gdbank: The beginnings of a corpus of dependency structures and type-logical grammar in Scottish Gaelic

Colin Batchelor
Royal Society of Chemistry, Cambridge, UK CB4 0WF
batchelorc@rsc.org

Abstract

We present gdbank, a small handbuilt corpus of 32 sentences with dependency structures and categorial grammar type assignments. The sentences have been chosen to illustrate as broad a range of the unusual features of Scottish Gaelic as possible, particularly nouns being used to represent psychological states where more thoroughly-studied languages such as English and French would prefer a verb, and prepositions marking aspect, as is also seen in Welsh and, for example, Irish Gaelic. We provide hand-built dependency trees, building on previous work on Irish Gaelic and using the Universal Dependency Scheme. We also provide a tentative categorial grammar account of the words in the sentences, based largely on previous work on English.

1 Introduction

Scottish Gaelic (usually hereafter Gaelic) is a Celtic language, rather closely related to Irish, with around 59,000 speakers as of the last UK census in 2011. As opposed to the situation for Irish Gaelic (Lynn et al., 2012a; Lynn et al., 2012b; Lynn et al., 2013; Lynn et al., 2014) there are no treebanks or tagging schemes for Scottish Gaelic, although there are machine-readable dictionaries and databases available from Sabhal Mòr Ostaig. A single paper in the ACL Anthology (Kessler, 1995) mentions Scottish Gaelic in the context of computational dialectology of Irish. There is also an LREC workshop paper (Scannell, 2006) on machine translation between Irish and Scottish Gaelic. Elsewhere in the Celtic languages, Welsh has an LFG grammar (Mittendorf and Sadler, 2005) but no treebanks. For Breton there is a small amount of work on morphological analysis and Constraint-Grammar-based machine translation (Tyers, 2010). Recent work on the grammar of Scottish Gaelic (for example (Adger and Ramchand, 2003; Adger and Ramchand, 2005), but there are many more examples) has largely focussed on theoretical syntactic issues somewhat distant from the more surfacy approaches popular in the field of natural language processing. This paper explores grammatical issues in Scottish Gaelic by means of dependency tagging and combinatory categorial grammar (CCG), which we see as complementary approaches. As such it is explicitly inspired by CCGbank (Hockenmaier and Steedman, 2007), which consists of dependency structures and CCG derivations for over 99% of the Penn Treebank. It is hoped that this corpus will be a useful adjunct to currently on-going work in developing a part-of-speech tagset and tagger for Scottish Gaelic.

Section 2 describes how the corpus was prepared, sections 3 and 4 give some context for the dependency scheme and categorial grammar annotations respectively, and the main part of the paper is section 5, which deals with language-specific features of the corpus.

2 Preparing the corpus

The corpus consists of a small handbuilt selection of sentences from the transcripts of *An Litir Bheag*, which is a weekly podcast from the BBC written by a native speaker and aimed at Gaelic learners, example sentences from (Lamb, 2003), the BBC's online news in Gaelic and the Gaelic column in the Scotsman newspaper. In order to illustrate as much of the interesting points of Scottish Gaelic as possible,

Proceedings of the First Celtic Language Technology Workshop, pages 60–65,
Dublin, Ireland, August 23 2014.

Dependency	Example	Gloss	GR
det	*gach latha* (det latha gach)	every day	det
dobj	*Ithidh i ìm* (dobj Ithidh ìm)	She eats butter	dobj
adpmod	*Tha piseag agam* (adpmod Tha agam)	I have a kitten	ncmod
adpobj	*às an eilean* (adpobj às eilean)	from the island	dobj
nsubj	*Tha mi a' dol* (Tha mi)	I am coming	ncsubj
prt	*Chan eil* (prt eil chan)	is not	ncmod
xcomp	*Tha mi ag iarraidh* (xcomp Tha iarraidh)	I want	xcomp
acomp	*Tha i breagha* (xcomp Tha breagha)	It is fine	xcomp
ccomp	*bheachd gun tigeadh e* (ccomp bheachd tigeadh)	thought he would come	ccomp
mark	*gun tigeadh e* (mark tigeadh gun)	that he would come	ncmod

Table 1: Examples of the UDS-based scheme in this paper mapped to the Briscoe and Carroll scheme.

we looked in particular for sentences describing psychological states and made sure that a reasonable number of the sentences used each verb for "to be", which we will illustrate in section 5.

The sentences are tokenized by hand using the following rules: (1) Punctuation which never forms part of a lexical item such as the comma, the full stop, the colon and the semicolon is always separated out from the previous word. (2) Strings connected by a hyphen, for example *h-Alba* in *Banca na h-Alba* (Bank of Scotland) or *t-Òban* as in *an t-Òban* (the town of Oban) are always kept together. (3) The apostrophe is kept together with the copula where it proceeds it, for example in *'S fhearr leam* (I like). (4) Because the past tense particle *do* is reduced to *dh'* before a vowel and before *f*, and this is always typographically closed up, we separate out past-tense *dh'* as its own token. These rules work for the small dataset described here but would clearly need to be expanded for work in the wild.

In this preliminary work the dependencies and types have been determined by a single, non-native speaker, annotator, according to a set of guidelines which were built up during the annotation process. This is clearly less than ideal, however, the guidelines are available along with the corpus and we hope to be able to get the input of a native speaker, not least for interannotator studies.

We use the CoNLL-X format (Buchholz and Marsi, 2006), leaving the POS and projective dependency fields empty and store the categorial grammar type under field 6, FEATS.

3 Dependency scheme

There are four dependency schemes that we consulted while preparing the corpus. The initial inspiration was provided by the C&C parser (Curran et al., 2007), which in addition to providing categorial grammar derivations for sentences provides a dependency structure in the GR (Grammatical Representation) scheme due to (Briscoe and Carroll, 2000; Briscoe and Carroll, 2002). This contains 23 types and was developed originally for parser evaluation. Another popular scheme is the Stanford Dependency scheme (de Marneffe and Manning, 2008; de Marneffe and Manning, 2013), which is more finely-grained with over twice the number of dependency types to deal specifically with noisy data and to make it more accessible to non-linguists building information extraction applications. A very important scheme is the Dublin scheme for Irish (Lynn et al., 2012a; Lynn et al., 2012b; Lynn et al., 2013), which is of a similar size to the Stanford scheme, but the reason for its size relative to GR is that it includes a large number of dependencies intended to handle grammatical features found in Irish but not in English. Lastly we mention the Universal Dependency Scheme developed in (McDonald et al., 2013), which we have adopted, despite its being coarser-grained than the Dublin scheme, on account of its simplicity and utility for cross-lingual comparisons and cross-training (Lynn et al., 2014).

Table 1 gives examples of the dependency relations used along with their mapping to the GR scheme.

4 Categorial grammar

Combinatory categorial grammar (CCG) is a type-logical system which was developed to represent natural languages such as English but has subsequently been extended to other systems such as chord se-

quences in jazz (Granroth-Wilding and Steedman, 2012). For a full description the reader is referred to (Steedman and Baldridge, 2003), but in order to follow the rest of this paper you merely need to know that the type N/N is a function which takes an argument of N to its right, returning N, and that the type N\N is a function expecting an argument of N to its left and that these are combined by application, composition, where A/B combines with B/C to yield A/C, and type-raising where N is converted to T/(N\T). Attractive features of CCG for modelling a less-well-studied language include that it is a lexical theory in which it is the lexicon contains the rules for how words are combined to make sense rather than an external grammar, that it allows all manner of unconventional constituents, which is particularly powerful for parsing coordinated structures in English, that it is equivalent to a weakly context-sensitive grammar and hence has the power of a real natural language. In Steedman and Baldridge (2003) there are examples of the application of multimodal CCG to Irish Gaelic. However, to the best of our knowledge this paper is the first application of CCG to Scottish Gaelic.

In gdbank, there is a single hand-built CCG derivation for every sentence. The notation is based on that in CCGbank with a small number of adaptations for Gaelic (see next section). The basic units that can be assembled into types are S (clauses), N (nouns), conj (conjugations), and PP (prepositional phrases). For subcategorization purposes and to help keep things clear for the annotator and the reader we mark prepositional phrases with the dictionary form of the preposition.

We have not yet investigated overgeneration and ungrammatical sentences, hence there is only one kind of modality in gdbank; however restricting the way words can combine to the way in which they actually do combine in Gaelic is an obvious and essential next step.

5 Language-specific features

Prepositional phrases in Gaelic are often single-word, fused preposition–pronouns, a part-of-speech found across the Celtic languages. An ambiguous case of this is the token *ris*, which can be either *ri* with the pronoun *e*, hence taking the CCG type PP[ri], or the pre-determiner form of *ri*, hence PP[ri]/N[b]. The other class of fused preposition–pronoun we need to consider is that in sentences like *Tha mi gad chluinntinn*, "I can hear you", where *gad* is *ag* fused with *do* "your". In this case it has type PP[ag]/S[n]. Adjectives as in CCGbank are treated as clauses, S[adj]. The verbal noun is labelled S[n] by analogy with Hockenmaier and Steedman (2007). In addition to declarative and interrogative clauses, S[dcl] and S[q], we take our lead from the fourfold division of preverbal particles and add negative clauses S[neg], usually introduced by *cha* or *chan*, and negative interrogative clauses, S[negq], introduced by *nach*.

There are two verbs for "to be" in Scottish Gaelic, *bi* and *is*. *Bi* is used for predicate statements about nouns, to forming the present tense and to describe some psychological states. It does not usually equate two NPs, with an exception we will come to. In the Dublin scheme the prepositional phrase headed by *ag* in *Tá sé ag iascaireacht* ("He is fishing.") is treated as being an externally-controlled complement of *Tá* (Gaelic *tha*) and we carry this analysis over into Scottish Gaelic where this is the most common way of expressing the present tense. Figure 1 demonstrates this, where *dhachaigh* is a non-clausal modifier of *dol*, the verbal noun for "to go". *Is* can be used as the copula between two NPs, and to express psychological states such as liking and preference. To say "I am a teacher", the Gaelic is *'S e tidsear a th' annam.* This, at least on the surface, equates pronoun *e*, with a noun described by a relative clause including the verb *bi*. Fig. 1 shows our dependency tree for this. Note that this is different from the scheme in Lynn et al. (2012b) because of a difference between the two languages. They treat the analogous sentence *Is tusa an múinteoir* "You are the teacher" as having a subject, "the teacher", and a clausal predicate, *tusa*, "you indeed".

The most straightforward way of expressing a preference is the assertive *is* followed by an adjective or noun, a PP marking the preferrer, and then the object. If you dislike music, you might say *Is beag orm ceòl*. There are exactly analogous constructions in Irish with *is* + adjective + PP[le] + object, for example *Is maith liom...* "I like...", which in (Uí Dhonnchadha, 2009) is treated as having the prepositional phrase as the subject and the adjective as predicate. We modify this to use adpmod as in the Universal Dependency Scheme as shown in Fig. 1.

(a) Tha mi a' dol dhachaigh (b) 'S e tidsear a th' annam (c) Is beag orm ceòl

Figure 1: Dependency trees for (a) "I am going home", (b) "I am a teacher" and (c) "I hate music".

Type	Count	Notes	Type	Count	Notes
N	104	noun	N\N	13	adjective/genitive noun
PP/N	41	preposition	PP/S[n]	10	*ag/a'/air* etc.
N/N	38	determiner	S[dep]/PP/N	8	*bi*, *is* (after particle)
.	31	.	(N\N)/S[dcl]	7	relative
S[dcl]/PP/N	25	*bi*, *is*	S[n]	7	intransitive verbal noun
PP	18	PP	(N\N)/(N\N)	7	genitive article

Table 2: Counts for most common types found in corpus. PP[air], PP[aig] and so on have been merged.

6 Conclusions and future work

In this paper we have presented a small handbuilt corpus of Scottish Gaelic sentences, their dependency structures and their CCG derivations. To the best of our knowledge this represents the first attempt to handle a range of real-life Scottish Gaelic sentences in such a way. gdbank itself and the guidelines used to build it are available from https://code.google.com/p/gdbank/ and we welcome feedback. We have of course only been able to illustrate a small number of constructions. Tables 2 and 3 list counts for the categorial types and dependency relations used. In 32 sentences there are a total of 406 tokens.

We have not yet on the other hand attempted to deal with the morphology of Scottish Gaelic, for example lenition and slenderization, beyond drawing the attention of the human annotator to these phenomena when they may affect the correct parsing of a sentence. Clearly for automated natural-language processing of Gaelic these will need to be treated programmatically. We also disregard case and gender, although we expect that these will be dealt with as part of a rather more ambitious project, that of the Lamb group at the University of Edinburgh to build a part-of-speech tagset and tagged corpus which we look forward to seeing.

Acknowledgements

The anonymous referees for their very constructive comments.

Relation	Count	Relation	Count	Relation	Count
adpmod	58	mark	23	amod	11
nsubj	47	nmod	18	advmod	9
adpobj	38	ccomp	17	acomp	7
det	34	prt	14	cc	6
p	33	dobj	13	rcmod	4
ROOT	32	xcomp	13	appos	2

Table 3: Counts for dependency relations in gdbank. Note the high number of adpmod relations which is significantly larger than adpobj because of fused preposition–pronouns in Gaelic.

References

David Adger and Gillian Ramchand. 2003. Predication and equation. *Linguistic Enquiry*, 34:325–359.

David Adger and Gillian Ramchand. 2005. Psych nouns and predications. In *Proceedings of the 36th Annual Meeting of the North East Linguistic Society*, Amherst, MA, October.

Ted Briscoe and John Carroll. 2000. Grammatical relation annotation. Online at http://www.sussex.ac.uk/Users/johnca/grdescription/index.html.

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands, Spain, May.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, New York, NY, June.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic, June. Association for Computational Linguistics.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK, August. Coling 2008 Organizing Committee.

Marie-Catherine de Marneffe and Christopher D. Manning. 2013. Stanford typed dependencies manual. Online at http://nlp.stanford.edu/software/dependencies_manual.pdf.

Mark Granroth-Wilding and Mark Steedman. 2012. Statistical parsing for harmonic analysis of jazz chord sequences. In *Proceedings of the International Computer Music Conference*, pages 478–485. International Computer Music Association, September.

Julia Hockenmaier and Mark Steedman. 2007. CCGBank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33:355–356.

Brett Kessler. 1995. Computational dialectology in Irish Gaelic. In *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, page 60, Dublin, Ireland, March.

William Lamb. 2003. *Scottish Gaelic, 2nd edn*. Lincom Europa, Munich, Germany.

Teresa Lynn, Ozlem Cetinoglu, Jennifer Foster, Elaine Uí Dhonnchadha, Mark Dras, and Josef van Genabith. 2012a. Irish treebanking and parsing: A preliminary evaluation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1939–1946, Istanbul, Turkey, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1189.

Teresa Lynn, Jennifer Foster, Mark Dras, and Elaine Uí Dhonnchadha. 2012b. Active learning and the irish treebank. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 23–32, Dunedin, New Zealand, December.

Teresa Lynn, 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.

Teresa Lynn, Jennifer Foster, Mark Dras, and Lamia Tounsi. 2014. Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *Proceedings of Celtic Language Technology Workshop 2014*, Dublin, Ireland, August.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ingo Mittendorf and Louisa Sadler. 2005. The Welsh PARGRAM grammar. In *12th Welsh Syntax Workshop*, Gregynog, Wales, July.

Kevin P. Scannell. 2006. Machine translation for closely related language pairs. In *Proceedings of the LREC 2006 Workshop on Strategies for developing machine translation for minority languages*, pages 103–107, Genoa, Italy, May.

Mark Steedman and Jason Baldridge. 2003. Combinatory Categorial Grammar. Online at http://homepages.inf.ed.ac.uk/steedman/papers/ccg/SteedmanBaldridgeNTSyntax.pdf.

F. M. Tyers. 2010. Rule-based Breton to French machine translation. In *Proceeedings of the 14th Annual Conference of the European Association of Machine Translation, EAMT10*, pages 174–181, Saint-Raphaël, France, May.

Elaine Uí Dhonnchadha. 2009. *Part-of-speech tagging and partial parsing for Irish using finite-state transducers and constraint grammar*. Ph.D. thesis, Dublin City University.