The Design and Development of *Corpas na Gaeilge Comhaimseartha* (Corpus of Contemporary Irish)

Katie Ní Loingsigh, Brian Ó Raghallaigh and Gearóid Ó Cleircín (Dublin City University, Ireland)

1 Introduction

Corpas na Gaeilge Comhaimseartha (Corpus of Contemporary Irish (CCI)) is a monolingual collection of Irish-language texts in digital format which was compiled in Fiontar & Scoil na Gaeilge, Dublin City University, and is available at the following link: www.gaois.ie/g3m/en/. CCI currently contains 14.9 million words. Fiontar & Scoil na Gaeilge is a school in the Faculty of Humanities and Social Sciences in Dublin City University which specialises in interdisciplinary teaching and research through the medium of Irish and has developed the resources Téarma.ie, Logainm.ie, Ainm.ie and Dúchas.ie as well as other projects in language technology and digital humanities (Ó Raghallaigh & Měchura, 2014).

2 Background

There are five major corpora containing Irish-language data available for research and linguistic analysis at present. The primary Irish-language corpus currently available is *Nua-Chorpas na hÉireann* (The New Corpus for Ireland, (NCI)) which contains *c*.30 million words and has been annotated using a morphological analyser and a part-of-speech tagger as developed by Uí Dhonnchadha (2009). The core of this corpus is taken from an 8 million word corpus of Irish that was developed by *Institiúid Teangeolaíochta Éireann* (the Linguistic Institute of Ireland) as part of the EU PAROLE project (Kilgarriff, Rundell & Uí Dhonnchadha, 2006, p.133).

In addition to NCI, the Royal Irish Academy has compiled two Irish-language corpora, *Corpas na Gaeilge*, *1600–1882* (Royal Irish Academy, 2004), which contains 7.2 million words and *Corpas na Gaeilge*, *1882–1926* which contains 7.1 million words. These two historical corpora were compiled as part of lexicographic work undertaken for the Historical Dictionary of Irish project (Foclóir Stairiúil na Gaeilge, 2017). *Corpas na Gaeilge*, *1882–1926* has been annotated and enriched using tools developed by Uí Dhonnchadha (2009) and Scannell (2009) and this work is still in progress (Uí Dhonnchadha et al. 2014). The fourth corpus, *Tobar na Gaedhilge* (http://www.smo.uhi.ac.uk/~oduibhin/tobar/), is freely available and contains a collection of 20th century Gaelic texts, primarily in Irish, but also a limited number of texts in Scots-Gaelic. This corpus contains over 5 million words. A multilingual corpus of over 18 million words, *CELT – Corpus of Electronic Texts* (http://www.ucc.ie/celt/), has been developed in University College Cork and includes texts in Irish, English, Latin and Anglo-Norman French and remains unannotated at present.

The Corpus of Contemporary Irish is the most recently developed Irish-language corpus and is currently unannotated. There are two main searches available to the user – a specific search ('This phrase as is') and a broad search, which provides approximate and related string matching. Results can be filtered according to collection.

3 Methodology and data collection

The primary aims of CCI is to provide a freely accessible linguistic resource which reflects current usage of the Irish language. The following design criteria were followed in the selection of material for inclusion in CCI[1]:

- Eligible texts must be available digitally;
- Eligible texts must have been edited, i.e. material from blogs, social media etc. is ineligible;
- Prose texts (fiction and non-fiction) written and published in Irish from the year 2000 onwards are eligible;
- Translated material is ineligible.

Once texts are selected, the following data is cleaned before import:

- footnotes, glosses, long strings of text in other languages;
- bullet points and internal marks such as †, *, ◇, etc.;
- tables and figures;
- bibliographies (internal references in the body of the text are included).

At the outset, an initial list of Irish-language publishers and other copyright-holders was compiled and a brief outline of the project detailing its nature was given to each individual or organisation. Permission was sought to obtain edited texts in digital format and to use the electronic copies of the texts as part of the CCI project. As a result of these interactions, permission along with the necessary texts were received in various formats, i.e. primarily in PDF, Quark and Word format. CCI segments and related metadata are stored in UTF-8 encoding in a relational database and delivered through a web-based search interface. Although source CCI texts are stored on disk in XML format, they are not freely available to download in any standard XML format due to copyright restrictions. The collection and preparation of the various corpus sources will be examined in the following section of this paper.

The first major stage of development involved the ingestion of texts obtained from the published archive of *Cois Life*[2], one of the foremost Irish-language publishers, along with the ingestion of online material from *Tuairisc.ie*[3], *Beo!*[4] and *Nuacht RTÉ*[5]. This collection amounted to an preliminary corpus of 5.3 million words. The online material was automatically imported into CCI. The *Cois Life* archive was available in PDF format and was imported semi-automatically. This pilot version of CCI was made publically available in April 2016 while further material for inclusion in the corpus was being collected and formatted. The relevant metadata in relation to each of the collected texts were recorded in the corpus database as follows:

[1] While every effort was made to stringently follow the above design criteria, there are still minor instances of strings of text in other languages, internal marks, etc. occurring in the corpus.
[2] https://www.coislife.ie/
[3] http://tuairisc.ie/
[4] http://www.beo.ie/
[5] http://www.rte.ie/news/nuacht/

Type of text	Metadata recorded
Page in published book	• Title; • Author; • Publisher; • Year of publication; • Page number.
Published articles from journals, newspapers, etc.	• Journal title; • Journal volume; • Journal number; • Year; • Month; • Article title; • Article subtitle (where applicable); • Author; • Page number(s).
Online material	• Title; • Subtitle (where applicable); • Author; • Publication date; • URL.

Table 1: Metadata recorded

The second major stage of development involved a significant expansion of the corpus in size and scope. New material was received from a number of significant copyright-holders and was manually analysed and prepared for inclusion in CCI. The current contents of the corpus are as follows:

Type[6]	Word count
Books	3,447,051
Print articles	5,517,040
Online articles	6,017,567
Total	14,981,658

In conjunction with the collection and preparation of new material, the existing contents of the corpus were re-examined and analysed. A number of these texts in their entirety along with specific parts of other texts were judged to fall outside the design criteria and were cleaned from the corpus, e.g. Irish-language quotations from

[6] Genre tagging of the texts in the corpus is currently ongoing. Once completed, this will provide a more detailed description of the various text types in CCI.

texts published during the 20th century or earlier, long strings of text in other languages, religious prayers, collections of proverbs, collections of poetry and grammatical texts among others. This manual re-examination served as a quality assessment of texts.

The manual conversion and preparation of texts proved to be the most labour-intensive aspect of the project. Additionally, the conversion of texts from various formats to .txt format created challenges as often the source text did not transfer exactly to .txt format. For example, the transfer of typesetting hyphens was a common issue in a significant number of texts converted to .txt format from PDF and Quark source files, e.g. *tábh-acht*, *athbheo-chan, beag-nach*. This occurred primarily in instances where restrictive columns in the source texts included hyphens to facilitate the design and layout of the text. These hyphens were manually deleted in these instances.

In addition to the inclusion of unnecessary hyphens, some non-printing control characters and also additional spaces between letters, words and paragraphs were transferred during the formatting of texts. Specific instances of letters being transferred incorrectly were also produced in certain texts, e.g. *chú**Lg**hogarnaíl* which is correctly spelled as *chúl**ch**ogarnaíl*. While every effort was made to ensure these minor inconsistencies did not transfer to the corpus, there are still, inevitably, a small number of examples of unnecessary hyphens, etc. present in CCI.

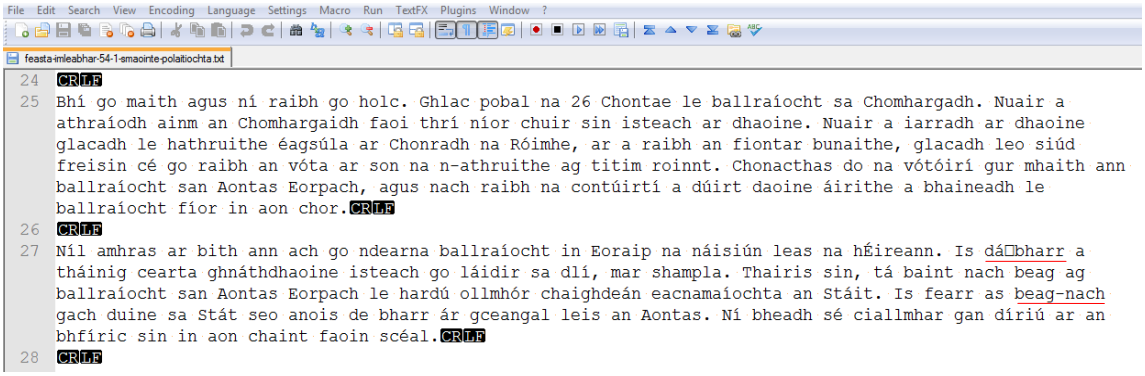

Figure 1: Example of additional unnecessary markings

4 Conclusion and future research

This paper presents the initial phases in the compilation of the Corpus of Contemporary Irish. Apart from material still being processed, all of the Irish-language material received from publishers, individuals and organizations is now available online and further material will be added as it becomes available. It is intended to use CCI in conjunction with a corpus query tool to undertake linguistic research and analysis of grammatical and collocational patterns. With the aid of corpus query tools, the data will form a valuable resource for linguistic research on contemporary written Irish in areas such as stylistics, terminology, lexicography, phraseology, discourse analysis, etc.

References

About CELT: The online resource for Irish history, literature and politics. (2016, December 8). Retrieved April 21, 2017, from http://www.ucc.ie/celt/about.html.

Foclóir Stairiúil na Gaeilge. (2017, April 20). Retrieved April 21, 2017, from https://www.ria.ie/research-projects/focloir-na-nua-ghaeilge.

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

Ó Duibhín, C. (2016, August 30). Tobar na Gaedhilge. Retrieved April 21, 2017, from http://www.smo.uhi.ac.uk/~oduibhin/tobar/.

Ó Raghallaigh, B., & Měchura, M. B. (2014). Developing high-end reusable tools and resources for Irish-language terminology, lexicography, onomastics (toponymy), folkloristics, and more, using modern web and database technologies. Proceedings of the First Celtic Language Technology Workshop, 66-70. doi:10.3115/v1/w14-4610.

Royal Irish Academy. (2004). *Corpas na Gaeilge 1600-1882: foclóir na Nua-Ghaeilge*. Baile Átha Cliath: Acadamh Ríoga na hÉireann.

Scannell, K. (2009). Standardization of corpus texts for the NEID. [PowerPoint slides]. Retrieved April 21, 2017, from http://borel.slu.edu/pub/naaclt09.pdf.

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.

Uí Dhonnchadha, E., Scannell, K., Ó hUiginn, R., Ní Mhearraí, É., Nic Mhaoláin, M., Ó Raghallaigh, B., Toner, G., Mac Mathúna, S., D'Auria, D., Ní Ghallchobhair, E. and O'Leary, N. (2014). Corpas na Gaeilge (1882–1926): integrating historical and modern Irish texts. In *Proceedings of the workshop Language resources and technologies for processing and linking historical documents and archives-Deploying Linked Open Data in Cultural Heritage, LRT4HDA, LREC 2014*, Reykjavík, Iceland. 12-18. Retrieved April 21, 2017, from http://www.lrec-conf.org/proceedings/lrec2014/index.html.