle of Man in 1948, and those recordings are now available digitally as well.[3]

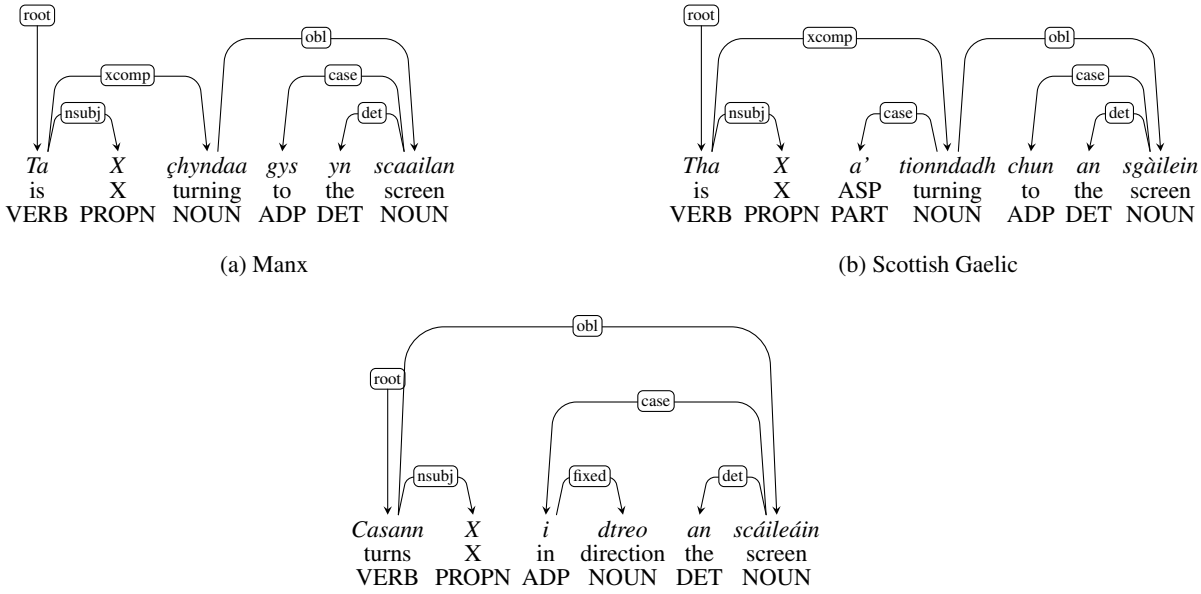
Outside of these corpus resources, very little advanced language technology exists for Manx. The lack of a lemmatizer and part-of-speech tagger makes it more difficult for language learners and linguistic researchers to search corpora of Manx texts. These difficulties are compounded by spelling variations in the traditional texts as well as the system of initial mutations. With these challenges in mind, we set out to lay the foundation for Manx language technology by developing a lemmatizer, tagger, and dependency parser. The Universal Dependencies framework (Nivre et al., 2016; Nivre et al., 2020) was ideal for our purposes, offering a unified annotation scheme which opens up the possibility of cross-lingual analysis, as well as a rich ecosystem of computational tools.

In §2, we present our new treebank for Manx, annotated according to the Universal Dependencies (v2) guidelines and released under an open source license. There are now UD treebanks for five of the six Celtic languages, with Manx joining Irish (Lynn and Foster, 2016; Lynn et al., 2017), Breton (Tyers and Ravishankar, 2018), Scottish Gaelic (Batchelor, 2019), and Welsh (Heinecke and Tyers, 2019) — only Cornish remains to be done. The Manx treebank is relatively small, consisting of 291 sentences and about 6000 tokens, and was annotated entirely by hand. Because of the close linguistic relationship between Manx and its sister languages of Irish and Scottish Gaelic, we were able to refer to the existing

[1] See https://sites.google.com/view/gailck-hasht/fockleyryn.

[2] See http://bible.learnmanx.com/.

[3] See https://www.imuseum.im/search/collections/archive/mnh-museum-676861.html.



(a) Manx

(b) Scottish Gaelic

(c) Irish

Figure 1: A trilingual example; "X turns to the screen"

UD treebanks for these languages while constructing the corpus. That said, the grammar of Manx does differ in a few important ways, and we discuss some of these divergent analyses in §3.

In §4, we evaluate the accuracy of several dependency parsing models on the Manx corpus. First, we trained UDPipe models using the corpus itself via 10-fold cross validation, achieving labeled accuracy scores of 65.20 on plain text input, and 76.29 when the parser was given access to the gold POS tags. We then evaluated the performance of delexicalized models trained on the Irish and Scottish Gaelic treebanks on the Manx corpus but were unable to achieve comparable results despite the close linguistic relationships. Other approaches to cross-lingual parsing such as annotation projection might give better results in the future, but these would depend on the development of machine translation engines and large parallel corpora between the Gaelic languages.

2 Corpus Development and Annotation

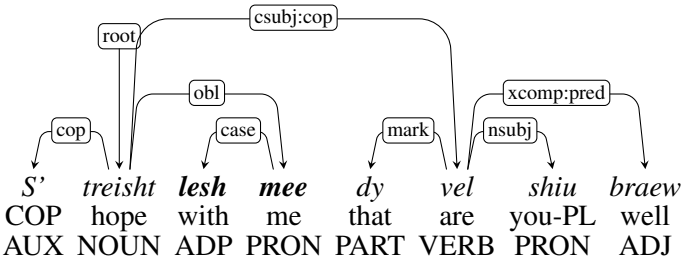
The grammar of Manx is very close to both Irish and Scottish Gaelic, sharing features such as VSO word order, initial consonant mutations, inflected prepositions, and extensive use of the verbal noun. In places where Irish and Scottish Gaelic diverge, Manx is typically more closely allied with Scottish Gaelic, e.g. in the use of a periphrastic construction involving the verbal noun to express the present tense; see Figure 1. For reasons of space we make no attempt to survey the grammar here, referring the interested reader to Draskau (2008) for a general overview. In terms of Universal Dependencies *per se*, we refer to Lynn and Foster (2016) and Batchelor (2019), as well as the detailed guidelines for those languages on the Universal Dependencies web site.[4]

The sentences in the treebank were taken from a web-crawled corpus of Manx consisting of more than eight million words of text. This corpus contains virtually all non-trivial Manx language texts on the open web, and therefore the treebank presented here is as close as possible to a "random sample" of Manx on the web. As such, the Bible is heavily represented, but there are also many sentences from modern sources: the Manx Wikipedia, news stories from Manx Radio, blog posts, etc.[5]

The web corpus was segmented by sentences and shuffled; then, 300 random sentences were chosen for annotation (nine of these require further analysis and did not make it into the initial release of the

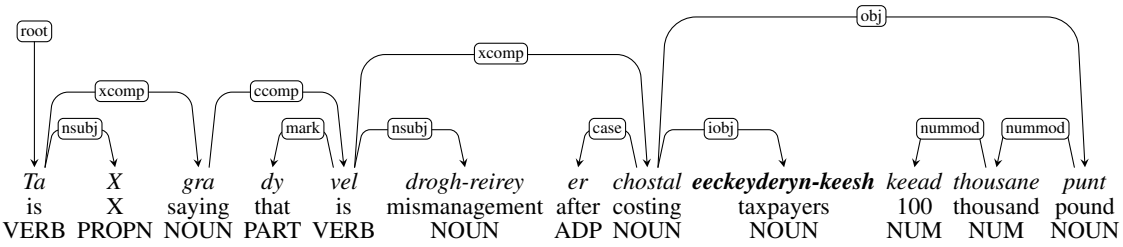
[4] `https://universaldependencies.org/`

[5] This design choice is not without controversy; there are significant differences between the traditional language and the Manx of the modern revival. By including samples of both in the corpus, we hope to develop tools that will be effective in processing both varieties.



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

Figure 2: Example of a decomposed inflected preposition



"X says that mismanagement has cost taxpayers one hundred thousand pounds"

Figure 3: Example of an indirect object in Manx. In Irish and Scottish Gaelic, where there are no indirect objects, verbs meaning "to cost" would express this argument via an oblique PP.

treebank). The UD part-of-speech tags and dependency annotations were added manually by the present author. We have not added detailed morphological information, although we plan to do this in future releases. Since there are only about 6000 tokens in the final corpus, we have released the data as a single file (without splitting into train/dev/test sets), per the recommendation of the UD maintainers.

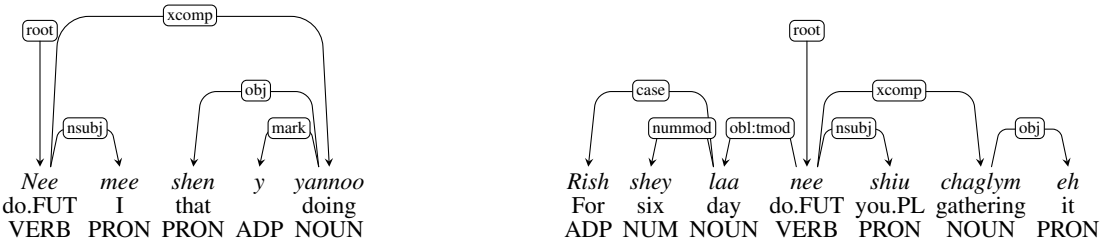
3 Comparison with Irish and Scottish Gaelic

In this section we highlight some differences between the Manx treebank and the Irish and Scottish Gaelic treebanks.

All of the Celtic languages have so-called "inflected prepositions," e.g. Manx *lhiam* "with me" (Ir. *liom*, Sc.G *leam*), Manx *lhiat* "with you" (Ir. and Sc.G. *leat*), etc. These are treated as single tokens in the Irish and Scottish treebanks, while we have instead followed the Breton (Tyers and Ravishankar, 2018) and Welsh (Heinecke and Tyers, 2019) treebanks by decomposing these into their constituent preposition and pronoun. Figure 2 shows the parse for sentence `aa_239` from the treebank: *S'treisht lhiam dy vel shiu braew* ("I hope that you are well"), in which the word *lhiam* has been decomposed into *lesh* "with" plus *mee* "me". Treating inflected prepositions in this way helps address ambiguities in some third person singular cases (e.g. *lesh* is also the inflected form meaning "with him"), and we hope it also makes it easier to learn syntactic generalizations since these prepositions behave like standard prepositional phrases.

Indirect objects do not occur in Irish or Scottish Gaelic, and, to the best of our knowledge, do not appear in traditional Manx texts either. There are examples of indirect objects in revived Manx, however, presumably under the influence of English. See Figure 3, which is a simplified version of sentence `aa_093` in the treebank.

Verbal nouns play an important role in all of the Celtic languages but particularly so in Manx. We follow the lead of Irish, Scottish Gaelic, and Welsh by tagging verbal nouns as `NOUN` and treating them syntactically as `xcomp` of the surrounding verb (in contrast with the Breton treebank which treats them as full-fledged verbs). Objects can come before or after the verbal noun in all three Gaelic languages, although there is a greater tendency for them to follow the verbal noun in Manx. An unusual feature of



(a) "I will do that"

(b) "For six days you will gather it"

Figure 4: Examples of verbal nouns as `xcomp` of *jean*

Manx is the frequency with which verbal nouns are used together with the verb *jean* ("do", Ir. *déan*, Sc.G. *dèan*) to express past or future tense, 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, as in the first example in Figure 4. One consequence of this is a much higher overall frequency for the lemmas *jean* and *jannoo* in the Manx treebank (occurring once out of every 64 tokens on average) than for the corresponding lemmas in Irish or Scottish Gaelic (occurring once per 197 and 224 tokens, respectively).

4 Parsing Experiments

All of the experiments in this section make use of the latest version[6] of UDPipe (Straka and Straková, 2017) with the default settings (hidden layer size of 200 and the projective parsing algorithm).

We began by evaluating the UDPipe lemmatizer, POS tagger, and parser via 10-fold cross validation on the Manx treebank itself. In each iteration, 261 sentences were used for training and 30 sentences were used for testing; the same splits were preserved across the experiments. In the first set of experiments we evaluated on plain text input, which is to say we made no use of the gold standard tokens, lemmas, or POS tags. The F_1 scores and standard deviations over the ten splits are presented in Table 1, with the parser accuracy reported as both unlabeled (UAS) and labeled (LAS) attachment scores. The results are comparable to the corresponding scores for Irish and Scottish Gaelic, despite the much smaller size of our corpus. Since those treebanks are large enough to have a standard train/test split, we used those in place of 10-fold cross validation, obtaining UAS of 77.91 and LAS of 68.91 for Irish, and UAS of 71.55, LAS of 63.86 for Scottish Gaelic (again making no use of gold-standard inputs).

Model	**Lemma**	**POS**	**UAS**	**LAS**
Manx 10-fold cross validation	87.43	89.06	72.83	65.20
Manx 10-fold standard deviation	1.55	1.13	2.74	2.96

Table 1: Manx lemmatization, part-of-speech tagging, and dependency parsing evaluated on plain text input

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. Doing this allows a fairer comparison with the results of the delexicalized cross-lingual parsing experiments below. The F_1 scores and standard deviations are presented in the first two rows of Table 2. We ran the analogous experiments for Irish and Scottish Gaelic, again using the standard train/test splits, obtaining UAS of 80.57 and LAS of 73.71 for Irish, and UAS of 81.60, LAS of 76.90 for Scottish Gaelic.

Given the close relationship between the three Gaelic languages, we thought it worth exploring the possibility of cross-lingual parsing, particularly in this context of a severely under-resourced language

[6]The version on the master branch on GitHub as of 1 August 2020: `https://github.com/ufal/udpipe/commit/a2e2ffa24fc8d9c487073dfa17472699f8c59134`.

and its better-resourced neighbors. To this end, we trained delexicalized parsers on the Irish and Scottish Gaelic UD treebanks and evaluated those models directly on the Manx treebank by providing the gold-standard Manx POS tags as the input. The results are presented in the third and fourth rows of Table 2. The scores were poor in comparison with the monolingual 10-fold cross-validation, and indeed worse than comparable results for language pairs that one might expect to be more difficult; see, for example, (Agić et al., 2014; Aepli and Clematide, 2018), and notably (Tiedemann, 2015) which includes results for a delexicalized Irish model transferred to several other European languages.

Model	**Lemma**	**POS**	**UAS**	**LAS**
Manx 10-fold cross validation	90.40	92.19	82.61	76.29
Manx 10-fold standard deviation	1.22	1.11	1.82	2.30
Irish delexicalized	-	-	40.43	31.59
Scottish delexicalized	-	-	28.71	19.66

Table 2: Evaluation scores for Manx lemmatization, part-of-speech tagging, and dependency parsing using gold-standard inputs

5 Conclusion and Future Work

In §2 and §3, we presented a new corpus for Manx Gaelic annotated according to version 2 of the Universal Dependencies guidelines, and gave several examples where the grammar of Manx differed from Irish and Scottish Gaelic.

In the previous section, we evaluated a lemmatizer, tagger, and dependency parser trained on the Manx corpus, and obtained encouraging results. We also trained delexicalized cross-lingual models on the Irish and Scottish Gaelic treebanks, but their performance on Manx was relatively poor. These results, together with the relative ease with which we were able to annotate almost 300 sentences,[7] seem to argue in favor of under-resourced language groups investing energy primarily into monolingual treebank development. It remains to be seen whether the parsing scores can be improved by augmenting the Manx training data with trees obtained by other means, e.g. annotation projection (Yarowsky et al., 2001; Tyers et al., 2018), making use of some existing parallel texts between Irish and Manx. Unfortunately, there are no machine translation engines from Irish or Scottish Gaelic into Manx, so cross-lingual methods like those described by Tiedemann et al. (2014) are not available to us.

We also plan to continue adding to the treebank. All of the sentences in the initial release were annotated manually, but given the strong results above we should be able to accelerate development by post-editing the output of the UDPipe parser. We will also add morphological information to bring our treebank in line with the structure of the Irish and Scottish Gaelic versions.

Finally, we hope to develop a small trilingual parallel treebank with Irish and Scottish Gaelic consisting of sentences from the respective translations of *Alice in Wonderland* (the only substantial text available in all three languages, the Bible excluded), and through that work further harmonize the annotation guidelines across the Gaelic language family.

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.

I am grateful to the staff at the Acadamh in Carna for their hospitality during my visit, to the Fulbright Program for the financial support which made it possible, and to Saint Louis University for a much-needed sabbatical leave.

Thanks to the anonymous reviewers for helpful comments and suggestions.

Finally, thanks to Teresa Lynn and Colin Batchelor for their invaluable work on the Irish and Scottish Gaelic treebanks, without which the present work would not have been possible.

[7] About three weeks for one annotator working full-time.

References

Noëmi Aepli and Simon Clematide. 2018. Parsing approaches for Swiss German. In *SwissText 2018*.

Željko Agić, Jörg Tiedemann, Kaja Dobrovoljc, Simon Krek, Danijela Merkler, and Sara Može. 2014. Cross-lingual dependency parsing of related languages with rich morphosyntactic tagsets. In *EMNLP 2014 Workshop on Language Technology for Closely Related Languages and Language Variants*.

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

Jennifer Draskau. 2008. *Practical Manx*. Liverpool University Press.

Johannes Heinecke and Francis M. Tyers. 2019. Development of a Universal Dependencies treebank for Welsh. In *Proceedings of the Celtic Language Technology Workshop*, pages 21–31, Dublin. European Association for Machine Translation.

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

Teresa Lynn, Jennifer Foster, and Mark Dras. 2017. Morphological features of the Irish Universal Dependency treebank. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, Bloomington, Indiana.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Milan Straka and Jana Straková. 2017. 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*, pages 88–99.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Eighteenth Conference on Computational Natural Language Learning (CoNLL 2014)*.

Jörg Tiedemann. 2015. Cross-lingual dependency parsing with Universal Dependencies and predicted PoS labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 340–349.

Francis M. Tyers and Vinit Ravishankar. 2018. A prototype dependency treebank for Breton. In *Actes de la conférence Traitement Automatique de la Langue-Naturelle, TALN*, volume 1, pages 197–204.

Francis Tyers, Mariya Sheyanova, Aleksandra Martynova, Pavel Stepachev, and Konstantin Vinogorodskiy. 2018. Multi-source synthetic treebank creation for improved cross-lingual dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 144–150, Brussels, Belgium, November. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8.