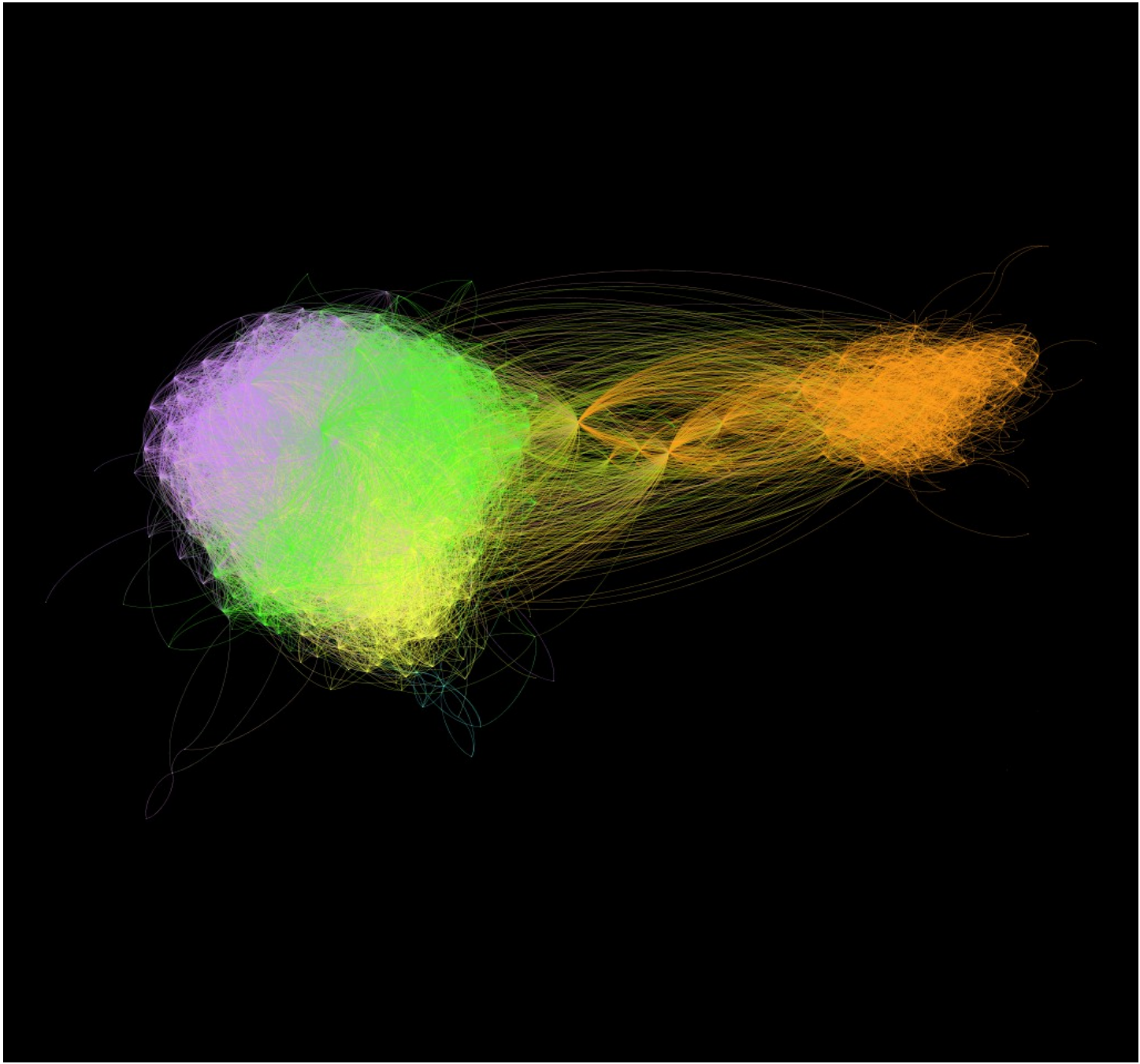


# Eadar-Ghaeilg: Scottish and Manx Gaelic resources for Irish speakers

Kevin Scannell  
Saint Louis University  
5 October 2015





# Eadar-Ghaeilg (sic)



“[Gàidhlig agus Gaeilge]... are not names for regional varieties Of Gaelic, they are regional names for the whole of Gaelic.”

Ciarán Ó Duibhín, <http://www.smo.uhi.ac.uk/~oduibhin/alba/ouch.htm>

# Why Machine Translation?

- One big community from 2 (or 3) small ones
- Break down communication barriers
- Open new markets for Gaelic writers
- Tools for learners, w/o going via English
- “Transfer” Irish language tech to Gaelic
- Enables comparative linguistic study
- Rehabilitate a much-maligned technology!
- Deliver MT results in creative, tailored ways

# Example 1

- gd: “Bha e fhéin 'na sheasamh a-measg a' bhuntàta an uair a chunnaic e iad le 'n gunnachan...”

# Example 1

- gd: “Bha e fhéin 'na sheasamh a-measg a' bhuntàta an uair a chunnaic e iad le 'n gunnachan...”
- ga: “Bhí sé féin ina sheasamh i measc na bprátaí nuair a chonaic sé iad lena gcuid gunnaí...”

# Example 2

- gd: “Chunnacas fo sgàil craobh na dòrainn a' coiseachd sràidean Pharais gu lòghmhor na seann siùrsaichean beaga breòite a chunnaic Baudelaire 'na ònrachd.”



# Example 2

- gd: “Chunnacas fo sgàil craobh na dòrainn a' coiseachd sràidean Pharais gu lòghmhor na seann siùrsaichean beaga breòite a chunnaic Baudelaire 'na ònrachd.”
- ga: “Chonacthas faoi scáth chrann an doilíosa ag siúl sráideanna Pháras go soilseach na striapaigh aosta bheaga bhuailte a chonaic Baudelaire ina uaigneas”
- Somhairle MacGill-Eain, aistr. Paddy Bushe

# Parallel corpus

- Aligned Irish-Scottish Gaelic texts
- A little bit of everything!
- Software translations, Bible texts, tweets
- Wikipedia articles, poems, prayers, ...
- 130k segments, ~1M words on each side
- Unusual in that very little is direct translation

# Statistical Machine Translation

- We extract translation pairs from parallel corpus
- Even easier when pairs are often cognates
- Model incorporates rule-based spelling changes
- sg- → sc, -chd- → -cht-, etc.
- Each SG word then maps to (usu. many) IG words

# How to translate?

- Work through the source sentence left-to-right
- Incorporate limited reordering via a “phrase table”
- e.g. “mun cuairt oirnn” → “inár dtimpeall”
- Each word/phrase has multiple possible translations
- Hence each sentence has *many* possible translations
- 20 word sentence, ~2 translations/word => ~1000000!
- Goal is to find the *most probable* translation
- Prune the possibilities via a “Markov model”

# Guessing Game, I

- Probability of word depends only on prev two
- “the two \_\_\_\_\_”
- $P(\text{men}|\text{the two}) = 0.0413$
- $P(\text{of}|\text{the two}) = 0.0338$
- $P(\text{countries}|\text{the two}) = 0.0298$
- $P(\text{sides}|\text{the two}) = 0.0204$
- $P(\text{groups}|\text{the two}) = 0.0164$

# Guessing Game, II

- “the fact \_\_\_\_\_”
- $P(\text{that}|\text{the fact}) = 0.8698$
- $P(\text{is}|\text{the fact}) = 0.0312$
- $P(\text{of}|\text{the fact}) = 0.0241$
- $P(\text{remains}|\text{the fact}) = 0.0092$
- $P(\text{was}|\text{the fact}) = 0.0050$
- $P(\text{they}|\text{the fact}) = 0.0043$

# Guessing Game, III

- “the united \_\_\_\_\_”
- $P(\text{states}|\text{the united}) = 0.5240$
- $P(\text{kingdom}|\text{the united}) = 0.3129$
- $P(\text{nations}|\text{the united}) = 0.0859$
- $P(\text{arab}|\text{the united}) = 0.0075$
- $P(\text{front}|\text{the united}) = 0.0061$
- $P(\text{democratic}|\text{the united}) = 0.0024$

# Guessing Game, IV

- “button fell \_\_\_\_\_”
- Doesn't appear at all in a 100M word corpus
- “Backoff smoothing”
- Estimate  $P(w|\text{button fell})$  using  $P(w|\text{fell})$
- Or get a bigger corpus!



# Web Corpora

- No linguistic analysis of source lang needed
- No statistics for source language either!
- Entirely driven by statistics of target (Irish)
- Of which there is A LOT online
- About 150 million words of Irish in total
- Some care needed, but basically more=better

# Disambiguation I: Lexical Gaps

- “ach coiseachd an iar tron Mhunadh Gheal”
- “am biodh a' ghaoth an iar leotha...”
- “air a' chosta an iar...”
- Give “siar”, “aniar”, “thiar” respectively

# Disambiguation II: Function Words

- gd: Bhitheamaid gun sgillinn ruadh nan dèanamaid sin
- “nan” -> na, ina, or dá
- ga: Bheimis gan pingin rua dá ndéanfaimis sin

# Disambiguation III: Initial mutations

- gd #1: “Chroch sibh an radan”
- ga: “{Chroch,gCroch} {sibh,tú} {a,an,in} {francach,fhrancach,bhfrancach}”
- gd #2: “Beir air an radan”
- ga: “{Beir,mBeir} {air,ar} {a,an,in} {francach,fhrancach,bhfrancach}”
- $2 \times 2 \times 3 \times 3 = 36$  possible translations in each case
- Best for #1: “Chroch sibh an francach” (or “tú”?)
- Best for #2: “Beir ar an bhfrancach”

# Deliverables

- Gàidhlig-Gaeilge dictionary: 16107 headwords
- Gàidhlig twitter stream for Irish speakers
- InterGaelic.com, with Michal Boleslav Měchura
- Integrated into Clilstore and Multidict
- Open source code and data for translator
- Translations via a web service
- Two popular Gàidhlig books being translated

# Evaluation

- Test set of 593 sentences translated directly
- WER (“word error rate”) 0.3740
- Baseline (do nothing!) 0.8809
- Still, naive evaluation is a bit “unfair”
- Initial mutations
- Structural differences:
- “Tha mi a' tuigsinn a-nis” vs. “Tuigim anois”

# Dictionary

## **bàta-slaodaidh**

---

**bàta-slaodaidh**, *fir*: tuga.

**bàta-smùide**, *fir*: galbhád.

**bàta-teasairginn**, *fir*: bád tarrthála.

**bàth**, *br*: báigh.

**bàthach**, *bain*: bóitheach.

**bàthadh**, *fir*: bá.

**bàthaich**, *fir* → *bàthach*, *bain*.



**bathais**, *bain*: éadan.

†**Bathamach**, *aid*: Bahámaíoch.

†**Bathamach**, *fir*: Bahámaíoch.

★**bathar**, *fir*: earra, táirge.

**bathar-bhog**, *fir* → *bathar-bog*, *fir*.

**bathar-bog**, *fir*: bogearra.

**bathar cruaidh**, *fir*: crua-earra.

**batharnach**, *fir*: stóras.

**bàthte**, *aid*: báite.

**bàt-iasgaich**, *fir* → *bàta-iasgaich*, *fir*.

**bàt'-iasgaich**, *fir* → *bàta-iasgaich*, *fir*.

**b' e** ba é.

**beach**, *fir*: beach.

†**beachair**, *fir*: beachaire.



**beachd**, *fir*: tuairim, barúil.


# intergaelic.com

FOCLÓIR

AISTRÍUCHÁN

Tha ceannard ionmhais NHS na Gàidhealtachd ag ràdh gur dòcha gun tèid iad £5m thairis air a' bhuidseat aca ro dheireadh na bliadhna-ionmhais, mura tèid aca air cosgaisean a ghearradh mar a tha còir.

A rèir Nick Kenton, tha cosgaisean a bharrachd an lùib

 Aistrigh »

Tha ceannard ionmhais NHS na Gàidhealtachd ag ràdh gur dòcha gun tèid iad £ 5m

Tá ceannaire chiste NHS na Gaeltachta ag rá gur dócha go rachaidh siad £ 5m

thairis air a' bhuidseat aca ro dheireadh na bliadhna-ionmhais, mura tèid aca air

thairis an bhuiséad acu roimh dheireadh na bliadhna-ionmhais, mura rachaidh acu ar

cosgaisean a ghearradh mar a tha còir.

táillí a ghearradh mar atá ceart.

A rèir Nick Kenton, tha cosgaisean a bharrachd an lùib seirbheisean a chumail ri

De réir Nick Kenton, tá costas sa bhreis ceangailte le seirbhísí a choinneáil le

euslaintich san iar-thuath agus Ospadal an Ràthaig Mhòir a' cur ris an uallach a th' orra.

othair san iarthuaisceart agus Ospidéal an Ràthaig Mhóir ag cur leis an ualach atá orthu.



# Twitter stream

- 
- BBCSpors** BBC Spòrs Gaelic  
<http://t.co/rtrxR8FFYd> Rugbaidh #PRO12 | Dùn Èideann v Ospreys |  
Tòiseachadh aig 5.50 uf air @bbcalba | @HughDan1956 @calum\_macaulay  
1 lá ó shin ↩ Freagra ↗ Atweetáil ☆ Réiltín
- 
- akerbeltzalba** Akerbeltz  
Duilich tha coltas gu bheil òstair an fhaclair \* air fad \* shìos, fiù an làrach  
aca-san. Bidh sinn air ais cho luath 's a ghabhas.  
1 lá ó shin ↩ Freagra ↗ Atweetáil ☆ Réiltín suíomh
- 
- bbcnaidheachdan** BBC Naidheachdan  
Aithris bhideo: Morair Mhinginis ag ràdh g'eil foghlam tro mheadhan na  
Gàidhlig a' toirt fein aithne air ais... <https://t.co/0A0HpeMjXr>  
1 lá ó shin ↩ Freagra ↗ Atweetáil ☆ Réiltín
- 
- bbcnaidheachdan** BBC Naidheachdan  
Aithris bhideo: Ceist mu dè a bu chòir tachairt do dh'aitreabh Sgoil t-Oib  
anns na Hearadh. <https://t.co/JcndcvZNKZ>  
1 lá ó shin ↩ Freagra ↗ Atweetáil ☆ Réiltín
- 
- BBCAimsir** BBC Aimsir  
Chuala sibh mu bogha frois ach dè mu dheidhinn bogha ceò? Abair dealbh  
inntinneach! <https://t.co/98rsSaOllx>  
1 lá ó shin ↩ Freagra ↗ Atweetáil ☆ Réiltín

# Clilstore

Clilstore Guthan nan Eilean Iain Tormod MacLeòid Alex, Kathleen & Mìcheal Unit info

## Foghlam Fad Beatha agus Sabhal Mòr Ostaig



Chaidh Sabhal Mòr Ostaig a stèidheachadh mar cholaiste Ghàidhlig ann an 1973 ann an Slèite san Eilean Sgitheanach. Thairis air na bliadhnaichean tha e air fàs gu luath, le togalaichean ùra aig Àrainn Ostaig an toiseach, agus an uair sin faisg air làimh aig Àrainn Chaluim Chille, le **seallaidhean** brèagha air an Linne Shleibhteach.

Multidict navigation frame Help About

Word to translate  Go Multidict will try these wordforms in rotation (on r

**sealladh** ← seallaidhean →

From ↔ To Dictionary Esc

Gàidhlig (gd) Gaeilge (ga) Intergaelic

---

INTERGAELIC

FOCLÓIR AISTRÍÚCHÁN

seall | fealladh | gealladh | mealladh

---

 **sealladh**

AINMFHOCAL FIRINSCNEACH  
faclair.com »

 **amharc**  
potafoal.com »

 **radharc**  
potafoal.com »

# Gaelg/Manx Gaelic

- Part of the Gaelic dialect continuum
- Orthography very different ga/gd
- Declared “extinct” by UNESCO
- Being revived now, including immersion school
- Probably less than 500 fluent speakers

# Three-Way Comparison

- ga: “Níl leigheas agat air sin, arú,” a dúirt an Cat. “Táimid go léir as ár meabhair anseo.”
- gd: “O, chan eil atharrach ann,” ars an Cat. “Tha sinn uile às ar ciall an seo.”
- gv: “Ogh, cha nel niart ayd er shen,” dooyrt y Kayt. “Ta shin ooilley keoi ayns shoh.”

# Another Parallel Corpus

- The full Manx Bible available!
- Alice in Wonderland
- ECRML
- Software translations, lexical material, ...
- 61k segments, ~850k words on each side

# Disambiguation

- Even more challenging b/c of reduced spelling
- “Diu y kayt ooilley'n bainney oor”
- Manx “diu” is either Ir. “d'ól (d'ibh)” or “daoibh”
- Manx “oor” is either Ir. “úr” or “uair”
- “D'ól an cat an bainne úr go léir”
- “Beeym ersooyl tree kerroo oor”
- “Beidh mé ar siúl trí cheathrú uair (an chloig)”

# Progress

- Parallel corpus complete
- Bilingual dictionary 48% complete
- Launch planned March 2016
- Same set of deliverables

# Ar Ghuailí na bhFathach

- Michael Bauer
- Caoimhín Ó Donnaíle
- Donncha King
- Ciarán Ó Duibhín
- Ciarán Dunbar
- Phil Kelly
- Michal Boleslav Měchura
- Eoin Ó Murchú
- Adrian Cain
- Brian Stowell
- Gearóid Ó Néill