Tiomsú Corpais don Taighde Foclóireachta: *Corpas Foclóireachta na Gaeilge (CFG2020)*

[Corpus Creation for Lexicographical Research: *Corpas Foclóireachta na Gaeilge (CFG2020)*]

Mícheál J. Ó Meachair, Ollscoil Chathair Bhaile Átha Cliath; Brian Ó Raghallaigh, Ollscoil Chathair Bhaile Átha Cliath; Úna Bhreathnach, Ollscoil Chathair Bhaile Átha Cliath; Gearóid Ó Cleircín, Ollscoil Chathair Bhaile Átha Cliath; Kevin Scannell, Saint Louis University

Achoimre

Leagtar amach sa pháipéar seo na céimeanna a leanadh le *Corpas Foclóireachta na Gaeilge 2020* (CFG2020), corpas aonteangach 77.3 milliún focal, a thiomsú. Mínítear comhthéacs an tionscadail agus na riachtanais a spreag na cinntí a tógadh lena linn. Déantar cur síos ansin ar chéim an tiomsaithe agus ar na céimeanna próiseála. Tugtar spléachadh ar inneachar an chorpais, ar an acmhainn a cruthaíodh lena chuardach, agus ar an gcineál anailíse agus taighde a cumasaíodh leis seo. Tiomsaíodh CFG2020 ar an tuiscint gur réamhchéim é ar thionscadal níos leithne corpais, is ar an gcúis sin a dhéantar moltaí i dtaca lena fheabhsú agus lena mhéadú.

Eochairfhocail: Corpas foclóireachta, Foclóireacht, Corpais, Gaeilge

Abstract

This paper sets out the steps followed in the compilation of Corpas Foclóireachta na Gaeilge 2020 (CFG2020)*, a monolingual 77.3 million word Irish-language corpus. The context and circumstances of the project are explained, along with the motivation for various decisions made. The compilation and processing stages are described in detail. The contents of the corpus are outlined and the resource created to query CFG2020 is presented, along with reference to the kinds of analysis and research which it enables. CFG2020 was created as a first step towards a proposed larger corpus project, and suggestions for improvement and expansion are therefore proposed.*

Keywords*: Lexicographic corpus, Lexicography, Corpora, Irish language*

1. Comhthéacs an Taighde

Ba é an tionscadal a bhfuil cur síos air in Kilgarriff et al. (2006) a chruthaigh an acmhainn dheireanach corpais don taighde foclóireachta Gaeilge. Rinneadh an obair sin i gcomhthéacs a bhí an-chosúil leis an gcomhthéacs ina ndearnadh obair an tionscadail atá faoi chaibidil sa pháipéar anseo, in achar teann ama:

> The other key requirements were that the corpus should form an adequate data source to support a major programme of lexicographic work, and that it should be collected and encoded with the one-year set-up phase for the new dictionary. Kilgarriff et al. (2006, 128)

Cruthaíodh dhá chorpas le linn an tionscadail bliana sin: corpas Béarla, agus Nua-Chorpas na hÉireann (NCÉ); corpas Gaeilge 29.7 milliún focal, ar ábhar as ceann de na trí mhórchanúint (Gaeilge na Mumhan, Gaeilge Chonnacht agus Gaeilge Uladh) a leath de agus ábhar nár bhain le haon chanúint ar leith an leath eile. Fuarthas breis agus 8 milliún de na focail Ghaeilge sin, an tríú cuid den líon focal iomlán, ó Institiúid Teangeolaíochta Éireann (ITÉ) a bhí tar éis corpas Gaeilge a fhorbairt mar chuid den tionscadal ilteangach Eorpach PAROLE (Kilgarriff et al., ibid). Bhain na téacsanna a bailíodh do NCÉ leis an tréimhse 1883–2005. Roghnaíodh an dáta tosaigh seo sa chaoi go leanfadh an corpas nua seo ar aghaidh ó chorpas an Acadaimh Ríoga (Royal Irish Academy, 2004), a raibh an bhliain deiridh 1882 luaite leis[1] ag an am. (1926 atá luaite leis anois). Saothar mór is ea é, ach ós rud é nár leanadh den bhailiú ábhair tar éis chríoch an tionscadail, chuaigh NCÉ as dáta. Tá cur síos déanta ar cheann de na bealaí ar imigh NCÉ as dáta in Ó Meachair (2020) agus an-chuid de théarmaíocht agus gnáthfhocail na mílaoise nua in easnamh air.

[1] URL: https://www.ria.ie/research-projects/focloir-stairiuil-na-gaeilge

Rinne údair an pháipéir seo anailís ar an líon focal a bailíodh do NCÉ de réir bliana le heolas grinn a fháil ar dháileadh an inneachair. Aimsíodh gur bhain 36% den líon focal iomlán leis na blianta 2002 agus 2004 amháin, agus gur bhain 51.72% den líon focal iomlán leis an tréimhse 2000–2004. Meastar gur mar seo a bhí toisc nach raibh Kilgarriff et al. ábalta tarraingt ar thacar mór sonraí a bailíodh cheana, seachas corpas PAROLE ITÉ, rud a d'fhág go raibh orthu an-chuid sonraí a bhailiú as an nua. Bíonn níos lú oibre i gceist, de ghnáth, le cáipéisí idirlín a thiomsú ina n-ábhar corpais ná mar a bhíonn le leabhair agus ábhar clóite eile. Ní raibh an oiread céanna ábhair Ghaeilge ar fáil ar líne ón tréimhse roimh an mbliain 2000, agus dá bhrí sin, ba ghá brath go mór ar an tréimhse chúng seo 2000 – 2004. D'éirigh le foireann NCÉ an sprioc a bhí rompu a bhaint amach agus corpas mór a thiomsú, i gcomhthéacs an ama sin, a bhí cothromaithe de réir na canúna. Má tá locht le fáil ar air, tá an locht céanna le sonrú ar neart tionscadal corpais eile nach é, agus sin nárbh acmhainn dó maireachtáil sách fada le go dtógfaí ar a raibh bainte amach agus chun go ndéanfaí nuashonrú rialta air. Mar chuid den tionscadal céanna, cruthaíodh corpas Béarla 225 milliún focal le tacú leis an taighde foclóireachta dátheangach. Bhí na taighdeoirí ábalta tarraingt ar na céadta milliún focal Béarla trí iarratais a chur chuig tionscadail fhadtéarmacha bhailithe corpas.[2]

Is léir ó thionscadail chorpais a reáchtáladh i gcás teangacha Eorpacha eile gur cur chuige rathúil é corpas a thiomsú trí ábhar a bailíodh cheana a thabhairt le chéile agus ábhar nua-bhailithe a chur leis. Go deimhin, feictear in Knight et al. (2020a, 2020b) agus tionscadal corpais náisiúnta na Breatnaise (CorCenCC) idir chaláin acu gur nós i gcónaí é corpais agus bailiúcháin éagsúla a thabhairt le chéile nuair atá corpas náisiúnta á thiomsú agus go bhfuil na

[2] San áireamh sa líon mór focal seo bhí an 100 milliún focal a bhí sa *British National Corpus* (BNC), 100 milliún focal eile a tógadh ón *Gigaword Corpus* leis an Linguistic Data Consortium le toirt a chur leis na sonraí Béarla agus 25 milliún focal a tiomsaíodh as an nua ó fhoinsí digiteacha le hionadaíocht a dhéanamh ar Bhéarla na hÉireann.

teicneolaíochtaí teanga atá ar fáil don teanga an-tábhachtach sa phróiseas seo. Is corpas cothromaithe é CorCenCC (Knight et al. 2020a, 2020b) ar tiomsaíodh é don taighde ginearálta teangeolaíochta agus mar bhunús le ríomhuirlisí a fhorbairt. Sa chaoi go bhféadfaí CorCenCC a úsáid chuige seo ba ghá go bhfágfaí ábhar áirithe eile ar lár. Is é sin, ábhar a tháinig salach ar choinníollacha na coibhéise agus ábhar nach bhféadfaí é a roinnt ar an bpobal i gcoitinne de bharr na gceangal cóipchirt. Fágann sé seo gur tiomsaíodh corpas oscailte agus inroinnte, seachas corpas ollmhór dúnta. Is nós le grúpaí taighde a thiomsaíonn corpais ollmhóra don taighde foclóireachta iad a choimeád dúnta, ach síntiús a bheith le híoc as iad a úsáid, cuir i gcás COBUILD, OED, agus an *Corpus of Contemporary American English* (COCA). I gCorpas na Seicise (*Český národní korpus*[3]) is féidir cuardach a dhéanamh ar ollchorpas amháin (os cionn 3.5 billiún focal), a tiomsaíodh le hábhar a bailíodh cheana. Tá briseadh síos ar na corpais éagsúla a tiomsaíodh ar fáil trí shuíomh an tionscadail[4], agus eolas ar na corpais arbh fhéidir cuardach a dhéanamh orthu ar fáil freisin[5]. I gcás Chorpas na Baiscise, *Egungo Testuen Corpusa*[6], is féidir cuardach a dhéanamh ar ollchorpas amháin 355 milliún focal agus tugtar rochtain don phobal agus do thaighdeoirí ar chorpais éagsúla tríd an suíomh gréasáin.

I bhfianaise na hoibre atá beartaithe ar thionscadal nua foclóireachta, thug Foras na Gaeilge maoiniú do ghrúpa taighde Gaois chun réamhobair a dhéanamh ar chorpas foclóireachta, a bhfuil cur síos air thíos. Tugadh faoin réamhobair sin ar an tuiscint gur cur chuige tiomsaithe den sórt céanna atá pléite thuas a bheadh feiliúnach, cur chuige a tharraingeodh le chéile acmhainní maithe corpais a bhí ann cheana agus a chuirfeadh leo.

[3] URL: https://www.korpus.cz/
[4] URL: https://wiki.korpus.cz/doku.php/en:cnk:uvod
[5] URL: https://www.korpus.cz/kontext/corpora/corplist
[6] URL: https://www.ehu.eus/etc/

2. Scóip an Tionscadail

Rinneadh an obair ar thionscadal CFG2020 a ndéantar cur síos air anseo sa tréimhse 1 Lúnasa–31 Nollaig 2020. Thart ar 7 mí oibre (coibhéis lánaimseartha) san iomlán a bhí i gceist. Ba é bunsprioc an tionscadail seo corpas a thiomsú a bheadh úsáideach agus tús á chur le cruthú foclóirí nua Gaeilge-Béarla agus Gaeilge-Gaeilge. Níl aon mhórfhoclóir aonteangach comhaimseartha ar fáil i gcás na Gaeilge agus foilsíodh an mórfhoclóir deireanach Gaeilge-Béarla ag deireadh na 1970idí (Nic Pháidín, 2008). Tá sé ar intinn ag Foras na Gaeilge an dá fhoclóir nua seo a fhorbairt i gcomhthráth sna blianta beaga amach romhainn agus beidh corpas mór aonteangach Gaeilge agus uirlisí cuí cuardaigh is anailíse riachtanach chun an sprioc sin a bhaint amach. D'aontaigh an dhá pháirtí (Foireann Thionscadal Foclóireachta Fhoras na Gaeilge agus grúpa taighde Gaois) ar na clocha míle seo a leanas (tá míniú níos cuimsithí ar na míreanna éagsúla thíos):

1. Comhéadan gréasánbhunaithe príobháideach a fhorbairt leis an ábhar corpais a óstáil agus a chuardach. Óstáil an ábhair sa néal, rochtain de réir na socruithe cóipchirt reatha a dheimhniú, agus leathanaigh eolais faoin ábhar a chur leis an suíomh.
2. Tiomsú ábhair ó cheithre fhoinse:
    a. Corpas na Gaeilge Comhaimseartha (CGC)
    b. Nua-Chorpas na hÉireann (NCÉ)
    c. Trascríbhinní Raidió na Gaeltachta (RnaG)
    d. Corpas an Chrúbadáin.
3. An t-ábhar ó na foinsí scríofa (CGC, NCÉ, An Crúbadán) a thabhairt chun rialtachta ó thaobh clibeanna, leagan amach agus formáidithe de. Clibeáil ranna cainte, leamaí agus caighdeánú ar an ábhar.
4. Liostaí minicíochta a chruthú.
5. Scóipeáil agus pleanáil i dtreo mórthionscadail chorpais. Cinneadh faoi fhoilsiú fadtéarmach an ábhair phíolótaigh (ar shuíomh nua, ar gaois.ie nó ar shuíomh eile de chuid an Fhorais).
6. Scóipeáil a dhéanamh ar chomhaontuithe cóipchirt shonraí an chorpais le gurbh fhéidir tacar substantach, ar a laghad, de a chur ar fáil trí chomhéadan foclóireachta dála SketchEngine.
7. Cáipéisiú ar an obair agus ar na torthaí.

Ní raibh sé i gceist go mbeadh CFG2020 sách mór ná sách forbartha ann féin leis an taighde foclóireachta a chur i gcrích, ach go líonfadh sé cuid den bhearna a fágadh nuair nár cuireadh le NCÉ go leanúnach ó 2005 i leith. Roghnaíodh na cheithre fhoinse atá luaite faoi chloch mhíle #2 ar an gcéad dul síos toisc go raibh rochtain ag na taighdeoirí orthu go saoráideach. Aithníodh freisin go raibh na sonraí ar ardchaighdeán agus go mbeadh na meiteashonraí so-láimhsithe dá réir sin, rud a laghdódh an t-am próiseála a bheadh le caitheamh. Aontaíodh go gcaithfí cur leis na foinsí úd amach anseo, ach acmhainní a bheith ar fáil chuige sin, agus tá anailís ar siúl ag foireann an tionscadail chun bearnaí ó thaobh réimse nó achar ama de a aimsiú. Cuid thábhachtach eile den tionscadal seo ná na sonraí a bheith leagtha amach ar bhealach a chumasódh agus a d'éascódh tuilleadh forbartha agus tiomsaithe corpais amach anseo. Is é sin le rá, go mbeifí in ann cur le méid an chorpais agus saibhriú a dhéanamh ar mheiteashonraí an chorpais.

3. Dearadh an Chorpais

Rinneadh gach iarracht an oiread réimsí agus ab fhéidir a chur san áireamh in CFG2020 sa chaoi go ndéanfadh sé samplaí maithe ionadaíocha den Ghaeilge chomhaimseartha a sholáthar don taighde foclóireachta. Ba ghá go mbeadh formhór mór an ábhair ar ardchaighdeán ó thaobh na heagarthóireachta de ach aithníodh freisin gur cuid thábhachtach den Ghaeilge chomhaimseartha an méid a mbítear á phostáil ar na meáin shóisialta agus ar bhlaganna, chomh maith le hábhar urlabhra, cé nach mbeadh eagarthóireacht den chineál céanna i bhfeidhm ar na foinsí sin. Is d'aon ghnó a cuireadh an bhliain 2020 le códainm an chorpais. Féachtar ar CFG2020 mar an chéad leagan de chorpas foclóireachta — corpas a gcaithfear é a fhorbairt ar bhealaí éagsúla le taighde foclóireachta a chur i gcrích atá

de réir na ndea-chleachtas idirnáisiúnta.

3.1 Bunfhoinsí na Sonraí

Ceithre phríomhfhoinse sonraí atá i gceist in CFG2020: Corpas na Gaeilge Comhaimseartha (2000 i leith), Nua-Chorpas na hÉireann (1883–2004), bailiúchán tras-scríbhinní Raidió na Gaeltachta (2013–2020), agus Corpas an Chrúbadáin (2000 i leith).

Tá Corpas na Gaeilge Comhaimseartha (CGC) á fhorbairt agus á óstáil ag grúpa taighde Gaois (DCU) ó 2015 i leith. Is éard atá ann corpas neamhchothromaithe de théacsanna Gaeilge a bhfuil eagarthóireacht déanta orthu agus a foilsíodh ó thús an 21ú haois i leith. Bhí 30 milliún focal ann ag tús thionscadal CFG2020 agus bailíodh 4 mhilliún focal sa bhreis lena linn. Tá CGC ar fáil go poiblí[7] ar bhealach teoranta simplí, agus in úsáid ag foireann Gaois don taighde. Déantar suas le 18,000 cuardach air gach mí agus tá fás seasta ar líon na gcuardach sin. Tá sé de bhuntáiste ag CGC thar chorpas ar bith eile atá forbartha don Ghaeilge go bhfuil sé suas chun dáta agus go bhfuil mórán dua caite le hábhar nach raibh ar fáil ar líne riamh a bhailiú agus an t-ábhar sin a chur in oiriúint do chorpas. Bhí clibeáil ar siúl ar CGC tráth tosaithe an tionscadail seo agus beartaíodh an próiseas clibeála céanna a chur i bhfeidhm ar na trí fhoinse eile dá réir sin.

Tá *Nua-Chorpas na hÉireann* (NCÉ), a pléadh thuas, á óstáil ag *LexicalComputing* ar son Fhoras na Gaeilge. Is féidir an 30 milliún focal seo de théacsanna ardchaighdeáin Gaeilge scríofa a chuardach trí leagan teoranta de SketchEngine ach clárú le Foras na Gaeilge ar dtús[8]. Cuireadh na sonraí ar fáil in ollchomhad amháin XML agus córas clibeála i bhfeidhm orthu nach raibh ar aon dul le haschur an chlibeálaí a bheadh in úsáid ar an tionscadal seo, is é sin an

[7] URL: https://www.gaois.ie/ga/corpora/monolingual/
[8] URL: http://corpas.focloir.ie/

clibeálaí a bhfuil cur síos déanta air in Uí Dhonnchadha (2009).

Is iad na tras-scríbhinní ar chláir Raidió na Gaeltachta an tríú foinse. Maoiníonn Foras na Gaeilge tras-scríobh ar chláracha áirithe raidió chun sprice taighde foclóireachta. Roghnaíodh na cláir *Barrscéalta*, *Iris Aniar*, agus *An Saol ó Dheas*, ar an tuiscint go bhfaightear iontu sampla ionadaíoch comhaimseartha d'urlabhra na dtrí mhórchanúint. Bhí 4.5 milliún focal tras-scríofa ar fáil ag tús an tionscadail. Ní raibh an t-ábhar tiomsaithe i bhformáid chorpais agus cuireadh ar fáil i bhformáid DOC nó DOCX é. Ní raibh sé ailínithe le comhaid fuaime agus ní raibh sé clibeáilte. Bhí na comhaid seo le tiontú go TXT agus bhí oiread ábhar iomarcach agus ab fhéidir le baint amach, dála teidil, intreoracha, agus fonótaí.

Fuarthas an ceathrú foinse, Corpas an Chrúbadáin (Scannell, 2007), ó thionscadal corpais atá á reáchtáil ón mbliain 2000 i leith. Bailítear ábhar scríofa go huathoibríoch ón idirlíon i gcás 2,200 teanga nach bhfuil na ríomhacmhainní cuí acu. Tá thart ar 200 milliún focal sa chorpas Gaeilge ann. Cuirtear céimeanna éagsúla próiseála i bhfeidhm ar na doiciméid uile le caighdéan an ábhair a dheimhniú. Ní bheadh sé d'acmhainn ag tionscadal CFG2020 Corpas an Chrúbadáin ina iomláine a úsáid. Ina ionad sin, beartaíodh go roghnófaí c. 10 milliún focal as chun bearna(í) aitheanta ó thaobh réimse de a líonadh, go háirithe maidir le hábhar ó bhlaganna agus ó na meáin shóisialta, dhá réimse nach bhfaightear sna bunfhoinsí eile. Mar théacs lom a bhí an t-ábhar seo stóráilte, gan aon chlibeáil ná anótáil ná meiteashonraí.

Nuair a chuirtear na ceithre fhoinse san áireamh, is féidir idirdhealú a dhéanamh idir c. 8.5 milliún focal a bhí le tiomsú as an nua (4m focal a bailíodh as an nua do CGC agus 4.5m focal ó thras-scríbhinní Raidió na Gaeltachta) agus ábhar a bhí ann cheana agus ar baineadh athearraíocht as. Ba ghá próiseáil agus clibeáil as an nua a dhéanamh ar fhormhór an ábhair sin, áfach, mar a fheicfear thíos.

Faoi mar a bheifí ag súil leis, tá socruithe éagsúla cóipchirt i bhfeidhm ar na ceithre fhoinse thuasluaite, socruithe a rinneadh ag tréimhsí éagsúla agus faoi laincisí éagsúla. Aontaíodh dá bhrí sin go mbeadh rochtain ar an gcorpas teoranta do líon beag úsáideoirí cláraithe trí thairseach ar shuíomh gaois.ie.[9] D'fhéadfaí an rochtain a fhairsingiú amach anseo, ach deis a bheith ann socruithe nua a dhéanamh leis na sealbhóirí éagsúla cóipchirt.

3.2 Próiseáil an Chorpais

Leagtar amach sa mhír seo na céimeanna réamhphróiseála a cuireadh i gcrích ar ábhar an chorpais, an chlibeáil a rinneadh ar an gcorpas, agus ansin an iarphróiseáil a rinneadh ar an ábhar clibeáilte sin.

Is éard atá i gceist le réamhphróiseáil ná na doiciméid agus a n-inneachar a ullmhú le go mbeidh an clibeálaí ábalta iad a phróiseáil mar is ceart leis an bpríomhchéim phróiseála a chur i gcrích go cruinn: clibeáil na ranna cainte, na leamaí, agus na sonraí moirfeolaíochta. Tá thart ar 80,000 comhad sa chorpas agus mar sin ní raibh sé réadúil na céimeanna glantacháin, eagarthóireachta, agus iarphróiseála a dhéanamh de láimh. Rinneadh an réamhphróiseail seo le ríomhchláir a scríobhadh sa teanga ríomhchlárúcháin Python. Tá soláthar ar ríomhuirlisí don Ghaeilge teoranta, i gcomparáid le teangacha eile. Tá dul chun cinn déanta ag taighdeoirí le sonraí Gaeilge a phróiseáil ar bhealaí nua-aimseartha a chuideodh le tuarthéacs nó meaisín-aistriúchán, mar shampla, ach níor seoladh aon chlibeálaí a dhéanfadh difear suntasach don chaoi a bpróiseáiltear sonraí Gaeilge le breis agus 10 mbliana anois. Tá sé seo fíor in ainneoin an riachtanas a bheith luaite ag Judge et al. (2012, lch. 17). B'fhacthas do na húdair le linn an tionscadail seo go bhfuil réimse níos leithne carachtar in úsáid anois ná mar a bhí deich mbliana

[9] URL: https://www.gaois.ie/ga/corpora/lexicography/in/

ó shin, tráth ar fhorbair Uí Dhonnchadha clibeálaí ranna cainte (Uí Dhonnchadha, 2009). Sampla de na carachtair seo ná emojíonna, agus carachtair nó siombail speisialta. Is cuid de chomhthéacs na habairte úsáid an emoji anois, pé acu ag gáire faoina dúradh atá an t-úsáideoir nó ag léiriú go bhfuil íoróin ag baint leis an méid a dúradh trí chaochadh súile a chur leis. Ní féidir an emoji a scriosadh mar sin — chaillfí cuid den chomhthéacs dá ndéanfaí sin. Beartaíodh neamhfhocail a chur in ionad gach emoji sa chaoi go mbeadh an clibeálaí ábalta na doiciméid a phróiseáil. Bunaíodh an t-ionadú seo ar fhothacar de na teaghráin chaighdeánacha a thagraíonn do chineálacha éagsúla emoji faoi mar atá siad leagtha amach don Ghaeilge ag an Common Locale Data Repository.[10] Mar shampla:

 [emoji-straoiseog-ag-gáire-le-súile-dúnta]

 [emoji-straoiseog-ag-gáire-le-fuarallas]

 [emoji-sna-trithí-gáire]

 [emoji-straoiseog-le-deora-áthais]

Roghnaíodh comhfhocal fada mar seo atá ina neamhfhocal sa chaoi nach gcuirfí isteach ar mhinicíocht leithéidí “straoiseog”, “gáire”, “fuarallas”, “deora”, agus araile i dtorthaí staitistiúla.

I measc na dtascanna eile a cuireadh i gcrích bhí cruthú agus pairseáil XML, rialú caighdeáin ar an gclibeáil agus ar leagan amach na gclibeanna, comhaireamh líonta focal, asbhaint samplaí de réir patrúin slonn rialta, agus ríomhchlár a d’eagraigh inneachar na dtéacsanna de réir na habairte. Úsáideadh ríomhchlár le hábhar nach raibh ionchódú ceart air a

[10] URL: http://cldr.unicode.org/

mharcáil le go bhféadfaí é a sheiceáil de láimh freisin, agus ceartaíodh gach earráid ionchódaithe a aimsíodh.

Próiseáladh gach comhad ina cheann agus ina cheann leis an gclibeálaí a forbraíodh le linn Uí Dhonnchadha (2009). Tá cur síos iomlán ar an bpróiseas clibeála le fáil in Uí Dhonnchadha (2009) go príomha, agus mínítear cuid den phróiseas in Kilgarriff et al. (2006). Tá ionchur samplach agus an t-aschur críochnúil a soláthraíodh sa tionscadal seo le feiceáil i bhFíor 1.

Fíor 1

Leagan Amach XML na Sonraí

Ionchur: “Bhreac an Sáirsint ina nótaí gur tháinig cuma fheargach ar William.”
Aschur:

```
<doc>
...
<word>
<token>Bhreac</token>
      <lemma>breac</lemma><POS>Verb VTI PastInd Len</POS>
</word>
<word>
<token>an</token>
      <lemma>an</lemma><POS>Art Sg Def</POS>
</word>
<word>
<token>Sáirsint</token>
      <lemma>sáirsint</lemma><POS>Noun Masc Com Sg DefArt</POS>
</word>
<word>
<token>ina</token>
      <lemma>i</lemma><POS>Prep Poss 3P Sg Masc</POS>
</word>
<word>
<token>nótaí</token>
      <lemma>nóta</lemma><POS>Noun Masc Com Pl</POS>
</word>
<word>
<token>gur</token>
      <lemma>gur</lemma><POS>Part Vb Cmpl Past</POS>
</word>
<word>
<token>tháinig</token>
      <lemma>tar</lemma><POS>Verb VI PastInd Len</POS>
</word>
<word>
<token>cuma</token>
      <lemma>cuma</lemma><POS>Noun Fem Com Sg</POS>
</word>
<word>
<token>fheargach</token>
      <lemma>feargach</lemma><POS>Adj Fem Com Sg Len</POS>
</word>
<word>
<token>ar</token>
      <lemma>ar</lemma><POS>Prep Simp</POS>
</word>
<word>
<token>William</token>
      <lemma>William</lemma><POS>Prop Noun Masc Com Sg</POS>
</word>
<word>
<token>.</token>
      <lemma>.</lemma><POS>Punct Fin</POS>
</word>
…
</doc>
```

Leagadh an XML amach ar an mbealach seo (Fíor 1) sa chaoi gur féidir forbairt a dhéanamh air. D'fhéadfaí eagarthóireacht a dhéanamh ar chlibeáil an chorpais trí chomhéadan eagarthóireachta freisin, bíodh an t-eagarthóir ag cur clibeanna breise leis nó ag ceartú clibe.

4. Torthaí na hOibre

Bhí 77.3 milliún focal i gcorpas CFG2020 faoi dheireadh an tionscadail. Bhí gach focal clibeáilte le leama, roinn chainte, agus sonraí moirfeolaíochta. Tá na sonraí sábháilte i bhformáid XML. Is é seo thíos an méid focal a tháinig as na foinsí éagsúla.

Tábla 1

Líon Focal de réir Foinse

Foinse	**Líon deiridh focal**
Corpas na Gaeilge Comhaimseartha (CGC)	34m
Nuachorpas na hÉireann (NCÉ)	29.7m
Tras-scríbhinní Raidió na Gaeltachta (RnaG)	4.5m
Corpas an Chrúbadáin	9.3m
Iomlán	**77.3m**

Cé nach bhfuil clibeáil bheacht déanta ar an ábhar ó thaobh réimse de go fóill, is féidir a rá go bhfuil ábhar as na réimsí seo a leanas ar fáil i CFG2020: saothair ficsin agus neamhfhicsin, nuachtáin agus irisí, tréimhseacháin (acadúla den chuid is mó), ábhar oifigiúil (e.g. doiciméid rialtais agus dlíthiúil), ábhar gréasáin, blaganna, postálacha ó na meáin shóisialta, agus ábhar

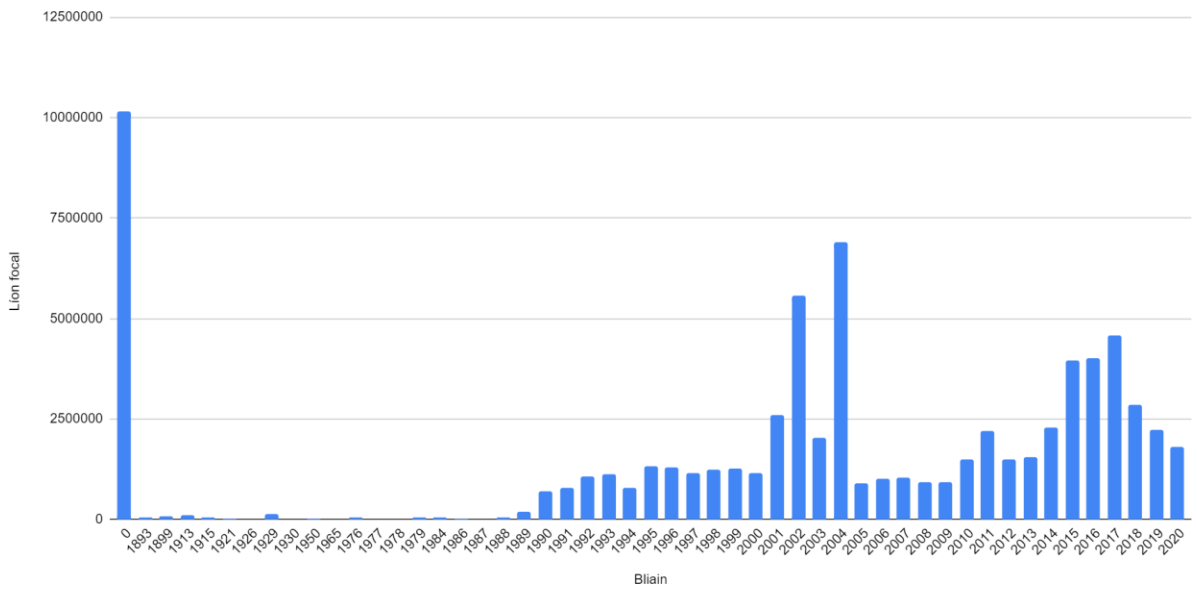
cráifeach. B'as NCÉ a tháinig formhór an ábhair (~12% den iomlán) nach mbaineann le foinsí nuachta ná le saothair litríochta, amhail an t-ábhar dlí agus an t-ábhar oideachasúil. Má tá deis ann CFG2020 a fhorbairt amach anseo moltar díriú go sonrach ar ábhar as réimsí nach réimsí na nuachta agus na litríochta iad, mar shampla spórt (rialacha cluichí Gaelacha agus agallaimh le himreoirí, mar shampla), dlí agus oideachas.

Ábhar urlabhra atá i 6% de na focail in CFG2020 bunaithe ar c. 320 uair a chloig de thaifeadtaí raidió. Ní dhearnadh aon tiomsú corpais ar an ábhar tras-scríofa atá in úsáid in CFG2020. Níl aon fhorluí idir an t-ábhar tras-scríofa in CFG2020 agus an t-ábhar a úsáideadh in Uí Dhonnchadha (2013). Ba ghá caitheamh leis an ábhar tras-scríofa seo mar bhailiúchán tras-scríbhinní seachas corpas urlabhra mar sin. Rinneadh obair mhór leis an mbailiúchán tras-scríbhinní seo a chruthú, agus infheistíocht nach beag, agus is foinse sonraí é arbh fhiú go mór forbairt a dhéanamh air.

Baineann téacsanna CFG2020 leis an tréimhse 1883–2020. É seo ráite, is beag téacs ó dheireadh an 19ú haois agus ó thús an 20ú haois atá sa chorpas (féach *Fíor 2: Líon focal in CGC agus NCÉ de réir na bliana*). Baineann na seansonraí sin le NCÉ amháin (cé gur tiomsaíodh formhór an chorpais sin ó shonraí a cruthaíodh ón mbliain 2000 i leith).

Fíor 2

Líon Focal in CGC agus NCÉ de réir Bliana



Tabharfaidh an léitheoir faoi deara go bhfuil an líon focal thuas measartha íseal don tréimhse 2005–2014 agus don tréimhse 2019–2020 i gcomparáid leis na tréimhsí 2000–2004 agus 2015–2018. Cé nach raibh bliain ar leith sannta leis na sonraí a tógadh ó Chorpas an Chrúbadáin i gcónaí, baineann 1.8 milliún focal ó na meáin shóisialta leis na blianta 2013–2017, baineann 1.5 milliún focal le hóráidí rialtais ón tréimhse 2003–2017, agus baineann 6 mhilliún focal le blaganna a foilsíodh idir na blianta 2005 agus 2020. Tá obair shaibhrithe de dhíth ar mheiteashonraí an ábhair seo ón gCrúbadán ó thaobh an dáta de ach is léir go gcuideoidh siad le cothromaíocht níos fearr a bhaint amach ó bhliain go bliain sa tréimhse 2000–2020. Anuas air sin, tá c. 10.5 milliún focal a tháinig ó CGC agus NCÉ nach bhfuil bliain ar leith luaite leo sna meiteashonraí ar chúiseanna éagsúla (tá an líon focal seo luaite leis an mbliain 0 in *Fíor 2* thuas). Fágann sé sin nach bhfuil bliain luaite le 13.8% den ábhar (breis agus 10 milliún focal). Tiomsaíodh thart ar leathmhilliún de na focail atá in NCÉ as foinsí a foilsíodh roimh an mbliain

1977, is é sin, sular foilsíodh *Foclóir Gaeilge-Béarla* (Ó Dónaill 1977). Tá sé i gceist ag na húdair anailís a dhéanamh ar an tionchar a d'imreofaí ar inneachar an chorpais dá mbainfí an t-ábhar seo (ar sciar an-bheag den chorpas iomlán é) as CFG2020.

Cruthaíodh liostaí minicíochta as na téacschomharthaí uile in CFG2020. Is iad seo a leanas na deich n-ainmfhocal, na deich mbriathar agus na deich n-aidiacht nó dobhriathar is airde minicíochta sa chorpas.

Tábla 2

Na Focail is Airde Minicíochta in CFG2020

Ainmfhocal	**Briathar**	**Aidiacht / Dobhriathar**
Gaeilge	bhí	eile
duine	tá	amach
daoine	atá	maith
chuid	bhfuil	mór
lá	raibh	anois
chéile	bheith	isteach
áit	dhéanamh	amháin
teanga	dúirt	nua
am	mbeadh	ansin
deireadh	dul	anseo

Ní haon ionadh go bhfuil formhór na mbriathra san aimsir chaite agus san aimsir láithreach, agus formhór an chorpais ag baint leis an litríocht agus le míreanna nuachta, réimsí ina mbíonn cur síos ar shuíomh agus chomhthéacs go minic. Bíonn ainmfhocail a thagraíonn do dhaoine agus do chúrsaí ama coitianta in an-chuid teangacha, ach tá liosta ainmfhocal seo na

Gaeilge éagsúil agus focal ann ("Gaeilge") a thagraíonn do chúrsaí teanga. Dar linn go bhfuil an focal "Gaeilge" ar an ainmfhocal is airde minicíochta ar chúpla cúis. Ós rud é gur mionteanga í an Ghaeilge tharlódh sé go mbíonn pobal na teanga ag cur is ag cúiteamh faoi chúrsaí teanga go measartha minic i gcomparáid le pobail teanga eile. Anuas air seo, tá eagraíochtaí agus ócáidí go leor ann a bhfuil an focal úd san áireamh ina n-ainmneacha (e.g. Foras na Gaeilge, Seachtain na Gaeilge) agus a bhíonn faoi chaibidil go rialta sna meáin. Féach an spléachadh seo de na téacschomharthaí is airde minicíochta a bhí rangaithe mar ainmfhocail san *Oxford English Corpus* (OEC) agus sa *Corpus of Contemporary American English* (COCA)[11].

Tábla 3

Na Focail is Airde Minicíochta in OEC agus COCA

Focal	**Rangú de réir minicíochta in OEC/COCA**
one	35/51
time	55/52
people	61/62
year	63/54
back	81/108
two	84/80
way	90/84
day	98/90

[11] Aithnítear nach ainmfhocail i gcónaí cuid acu seo, ach rangaigh na corpais úd mar ainmfhocail ar dtús iad nuair a bhí an céad focal is airde minicíochta á bhfoilsiú acu. Bíonn an deacracht chéanna ann sa Ghaeilge le leithéidí "dó", "déanamh", "is", agus focail eile nach iad.

I réimse na teangeolaíochta corpais déantar tomhas ar shaibhreas léacsach corpais trí chóimheas cineál:téacschomhartha (*type:token ratio*) corpais éagsúla a ríomh agus a chur i gcomparáid lena chéile. Is tomhas an-simplí é a úsáideadh den chéad uair in Johnson (1944) agus a úsáidtear go fóill, go deimhin bhí an coibhéas in úsáid ag Biber ina fhoilseachán ceannródaíoch (Biber, 1988). Tá scór ard níos saibhre ná scór íseal ach is tomhas an-simplí é nach n-insíonn an scéal ar fad. Is fiú cuimhneamh nach n-áiríonn tomhas cineál:téacschomhartha an saibhreas ilfhoclach atá ag corpas, agus gheobhfadh corpas a luann an-chuid ainmneacha éagsúla daoine scór saibhris níos airde toisc gur focail éagsúla na hainmneacha sin ar fad — rud atá le tabhairt san áireamh agus ábhar ó na meáin shóisialta agus ó Dháil Éireann sna sonraí ó Chorpas an Chrúbadáin.

Tábla 4

Torthaí na hAnailíse Cóimheas Cineál:Téacschomhartha

Foinse	**Scór Cineál:Téacschomhartha**	**Líon Cineál sa Chorpas**
CFG2020	1.121%	866,696
Fothacar de Chorpas an Chrúbadáin	4.537%	422,046
NCÉ	1.206%	355,951
CGC	0.874%	297,046
Tras-scríbhinní RnaG	1.529%	68,805

Léamh amháin ar na staitisticí seo ná go mbaineann na corpais a tharraingíonn ar réimsí éagsúla an scór is airde amach. Tá clúdach níos fairsinge réimsí i gCorpas an Chrúbadáin seo agus in NCÉ i gcomparáid le CGC, mar shampla. Is fiú cuimhneamh go n-éiríonn sé níos deacra agus níos deacra scór ard a choimeád de réir mar a éiríonn corpas níos mó. Féach scór CFG2020

i gcomparáid le scór fhothacar an Chrúbadáin, atá ina chuid de CFG2020 dar ndóigh. Meastar, áfach, go bhfuil CGC saibhir ar bhealaí eile nach raibh deis iad a thomhas le linn an tionscadail seo. Ceaptar go bhfuil CGC an-saibhir ag leibhéal ilfhoclach, cuir i gcás, le comhlogaíochtaí agus frásaí a nochtadh. Bunaíodh an tátal seo ar an liosta 2-gram go 3-gram a gineadh as na foinsí uile, agus a bhfuil cur síos orthu i dTábla 5. Meastar go n-aimseofar struchtúir in ábhar urlabhra thras-scríbhinní RnaG nach n-aimseofaí in ábhar scríofa, agus go mbeidh idir shaibhreas agus luach breise ag baint leo siúd.

Táble 5

Torthaí na hAnailíse n-Graim

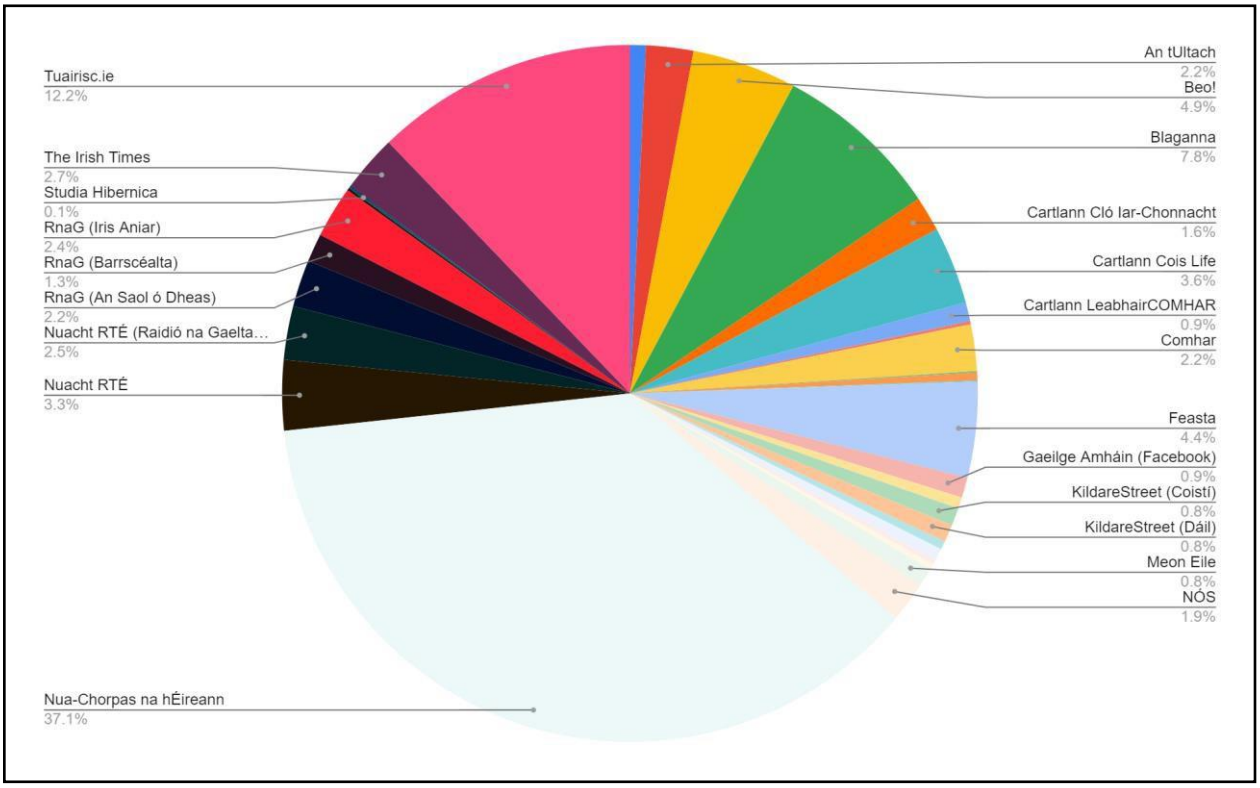
Foinse	**Líon 2-3gram uathúil sa bhfoinse**	**Líon 2-3gram uathúil sa bhfoinse normalaithe de réir an mhilliún focal**
CFG2020	41,821,073	568,994
CGC	21,053,393	619,217
NCÉ	18,442,855	620,971
Tras-scríbhinní RnaG	1,272,910	282,882
Fothacar de Chorpas an Chrúbadáin	1,062,221	114,217

Níl meiteashonraí uile na bhfoinsí corpais ag teacht le chéile, ar ndóigh. Bíonn ábhar CGC agus NCÉ sannta ar an tslí chéanna ó thaobh ábhar nuachta de, is é sin, go bhfuil gach doiciméad as an iris Nós in CGC luiate le Nós agus bíonn gach doiciméad as an iris Feasta in NCÉ luaite le Feasta. Ní mar seo atá i gcás na leabhar, in CGC bíonn leabhair sannta leis an teach foilseacháin agus in NCÉ bíonn gach leabhar sannta leis an údar. D'fhág sé sin go raibh ceithre bhailiúchán is fiche luaite le CGC, trí bhailiúchán le tras-scríbhinní Raidió na Gaeltachta

agus trí mhór-bhailiúchán[12] le Corpas an Chrúbadáin. Ní raibh de rogha ann i gcas NCÉ, áfach, ach caitheamh leis mar mhórbhailiúchán amháin go fóill. Tá na bailiúcháin seo léirithe i bhFíor 2 thíos.

Fíor 2

CFG2020 de réir Bailiúcháin



Is fiú a lua freisin go bhfuil meiteashonraí sannta le hábhar NCÉ a dhéanann cur síos ar chanúint an údáir, nach mbíonn luaite le foinse ar bith eile in CFG2020. Anuas air seo, tá taifeadadh déanta ar na húdair a bhí ina gcainteoirí dúchais i meiteashonraí NCÉ agus ar na téacsanna ar aistriúcháin iad freisin. Níl aon ábhar aistrithe i measc ábhar CGC.

[12] D'fhéadfaí na blaganna aonair a scagadh ina gceann agus ina gceann, nó caitheamh leo uile mar bhailiúchán amháin a dhéanann ionadaíocht ar ábhar ón mblagaisféar.

Rinneadh anailísí iniúchta na ranna seo le spléachadh a fháil ar inneachar CFG2020. Is ar mhaithe leis an bpleanáil agus le seiceáil na hoibre a rinneadh iad — cur chuige a bhí fiúntach agus torthúil.

5. Suíomh Gréasáin don Chuardach

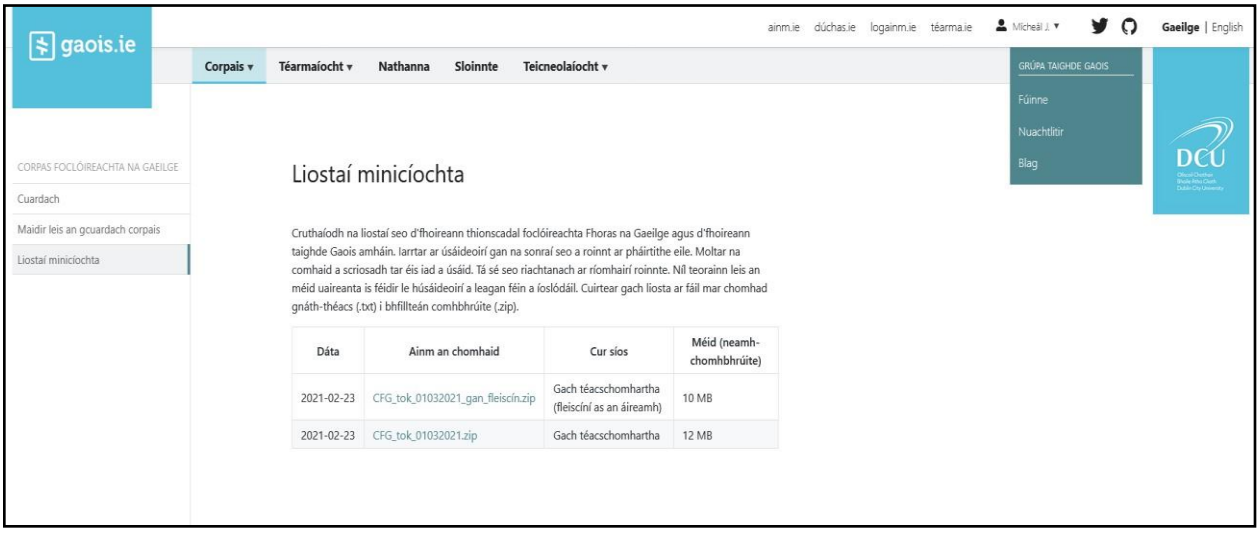
Is ar shuíomh CGC (www.gaois.ie/ga/corpora/monolingual/) a tógadh an suíomh nua príobháideach a forbraíodh don tionscadal seo, tá an chuma chéanna ar an dá leathanach cuardaigh ach caithfear logáil isteach ar shuíomh Gaois le húsáid a bhaint as CFG2020.

Físor 3

Comhéadan Cuardaigh CFG2020

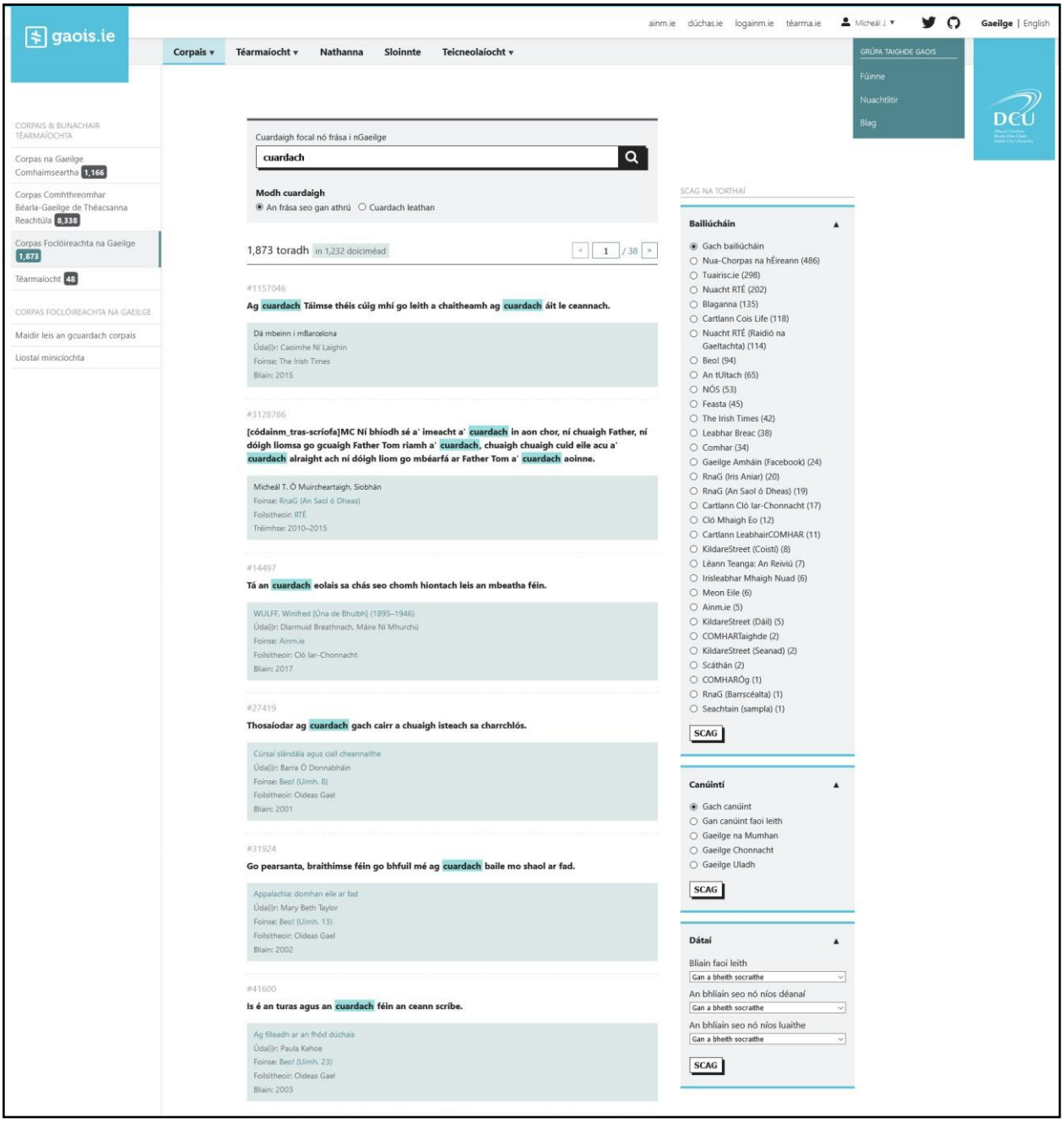
Físor 4

Leathanach le Liostaí Minicíochta agus Seachtháirgí a Íoslódáil



Fíor 5

Cuardach ar an bhFocal "cuardach"



Mar chuid de CGC, forbraíodh bunachar sonraí coibhneasta SQL Server chun bailiúcháin, doiciméid agus deighleáin an chorpais a stóráil agus a innéacsú ar mhaithe le luas

cuardaigh, agus comhéadan cuardaigh gréasáin chun na deighleáin seo a chuardach thar an idirlíon. Cuireann an comhéadan seo ar chumas an úsáideora cuardach lántéacs a dhéanamh sa chorpas trí théarma cuardaigh aonair nó frása ilfhoclach a chlóscríobh isteach i mbosca cuardaigh. Tá dhá mhodh cuardaigh ar fáil, ceann a aimsíonn torthaí ina bhfuil na téarmaí cuardaigh le fáil leis an litriú céanna agus san ord céanna inar clóscríobhadh iad sa bhosca cuardaigh, agus ceann a aimsíonn torthaí ina bhfuil leaganacha infhillte agus leaganacha malartacha de na téarmaí cuardaigh. Ní gá gurb ionann ord na bhfocal sna torthaí agus ord na bhfocal sna téarmaí cuardaigh sa chuardach seo. Tugtar áiseanna scagtha de réir bailiúcháin agus de réir foirmeacha na bhfocal don úsáideoir. Tá an scagaire foirmeacha focal bunaithe ar oll-liosta péirí leamaí-téacschomharthaí.[13]

I leagan CFG2020 den chomhéadan, tá na feidhmeanna cuardaigh agus scagtha céanna ar fáil is atá ar fáil i gcomhéadan poiblí CGC. Anuas orthu seo, forbraíodh dhá scagaire breise do CFG2020, is é sin scagaire canúintí (le gach canúint nó canúint faoi leith a roghnú) agus scagaire blianta (le bliain faoi leith nó raon blianta a roghnú). Rinneadh optamú suntasach ar an mbunachar chun go mbeadh luas an chuardaigh fós sásúil agus an líon focal ardaithe ó 33.5m i CGC go 77.3 milliún focal i CFG2020. Rinneadh roinnt feabhsuithe i gcur i láthair na meiteashonraí freisin. Déanfar na forbairtí seo ar fad a chur i bhfeidhm ar chomhéadan CGC ar ball, le go mbeidh an dá chomhéadan mar a chéile. Cruthaíodh leathanach breise ar chomhéadan príobháideach CFG2020 ar a gcrochtar na liostaí minicíochtaí agus aon seachtháirgí eile a bhaineann leis an tionscadal seo.

[13] URL: https://github.com/michmech/lemmatization-lists

6. Conclúid agus Pleananna Forbartha

Léiríonn an tionscadal seo gur féidir corpas fiúntach a thiomsú in achar gearr trí leas a bhaint as sonraí corpais ó fhoinsí éagsúla idir shean agus nua. Bhí dúshláin le sárú chun comhleanúnachas a dheimhniú i dtaca le formáidí, córais chlibeála agus cúinsí cóipchirt ach tá an toradh an-sásúil: corpas 80 milliún focal, é clibeáilte de réir ranna cainte agus inchuardaithe ar go leor bealaí. Ní hin le rá, áfach, nach bhféadfaí é a fheabhsú agus creideann na húdair go bhfuil an scóip agus an mhais chriticiúil in CFG2020 anois chun tionscadal náisiúnta corpais don Ghaeilge a bhunú air.

Aithníodh bearnaí áirithe in CFG2020 le linn an tionscadail agus d'aimsigh an fhoireann taighde foinsí a líonfadh na bearnaí sin, agus foinsí a chuirfeadh le méid agus caighdeán an chorpais freisin. Creideann na húdair go bhféadfaí corpas ardchaighdeáin clibeáilte Gaeilge 150-200 milliún focal a sholáthar faoin mbliain 2025, ceann a bheadh sách téagartha agus sách ionadaíoch chun a bheith mar bhunchloch do réimse leathan tionscadal taighde. Chuige sin theastódh infheistíocht fhadtéarmach chun líon na dtéacsanna a mhéadú, chun caighdeán an ábhair a fheabhsú, chun uirlisí nua próiseála agus cuardaigh a fhorbairt agus chun modhanna éagsúla rochtana a chruthú a chuirfeadh an t-ábhar ar fáil don oiread páirtithe leasmhara agus is féidir.

Bhí tionscadal CFG2020 dírithe ar riachtanais na foclóireachta agus fiú dá gcloífí leis an sprioc sin amháin bheadh sé riachtanach an corpas a mhéadú a thuilleadh agus é a choinneáil cothrom le dáta le téacsanna úra chun bearnaí a líonadh agus chun cothromaíocht a chothú idir réimsí agus canúintí, idir ábhar scríofa agus ábhar urlabhra. Bheadh sé inmhianaithe, chomh maith, áiseanna breise cuardaigh corpais a chur le comhéadan Gaois a fhreagródh níos fearr do riachtanais an taighde foclóireachta. Anuas ar an gcuardach simplí agus frásaí atá ann cheana,

theastódh cuardach leamaí, cuardach ranna cainte, agus cuardach casta le carachtair speisialta nó sloinn rialta. Maidir le cur i láthair na dtorthaí, theastódh feidhmeanna cuardaigh de réir na comhchordachta le nascacht inlíne chuig meiteashonraí le samplaí maithe foclóireachta a léiriú.

Creideann na húdair, áfach, nár ghá don tionscadal seo a bheith teoranta don fhoclóireacht amháin agus gur cheart féachaint leis an leas is fearr agus is fairsinge a bhaint as an ábhar. D'fhéadfaí tacair éagsúla sonraí ón gcorpas a phacáistiú mar tháirgí ar leith agus iad a chur ar fáil d'úsáideoirí de réir a gcuid riachtanas — an t-ábhar urlabhra amháin mar shampla. Bheadh na sonraí nó seachtháirgí corpais a thiomsófaí i dtionscadal náisiúnta mar seo ar fáil do thaighdeoirí agus do thionscadail taighde eile amach anseo. Creideann na húdair go spreagfadh sé sin taighde idirdhisciplíneach corpais ar ghnéithe éagsúla den Ghaeilge chomhaimseartha, lena n-áirítear an teangeolaíocht, teagasc na Gaeilge, an ríomhaistriúchán agus teicneolaíochtaí teanga, an téarmeolaíocht agus, ar ndóigh, an fhoclóireacht, rud a chinnteodh go bhfaighfí luach ar airgead as an infheistíocht a bheadh de dhíth chun tionscadal mar seo a bhunú agus a bhuanú. Is fada tionscadal náisiúnta corpais de dhíth ar an nGaeilge agus tá súil ag na húdair gur céim bheag i dtreo na sprice sin é CFG2020.

Nóta buíochais

Táimid buíoch de na heagarthóirí agus na hintéirnigh i ngrúpa taighde Gaois a chabhraigh leis an obair ar fad a rinneadh ar an tionscadal seo, chomh maith le Ronan Doherty a d'oibrigh ar an tionscadal mar chonraitheoir teicniúil. Táimid buíoch de na foilsitheoirí agus sealbhóirí cóipchirt ar fad a sholáthar na téacsanna a chuir ar ár gcumas an obair seo a dhéanamh. Is iad Foras na Gaeilge a mhaoinigh an tionscadal agus bhí comhoibriú dlúth i gceist leis an Eagarthóir

Foclóireachta, an Dr Pádraig Ó Mianáin agus leis an mBainisteoir Tionscadail i Rannóg na Foclóireachta, Cormac Breathnach.

Tagairtí

Biber D. (1988). *Variation across speech and writing*. Cambridge University Press.

Johnson, W. (1944). Studies in language behavior: I. A program of research. In *Psychological Monographs 56*(2): 1–15.

Judge, J., Ní Chasaide, A., Ní Dhubhda, R., Scannell, K. P. & Uí Dhonnchadha, E. (2012). *The Irish language in the digital age: An Ghaeilge sa ré dhigiteach*. https://tinyurl.com/h6hkdpc

Kilgarriff, A., Rundell, M. & Uí Dhonnchadha, E. (2006). Efficient corpus development for lexicography: Building the New Corpus for Ireland. *Language Resources & Evaluation*. 40. 127-152. https://www.tcd.ie/slscs/assets/documents/staff/nci_nlej.pdf

Knight, D., Loizides, F. & Neale, S. (2020a). Developing computational infrastructure for the CorCenCC corpus: The National Corpus of Contemporary Welsh. *Language Resources & Evaluation*. https://doi.org/10.1007/s10579-020-09501-9

Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., & Thomas, E. M. (2020b). The National Corpus of Contemporary Welsh: Project Report | Y Corpws Cenedlaethol Cymraeg Cyfoes: Adroddiad y Prosiect. https://arxiv.org/abs/2010.05542

Nic Pháidín, C. (2008). Corpus planning for Irish: Dictionaries and terminology. In *A new view of the Irish language*. (lgh. 93–107). Cois Life.

Ó Dónaill, N. (1977). *Foclóir Gaeilge-Béarla*. An Gúm.

Ó Meachair, M. J. (2020). *The creation and complexity analysis of a corpus of educational materials in Irish (EduGA)*. [Tráchtas neamhfhoilsithe dochtúireachta]. Coláiste na Tríonóide, Baile Átha Cliath. URL: http://www.tara.tcd.ie/handle/2262/92421

Ó Mianáin, P. (2012). *The New English-Irish Dictionary*. European Federation of National Institutions for Language, Budapest. http://www.efnil.org/documents/conference-publications/budapest-2012/14-EFNIL-Budapest-OMianain-Final.pdf

Royal Irish Academy (2004). *Corpas na Gaeilge 1600-1882: The Irish Language Corpus 1600-1882*. [CD-ROM]. ISBN: 9780954385545.

Scannell, K. (2007). The Crúbadán project: Corpus building for under-resourced languages. Building and exploring web corpora. *Proceedings of the 3rd Web as corpus workshop*. https://go-pdf.online/proceedings-of-the-3rd-web-as-corpus.pdf

Uí Dhonnchadha, E. (2009). *Part-of-speech tagging and partial parsing for Irish using finite-state transducers and constraint grammar*. [Tráchtas neamhfhoilsithe dochtúireachta]. Ollscoil Chathair Bhaile Átha Cliath. URL: http://doras.dcu.ie/2349/

Uí Dhonnchadha, E. & Frenda, A. (2013). *Comhrá: Corpas na Gaeilge labhartha*. [Tacar sonraí]. Coláiste na Tríonóide, Baile Átha Cliath. https://www.scss.tcd.ie/~uidhonne/comhra/index.utf8.html

Unicode Consortium (2020). Common Locale Data Repository (Leagan 38.1.) [Tacar Sonraí]. http://cldr.unicode.org/index/downloads

Wynne, M. (ed.) (2005). *Developing linguistic corpora: A guide to good practice*. ISSN 1463 5194. http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf