Corpas Náisiúnta na Gaeilge 2022-2029: A Project Overview

Ó Meachair, M. J., Bhreathnach, Ú., Ó Raghallaigh, B., Ó Cleircín, G., Měchura, M., Scannell, K.

Dublin City University (DCU), Fiontar & Scoil na Gaeilge, Droim Conrach, D09 N920
{micheal.omeachair, una.bhreathnach, brian.oraghallaigh, gearoid.ocleircin, michal.boleslav.mechura}@dcu.ie, kscanne@cadhan.com.

Abstract

This paper reports the latest developments, planned works, and issues of the Corpas Náisiúnta na Gaeilge (henceforth: CNG, translation: *the National Corpus of Irish*) project, detailing the work that has been completed to date, current work, and planned future work. This report details the compilation of corpora, development of a project website and part-speech tagger, the challenges of expanding existing corpora, and the addition of historical and legal corpora. We also present the training and outreach activities of the project.

Keywords: corpus linguistics, NLP, low-resource language

The CNG project is being administered by the Gaois (www.gaois.ie) research group with funding from the Department of Rural and Community Development and the Gaeltacht and the National Lottery. Gaois is a research group in Fiontar & Scoil na Gaeilge, DCU, comprising lecturers, researchers and postgraduate students. Our aim is to sustain and transform Irish language and culture through the development of innovative and trusted resources. These resources include the National Terminology Database for Irish (www.tearma.ie), the Placenames Database of Ireland (logainm.ie), among others, and since 2024 this also includes the CNG project (www.corpas.ie). CNG built on the Corpus of Irish for Lexicography which was a proof-of-concept and is detailed in Ó Meachair, et al (2021).

1. Project Phases

In this section we introduce the three phases of the CNG project. This section concludes by reporting on a selection of challenges that arose with the compilation and provision of data.

1.1. Phase 1: 2022-2024

Corpus name	Size
Corpas Náisiúnta na Gaeilge, CNG (The National Corpus of Irish)	101 million words
Corpas Monatóireachta na Gaeilge, CMG (The Monitor Corpus of Irish)	1 million words *per annum*
Corpas na Gaeilge Labhartha, CGL (The Corpus of Spoken Irish)	9 million words
Corpas na Gaeilge Scríofa, CGS (The Corpus of Written Irish)	131 million words

Table 1: Corpora compiled during Phase 1

In brief, CNG is a balanced representative corpus of Irish language for the period 2000-2024. We included written data from both online and printed sources (for example: literature, news, academic, blogs), from as many genres and registers as possible. We also included spoken data from radio and television, from speeches and lectures, as well as creative spoken works such as songs and stage dramas.

CMG includes samples from genres that reliably publish in Irish every year: news, novels, legal texts, annual governmental reports and business reports.

CGL includes a variety of data from radio, television, and in-person contexts. Some of the data are transcribed and some are written to be spoken, such as scripts and lectures.

CGS includes written data that has gone through an editorial process, thus excluding blogs and social media posts, as well as some self-published documents. No balancing has been done to reduce the volume of legal texts or the more prolific news agencies.

During Phase 1 the www.corpas.ie website was developed, leveraging NoSketchEngine (Natural Language Processing Centre, 2025; Rychlý, P. 2007) for concordancing purposes, and a part-of-speech tagger was developed that built on UD-Pipe technologies (Straka and Straková, 2017) and used the PAROLE tagset for Irish (Uí Dhonnchadha, 2009).

1.2 Phase 2: 2025

Phase 2 lasted one year (2025) and saw delivery of the following outputs:

Outputs
Addition of 1-million words to CMG, for the year 2025.
Addition of 1-million words to CGL from the period 2000-2025

Generation of updated N-gram and word-frequency lists
Additions, improvements, and corrections to the tagging hitherto completed on all corpora

Table 2: Outputs completed during Phase 2

These outputs were added to the project website in Q1, 2026.

1.3 Phase 3: 2026-2029

Phase 3 commenced in January 2026, and has the following outputs as its stated objectives.

Output	**Metric**
Addition of the Parallel English-Irish corpus of legislation from Ireland and the European Union to the project website.	69.3m in Irish, 63.7m in English
Addition of a further four years, 4 million words, to CMG.	4-million words
Addition of collected data to CGS	Not stated
Addition of collected data to CGL	Not stated
Addition of the *Bailiúchán na Scol* (1937-1939) Collection to the project website.	10-million words of Irish
Improvements to search functionality and website features, and updates to word lists.	Subject to testing

Table 3: Planned outputs from Phase 3

1.4 Compilation and provision challenges

Our corpora depend on licences for the scientific usage of copyrighted data, some of which are long-standing while other new sources are consistently sought out. News agencies and publishing houses are the biggest contributors of data, and it is recognised that sharing their works is not a core part of their workflows. This is not an issue for low-resourced languages alone, as documented in Kupietz and Lungen (2014). This can relate to API access changing in some manner or to fiction and non-fiction publications being produced as stand-alone projects, rather than archived files complete with metadata. Furthermore, not all authors have the same contracts with their publishers and looking into these contracts is very time consuming for publishing houses.

We work hard to explain our scientific uses in plain non-technical language so as not to complicate this process, assuring other parties we cannot and will not re-share any publication in a way that would infringe on Irish copyright law. We do share some derived data via our project website (www.corpas.ie) or by special request - this is done in accordance with legal research in this area and both Irish and EU copyright law (Iacino, G. et. al. (2024), Kamocki, P. (2021), Directive (EU) 2019/790). Word-embeddings and n-gram data can be found on the project website. *Bone fide* researchers can contact the Gaois group to request a shuffled subset of the corpus data. This does not include spoken data (because it can be more personal and recognisable) creative works (novels, poetry, songs, *etc*); a consideration that reflects the respect we have for our creative colleagues and contributors in the dawning age of AI-generated publications. We recognise how the shuffling diminishes the context of the language use, but the legal and ethical considerations supersede this and we remind users context can be linguistically explored via www.corpas.ie. The shuffled text includes roughly one sentence per line and the shuffling is done automatically at least 1000 times (publication forthcoming).

2. Training and outreach

The authors of this paper have also contributed to training and outreach in a number of areas. A selection of which are introduced below.

2.1 Training

The module "Taighde Corpais" (Corpus Research) worth 10 ECTS and taught at NFQ Level 9 has been administered since 2022. This is a mandatory module for students completing an MA in Irish with Fiontar & Scoil na Gaeilge.

Seminars and social-media campaigns have also been run with a view to training the Irish-speaking community to both use the www.corpas.ie website and to engage in corpus-linguistic research.

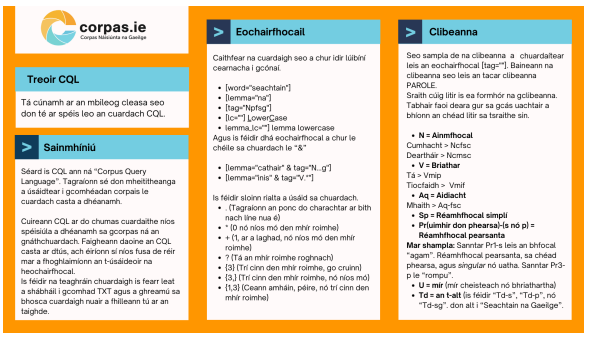

Figure 1. A cheat sheet for using CQL and PAROLE on www.corpas.ie.

2.2 Outreach and conferences hosted

The authors of this paper have been engaged in a number of outreach activities, inter-institutional publications, consultancy and provision of data for Irish-language NLP research, as well as participation in the CLARIN-ERIC network as

ambassador for corpus-linguistic research in Ireland.

2.2.1 Project launch[1] (November 29, 2024)

In November 2024 we hosted a series where guest speakers came to Fiontar & Scoil na Gaeilge, DCU to launch our project. Contextualisation of corpus use, use cases for national corpora in other countries, and technological developments were among the topics presented.

Figure 3. Michal Křen (Charles University, Prague) presenting the Czech National Corpus and its variety of uses

2.2.2 Corpus Linguistics for the Languages of Ireland (Nov 13-14, 2025)

The Gaois research group hosted workshops and a conference for researchers and practitioners working on corpus research for the languages of Ireland.

Workshops titles[2]:

- 'Common Language Resesources and Technology Infrastructure' - Dr. Martin Wynne (Oxford University)
- '*Quo Vadis* Corpas.ie' - Dr. Mícheál J. Ó Meachair & Dr. Michal Měchura.

Conference keynote speaker and contributors[3]:

- 'Corpus Linguistics and Language in Ireland: A promising future?' - Prof. Raymond Hickey.
- 19 researchers presented completed work and ongoing research.

3. Conclusions

We are working to stay true to the principles of the Gaois research group while delivering the CNG project and serving the Irish-language community. With a view to expanding our user base we will add to our training and outreach efforts by conducting and disseminating in-depth linguistic research, by adding specialized corpora and collections that are widely known, and by continuing to host conferences and workshops.

4. Bibliographical References

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC

Iacino, G., Kamocki, P., Du, K., Schöch, C., Witt, A., Genêt, P. and Calvo Tello, J. (2024). Legal status of Derived Text Formats–2nd deliverable of Text+ AG Legal and Ethical Issues –. RuZ - Recht und Zugang. 5. 149-172. 10.5771/2699-1284-2024-3-149.

Uí Dhonnchadha, E. (2009) Part-of-speech tagging and partial parsing for Irish using finite-state transducers and constraint grammar. PhD thesis, Dublin City University..

Kupietz, M. and Lüngen, H. 2014. Recent developments in DeReKo. In Proceedings of the ninth conference on international language resources and evaluation (LREC'14), pages 2378–2385, Reykjavik, Iceland. ELRA.

Kamocki, P. (2021). When Size Matters. Legal Perspective(s) on N-grams. CLARIN Annual Conference. 122-128. 10.3384/ecp18014.

Natural Language Processing Centre (NLP Centre) at the Faculty of Informatics, Masaryk University (2025) "NoSketch Engine" *Available at:* https://nlp.fi.muni.cz/trac/noske. (Accessed 8 March 2026)

Ó Meachair, M. J., Ó Raghallaigh, B., Bhreathnach, Úna, 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.

Rychlý, P. (2007) Manatee/Bonito-A Modular Corpus Manager. In: RASLAN. p. 65-70.

Straka, M., and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of CoNLL 2017.

[1] https://www.gaois.ie/en/blog/seoladh-corpas-ie

[2] Day one blogpost: https://www.gaois.ie/en/blog/an-teangeolaiocht-chorpais-la-1

[3] Day two blogpost: https://www.gaois.ie/en/blog/an-teangeolaiocht-chorpais-la-2