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

Colin Batchelor
Royal Society of Chemistry

Motivation

- There has been no thorough application of Combinatory Categorial Grammar (CCG, Steedman and Baldridge 2003) to a Celtic language such as Scottish Gaelic.
- For NLP purposes, Scottish Gaelic, with 60 000 speakers in Scotland, is an under-resourced language.
- Dependency grammar provides a complementary perspective, and there is existing work, see for example Lynn *et al.* (2014), on Irish Gaelic, which is very closely related.

The corpus

Each token in each sentence has a categorial type assignment and a dependency relation to another token or the root.

Source	Number of sentences	Tokens
MacIlleathain (2014)	25	349
The Scotsman	6	151
Robertson and Taylor (2005)	1	4
Lamb (2003)	5	29
Crawford and Imlah (2006)	3	90
Total	40	623

Phenomena covered so far:

- Embedded interrogative clauses (*carson a tha...*).
- Fronting with *is* + *ann*.
- Indirect speech.
- Non-constituent coordination.
- Obligation with *feum* and *bi* + *aig* + *ri*.
- Preposed prepositional phrases.
- Psych nouns with *bi* (*beachd*, *cuimhne*, *eagal*, *fios*, *gràin*) and *is* (*còir*, *toigh*).
- Superlatives including past (*na bu dlùithe*).

The corpus itself is a plain-text, UTF-8 encoded file in CoNLL-X format (Buchholz and Marsi 2006). We put the categorial grammar annotations in column 6. The guidelines are available alongside and consist of over 80 rules. See `http://gdbank.googlecode.com`.

Categorial grammar

Categorial grammar is based on clauses (`S`), and nouns (`N`). Every word has a type, which can be one of these two atomic types or a function which has arguments immediately adjacent to it on either side. Feature structures can be used to handle persons, number, lenition and so forth. Parsing involves applying a very small set of operations: application, composition, substitution, type-raising and type-changing.
Forward and backward application rules:

$$X/Y : f \qquad Y : a \Rightarrow_{>} X : f(a) \tag{1}$$

$$Y : a \qquad X\backslash Y : f \Rightarrow_{<} X : f(a) \tag{2}$$

Here is an example derivation using forward application only with neo-Davidsonian semantics:

Tha	*Màiri*	*a'*	*dol*
BI	Mary	PROG	going
$\mathrm{S/S/N/N} : \lambda f.\lambda x.\exists e.f(e,x)$	$\mathrm{N} : \mathit{màiri}'$	$\mathrm{S/N/S/N} : \lambda f.\lambda e.\mathit{prog}(e) \wedge f(e)$	$\mathrm{S/N} : \mathit{goes}(e) \wedge \mathit{agent}(e,x)$
$\mathrm{S/S/N} : \lambda f.\exists e.f(e, \mathit{màiri}')$ (>)		$\mathrm{S/N} : \mathit{prog}(e) \wedge \mathit{goes}(e) \wedge \mathit{agent}(e,x)$ (>)	
$\mathrm{S} : \exists e.\mathit{prog}(e) \wedge \mathit{goes}(e) \wedge \mathit{agent}(e, \mathit{màiri}')$ (>)			

Long-distance dependencies can be handled through type-raising. The unary type-changing rules can turn predicative adjectives into attributive adjectives, or turn PPs into modifiers of either clauses or nouns. The choices below are for Gaelic but should work for other Celtic languages:

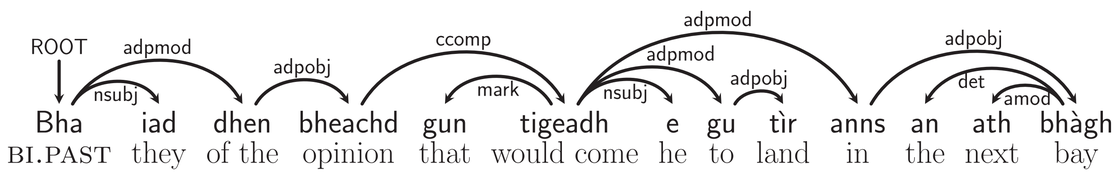
- The verb takes arguments to its right: `S/N/N`.
- Verbal nouns are treated as verbs: `S[n]/N/N`.
- Determiners (*an*, *gach*) take NPs to their right: `N/N`.
- Most attributive adjectives take nouns to their left: `N\N`.
- Predicative adjectives are clausal: `S[adj]/N`.
- Prepositions that take an argument are `PP/N` but fused preposition–pronouns (*orm*, *agad*, *annaibh*, *aiste*) are simply `PP`.
- Particles are feature-changing functions. *Ag* turns a verbal noun `S[n]/N` into an aspect-marked clause `S[asp]/N`.

Dependency scheme

Like Lynn *et al.* (2014) we use the Universal Dependency Scheme (McDonald 2013), which is based heavily on the Stanford scheme (de Marneffe and Manning 2008). We do not use all of the dependency types.

Label	Dependency
`ROOT`	Identifies head of sentence
`nsubj`	Subject of verb
`dobj`	Object of verb
`adpmod`	Introduces prepositional phrase
`adpobj`	Object of preposition
`det`	Determiner
`p`	Punctuation
`mark`	Marks a clausal complement
`nmod`	Nominal modifier
`xcomp`	Externally-controlled clausal complement
`ccomp`	Clausal complement
`prt`	Particle before verb
`amod`	Adjectival modifier
`acomp`	Adjectival complement
`advmod`	Adverbial modifier
`cc`	Conjunction
`rcmod`	Relative modifier
`appos`	Apposition

Example:



Work in progress

- Most important: implementing the grammar in openccg (`https://github.com/OpenCCG/`).
- More detailed semantics.
- Adding more sentences.
- Adding coarse-grained part-of-speech tags from the Universal POS scheme of Petrov *et al.* (2012).

Bibliography

- Sabine Buchholz and Erwin Marsi (2006), CoNLL-X shared task on multilingual dependency parsing, *Proceedings of the Tenth Conference on Computational Natural Language Learning*, New York.
- Robert Crawford and Mick Imlah, eds (2006), *New Penguin Book of Scottish Verse*, Penguin, Harmondsworth.
- William Lamb (2003), *Scottish Gaelic*, 2nd edn, Lincom Europa, Munich, Germany.
- Teresa Lynn, Jennifer Foster, Mark Dras and Lamia Tounsi (2014), Cross-Lingual Transfer Parsing for Low-Resourced Languages: An Irish Case Study, *Proceedings of Celtic Language Technology Workshop 2014*, Dublin, Ireland.
- Ruairidh MacIlleathain (2014), *An Litir Bheag*, podcast online at `http://www.bbc.co.uk/alba/foghlam/learngaelic/anlitirbheag/`
- Ryan McDonald *et al.* (2013), Universal Dependency Annotation for Multilingual Parsing, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria.
- Slav Petrov, Dipanjan Das and Ryan McDonald (2012), A Universal Part-of-Speech Tagset, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.
- Boyd Robertson and Iain Taylor (2005), *Teach Yourself Gaelic*, Teach Yourself Books, London.
- Mark Steedman and Jason Baldridge (2003), "Combinatory Categorial Grammar" online at `http://homepages.inf.ed.ac.uk/steedman/papers/ccg/SteedmanBaldridgeNTSyntax.pdf`

Thomas Graham House, Milton Road, Cambridge, UK CB4 0WF
e-mail: batchelorc@rsc.org
twitter: @my_disposition