Automatic derivation of categorial grammar from a part-of-speech-tagged corpus in Scottish Gaelic

Colin Batchelor
Royal Society of Chemistry, Cambridge, UK CB4 0WF
colin.r.batchelor@gmail.com

RÉSUMÉ

Grammaire catégoriale dérivé automatiquement d'un corpus des textes en gaélique écossais avec annotations syntaxiques

Nous presentons une grammaire catégoriale preliminaire pour le gaélique écossais qui nous avons derivé automatiquement du corpus de texte ARCOSG (*Annotated Reference Corpus of Scottish Gaelic*) de l'Université d'Édimbourg, qui contient plus que 80 000 des entités lexicales en plusieurs genres avec annotations syntaxiques. Nous discutons nos méthodes pour la dérivation de cette grammaire, les traits distinctifs du gaélique écossais et du corpus, l'analyse lexicale categoriale, et dont on a besoin pour une évaluation rigoureouse et systématique d'une telle grammaire.

ABSTRACT

We present a preliminary categorial grammar for Scottish Gaelic derived automatically from the University of Edinburgh's Annotated Reference Corpus of Scottish Gaelic (ARCOSG), which contains over 80 000 tokens of part-of-speech-tagged text in multiple genres. We discuss our methods for deriving this grammar, the distinctive features of Scottish Gaelic and of the corpus, parsing CCG, and set out what is needed for a rigorous and systematic evaluation of the work presented here.

MOTS-CLÉS : gaélique écossais, grammaire categoriale, CCG.

KEYWORDS: Scottish Gaelic, categorial grammar, CCG.

1 Introduction

Scottish Gaelic, like the other Celtic languages, is marked by VSO word order, fused preposition–pronouns, word-initial mutation and extensive use of periphrastic constructions (Lamb, 2003). As in Irish the copula and verb "to be" are separate, and psychological states are typically expressed with a combination of either of those, prepositional phrases and nouns. As such it is a challenging language for automatic processing, a situation which is not helped by its having historically been an under-resourced language for natural language processing, but this started to change at the first Celtic Language Technology Workshop in Dublin in 2014 with the publication of three papers by Lamb & Danso (2014), Scannell (2014) and Batchelor (2014). Subsequently the University of Glasgow has launched the *Corpas na Gàidhlig* 'Corpus of Gaelic' as part of the Digital Archive of Scottish Gaelic (DASG) (University of Glasgow, 2016). The potential for developing resources for Scottish Gaelic has been strengthened by a recent flurry of activity in Irish, which is very closely related, the two having shared a common literary form until the 18th century. Irish now boasts a dependency treebank (Lynn, 2016), a mapping of this Irish Treebank annotation scheme to the scheme in the

Universal Dependencies Project (Nivre *et al.*, 2015), [1] and tools for POS-tagging tweets (Lynn *et al.*, 2015). In this paper we present a Scottish Gaelic categorial grammar bank derived, in contrast to our small hand-built grammar presented in Batchelor (2014), wholly automatically from a part-of-speech tagged corpus, the Annotated Reference Corpus of Scottish Gaelic (ARCOSG) (Lamb *et al.*, 2016), the longer-term background to which is described in Lamb (2008).

2 Methods

2.1 Categorial grammar

Combinatory categorial grammar (CCG) (Steedman & Baldridge, 2003) is a fully-lexicalized theory. This means that all of the grammar resides in the lexicon and that parsing involves applying those rules stored within the lexical entries. Each lexical entry, or word, has a type which may either be atomic or composite. As is standard we work with a small set of atomic types, which in this exercise are the clause (S), the noun phrase (N) and the prepositional phrase (PP). The composite types are functions and are written with slashes indicating whether their arguments are to their right or to their left. To take a simple example, intransitive verbs in Scottish Gaelic have type S/N, indicating that they expect a noun phrase to their right, and attributive adjectives have type N\N, indicating that they expect a noun phrase to their left. Parsing in its simplest form then involves function **application** using the rules :

$$\mathtt{A/B} \quad \mathtt{B} \quad \rightarrow_{>} \quad \mathtt{A} \tag{1}$$

$$\mathtt{B} \quad \mathtt{A\backslash B} \quad \rightarrow_{<} \quad \mathtt{A} \tag{2}$$

To give a concrete example, the phrase *Thàinig corra-ghridheach ghiùigeach* 'A demure heron came' parses as follows :

$$\dfrac{\dfrac{\textit{Thàinig}}{\mathtt{S/N}} \quad \dfrac{\dfrac{\textit{corra-ghridheach}}{\mathtt{N}} \quad \dfrac{\textit{ghiùigeach}}{\mathtt{N\backslash N}}}{\mathtt{N}}<}{\mathtt{S}}> \tag{3}$$

the N\N of *ghiùigeach* combines backwards with the N of *corra-ghridheach* to yield an N, which is then consumed by the S/N of the verb *thàinig* to yield a complete clause.

In addition to application, there are also **harmonic composition** operations.

$$\mathtt{X/Y} \quad \mathtt{Y/Z} \quad \rightarrow_{>\mathbf{B}} \quad \mathtt{X/Z} \tag{4}$$

$$\mathtt{Y\backslash Z} \quad \mathtt{X\backslash Y} \quad \rightarrow_{<\mathbf{B}} \quad \mathtt{X\backslash Z} \tag{5}$$

Operation (4) enables us to use types such as N/S[gu] for "propositional" nouns such as *dùil* 'expectation' or *dòchas* 'hope' so that they can combine with clauses that begin with the word *gu* 'that'.

1. http://universaldependencies.org/

2.2 Assigning types

The usual process for generating a categorial grammar bank, as exemplified for English (Hockenmaier & Steedman, 2007), and Chinese (Tse & Curran, 2010), is to take a pre-existing set of context-free grammar parse trees, to convert any non-binary nodes to binary node, and to assign a category to every node. For German, Hockenmaier Hockenmaier (2006) describes an analogous process based on the TIGER dependency treebank.

However, there being no treebanks for Scottish Gaelic, we need to take a different approach. The main resource for Scottish Gaelic is ARCOSG, which is a corpus of 76 texts from a variety of genres. These have been part-of-speech tagged by hand according to a tagging scheme described in Naismith & Lamb (2014). What we can do, therefore, is to build a categorial grammar in which each lexical entry contains a category that is assigned purely on the token and tag information for a given word in ARCOSG. This is similar to supertagging (Bangalore & Joshi, 1998), an approach which is usually the first step in CCG parsing, in which all of the possible CCG categories are applied to each word in the text and the CCG parser then attempts to find the best overall parse. The difference here is that we are doing this on the level of the original corpus itself, in order to generate a grammar.

The initial version of the mapping was based on the scheme in Batchelor (2014), which is itself largely based on Hockenmaier & Steedman (2007) with adjustments for VSO order in Gaelic. This was refined first by ensuring that there was complete coverage of all of the parts of speech in ARCOSG, and then that it was possible to parse the corpus itself. A summary is given in Table 1.

There are some subtleties which we shall discuss here. The ARCOSG tagset is based closely on the PAROLE tagset used by Uí Dhonnchadha (2009). (Lynn, 2016) describes in detail how the PAROLE tagset is not completely appropriate for her work in dependency grammar. Many of these are familiar topics in Celtic linguistics and are also relevant toe our categorial grammar treatment.

In ARCOSG the prepositional pronouns, for example *orm*, *ort* ("on me", "on you") are treated as pronouns whereas for verbal subcategorization they should be treated in the same way as prepositions. We treat transitive verbal nouns as `S[small]/N/N` and the aspectual particles *a'*, *ag*, *air*, *gu* and *ri*, which precede verbal nouns and are in most cases identical to prepositions, as type-changing particles. [2] *Airson* is tagged as a fossilized noun (*Nf*) in ARCOSG, whereas we treat it here as a preposition (`PP/N`). If a word in ARCOSG is in the "wrong" case according to the accepted grammar of Scottish Gaelic, then it will be tagged with the correct case and the part of speech marked with an asterisk. In these cases we disregard the asterisk and treat the word as a variant.

If we allow dashes and commas to act as noun-coordinators and noun-postmodifiers then we can handle apposition introduced by punctuation. More difficult are plural genitives, which are often identical to either the singular or plural nominative and may be tagged as such.

2. One longer-term reason for doing this is to make the semantics more transparent. First consider the verbal nouns as a whole :

— Intransitive verbs : `S[small]/N`: $f(e) \wedge agent(e, x)$

— Transitive verbs : `S[small]/N/N`: $f(e) \wedge agent(e, x) \wedge patient(e, y)$.

The particles that are unmarked for person, *a'/ag*, *gu*, *ri* and *air*, supply the aspect, hence *a' cluinntinn* ("hearing") gives us

$$progressive(e) \wedge hears'(e) \wedge agent(e, x) \wedge patient(e, y). \quad (6)$$

gam, *gad* and so forth supply not only the aspect but also the patient, hence *gad chluinntinn* ("hearing you") :

$$progressive(e) \wedge hears'(e) \wedge agent(e, x) \wedge patient(e, thu'). \quad (7)$$

ARCOSG	CCG	Comments	Example
Ap	S[adj]/N	predicative adjective	
Aps	(S[adj]/N)/N	second comparative	*feairrde*
Aq	N*N	attributive adjective	
Ar	N/*N	premodifying adjective	*droch*, *seann*
Av	N*N	past participle	
Cc	N*N/N, S*S/S	coordinators	*agus*, *ach*
Cs	S*S/S	subordinators	
Csw	S[gu]/N/N	*gur*	
D	N/*N	determiners	
Fq	S/*S	open quote	
all other *F*	S*S	punctuation	
Mc	N	cardinal numbers	
Mo	N/*N	ordinal numbers	
Nf	N	fossilized noun	
except *airson*	PP[airson]/N	preposition	
Nn-mn	N/*N	forename	
Nv	as verbs	verbal noun	
N...g	N*N	genitive noun	
N...v	S/S	vocative noun	*a Sheumais*
all other *N*	N	nouns	
Pn	N	numerical pronouns	*ceithir*
Pp	N	pronouns	*mi*, *mise*, *i*, *iad*
Pr	PP	personal prepositions	
Q	S[x]/S[y]	clause feature value changers	*cha*, *do*, *gu*
except *Q-s*	(S*S)/S[dep]	"if"	*nam*, *nan*
R	S*S	adverbs	
Sa	S[asp]/N/S[small]/N	aspect	*a'*, *air tighinn*
	S[asp]/N/S[inf]/n		*air a chumail*
Sap	S[asp]/S[small]/N	personal aspect	*gad*, *gam*
Sp	PP/N	prepositions	
T...n, *T...d*	N/*N	articles	
T...g	(N*N)/(N*N)	genitive articles	
Uf	N	fossilized noun	*dòcha*, *urrainn*
Ug	S[inf]\N/S[small]/N/N	agreement particle	
Uv	(S/*S)/(S/*S)	vocative particle	*a Sheumais*
V	varies	verbs	
W	varies	copula	
Xfe	N	foreign words	
Xsc	S/*S	marks a speaker	

TABLE 1 – The most important part-of-speech classes from ARCOSG and the types they map to in our categorial grammar treatment.

ARCOSG POS	Description	Procedure
Nv	verbal noun	see Table 3
all *W*	copula	*is*
*V*s*	past tense	delenite
Vm-1p	1p imperative	remove *-eamaid* or *-amaid*
Vm-2s	singular imperative	preserve
Vm-2p	plural imperative	remove *-ibh* or *-aibh*
V-h, *Vm-3*	conditional, 3p.imp.	delenite, remove *-eadh* or *-adh*
*V.*d*	dependent form	delenite
*V.*f*	future tense	remove *-idh* or *-aidh*
*V.*r*	relative	remove *-eas* or *-as*
V-s0	past impersonal	delenite, remove *-eadh* or *-adh*
V-p0	present impersonal	remove *-ear* or *-ar*

TABLE 2 – Operation of the lemmatizer on verbs. In each case the slenderized form of the suffix is given first.

For determiners, conjunctions and adjectives we use the non-associative, non-permutative slash $/_\star$ from multimodal combinatory categorial grammar (Baldridge & Kruijff, 2003). We ban forward-crossed composition, though this may prove to be unnecessary if we make full use of the multimodal slash repertoire.

2.3 Lemmatization

The ARCOSG tagset marks nouns and articles for number and case, verbs and prepositions and pronouns for person and number, and verbs for tense and whether they are the independent, dependent or relative form of the verb. These are incorporated as features ; for example the verb *thòisich* with the tag *V-p* gets the tense feature `pres`.

However, it does not mark them for transitivity or which prepositional phrases they subcategorize with. This is clearly beyond the scope of a POS tagger, especially one for a corpus of this size, and a full treatment requires a larger dictionary. For this we require a lemmatizer for verbs. We are not aware of any publications about a verb lemmatizer for Scottish Gaelic. Lemmatizers for Irish have previously been presented by Uí Dhonnchadha & Van Genabith (2005) and Měchura (2014). The lemmatizer requires the surface form of the verb and a part-of-speech tag, but Gaelic, while morphologically rich, is largely systematic and it mostly proceeds by delenition [3] where necessary and removing endings. [4] The procedure for this, which covers all of the grammatical categories for verbs found in ARCOSG, is listed in Table 2. The irregular verbs *bi*, *abair*, *beir*, *cluinn*, *dèan*, *faic*, *faigh*, *rach*, *ruig*, *thoir*, *thig* and all verbal nouns are treated separately, the irregular verbs by means of a lookup table and verbal nouns by deleniting where necessary and following the procedure in Table 3.

3. In contrast to the mutations in Welsh, Cornish and Breton, lenition in Irish and Scottish Gaelic is marked orthographically by inserting an *h* after the initial consonant.

4. The endings take different forms according to whether they follow a 'slender' consonant or a 'broad' consonant. These are marked in the orthography as follows : a slender consonant has the vowels *i* or *e* as neighbours ; a broad consonant has the vowels *a*, *o* or *u*. There are occasional exceptions, usually compound words such as *airson* and *rudeigin*, but they do not affect the algorithm.

Input	Output	Example
in list	return dictionary form	*tuiteam* → *tuit*
-sinn, *-s'*, *-tainn*	strip ending	*creidsinn* → *creid*
-eachadh	*-ich*	*cruinneachadh* → *cruinnich*
-achadh	*-aich*	*sgioblachadh* → *sgioblaich*
-gladh	*-gail*	*fosgladh* → *fosgail*
-eadh	strip ending	*tilleadh* → *till*
-adh	strip ending	*glanadh* → *glan*
otherwise	preserve	*ruith* → *ruith*

TABLE 3 – Operation of the lemmatizer on verbal nouns.

#	Rule	Explanation
1	$\texttt{PP} \rightarrow \texttt{N}\backslash_\star\texttt{N}$	PPs modifying noun phrases
2	$\texttt{PP} \rightarrow \texttt{S}\backslash_\star\texttt{S}$	postposed PPs modifying clauses
3	$\texttt{S}\backslash_\star\texttt{S} \rightarrow \texttt{S}/_\star\texttt{S}$	preposed adverbials
4	$\texttt{N[place]} \rightarrow \texttt{N}\backslash_\star\texttt{N}$	placenames used attributively
5	$\texttt{S[small vowel]}/\texttt{N} \rightarrow \texttt{S[inf]}$	intransitive verbal nouns without agreement particle
6	$\texttt{S[small vowel]}/\texttt{N} \rightarrow \texttt{S[inf]}\backslash\texttt{N}$	transitive verbal nouns without agreement particle
7	$\texttt{PP}/\texttt{N} \rightarrow \texttt{PP}/\texttt{S[int]}/\texttt{PP}$	relative clauses in PPs

TABLE 4 – Type-changing rules.

2.4 Unary rules

The binary combinators listed above are insufficient for practical CCG parsing. We have a set of systematic **type-changing rules**, listed in Table 4, which serve to reduce the size of the lexicon by minimizing the number of types. Rules 1 and 2 convert the atomic category PP into a modifier of both NPs and clauses. Rule 3 allows clausal modifiers, be they PPs or adverbial phrases, to go before the clause. Rule 4 deals with placenames, which are not marked for case, when they are used attributively.

Passive-type constructions using the verb *rach* 'to go' are common in the news genre. For example, *Chaidh barrachd dhaoine a mharbhadh* 'more people were killed', where *chaidh* is the past tense of *rach* 'go', *marbhadh* is the verbal noun 'killing', and *a* is an "agreement particle". To handle the case where the verbal noun begins with a consonant, we use **type-raising rule** 1 in Table 5 and give the agreement particle the type $\texttt{S[inf]}\backslash\texttt{N}/\texttt{S[small]}/\texttt{N}/\texttt{N}$.

$$\dfrac{\dfrac{\textit{chaidh}}{\texttt{S[dcl]}/\texttt{S[inf]}} \quad \dfrac{\dfrac{\dfrac{\textit{barrachd dhaoine}}{\texttt{N}}}{\texttt{S}/\texttt{S}\backslash\texttt{N}}\,{>}\mathbf{T} \quad \dfrac{\dfrac{\textit{a}}{\texttt{S[int]}\backslash\texttt{N}/\texttt{S[small]}/\texttt{N}/\texttt{N}} \quad \dfrac{\textit{mharbhadh}}{\texttt{S[small]}/\texttt{N}/\texttt{N}}}{\texttt{S[inf]}\backslash\texttt{N}}\,{>}}{\texttt{S[inf]}}\,{>}}{\texttt{S[dcl]}}\,{>} \tag{8}$$

Rules 5 and 6 in Table 4 handle the analogous case where the verbal noun begins with a vowel, so there is no agreement particle. Rule 7 deals with prepositional relative clauses, for example *leis a bheil an taigh* 'who owned the house'. Unlike conventional relative clauses, which take a declarative

#	Rule	Explanation
1	$\mathtt{N} \rightarrow_{>\mathbf{T}} \mathtt{S/S\backslash N}$	For the *rach* passive
2	$\mathtt{PP} \rightarrow_{<\mathbf{T}} \mathtt{S\backslash S/N}$	For relative clauses
3	$\mathtt{S[adj]/N} \rightarrow_{<\mathbf{T}} \mathtt{S\backslash S/S[adj]/N}$	For relative clauses

TABLE 5 – Type-raising rules

or relative future form of the verb after the relativiser *a*, these take the interrogative form of the verb, for example *a bheil* 'is ?'. We then use forward composition (eqn. 4)

The other type-raising rules in Table 5 enable us to form relative clauses with *a*. To take the example NP *an gille a tha bochd* 'the boy who is ill' :

$$\dfrac{\dfrac{\textit{an gille}}{\mathtt{N}} \quad \dfrac{\textit{a}}{\mathtt{N\backslash N/S/N}} \quad \dfrac{\textit{tha}}{\mathtt{S[dcl]/(S[adj]/N)/N}} \quad \dfrac{\dfrac{\textit{bochd}}{\mathtt{S[adj]/N}}}{\mathtt{S\backslash S/S[adj]/N}}\,{<}\mathbf{T}}{\mathtt{S[dcl]/N}}\,{<}\mathbf{B}_{\times}}{\mathtt{N\backslash N}}\,{>}}{\mathtt{N}}\,{>} \tag{9}$$

we use the additional backward crossed composition operation

$$\mathtt{Y\backslash Z} \quad \mathtt{X\backslash Y} \rightarrow_{<\mathbf{B}_{\times}} \mathtt{X/Z}. \tag{10}$$

in addition to type-raising rule 3.

3 In practice

3.1 Pre-processing

The POS-tagged text in ARCOSG treats multiword expressions such as toponyms *e.g. Beinn na Faoghla* 'Benbecula', multiword prepositions such as *an aghaidh* 'against' and fixed phrases such as *Gu sealladh ort !* 'Heaven preserve you !' as single tokens. For simplicity we apply a preprocessing step to ARCOSG where lexical entries containing spaces have them replaced with underscores in place of spaces, thus `ann_an` instead of `ann an`.

3.2 Parsing

Out of the available CCG parsers, we chose OpenCCG, a categorial grammar parsing and realization toolkit, [5] to parse Gaelic text taken from ARCOSG. The key strengths of OpenCCG for rapid prototyping and development of categorial grammars are that it has an interactive mode and a transparent syntax (dotccg format (Baldridge *et al.*, 2007)) for specifying grammars, and an efficient chart parser. One weakness is that by default it doesn't handle out-of-vocabulary text. We also considered the CCG parser in the NLTK ; however the version in NLTK 3.1 (October 2015) doesn't

5. `http://openccg.sourceforge.net/`

support features, such as the type of clause, gender or tense, and as such it is not usable for our purposes. Otherwise the excellent and well-established C&C parser (Curran *et al.*, 2007) is too closely entangled with the underlying CCGbank to be used for this sort of development work.

For the word *ann* 'in it', 'there', 'in him', the OpenCCG parser produces seven parses for which we list here the final result without the full derivations :

```
Parse 1: pp/n
Parse 2: pp
Parse 3: pp<1>/(s{clause=int}/pp<1>)
Parse 4: n<2>\n<2>
Parse 5: s<3>\s<3>
Parse 6: s<6>\@i(s<6>/@ipp)
Parse 7: s<11>/s<11>
```

The first parse comes from the phrase *ann a bhith* 'in which... is', which appears several times in the corpus, and the others are from the type-raising and type-changing rules we have discussed before. Clearly there is no one correct parse for a single word. The correct full derivation (out of six found by OpenCCG for our grammar) for *tha i fliuch* 'it is wet' (used usually of the weather) is :

```
(lex)  tha :- s{clause=dcl, phon=cons, tense=pres}/(s{clause=bi_arg}/n)/n
(lex)  i :- n{ont=pron}
(>)    tha i :- s{clause=dcl, phon=cons, tense=pres}/(s{clause=bi_arg}/n)
(lex)  fliuch :- s{clause=adj}/n
(>)    tha i fliuch :- s{clause=dcl, phon=cons, tense=pres}
```

In the grammar `bi_arg` stands for a `clause` feature value of either `asp` or `adj`, indicating which sorts of clause can be an argument for the verb *bi*.

For development purposes we use the interactive parser `tccg`.

3.3 Towards evaluation

Clark and Hockenmaier (Clark & Hockenmaier, 2002), in the context of CCGbank, compare methods for evaluating the performance of a CCG system. These involve the CCG system being able to output dependencies, whether they be the Universal Dependencies mentioned earlier or ones obtained directly from the steps in a CCG derivation, and comparing those dependencies to a gold standard. This allows for a systematic check of not only whether the correct parts of speech have been assigned, but also, for example, subjects, objects and PP attachment. In contrast, the default testing framework for OpenCCG involves counting the number of parses for a given sentence and comparing it with the expected number. This is useful for pedagogical reasons, but knowing that the correct number of parses has been returned for a sentence is less helpful than knowing how much of it was assigned correctly. A further difficulty is that parsing a sentence in CCG is equivalent to deriving a proof, and if that proof fails for whatever reason, then there is no way of recovering the partial parses to award partial credit to the parser. Hence the program both flatters successful parses and unduly penalizes unsuccessful ones, and so we have not been able to provide a sensible evaluation of the parsing performance. Lastly, because the CCG parser doesn't handle out-of-vocabulary text, we cannot have separate training and testing data.

We can, however, give a qualitative account of the situations where more work is needed. Our examination has focussed on the section of ARCOSG consisting of news scripts from Radio nan Gàidheal, a genre which has been described in detail by Lamb (1999). This section has 11354 tokens and is about 13% of the total 87038. It is amenable to automatic sentence-splitting and does not contain interjections or direct speech, which make parsing harder. The grammar works accurately on simple clauses based on transitive and intransitive verbs, relative clauses and passives formed with the verb *rach*.

Apposition, despite the measures above to deal with punctuation, is still not fully handled. *Rùnaire Èirinn a Tuath Mo Mowlam* 'Northern Ireland Secretary Mo Mowlam', for example, doesn't parse. Similarly if there is a sequence of words tagged as 'foreign', which are treated as nouns for simplicity, then the whole parse will fail. Sequences of nominative nouns also occur in temporal and spatial expressions and chains of possession where only the last noun is grammatically marked as genitive.

Cosubordination, a sort of coordination where the coordinated clause can express, among other things, reason, *dh'fhalbh Alasdair agus i 'na suain*—"Alasdair left because she was fast asleep" or time, is, contrary to initial suspicions, attested in the news subcorpus. *Chaidh bratach Bhreatainn a thoirt a-nuas ann an seirbheis taobh muigh an taighe, 's an Last Post ga chluiche* 'The British flag was taken down in a service outside the house as the Last Post was played' exemplifies this. The conjunction *'s* 'and' joins a *rach* passive to a non-constituent. We anticipate that it should be possible to handle this elegantly in CCG using type-raising rules such as we have seen previously, but this is future work.

4 Conclusions and future work

We have produced a medium-coverage categorial grammar of Scottish Gaelic using all of the Annotated Reference Corpus of Scottish Gaelic and where every type is assigned based solely on the token value and its POS tag. The key difficulty has been in providing a convincing evaluation of the foregoing. To this end we need firstly a gold standard corpus of dependencies, of the sort we previously presented in Batchelor (2014) which can be used to evaluate successful parses. The other key requirement is to migrate to a statistical approach, ensuring that there are some successful parses to evaluate. A conventional CCG workflow involves a statistical supertagging stage prior to parsing. Supertagging is similar to POS-tagging but typically uses a larger tagset. Whereas the focus in the ARCOSG POS set is on morphological features, supertags can indicate subcategorization, whether a PP modifies a noun or a clause, or whether a comma is appositive or not, among other functions. The C&C supertagger for English uses around 500 supertags as opposed to 50 Penn Treebank POS tags. As such, the problems described in Lamb & Danso (2014) with ordinary POS-tagging in Scottish Gaelic will be harder for supertagging, but it seems plausible that because of different focus, the number of supertags required for Gaelic will be similar to that for English. A working solution to this would also handle the problems of out-of-vocabulary text and foreign words described in the section above. The code, a small set of Python scripts is available at `https://github.com/colinbatchelor/gdbank/`.

Acknowledgements

Many thanks to William Lamb for a preview copy of ARCOSG, to Teresa Lynn for a critical reading of the manuscript, and to the anonymous referees for their very helpful suggestions.

Références

BALDRIDGE J., CHATTERJEE S., PALMER A. & WING B. (2007). DotCCG and VisCCG : Wiki and Programming Paradigms for Improved Grammar Engineering with OpenCCG. In T. H. KING & E. BENDER, Eds., *Proceedings of the GEAF 2007 Workshop* : CSLI Publications, Stanford, CA.

BALDRIDGE J. & KRUIJFF G.-J. M. (2003). Multi-Modal Combinatory Categorial Grammar. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary.

BANGALORE S. & JOSHI A. (1998). Supertagging : an approach to almost parsing. *Computational Linguistics*, **22**, 1–29.

BATCHELOR C. (2014). gdbank : The beginnings of a corpus of dependency structures and type-logical grammar in Scottish Gaelic. In *Proceedings of the First Celtic Language Technology Workshop*, p. 60–65, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

CLARK S. & HOCKENMAIER J. (2002). Evaluating a Wide-Coverage CCG Parser. In *Proceedings of the LREC 2002 Beyond Parseval Workshop*, p. 60–66, Las Palmas, Spain.

CURRAN J., CLARK S. & BOS J. (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*, p. 33–36, Prague, Czech Republic : Association for Computational Linguistics.

HOCKENMAIER J. (2006). Creating a CCGbank and a wide-coverage CCG lexicon for German. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, p. 505–512, Sydney, Australia.

HOCKENMAIER J. & STEEDMAN M. (2007). CCGBank : A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. *Computational Linguistics*, **33**, 355–356.

LAMB W. (1999). A diachronic account of Gaelic news-speak : The development and expansion of a register. *Scottish Gaelic Studies*, **XIX**, 141–171.

LAMB W. (2003). *Scottish Gaelic, 2nd edn*. Munich, Germany : Lincom Europa.

LAMB W. (2008). *Scottish Gaelic Speech and Writing : Register Variation in an Endangered Language*. Belfast : Cló Ollscoil na Banríona.

LAMB W., ARBUTHNOT S., NAISMITH S. & DANSO S. (2016). *Annotated Reference Corpus of Scottish Gaelic (ARCOSG), 1997–2016 [dataset]*. Rapport interne, University of Edinburgh ; School of Literatures, Languages and Cultures ; Celtic and Scottish Studies. `http://dx.doi.org/10.7488/ds/1411`.

LAMB W. & DANSO S. (2014). Developing an Automatic Part-of-Speech Tagger for Scottish Gaelic. In *Proceedings of the First Celtic Language Technology Workshop*, p. 1–5, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

LYNN T. (2016). *Irish Dependency Treebanking and Parsing*. PhD thesis, Dublin City University and Macquarie University.

LYNN T., SCANNELL K. & MAGUIRE E. (2015). Minority Language Twitter : Part-of-Speech Tagging and Analysis of Irish Tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, p. 1–8, Beijing, China : Association for Computational Linguistics.

MĚCHURA M. B. (2014). Irish National Morphology Database : a high-accuracy open-source dataset of Irish words. In *Proceedings of the First Celtic Language Technology Workshop*, p. 50–54, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

NAISMITH S. & LAMB W. (2014). Scottish Gaelic Part-of-Speech Annotation Guidelines. Celtic and Scottish Studies, University of Edinburgh.

NIVRE J. *et al.* (2015). Universal Dependencies 1.2. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

SCANNELL K. (2014). Statistical models for text normalization and machine translation. In *Proceedings of the First Celtic Language Technology Workshop*, p. 33–40, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

STEEDMAN M. & BALDRIDGE J. (2003). *Combinatory Categorial Grammar*. Rapport interne, University of Edinburgh. `http://homepages.inf.ed.ac.uk/steedman/papers/ccg/SteedmanBaldridgeNTSyntax.pdf`.

TSE D. & CURRAN J. R. (2010). Chinese CCGbank : extracting CCG derivations from the Penn Chinese Treebank. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, p. 1083–1091, Beijing, China.

Uí DHONNCHADHA E. (2009). *Part-of-speech tagging and partial parsing for Irish using finite-state transducers and constraint grammar*. PhD thesis, Dublin City University.

Uí DHONNCHADHA E. & VAN GENABITH J. (2005). Scaling an Irish FST morphology engine for use on unrestricted text. In *Fifth International Workshop on Finite-State Methods in Natural Language Processing*, Helsinki.

UNIVERSITY OF GLASGOW (2016). Corpas na Gàidhlig, Digital Archive of Scottish Gaelic (DASG), `http://www.dasg.ac.uk/corpus/`. accessed 15 April 2016.