

# The language communities as active partners in technology provisions: the Irish ABAIR experience

Ailbhe Ní Chasaide<sup>1</sup>, Neasa Ní Chiaráin<sup>1</sup>, Harald Berthelsen<sup>1</sup>, Andrew Murphy<sup>1</sup>, Liam Lonergan<sup>1</sup>,  
John Sloan<sup>1</sup>, Christoph Wendler<sup>1</sup>, Connor McCabe<sup>1</sup>, Emily Barnes<sup>2</sup>, Christer Gobl<sup>1</sup>

<sup>1</sup>Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences,  
Trinity College Dublin, Ireland

<sup>2</sup>School of Education, Trinity College Dublin, Ireland

anichsid@tcd.ie, nichiar@tcd.ie cegobl@tcd.ie

## Abstract

The impact of speech and language technologies for the endangered languages depends on the extent to which the language community engage with process. In this paper, the range of speech technologies and applications developed for Irish in the ABAIR initiative in Trinity College Dublin is presented, along with reflections on the many ways in which the language community has come to play an increasingly central role. Community involvement and buy-in is essential for all aspects – not only for the development of core technologies, such as TTS and ASR systems, but to prompt development priorities and to collaborate in the provision of the most important and urgently needed applications. In order for technology to achieve its potential for the endangered language, developers need a knowledge of the language and an understanding of the sociolinguistic context in which the technologies will be used.

**Index Terms:** language community, partnership, endangered language, Irish, TTS, ASR, education, disability

## 1. Introduction

For an endangered language, the nature of the partnership between speech technology developers and the language community sets the bedrock upon which technologies can be built that are appropriate for the community and that can truly impact the language's survival and maintenance. This paper outlines from this perspective the linguistic and speech resources being developed for Irish in the ABAIR initiative at Trinity College Dublin, discussing how the interaction with the language community has evolved with ABAIR's evolution and how it is now shaping research and development – as an active partner in the enterprise, rather than as a passive recipient of technologies.

## 2. The language and the language community

Irish, a member of the Q-Celtic branch of the Celtic languages, is closely related to Scottish Gaelic and to Manx (now extinct), and more distantly related the P-Celtic languages, Welsh, Breton and Cornish (also extinct). The language declined over centuries of colonial rule, with a precipitous acceleration following the great famine of the mid nineteenth century, and the extensive immigration that followed. Today, it is spoken as a community language in Gaeltacht regions, mostly in remote parts of the western seaboard, illustrated in Figure 1, and it has

been classified by UNESCO as “definitely endangered” [1]. Although the repository of the rich cultural heritage of song, poetry and story, these Gaeltacht communities are fragile: it is estimated that only 24% (23,175 people) speak Irish on a daily basis outside of the education system [2]. As the language receded, the Irish speaking pockets became more isolated from each other, reducing the critical density of speakers in any one area and contributing to the evolution of three rather different dialect groups of Ulster, Connacht and Munster, diverging considerably in pronunciation, lexicon, and grammar. As with many endangered languages, the legacy of history results in there being no standard spoken dialect: each is held as a gold standard.

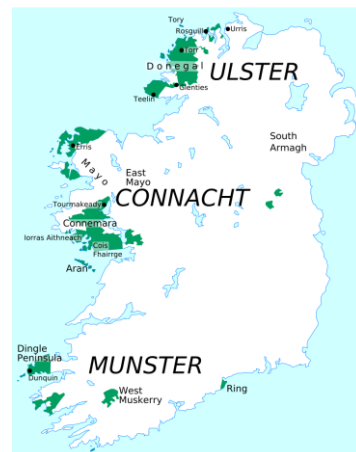


Figure 1: *The official Gaeltacht (Irish speaking) regions.*

Irish is recognised as the first national language in Ireland, and as an official EU language since 2007. It is a compulsory subject in primary and secondary school and there is a growing demand for Irish-medium education. There are pockets of Irish-speaking families outside the Gaeltacht, especially in Dublin and Belfast, with small but vibrant communities of speakers and learners of Irish across Ireland and also among Irish diaspora across the globe. Effective language teaching is seen as essential to the transmission of the language, both in the Gaeltacht and beyond. There are however many challenges: learners typically lack access to native speaker models of the language, to the detriment of L2 acquisition of native-like pronunciation. Teachers lack resources, and often, confidence in their own proficiency. Those with disabilities, are largely

excluded from Irish language classes and Irish medium education. Nonetheless, despite its geographical scattered nature and the many challenges, the different sectors of the language community are a potential source of strength. ABAIR aspires to playing an enabling role, connecting communities and empowering the native speaker communities.

### 3. ABAIR's development of speech and linguistic resources

The ABAIR initiative has evolved over many years from fledgling projects and now involves a fairly broad canvas of research and development of linguistic and speech resources for Irish. The technologies developed are viewed as belonging to the community, and once developed are made available to the public on the website [www.abair.ie](http://www.abair.ie). The site also provides an overview of the work-in-progress, and it includes:

- (i) core technologies, TTS and ASR systems and underpinning linguistic resources;
- (ii) applications which exploit both these speech technologies and our linguistic resources and knowledge of the language. The applications are targeted at three overlapping cohorts: firstly, the general public; secondly, learners and teachers of Irish; thirdly, those with disabilities.
- (iii) a research page, to share the laboratory's basic and applied research.
- (iv) a mechanism for connecting to the language communities, facilitating their participation in many aspects of the research and development.

An accompanying SIGUL submission demonstrates the website and the technologies that can be accessed on it, and some specific technologies/applications are presented more fully in further submissions to this workshop.

A broader outline of the different aspects of ABAIR's activity, and the community's growing role in it are discussed in the following sections.

### 4. Core Technologies appropriate to the community

At the outset, the goal was to develop TTS for Irish, along with the linguistic resources that it required. More recently ASR has also been developed.

#### 4.1. Text-to-Speech

Building a TTS system may seem a straightforward matter of: find a voice talent, record X hours, build a system. However building TTS for Irish is not straightforward, and neither is it likely to be for many endangered languages. There is no standard variety, and imposing one dialect on the others would be unacceptable. Connemara speakers would be reluctant to use a Donegal (Ulster) synthetic voice and might have considerable difficulty understanding it. A non-native speaker voice is also inappropriate for native speaker communities.

A multidialect facility was thus envisioned from the outset. The initial voice developed was for the Donegal dialect, with subsequent extension to the dialects of Connacht and Munster (Figure 1). To facilitate this, the linguistic components of the system (e.g., letter-to-sound rules) were developed in two parts: 'global' modules that capture features common to the dialects, and 'dialect-specific' modules to enable adaptation to new dialects. Separate corpora were designed for the recordings of

the different dialects, based on dialect-specific materials. The earliest recordings were carried out at Trinity College, using voice talent sourced in Dublin. This proved too limiting, and voice talent scouting and the recording shifted to Gaeltacht locations, with recording equipment set up in local centres or even in a speaker's home. Recordings for TTS in a further dialect are currently ongoing in Ring, Waterford (Figure 1) with the help of the community.

This approach contrasts sharply with the practice of 'Big Tech' and commercial companies: Irish synthetic voices have been recently developed, based on L2 speech, with no provision for or awareness of the native-speaker community. What might appear to be a positive step can, in our opinion result in unintended negative outcomes. To give one example: one of the major advantages of having the ABAIR voices easily accessible in educational applications is that it brings the native speaker speech into the classroom and the into the learner's home – alleviating the longstanding difficulty of learners having little access to native speaker models of the language. It would in our view be unfortunate if the provision of L2 voices were to become the de-facto model for language learners. An L2 voice in other technology applications sends subliminal messages also to the native speaker communities, which can further erode the delicate ecology of the endangered language.

#### 4.2. Speech recognition

Building speech recognition has been very much a community effort. As with TTS, the diversity of dialects is an important consideration. A system that fails (more often) with one or other dialect would be poorly received. For ABAIR's initial ASR system ÉIST [3], tracking the performance of the system across dialects was closely monitored. Even when corpora are balanced across dialects, the performance may be skewed – demonstrating a need for 'positive action to ensure equitable cross-dialect performance [4, 5]. It is envisaged that the ASR architecture for Irish will entail a dialect recognition component to optimise performance and cater for cross dialect differences.

Corpus collection for ASR relies on large-scale community buy in. Here, in addition to dialect-specific corpora, L2 corpora are essential, as many envisaged applications (see below) are directed at L2 speakers and learners. Given the scale and diversity of corpora needed, an online recording tool *MíleGlór* (A Thousand Voices) was developed for both online and in-the-field recording. It is dialect-sensitive: information on the speakers dialect, age, L1/L2 status etc. is elicited, and the materials presented for recording are tailored accordingly.

To reach the native-speaker communities (who are the least likely to be online) the ABAIR team has increasingly presented its work at public events, such as *Oireachtas na Samhna*, a major, yearly Irish-language gathering. This allows considerable interaction and discussion with the wider language community, an opportunity to explain why the *MíleGlór* recordings are needed and the opportunity for extensive in-person recordings of native speakers.

The response has been phenomenal, and this interaction has yielded a considerable body of speech data for the different dialects and it has helped create a greater awareness of ABAIR's resources. People relate particularly to the need for applications for disability and access, as well as for educational resources and supports. Those offering their voices for recording are encouraged to continue recording online with *MíleGlór* at home, and to get family, friends and neighbours to contribute

short recordings also. The public buy-in is clear from the steadily growing traffic on the ABAIR website.

## 5. Applications with active participation of user groups

The impact of these technologies for the community derives mainly from the applications that deploy them. From the very first attempt to build a TTS system, requests came for applications.

### 5.1. Applications for disability and access

The first approaches concerned children with disabilities. A screen-reading facility was urgently sought by parents, grandparents of children with visual impairment attending Gaeltacht and Irish-medium schools. Visits with their children to the Lab brought home the severe disadvantages they were under and the threat to their continued participation in school. Approaches by visually impaired adults also highlighted the extent of their exclusion from the language (there being only 3 Irish Braille books available in Irish) and led to a skilled blind programmer joining the team collaborate in the development of a screen-reading facility (an Irish language plugin for the open source NVDA screen-reader). As a blind researcher, he collaborated on system design, advised on the features needed from a user's perspective. He also tested the system with visually impaired schoolchildren by networking with teachers of the blind [6]. He also further developed the screen reader system to work with a Liblouis Braille reader, allowing simultaneous speech and Braille output. He has subsequently continued on a voluntary basis to maintain and update the screen-reading facility and is heads a network for the visually impaired that can advise ABAIR (more below).

Parents and teachers of pupils with dyslexia also sought technology supports. This prompted basic research with teachers and school children, examining the largely non-existent provision for dyslexia assessment and literacy intervention for Irish [7] and an exploratory development of screening [8,9,10] procedures. These requests have also prompted the ongoing development of a platform to train phonological awareness and early literacy [11].

AAC for Irish is also being developed as a response to urgent needs – triggered by a parent of non-speaking autistic children attending Irish-medium education. AAC is an assistive communication device, where sentences can be composed via strings of images/words and spoken out with a synthetic voice. An Irish facility was urgently needed to enable the children to continue in their Irish-medium school and to communicate with their Gaeltacht-based extended family (see parallel SIGUL submission). Collaboration with the wider AAC community is vital here and the parent in question, now a PhD student in ABAIR's research group is involved in an online network, which includes an Irish-learning AAC user and therapists from the U.S. A prototype AAC system is under development [12, 13].

### 5.2. Educational applications

Speech technology has the potential to revolutionise the teaching and learning of Irish and educational applications feature large in ABAIR's application building. Since the founding of the State, education was viewed as key to the maintenance and transmission of the language. Speech technology is for obvious reasons particularly crucial for the

endangered language, as it places the spoken language center stage in all language learning activities. Given the limited access of most learners to native speaker models of the language, having native-speaker voices at your fingertips is enormously helpful.

Even before the first TTS system (for Donegal) was publicly released, a pre-release to a few people was leaked and went viral. Enthusiastic feedback from learners came in from learners – including from the States, Czechoslovakia and Brazil – bringing home how important was access to the authentic spoken language. The fact that the writing system is rather complex and opaque is probably a factor here. The ABAIR website, with ready access to the TTS and ASR facility, is widely used by school children and their parents, as well as by individuals who find it useful to find the pronunciation, to read text aloud etc.

There is high demand for educational applications, games etc. To this end different platforms are being developed, where integration of the ABAIR voices ensures learners are constantly exposed to native speaker speech. *An Scéalaí* [14, 15] is a system, aimed at second and third level students and which can be classroom based or used by the autonomous learner. It features both spoken and written corrective feedback and is now being tested on large numbers of learners. A further platform still under development, *Mol an Óige*, is aimed at the younger learner and targets the training of phonological awareness and early literacy [11]. While these applications are meant for the learner population at large – in the Gaeltacht and in the wider community – they are designed to ensure that they are maximally accessible.

### 5.3. Linguistic knowledge underpinning

In the development of all these applications, linguistic knowledge and an understanding of the structure of the Irish dialects was critical. One the face of it one might imagine that some of the application might require only an interface to the synthetic voices. However, that is only a small part of the task. For example, in developing the AAC prototype, features of language structure must be fully appreciated to come up with reasonable ways to output image to text. Irish has a different word order from English and is a highly inflected language – so that words appear in many forms depending on the grammatical context. This informed all aspect of the AAC design – how the images are arranged and embedded in the displays – and where external grammatical modules were also needed to can take root forms of words and generate the appropriate grammatical forms in the sentence context.

### 5.4. Collaborative user-oriented networks

The active intervention and participation of users and interested parties is the common thread in all of ABAIR's application development.

In the case of the disability related applications, the networking with the users and interested parties was until recently rather informal. More recently a more formal network has been established NEARTÚ (meaning STRENGTHEN) with an online presence on the ABAIR website, where users are invited to join. This network comprises sub- groups with a stake in Irish accessibility applications: members may include users, parents, carers, therapists, specialist teachers, representative of disability organisations etc. NEARTÚ serves as a support network to share information and to advocate for the needs of those with different disabilities. The group will work with and

assist the team in a variety of ways: identifying development priorities, advising on the design of applications, testing prototypes, providing feedback, and disseminating outputs when sufficiently robust for public use. It is envisaged that members of these networks will join the laboratory, working with the team to ensure applications meet the users' needs.

Educational applications equally require extensive collaboration with teachers and educational specialists. Networks are at this point informal, but a more formal network is envisaged in the future – to similarly advise on developments, test platforms, provide feedback and, increasingly, to contribute content.

## 6. Research hub with community input

The website has a page for research output, presenting basic and applied publications at present. There are however many potential research possibilities. The *MíleGlór* recording setup can be exploited to carry out recordings of carefully planned materials to illuminate aspects of the language structure, , allowing empirical research and new insights into Irish and the cross-dialect differences.

Most of the applications for access and education need to be solidly founded on research the language as currently spoken and on language and literacy development. A widening of current basic research would be desirable.

Furthermore the applications (educational platforms and some accessibility applications) yield rich user data. These will be harvested and will provide sources for research. This will not only enrich our understanding of L1 and L2 acquisition, but serves the more practical purpose of guiding future iterations of the applications.

There is potential for a collaborative research hub that will attract researchers from the wider community helping to make the language better equipped for the digital age, and providing new insights into the structure of the language and its acquisition.

## 7. Conclusions

Developing speech and language technologies is not a matter of building a series of technologies, ticking them off as 'done'. Rather, as state-of-the-art technologies emerge, the major task is to make the workable for the communities that use them. The technology itself is rapidly evolving, and with it, new vistas and opportunities will need to be grasped.

For the endangered language in particular, we argue that the ideal is for the language communities to take ownership of the technology and for developers to take every opportunity to foster community partnership in the enterprise. Development by outside agencies/companies who are remote from the language communities may inadvertently be providing technologies that can disempower native-speaker communities and miss important opportunities for the wider language communities. Ultimately, one would ideally strive where possible for the establishment of an indigenous speech and language technology sector, and train members of the language community to become the researchers and developers of the future: that would be the greatest safeguard.

The ABAIR experience has also brought home the critical need for technology developers to have linguistic expertise in the language. Nowadays, with the explosion of deep neural technologies, one gets the impression that knowledge of the

language itself is no longer relevant or necessary. This is not at all true for the endangered language, if only because the vast corpora needed are not available. It is certainly the case that none of the applications we have developed, or are now developing, could be sensibly accomplished without that expertise in Irish linguistics as well as an understanding of cross-dialect variation.

Within development teams, interdisciplinary skills are needed – from the technical expertise to linguistic skills and of course the skills and experience of the professionals and users of whatever technology applications are targeted.

The ABAIR experience is but one story. Of course, the endangered language communities vary a great deal in terms of their specific contexts and circumstances – they vary in the extent to which the language is endangered, in the extent that their language is recognised and receives support, in the opportunities that present for local technology development.

Yet despite differences in the constraints and circumstances, we propose that the underlying principles pertain: that the language, its corpora and resources rightly belong to the language communities and should be freely available to them; that best outcomes stem from the deepest collaboration possible with the communities; that an ideal would be the inclusion of researchers who are themselves from the community, or, failing that, researchers and developers who are attuned to the language, the sociolinguistic context and who are equipped to deliver technologies that are appropriate and sensitive the real needs of the language communities.

## 8. Acknowledgements

The 'ABAIR initiative' covers a number of related ongoing and past funded projects which have supported different aspects of the work reported. The principle source of funding has come from An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meáin; parts of the developments reported have been supported by An Chomhairle um Oideachas Gaeltachta & Gaelscolaíochta (COGG); the National Council for the Blind of Ireland (NCBI) and Foras na Gaeilge.

## 9. References

- [1] Moseley, C. (ed). 2010. Atlas of the world's languages in danger (3rd ed.). Paris: UNESCO Publishing. Retrieved from <http://www.unesco.org/culture/en/endangeredlanguages/atlas>
- [2] CSO. 2011. Daonáireamh na hÉireann: Cainteoirí Gaeilge. Central Statistics Office. Retrieved February 25, 2013 from [http://www.cso.ie/en/media/csoie/census/documents/census2011profile9/Profile\\_9\\_Irish\\_speakers\\_-\\_Combined\\_document.pdf](http://www.cso.ie/en/media/csoie/census/documents/census2011profile9/Profile_9_Irish_speakers_-_Combined_document.pdf)
- [3] Lonergan, L., Qian, M., Berthelsen, H., Wendler, C., Murphy, A., Ní Chiaráin, N., Gobl, C. and Ní Chasaide, A. (2022). Automatic Speech Recognition for Irish: the ABAIR-ÉIST System. *Proceedings of the 4th Celtic Language Technology Workshop at LREC 2022 (CLTW 4)*. European Language Resources Association (ELRA), Marseille, pp. 47-50. URL: <https://www.pure.ed.ac.uk/ws/portalfiles/portal/281144082/Fran-senEtal2022LREC2022Proceedings.pdf#page=60>
- [4] L. Lonergan, M. Qian, N. N'í Chiaráin, C. Gobl, and A. Ní Chasaide, "Cross-dialect lexicon optimisation for an endangered language ASR system: the case of Irish," in proceedings of INTERSPEECH, 2022. [Online]. Available: <https://doi.org/10.21437%2Finterspeech.2022-838>
- [5] Lonergan, L., Qian, Ní Chiaráin, N., Gobl, C. and Ní Chasaide, A. (2022). , "Towards dialect-inclusive recognition in a low-resource language: are balanced corpora the answer?" in Proceedings of INTERSPEECH, Dublin, Ireland, in press

- [6] McGuirk, R. (2015). Exploration of the Use of Irish Language Synthesis with a Screen Reader in the Teaching of Irish to Pupils with Vision Impairment, MPhil Dissertation, School of Linguistic, Speech and Communication Sciences, Trinity College, Dublin, Ireland.
- [7] Barnes, E. (2017) Dyslexia Assessment and Reading Interventions for Pupils in Irish- Medium Education: Insights into current practice and considerations for improvement, M.Phil. Dissertation, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland
- [8] Barnes, E. (2021) Predicting dual-language literacy attainment in Irish-English bilinguals: language-specific and language-universal contributions, PhD Thesis, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland
- [9] Barnes, E., Ní Chasaide, A. and Ní Chiaráin, N. (2018) The design and pre-testing of literacy and cognitive tasks in Irish and English, in *Proceedings of Literacy Association of Ireland 42nd International Conference*, Dublin
- [10] Barnes, E., Ní Chasaide, A., Ní Chiaráin, N. (2021) Bilingual phonological awareness: when interdependence becomes interference, in *Proceedings of the 15th Congress of the International Association for the Study of Child Language*.
- [11] Ní Chasaide, A., Ní Chiaráin, N., Berthelsen, H., Wendler, C., Murphy, A., Barnes, E. and Gobl, C. (2019). Leveraging phonetic and speech research for Irish language revitalisation and maintenance. *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, pp. 994-998.
- [12] Barnes E., Morrin, O., Ní Chasaide, A., Cummins, J., Berthelsen, H., Murphy, A., Nic Corcráin, Claire O’Neill, C., Gobl, C. and Ní Chiaráin, N. (2022). AAC don Ghaeilge: the Prototype Development of Speech-Generating Assistive Technology for Irish. *Proceedings of the 4th Celtic Language Technology Workshop at LREC 2022 (CLTW 4)*. European Language Resources Association (ELRA), Marseille, pp. 127-132. URL: <https://www.pure.ed.ac.uk/ws/portalfiles/portal/281144082/FransenEtal2022LREC2022Proceedings.pdf#page=140>
- [13] Ní Chasaide, A., Barnes, E., Ní Chiaráin, N., McGuirk, R., Morrin, O., Nic Corcráin, M., Cummins, J. (2022). Challenges in assistive technology development for an endangered language: an Irish (Gaelic) perspective. *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, Association for Computational Linguistics, Dublin, Ireland, pp. 80-87
- [14] Ní Chiaráin, N., Ní Chasaide, A., (2019) An Scéalaí: autonomous learners harnessing speech and language technologies, *SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, Graz, Austria
- [15] Ní Chiaráin, N. and Ní Chasaide, A. (2020). The Potential of Text-to-Speech Synthesis in Computer-Assisted Language Learning: A Minority Language Perspective. In Alberto Andujar (ed.), *Recent Tools for Computer- and Mobile-Assisted Foreign Language Learning*, Hershey, PA, IGI Global, pp. 149-169.
- [16] S. B. Davis and P. Mermelstein, “Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.
- [17] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
- [18] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*. New York: Springer, 2009.
- [19] J. Smith, F. Lastname2, and F. Lastname3, “A really good paper about Dynamic Time Warping,” in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 100–104.