Seimineár Uí Bhuachalla
Ollscoil Notre Dame
12 Aibreán 2025

I'm very happy to be back at Notre Dame, the center of Irish language studies in the United States.

The final speaker at last night's event talked about normalizing Potawatomi in all aspects of their lives. It struck me that this is a good summary of the work I've been doing for more than 25 years now — specifically, normalizing the Irish language on the computer and in online spaces. This means both its presence as an interface language (menus, buttons, etc.) and its use as a language for communicating online, in social media, blogs, etc.

What I don't normally share publicly is that this work has kept me in a constant state of anxiety for 25 years. We're all aware that the computing landscape changes incredibly quickly. The group I work with has invested thousands of hours of work into software sometimes used by less than 100 people. Our successes grow obsolete faster than I'd like to admit. But mostly the anxiety comes from struggling to figure out what's coming next, and how to not let the language fall further behind with every new technological advance.

Andy Caomhánach was kind enough to give me a shout-out last night at the event, but it's only fair that I point out almost all of the translation work I've done has been in collaboration with an all-volunteer team of 10-12 software localizers, including people like Brian Ó Broin who is here with us today, as well as pioneers like Caoimhín Ó Donnaíle, and the late Marion Gunn.

Andy also mentioned that open source software has been a winning strategy for the Irish language, which is certainly true. We began our localization work with projects like Mozilla Firefox/Thunderbird and OpenOffice not out of ideology, but because we had very few other options. With open source, we didn't need to seek permission from a corporation to translate — we could just do the work, build the software, and distribute it. In the early 00's, we sent a lot of emails to contacts in the big tech companies offering our (volunteer!) localization services but received few answers, much like the Potawotami's overtures to Duolingo and Rosetta Stone. There were a few significant wins — GMail, Twitter/X, Bluesky, WhatsApp are all now available in Irish thanks to our team. But the landscape has changed quickly and radically. Nowadays, most people would rather use applications in the cloud, e.g. GMail vs. Mozilla Thunderbird, or Google Docs vs. OpenOffice. I'll return to this point later, but the key point is that we need to go where our users are. That a piece of software is available in Irish isn't enough to convince most people to switch to it. This puts us back in the position of needing to "ask permission" from the Microsofts, Googles, Apples, and Facebooks of the world.

But I'm really here to talk about AI. AI is changing everything already, and will continue to do so, for better or worse. Those of us who are teachers have been on the front lines of this.

There have been some mentions today of machine translation, automatic transcription software, OCR, etc. You might not think of these as “AI”, but in fact the same basic neural network technologies that underlie AI chatbots like ChatGPT are also the reason why Google Translate has gotten so much better in recent years, and why voice recognition tech has improved for many languages as well.

For the non-experts, all that you need to know is that these approaches are purely data-driven, and involve no language-specific expertise. Google Translate now supports 249 languages. In case it’s not obvious, the Google Translate team does **not** have (or need) speakers of all of these languages (and specifically, there are no Irish speakers on the team, as far as I'm aware).

What this means is that Google, OpenAI, Meta, and friends need “our” data to succeed, as a proxy for actually hiring members of the language community. This has led some indigenous language communities, especially the Māori, to the notion of “data sovereignty”. That is, the view that language data is a precious commodity belonging to the community, and that it should not simply be given away to big tech.

Maintaining data sovereignty requires that the community itself be able to develop the necessary technological resources. This presents a major challenge in many indigenous communities. Fortunately, in the case of Irish, we have no shortage of skilled practitioners:

- The speech technology group at TCD that has produced abair.ie and related tools for processing spoken Irish
- Researchers at the Adapt Centre at DCU who have developed machine translation engines that can outperform Google Translate
- The Gaois group at Fiontar DCU who recently developed Corpas Náisiúnta na Gaeilge, which gathers together a representative sample of Irish written since the year 2000.
- My own work on LLM-based proofing tools (spelling and grammar checkers)
- Andy Caomhánach and collaborators at Acadamh na hOllscolaíochta Gaeilge, who have produced speech-to-text technology that we will hear about later today

Part of this success has been driven by a democratization of AI tools, including open source “foundation models” that anyone can use or fine-tune for their own applications without having to go through the slow (and costly) process of training from scratch, as well as user-friendly open source software for training, testing, and deploying AI models.

The examples above show that the Irish language community can do better than big tech in a number of ways:

- We win by having actual experts in the language, who understand the linguistic structure, fine points of grammar, and dialect variation that are critical to creating high-quality resources
- We win by having better language data — by curating high-quality datasets that reflect Irish as a living language. The standard datasets used by big tech to train models are

gathered by crawling more or less everything on the web, and therefore include significant amounts of Irish language text. On the surface, that's good news. Unfortunately, much of what is included in these datasets is spammy, Google-Translated Irish content, which then gets fed back into the next generation of models.

- We win by building tools that Irish speakers actually want and need, vs. things that no one in the community wants or needs (and which potentially do harm — how many times have you seen signs that are clearly Google-Translated gibberish?)
- Finally, we win by doing this work in a way that respects the Irish speakers who produce the training data that powers these models, The Gaois group took great care in securing permission from contributors to the corpus, in stark contrast to Google, Meta, etc. who gather everything they can, no matter the quality or copyright status. We know now that in the case of Meta this includes pirated copies of millions of books and academic papers — including work written by almost everyone in this room.

This all sounds promising, and indeed the groups I mentioned above have a track record of producing high-quality resources for the language. But we are left with one major challenge that keeps me up at night — namely, "platforming" these resources and getting them in the hands of users where they already are. It doesn't matter if the Adapt Centre can produce a better translation engine if everyone simply turns to Google Translate out of convenience, or because it's embedded in the Chrome browser. And it doesn't matter if the Acadamh or TCD produce high-quality speech-to-text engines if they're not available on my phone — I want to be able to ask Siri "cá bhfuil an bia Iodálach is fearr in South Bend?" and get an answer back in fluent Conamara Irish!

So there is no escaping some form of collaboration with the big tech companies. Fortunately, many of them have a strong presence in Ireland, and I believe the Irish government can play a role in encouraging (or indeed requiring) these collaborations in support of the first national language.