*INTERSPEECH 2017*
August 20–24, 2017, Stockholm, Sweden

# The ABAIR initiative: Bringing Spoken Irish into the Digital Space

*Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Christoph Wendler, Harald Berthelsen, Andy Murphy, Christer Gobl*

The Phonetics and Speech Lab., School of Linguistic, Speech and Communication Sciences, Trinity College, Dublin

anichsid@tcd.ie, neasa.nichiarain@tcd.ie

## Abstract

The processes of language demise take hold when a language ceases to belong to the mainstream of life's activities. Digital communication technology increasingly pervades all aspects of modern life. Languages not digitally 'available' are ever more marginalised, whereas a digital presence often yields unexpected opportunities to integrate the language into the mainstream. The ABAIR initiative embraces three central aspects of speech technology development for Irish (Gaelic): the provision of technology-oriented linguistic-phonetic resources; the building and perfecting of core speech technologies; and the development of technology applications, which exploit both the technologies and the linguistic resources. The latter enable the public, learners, and those with disabilities to integrate Irish into their day-to-day usage. This paper outlines some of the specific linguistic and sociolinguistic challenges and the approaches adopted to address them. Although machine-learning approaches are helping to speed up the process of technology provision, the ABAIR experience highlights how phonetic-linguistic resources are also crucial to the development process. For the endangered language, linguistic resources are central to many applications that impact on language usage. The sociolinguistic context and the needs of potential end users should be central considerations in setting research priorities and deciding on methods.

**Index Terms**: endangered languages, Irish Gaelic, speech technology, user applications, knowledge-based, linguistic-phonetic

## 1. Introduction

In her poignant account of the processes of language death, Dorian describes how, when a language ceases to belong to mainstream activities and ceases to be integrated into the dominant social and power networks, it looses prestige and relevance, and eventually dies [1]. These processes described many decades ago for the Scottish Gaelic dialect of East Sutherland are sadly the new reality for many languages today. It has been estimated that between 60-90% of the world's languages may be extinct within the next 100 years [2]. One's language is arguably one's greatest cultural artefact and the current erosion of our linguistic heritage is a cultural analogue of the ongoing ecological loss of habitats and species.

Languages are not technology-neutral. Digital technology increasingly pervades every aspect of our lives, how we work, communicate, socialise, learn, etc. and if Dorian were writing today, digital technology would likely loom large in the equation. But, although the digital revolution is undoubtedly contributing to the globalisation of a monolingual culture and thus to language loss, this same digital technology offers unprecedented opportunities to counteract this trend of language attrition. By bringing the language into the digital space it can provide the very mechanism that integrates it back into the mainstream. Precisely because the technology is so tightly interwoven with every aspect of our lives, embedding one's language into the technology core can greatly increase the domains in which it can be used, making it a normal part of the average person's day-to-day activities.

Endangered, minority, and the widely spoken but under-resourced languages, all lack the commercial incentives that have driven speech and language development in the major world languages. The processes of development, however, have become widely available and are now less dependent on large-scale financing. It is imperative for the survival of many of these languages that the new development opportunities are grasped: failure to do so risks increasing their marginalisation in the digital world.

It is within this context that ABAIR is presented, an initiative to develop speech technology and resources for the Irish language, which has progressed for the most part with modest financial resourcing. ABAIR's experience demonstrates how even small beginnings can stimulate a chain reaction and can yield impact in unexpected ways. Largely similar kinds of challenges confront minority and otherwise under-resourced languages and this paper is offered in the hope that the ABAIR experience - the difficulties presented and the 'solutions' adopted - may be instructive for other language communities who share this journey.

The ABAIR initiative has evolved to embrace three key aspects illustrated in Figure 1: (i) the development of technology-oriented linguistic-phonetic resources, (ii) the development of core technology (text-to-speech, TTS, to date), and (iii) the development of applications that use the technology and the underpinning linguistic resources to enable their integration into mainstream activities. These applications are geared to the public, to language learners, and to the inclusion of those with disabilities. As the image in Figure 1 attempts to convey, although the impact of speech and language technologies are primarily realised in terms of their applications, these depend on the solidity of the core technologies and particularly on the underpinning of the roots, which are the linguistic resources.

## 2. The sociolinguistic context and goals

Irish, a Q-Celtic language closely related to Scottish Gaelic, and has been classified by UNESCO as 'definitely endangered' [3]. It is spoken as a community language in Gaeltacht areas (Irish-speaking pockets), mostly situated in the West of

http://dx.doi.org/10.21437/Interspeech.2017-1407

Ireland. Even within the Gaeltacht, it is estimated that only 24% of the population speak Irish on a daily basis, although the precise numbers are hard to come by and depend on what is officially designated a *Gaeltacht*. The critical mass of native speakers is further dissipated by virtue of the fact that the Gaeltacht areas are situated in locations which are remote from each other.

The situation of Irish is somewhat anomalous in that, although endangered, it is recognised as the first national language in the Republic and has, since 2007, been designated an official EU language. Despite the demographics, there is a reasonably vibrant community of Irish speakers outside the Gaeltacht, particularly in urban areas. It is compulsory to study Irish up to school-leaving age and there is a growing demand for Irish-medium education. It is evident that the future of the language will depend on its transmission, both in the Gaeltacht and outside.

Education is a key factor and presents many challenges for language teaching. One particular challenge in the area of education is the lack of native speaker models and most school children do not hear the language spoken in its native form or as a community language. Teaching resources are often dated and can reflect a rather poor understanding of the linguistic structure. In an era where modern methodologies incorporating computer-assisted language learning (CALL) are on the increase, it is particularly important that Irish is poised to avail of the opportunities that the digital revolution is bringing about in education. For the general public who has an interest in Irish, social media incorporating speech and language technologies have the potential to consolidate geographically disparate communities. Inclusion of those with disabilities is a further dimension where these technologies are vitally needed.

The sociolinguistic context is a major consideration in setting goals and priorities for speech and language technology development. Understanding of the linguistic challenges of the specific language is a further central consideration in order to ensure that the technology and end-applications are appropriate to the needs of end-users. Some of the challenges of linguistic resourcing are outlined in the following section.

## 3. Building linguistic resources

In the early stages, much of the work was directed at building the linguistic resources to underpin technology development, such as annotated corpora, pronunciation lexica and letter-to-sound (LTS) rules, which were either unavailable, or not in a form that was adapted for technology.

### 3.1. A multi-dialect approach

As is typical of endangered languages, there is no spoken standard (there is a written standard), and a multi-dialect facility was envisaged from the outset. While research initially focused on resources for the Ulster dialect of Donegal, components (phonetic, lexical, etc.) were developed in a modular fashion: *core* modules encompassing what was common to all dialects and *local* modules, containing dialect-specific rules.

### 3.2. Linguistic structure and challenges

Linguistic resources required for a TTS system include a phonetiser, which translates the written letters (or groups of letters) into their corresponding sounds. The sound sequences need to be marked for stress and syllabification, in single words and compounds. Furthermore, postlexical rules are required to cater for the fact that these stress and syllabification patterns may work differently in running speech than in citation form. Some features of Irish present particular challenges.

#### *3.2.1. Phonology, Orthography: Letter-to-sound rules*

The phonological system of Irish differs from other Western European languages in that the consonantal system is over twice the size: an opposition of palatalized and velarized consonants effectively doubles the inventory vis-a-vis a language like English e.g., /bʲi:/ *bí* (be) vs. /bˠi:/ *buí* (yellow). The fact that the Roman alphabet doesn't cater to this feature has resulted in convoluted, complex orthographic conventions - see illustrative example in Figure 2 of the diversity of orthographic representations for the phoneme /i:/ depending on the quality of preceding or following consonants.

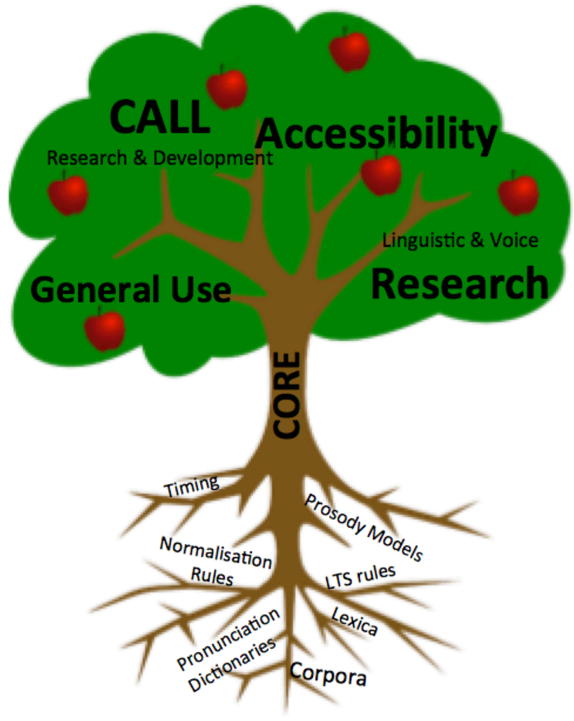


Figure 1: *The ABAIR Framework for Development incorporating (1) linguistic resource development, (2) core technology and (3) applied technology*

This type of difficulty is compounded by the fact that the writing system is ancient and many letters represent sounds that have either mutated, weakened or disappeared. (Unsurprisingly, the phonics of Irish are poorly grasped by learners, and often by teachers, greatly increasing the difficulty learners experience in acquiring pronunciation and literacy skills – a point returned to below).

| Sound: | Orthography: | Example: |
|---|---|---|
| /i:/ | í | => sí /sʲi:/ |
| | ío | => díol /dʲi:lˠ/ |
| | aí | => caí /kˠi:/ |
| | ao | => saol /sˠi:lˠ/ |
| | oí | => cloí /klˠi:/ |
| | uí | => buí /bˠi:/ |
| | aío | => maíomh /mˠi:wˠ/ |
| | aoi | => saoil /sˠi:lʲ/ |
| | uío | => buíon /bˠi:nˠ/ |

Figure 2: *sample of phonological -> orthographic mapping*

### 3.2.2. *Lexicon, Morphology, Grammar*

Irish morphology is also complex. Initial consonants undergo 'mutations' (consonantal alternations, e.g. of stops and fricatives) in specific grammatical contexts, a feature common to all Celtic languages. Verbal forms can have as many as 42 inflected forms. A complex system of numerals (differentiating between animate and non-animate) interacts with the inflectional system in multiple ways.

A pronunciation lexicon also needed to be developed and this, like the other components, was hand-crafted, since machine learning yielded poor results. An existing pocket lexicon *Foclóir Póca* [4] was unsuitable, as the forms contained were a mix of all dialects, and thus not matched to the speech of any one speaker. As with other linguistic hand built components, the time spent on their development paid a rich dividend when it came to the inclusion of new dialects, as relatively simpler adaptations were required. As will be clear in the following sections, these knowledge-based linguistic resources are particularly important for language-learning applications.

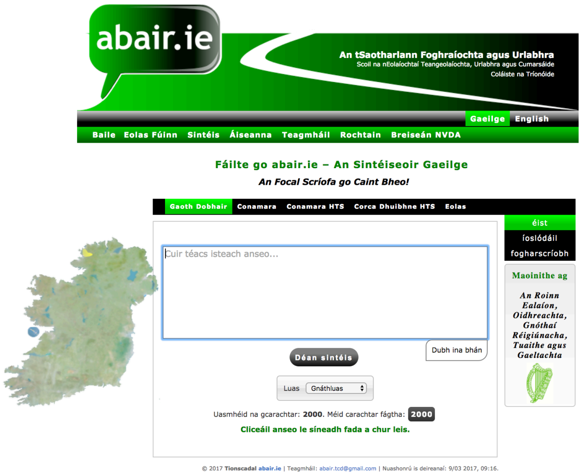
## 4. Core TTS development

TTS has been the core technology under development to date. Synthetic voices for the three main dialects have been developed and made available online at www.abair.ie. As shown in Table 1, two types of TTS systems have been developed: unit selection systems using the Festival speech synthesis system [5] and hidden Markov model-based speech-synthesis system (HTS) [6]. These are continuously being refined.

Table 1: *Synthetic voices available at www.abair.ie.*

| Dialect/Synthesis | Sample Rate | No. Utts | Sex/Age |
|---|---|---|---|
| Ulster; unit-sel. | 44.1kHz | 4,014 | F / 20s |
| Ulster; HTS | 32 kHz | 4,200 | |
| Connaught; unit-sel. | 44.1kHz | 3,170 | M / 50s |
| Connaught; HTS | 32 kHz | 3,300 | |
| Munster; unit-sel. | 44.1kHz | 2,622 | F / 50s |
| Munster; HTS | 32 kHz | 2,700 | |

Figure 3 shows the public webpage, which can be navigated in Irish or in English (Irish version shown).



An tSaotharlann Foghraíochta agus Urlabhra
Scoil na nEolaíochtaí Teangeolaíochta, Urlabhra agus Cumarsáide
Coláiste na Tríonóide

Fáilte go abair.ie – An Sintéiseoir Gaeilge

*An Focal Scríofa go Caint Bheo!*

Maoinithe ag

*An Roinn Ealaíon, Oidhreachta, Gnóthaí Réigiúnacha, Tuaithe agus Gaeltachta*

Uasmhéid na gcarachtar: **2000**. Méid carachtar fágtha: **2000**

Cliceáil anseo le síneadh fada a chur leis.

© 2017 **Tionscadal abair.ie** | Teagmháil: abair.tcd@gmail.com | Nuashonrú is deireanaí: 9/03 2017, 09:16.

Figure 3: *the ABAIR public webpage*

The user chooses the dialect, the synthesis mode and can further manipulate the speed of the speech output. The audio files can be saved in .mp3 or .wav formats. Certain linguistic components are made available for users. For example, there is direct access to the system's phonetiser, providing the phonetic forms, and users can also access a part-of-speech tagger, developed by Uí Dhonnchadha [5]. Current interest focuses on the provision of more voices, e.g. childrens' voices, needed for many applications. More dialects are being targeted, particularly the highly endangered varieties. Having a live synthesiser for a dialect that is on the verge of disappearance fulfills an important documentation and preservation function [6] and may help prevent its irretrievable loss.

More basic research on speech synthesis is also being conducted with a view to downstream innovations geared at many of the applications envisaged. An important aspect will be the incorporation of mechanisms to control the synthetic voice quality to enable expressiveness [7], [8], as is desirable in many of the applications described below, such as interactive multimodal language learning games. In the future, further core technologies are also envisaged, such as ASR and spoken dialogue systems.

## 5. Applied technologies

As mentioned, it is through the applications that the technology impacts on the life of the language, by enabling its use in peoples' daily lives. The areas of application that stemmed from the initial TTS development have targeted the areas of public usage, education and disability/access.

### 5.1. Public usage

Within days of an ABAIR voice being available on the web and with no publicity (intended as a limited trial run) it drew a wide reaction that revealed an unexpected community of Irish users from all corners of the globe, digitally connected through the site. Half of all users were from North America but there was coverage in virtually every country worldwide. Feedback from learners in particular stressed how they could finally match the spoken word to written Irish texts.

A web reader has been developed more recently to facilitate access to the spoken output of all online text, e.g. online dictionaries (pronunciation dictionaries only give access to headwords). Interest from SMEs and other public bodies is considerable and the API for Irish synthesis will enable the integration of the language into many domains where users expect spoken output to be available.

### 5.2. Language learners & CALL

The highest area of demand is coming from the educational sector. In parallel collaborative projects, language learning applications are being developed that incorporate both the ABAIR voices and the phonetic-linguistic resources.

**CabairE:** tools to aid literacy acquisition are under development and these draw on the phonological components of ABAIR to promote phonological awareness, and the LTS rules to assist in phonics training. The next step will include materials for the acquisition of morphophonemic and grammatical aspects of the language, which are often poorly acquired.

### 5.3. Educational Language Learning Games:

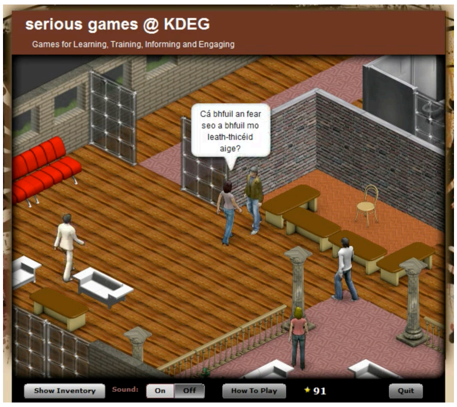

serious games @ KDEG
Games for Learning, Training, Informing and Engaging
Cá bhfuil an fear seo a bhfuil mo leath-thicéid aige?
Show Inventory
Sound:
On
Off
How To Play
91
Quit

Figure 4: *the* Digichaint *educational game*

ABAIR voices are also used to build educational games, which involve immersive multimodal interaction in a virtual reality space, in collaboration with game design researchers (e.g. KDEG, Trinity College, Dublin). A number of different games have been piloted including an adaptive, interactive game, *Digichaint,* [9] (see Figure 4) based on the adaptive ALIGN architecure developed originally as part of the European project ELEKTRA, [10], [11]. A second platform, *Fáilte go TCD,* featuring a virtual reality scene set at Trinity College, Dublin [12]. This arose from a collaboration with the Metropolis project [13]. A third platform consisted of a chatbot, *Taidhgín*, featuring a talking monkey [14]. *Taidhgín* was built using Artificial Intelligence Markup Language (AIML), an XML-based open-source programming language developed by Richard Wallace and the Alicebot free software community. Taidhgín is hosted and run from Pandorabots which is a 'free open-source-based community web service which enables you to develop / publish chatbots on the web' (pandorabots.com).

Tests carried out nationwide on 16-year old learners of Irish (N=250) yielded very positive responses, particularly for the monkey-chatbot [15]. There is enormous scope to develop such educational games incorporating both simple spoken interaction and game-based training on specific phonological and grammatical language goals. Following on the pilot work, it is clear that even simple dialogue systems will be a very powerful tool in eliciting spoken interaction and in providing an engaging framework where more potentially 'boring' aspects of linguistic structure can be trained in a fun way. As the linguistic resources available come on stream, so too will the sophistication that can be built into such dialogue-based games. A recent example involves the use of *WordNets as Gaeilge* (*WordNets* in Irish) [16] to allow the monkey-chatbot to detect the appropriateness of the use of the copula in Irish.

The intention is that ABAIR will serve as a virtual resource centre for web-based educational tools, games, etc. that incorporate speech and language technology and linguistic knowledge. This will entail extensive collaboration with educationalists in content provision. One of the attractions of such web-based resources is their global outreach potential, catering to local and foreign learners at various language levels.

### 5.4. Disability and access

A high demand for speech technology applications comes from the area of disability and access. Those with disabilities are often excluded or discouraged from participation in Irish language education. The visually impaired lack access to written and online materials. In the case of children with dyslexia, the lack of materials for screening and intervention led to an almost automatic exclusion also. The addition of Irish synthesis to the open source NVDA screenreading facility for the visually impaired has been developed, along with the facility for Braille output in Irish and is freely available to the public [17]. Other initiatives include the development of speech-enabled school textbooks (DAISYBooks) for the visually impaired, including highlighting and magnifying of text, easy speed control for the spoken output, etc. In a separate new initiative, tools targeting phonological awareness, phonics and literacy development will provide a foundation for screening and intervention in the case of dyslexia [18], [19].

## 6. A Digital Plan for Irish

There is a growing awareness in Ireland as elsewhere that national plans need to be put in place to ensure that the national languages, whether minority or majority, are not left behind in the digital space. Work is ongoing on a *Digital Plan for Irish*, to tie in with the Irish Government's *20-Year Strategy for the Irish Language (2010-2030)*. The case for developing speech and language technologies is strengthened by the fact that Irish has been designated as an official language of the EU.

This *Digital Plan* embraces both Speech and Language technologies, some at very initial stages of development. The plan acknowledges the importance of a holistic vision that encompasses both linguistic resource development, core technology building and applications in areas such as those outlined above (see Figure 1). Fundamentally, a fusion of linguistic knowledge-based and machine learning approaches is envisaged to yield the best outcomes for language communities.

## 7. Conclusions

Future ABAIR developments will include recognition (ASR) and dialogue systems. Although full systems will take time to achieve, even limited versions can initially be usefully deployed in user applications.

A clear vision of end user applications helps to drive and define the priorities for technology development and necessitates goals that entail quite basic aspects of research. For example, in ASR research, it is clear that childrens' voices will be important, and this, along with the diversity of dialects to be catered for, presents numerous challenges.

Although many of the tangible outputs are of a practical nature, and intended for immediate public use, this work is ongoing in parallel with and building on basic research, both in linguistics and in fundamental aspects of the technology. An important goal in our synthesis development will be the incorporation of expressive nuancing of the synthetic voice, as will be needed to create believable and engaging interactive dialogue partners and game characters, as well as for disability applications [9], [12], [20]. The applied research on language educational tool development is stimulating fundamental research on dyslexia and on how the phonological and phonics system of the language impacts on literacy acquisition.

## 8. Acknowledgements

This work is supported by An Roinn Ealaíon, Oidhreachta, Gnóthaí Réigiúnacha, Tuaithe agus Gaeltachta, *ABAIR* initiative, and COGG, *CabairE project*. Funding in the past was provided by Foras na Gaeilge, the National Council for the Blind of Ireland, and the EU-INTERREG programme.

# 9. References

[1] N. C. Dorian, *East Sutherland Gaelic: the dialect of the Brora, Golspie, and Embo fishing communitites*. Dublin Institute for Advanced Studies, 1978.

[2] S. Romaine, "Preserving endangered languages," *Lang. Linguist. Compass*, vol. 1, no. 1–2, pp. 115–132, 2007.

[3] C. Moseley, Ed., *Atlas of the world's languages in danger*, 3rd ed. Paris: UNESCO Publishing, 2010.

[4] D. Ó Baoill, Ed., *Foclóir Póca*. Dublin: An Gúm, 1986.

[5] R. A. J. Clark, K. Richmond, and S. King, "Festival 2-build your own general purpose unit selection speech synthesiser," in *Proceedings of the Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 173–178.

[6] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *IEEE Speech Synthesis Workshop*, 2002, pp. 227–230.

[7] C. Gobl, E. Bennett, and A. Ní Chasaide, "Expressive synthesis: how crucial is voice quality," in *Proceedings of the IEEE Workshop on Speech Synthesis*, 2002, pp. 1–4.

[8] A. Murphy, I. Yanushevskaya, C. Gobl, and A. Ní Chasaide, "Rd as a control parameter to explore affective correlates of the tense-lax continuum," in *Proceedings of Interspeech 2017*, forthcoming.

[9] N. Ní Chiaráin and A. Ní Chasaide, "The Digichaint interactive game as a virtual learning environment for Irish," in *CALL communities and culture – short papers from EUROCALL 2016*, S. Papadima-Sophocleous, L. Bradley, and S. Thouësny, Eds. Research-publishing.net, 2016, pp. 330–336.

[10] N. Peirce and V. Wade, "Personalised learning for casual games: The 'Language Trap' online language learning," in *Proceedings of the Fourth European Conference on Game Based Learning (ECGBL 2010)*, 2010, pp. 306–315.

[11] N. Peirce, O. Conlan, and V. Wade, "Adaptive Educational Games: Providing Non-invasive Personalised Learning Experiences," in *Second IEEE International Conference on Digital Games and Intelligent Toys Based Education*, 2008, pp. 28–35.

[12] N. Ní Chiaráin and A. Ní Chasaide, "Evaluating Synthetic Speech in an Irish CALL Application: Influences of predisposition and of the holistic environment," in *SLaTE 2015: 6th Workshop on Speech and Language Technologies in Education*, 2015, pp. 149–154.

[13] C. O'Sullivan and C. Ennis, "Metropolis: multisensory simulation of a populated city," in *Proceedings of the Third International Conference on Games and Virtual Worlds for Serious Applications*, 2011, pp. 1–7.

[14] N. Ní Chiaráin and A. Ní Chasaide, "Chatbot Technology with Synthetic Voices in the Acquisition of an Endangered Language: Motivation, Development and Evaluation of a Platform for Irish," in *the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[15] N. Ní Chiaráin, "Text-to-Speech Synthesis in Computer-Assisted Language Learning for Irish: Development and Evaluation," (Doctoral thesis, CLCS, Trinity College, Dublin), 2014.

[16] J. O'Regan, K. Scannell, and E. Uí Dhonnchadha, "lemonGAWN: WordNet Gaeilge as Linked Data," in *LDL 2016 – 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, 2016, pp. 36–40.

[17] R. McGuirk, "Exploration of the Use of Irish Language Synthesis with a Screen Reader in the Teaching of Irish to Pupils with Vision Impairment," Trinity College, Dublin, Ireland, 2015.

[18] F. Nic Pháidín, N. Ní Chiaráin, and A. Ní Chasaide, "Assessing Irish literacy resources: guidelines for teacher and publishers," in *Exploring the Literacy Landscape: Celebrating 40 YEars of Research and Practice: Conference Proceedings of the Literacy Association of Ireland's 40th International Conference, forthcoming*.

[19] E. Barnes, N. Ní Chiaráin, and A. Ní Chasaide, "Departures from the 'norm': how the phonology, morphology and orthography of the Irish language impact on literacy instruction and acquisition," in *Exploring the Literacy Landscape: Celebrating 40 Years of Research and Practice: Conference Proceedings of the Literacy Association of Ireland's 40th International Conference, forthcoming*.

[20] N. Ní Chiaráin and A. Ní Chasaide, "Chatbot technology with synthetic voices in the acquisition of an endangered language: motivation, development and evaluation of a platform for Irish," in *10th edition of the Language Resources and Evaluation Conference, 23-28 May 2016*, 2016, pp. 3429–3435.