Developing a Spoken Corpus and a Synthesiser for Irish (Gaelic)

**John Wogan, Brian Ó Raghallaigh, Áine Ní Bhriain,
Eric Zoerner, Harald Berthelsen, Ailbhe Ní Chasaide, Christer Gobl**

Phonetics and Speech Laboratory, School of Linguistics, Speech and Communication Sciences
Trinity College Dublin, Ireland

{woganj, oraghalb, anibhri, zoernee, berthelh, anichsid, cegobl}@tcd.ie

Abstract

As one of the 'lesser resourced' languages, Irish (Gaelic) shares the disadvantage of other minority languages in lacking many of the resources that would be needed for the development of speech and language technology. This paper describes some of the linguistic, technical and practical difficulties presented in trying to put in place annotated corpora and resources for speech synthesis, and outlines how the developments took account of these challenges. Although the research is focussed on the provision of specific resources for a single dialect, the need for the long-term perspective is emphasised and the need to ensure maximum reusability of resources.

1. Introduction

Irish (Gaelic) shares the disadvantage of other minority languages in having almost no provision for the emerging speech and language technology, such as machine translation, speech synthesis and recognition, etc. Although these technologies are rapidly advancing and increasingly well understood, harnessing them for a new language requires many prior resources that are often lacking.

This paper deals with speech technology development, and thus relates to this workshop in the broader setting of what would be required for speech-to-speech machine translation. An important first step in the provision of speech technology for Irish is to develop a text-to-speech system. In principle, it should be a relatively straightforward matter to deliver such a system. In practice however, there are many perquisites that need to be in place and linguistic and practical issues that have to be resolved to enable the development of a synthesis system. Today's commercial-grade synthesis systems for the 'major' languages are based on prior research and extensive spoken corpora, which have been fully annotated at the phonetic level, as well as pronunciation lexica, letter-to-sound rules, and often also sophisticated models of the prosodic and segmental features of the language. Thus, if one is not to sacrifice the quality of the eventual system, one needs to invest considerable time and effort to put these resources in place, either from scratch or by or adapting and upgrading existing materials.

These considerations informed our targets for Irish, within the Welsh-Irish project WISPR, funded by the EU-Interreg IIIA programme (www.tcd.ie/CLCS/phonetics/projects/prosody.html). Our objective was to develop a spoken corpus of Irish as a basis for synthesis, and to develop in parallel some of the other prerequisites for a text-to-speech system. Within the project, parallel work has also been carried out on Welsh, but in this paper we will describe the problematic issues needed to be considered for the Irish developments, and how during the development work the strategies adopted were geared towards extenuating these difficulties. Throughout, the emphasis has been less on the single application at hand, as on building a solid basis for the long-term development.

2. The situation of Irish

Irish is a Celtic (Goidelic) language spoken in Ireland. The Constitution of Ireland accords Irish the status as the first national language of the State. Under the Good Friday Agreement, Irish received formal recognition in Northern Ireland. As of 2007, Irish will also be recognised as an official language of the European Union. Despite the status afforded the language, the population of native speakers is small and decreasing, and there is little commercial incentive to develop speech technology resources.

Speech technology is particularly crucial for minority languages such as Irish whose future is precarious, and could contribute importantly to their preservation. Furthermore, as this technology is becoming increasingly crucial for education and access, particularly for people with disabilities, speakers of minority languages are becoming even more marginalised. Blind users of Irish have no access to electronic material in the language, and there is currently no way of providing synthesis-based communication devices to those that require them. In the pedagogical domain, speech technology could increasingly play a role in enabling the teaching and learning of the language. In an increasingly technological society, it would facilitate its widespread use among the broader Irish population. The lack of speech technology facilities in a minority language given their increasing availability and widespread use in the competing majority language will undoubtedly impact negatively on the sustainability of the former.

The difficulties that present for speech technology development may stem from a variety of sources. Many of the issues are specifically linguistic, such as the inherent complexity of the sound structure, the opaque orthographic system. Some others have to do with the deficit in the provision of necessary prerequisites: e.g., the lack of prior corpus collection, unavailability of resources such as pronunciation lexica (or in our case, lexica appropriate to the task), the availability of suitable prosodic and segmental analyses of the dialect, which are adapted to the needs of the envisaged developments.

Other issues arise from the present or historical context of the language. The fact that there is no standard dialect of modern Irish obliges us to consider long-term multi-dialect provision (see next section). Code switching

between Irish and English also needs to be borne in mind: as most speakers of Irish are bilingual, and as English has such a dominant presence, code switching is a fairly typical feature for many native speakers, and embedded English words and phrases may be common in certain kinds of texts. Thus to deliver eventual Irish synthesis, we considered it wise to anticipate a future need for our synthesiser to also be able to 'speak' Irish English. Provision was therefore made for this in planning the corpus collection.

In the following sections we outline some of the major challenges to the intended development, and describe how the eventual research attempted to take account of them.

3. Reusability of existing resources: the challenge of dialect

There is no standard spoken dialect of Irish. There are approximately 150,000 native speakers living in pockets spread largely on the western fringes of the island (see Figure 1). The Gaeltachts (Irish speaking areas) can be broadly divided into three dialects which largely coincide with the provinces, i.e. Ulster (Donegal), Connaught (Connemara and Mayo) and Munster (Kerry, Cork). Any of the three main dialects would be an equally acceptable choice in the initial development of resources.

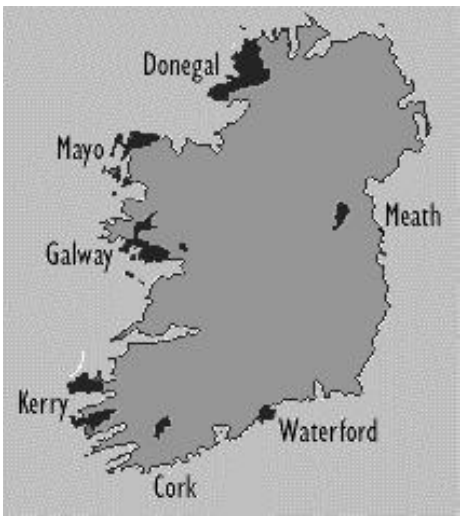

Figure 1: Map showing in black, the main regions where Irish is spoken.

The Donegal dialect was chosen for this project. From the outset, it was clear that whatever the initial dialect chosen, our long-term programme must encompass all three main dialects. Indeed, the Welsh-Irish grouping envisage a future extension of this research to the other Celtic languages and would furthermore strive to streamline methodologies, to enable them to be of wider use for similar developments with other minority languages. Thus, while the work is centred on a single dialect, many of the decisions made were at least partially guided by these long-term aspirations, and aimed to ensure reusability of resources and the establishment of an infrastructure for future developments.

4. Phonetic/phonological issues

For the task of speech synthesis, it is necessary to phonetically annotate a speech corpus. Before this activity can be undertaken, one must establish the system of contrasts used by the chosen speaker. While there is a long traditional of dialect studies of Irish, many of the studies based their analyses on the speech of older informants. As such, these resources may not represent the present day facts of the particular dialect. Questions arise in Irish dialects particularly about the inventory of laterals and nasals, and sometimes about the precise number of vocalic contrasts. This necessitates careful analysis of the chosen dialect as well as of the idiolect of the chosen informant.

The complexity of the phonetic and phonological systems of Irish presents particular challenges for corpus collection. Irish is complex from both phonetic (Ní Chasaide, 1999) and phonological (Ní Chiosáin, 1991) perspectives. Irish, possessing between 55 and up to 65 contrasts, depending on dialect and on the phonological approach adopted. A spoken corpus of any language needs to contain adequate coverage of all possible diphones, in as varied environments as possible. Consequently, for Irish, the corpus needs to be quite large if one is to be reasonably confident of ensuring adequate diphone coverage.

The large phoneset required for Irish is mainly due to the contrast of palatalised and velarised segments throughout the consonantal system. The secondary articulation of consonants has a major effect on the realisation of adjacent vowels. For example, Figure 2 shows a spectrogram of the word *buí* 'yellow' [bʸi:].

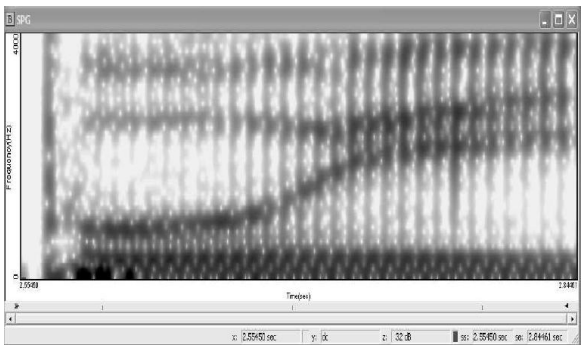

Figure 2: Spectrogram of the Irish word *buí* [bʸi:] (yellow).

The *F*2 glide from the plosive release of the velarised [bʸ] to the steady state of the vowel [i:] is particularly noticeable, and this glide needs to be appropriately segmented as being part of the vowel. Such transitions are common when the vowel quality is particularly different from the (secondary) consonant quality, i.e. transitions from velarised consonants to front vowels and from palatalised consonants to back vowels. Thus, nearly all monothongs are potentially realised as diphthongs in this dialect (the degree of diphthongisation is somewhat variable across dialects). This feature of Irish presents quite a challenge to ensure full coverage of variants (see discussion below on the diphone corpus). It also poses a potential difficulty to the concatenation process in synthesis: as pointed out by Syrdal (2001), discontinuities are perceived more often in diphthongs than in monothongs.

Another issue that impinges on our efforts concerns particular gaps in prior linguistic analysis of the language. Despite the strong tradition of dialectology, all past descriptions have focussed on the segmental level and

there has been virtually no coverage of the suprasegmental level. Having a good prosody model is vital to the provision of high quality text-to-speech. This gap in knowledge of prosody is being addressed in a concurrent project, Prosody of Irish Dialects (www.tcd.ie/CLCS/phonetics/projects/prosody.html). Within WISPR, while the lack of a prior prosodic model meant that we were not in a position to prepare even a demo diphone synthesiser, we have prepared the ground for such a development in the future, when the prosodic model is sufficiently elaborated.

5. A dialect-specific lexicon

The development of a lexicon for a minority language can also prove to be problematic. The only previously existing pronunciation dictionary was *An Foclóir Póca* (Rialtas na hÉireann, 1986). Since this dictionary provides some 15,000 words phonemically transcribed, one might expect that *An Foclóir Póca* could be used to bootstrap an automatic annotation process as well as the development of letter-to-sound rules. However, the forms and pronunciations in *An Foclóir Póca* do not correspond exactly to any one of the three spoken dialects, but rather represent an attempt at establishing a *Lárchanúint,* a 'Middialect' or a standard form, which compromises between the forms of all three.

In order to develop a dialect-specific lexicon for Donegal Irish, it was decided to adapt *An Foclóir Póca* to Donegal forms. To begin with, a short (20 min) corpus was carefully hand transcribed, and the words (orthographic and phonetic forms) were used to form a mini-lexicon for the dialect. This mini-lexicon was then compared to the forms in *An Focláir Póca*, and *sound-to-sound* rules were mapped, using the WAGON tool (www.cstr.ed.ac.uk/projects/speech_tools). Normally used to generate statistically based letter-to-sound rules, WAGON was in this case used to map between two sets of phonetic forms. The output rules were then applied to *An Focláir Póca*, in stages, beginning with the most common 500 headwords, to produce a Donegal lexicon. Before being added to the Donegal lexicon, the rule-predicted pronunciations were corrected by hand, to ensure that they were indeed in conformity with Donegal pronunciations. The Donegal-lexicon eventually included all 15,000 words of *An Focláir Póca*, plus 1,000 additional words gleaned from the 20-minute corpus.

Once this process was completed, lexicon development progressed alongside the automatic segmentation of the unit selection corpus and the development of letter-to-sound rules. Automatic segmentation involves forced alignment, and to do this the lexicon must contain all the words found in that corpus. Thus all words of the corpus, not yet in the lexicon needed to be added, along with their pronunciations. The Donegal letter-to-sound rules were therefore run on these new words, hand checked for accuracy and then added to the lexicon. This process was, reiterated until all the words of the corpus were entered, giving a total of 24,000 entries.

6. Letter-to-sound rules

Different methodologies can be used to generate letter-to-sound rules. As just mentioned, statistically based rules may be generated using the WAGON tool to map the correspondences between orthographic forms and the pronunciations in the Donegal-adapted lexicon. Through reiteration, with successive versions of the lexicon, one would expect increasingly accurate letter-to-sound mapping. The results using this approach turned out to be disappointing, yielding an unacceptably high error rate in the phonetic forms predicted. It may be that this approach may not in be well suited to the orthography of Irish.

A problem with statistically based letter-to-sound rules is that the rules themselves are not available in any explicit form. They can therefore not be scrutinised and corrected. Quite apart from this, as our long-term aspiration is to develop multi-dialect synthesis, we would in principle want to develop explicit letter-to-sound rules at some stage. Ideally this should be done in such a way as to differentiate between the common core of rules that hold across all dialects, and those that pertain to particular dialects, which might be viewed as a further layer of rules. This should in principle be the most interesting approach, not only as it yields important new linguistic information (both synchronic and diachronic), but also because it maximises the reusability of the resources we develop. As discussed earlier, though the quickest route to achieving a particular short-term goal is not necessarily be the best one: strategically 'reusable' approaches are important to our long-term interests.

For all of those reasons, handwritten letter-to-sound rules were subsequently encoded for use within Festival (www.cstr.ed.ac.uk/projects/festival). These rules were based in the first instance on *An Foclóir Póca* and Ó Baoill (1986), with rule adjustment for the Donegal dialect. The results of these handwritten rules have been encouraging, and the accuracy of their output has been much better than that achieved using the statistical approach. Consequently, the handwritten rules have replaced the latter, and now form the basis of the transcribed corpora and of the demo voice that have been developed. It is our expectation that only relatively minor further adjustments of these rules will allow us to begin the transcription of other dialects.

7. Developing annotated spoken corpora

Three corpora were collected to cater not only our immediate, but also our long-term objectives. They included a large corpus aimed at the eventual provision of a unit selection concatenative synthesis, an extended diphone corpus which will eventually enable the development of a diphone concatenative synthesis system, and finally, an Irish English recording of the ARCTIC (http://festvox.org/cmu_arctic/cmu) corpus, with a view to allow code switching in our synthetic voice. All corpora were recorded with the same speaker, a young female speaker of the Gaoth Dobhair dialect of Donegal. The recording conditions were also the same throughout, ensuring compatibility of corpora for eventual use within a single synthesis system.

The primary (unit selection) corpus involved 15 hours of recorded speech. Ideally, such a corpus should be designed to provide full coverage of all possible diphones of the language, in as many as possible environments, in the minimum of recording time. Clearly it was not possible at the outset to design such a corpus for Irish. Although we can figure out what coverage we need of distinct phones and can calculate how many diphones we need to cover, in the absence of any prior annotated data, or automatic segmentation facility for Irish, we could not

calculate their frequency in any given text, or ascertain which were lacking. Thus, a large recording was in fact required: as the inventory of sounds is large and their interaction complex, the larger the corpus the greater the likelihood of all necessary sounds being captured along with their combinations in different contexts.

The texts for the unit-selection corpus were intentionally chosen from writings from this locality. This was deemed important, as it was feared that texts with forms and structures from other dialects might trigger dialect switching on the part of the reader, something that would introduce inconsistencies into our corpus. The novels of the Donegal author Séamus Mac Grianna (Máire) were the primary source. These were not available in electronic form, and were therefore scanned. However, as there is no optical character recogniser for Irish, the scanning resulted in numerous errors, and hand-correction was required.

The diphone corpus was recorded with a view to the future development of a diphone synthesis system, when a full prosodic model for the dialect has been elaborated. As the diphone corpus was recorded with the same voice and recording conditions as the unit selection corpus, it was also the intention to use it to extend the latter. By combining them, we wished to ensure complete coverage of all occurring sound sequences. This should prevent catastrophic failures of the eventual unit selection synthesis system that could conceivably arise from gaps in the coverage of the larger corpus. It should also be noted that for the purposes of building a diphone synthesiser, the large unit selection corpus will allow us to extend the diphone corpus further, by extracting sound combinations in more prosodic contexts.

Traditionally, diphones are recorded using nonsense words containing the diphones of interest, e.g. [daʃaʃa], is intended to capture the [ʃa] diphone and the [aʃ] diphone. These are elicited in citation form. We decided however against a 'citation form' diphone recording, given our interest in combining the two corpora. If citation-elicited diphones were concatenated with materials originally from the unit-selection corpus, it was feared that they would be temporally and prosodically out of kilter. To minimise this potential problem, diphones in the enhanced diphone corpus were elicited in sentence frames.

The diphone corpus recorded in WISPR was substantially enriched beyond what is usually included in a diphone set, to take account of the complexity of the sound system of Irish. It includes, to begin with, a full set of cross word-boundary diphones. It also includes Consonant$_1$ – Vowel – Consonant$_2$ sequences, where all C_1V sequences were elicited in contexts where C_2 was either palatalised or velarised, so as to allow for the very different vowel allophones that arise in these contexts. Likewise for VC_2 diphones, they were also elicited to allow versions where C_1 was either palatalised or velarised. Syllables were also recorded to include clusters of the form CCV, CCCV, VCC. Although a minimalist approach would suggest 55 phonemes and about 3,000 diphones for Irish, the enriched diphone corpus amounted to over 11,500 units.

The third corpus was an Irish English recording of the ARCTIC corpus, a compact corpus designed to yield coverage of the phonemes of English. This was to allow for code switching which is common in Irish speech. Although less prevalent in texts, it is nonetheless frequent enough in texts which purport to be representations of true daily conversational styles.

8. Conclusions

Using the corpora and the resources developed under WISPR we have put together a first demo of an Irish speaking synthetic voice, using the Multisyn voice building facilities in Festival. To this is added an Irish English voice, based on the ARCTIC corpus.

We would emphasise that neither the corpora nor the synthetic voices are in any way complete. Many aspects of a full Irish text-to-speech system remain to be done (e.g., tokenisation, implementation of a prosody model). The Irish corpora will also require more work to eliminate errors in the segmentation. These segmentation errors are clearly highlighted in the demo Irish voice.

Nonetheless, the demo voice also amply shows that high quality text-to-speech is well within our reach. Clearly, the work we report on here is, we hope, the beginning of an extended programme to provide speech technology facilities for the dialects of Irish. By extending the Welsh collaborations fostered by WISPR, we aspire to collaborations involving other Celtic languages, including Scottish Gaelic and Manx (near relatives of Irish) as well as Breton and Cornish (near relatives of Welsh). It is also our hope that our experiences and solutions to the linguistic and resource difficulties encountered, will be of use to others who envisage similar work on their languages. The most important message, perhaps, is to take the long view. To maximise the return on our efforts, we need to ensure that the research done provides not necessarily the *shortest* route towards the development of a specific application, but prioritises rather the establishment of an infrastructure for further developments.

9. Acknowledgments

This research was funded by the EU Interreg IIIA Community Initiative Programme.

10. References

ARCTIC, http://festvox.org/cmu_arctic/cmu

Festival, http://www.cstr.ed.ac.uk/projects/festival

Ní Chasaide, A. (1999). Irish. In *The Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press, pp. 111–16.

Ní Chiosáin, M. (1991). Topics in the Phonology of Irish. Ph. D. dissertation, University of Massachusetts.

Ó Baoill, D. P. (1986). *Lárchanúint donGhaeilge*. Dublin: Institiúid Teangeolaíochta Éireann.

Prosody of Irish Dialects, IRCHSS-funded, http://www.tcd.ie/CLCS/phonetics/projects/prosody.html.

Rialtas na hÉireann (1986). *An Foclóir Póca*, Dublin: An Gúm.

Syrdal, A. (2001). Phonetic effects on listener detection of vowel discontinuities. In *Proc. of Eurospeech* 2001, pp. 979-982.

WAGON, http://www.cstr.ed.ac.uk/projects/speechtools

WISPR, Welsh and Irish Speech Processing Resources, EU/Interrreg IIIA project, 2004-2005, http//:www.tcd.ie/CLCS/phonetics/projects/prosody.html