Ollscoil na Gaillimhe
University of Galway

Claonadh Inscne, Líonraí Néaracha, agus an Ghaeilge

Title	Claonadh Inscne, Líonraí Néaracha, agus an Ghaeilge
Author(s)	Scannell, Kevin P.
Publication Date	2020
Publisher	Acadamh na hOllscolaíochta Gaeilge, Ollscoil na hÉireann, Gaillimh
Repository DOI	10.13025/6f23-sq24

LÉANN TEANGA
AN REIVIÚ

Claonadh Inscne, Líonraí Néaracha, agus an Ghaeilge

Kevin P. Scannell

Réamhrá

Baineann an t-alt seo leis an **meaisínfhoghlaim**, réimse ríomheolaíochta atá go mór i mbéal an phobail sa lá atá inniu ann. Is éard atá i gceist leis an meaisínfhoghlaim ná raon leathan teicnící a ligeann don ríomhaire scileanna nua a fhoghlaim as a stuaim féin. Mar shampla, chun córas ríomhaistriúcháin ó Bhéarla go Gaeilge a fhorbairt, is leor bailiúchán abairtí dátheangacha (na **sonraí traenála**) a sholáthar don ríomhaire sa chaoi gur féidir leis ginearálú a dhéanamh ó na sonraí sin agus abairtí nua a aistriú. An rud is tábhachtaí le tuiscint faoin chur chuige seo ná nach mbaineann sé úsáid as aon saineolas teangeolaíochta ar bith, seachas an saineolas intuigthe atá leabaithe sna sonraí traenála.

Sa chuid eile den alt seo, beimid ag plé le teicníc amháin sa mheaisínfhoghlaim darb ainm **líonraí néaracha**. Is creatlach mhatamaiticiúil í líonra néarach a mbaineann an ríomhaire úsáid aisti chun rudaí nua a fhoghlaim ó shonraí traenála. Tá siad in úsáid go forleathan i réimsí éagsúla, ón ríomhaistriúchán go gluaisteáin uathrialaitheacha go galaraithint. Tá na mionsonraí teicniúla thar a bheith casta, ach ar an dea-uair níl na mionsonraí sin de dhíth chun an chuid eile den alt seo a thuiscint.

Le roinnt blianta anuas, tá mé ag baint úsáid as líonraí néaracha chun córais nua a fhorbairt atá in ann an Ghaeilge a phróiseáil ar bhealaí éagsúla: aistriúchán a dhéanamh, gramadach a sheiceáil, anailís a dhéanamh ar sheintimint, ceisteanna a fhreagairt, agus araile. I ngach cás, bainim úsáid as corpas téacs aonteangach nó dátheangach mar shonraí traenála. Tuigtear le fada go bhfuil claontaí de chineálacha éagsúla leabaithe in aon chorpas traenála den sórt seo: claonadh bunaithe ar inscne, ar chine, nó ar ghnéaschlaonadh mar shampla. Chomh maith

leis sin, tá go leor taighde ann a thaispeánann an tionchar a imríonn na claontaí seo ar na líonraí néaracha a fhaightear amach as an bpróiseas traenála, agus ar na córais a bhaineann úsáid astu: córais a dhéanann ríomhaistriúchán (Prates, Avelar & Lamb, 2018; Kuczmarski, 2018), anailís ar sheintimint (Kiritchenko & Mohammad, 2018), réiteach comhthagartha (Lu *et al.*, 2018; Rudinger, 2018), aithint cainte (Tatman, 2017), agus tuairisciú ar theanga mhaslach nó ghangaideach (Dixon *et al.*, 2018; Park, Shin & Fung, 2018).

Tá sé an-éasca samplaí de chlaonadh inscne a aimsiú sa réimse ríomhaistriúcháin, mar shampla nuair a dhéantar aistriúchán ó theanga nach bhfuil inscne marcáilte inti go teanga eile ina bhfuil inscne marcáilte. Sa chás sin, is minic go roghnaíonn córais ar nós Google Translate focail fhirinscneacha sa sprioctheanga, go háirithe in abairtí a dhéanann tagairt do ghairmeacha ar nós 'ollamh', 'innealtóir', nó 'ríomhchláraitheoir'. Tá go leor samplaí den sórt seo ar fáil fiú ó Bhéarla go Gaeilge: faoi láthair, aistríonn Google Translate an abairt 'My PhD advisor is a full professor' mar 'Is ollamh iomlán ***é*** mo chomhairleoir PhD', cé nach bhfuil aon leid maidir le hinscne an chomhairleora sa bhuntéacs. Anois, tá sé fíor go bhfuil níos mó fear ná mná ina n-ollúna nó ina léachtóirí ag an tríú leibhéal sa lá atá inniu ann (An Phríomh-Oifig Staidrimh, 2017, §4.7), agus dá bharr sin is dócha go mbeadh níos mó tagairtí d'ollúna fireanna ná baineanna in aon chorpas traenála. Foghlaimíonn an ríomhaire an claonadh staitisticiúil sin ón gcorpas, agus feicimid an toradh san aschur ón gcóras aistriúcháin.

Cúpla rud tábhachtach le tabhairt faoi deara faoin sampla seo. Ar dtús, feictear an éifeacht seo fiú in éagmais ráiteas neamhbhalbh claonta in aghaidh grúpa ar leith (ar nós 'tá fir níos fearr ná mná ag X nó Y'). Ina áit sin, is éard atá i gceist ná treochtaí leathana staitisticiúla sna sonraí traenála a eascraíonn as claontaí struchtúrtha sa tsochaí i gcoitinne (mar shampla, níos mó fear ná mná ina n-ollúna). Agus dá bharr sin, ní leor ráitis den sórt seo a scagadh amach, ná traenáil a dhéanamh ar shonraí 'neodracha' amháin (scéalta nuachta, srl.) chun an fhadhb a mhaolú (Bolukbasi, *et al.*, 2016). An dara pointe ná gur féidir le córas meaisínfhoghlama claontaí sna sonraí traenála a *mhéadú,* mar shampla trí fhoirmeacha firinscneacha a roghnú níos minice ná mar a roghnófaí bunaithe ar shonraí déimeagrafacha amháin (Prates, Avelar & Lamb, 2018), nó fiú nuair a bhíonn fianaise láidir ann gur bean í an t-ábhar (mar shampla, má chuireann tú ainm baininscneach leis an abairt thuas, ar nós 'Madeleine, my Ph.D. advisor, is a full professor', gheobhaidh tú 'é' san aistriúchán mar sin féin).

San alt seo, díreoidh mé ar chlaonadh inscne i gcorpas mór Gaeilge agus an tionchar a imríonn sé ar theicneolaíochtaí teanga. I dtús báire, déanaim anailís corpasbhunaithe ar phoist agus ar ghairmeacha i bhfrásaí ina bhfuil inscne an duine marcáilte, mar shampla athruithe tosaigh firinscneacha vs. baininscneacha i bhfrásaí ar nós *ina b(h)unmhúinteoir* nó *ina (h)ollamh*. Ansin, taispeánaim go bhfuil claontaí inscne le feiceáil sa **leabú focal** a úsáidtear go minic i bpróiseas réamhthraenála ar líonraí néaracha níos casta. Ansin, léirím an drochthionchar a imríonn na

claontaí inscne seo ar líonra néarach ar leith, ceann a thuarann athruithe tosaigh sa Ghaeilge. Sa rannán deireanach, déanaim iarracht na héifeachtaí seo a mhaolú trí athruithe sa phróiseas traenála.

Toisc go bhfuil an taighde san alt seo bunaithe ar struchtúir ghramadaí Gaeilge atá go láidir dénártha ó thaobh inscne, ní raibh mé in ann anailís chorpasbhunaithe a dhéanamh ar chlaonadh in aghaidh daoine trasinscneacha nó neamh-dhénártha mar shampla. Bheadh sé spéisiúil filleadh ar an gceist seo sa todhchaí.

Tá an t-alt seo tiomnaithe do m'iníon Maddy.

Claonadh Inscne i gCorpas Gaeilge

Thosaigh mé ag bailiú téacsanna Gaeilge ón Idirlíon sna 1990í, agus tá níos mó ná 140 milliún focal sa chorpas Idirlín seo faoi láthair, ábhar ó na meáin shóisialta, blaganna, liostaí r-phoist, nuachtáin, agus cáipéisí oifigiúla mar shampla. Tá na torthaí go léir sa rannán seo bunaithe ar an gcorpas seo.

Níl sé ró-éasca uimhir chruinn a chur ar an iomlán, ach tá sé soiléir go bhfuil an chuid is mó den Ghaeilge ar líne scríofa ag fir nó ag daoine a úsáideann ainm firinscneach. Mar shampla, ar an suíomh micreabhlagadóireachta Twitter, i measc na 803 úsáideoir is gníomhaí i nGaeilge a úsáideann a bhfíorainm, tá ainm baininscneach ag 285 acu (35.5%). Nó, i measc na 243 duine a scríobh ailt nuachta ar Tuairisc.ie idir Deireadh Fómhair 2014 agus Nollaig 2019, tá ainmneacha baininscneacha ag 68 acu (28%), agus níl na scríbhneoirí sin freagrach as ach 19.5% den ábhar ar an suíomh, de réir líon focal. Tá an scéal mar an gcéanna maidir le hábhar na scríbhneoireachta ar líne: as 1769 beathaisnéisín ar an suíomh ainm.ie, ba mhná iad 176 (10%) i Samhain 2019 (Fiontar, 2019). Tá cúrsaí beagáinín níos fearr ar Vicipéid na Gaeilge: is faoi mhná 1804 as na 10044 beathaisnéisín ar an suíomh sin (18%) in Eanáir 2020.

Tá tuilleadh fianaise maidir le claonadh inscne i gcorpas na Gaeilge ar fáil toisc go bhfuil inscne marcáilte ar mhórán struchtúr gramadaí sa teanga. Díreoimid ar dhá struchtúr den sórt seo anois. An chéad cheann ná frásaí san fhoirm *ina* ***X***, nuair is post nó gairm é ***X***, ar nós *múinteoir*, *saighdiúir*, nó *ollamh*, mar shampla san abairt *Tá Seán ina mhúinteoir*. An dara ceann ná frásaí san fhoirm ***COP X*** *é/í*; anseo is post nó gairm é ***X*** arís, agus is leagan éigin den chopail é ***COP*** (*is, gur, ba,* srl.), mar shampla *Is bainisteoir í Máire*. An bunsmaoineamh atá ann ná athruithe tosaigh a chomhaireamh sa chorpas sa chéad chás, agus na forainmneacha *é* vs. *í* a chomhaireamh sa dara cás; is é sin le rá go nglacfaimid gur fear atá i gceist le gach *ina mhúinteoir*, *ina ollamh*, nó *is dochtúir é*, agus bean atá ann le gach *ina múinteoir*, *ina hollamh*, nó *is dochtúir í*. Bhí mé leathanaigeanta ó thaobh cad is 'gairm' ann san anailís seo; tá formhór mór na samplaí sa gcorpas cosúil leis na cinn thuasluaite, ach ghlac mé freisin le focail ar nós *pótaire*, *príosúnach*, *teifeach*, agus *tírghráthóir* atá sách coitianta sna struchtúir seo, cé nach 'gairmeacha' iad i ndáiríre. Níor chuir mé creideamh ná náisiúntacht san áireamh (*is*

Éireannach é, bhí sé ina Phrotastúnach), ná aon fhocal a bhfuil inscne ceangailte leis *a priori* (i mo thuairimse) ach oiread: *athair*, *máthair*, *buachaill*, *cailín*, *rí*, *banríon*, srl. Caithfear a bheith cúramach mar gheall ar roinnt deacrachtaí eile freisin:

1. **Abairtí diúltacha.** Is annamh a fheictear frásaí ar nós *ina heaspag* nó *ina Taoiseach* sa gcorpas, ach tá siad ann, go hiondúil in abairtí diúltacha, mar shampla *Níl bean ar bith ina heaspag go fóill* nó *ní raibh bean ina Taoiseach riamh*.
2. **Botúin ghramadaí.** Tá botúin ghramadaí ann freisin: *tá Aoife ina chéimí ollscoile* nó *bhí Pádraig ina Teachta Dála*.
3. **Ábhar neamhdhaonna.** Anois is arís is logainm nó comhlacht atá i gceist sna struchtúir seo, mar shampla *Tá Guinness ina ambasadóir gnó* nó *Tá Éire ina sínitheoir ar an gCoinbhinsiún* agus ní chuirimid samplaí den sórt seo san áireamh san anailís thíos.
4. **Consain neamhathraithe.** Ar ndóigh ní féidir inscne a oibriú amach in abairtí *ina* ***X*** nuair a thosaíonn an t-ainmfhocal ***X*** le litir *l*, *n*, nó *r*: *ina luibheolaí*, *ina naoscaire*, nó *ina ríomhchláraitheoir*.
5. **Úsáid neamh-aisfhillteach.** Uaireanta, ní hionann an sealbhóir agus an t-ainmfhocal ***X*** i bhfrása *ina* ***X***; mar shampla *chuir sí a muinín ina dochtúir* nó *bithiúnach atá ina dochtúir*.

Tar éis dúinn déileáil leis na deacrachtaí seo, tá 446 gairm sa chorpas atá luaite ar a laghad seacht n-uaire i gceann den dá struchtúr thuas. An staitistic is suntasaí ná nach bhfuil ach 12 (sic) as an 446 gairm luaite níos minice san fhoirm bhaininscneach ná san fhoirm fhirinscneach: *altra*, *aoi-léachtóir*, *comhláithreoir*, *comhordaitheoir*, *comhstiúrthóir*, *cuiditheoir*, *damhsóir*, *fidléir*, *gnáthurlabhra*, *mainicín*, *striapach*, agus *tréidlia*; tá na fir chun tosaigh i ngach cás eile. Sa chorpas ina iomláine, úsáidtear na struchtúir seo 33750 uair, le 5255 acu (15.6%) san fhoirm bhaininscneach, agus tá an céatadán de mhná fiú níos lú i réimsí áirithe; féach Tábla 1.

Réimse	**% bain.**	**Samplaí**
Cúrsaí míleata	3.3%	saighdiúir, óglach, ginearál
Creideamh	4.8%	sagart, ministir, misinéir
Spórt	5.6%	peileadóir, traenálaí, tosaí
Ollscolaíocht	8.2%	ollamh, acadamhaí, léachtóir
Polaitíocht	10.9%	príomh-aire, taoiseach, seanadóir
Dlí	14.2%	dlíodóir, breitheamh, garda
Litríocht	19.4%	file, drámadóir, úrscéalaí
Eolaíocht/teicneolaíocht	20.2%	eolaí, fisiceoir, innealtóir
Físealaín	22.3%	péintéir, dealbhóir, ealaíontóir
Ceol	24.3%	ceoltóir, amhránaí, píobaire

Tábla 1: Céatadán baininscneach i réimsí áirithe

Taispeánann na figiúirí seo go bhfuil claonadh inscne suntasach sa chorpas seo, ach tá an anailís bunaithe ar chur chuige sách simplí, comhaireamh a dhéanamh ar struchtúir ghramadaí. Na ceisteanna atá romhainn anois ná an imríonn an claonadh seo aon éifeacht ar líonraí néaracha a thraenáiltear ar an gcorpas, nó ar aon fheidhmchlár a bhunaítear ar líonraí néaracha? Déanaim iarracht freagra a thabhairt ar na ceisteanna seo sa chuid eile den alt.

Tionchar ar Leabú Focal

Is éard atá i **leabú focal** ná mapáil ó na focail i dteanga go dtí veicteoirí[1] i spás ardtoiseach, de ghnáth, idir 150 agus 300 toise. Baineann cúpla buntáiste le húsáid veicteoirí inár gcomhthéacs: ar dtús, is féidir an *fad* idir dhá veicteoir a thomhas, agus mar sin, an fad idir dhá *fhocal* a thomhas freisin trí úsáid a bhaint as an leabú focal. Chomh maith leis sin, is féidir matamaitic shimplí a dhéanamh ar veicteoirí (suimiú, dealú, srl.), agus oibríonn an mhatamaitic ar fhocail freisin arís mar gheall ar an leabú focal. An mhíorúilt atá ann ná go bhfuil baint idir oibríochtaí matamaitice ar na veicteoirí seo agus cúrsaí séimeantacha. Thug (Mikolov, Yih, & Zweig, 2013) an sampla is clúití den fheiniméan seo:

rí - fear + bean ≈ banríon

Sa 'chothromóid' thuas agus sa chuid eile den alt, glacaim leis go bhfuil leabú focal roghnaithe, agus gurb ionann focal (i gcló trom) agus a veicteoir faoin leabú. Ní fíorchothromóid é seo; ní chiallaíonn an tsiombail ≈ gurb *ionann* na veicteoirí ar an dá thaobh, ach ina áit sin go bhfuil siad *i ngar dá chéile*, nó níos fearr gurb é **banríon** an veicteoir is gaire don veicteoir ar thaobh na láimhe clé.

Tá cúpla teicníc ann le leabú focal éifeachtach a chruthú. An sprioc atá ann i ngach cás ná an leabú a roghnú sa chaoi go mbeidh na veicteoirí a bhaineann le dhá fhocal i ngar dá chéile nuair a fheictear an dá fhocal sin sna comhthéacsanna céanna i gcorpas traenála. Is é sin le rá gur iarracht é gaolta séimeantacha a shamhaltú ar bhealach matamaiticiúil. Is minic go n-úsáidtear leabú focal mar ionchur i líonraí néaracha níos casta (cinn a dhéanann aistriúchán uathoibríoch, mar shampla) sa chaoi nach gá dóibh na gaolta séimeantacha céanna a fhoghlaim arís as an nua. Is sampla amháin é seo de theicníc an-tábhachtach sa mheaisínfhoghlaim darb ainm **réamhthraenáil**, agus seo é ceann de na bealaí a shleamhnaíonn claontaí den chineál a fheicfimid sa rannán seo isteach i dtáirgí atá dírithe ar úsáideoirí, ar nós ríomhaistriúcháin.

Bhain mé úsáid as an mbogearra word2vec[2] chun leabú focal Gaeilge a chruthú. Bhí an leabú seo traenáilte ar chorpas a bhfuil 129 milliún focal ann, leis na hathruithe tosaigh scriosta go hiomlán (sa chaoi nach mbeadh ach veicteoir amháin i gceist do *dochtúir*, *dhochtúir*, agus *ndochtúir*,

1 Is ainm galánta ar liosta uimhreacha é an focal *veicteoir*, sin an méid. Is é *toise* an veicteora fad an liosta. Is veicteoir tríthoiseach é (3,-2,11) mar shampla.

2 https://code.google.com/archive/p/word2vec/

mar shampla).[3] Ansin, rinne mé iarracht comparáid a dhéanamh idir focail a bhfuil inscne ceangailte leo *a priori* (*fear, bean, máthair, athair,* srl.) agus na gairmeacha luaite sa rannán roimhe seo. Chun na críche seo, bhain mé sochar as an struchtúr matamaiticiúil sa spás veicteoireach, ag leanúint Bolukbasi *et al.* (2016). I dtús báire, roghnaigh mé N=20 péire focal firinscneach/baininscneach: (f_1,b_1) = (**fear**, **bean**), (f_2,b_2) = (**athair**,**máthair**), (f_3,b_3) = (**sé**,**sí**), (f_4,b_4) = (**buachaill**,**cailín**), (f_5,b_5) = (**dearthár**,**deirfiúr**), srl. Sainmhíníonn gach péire 'treo' nó 'ais' inscne tríd an difríocht b_i - f_i agus tugaim $\boldsymbol{\mathcal{J}}$ ar mheán na ndifríochtaí seo (Zhao, *et al.*, 2018):

$$\boldsymbol{\mathcal{J}} = \frac{1}{N}\sum_{i=1}^{N}(b_i - f_i).$$

Ansin, is féidir veicteoir ar bith **w** a theilgean ar an veicteoir $\boldsymbol{\mathcal{J}}$:

$$\beta(\mathbf{w}) = (\mathbf{w}\cdot\boldsymbol{\mathcal{J}})/||\boldsymbol{\mathcal{J}}||.$$

Is uimhir í $\beta(\mathbf{w})$. Nuair a thagann **w** ó leabú focal, tá luach deimhneach níos airde ar $\beta(\mathbf{w})$ má tá an gaol idir **w** agus na focail bhaininscneacha $b_1,\ldots,b_N$ níos láidre ná leis na focail fhirinscneacha $f_1,\ldots,f_N$, agus luach níos diúltaí ar mhalairt de chás. Seo iad na deich bhfocal a bhfuil an luach $\beta(\mathbf{w})$ is airde acu agus na deich gcinn a bhfuil an luach is ísle (is diúltaí) acu is measc an 20000 focal is coitianta sa chorpas (ag fágáil $b_1,\ldots,b_N$, $f_1,\ldots,f_N$ agus ainmneacha dílse as an áireamh): *torrach, cíoch, cíche, brídeog, fúithi, mainicín, toircheas, bábóg, ceirbheacs, pógadh,* agus ar an taobh eile, *óglach, ginearál, gaiscíoch, canónach, cath, díorma, ridire, ardsagart, staraí, maor*. Ag díriú ar na gairmeacha amháin, faightear amach go gcloíonn na luachanna $\beta(\mathbf{w})$ go mór le steiréitíopaí traidisiúnta; is é sin le rá gur imir na steiréitíopaí céanna tionchar láidir ar an leabú focal. Féach Tábla 2 agus Léaráid 1.

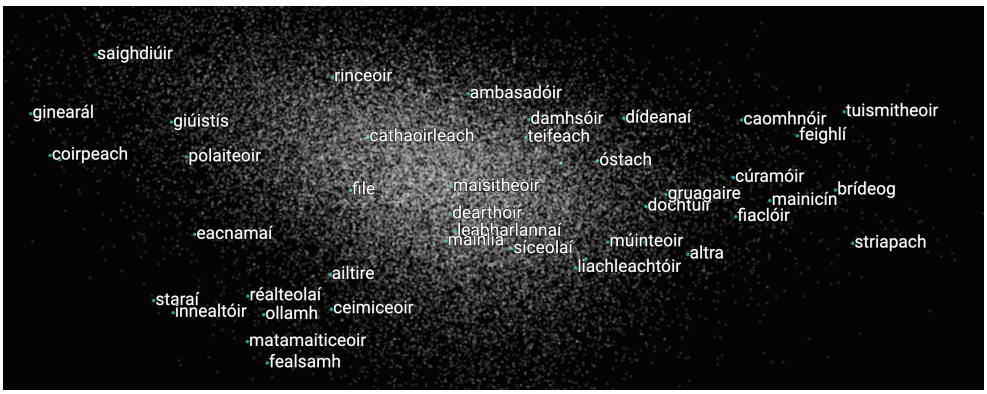
Focal w	**β(w)**
mainicín	0.3486
striapach	0.2972
feighlí	0.2625
gruagaire	0.2110
altra	0.2107
tuismitheoir	0.2103
teiripeoir	0.1747
glantóir	0.1700
cúramóir	0.1687
caomhnóir	0.1611

Tábla 2a: Na luachanna β(w) is airde

Focal w	**β(w)**
óglach	-0.4098
ginearál	-0.3624
gaiscíoch	-0.3144
canónach	-0.3121
ridire	-0.3013
ardsagart	-0.3000
staraí	-0.2975
maor	-0.2933
easpag	-0.2879
leifteanant	-0.2834

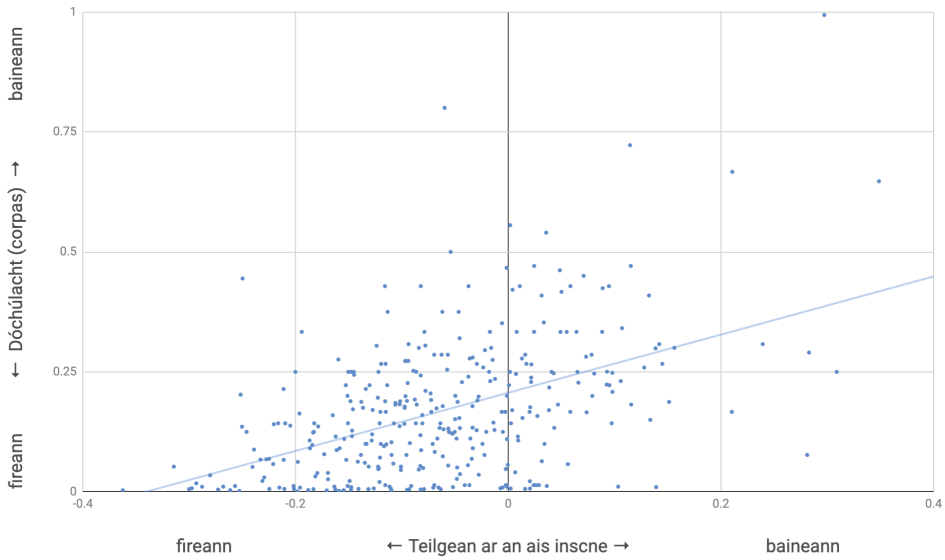
Tábla 2b: Na luachanna β(w) is ísle

3 Do na saineolaithe ar mhaith leo na torthaí seo a atáirgeadh, d'úsáid mé an tsamhail CBOW leis na socruithe seo a leanas: -size 200 -window 8 -negative 25 -hs 0 -sample 1e-4 -threads 20 -iter 15



Léaráid 1: An leabú focal arna theilgean ar spás déthoiseach; sampla gairmeacha ar taispeáint. Is é an *x*-ais an teilgean ar an veicteoir 'bean – fear': focail níos gaire do 'bean' ar dheis agus focail níos gaire do 'fear' ar chlé. Tá na focail 'neodracha' go díreach i lár na léaráide.

Bhí an anailís chorpasbhunaithe sa rannán roimhe seo thar a bheith simplí: ní raibh ach cúpla struchtúr gramadaí i gceist ansin, i gcomparáid leis an leabú focal a chuireann gach comhthéacs ó chorpas traenála san áireamh. Ach mar sin féin, bhíomar ag iarraidh claonadh den sórt céanna a lorg sa dá chás, agus dá bharr sin b'fhiú comparáid idir an dá chur chuige a dhéanamh. Bhí 346 gairm i measc an 20000 focal is coitianta, agus rinne mé scaipghram le $\beta(\mathbf{w})$ ar an *x*-ais agus an 'dóchúlacht bhaineann' ar an *y*-ais (an líon struchtúr baininscneach ina bhfuil **w**, roinnte ar an líon iomlán ina bhfuil **w**). Ní bheifí ag súil le comhghaol foirfe, ach is dea-chomhartha é go bhfuil comhghaol deimhneach ann (R=0.479); féach Léaráid 2.



Léaráid 2: Scaipghram le $\beta(\mathbf{w})$ ar an *x*-ais, dóchúlacht bhaineann ar an *y*-ais.

Tionchar ar Líonra Néarach ar leith

Ní dhéanann leabú focal dochar d'éinne é féin; is rud teibí é. Tarlaíonn an fhadhb nuair a úsáidtear leabú focal nó líonra néarach eile a bhfuil claonadh de shaghas éigin ann mar chuid de chóras atá dírithe ar úsáideoirí. Níl sé ródheacair drochthorthaí a shamhlú, mar shampla córas meaisínfhoghlama a dhéanann réamhscagadh ar iarratasóirí ar phost, nó a thaispeánann fógraí ar fholúntas poist do dhaoine ar líne; an dtabharfadh córas mar sin cothrom na Féinne do mhná i gcás poist atá go láidir ar an taobh 'firinscneach' san anailís thuas? D'fhéadfadh córas mar sin dochar a dhéanamh do dhaoine san fhíorshaol, agus caithfimid é sin a sheachaint (Lambrecht & Tucker, 2019).

Déanfaidh mé léiriú ar an bhfadhb le líonra néarach níos simplí, ceann a bhí traenáilte chun athruithe tosaigh sa Ghaeilge agus i nGaeilge na hAlban a thuar (Scannell, 2020). Tá an líonra seo thar a bheith cruinn, an freagra ceart (séimhiú, urú, srl.) á thabhairt níos mó ná 98% den am. Cé gurb é seo ceann de na struchtúir ghramadaí is simplí sa Ghaeilge, feicfimid go n-imríonn claonadh inscne sa chorpas traenála drochthionchar ar na torthaí mar sin féin. Chun é seo a dhéanamh, chuir mé an líonra is cruinne ó Scannell (2020) i bhfeidhm ar an gcorpas forbartha ón bpáipéar céanna (1 milliún focal in 39628 abairt), agus bhailigh mé gach sampla den struchtúr *ina* ***X*** (le ***X*** ina ghairm mar a bhí sé thuas) ó na torthaí. Bhí 237 sampla den chineál seo ann, agus thuar an líonra an t-athrú ceart 218/237 uair (92.0%). Níl sé sin rómhaith i gcomparáid leis an gcruinneas iomlán, ach tá sé seo ar cheann de na cásanna is deacra don líonra (aon rud a bhaineann leis an aidiacht shealbhach *a, ina, faoina,* srl.), toisc gur minic go mbraitheann an t-athrú tosaigh ceart ar leideanna comhthéacsúla caolchúiseacha. Ach is é an rud is suntasaí ná an difríocht idir na torthaí ar shamplaí firinscneacha (F-Scór 0.9570) agus ar shamplaí baininscneacha (0.7733). Tá na torthaí go léir leagtha amach i dTábla 3.

		Lipéad tuartha					
		Fir.	**Bain.**	**Iolra**	**Cruinneas**	**Clúdach**	**F-Scór**[4]
Fíorlipéad	Fir.	189	7	2	0.9594	0.9545	0.9570
	Bain.	8	29	2	0.8056	0.7436	0.7733

Tábla 3: Tuar ar athruithe tosaigh firinscneacha vs. baininscneacha

4 Cruinneas = precision (P), clúdach = recall (R), F-scór = 2PR/(P+R); féach https://en.wikipedia.org/wiki/Precision_and_recall

Seo botún amháin a rinne an líonra, inar thuar sé an t-athrú tosaigh firinscneach (séimhiú anseo) ar an bhfocal *ceannródaí*:

> *Bhí cáil riamh ar Samantha Cameron, nó 'Sam Cam' mar a tugadh uirthi, as a dea-ghléasadh agus as a bheith ina ceannródaí faisin.*[5]

Bheadh an ceann seo thar a bheith éasca d'eagarthóir daonna toisc go bhfuil go leor leideanna ann gur bean atá i gceist. É sin ráite, tá an focal *ceannródaí* go láidir ar an taobh firinscneach san anailís thuas, agus seans gur cuireadh an líonra ar strae mar gheall ar na hainmneacha 'neodracha' *Cameron* agus *Sam* freisin.

Cúpla pointe teicniúil b'fhiú a lua freisin. Ar dtús, ní thugann an líonra seo aon aird ar chomhthéacs ar thaobh na láimhe deise den spriocfhocal (mar gheall ar shocruithe teicniúla sa pháipéar tagartha). Dá dtabharfadh, b'fhéidir go mbeadh na torthaí beagáinín ní b'fhearr, ach gan dabht bheadh bearna mhór fós ann idir firinscneach agus baininscneach ó thaobh cruinnis de. Chomh maith leis sin, ní raibh aon réamhthraenáil i gceist nuair a d'fhorbair mé an líonra néarach seo. Tá leabú focal ann, ach d'fhoghlaim an líonra an leabú mar chuid dá phróiseas traenála féin. Is é sin le rá nár tháinig aon chlaonadh inscne ó leabú focal réamhthraenáilte sa chás seo, ach maireann an fhadhb mar sin féin.

Claonadh Inscne a Mhaolú

Tá go leor taighde foilsithe le cúpla bliain anuas a bhaineann le claonadh inscne i leabuithe focal Béarla agus teicnící chun an claonadh sin a mhaolú (Bolukbasi, *et al.*, 2016; Ethayarajh, Duvenaud, & Hirst, 2019; Gonen & Goldberg, 2019; James-Sorenson & Alvarez-Melis, 2019; Prost, Thain, & Bolukbasi, 2019; Zhao, *et al.*, 2018). Sa straitéis is coitianta, ceann a tugadh isteach in (Bolukbasi, *et al.*, 2016), déantar iarracht an claonadh a laghdú trí na veicteoirí sa leabú a athrú. Baintear úsáid as an ais inscne $\mathcal{J}$ a shainmhínigh mé thuas (nó ceann eile cosúil leis), agus déantar teilgean ar an spás ortagánach le $\mathcal{J}$ sa chaoi go mbeidh $\beta(\mathbf{w})=0$ d'aon fhocal $\mathbf{w}$ ar chóir dó a bheith neodrach ó thaobh inscne. Faraor, tá fianaise ann nach bhfuil an cur chuige seo éifeachtach mar nach ndéanann sé ach an claonadh inscne a chur i bhfolach seachas é a ruaigeadh as an leabú mar ba cheart (Gonen & Goldberg, 2019).

Mar gheall ar an deacracht seo, rinne mé iarracht déileáil go díreach leis an gclaonadh sa líonra néarach sa rannán roimhe seo. Bhí an cur chuige thar a bheith simplí: chuir mé na gairmeacha $\boldsymbol{X}$ i bhfrásaí *ina* $\boldsymbol{X}$ i bhfolach sna sonraí traenála ionas go mbeadh sé riachtanach don líonra néarach fianaise eile san abairt a úsáid chun an t-athrú tosaigh ceart a thuar, in áit a bheith ag brath ar ghnéithe an fhocail $\boldsymbol{X}$. Chaomhnaigh mé an chéad litir den fhocal $\boldsymbol{X}$ (toisc gur comhartha tábhachtach é sin agus ní dócha go mbeadh aon chlaonadh leabaithe

5 https://tuairisc.ie/lipead-nua-faisin-seolta-ag-sam-cam/

i litir amháin), ach bhí an chuid eile den fhocal folaithe ón líonra. Tugtar **mascadh sonraí** ar an teicníc seo, agus imríonn sé ról tábhachtach i dteicneolaíochtaí nua mór le rá ar nós BERT (Devlin, *et al.*, 2019). Tar éis dom an líonra céanna ó Scannell (2020) a thraenáil leis na sonraí masctha seo, bhí athrú beag ar na torthaí; féach Tábla 4.

		Lipéad tuartha					
		Fir.	**Bain.**	**Iolra**	**Cruinneas**	**Clúdach**	**F-Scór**
Fíorlipéad	Fir.	187	11	0	0.9541	0.9444	0.9492
	Bain.	9	28	2	0.7179	0.7179	0.7179

Tábla 4: Tuar ar athruithe tosaigh firinscneacha vs. baininscneacha; sonraí masctha

I ndeireadh na dála, is léir gur toradh diúltach é seo. Cé gur tháinig méadú ar na dóchúlachtaí baininscneacha i bhformhór na gcásanna, chloígh an líonra leis an 'geall slán' (firinscneach) den chuid is mó agus tá na scóranna go léir níos ísle. Go minic ní raibh fianaise láidir sna cásanna baininscneacha gur bean a bhí i gceist aon áit i ngar don spriocfhocal. Mar sin, tá cur chuige nua ag teastáil atá in ann an fhianaise scaipthe sin a aimsiú agus a úsáid go héifeachtach.

Conclúid

San alt seo, rinne mé iarracht solas a chaitheamh ar fhadhb thromchúiseach sa réimse meaisínfhoghlama, is é sin claonadh inscne atá leabaithe i gcorpas mór na Gaeilge. Chonaic muid an tionchar a imríonn an claonadh seo ar leabú focal ar an gcéad dul síos, agus ansin ar líonra néarach a thuarann athruithe tosaigh sa Ghaeilge. Cé go bhfuil mórán taighde ann ar theicnící éagsúla chun claonadh inscne a mhaolú i leabú focal, níl sé soiléir go bhfuil aon cheann acu go hiomlán éifeachtach. Mar sin, sa rannán deireanach, bhain mé úsáid as cur chuige *ad hoc* a rinne mionathrú ar na sonraí traenála, ach níor éirigh liom an claonadh inscne sa líonra néarach a laghdú. Tá i bhfad níos mó taighde de dhíth ar na ceisteanna seo i gcomhthéacs na Gaeilge, go háirithe toisc go bhfuil líonraí néaracha den sórt seo in úsáid go forleathan cheana féin. Gan idirghabháil fheiliúnach, tá an chontúirt ann go ndéanfadh na huirlisí seo dochar do dhaoine trí chlaontaí de chineálacha éagsúla a scaipeadh agus fiú a threisiú, agus é a dhéanamh faoi bhréagriocht na hoibiachtúlachta.

Leabharliosta

Bolukbasi, T. *et al.*, (2016) 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings'. *Advances in Neural Information Processing Systems*, Barcelona, 5-10ú Nollaig. 4349–57.

Devlin, J., *et al.*, (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Imleabhar 1, Minneapolis, 2–7ú Meitheamh. 4171–86.

Dixon, L. *et al.*, (2018) 'Measuring and mitigating unintended bias in text classification'. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, New Orleans, 2–3ú Feabhra. 67–73.

Ethayarajh, K., Duvenaud, D. & Hirst, G., (2019) 'Understanding Undesirable Word Embedding Associations'. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Flórans na hIodáile, 28 Iúil–2ú Lúnasa. 1696–705.

Fiontar, (2019) 'Beathaí Ainm agus ceisteanna comhionannais inscne'. *Blag Thionscadail Fiontar & Scoil na Gaeilge*, 9 Nollaig. Le fáil ag: https://www.gaois.ie/blg/ga/beathai-ainm-agus-ceisteanna-comhionannais-inscne/ (Íoslódáilte: 13 Nollaig 2019).

Gonen, H. & Goldberg, Y., (2019) 'Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them'. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Imleabhar 1, Minneapolis, 2–7ú Meitheamh. 609–14.

James-Sorenson, H. & Alvarez-Melis, D., (2019) 'Probabilistic Bias Mitigation in Word Embeddings'. Le fáil ag: https://arxiv.org/pdf/1910.14497.pdf (Íoslódáilte: 22 Eanáir 2020).

Kiritchenko, S. & Mohammad, S.M., (2018) 'Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems'. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, New Orleans, 6–7ú Meitheamh. 43–53.

Kuczmarski, J., (2018) 'Reducing gender bias in Google Translate', *The Keyword*, 6 Nollaig. Le fáil ag: https://www.blog.google/products/translate/reducing-gender-bias-google-translate/ (Íoslódáilte: 16 Eanáir 2020).

Lambrecht, A. & Tucker, C.E., (2019) 'Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads', *Management Science*, 65 (7). 2966–81.

Lu, K. *et al.*, (2018) 'Gender bias in neural natural language processing'. Le fáil ag: https://arxiv.org/pdf/1807.11714 (Íoslódáilte: 16 Eanáir 2020).

Mikolov, T., Yih, W.T. & Zweig, G., (2013) 'Linguistic regularities in continuous space word representations'. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, 9–14ú Meitheamh. 746–51.

Park, J.H., Shin, J. & Fung, P., (2018) 'Reducing Gender Bias in Abusive Language Detection'. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, An Bhruiséil, 31ú Deireadh Fómhair–4ú Samhain. 2799–804.

Prates, M.O.R., Avelar, P.H.C. & Lamb, L.C., (2018) 'Assessing gender bias in machine translation — a case study with Google Translate'. *Neural Computing and Applications*. 1–19. https://doi.org/10.1007/s00521-019-04144-6

An Phríomh-Oifig Staidrimh, (2017) 'Women and Men in Ireland 2016: Education'. Le fáil ag: https://www.cso.ie/en/releasesandpublications/ep/p-wamii/womenandmeninireland2016/education/ (Íoslódáilte: 17 Eanáir 2020).

Prost, F., Thain, N. & Bolukbasi, T., (2019) 'Debiasing Embeddings for Reduced Gender Bias in Text Classification'. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, Flórans na hIodáile, 2ú Lúnasa. 69–75.

Rudinger, R., *et al.*, (2018) 'Gender Bias in Coreference Resolution'. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, Imleabhar 2 (Gearrpháipéir), New Orleans, 1–6ú Meitheamh. 8–14.

Scannell, K.P., (2020) 'Neural Models for Predicting Celtic Mutations'. Faoi athbhreithniú.

Tatman, R., (2017) 'Gender and dialect bias in YouTube's automatic captions'. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, Valencia, 4ú Aibreán. 53–9.

Zhao, J., *et al.*, (2018) 'Learning Gender-Neutral Word Embeddings'. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, An Bhruiséil, 31ú Deireadh Fómhair–4ú Samhain. 4847–53.