New Frontiers in Language Technology for Minority Languages

Kevin Scannell
Saint Louis University

Linguistic Diversity: The Numbers

- About 7100 languages spoken in the world (Ethnologue)
- Almost half are “endangered” (UNESCO)
- 2500-3000 have some online presence
- < 1000 written by the language community
- Wikipedias for about 300 languages
- Tweets in about 275 languages
- Spell checkers for about 180
- Google search interface translated into 150
- Google Translate in about 100

Language Technology: Examples

- Machine translation
- Speech recognition
- Predictive text
- Search engines
- Dialogue systems
- Spelling/grammar checking
- Text normalization (e.g. modernization)
- Optical character recognition

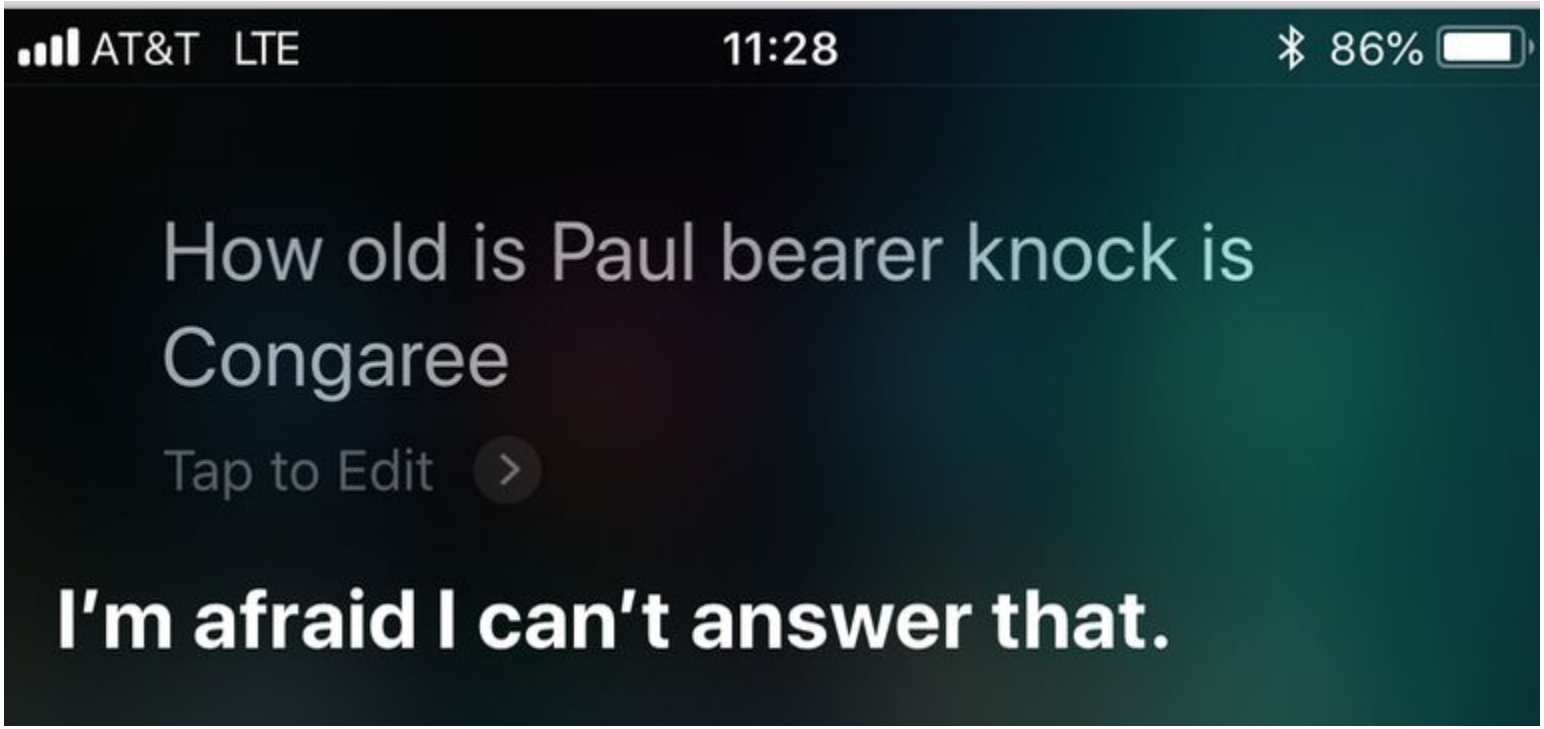
“Praistriúchán”

New Irish portmanteau word: “praiseach” = “a mess, a botch job”, “aistriúchán” = “translation”

Google used to translate English to Irish on 1916 commemorations website

Updated / Thursday, 13 Nov 2014 **23:48**

Siri in Irish?



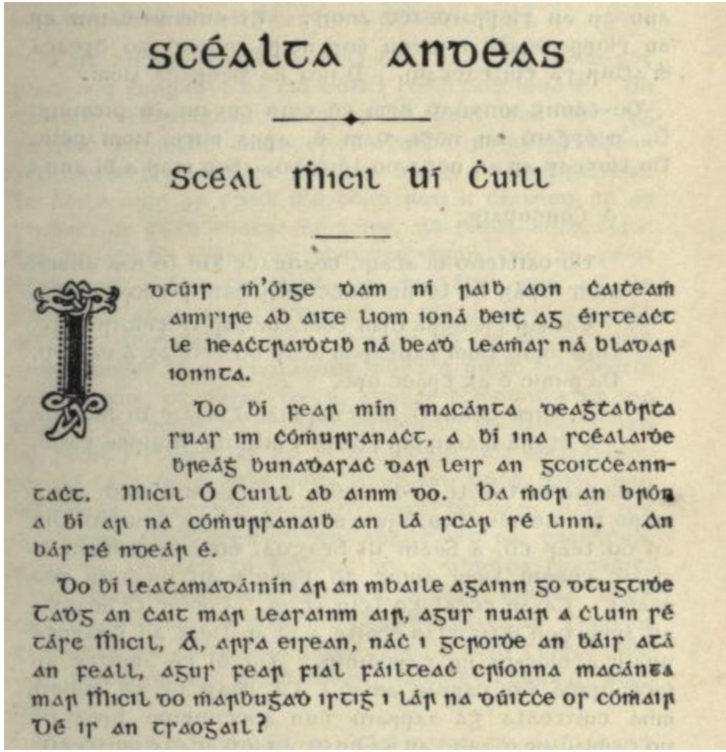
Google Search

FO/GRAI/ FOSTAI/OCHTA <EMPLOYMENT NOTICES> (ANOIS 2-3,9-10.xi.91)

1. post: Mu/inteoir (Ionadai/) le Mata & Eolai/ocht
   tre/imhse: Samhain go Ma/rta
   fo/n: +353-01-2825872
   seoladh: An Pri/omhoide
   Cola/iste Ra/ithi/n
   Bo/thar Florence
   Bre/
   Co.Chill Mhantain.

2. post: Pri/omhiniu/cho/ir Inmhea/nach
   spriocdha/ta: 15.11.91
   seoladh: An Ceannasai/ Riaracha/in Foirne
   An Post
   Ardoifig an Phoist
   Sra/id Ui/ Chonaill
   Baile A/tha Cliath 1.

Google Books


Scéalta Andeas

Scéal Ṁicil Uí Ċuill

I dTúis m'óige dam ní raib aon ċaiteaṁ aimsire ab aite liom ioná beiṫ ag éisteaċt le heaċtraiḋtib ná beaḋ leaṁas ná bladar ionnta.

Do bí fear mín macánta deaġtabṗta suas im ċóṁursanaċt, a bí ina scéalaiḋe breáġ bunaḋasaċ dar leis an gcoitċeann-taċt. Micil Ó Cuill ab ainm do. Ba mór an brón a bí ar na cóṁursanaib an lá scar sé linn. An bás fé ndeár é.

Do bí leaṫamaḋáinín ar an mbaile againn go dtugtiḋe Taḋg an ċait mar leasainm air, agus nuair a ċluin sé tásc Ṁicil, Á, arsa eisean, náċ i gcroiḋe an báis atá an feall, agus fear fial fáilteaċ críonna macánta mar Ṁicil do ṁarbuġaḋ istiġ i lár na ḋúiṫċe os cóṁair Dé is an tsaoġail?

```
SCéAlCA AnXDeAS

scéAi micit m ctiitt

(t^^S^ T)cúif tfi'óise 'óAni tií fAib Aon óAiteAtfi

^SP^ Aimpife At> xMce tiom lonÁ DeiC ^5 éifce-Aóc

jl: ^ iieAócf Ai'ócib nÁ beAt) teAtfiAf nÁ t)tAX>At^

jllj lonnCA.

^SS) T)o t)í peAt^ mín mACÁncA •oeA$t-At>t\tA

^^ fUAf im óórtiut^fAnAóc, -a bí inA fcéAtAi"óe

t)feÁ5 GunA'óAf Aó "OAt^ teif An ^coicóeAnii-

CAóc. ITIicit Ó Cuitt aX) -Ainm "oo. t)A tfiót^ ^n bttóu

A ttí A^A nA cómtit^fAnAiO An lÁ ycÁ\< fé tinn. Ati

bÁf f é nT)eÁt\ é.

"Oo t)í leACAmA'oÁinín At^ An mbAite Aj^Ainti 50 ■ociisci'óe
Ua'ós An ÓAIC m^t^ teAf Ainm -áit^, A^tif ntiAit^ a Otuin fé
cÁfc Thicit, A, AttfA eifeAn, nÁó 1 scttofóe Art bÁif acá
An feAtt, A^tif fe^tt fiAt fÁitceAó cttíonnA mACxinsA
m^tt fhicit 'oo tfiAttbuSAt) ifci$ 1 tÁf nA "oúitóe of cótfiAitt
"Oé if An cfAO^Ait?
```

What’s needed?

- **Bigger, better datasets**
- **ML models tailored to the linguistics**
- Technical capacity within communities
- Collaboration with Google, Facebook, Twitter, etc.

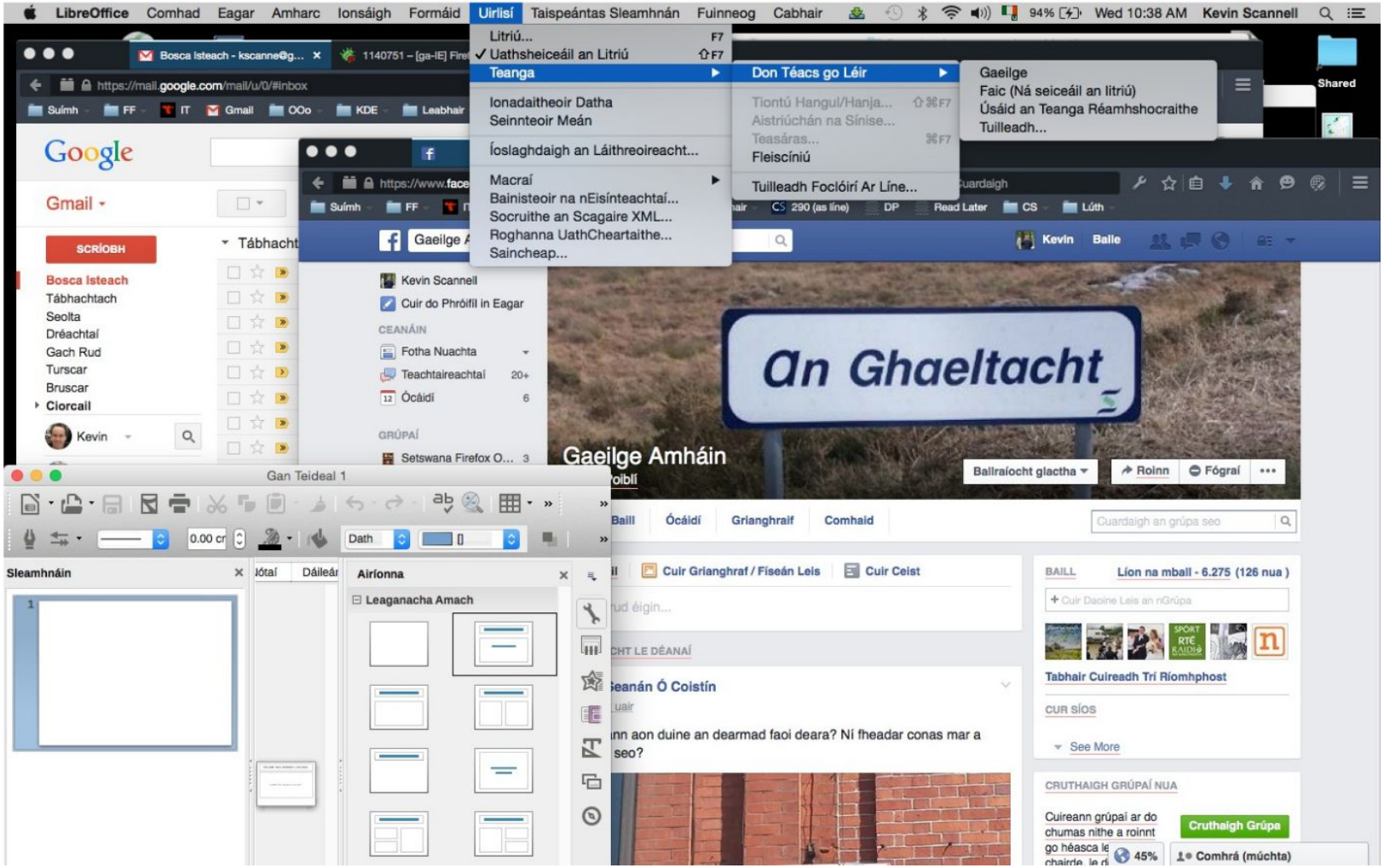
Collecting “everything”

- The Crúbadán project (crubadan.org), c. 2000 - present
- Indigenous Tweets (indigenoustweets.com), 2011 - present
- RSS feeds (e.g. chuala.me)
- Public Facebook posts
- Feedback loops + crawling
- At least 250 million words of Irish online, before cleaning


Basic Human Needs
Self-actualization
Creativity. Problem solving. Authenticity. Spontaneity
Esteem
Self-esteem. Confidence. Achievement
Social Needs
Friends. Family
Safety and Security
Physiological Needs (survival)
Air, Shelter, Water. Food
WiFi
Battery

Digression: Linguistic landscape

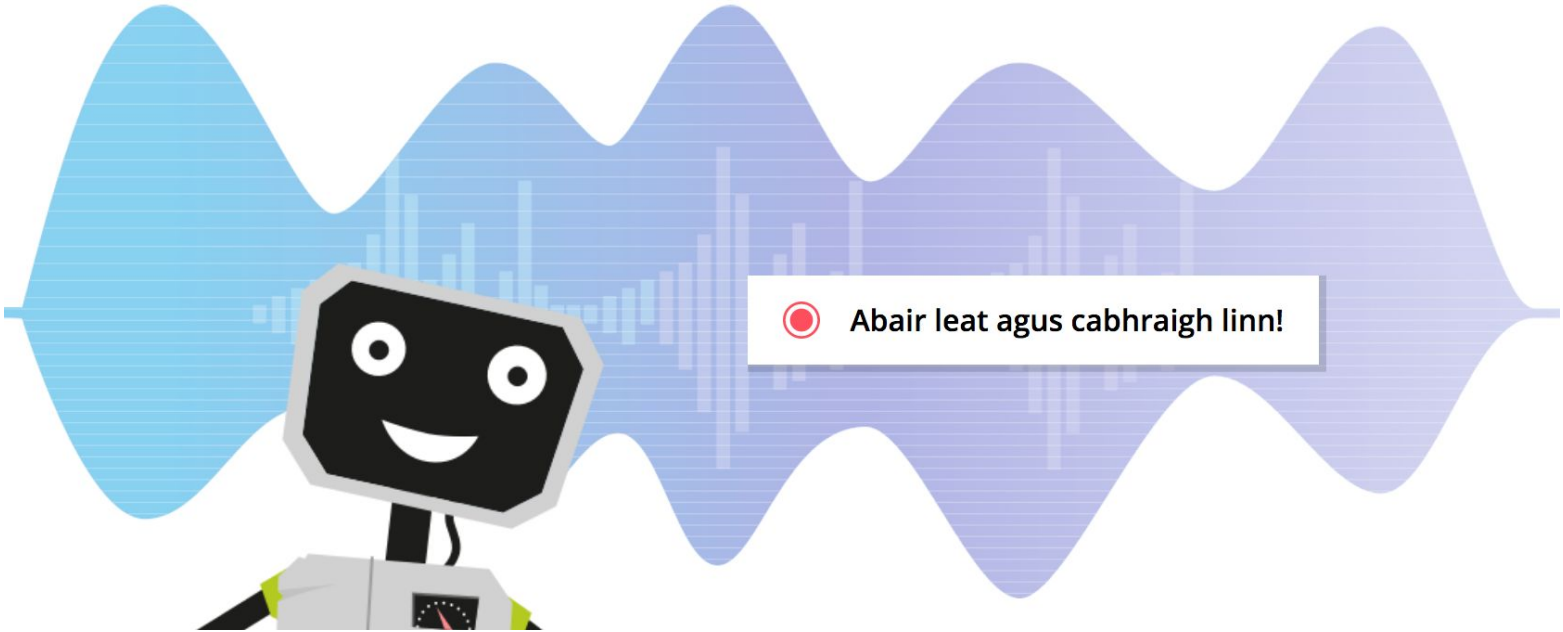
- Fully-localized user experience on computer and mobile devices
- Normalization of the language on the computer
- Long but spotty history for Irish!
- Early success via open source software
- Shifting landscape (Thunderbird to GMail, OpenOffice to Google Docs, etc.)
- “Transploitation”
- Uneven and sometimes uneasy collaboration with big companies
- Not sustainable or scalable
- An Ríomhacadamh (riomhacadamh.wordpress.com)


LibreOffice Comhad Eagar Amharc Ionsáigh Formáid Uirlisí Taispeántas Sleamhnán Fuinneog Cabhair
Litriú... F7
Uathsheiceáil an Litriú ⇧F7
Teanga
Ionadaitheoir Datha
Seinnteoir Meán
Íoslaghdaigh an Láithreoireacht...
Macraí
Bainisteoir na nEisínteachtaí...
Socruithe an Scagaire XML...
Roghanna UathCheartaithe...
Saincheap...
Don Téacs go Léir
Tiontú Hangul/Hanja... ⇧⌘F7
Aistriúchán na Sínise...
Teasáras... ⌘F7
Fleiscíniú
Tuilleadh Foclóirí Ar Líne...
Gaeilge
Faic (Ná seiceáil an litriú)
Úsáid an Teanga Réamhshocraithe
Tuilleadh...
Google
Gmail
SCRÍOBH
Bosca Isteach
Tábhachtach
Seolta
Dréachtaí
Gach Rud
Turscar
Bruscar
Ciorcail
An Ghaeltacht
Gaeilge Amháin
Gan Teideal 1
Sleamhnáin
Airíonna
Leaganacha Amach
Ballraíocht glactha
Roinn
Fógraí
Baill
Ócáidí
Grianghraif
Comhaid
Cuir Grianghraf / Físeán Leis
Cuir Ceist
Tabhair Cuireadh Trí Ríomhphost
CUR SÍOS
See More
CRUTHAIGH GRÚPAÍ NUA
Cruthaigh Grúpa

Common Voice
moz://a

0
141

Seo é Common Voice – córas de chuid Mozilla a mhúineann do ríomhairí an chaoi a labhraíonn daoine.



Language modeling

- Chomsky: “... the notion of ‘probability of a sentence’ is an entirely useless one, under any known interpretation of this term.”
- Let S = “this is an entirely useless notion”
- P(S) = P(this|^) P(is|this) P(entirely|this is) ... P(notion|this is an entirely useless)
- Usually formulated and computed this way (word prob. given history)
- Humans are pretty good at estimating these, at least
- P(Friday|My party is this coming) > P(Tuesday|My party is this coming)
- P(is|The man with the glasses) > P(are|The man with the glasses)

Applications

- Virtually all of the language technologies mentioned above!
- Can be used generatively
- MT, ASR, etc. fundamentally generate text, conditioned on input
- Traditional probabilistic models (“noisy channel”): $P(T \mid S) = P(S \mid T)\, P(T) / P(S)$
- Baked into many neural models for MT, etc.
- Better language models give better end-to-end performance, generally

Neural language models

- A flood of recent papers on neural language modeling, big leaps forward
- Almost 100% (and implicitly!) focused on English
- Word “English” doesn’t even appear in Google Brain’s landmark 2016 paper
- Same models applied to Irish show only small gains at best

Incorporating linguistic knowledge

- Celtic languages have “initial mutations”
- Almost always predictable from previous two words
- *bád seoil* “sailboat”, *mo b**h**ád seoil* “my sailboat”, *ár **m**bád seoil* “our sailboat”
- N-gram models don’t “see” that these are all the same word
- If most training examples are first type, say, harder to predict collocation
- (Google even gets this wrong in previous image: *tríd an **m**bóthar*)
- Easy enough to use a *factored language model* to get better results for Irish
- Examples like this abound…

Some success stories

- Intergaelic MT engines
  - Google Translate-like interface
  - Annotated novels and short stories
  - Social media translations, mouse-over help
- Irish standardization
  - Foclóir Stairiúil na Gaeilge
  - Corpas na Gaeilge 1600-1926
  - New English-Irish dictionary project

Tha crois margaid sgoinneil ann bhon 15mh linn. Thog John Cockburn Ormiston mar bhaile modail ann an 1735 mus do lorgadh gual faisg air anns an 19mh linn. Dh'fhàs an cunntas-sluaigh còmhla ris a' ghnìomhachas mèinnearachd. Dùineadh na mèinnean anns na 1960an. Chaidh Ormiston Hall a thogail airson John Cockburn ann an 1745. Cheannaich an Iarla Hopetoun e agus

Aistrigh »

Tha crois margaid sgoinneil ann bhon 15mh linn. Thog John Cockburn Ormiston mar bhaile

Tá cros margadh iontach ann ón 15ú haois. Thóg John Cockburn Ormiston mar bhaile

modail ann an 1735 mus do lorgadh gual faisg air anns an 19mh linn. Dh'fhàs an

samhail i 1735 sular aimsíodh gual i ngar dó sa 19ú haois. D'fhás an

cunntas-sluaigh còmhla ris a' ghnìomhachas mèinnearachd. Dùineadh na mèinnean anns na

daonáireamh in éineacht leis an ghnó mianadóireachta. Dúnadh na mianaigh sna

1960an. Chaidh Ormiston Hall a thogail airson John Cockburn ann an 1745. Cheannaich an

1960í. Chuaigh Ormiston Hall a thógáil ar son John Cockburn i 1745. Cheannaigh an

Iarla Hopetoun e agus Ormiston fhèin agus leasaicheadh an talla ann an 1772. Sgrios teine e

Iarla Hopetoun é agus Ormiston féin agus feabhsaíodh an halla in 1772. Scrios tine é

anns an Dàrna Cogadh.

sa Dara Cogadh.

SKEEALAGHT

YN VIBLE VOOAR

Tra va mee gobbragh ayns Purt ny hInshey ec oik ny barroosyn, va shenn dooinney cheet dys Purt ny hInshey ny keayrtyn voish Skeristal t'er y raad dys Rhumsaa mysh tree ny kiare meeilaghyn voish Purt ny hInshey. V'eh enmyssit Tommy Kneale as v'eh baghey ayns Skeristal, as v'eh ny ghooinney beg as va symm echey er y Ghaelg as er ny shenn skeealyn. As dooyrt eh dy row bible sy Ghaelg echey ayns y thie as v'ee foddey ny smoo na ny bibleyn cadjin — v'ee ny smoo na daa hrie er e lhurid as foddey ny smoo na ny bibleyn elley, as hug eh cuirrey dooys cheet dys y thie echey dy yeeaghyn urree traa erbee, as va shen goll er daa vlein agh cha row rieau traa aym dy gholl.

Aghterbee, cha jagh mee rieau lesh shilley er y vible shoh; as eisht cheayll mee dy row eh er ngeddyn baase as va mee boirit mychione shen, as cheayll mee myrgeddin dy row ooilley e stoo-thie tilgit magh ass y thie as aile currit orroo as va ooilley caillt. Aghterbee, va mee shickyr dy row yn vible shoh, yn vible neu-chadjin, dy row ee loshtit as stroit as currit mow. Agh cha row mee my vooinjer da as cha row cair erbee aym dy vannoo briaght mychione y vible, agh tammylt ny lurg

AN BÍOBLA MÓR

Nuair a bhí mé ag obair i bPort na hInse ag oifig na mbusanna, bhí seanduine ag teacht go Port na hInse uaireanta ó Skeristal atá ar an mbealach chuig Rhumsaa timpeall trí nó ceithre mhíle ó Phort na hInse. Bhí sé ainmnithe Tommy Kneale agus bhí sé ina chónaí i Skeristal, agus bhí sé ina dhuine beag agus bhí suim aige sa Mhanainnis agus sna seanscéalta. Agus dúirt sé go raibh bíobla sa Mhanainnis aige sa teach agus bhí sí i bhfad níos mó ná na bíoblaí coitianta — bhí sí níos mó ná dhá throigh ar a fhad agus i bhfad níos mó ná na bíoblaí eile, agus thug sé cuireadh domsa teacht go dtí an teach aige le féachaint uirthi am ar bith, agus bhí sin ag dul ar dhá bhliain ach ní raibh riamh am agam le dul.

Ar aon chuma, ní dheachaigh mé riamh le cuairt ar an bhíobla seo; agus ansin chuala mé go raibh sé tar éis bás a fháil agus bhí mé buartha faoi sin, agus chuala mé freisin go raibh uile a throscáin caite amach as an teach agus tine curtha orthu agus bhí uile caillte. Ar aon chuma, bhí mé cinnte go raibh an bíobla seo, an bíobla neamhghnách, go raibh sé dóite agus scriosta agus scriosta. Ach ní raibh mé i mo mhuintir dó agus ní raibh ceart ar bith agam le déanamh fiosraithe faoin bhíobla.

Standardization / indexing: corpas.ria.ie

Gaeilge Gaedhilge Gaeilige Gaelige Gailge Gaoidhilge Gaoidheilge Gaelge
Gaidhlige Gaedheilge Gaoidhelge Gailege Gaielge Gaodhailge Gaedilge
Gaeidhilge Gaedhilige Gaoidilge Gaeilgele Gaedhlige Gaédhilge Gaoilge
Gaeillge Gaeilga Gaidhilge Gaelilge Gaodheilge Gaeiilge Gaedhilghe
Gadhilge Gaheilge Gaellge Gaoilaige Gaodhilge Gaedhilgé Gaeilege Gaaeilge
Gailige Gaeilgé Gaeghilge Gaedhailge Gaoidhlige Gaelgie Gaeiloge Gaeilgle
Gaeilghe Gaeilge Gaeidhlge Gaeidheilge Gaeeilge Gaoilige Gaóilge
Gaoilaga Gaoigheilge Gaoidhlge Gaoidelge Gaoideilge Gaodhéilge Gaieilge
Gaeulge Gaeuilge Gaeolge Gaeoidheilge Gaeilgi Gaeilgee Gaeílge Gaeidlge
Gaeidilge Gaeidhelge Gaehilge Gaeeilgee Gaedhlge Gaedhiilge Gaedhelga
Gaédhailge Gaedgilge Gadehilge Gaddhilge Gaoghailge Gaileige Gaidlhlige
Gaidhlge Gaeliage Gaelga Gaéilge Gaedilghe Gaedhulge Gaedhealg Gaedheilg
Gaédheilg Gaedhilg Gaedhilig Gaeilg Gaoidhealg Geadhilge Geailge