Universal dependencies for Scottish Gaelic: syntax

Colin Batchelor

Royal Society of Chemistry, Cambridge, UK

19th August 2019
Overview

Motivation

Scottish Gaelic

Using the Annotated Reference Corpus of Scottish Gaelic

Dependency grammars

Method, examples and results

Cosubordination

Conclusions and next steps
Motivation

- No validated syntactic parser for Scottish Gaelic.
- Contributing to a multi-language project enables us to concentrate on the data while other people concentrate on the tools (tokenisers, sentence-splitters, parsers, and so on.)
Scottish Gaelic

- Around 60,000 speakers.
- Separate written form from Irish since ca. the Jacobite rebellion.
- VSO (cf. Breton).
- Usual way of forming the present tense is \( bi + \) a verbal noun.
- Compound/inflected prepositions like other Celtic languages.
- Multiple ways of forming copular statements.
- Cosubordination and other unusual uses of coordinators \( agus/\text{is}'s. \)
- Morphology: orthographically-marked lenition (except for l, n and r) and slenderisation.
- Resources: mainly dasg.ac.uk, Am Faclair Beag and other work by Michael Bauer, work by Kevin Scannell, Caoimhín Ó Donnaíle’s resources at SMO and ARCOSG.
ARCOSG (Lamb et al. 2014–2016)

First machine-readable corpus of Scottish Gaelic.

- Over 80,000 tokens in 77 documents.
- Over 1300 utterances; around 1800 sentences (sentence/utterance boundaries not supplied).
- Categories: conversation, interview, sports commentary, narrative, news scripts, fiction, formal writing, popular writing.
- We also know the authors of the written text. (cf. https://www.aclweb.org/anthology/P19-1339)
- Identifying dialect takes a bit more digging.
- Tagged with parts of speech based on the PAROLE scheme.

Online here: https://doi.org/10.7488/ds/1411
From ARCOSG to parsing

- We follow the v2 Universal Dependencies guidelines and Lynn et al.’s treatment of Irish, with some special cases for Gaelic.
- We assign a UPOS based on the ARCOSG tag, largely following the Irish scheme.
- Some deviations from the PAROLE scheme:
  - Personal prepositions (agam, agad) count as ADPs.
  - Aspect markers a'/ag, ri, gu are also ADPs (but cf. Johannes and Fran’s talk).
  - Specials like airson ‘for’ are handled separately.
- Python scripts to lemmatise and assign features.
- In our initial pass we split sentences on full stops and subdivide them more accurately as part of the annotation process.
Universal Dependencies

Requirements set out by Christopher Manning:

▶ UD needs to be satisfactory on linguistic analysis grounds for individual languages.
▶ UD needs to be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
▶ UD must be suitable for rapid, consistent annotation by a human annotator.
▶ UD must be suitable for computer parsing with high accuracy.
▶ UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing.
▶ UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, ???).
Universal Dependencies

Some consequences:

- Content words are heads (in PPs the heads are nouns, not the prepositions).
- No distinctions between arguments and modifiers.
- Copula (x is a y) has its own relation. Languages like Russian and Chinese have zero copula.

Specific to this work:

- Treat coordinative phrases like *air sgàth ’s gun*... (‘on the cause and that...’ ≈ ‘because...’), *fiu ’s ged a*... (‘worthy and although...’ ≈ ‘even if...’), *còrr is*... (‘extra and...’ ≈ ‘more than...’) strictly as coordination.
- In NPs ending with a noun in the genitive, say *beagan mhionaidean* ‘a few minutes’, treat *beagan* as the head and *mhionaidean* as a modifier.
Method

- Convert ARCOSG to CoNLL-U format.
- Put aside sports commentary, conversation, interview and narrative files with no obvious sentence boundaries for now.
- Annotate around a thousand tokens from shortish sentences.
- Train up a parser (initially MaltParser, then udpipe) to annotate the rest.
- Fix errors and iterate, starting with the shortest sentences.
Copula: nouns as roots

'S i Morag a rinn a' bhanaís
AUX PRON PROPN PART VERB DET NOUN
Wp-i Pp3sf Nn-fn Qa Tdsf V-s Ncsfn
COP PRED Morag that did the wedding

'S ise an dotair
AUX PRON DET NOUN
Wp-i Pp3sf-e Tdsm Ncsmn
COP she-EMPH the doctor
Copula: contrastive use

Examples from Cox (Geàrr-Ghràmar na Gàidhlig, 2017):

'S ann an-diugh a cheannaich mi bainne
AUX ADP ADV PART VERB PRON NOUN
Wp-i Pr3sm Rt Q-r V-s Pp1s Ncsmn
COP PRED today REL bought I milk

'S ann a cheannaich mi bainne
AUX ADP PART VERB PRON NOUN
Wp-i Pr3sm Q-r V-s Pp1s Ncsmn
COP PRED REL bought I milk
Results

1021 sentences. 20 031 words. 20 021 tokens.

<table>
<thead>
<tr>
<th>Transition system</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>projective</td>
<td>0.796 [0.752, 0.839]</td>
</tr>
<tr>
<td>swap (non-projective)</td>
<td>0.789 [0.747, 0.825]</td>
</tr>
<tr>
<td>link2 (non-projective)</td>
<td>0.792 [0.750, 0.835]</td>
</tr>
</tbody>
</table>

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.
## Statistics: parts of speech

<table>
<thead>
<tr>
<th>UPOS</th>
<th>Count</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUN</td>
<td>4567</td>
<td>incl. verbal noun</td>
</tr>
<tr>
<td>ADP</td>
<td>2711</td>
<td></td>
</tr>
<tr>
<td>PUNCT</td>
<td>2367</td>
<td></td>
</tr>
<tr>
<td>VERB</td>
<td>2029</td>
<td></td>
</tr>
<tr>
<td>PART</td>
<td>1848</td>
<td></td>
</tr>
<tr>
<td>DET</td>
<td>1450</td>
<td></td>
</tr>
<tr>
<td>ADJ</td>
<td>850</td>
<td></td>
</tr>
<tr>
<td>ADV</td>
<td>747</td>
<td></td>
</tr>
<tr>
<td>CCONJ</td>
<td>646</td>
<td></td>
</tr>
<tr>
<td>NUM</td>
<td>243</td>
<td></td>
</tr>
<tr>
<td>AUX</td>
<td>230</td>
<td><em>is, gur</em></td>
</tr>
<tr>
<td>SCONJ</td>
<td>218</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>INTJ</td>
<td>38</td>
<td></td>
</tr>
</tbody>
</table>
Statistics: Dependency relations

<table>
<thead>
<tr>
<th>Deprel</th>
<th>Count</th>
<th>Comments</th>
<th>Deprel</th>
<th>Count</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>punct</td>
<td>2367</td>
<td></td>
<td>flat</td>
<td>390</td>
<td></td>
</tr>
<tr>
<td>case</td>
<td>2308</td>
<td></td>
<td>xcomp:pred</td>
<td>263</td>
<td>adjectives</td>
</tr>
<tr>
<td>obl</td>
<td>2145</td>
<td></td>
<td>cop</td>
<td>226</td>
<td></td>
</tr>
<tr>
<td>nsubj</td>
<td>1761</td>
<td></td>
<td>acl:relcl</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>mark</td>
<td>1565</td>
<td></td>
<td>parataxis</td>
<td>195</td>
<td></td>
</tr>
<tr>
<td>root</td>
<td>1021</td>
<td></td>
<td>nummod</td>
<td>190</td>
<td></td>
</tr>
<tr>
<td>xcomp</td>
<td>832</td>
<td>verbal nouns</td>
<td>nmod:poss</td>
<td>171</td>
<td></td>
</tr>
<tr>
<td>nmod</td>
<td>724</td>
<td></td>
<td>csubj:cop</td>
<td>111</td>
<td></td>
</tr>
<tr>
<td>obj</td>
<td>651</td>
<td></td>
<td>appos</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>advmod</td>
<td>625</td>
<td></td>
<td>compound</td>
<td>82</td>
<td>fhèin, fhìn</td>
</tr>
<tr>
<td>fixed</td>
<td>547</td>
<td>'s e, an dèidh</td>
<td>obl:tmod</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>conj</td>
<td>543</td>
<td></td>
<td>vocative</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>ccomp</td>
<td>529</td>
<td></td>
<td>case:voc</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>amod</td>
<td>517</td>
<td></td>
<td>obl:smod</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>
Cosubordination!

Cosubordination is a distinctive feature of Scottish Gaelic found in all registers. This is where *agus*, *is* and ’s ‘and’ coordinate a sentence with an incomplete clause where the tense and other features come from the matrix clause (examples from corpus):

- *Chaidh bratach Bhreatainn a thoirt a-nuas ann an seirbheis taobh muigh an taighe, ’s an Last Post ga chluiche.*
- *... tha d` a she` orsa eadar-dhealaichte ann de fhuil A; ’se sin A1 agus A2 is iad fo smachd dors aileal eadar-dhealaichte.*
- *Tha bean-gairm SAND, Lorrainne Mann, ag ràdha gu feuch iad a-nise ri barrachd taic fhaighinn, ’s iad a’ dol a sgrìobhadh gu comhairlean coimhearsnachd anns an sgìre.*

Cannot be fronted or clefted. Also seen in Irish, where it can be fronted, but couldn’t find it in Irish treebank.
He is fishing
Depictive approach

Chaidh e air chall ann an ceò, 's e iasgach.

Went he fishing in mist, and he was lost in mist.
Elliptical approach

(In the paper I used orphan, but it should be promoted to xcomp. Use conj:cosub?)
Auxiliary approach

Chaidh went he asp lost in mist, and he asp fishing
Comparison

With *chaidh* and *bi* as heads, we need to choose:
- The depictive approach coordinates like with like.
- The elliptical approach is easier to make projective.
- The elliptical approach implies that there is *something* that has been elided.

With the verbal noun as head the approach is obvious, but doesn’t work for cases with PPs and adjectives.
Conclusions and next steps

- Decide about cosubordination.
- Address conversations.
- Address test/train/dev split. (*qv.* https://www.aclweb.org/anthology/P19-1267)
- Write up treatment of comparatives.
- Will go live in v2.5 of Universal Dependencies this November.
- universaldependencies.org
- Code (and partly-processed trees) here: https://github.com/colinbatchelor/gdbank/
- *Mòran taing dhan a h-uile duine!*
- *Ceist sam bith?*