

Machine Learning and Technology for Endangered Languages

Kevin Scannell
Saint Louis University
16 November 2015

Web as Linguistic Archive

- Billions of unique resources, +1
- Raw data only, no annotations, -1
- Wide range of languages represented, +1
- Very little material in endangered languages, -1
- Full text searchable, +1
- Can't search archive by language, -1
- Snapshots archived periodically, +1
- Resources removed or changed at random, -1
- Free to download, +1
- Usage rights unclear for vast majority of resources, -1

An Crúbadán: History

- First attempt at crawling Irish web, Jan 1999
- 50M words of Welsh for historical dict., 2004
- ~150 minority languages, 2004-2007
- ~450 languages for WAC3, 2007
- Unfunded through 2011
- Search for “all” languages, started c. 2011

How many languages?

- Finishing a 3-year NSF project
- Phase one: aggressively seek out new langs
- Phase two: produce free+usable resources
- Current total: 2168
- At least 200 more queued for training
- 3000?

Design principles

- Orthographies, not languages
- Labelled by BCP-47 codes
- en, chr, sr-Latn, de-AT, fr-x-nor, el-Latn-x-chat
- Real, running texts (vs. word lists, GILT)
- Get “everything” for small languages
- Large samples for English, French, etc.

How we find languages

- Lots of manual web searching!
- Special code monitors WP, JW, UN, bible.is
- Typing/OCR of scanned or offline texts
- Build statistical language ID models
- Special thank to Ed Jahn, George Mason
- NSF grant 1159174

Adding Value

- Separating orthographies/dialects
- Clean boilerplate text
- Convert to UTF-8 text + normalize
- Sentence segment and tokenize
- Produce formats useful for NLP












Three modules

- Traditional web crawler
- Twitter crawler
- Blog tracker

Twitter crawler

- Twitter's REST API
- Seed searches with words from web corpora
- Language ID particularly challenging
- Crawl social graph to find new tweets
- <http://indigenoustweets.com/>



Erabiltzailea	Euskara	Guztira	% Euskara	Jarraitzaileak	Jarraitzen	Azken tuita
1 berria 	42771	45884	93.2	18486	708	2013-08-05 15:47:30
2 txargain 	38964	53350	73.0	953	338	2013-08-05 23:17:30
3 euskalherrian 	29203	57040	51.2	4672	76	2013-08-06 04:15:58
4 toki_kom 	28922	31541	91.7	476	51	2013-08-06 01:14:02
5 eitbcomBerriak 	25525	26960	94.7	5088	188	2013-08-05 19:59:26
6 joseba01 	17330	29673	58.4	435	502	2013-08-06 00:12:08
7 theklaneh 	16242	33797	48.1	1795	355	2013-08-05 03:50:26
8 argia 	15284	16818	90.9	9598	1394	2013-08-05 21:15:57
9 euskaljakintza 	14225	24857	57.2	5176	1225	2013-08-05 07:24:48
10 joxe 	12976	22632	57.3	1578	900	2013-08-05 23:17:13
11 goiena 	12576	13932	90.3	1960	341	2013-08-05 14:07:54

Ag feitheamh le a0.twimg.com...

INDIGENOUS TWEETS.COM

Euskara

Pil-pilean:

- kariiiiis
- antraxa
- #goraligoteosana
- Carrow
- apurtzeraaaaaaaa
- zantzelekn
- ceditutie
- #hiroshimagogoan
- erritmoen
- mortaaaaaaaaaaaaa

Norbait falta da?
Twitter erabiltzaile-izena:

@ Bidali

Esaiozu munduari hemen zaudela:



Blog

Indigenous Blogs

Kevin Scannell



Blog tracker



- Blogger platform only
- Works hand-in-hand with traditional crawler
- Registers all blogs with an in-language post
- Tracks all past and future posts
- <http://indigenusblogs.com/>

Deliverables

- See <http://crubadan.org/>
- Word and bigram frequency lists
- Character trigram frequencies
- Lists of URLs in each language
- Discoverable as an OLAC repository

Spelling and grammar checkers

- Corpus-based Irish spellchecker, 2000
- Grammar checker, 2003
- 28 new spellcheckers since 2004
- Collaborations with native speakers
- All under open source licenses

Computational Morphology

- Finite-state transducers (Xerox FST, foma)
- Very fast + bidirectional
- Cover most morphology of human langs
- Hunspell uses “two-fold affix stripping”
- Morphological analysis only
- Not as powerful, theoretically
- But, simpler formalism, better user support

Powerful enough?

- Hungarian
- Northern Sámi
- Basque
- Lingala, Kinyarwanda, Swahili, Chichewa
- Nishnaabemwin

Language ID

- Component *and* an application of Crúbadán
- Character n-grams + word models
- NLTK 3-gram data set
- Indigenous Tweets and Blogs

Predictive text

- T9 input
- Adaptxt
- Firefox OS
- Dasher



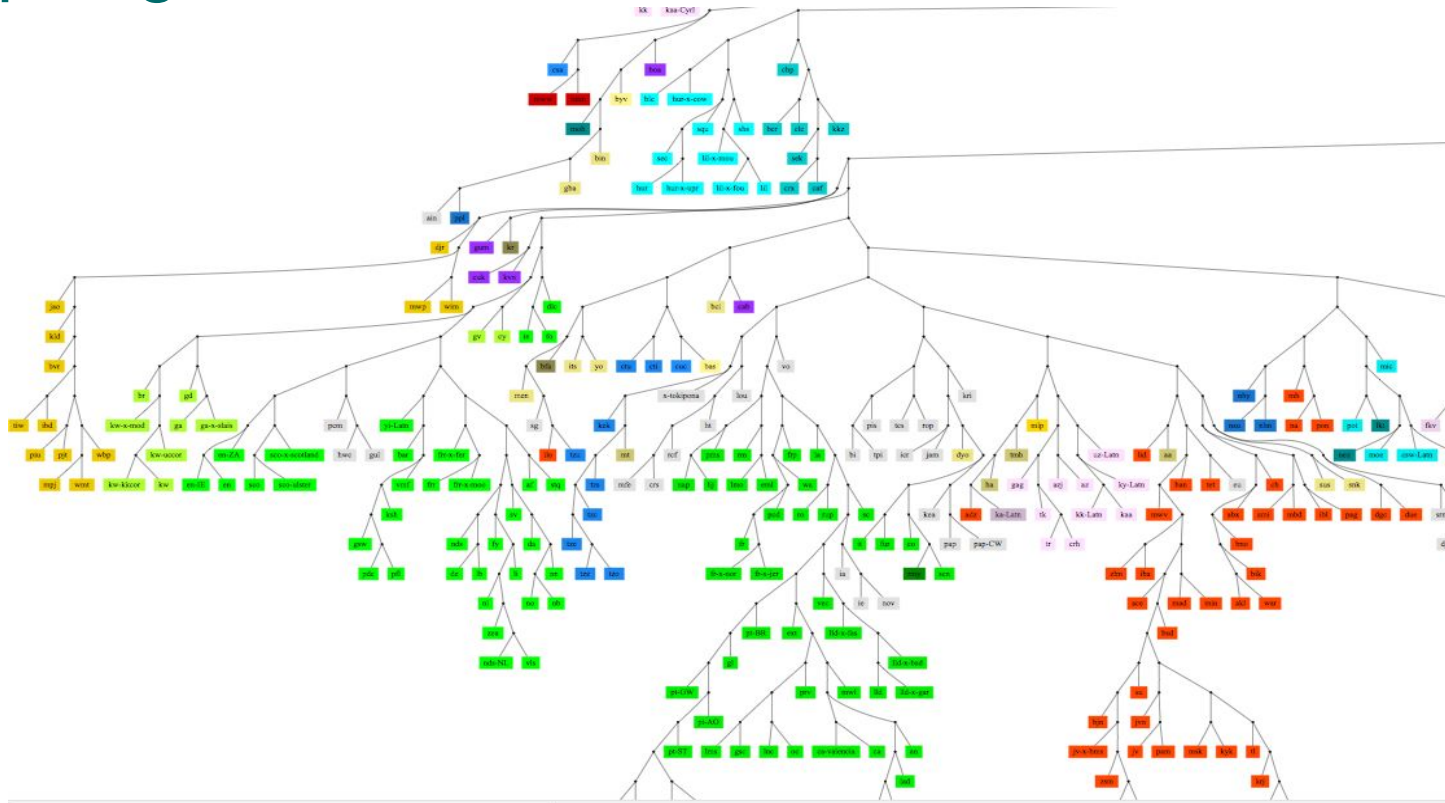
accentuate.us



- Web service for diacritic restoration
- Eni kookan lo ni eto si omi nira lati ni imoran ti o wu u, ki o si so iru imoran bee jade
- Ènì kòòkan ló ní ètọ sí òmì nira láti ní ìmọ̀ràn tí ó wù ú, kí ó sì sọ irú ìmọ̀ràn bẹ̀ẹ̀ jáde
- End-user clients for Firefox, LibreOffice
- Perl, Python, Haskell libraries
- Joint work with Michael Schade

Orthotree

- <http://indigenoustweets.blogspot.com/2011/12/>
- <https://github.com/kscanne/orthotree>



N-gram language models

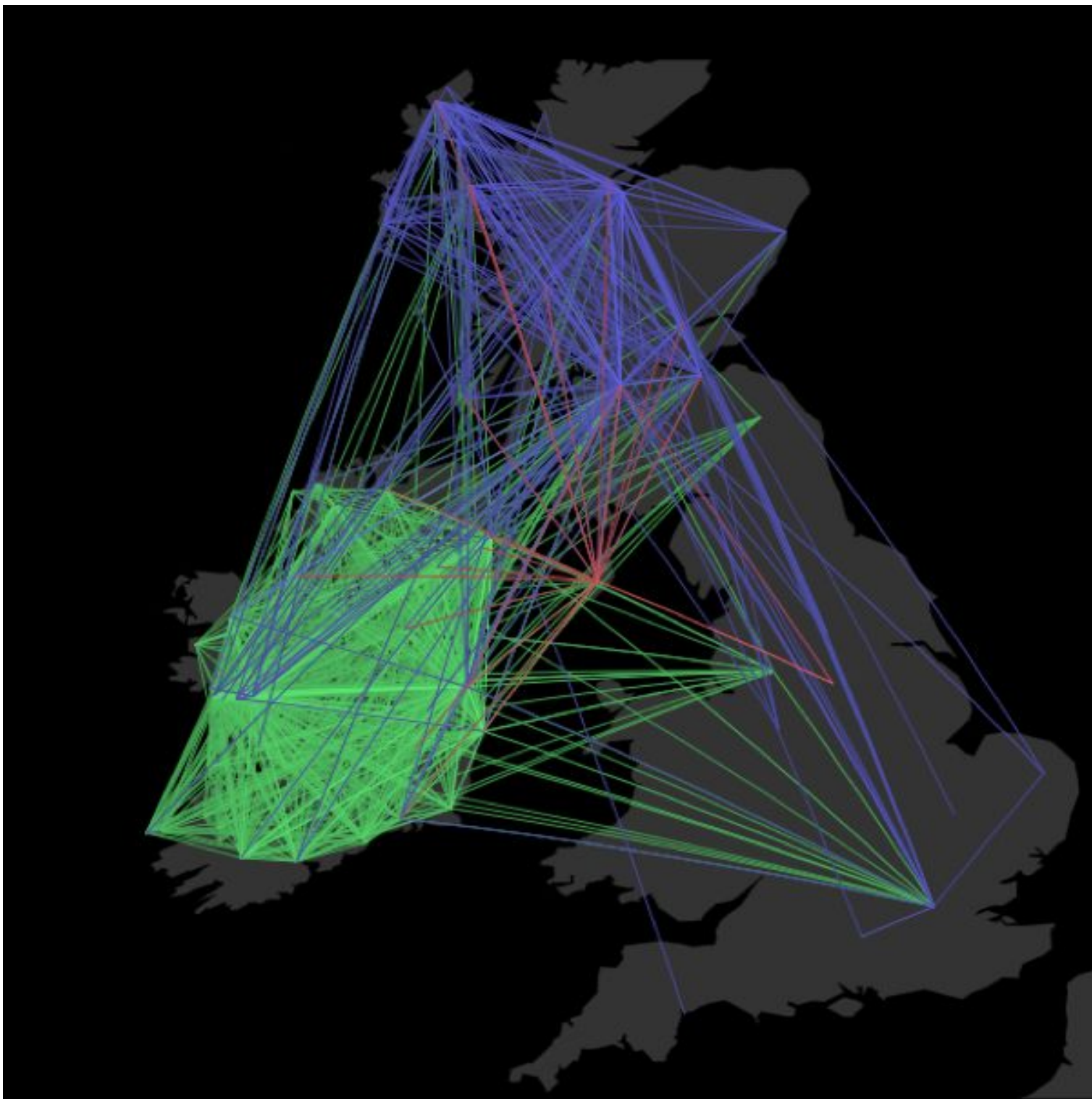
- Opens NLP to 100's of languages
- Noisy channel model
- Spelling and grammar checking
- OCR correction (e.g. old Irish fonts)
- Speech recognition
- Machine translation and related tasks

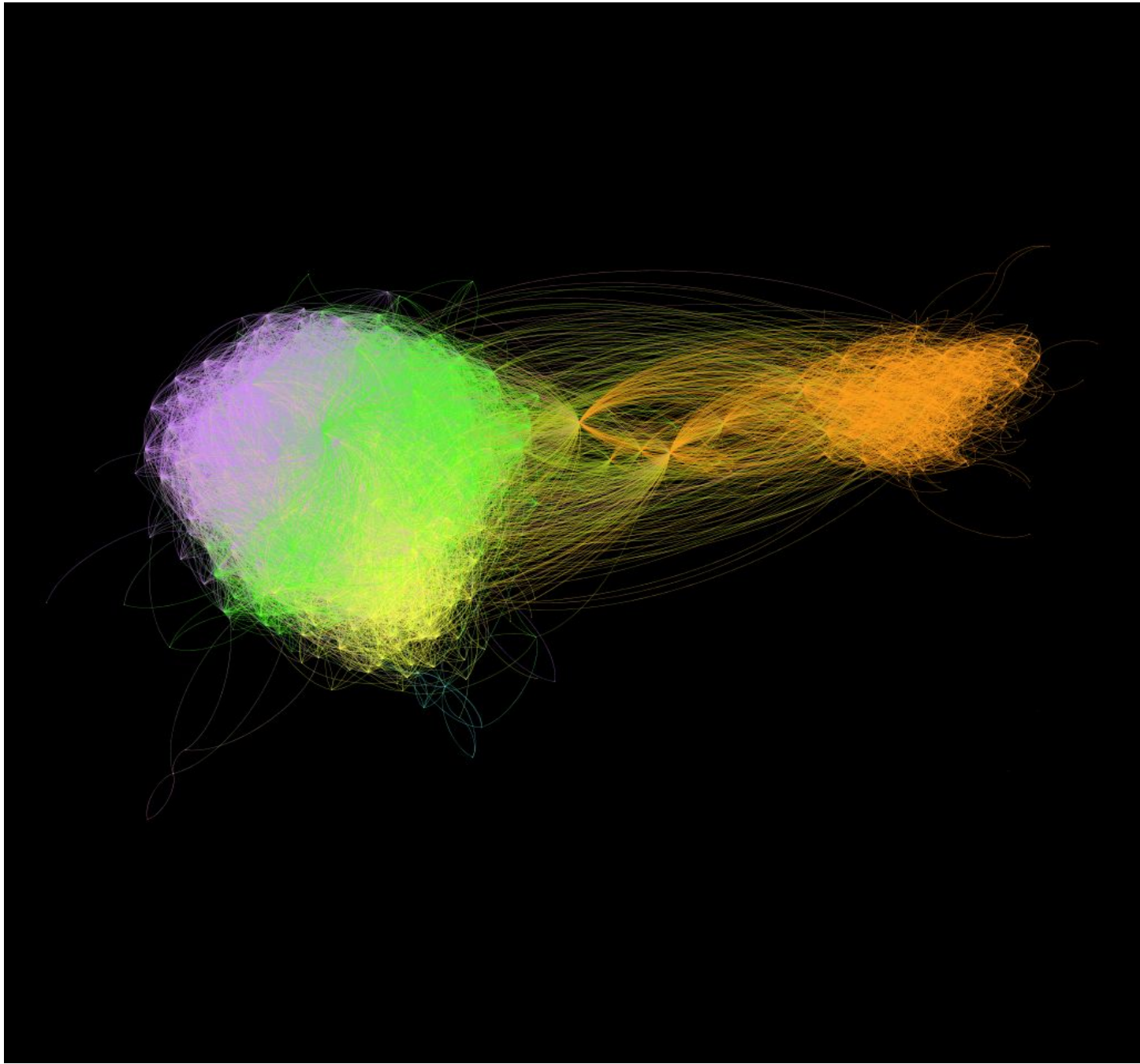
Irish standardization

- Irish spelling/grammar reform in the late 40s
- Presents huge problems for Irish NLP
- Search (web texts, corpora for lexicography)
- Using old texts for NLP (parallel corpora)
- Solution is to treat this as an MT problem
- “IBM Model 1” + spelling changes
- In use at two big Irish dictionary projects

Intergaelic

- “Q-Celtic”: Irish, Scottish, Manx Gaelic
- Scottish to Irish similar to standardization
- Same engine works
- Irish is strongest of the 3; big n-gram models
- Facilitates communication in social media
- Open to big “market” of Irish speakers





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

Cilstore

Cilstore 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 Chalum 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

sealladh

seall | fealladh | gealladh | mealladh

sealladh

AINMFHOCAL FIRINSCNEACH
faclair.com »

amharc
potafoocal.com »

radharc
potafoocal.com »

intergaelic.com

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 Mhòir ag cur leis an ualach atá orthu.