SPEECH TECHNOLOGY AS DOCUMENTATION FOR ENDANGERED LANGUAGE PRESERVATION: THE CASE OF IRISH

Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Harald Berthelsen, Christoph Wendler, Andrew Murphy

Phonetics and Speech Laboratory, Trinity College Dublin, Ireland

ABSTRACT

Developing speech technology such as text-to-speech (TTS), requiring as it does a raft of phonetic and linguistic resources, can provide a powerful way to document endangered languages. Drawing on the experience of the ABAIR initiative, developing such resources for Irish [1], we illustrate how both the technology and the underpinning resources can be exploited in a variety of ways that can contribute to the preservation and revitalisation of these languages. By enabling new avenues of application, they can further help address the particular challenges that face the language users and learners. To maximise the immediate and downstream impact, resource development should ideally involve linguistically transparent, rule-based approaches, rather than the machine learning approaches typical of the commercially driven TTS systems for major world languages.

Keywords: Speech technology, TTS synthesis, language maintenance and preservation, Irish, Gaelic.

1. INTRODUCTION

It is widely predicted that between 60 to 90% of the world's some 6900 languages may become extinct within the next hundred years [2]. Many factors are contributing to this process. Among these is the globalisation of culture, accelerated by the rapidly developing web and communication technologies, which are feeding the trend towards dominance of a limited number of world languages. The field of phonetics has always played an important role in documenting endangered languages. As pointed out further in [2] documentation and revitalisation activities tend to go hand in hand: the documentation efforts of many, if not all, phoneticians are prompted not only by the desire of archiving for posterity, but also of contributing to the maintenance of the language.

It is striking that some of the same technological and globalisation trends that are contributing to the loss of languages can potentially be harnessed to stem this erosion. In this paper, we discuss, on the basis of our experience in developing speech resources and technology for Irish, the role these can play in providing extensive phonetic and linguistic documentation of endangered languages and dialects, while simultaneously supporting the maintenance of these languages. Clearly, while every endangered language has its specific context and challenges, many of the problems are shared, and it is hoped that the experience of the ABAIR initiative for Irish may provide useful pointers for other such languages.

2. THE IRISH LANGUAGE CONTEXT

Irish is a Celtic language, which together with Scottish Gaelic and Manx (now extinct as a first language), traces itself to the Q-Celtic branch, spoken by the Celtic people who came to Ireland about 500 BC. Classified by UNESCO as being in its "definitely endangered" category [3], it is spoken as a community language only in the Gaeltacht, i.e., Irish speaking regions, illustrated in Figure 2. Even in these areas, the size of the Irish speaking population and the rate of transmission to the next generation is diminishing. It is estimated that only 24% (23,175 persons) of those who live in Gaeltacht areas speak Irish on a daily basis outside the education system [4]. Given that this Irish speaking population is dispersed among the different Gaeltacht regions (see Figure 2), the issue of critical mass is even more acute. There are few, if any monolingual speakers.

On the more positive side, and unusually for an endangered language, Irish enjoys considerable State support. It is recognised as the first national language in the Republic, and has since 2007 had official status as an EU language. It is a compulsory subject of study for all pupils attending primary and second level schools in the Republic. Reasonably large numbers of families speak Irish families in Dublin, Belfast and other urban locations. Outside of the Gaeltacht, there is also a growing demand for Irish-medium education, particularly at primary level, and a consequent growth in the number of such schools. Thus, the future maintenance of the language will depend, not only on its transmission as an L1 in the Gaeltacht, but also on the effectiveness with which it can be transmitted as a second language outside the Gaeltacht through education.

Irish language education presents many challenges. One difficulty is the lack of exposure to native

speaker models for the learner: most teachers are non-native speakers. The acquisition of pronunciation is particularly problematic. Central to the sound system of Irish is the contrast of palatalised and velarised consonants [5] and many grammatical processes, such as the formation of plurals, and the marking of case exploit this contrast (see examples (a & b) in Figure 1. Further morphophonemic processes involving alternations of the initial consonants of words (termed mutations) are also pervasive and carry important grammatical functions (e.g., c, d and e in Figure 1).

Figure 1: Use of palatalised vs. velarised contrast and of initial consonantal alternations in Irish grammatical processes.

	Transcription	Orthography	Translation	Grammatical Information
(a)	/bˠ a: **dˠ**/	Bád	Boat	Nominative case
(b)	/b a: **dʲ**/	Báid	Boats	Plural (or genitive sg.case)
(c)	/ə **bˠ** a: dˠ/	a bád	her boat	Unchanged
(d)	/ə **wˠ** a: dˠ/	a bhád	his boat	Lenition mutation
(e)	/ə **mˠ** a: dˠ/	a mbád	their boat	Eclipsis mutation

These aspects are often only partially acquired: a recent study [6] demonstrated that urban (non-Gaeltacht) speakers had a poor mastery and unstable realisations of these kinds of phonological and morphological processes, compared to their Gaeltacht counterparts. Failure to produce as appropriate the many realisations of root forms (as illustrated for the word 'bád' *boat* in Figure 1) impacts on every level of the linguistic system.

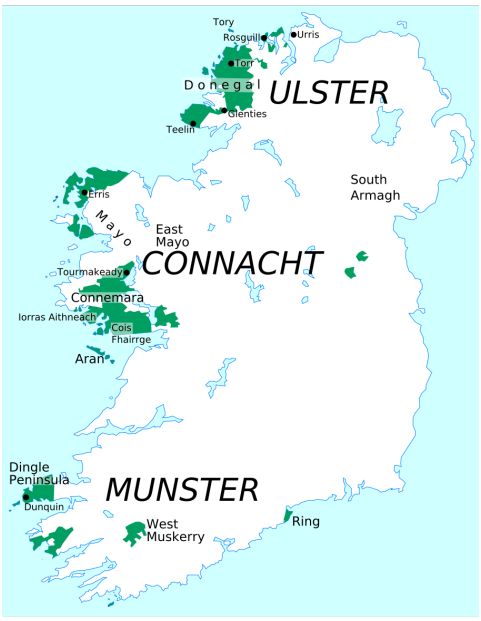
The problem is compounded by the archaic and rather opaque writing system of Irish, which presents a major stumbling block for many learners. Mastering the mapping from phoneme to grapheme is not easy if a learner has not grasped the fundamental phonological and morpho-phonological processes of the language. The problem is exacerbated when, as so often for Irish, the learner has limited access to a native speaker model.

There are other issues. As with many minority and endangered languages, there is no spoken standard dialect. Rather, there are three main dialects (see Figure 2). Within these three main dialects, one finds smaller dialect pockets, and these are particularly under threat of disappearing.

For Irish language education, learners' attitudes and motivation can be an issue. As pointed out by [7], endangered languages suffer by not having the attractive media content available to major world languages. It has been frequently observed that the lack of modern resources in Irish language teaching makes it more difficult to connect to the younger generation of the digital age [8].

In the following sections we will describe an ongoing research initiative (ABAIR) which targets documentation/resource building along with TTS development. It has so far yielded often unexpected dividends: indications are that these developments are helping to alleviate some of the problems under discussion here, and suggest a positive role in language preservation and revitalisation.

Figure 2: Official Gaeltacht regions in Ireland.



3. ABAIR: DOCUMENTATION AND RESOURCE BUILDING

The ABAIR initiative has multiple goals. The major ostensible goal is the development of text-to-speech synthesis systems for Irish. But this is not intended to be simply the provision of a piece of technology, attractive as that would be. Rather, a prime motivation for developing TTS for Irish is that it was seen as a way of documenting the Irish dialects. TTS systems depend on the availability of phonetic and linguistic resources, many of which were not available for Irish. Consequently, ABAIR aims to develop these resources so as to maximise their downstream applications and their usefulness for the maintenance and revitalisation of Irish.

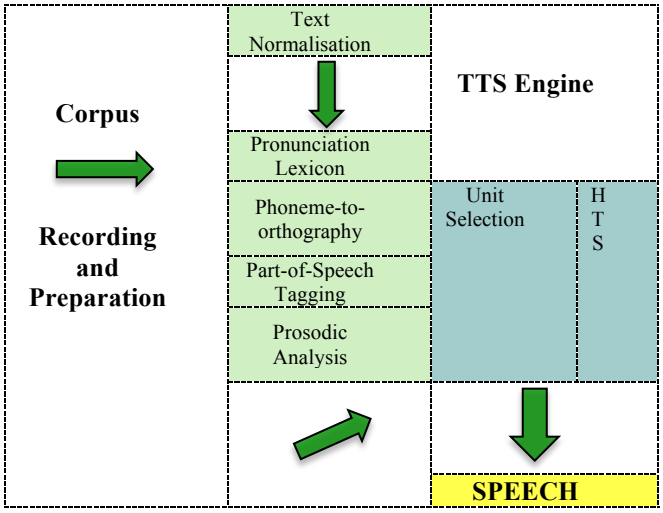
As there is not a spoken standard dialect of Irish, the need to cater for multiple dialects was acknowledged from the outset. This consideration guided many of the initial choices made in how the TTS and resource building was approached.

Figure 3 illustrates resources that are used in the building of a TTS system. In addition to annotated speech corpora (left panel), the boxes in the central column show the phonetic, phonological and linguistic resources that are required: pronunciation lexica; rules specifying the mapping between the phonemic and orthographic forms; part-of-speech tagging; and prosodic descriptions. Text normalisation rules provide for the expansion of symbols etc. This includes, for example, the rules for the realisation of numerals in different contexts, something which is quite complex in Irish.

These components, needed for TTS building, provide an opportunity to effectively document many different aspects of a language's structure, in ways that differ somewhat from the traditional forms of linguistic documentation. Depending on the language, some resources might already be available (e.g., in our case, the part of speech tagger [9]). However, even where specific resources appear to be available, the needs of TTS can highlight gaps. For example, in the case of Irish, the fact that there was already a pronunciation dictionary available [10] would suggest that this resource was already in place. However, this dictionary provides 'standardised' forms, which draw on forms taken from the three main dialects. As such it does not reflect the speech of any existing speaker, and could not be used for current speech synthesis techniques, which rely on recorded corpora from an actual speaker.

Current TTS systems have been mostly developed for major world languages and target large commercial markets. In such modern, commercially driven environments, components such as the letter-to-sound rules are usually derived by the use of statistical, machine-learning techniques. Statistical approaches may help speed up the technological development, but the solutions are in the form of a 'black box' which is impenetrable. For our purposes, a knowledge-based (handwritten) rule system was preferred for a number of reasons. From our perspective, the letter-to-sound rules are a system we need to understand and to maintain in a transparent form. Although knowledge-based rule-building is a slower process, it generates a reusable resource that allows for incremental extension and refinement. The choice was also relevant to the multidialect goals of ABAIR: it made sense to develop rules for the first dialect in a modular fashion that would facilitate the development of rules for further dialects. When working on the first dialect, we differentiated between the 'common core' ruleset, which would work for all dialects, and the 'local' rules pertaining to the specific pronunciation of that dialect. This effectively means that in developing rulesets for a new dialect the common core was already available and it sufficed to specify the local ruleset. Furthermore, a transparent ruleset has other downstream applications, such as the devising of explicit spelling rules for the learner, an important consideration for the endangered language of (more on this below).

Figure 3: Resources used in TTS synthesis



Developing phonetic and linguistic documentation in the framework of a technology (here TTS) serves to (i) extend the descriptive paradigms we have traditionally worked with, and to (ii) broaden of the domains of research, requiring us to focus particularly on the gaps in our knowledge. In our case, it has motivated cross-dialect prosodic research [11,12,13 and 14] as well as investigations of segmental phonetic/phonological divergences among the dialects, through the formulation of their letter-to-sound rules and their pronunciation lexica.

4. ROLE IN LANGUAGE PRESERVATION

The ABAIR experience illustrates how speech technology (here TTS), and the phonetic and linguistic resources that underpin it can contribute to the preservation of endangered languages.

4.1. Direct use of technology and resources

Synthetic voices for the Ulster and Connaught dialects are available on a public website [1] and the Munster dialect will soon be added. A textbox allows the user to type or copy in text and have it read aloud in the dialect of choice. Additionally the site gives access to facilities based on the individual resources. For example, a 'phonetisation' option provides the phonetic transcription corresponding to the text; users can also access grammatical information provided by the part-of-speech tagger.

The immediate reaction to the website was enormously positive, and demonstrated that globalisation can work both ways for the endangered language. The site revealed a (to us) hidden source of language strength: a global community of users/learners. From the outset, it was accessed across the globe, with particularly large numbers from the US and Britain. Feedback underscored that for the majority of users, being able to 'finally' know how to pronounce the written word was paramount. In Ireland it is widely used by school learners and their parents/teachers. The TTS appears to serve as a portable native speaker, helping to overcome the problem of access to native speaker models, and the barrier of the complex orthography.

4.2. Tailoring for specific purposes/user groups

Apart from their direct use, further applications are exploiting these facilities, extending their uses and their user-base. The plan is to make such further applications available on the website, building a virtual resource centre for the language.

CabairE is being developed as a specific literacy and pronunciation training tool, aiming to exploit the linguistic resources (and the voices) to make learners aware of the phonological and morphophonological processes of the language. These processes, and the regularities of phoneme-to-grapheme mapping can be subliminally shown through differential highlighting of the orthographic text. For example, the use of differently coloured letters to show the orthographic consonant/vowel strings that differentiate velarised and palatalised consonants should help make the learner more aware of the distinction, and of its orthographic realisations. As text is spoken out, the current word is highlighted (against the phrase, which is highlighted in a different colour). The user chooses the dialect, the rate of speech output and the degree of text magnification. The potential uses are open-ended, and CabairE is envisaged as a tool for teachers and materials developers, generating content which can be pooled on the website.

Screenreaders were urgently requested by the visually impaired, who previously could access online materials through English language TTS. The output was effectively a third language that had to be mastered. An NVDA screenreading facility is now available. A linked development, a **Web reader**, allow the public to hear the content of websites, online dictionaries and other e-resources.

CALL: multimodal interactive language learning games have been developed to the proof-of-concept level [15,16]. Both the synthetic speech and the gaming/design elements of these games have been tested with second level learners of Irish (N=252). Results on the whole were very positive [16,17], confirming that such modern digital resources are indeed attractive to younger learners, providing a fun painless route to serious learning. They also underscored that such materials are particularly needed for the endangered language, creating a 'virtual' native speaker environment.

Further uptake has included an Irish language social media website and small start-up companies featuring Irish content.

5. CONCLUSIONS

These resources and technologies provide powerful tools that can help address the many challenges confronting endangered languages. The potential symbiotic relationship between phonetic/linguistic documentation and technology development is maximised if developments are undertaken in such a way that they provide for future extension, refinement and multiple kinds of exploitation. These must be part of the planning process for the future of endangered languages [18].

A synthetic voice is in a way the ultimate form of documentation, preserving as it does a virtual native speaker of the language. A future goal for ABAIR will be the development of voices for those dialects that are near extinction. Given the modularity of the resources, and the linguistic proximity of Scottish Gaelic, it would also be interesting to extend ABAIR to these dialects. In fact some users in Scotland already use the Ulster ABAIR voice for Scottish Gaelic: orthographic adjustments have been provided. A synthetic reconstruction of Manx would also be of interest.

Speech and communication technologies are developing with an ever-increasing pace, and so also are the opportunities for their positive exploitation for the endangered language. Being able to harness this potential is likely to be one important determinant of their survival.

6. ACKNOWLEDGEMENTS

The 'ABAIR initiative' covers a number of related ongoing and past funded projects which have supported different aspects of the work reported. These include: ABAIR (An Roinn Ealaíon, Oidhreachta agus Gaeltachta); CabairE An Chomhairle um Oideachas Gaeltachta & Gaelscolaíochta (COGG); NCBI-ABAIR (National Council for the Blind of Ireland) CABÓGÍN/CABÓGAÍ (Foras na Gaeilge), and WISPR (EU Interreg).

7. REFERENCES

[1] "ABAIR - An Sintéiseoir Gaeilge." [Online]. Available: www.abair.ie. [Accessed: 19-Jan-2015].

[2] Romaine, S. 2007. Preserving Endangered Languages. *Language and Linguistics Compass 1/1-2*. 115-132.

[3] 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

[4] 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

[5] Ní Chasaide, A. 1999. Irish. In *Handbook of the International Phonetic Association*, Cambridge: Cambridge University Press, 111-116.

[6] Ó Broin, B. 2014. New urban Irish: pidgin, creole, or bona fide dialect? The phonetics and morphology of city and Gaeltacht speakers systematically compared. *J. CeltLing 15*. 24, 69-91.

[7] Edwards, V., Pemberton, L., Knight, J., and Monaghan, F. 2002. Fabula: A bilingual multimedia authoring environment for children exploring minority languages. *Language Learning & Technology* 6(2), 59-69.

[8] Judge, J., Ní Chasaide, A., Ní Dhubhda, R., Scannell, K. P. and Uí Dhonnachadha, E. 2012. *The Irish language in the digital age*, Springer.

[9] Ó Baoill, D. (ed). 1986. *Foclóir Póca.* An Gúm, Dublin.

[10] Uí Dhonnchadha, E. 2010. Natural language processing tools: Developing a part-of-speech tagger and partial dependency parser for Irish. Saarbrücken: LAP Lambert Academic Publishing.

[11] Dorn, A., O'Reilly, M. and Ní Chasaide, A. 2011. Prosodic signalling of sentence mode in two varieties of Irish (Gaelic), *Proc. 17th ICPhS*, *Hong Kong*, 611- 614.

[12] Dalton, M. and Ní Chasaide, A. 2006. Tonal alignment in Irish Dialects, *Language and Speech* 48(4), 441- 464.

[13] Dalton, M. and Ní Chasaide, A. 2007. Melodic alignment and micro-dialect variation in Connaught Irish. In: Gussenhoven, C. and Riad, T. (eds) *Tones and Tunes: Studies in Word and Sentence Prosody, Vol. 2*, Berlin, Mouton de Gruyter, 293-315.

[14] O'Reilly, M., Dorn, A. and Ní Chasaide, A. 2010. Focus in Donegal Irish (Gaelic) and Donegal English bilinguals, *Proc. 5th International Conference on Speech Prosody*, *Chicago, Illinois*, 4 pp.

[15] Ní Chasaide, A., Ní Chiaráin, N., Wendler, C., Berthelsen, H., Kelly, A., Gilmartin, E., Ní Dhonnchadha, E. and Gobl, C. 2011. Towards personalised, synthesis-based content in Irish (Gaelic) language education. *Proc. ISCA Special Interest Group link (SIG) on Speech and Language Technology in Education (SlaTE)*, Venice, 4 pp.

[16] Ní Chiaráin, N. and Ní Chasaide, A. 2014. Evaluating text-to-speech synthesis for CALL applications, *Proc. International CALL Research Conference*, *Antwerp*, University of Antwerp, 104-110.

[17] Ní Chiaráin, N. 2014. *TTS synthesis in CALL: The Development and Evaluation of Irish Language CALL Platforms*. Unpublished doctoral dissertation, Trinity College, Dublin.

[18] Government of Ireland, 2010. *20-year strategy for the Irish language 2010-2030*. Retrieved from http://www.ahg.gov.ie/en/20-YearStrategyfortheIrishLanguage2010-2030/Publications/20-Year%20Strategy%20-%20English%20version.pdf