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

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

[1]Phonetics & Speech Laboratory, [2]School of Education, Trinity College, Dublin, Ireland

anichsid@tcd.ie, nichiarn@tcd.ie cegobl@tcd.ie

Abstract

The impact of speech and language technology for the endangered language depends on the extent to which the language community engages with its development. 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 speech synthesis and recognition systems – but to prompt development directions, to determine priorities for the most important and urgently needed applications and to collaborate in their provision. 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 ongoing work of the ABAIR initiative at Trinity College Dublin, which is developing linguistic resources, speech technologies and applications for Irish (see www.abair.ie). The paper discusses 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 as a community language), and more distantly related to the P-Celtic languages, Welsh, Breton and Cornish (also extinct). Irish declined over centuries of colonial rule, with a precipitous acceleration following the great famine of the mid-nineteenth century, and the mass emigration it unleashed. 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]. These Gaeltacht communities, while proud of their language and of the rich cultural heritage of song, poetry and story that comes with it, are at the same time keenly aware of how fragile they are as a language community. It is estimated that only 24% (23,175 people) speak Irish on a daily basis outside of the education system [2]. With the loss of the language over time, 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, which diverge 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 regarded as a gold standard.

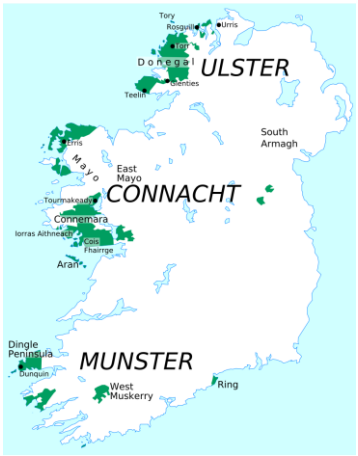

Figure 1: *The official Gaeltacht (Irish speaking) regions.*

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 Irish-speaking families outside the Gaeltacht, especially in Dublin and Belfast, and small but vibrant communities of speakers/learners of Irish across Ireland and among the Irish diaspora across the globe.

Effective language teaching is seen as essential to the survival 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 geographically scattered nature and the many challenges, the

different sectors of the language community are a potential source of strength.

Speech technology can potentially play an important role as part of the wider strategies needed [3, 4, 5] in the revitalisation of the endangered language. ABAIR aspires to playing an enabling role, helping to connect different sectors, empower native speaker communities and facilitate all who wish to make the language and its wider cultural heritage part of their lives.

3. ABAIR's 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. The site also provides an overview of its work-in-progress, and it includes:

(i) Core technologies, text-to-speech (TTS) and speech recognition (ASR).

(ii) Applications which harness these speech technologies, along with the linguistic resources that underpin them and the group's expertise in Irish linguistics. Applications target three overlapping cohorts: firstly, the general public; secondly, learners and teachers of Irish; thirdly, those with disabilities.

(iii) A glimpse at the laboratory's basic and applied research.

(iv) Mechanisms that help connect the research group to the language communities, facilitating their participation in different aspects of the research and development. It also details the group's activities, participation in public events, etc.

A parallel SIGUL 2023 paper [6] demonstrates the website. It and [7] provide technical details of the TTS and ASR systems and information on the corpora with which they are built. Descriptions of specific applications are also presented in parallel papers [8, 9]. The following sections provide an outline of ABAIR's work and of the community's growing role in it.

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 of Ulster, 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 (involving read speech) 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.

The number of voices and the technologies with which they are built have evolved over the years (see [6] for technical details). The focus to date has been on coverage of the three main dialect groups: the current provision is for Ulster Irish (1 female); Connacht Irish (1 female, 1 male), Munster Irish (1 female, 2 male).

As the synthetic voice preserves a 'virtual native speaker' of a dialect, we attach importance to extending coverage to the more endangered sub-dialects. These voices, coupled to the applications that use them (more below) will assist local communities in their efforts to preserve, maintain and revitalise their distinct dialects. As a step in this direction, recordings are currently ongoing in Ring, Co. Waterford (Figure 1) with active community involvement in recruiting and recording speakers.

This approach contrasts sharply with the practice of 'Big Tech' and commercial companies. For many years the ABAIR voices were the only TTS systems available, but more recently, commercially based synthetic voices have been developed. These have been built using 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 is not appropriate for native speaker communities either, and sends subliminal messages which can only erode the status of the endangered native language/dialect, a factor contributing to its loss, as demonstrated in the work of Dorian [10, 11].

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 [12, 13], tracking the performance of the system across dialects is being 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 [14]. A dialect recognition component to the ASR architecture is envisaged, to optimise performance and cater for cross dialect differences [14, 15].

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 are directed at L2 speakers and learners. Given the scale and diversity of corpora needed, an online recording tool *MíléGlór* ('A Thousand Voices') was developed and is being used 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 (the least likely to be online) the ABAIR team is increasingly focussed on large-scale live recordings (for speech recognition and synthesis) in the Gaeltacht communities. These are supplemented by other available recorded materials, harvested from various sources.

The active participation of the community has been fostered by presenting ABAIR's work at public events, such as *Oireachtas na Samhna*, a major, yearly, weeklong Irish-language gathering. This has allowed considerable interaction and discussion with the wider language community, an opportunity to explain why the *MíleGlór* recordings are needed and the opportunity to carry out extensive in-person recordings of native speakers. This interaction has yielded a considerable body of speech data for the different dialects, essential to our ASR system to date. 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. These kinds of interaction have created greater awareness of ABAIR's resources and established networks that facilitate the ongoing work of ABAIR within local communities. Once demonstrated, local communities are requesting their voices to be included, and they relate particularly to the need for applications for disability and access, as well as for educational resources and supports.

5. Applications with active participation of user groups

The impact of these technologies for the community derives mainly from the applications that deploy them. Specific requests came with the first mention that a TTS system was being developed, and long before it was available.

5.1. Applications for disability & access

The first approaches concerned children who required a screen-reading facility. Parents and grandparents of visually impaired children attending Gaeltacht and Irish-medium schools visited the lab with the children, bringing home the severe disadvantages they were under and the threat to their continued participation in school. Approaches by visually impaired adults also stressed the extent of their exclusion from the language (there being only three Irish Braille books available in Irish). This led to a skilled blind programmer joining the team to collaborate on the development of a screen-reading facility (an Irish plugin for the open-source NVDA screen-reader). He collaborated on system design, advised on features needed from a user's perspective, and tested the system with visually impaired school children by networking with teachers of the blind [6]. He further developed the system to work with the Liblouis Braill reader, to allow simultaneous speech and Braille output. He continues to collaborate with ABAIR on a voluntary basis, to maintain and update the facility. He also heads a user-network for the visually impaired to guide ABAIR (see below).

Parents and teachers of pupils with dyslexia also sought speech 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 [17] and an exploratory development of screening procedures [18, 19, 20]. This also prompted the ongoing development of a platform to train phonological awareness and early literacy [9].

AAC for Irish is also being developed in response to urgent needs. AAC is an assistive communication device, where sentences can be composed via strings of images/words, and spoken out with a synthetic voice. The development was initially triggered by a parent of non-speaking autistic children attending Irish-medium education, and her network of AAC users. An Irish facility was urgently requested to enable the children to continue in their Irish-medium school and to communicate with their Gaeltacht-based extended family. The development of an Irish prototype AAC system is described in [8] (see also earlier attempts [21]). The parent in question, now a PhD researcher with the ABAIR group, is involved in the design, development and testing of the system. She leads an online AAC advisory network, with members in Ireland and the United States.

5.2. Educational applications

Since the founding of the State, education was viewed as key to the maintenance and transmission of the language. Speech technology has the potential to revolutionise the teaching and learning of Irish, and educational applications feature large in ABAIR's application-building. 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 United States, Czechia and Brazil – brought home the importance of access to authentic spoken rendition of text. The fact that the sound system is complex, the writing conventions are opaque and both differ a great deal from English are probably a factor here (see [9]). The ABAIR website, with ready access to the TTS, ASR and applications using them, is widely used by school children, their parents, and anyone who wants to read text aloud, pronounce Irish names, 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* [21] is a system, aimed at second and third level students and which can be classroom based or used by the autonomous learner. It features spoken and written corrective feedback and is being extensively tested on 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 [9]. While these applications are meant for the learner population at large – in the Gaeltacht and in the wider community – they are designed in a way to ensure that they are maximally accessibile.

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. On the face of it one might imagine that some of the applications might require only an interface to the TTS systems. 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 informs all aspect of AAC design: how images are arranged and embedded in display boards; external modules,

needed to generate grammatical forms of lemmas, appropriate to the sentence context.

5.4. Evolving 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, networking was mostly informal, but the recently established *NEARTÚ* ('STRENGTHEN') network makes the collaboration more formal, with an online presence on the ABAIR website and an open invitation to join, for all interested individuals and groups with a stake in Irish accessibility applications. Members may include users, parents, carers, therapists, specialist teachers, representative of disability organisations, etc. NEARTÚ will guide ABAIR's accessibility application development. It will also provide information and support for those with disabilities, advocating for their needs and rights with regard to Irish speech technologies. NEARTÚ directly assists ABAIR by 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 continue to join the research group, helping to ensure that applications fully meet users' needs.

Educational applications equally require extensive collaboration with teachers and educational specialists. The current networks are extensive, but informal. A more formal network is envisaged – to similarly advise on developments, test platforms, provide feedback and, increasingly, to directly contribute content.

6. Research: technology and linguistics

ABAIR technology development is underpinned by linguistic knowledge of Irish, and sensitive to the sociolinguistic realities of the language. The website currently features ABAIR's publications and there is considerable potential for broadening its basic and applied research. Critical to Irish as an endangered language is the symbiotic relationship between technology development and linguistic research: the former opens up exciting new vistas for the latter, and the benefits are mutual. For example, the spoken corpora for the various dialects allow quantitative empirical research of a kind and scale not hitherto possible for Irish, yielding new insights into the language, cross-dialect differences and how dialects are evolving in this challenging context. The *MíleGlór* recording facility can be used to collect materials designed to elucidate very specific issues, e.g., aspects that are little understood, or normative data for therapy and teaching. ASR and forced alignment facilitate otherwise arduous transcription tasks and open up archival materials. All linguistic corpora and knowledge feed directly into better core technologies and more adequate applications.

The applications under development for access and education are already stimulating basic research on language and literacy development. These harvest user's data (with their consent), generating corpora that will enrich our understanding of L1 and L2 acquisition, while enabling future iterations of these applications that can 'intelligently' adapt to the user.

7. Conclusions

Developing speech and language technologies is not a matter of building a series of technologies, ticking them off as 'done'. Rather, the technology is rapidly evolving, and with it, new vistas and opportunities arise. As state-of-the-art technologies emerge, the goal is to maximise the ways in which they are deployed in applications useful to the language communities. The impact of ABAIR's current offerings is difficult to measure, but can be indirectly inferred from the growing traffic on its website, with currently c. 2000 visits per day.

For the endangered language in particular, 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 and these communities may inadvertently be providing technologies that can disempower native-speaker communities and miss important opportunities for the wider language community.

The *Digital Plan for the Irish Language* [3] envisages the establishment of an indigenous speech and language technology sector. Towards this, the education of young members of the language community is key, to ensure the prerequisite multidisciplinary expertise that will equip them as researchers and developers of the future. ABAIR's current linkages to the language community, e.g., corpus collection in the Gaeltacht with community involvement, stakeholder participation in application development, etc. contributes to this end.

The ABAIR experience highlights the need for technology developers to have linguistic expertise in the language. Nowadays, with the explosion of deep learning and AI 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. Furthermore, development teams require interdisciplinary skills – from technical expertise to linguistic knowledge and – for application development – the skills of experienced professionals and input from users in the area of application.

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 to which 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 common underlying principles pertain: that the language, its corpora, the resources and technologies that are built from them 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, where that is not (yet) possible, 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

This work is supported by An Roinn Turasóireachta, Cultúir, Ealaíon, Gaeltachta, Spóirt agus Meáin with funding from the National Lottery as part of *An Stráitéis 20 Bliain don Ghaeilge, 2010 – 2030*. It is also partially funded by An Chomhairle um Oideachas Gaeltachta & Gaelscolaíochta (COGG).

9. References

[1] C. Moseley (ed), "*Atlas of the world's languages in danger"* (3rd ed.). Paris: UNESCO Publishing. Retrieved from https://www.unesco.org/culture/en/endangeredlanguages/atlas, 2011.

[2] Central Statistics Office, "*Daonáireamh na hÉireann: Cainteoirí Gaeilge, 2011*", Retrieved from https://www.cso.ie/en/media/csoie/census/documents/census2011profile9/Profile Irish speakers - Combined document.pdf

[3] Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media, "*The Digital Plan for the Irish Language*", Government Publication, Dec. 2022. https://www.gov.ie/ga/foilsiuchan/4174d-plean-digiteach/

[4] An Roinn Cultúir, Oidhreachta agus Gaeltachta, "*20 Years Strategy for the Irish Language 2010 – 2030*", Government Publication, 2010.

[5] An Roinn Cultúir, Oidhreachta agus Gaeltachta, "*The Action Plan for the Irish Language, 2018-2022",* Government Publication, 2018.

[6] A. Murphy, L. Lonergan, M. Qian, H. Berthelsen, C. Wendler, N. Ní Chiaráin, A. Ní Chasaide, C. Gobl, "ABAIR & ÉIST: A demonstration of speech technologies for Irish", *Proceedings of SIGUL 2023 Workshop at INTERSPEECH 2023*, Dublin, Ireland, 2023.

[7] L. Lonergan, M. Qian, H. Berthelsen, C. Wendler, A. Murphy, N. Ní Chiaráin, C. Gobl, and A. Ní Chasaide, 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, 2022.

[8] E. Barnes, J. Cummins, R. Errity, O. Morrin, H. Berthelsen, C. Wendler, A. Murphy, H. Husca1, N. Ní Chiaráin, A. Ní Chasaide, "*Geabaire, the first Irish AAC system: voice as a vehicle for change", Proceedings of SIGUL 2023 Workshop at INTERSPEECH 2023*, Dublin, Ireland, 2023.

[9] A. Ní Chasaide, N. Ní Chiaráin, R. Errity, O. Mroz, O. Ní hAonghusa, S. Ní Chasaide, A. Giovannini, E. Barnes, "*Mol an Óige*: a phonological awareness and early literacy platform for Irish", *Proceedings of SIGUL 2023 Workshop at INTERSPEECH 2023*, Dublin, Ireland, 2023.

[10] N. Dorian, "*East Sutherland Gaelic: the dialect of the Brora, Golspie, and Embo fishing communities*", Dublin Institute for Advanced Studies, Dublin Ireland, 1978.

[11] N. Dorian, "*Language Death: The Life Cycle of a Scottish Gaelic Dialect",* University of Pennsylvania Press, 1981.

[12] L. Lonergan, M. Qian, N. Ní 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,* Incheon, Korea, 2022.

[13] L. Lonergan, M. Qian, N. Ní Chiaráin, C. Gobl, and A. Ní Chasaide, "Towards dialect-inclusive recognition in a low-resource language: are balanced corpora the answer?" in *Proceedings of INTERSPEECH 2023*, Dublin, Ireland, 2023.

[14] L. Lonergan, M. Qian, N. Ní Chiaráin, C. Gobl, A. Ní Chasaide, "Towards spoken dialect identification of Irish", *Proceedings of SIGUL 2023 Workshop at INTERSPEECH 2023*, Dublin, Ireland, 2023.

[15] R. McGuirk, "Exploration of the Use of Irish Language Synthesis with a Screen Reader in the Teaching of Irish to Pupils with Vision Impairment" unpublished M.Phil. Dissertation, School of Linguistic, Speech and Communication Sciences, Trinity College, Dublin, Ireland, 2015.

[16] E. Barnes, "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, 2017.

[17] E. Barnes, "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, 2021.

[18] E. Barnes, A. Ní Chasaide, and N. Ní Chiaráin, "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, 2018.

[19] E. Barnes, A. Ní Chasaide, N. Ní Chiaráin, "Bilingual phonological awareness: when interdependence becomes interference", in *Proceedings of the 15th Congress of the International Association for the Study of Child Language,* 2021.

[20] E. Barnes, O. Morrin, A. Ní Chasaide, J. Cummins, H. Berthelsen, A. Murphy, M. Nic Corcráin, C. O'Neill, C. Gobl, and N. Ní Chiaráin, "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, 2022.

[21] 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.