An Corpas Cliste: creating a learner corpus for Irish from a new, purpose-built iCALL platform

Neasa Ní Chiaráin[1]

Abstract. *An Corpas Cliste* (‘*Clever Corpus’*) is an Irish language learner corpus. The corpus data comes from a purpose-built intelligent Computer Assisted Language Learning (iCALL) platform called *An Scéalaí* (‘*the Storyteller*’) and comprises both audio and text, produced by second and third level learners of Irish. Metadata (e.g. L1, level of Irish, dialect preference, age) is saved with every learner account, along with data on platform engagement (e.g. speech/language technologies employed, time spent on task). This paper illustrates how *An Corpas Cliste* is structured and is being prepared for analysis and the methodologies and resources that are being used to exploit it with a view to enhancing the learning experience.

Keywords: learner corpus, iCALL, Irish, speech technologies.

1. Introduction[2]

An Corpas Cliste (*‘Clever Corpus’*) is a learner corpus that is being collected within the framework of an Irish (Gaeilge) iCALL platform, *An Scéalaí* (‘*the Storyteller*’). Two types of learner production data are being harvested within *An Scéalaí* – audio and text. The current discussion centres on learners’ written language, which currently consists of 44,093 learner stories – a total of 5,732,397 words at an average of 130 words per story (July 2022). Analysis of the spoken language will follow at a future date.

1. Trinity College Dublin, Dublin, Ireland; neasa.nichiarain@tcd.ie; https://orcid.org/0000-0002-4669-5667
2. Funding from An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta and the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media is gratefully acknowledged.

How to cite this article: Ní Chiaráin, N. (2022). An Corpas Cliste: creating a learner corpus for Irish from a new, purpose-built iCALL platform. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Friðriksdóttir, H. Garðarsdóttir, S. Thouësny, & M. J. Whelpton (Eds), *Intelligent CALL, granular systems, and learner data: short papers from EUROCALL 2022* (pp. 297-301). Research-publishing.net. https://doi.org/10.14705/rpnet.2022.61.1474

An Corpas Cliste represents a novel approach to the design of language learning technologies for Irish. The learner corpus being collected is intended (1) to serve as a sound empirical basis for linguistic research into the process and stages of Irish language acquisition (e.g. investigating the most common errors made by learners at various stages of proficiency), and (2) to enhance the *An Scéalaí* iCALL platform itself, making it more effective for future users. Analytic results from *An Corpas Cliste* are intended to be the driver of the personalised corrective feedback, which is at the heart of *An Scéalaí*.

To appreciate the nature of the corpus and the ways in which it will be deployed, the background of *An Scéalaí* is explained.

2. Background: *An Scéalaí*

The *An Scéalaí* iCALL platform is freely available as a desktop web application (see https://abair.ie/scealai). It has been designed from the ground up and is aimed at the individual language learner; 4,896 account registrations are logged, with circa 2,736 'active' users. It is part of the *ABAIR* initiative, which is developing speech technology for Irish (particularly Text-To-Speech (TTS) and Automatic Speech Recognition (ASR) systems – see https://abair.ie).

In the platform, learners write content, such as reflective diaries, stories, and essays, and are given feedback via (1) the latest versions of the ABAIR technologies, (2) grammar checker/dictionaries, and (3) teacher feedback (see Ní Chiaráin et al., 2022). The platform provides exposure to native speaker models of the language (critical in the Irish context), allowing proof listening and correction by choosing a preferred voice from three main (very different) dialects. Another key technology being employed, which learners report as extremely useful, is the grammar checker *An Gramadóir* (Scannell, 2013). Dictionaries (see teanglann.ie) are integrated and, when used in a class setting, learners can obtain feedback from their teacher via written/voice notes. Learners can record themselves and compare recording to native speaker (synthetic) models. By integrating the technologies, the multiple skills of writing, reading, listening, and speaking are promoted in tandem.

A major part of the rationale for building *An Scéalaí* was to have our own configurable tool to conduct Second Language Acquisition (SLA) research on Irish, to enable the harvesting, analysis, and exploitation of the types of learner data the platform is designed to yield.

3. *An Corpas Cliste*

Rationale: our aim is to use *An Corpas Cliste* as a mechanism for tracking how Irish language learning happens over time. By using the corpus to investigate learning gains over a long-term at both macro and micro levels (i.e. both the broad collection of language learners and individual language learning journeys), we hope not only to feed this information back into the iCALL platform to benefit our learners but also to inform curriculum development on a wider scale.

To achieve this, we are capturing an ordered series of events by language learners. In the current setup, learners' stories are saved after each edit made (if the user stops typing for 1.5 seconds, their story is saved). This yields large amounts of very detailed data. This data granularity presents challenges from an analytic point of view but will be extremely useful, e.g. to develop pedagogical strategies to deliver appropriate content to learners. By also saving learner engagement we gain insights into learners' usage of individual speech and language technologies (TTS/ grammar checkers/ASR systems).

The corpus data will in future lend itself well to machine learning methods, as we identify patterns in learning behaviour both at individual and group level. However, as illustrated below, extracting meaningful information from the corpus is crucial and we are not yet at the stage that we can depend on automatic methods to give reliable information.

4. Content management

The open-source NoSQL database management system MongoDB is being used to manage the *An Corpas Cliste* content. It is being used not only to store learners' final stories and accompanying feedback (both text and audio), but also to store learners' story composition journeys, including running TTS, listening back, performing grammar checks, receiving teacher feedback, and dictionary lookups. Timestamps are saved with each engagement as well as metadata provided by the learners.

To date, a considerable amount of effort has been devoted to the processing of data. Stored in several collections of a MongoDB database, the data has been processed with Python to produce several categorical DataFrames. These DataFrames serve as an attempt to give researchers a full overview of the learner corpus. The current structure allows us to explore the data and answer some initial research questions.

Data processing is an ongoing exercise. Restructuring will be an iterative process, depending on the data needed to be extracted (e.g. specific age groups/time frames/ levels of Irish) to give insights on learner engagement/language acquisition processes. Ongoing collection of data is being used to calculate relevant statistics, (e.g. numbers of learners/stories), as well as measures of SLA progress, (e.g. changes in grammar/writing errors over time). These statistics, along with the regular manual analyses conducted by linguistic researchers serve as a valuable resource for creating learner models in the future, a major goal of the current initiative.

5. Corpus analysis – preliminary findings

A preliminary analysis was conducted on the entire corpus to investigate the most common errors being made by learners. Two different approaches were employed – an automatic analysis using *An Gramadóir* and a manual one, where a linguist identified the most frequent types of errors by examining the written text.

The results of the first (automatic analysis) approach yielded the following as the 'Top 5' error categories, (terms as encoded in the grammar checking software):

- incorrect;
- unknown word;
- non-standard form;
- lenition missing; and
- possibly a foreign word.

Unfortunately, the description of the errors detected is often not helpful to the learner – the most flagged error is 'incorrect', which is an umbrella for many different items where the spelling deviates from the expected. These reflect a variety of grammatical errors, as well as simple typos, etc.

The result of the second (manual analysis) approach indicates that the 'Top 5' error categories involve issues with:

- genitive case;
- the lenition associated with the possessive adjective 'a';
- relative clause construction;
- missing acute accents (indicate long vowels); and

- misspellings associated with broad/slender vowel agreement (pronunciation-related).

Although the manual analysis is preliminary and not yet the result of a rigorous scientific analysis, it is nonetheless clear that the two approaches yield completely different results from one another (note: a thorough error analysis of a corpus subset (n=172 learners) is ongoing).

The take home point is that the grammar checker, as currently configured, is not fit for the present purpose. It simply does not differentiate adequately among different kinds of errors that show up in learners' written forms. In this way it does not align with what the teacher would want to focus on in terms of individual grammatical learning targets. Furthermore, the *An Gramadóir* output, as with most generic grammar checkers, is not in any case designed for the learner. The type of grammar checker needed for an L2 learner would ideally be designed to provide the type of personalised feedback that is user-friendly for learners at different stages.

6. Conclusions

Current developments for *An Corpas Cliste* include the building of independent grammar checking modules for specific aspects of the language. Analysis of the spoken language corpora will require a different set of analytic skills. For all these efforts, the training of language experts is crucial.

As mentioned above, we aim to deploy state-of-the-art machine learning approaches to deliver personalised corrective feedback via *An Scéalaí*. However, given the frequent assumption that machine learning approaches can of themselves do everything, we advise caution as we are reminded that the 'i' in iCALL will continuously demand human linguistic and pedagogical knowledge in the loop.

References

Ní Chiaráin, N., Nolan, O., Comtois, M., Gunning, N., Berthelsen, H., & Ní Chasaide, A. (2022). Using speech and NLP resources to build an iCALL platform for a minority language: the story of *An Scéalaí*, the Irish experience to date. *Proceedings of ComputEL-5, ACL 2022* (pp. 109-118).

Scannell, K. (2013). *An Gramadóir: an open-source grammar checking engine, Version 0.70*. https://cadhan.com/gramadoir/

Published by Research-publishing.net, a not-for-profit association
Contact: info@research-publishing.net

Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022
Edited by Birna Arnbjörnsdóttir, Branislav Bédi, Linda Bradley, Kolbrún Friðriksdóttir, Hólmfríður Garðarsdóttir, Sylvie Thouësny, and Matthew James Whelpton

Publication date: 2022/12/12

Typeset by Research-publishing.net
Cover photo by © 2022 Kristinn Ingvarsson (photo is taken inside Veröld – House of Vigdís)
Cover layout by © 2022 Raphaël Savina (raphael@savina.net)

ISBN13: 978-2-38372-015-7 (PDF, colour)

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2022.