*Proceedings of the CLTW 4 @ LREC2022* , pages 99–103
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

# Introducing the National Corpus of Irish Project

**Mícheál J. Ó Meachair, Úna Bhreathnach, Gearóid Ó Cleircín**
Dublin City University
Dublin, Ireland
{micheal.omeachair, una.bhreathnach, gearoid.ocleircin}@dcu.ie

**Abstract**

Abstract This paper introduces the National Corpus of Irish, an initiative to develop a large national corpus of written and spoken contemporary Irish as well as related specialised corpora. The newly-compiled corpora will be hosted at `corpas.ie`, in what will become a hub for corpus-based research on the Irish language. Users will be able to search the corpora and download data generated during the project from the `corpas.ie` website and appropriate third-party repositories. Corpus 1 will be a balanced general-purpose corpus containing c. 155m words. Corpus 2 will be a written corpus consisting of c. 100m words. Corpus 3 will be a spoken corpus containing 6.5m words. Corpus 4 will be a monitor corpus with a target size of 1m words per year from 2000 onwards. Token, lemma, and *n*-gram frequency lists will be published at regular intervals on the project website, and language models will be published there and on other appropriate platforms during the course of the project. This paper focuses on the background and crucial scoping stage of the project, and examines user needs as identified in a survey of potential users.

**Keywords:** corpora, Irish, resource evaluation

## 1. Introduction

This paper introduces the National Corpus of Irish project, an initiative to develop a large national corpus of written and spoken contemporary Irish as well as related specialised corpora. The project is being undertaken by the Gaois research group, Fiontar & Scoil na Gaeilge, DCU, with funding for the period 2022–24 from the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media, with support from the National Lottery.

The corpora will be hosted at `corpas.ie`, in what will become a hub for corpus-based research on the Irish language. The contents of the corpora will be presented in a way that facilitates use by both researchers and non-experts through the provision of simple and more complex searches. Users will be able to search the corpora and download data generated during the project from the `corpas.ie` website and appropriate third-party repositories. Comprehensive documentation pertaining to the data will also be available on the project website.

The following are the projected sizes of the corpora:

- Corpus 1: the National Corpus of Irish (CNG): c. 155 million words;
- Corpus 2: the Corpus of Written Irish: c. 100 million words;
- Corpus 3: the Corpus of Spoken Irish: c.6.5 million words;
- Corpus 4: the Monitor Corpus of Irish: one million words per annum from the year 2000 onwards

Corpus 1 will be a balanced general-purpose corpus that contains a wide and representative sample of Irish from the year 2000 to 2024. Corpus 2 will be a corpus that is focused on a higher register and will likely be of use to translators and terminologists, among other researchers. Corpus 3 will contain spoken data that will be of interest to phoneticians and researchers in the speech sciences, among others. Corpus 4 will include samples of similar sizes from the same domains for each of the included years, and is expected to be suitable for tests on language change as well as having limited general-purpose applications.

Token, lemma, and *n*-gram frequency lists will be published at regular intervals on the project website, and language models will be published there and on other appropriate platforms during the course of the project.

This project has 2 FTE staff, as well as benefiting from the technical and editorial expertise of the Gaois research group; who also developed the National Terminology Database for Irish, the Corpus of Contemporary Irish and the Corpus of Irish for Lexicography among other projects. Experts in software development (Kevin Scannell, Saint Louis University) and spoken corpora (Elaine Uí Dhonnchadha, TCD) are acting as consultants to the project. An Advisory Committee made up from a group of subject experts, drawn mainly from across the university sector, has been appointed to advise on both best practice and emerging research needs from the fields. These include, in alphabetical order, the fields of Computer-assisted Language Learning (CALL), Corpus Linguistics, Irish-language studies and analysis, Language Learning and Education, Lexicography, Linguistics, Natural Language Processing, and Terminology. Members of this committee are credited on the project website.

## 2. Background

Large corpora are one of the core digital resources needed for language technologies the world over. Corpora provide both linguistic knowledge and examples for researchers seeking to create dictionaries, term databases, and other knowledge bases. They are also essential in the creation of language tools, whether these are created using machine-learning techniques or from rules and conventions that have been extracted from said corpora.

The first large and publicly-available corpus of contemporary Irish was compiled in a one-year project in 2006 (Kilgarriff et al., 2006). It is known as *Nua-Chorpas na hÉireann*[1], comprising 30 million words, as well as both document- and word-level annotation. The corpus has not been maintained or supplemented since 2006 and is therefore out of date and unsuitable for research into contemporary language use (Ó Meachair, 2020).

The Gaois research group had hoped to secure funding for a larger corpus project for some time. The groundwork for this began in 2016 with the publication of *Corpas na Gaeilge Comhaimseartha* [2] (CGC). When first published CGC contained 5.3 million words; it now contains over 36 million words. CGC can be searched by members of the public without subscription or registration; 159,000 searches of CGC were recorded in 2021. The compilation of this corpus has been ongoing, but it will likely be subsumed into the National Corpus of Irish project with Corpus 2 taking its place.

The Gaois research group has also published a parallel corpus of Irish-English legal texts that the public can search without subscription or registration[3]. This parallel corpus currently contains 58.5 million words, 28.5 million in Irish and 30 in English. The compilation of this corpus is ongoing and it remains to be seen whether or not the entire corpus will be subsumed into the National Corpus of Irish project.

Subsequent to these projects another corpus compilation project called *Corpas Foclóireachta na Gaeilge*[4] (CFG), or the Corpus of Irish for Lexicography, was conducted by the Gaois research group with funding from Foras na Gaeilge in 2020–21. The aim of the CFG project was to compile a corpus of 100-million words from high-quality sources, tagged for part of speech and lemma, and to make this corpus searchable. It was not within the scope of the project to balance the corpus. Considerable metadata was developed to accompany the CFG corpus, however, this did not include all the metadata required in a national corpus. CFG includes all data from the CGC corpus and an additional 65 million words from a variety of other sources. It is a research corpus that is only available to members of the dictionary team at Foras na Gaeilge and researchers in the Gaois Research Group at present. A special agreement was made during the CFG project to facilitate the creation of by-products that could be used by Irish-language research groups namely, an ARPA language model, 1–5gram frequency lists, and an RNN language model suitable for Irish-language speech recognition. The CFG project's aims and objectives, as well as results from the first half of the project, are detailed in Ó Meachair et al. (2021).

The projects outlined above have contributed data, tools, practices, and experience that will benefit CNG from beginning to end. It is worth noting that numerous other Irish-language corpora have been compiled in the last ten years, but they were compiled for research purposes only and cannot be accessed by the public or registered users. These include (but are not limited to) Ní Ghloinn (2020), Ó Meachair (2020), Uí Dhonnchadha and Frenda (2013), and Scannell (2007).

## 3. Project planning

The CNG project began in January 2022 with a focused scoping and auditing step, to define all core deliverables, particularly workflow, technologies, and data sources.

A survey of future users was carried out as part of this step. This survey was publicised via social media and Gaois websites and was circulated to parties known to be interested. The survey, which received 27 responses, gave the respondents considerable scope to elaborate on their particular requirements. It served to test and elaborate on the research that had been conducted on use cases for Irish-language corpora during the application stage. The fields of research that had been anticipated at the application stage remained the same (e.g. linguistics, education, NLP, lexicography, terminology, and translation), but specific types of corpus searches came to light in survey responses. For example: it had been predicted that CQL searches were a requirement, as well as domain and publication-related metadata filtering. The way in which researchers intend to use the search functions for re-use in longer-term projects was also noted.

> "Deis ag gach úsáideoir fochorpais dá gcuid féin a chruthú, nó ar a laghad, na critéir chuardaigh i gcuardach casta a shábháil go mbeadh sé / sí in ann filleadh ar an 'bhfochorpas' sin go rialta. Bheadh sé seo tábhachtach dá mbeinn ag iarraidh a bheith ag obair i gcorpas iata. M.sh má tá obair ar siúl agam i 'bhfochorpas' thar thréimhse níor mhaith liom go dtarlódh sé go gcuirfí le hinneachar an mhórchorpais agus go gcuirfear le hinneachar m'fhochorpais gan choinnne—rud a chuirfeadh mo chuid staitisticí as riocht."

[1] https://focloir.sketchengine.co.uk/run.cgi/index

[2] https://www.gaois.ie/ga/corpora/monolingual/

[3] https://www.gaois.ie/ga/corpora/parallel/

[4] https://www.gaois.ie/ga/corpora/lexicography/login/

> [Translation: *An opportunity for users to create their own sub-corpus, or at least, to store the criteria of their advanced search so that he / she can return to their "sub-corpus" frequently. This would be important if I wanted to work in a closed corpus. For example: If I were to be working in a "sub-corpus" over an extended period of time I would not like for the contents of the large corpus to change, thus changing the contents of my sub-corpus and distorting my statistics.*]

The need to cater for specific research projects and their annotation needs was also raised, with the following respondent outlining a potential use for a customisable tagging tool:

> "Mura mbeadh sé clúdaithe faoin gclibeálaí séimeantaice, ba mhaith an rud uirlis anótála a bheith ar fáil go bhféadfadh úsáideoir torthaí a chuardaigh a anótáil (téarmaí / frásaí a aibhsiú; naisc idir téarmaí a léiriú), agus torthaí anótáilte a easportáil nó a chóipeáil agus a ghreamú go clár eile i.e. Word, PDF, Excel mar shampla."
>
> [Translation: *If it weren't to be covered by the semantic tagger, it would be good if a tagging tool were available to users so that they could tag the results of their searches (terminology / highlight phrases; display links between terms), and export these annotated results or have the ability to copy and paste them into another programme i.e. Word, PDF, Excel for example.*]

Another noteworthy point of feedback from the survey was the number of translators who reported using existing corpus searches as part of their verification process, and expressed a desire to continue this practice. A number of interested parties also expressed a desire for downloadable corpora or sub-corpora. It has subsequently come to light that sophisticated search functionalities on the project website would be more useful to these interested parties, rather than downloading the data only to use it in another corpus-querying tool. NLP practitioners will continue to want data to be available to them, because they have the computational skills and expertise required to manipulate the data for their specific needs, particularly larger datasets.

An Advisory Committee was assembled from as many of the fields related to the CNG project as possible. This includes experts from the fields of corpus linguistics and linguistics, education and language learning, phonetics, computer programming and natural language processing, lexicography and terminology. It is hoped that this committee will advise the research team throughout the project, advising on best practices and recent developments from their respective fields, as well as briefing the research team as to the specific corpus-based needs of linguists, lexicographers, or educators—for example.

## 4. Workflow and technologies

Workflow practices have largely been established during the previous corpus projects conducted by the Gaois research group. These practices include documentation of all computational processes, as well as the storage of both raw data and the corpus-ready versions of these data. A receipt system for data handovers was established, summarising the number of files being sent or the wordcount of the files sent, in order to ensure forks are avoided and data is not lost. This receipt system is inspired in some respects by the way GitHub[5] manages push requests. Database and file-storage specifications have also been audited to ensure the data is safe from a security point of view, and to ensure formats and encoding are maintained.

While the technology selection process is ongoing, it is clear that it will be more time- and cost-efficient to repurpose existing bona fide technologies than start developing our own technologies from scratch. Numerous examples of suitable existing technologies were gathered by the Gaois research group and presented to the appropriate experts from the committee in a number of meetings. During these meetings the pros and cons of each technology was discussed with a view to identifying which technologies work together best and which technologies, despite appearing suitable at first, were not suitable. For example: a considerable amount of pre-processing is done using Python, so the pros and cons of using Python in the project website as well were discussed. Ultimately, it was discovered that this was not necessary and was potentially limiting. For the purposes of the corpus project the technologies used to develop the website and its search functions simply needed the data to be processed appropriately and consistently, rather than those technologies needing to be integrated with the pre-processing and processing technologies.

The technologies that have been selected are still under development, or are still being adapted to the needs of the project, and may be subject to change. It is therefore too early in the project to provide specifications for them. It was agreed in the funding award that the Gaois research group would package a part-of-speech tagger, such that it would be usable for the present project and by others thereafter, and, if possible, a semantic tagger would also be packaged and made publicly available.

## 5. Data sources

Previously collected data were audited in order to identify gaps or imbalances in our corpora. This was done by calculating token counts for all definable subcorpora of the CFG project for year, genre (for example: news, literature, academic), and publisher. The

[5] `https://github.com/`

most useful results at this stage were found in the token counts for year and genre, with publisher counts being too variable to yield any concrete conclusions. Approximately 55 million of the 100 million words included in CFG were published after the beginning of the year 2000 and/or were not restricted by copyright agreements, therefore qualifying them for re-use in CNG; a practice that was successfully employed in the compilation of CFG (Ó Meachair et al., 2021) and CorCenCC (Knight et al., 2021b) to good effect.

The new potential sources were collated in an Excel spreadsheet along with information regarding the domain or mode of the language (sports, spoken, informative, educational, etc.), and the collection method (scraping of webpages, downloadable PDFs, request and/or collection required, etc). The focus here is on the collection of a wide variety of text types, and examples of language use in different contexts, as is desirable in modern corpora (Sinclair, 2003).

An additional consideration in prospecting for previously unfound Irish-language data was the collation of genres (news, governmental, literary, pop science, etc) and language modes (written, spoken, *e*-language / web data, etc) in order to ensure the corpus accords with best practices for national corpora, as much as is practicably possible in the minority-language context. 51 large corpora, most of which are national corpora, were surveyed for their size, balancing considerations, domains included, sampling methods, and the technologies used to deliver corpus searches—where possible. Notable among the findings were that the majority of the newest larger corpora sought to include as much data as possible (containing upwards of 500 million words). Rather than imposing prescribed balancing methodologies, users were expected to use search filters to create virtual sub-corpora, therefore tailoring their searches to their research aims (Davies, 2018; Kupietz et al., 2010).

The design of these very large corpora are of course different to each other in many ways, but empirical research can be conducted with them using very similar methodologies. Where this approach was not used, a significant number of the other corpora adopted an approach that was informed by BNC 1994 (Burnard, 1995), and then adapted it to fit their own language and/or research needs (Knight et al., 2021a; McEnery et al., 2006; Aksan and Aksan, 2009).

The exact design of Corpas Náisiúnta na Gaeilge will be informed by these approaches and influenced to some degree by the availability of texts. It is necessary in the minority-language context that a certain amount of collection be completed before the corpus design is finalised, because available texts are fewer overall than those written in languages that are more widely spoken and some domains are absent entirely (For example: biology and chemistry publications in Irish, instruction manuals).

## 6. Conclusion

The National Corpus of Irish project is only six months old, and is likely to evolve considerably before its conclusion at the end of 2024. This is an interdisciplinary project with many interdependent elements—newly-developed technologies, data from a variety of sources, legal agreements, and language expertise. Therefore, the planning and scoping stage described in this paper will be crucial to its success and timely completion.

## 7. Bibliographical References


Aksan, Y. and Aksan, M. (2009). Building a national corpus of Turkish: Design and implementation. *Working Papers in Corpus-based Linguistics and Language Education. Tokyo: TUFS*, (3):299–310.

Burnard, L. (1995). *Users Reference Guide for the British National Corpus. version 1.0*. Oxford University Computing Services.

Kilgarriff, A., Rundell, M., and Uí Dhonnchadha, E. (2006). Efficient corpus development for lexicography: building the New Corpus for Ireland. *Language resources and evaluation*, 40(2):127–152.

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

Knight, D., Morris, S., Arman, L., Needs, J., and Rees, M. (2021a). *Building a National Corpus: A Welsh Language Case Study*. Palgrave Macmillan Cham.

Knight, D., Morris, S., and Fitzpatrick, T. (2021b). *Corpus design and construction in minoritised language contexts - Cynllunio a chreu corpws mewn cyd-destunau Ieithoedd lleiafrifoledig: The National Corpus of Contemporary Welsh - Corpws Cenedlaethol Cymraeg Cyfoes*. Palgrave Macmillan Cham.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). The German reference corpus DeReKo: A primordial sample for linguistic research. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge.

Ní Ghloinn, A. (2020). Corpas Foghlaimeora TEG agus an Próifíliú Cumais sa Ghaeilge. In Eoghan Ó Raghallaigh, editor, *Léachtaí Cholm Cille L: Téamaí agus Tionscadail Taighde*, pages 119–156, Maigh Nuad. An Sagart.

Scannell, K. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 5, pages 5–15.

Sinclair, J. M. (2003). Corpora for lexicography. In Piet van Sterkenburg, editor, *A Practical Guide to Lexicography*, pages 167–178. John Benjamins, Amsterdam.

Ó Meachair, M. J., Ó Raghallaigh, B., Bhreathnach, Ú., Ó Cleircín, G., and Scannell, K. (2021). Tiomsú Corpais don Taighde Foclóireachta: Corpas Foclóireachta na Gaeilge (CFG2020). *TEANGA, the Journal of the Irish Association for Applied Linguistics*, 28:278–305.

Ó Meachair, M. J. (2020). *Complexity Analysis of a Corpus of Educational Materials in Irish (EduGA)*. Ph.D. thesis, School of Linguistic, Speech and Communication Sciences, Trinity College, Dublin.

## 8. Language Resource References

Mark Davies. (2018). *iWeb: The Intelligent Web-based Corpus*. `https://www.english-corpora.org/iweb/`.

Uí Dhonnchadha, Elaine and Frenda, Alessio. (2013). *Comhrá: Corpas na Gaeilge Labhartha*. The Centre for Speech and Language Technology for Irish, Coláiste na Tríonóide, School of Linguistic, Speech and Communication Sciences, TCD. `https://www.scss.tcd.ie/~uidhonne/comhra/index.utf8.html`.