Celtic Language Technology and Open Data

Kevin Scannell
Saint Louis University
23 August 2014

Manx Gaelic

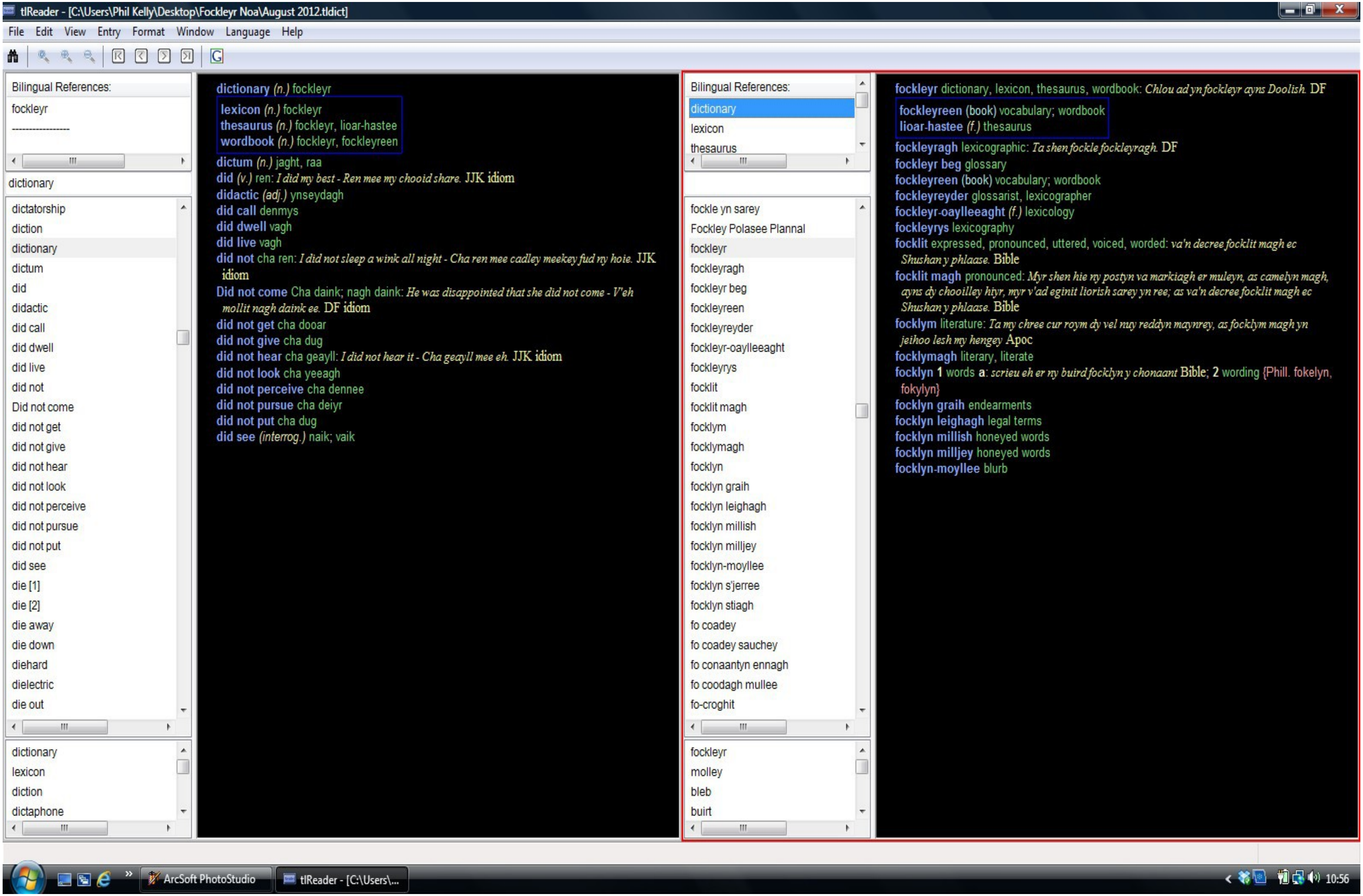
- Q-Celtic language
- Similar to Irish and Scot Gaelic
- With its own “idiosyncratic” orthography
- Probably less than 500 fluent speakers
- Immersion schools reviving the language

Example

- Manx: As tá mee myrgeddin grá rhyt,
  Dy nee uss Peddyr, as dy nee er y chreg shoh
  trog-ym's my agglish
- Irish: Agus deirimse leatsa
  gur tú Peadar agus is ar an gcarraig seo
  a thógfaidh mé m'eaglais
- Scots: Agus tha mise ag ràdh riut,
  Gur tusa Peadar, agus air a' charraig so
  togaidh mise m'eaglais

HLTs: Dictionaries

- Lexical database (TshwaneLex) by Phil Kelly
- http://homepages.manx.net/gaelg/
- iPhone, iPad: “Idioma” app
- Android: “GoldenDict” (free, open source)
- tlReader and GoldenDict on Windows



dictionary *(n.)* fockleyr

lexicon *(n.)* fockleyr
thesaurus *(n.)* fockleyr, lioar-hastee
wordbook *(n.)* fockleyr, fockleyreen

dictum *(n.)* jaght, raa
did *(v.)* ren: *I did my best - Ren mee my chooid share.* JJK idiom
didactic *(adj.)* ynseydagh
did call denmys
did dwell vagh
did live vagh
did not cha ren: *I did not sleep a wink all night - Cha ren mee cadley meekey fud ny hoie.* JJK idiom
Did not come Cha daink; nagh daink: *He was disappointed that she did not come - V'eh mollit nagh daink ee.* DF idiom
did not get cha dooar
did not give cha dug
did not hear cha geayll: *I did not hear it - Cha geayll mee eh.* JJK idiom
did not look cha yeeagh
did not perceive cha dennee
did not pursue cha deiyr
did not put cha dug
did see *(interrog.)* naik; vaik

fockleyr dictionary, lexicon, thesaurus, wordbook: *Chlou ad yn fockleyr ayns Doolish.* DF

fockleyreen (book) vocabulary; wordbook
lioar-hastee *(f.)* thesaurus

fockleyragh lexicographic: *Ta shen fockle fockleyragh.* DF
fockleyr beg glossary
fockleyreen (book) vocabulary; wordbook
fockleyreyder glossarist, lexicographer
fockleyr-oaylleeaght *(f.)* lexicology
fockleyrys lexicography
focklit expressed, pronounced, uttered, voiced, worded: *va'n decree focklit magh ec Shushan y phlaase.* Bible
focklit magh pronounced: *Myr shen hie ny postyn va markiagh er muleyn, as camelyn magh, ayns dy chooilley hiyr, myr v'ad eginit liorish sarey yn ree; as va'n decree focklit magh ec Shushan y phlaase.* Bible
focklym literature: *Ta my chree cur roym dy vel nuy reddyn maynrey, as focklym magh yn jeihoo lesh my hengey* Apoc
focklymagh literary, literate
focklyn 1 words **a**: *scrieu eh er ny buird focklyn y chonaant* Bible; **2** wording {Phill. fokelyn, fokylyn}
focklyn graih endearments
focklyn leighagh legal terms
focklyn millish honeyed words
focklyn milljey honeyed words
focklyn-moyllee blurb

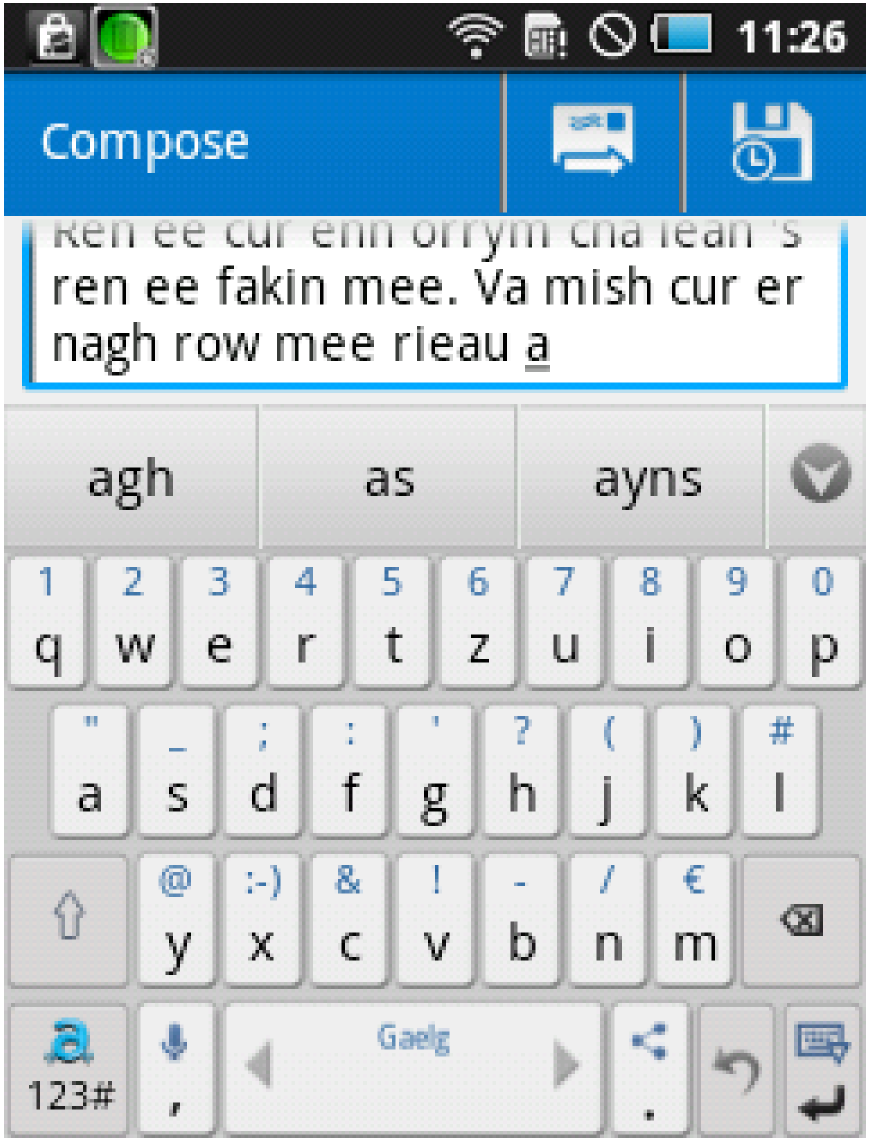
HLTs: Corpus

- Full Bible online (the best parallel material)
- 4791 articles in Manx Wikipedia, ~400k words
- 81405 tweets in Manx, 1.5M words
- Some blogs; about 120k words
- Texts/recordings/videos at learnmanx.com
- Skeealyn Vannin; transcribed audio
- In all, maybe 3M words of electronic text

HLTs: First Steps

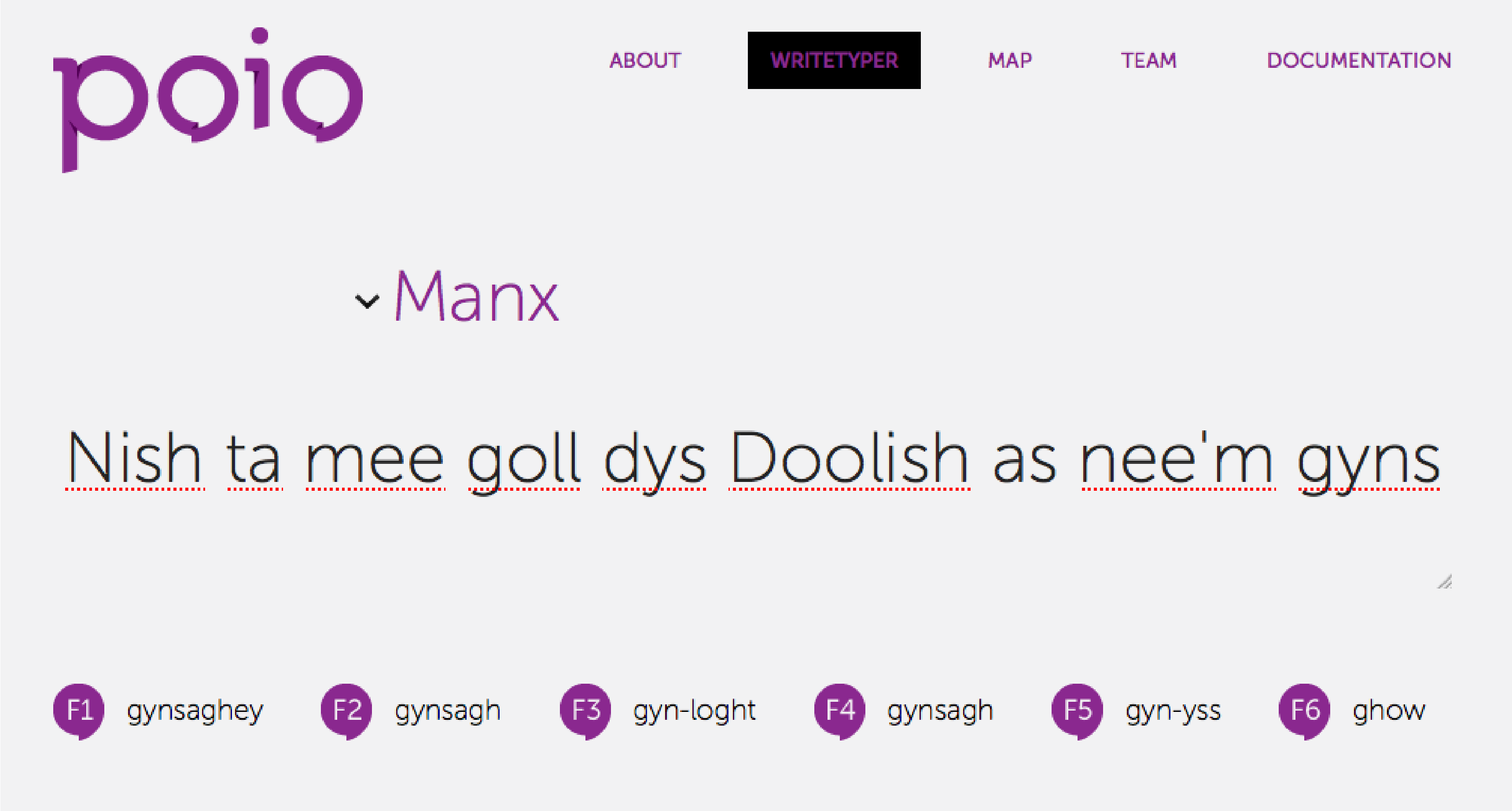
- Open source spellchecker
- Built using Phil's dictionary + web corpus
- Predictive text with word/character freqs
- Firefox OS, Adaptxt, Poio
- Thanks to Michael Bauer, Adrian Cain

Adaptxt



Ren ee cur enn orrym cha lean 's
ren ee fakin mee. Va mish cur er
nagh row mee rieau a

poio.eu



Nish ta mee goll dys Doolish as nee'm gyns

HLTs: Pan-Celtic

- Cognate induction, work by Gearóid Ó Néill
- Celtic cognates database, C. Ó Donnaíle
- Irish-Manx dict for Apertium w/ Josh Glatt
- Huge potential to bootstrap advanced tools

Example: WordNet

- Líonra Séimeantach na Gaeilge
- http://borel.slu.edu/lsg/
- 36k headwords, 77k linked word senses
- 1000 page PDF “thesaurus” for end-users
- English WordNet, bilingual dict, parallel corpus
- Open source (as is the English WordNet)
- Critical for semantic analysis of Irish text
- Have same ingredients for Irish → Manx!

A Culture of Openness

- The fruits of labor(s) of love
- Brian Stowell, Phil Kelly, Adrian Cain, others
- Each has shared all resources freely
- More advanced tools built on that foundation
- Same is true about core NLP tools in Irish
- Elaine's analyser and tagger
- Michal: INMD and Gramadán
- Apertium

Open Linguistic Data

- Critical in context of endangered languages
- No reinventing of the wheel
- Future-proofing for new formats, paradigms
- Other people will do cool stuff with your data!
- cf. Monica Ward's work with Gramadóir
- 28 new spellcheckers from web freq. lists
- Much much more...

Linked Open Data

- Put data online, with an open license, as text
- CSV, txt, JSON
- You can do more too...
- “Open Linguistics Working Group”
- Semantic web tech applied to linguistics/NLP
- RDF/OWL triples as data representation format
- Your data linked with existing open datasets
- Good tools for querying/reasoning

Quo Vadis?

- Siri as Gaeilge anyone?
- Entire open source pipeline for doing this
- Open Speech Initiative (speech.kde.org)
- I have huge language models
- CMU/Sphinx (and others) for training/decoding
- Crowdsource and host data at voxforge.org