Creating an ongoing research capability in speech technology for two minority languages: experiences from the WISPR project

*Briony Williams, *Delyth Prys and **Ailbhe Ní Chasaide

*Canolfan Bedwyr, University of Wales, Bangor, UK
**Centre for Language and Communication Studies, Trinity College, Dublin, Ireland
b.williams@bangor.ac.uk, d.prys@bangor.ac.uk, anichsid@tcd.ie

Abstract

This paper reports on efforts to set up a research capability in speech technology for two minority languages (Welsh and Irish), where the focus is on ensuring that this capability will outlive the project that provided the initial impetus (the WISPR project). The existing situation at the start of the project is summarized, together with the challenges (setting up plant and equipment from scratch, training researchers in specialised skills). Some innovative strategies are examined: remote working by researchers with regular intensive on-site work weeks, and the use of Internet technology as a medium for collaborative work patterns. It is hoped that these experiences may be useful to researchers in other minority languages who wish to set up a similar research capability.

1.Introduction

The WISPR project (Welsh and Irish Speech Processing Resources) [1] is an EU-funded project for developing speech corpora and text-to-speech synthesizers for Welsh and Irish, two lesser-used languages belonging to the Celtic language family, and spoken in parts of Wales and Ireland respectively.

The WISPR project is the first such project for these two languages, and hence a great deal of infrastructure work has been necessary. It is hoped that this investment in the undergirding infrastructure may outlive the WISPR project (which terminates at the end of 2005).

1.1.The existing situation for Welsh and Irish

Previous to the WISPR project, a diphone-based text-to-speech synthesiser had been developed for Welsh at the University of Edinburgh [2, 3]. This system was later ported to the "Festival" speech synthesis development framework by Alan Black [4]. Resources developed for this initial system included a small phonemic lexicon and hand-written letter-to-sound rules.

A further project then developed a small Welsh speech database, with a smaller subset hand-annotated at a fine phonemic level and at higher linguistic levels [5].

Meanwhile, at the University of Wales, Bangor, work was proceeding on a spelling checker for Welsh ("CySill"), together with a text corpus and bilingual Welsh-English online dictionary [6]. Extensive work was also carried out (and continues to the present day) on the standardisation of Welsh technical terminology, for use by professionals and school students. This work has evolved into a Welsh terminology centre, "e-Gymraeg", at Canolfan Bedwyr ("The Bedwyr Centre"), at the University of Wales, Bangor. The terminology and lexicographic work also provided input to the EU's "MELIN" project (Minority European Language Information Network") [7]. This project synchronized existing and new resources for Welsh, Irish, Catalan and Basque, creating a master online multilingual dictionary.

Despite such extensive developments in lexicography, terminology and text databases, however, there had been no work in Welsh speech technology located in Wales itself.

In the case of Irish, there has been no comparable development of text-based resources and tools, and no ongoing infrastructure outlasting a single project. In the case of speech technology, there has been no known Irish speech synthesiser or speech database. The result is that there are no existing letter-to-sound rules for Irish, and no existing phonemic lexicon in online form.

However, in recent years an informal group has formed, known as the "Irish Speech Group", to facilitate networking and discussion, and to work together to initiate research into Irish speech technology [8]. Its website concisely states the problem: "Previously, there have been a handful of research projects in Irish Universities which have each had to face the lack of resources and create their own tools suitable for the experimental issues at hand. Funding of individual and unconnected research projects has led to a dilution of effect and a squandering of opportunity for collaboration."

1.2.Steps towards a solution: the WISPR project

The WISPR project aims to be the first step in the solution to this problem. With the focus on common standards and procedures (to avoid duplicating effort), it aims to ensure that essential research is no longer lost or wasted. The project's longer-term goals include the following:

- Open-source software will be used wherever possible. In the case of text-to-speech synthesis, the platform used is the "Festival" development and run-time system [9]. Further documentation and development scripts from the Festvox project have also been used [10].
- It is important that the software produced by WISPR should be available at no cost to users. It is then the task of commercial resellers to incorporate this technology into proprietary products for sale.
- It is also important to ensure cross-platform compatibility (Windows, Mac, Linux/Unix, etc) in order to ensure the maximum possible number of users.

Reaching these aims has involved the WISPR project in several logistical challenges. Researchers in minority languages are faced with the same challenges as researchers in major languages (lack of funding, shortage of skilled researchers), but to an even greater degree. In addition, there are further challenges for work on minority languages.

The remainder of this paper will outline first the specific challenges faced, and then the steps adopted by the WISPR researchers to overcome or mitigate those challenges.

2.Challenges for speech technology research in minority languages

2.1.Researchers: training and funding

Compared to work on major languages, there is a chronic shortage of research funding for speech technology in minority languages, even in the case of relatively favoured languages such as Welsh and Irish. Some of the funding sources which are available may not have funds at the level required for speech technology research, where plant and equipment can be a greater issue than for text-based work.

In addition, there is a chronic problem of lack of continuity of funding (especially for languages which lack official state encouragement, such as Breton), and this leads to wasteful duplication of work, and even loss of work.

However, it is not only a matter of funding: for many minority languages, there is an acute lack of researchers with the necessary technological skills who also speak the language. For example, in the 1990's a promising research proposal to develop a speech synthesiser for Scottish Gaelic was approved by a UK Research Council, but had to be abandoned after the failure to locate a suitably-qualified speech researcher who also knew Gaelic.

2.2.Ownership of the field

A problem common to many minority languages is the fact that responsibility for speech and language technology (SLT) may either be split between different bodies (resulting in waste and inefficiency), or not assigned to any institution, especially where state support for the language is lacking. In such a case, there is a need for an organisation that will take responsibility for driving future developments in the field, and serve as a guarantor of continuity across time.

2.3.Loss of previous work and expertise

In the case of minority languages, it is very possible for past work to be lost, or nearly lost, when the researcher has to abandon it. In addition, scarce expertise has been lost (or nearly lost) when key researchers have been forced to leave the research field.

In the recent past, this has happened with a Festival-based Scottish Gaelic diphone-based synthesiser [11]. The materials for this work seem to have been lost, and only the MSc thesis remains (though without the diagrams). The researcher concerned no longer works on Gaelic.

The raw materials of the "SpeechDat Cymru" corpus of Welsh telephone speech [12] are available through ELRA (see http://www.elra.info). However, the expertise of the principal researcher (and only person previously known to attempt Welsh speech recognition) had been lost to the field, as he had been constrained to pursue an alternative career. Fortunately, he is now active in the WISPR project.

Similarly, the materials for the Welsh diphone-based TTS synthesiser mentioned earlier ([2], [3]) had been lost, and the researcher concerned was forced to pursue an alternative career – however this researcher is likewise now a member of the WISPR project.

This loss of software, tools and expertise represents an unacceptable waste of resources and a severe obstacle to progress. While this fact is of course true for work in major languages also, the problem is particularly acute in the case of minority languages, where there is no "critical mass" of research effort, and where the loss of just one researcher can represent a major setback for work in that language.

2.4.Geographical dispersion

Another challenge in the case of minority language speech research concerns the geographical dispersion of researchers. This is particularly true of the Welsh part of the WISPR project, where the lack of long-term prospects has made it unrealistic for three of the researchers to re-locate to Bangor in North Wales.

This has led to the situation where three staff members (all part-time on WISPR) are based at Bangor, while others are located as follows: one full-time researcher in Scotland; one full-time researcher in South Wales; and one part-time researcher in Bristol.

A working pattern has been adopted whereby the remote workers gather in Bangor for one week in every month, for an intensive week of work, and occasional WISPR-wide project days with seminars.

There has been a need to develop remote methods of working together, using such technology as email, Internet-based discussion lists, instant messaging (Internet text conversations), a shared file area accessed using FTP, and of course the telephone. Other avenues to be explored include the possibility of desktop-based videoconferencing (using NetMeeting or similar), a shared server, possibly a Virtual Private Network. However, all these are necessary substitutes for face-to-face working, and can tend to slow down the interactions between researchers.

Some specialized tools cater for the needs of the tele-researcher. These include such facilities as borrowing rights at university libraries (the "SCONUL Research Extra" scheme), online access to journals (the "Athens" scheme), and resources such as CiteSeer (http://citeseer.ist.psu.edu), the scientific literature digital library. Some of these resources are also available to researchers in other countries, and could form a starting-point for other similar teams.

However, for each remote worker, there is the need to maintain computing equipment at their own expense. With little computing support, a tele-researcher needs to be an advanced computer user, able to carry out such tasks as: install Linux, compile Festival, maintain backups and OS patches, transfer files using FTP. All of this represents a significant overhead in time, effort and complexity.

However, there are some compensations: one colleague acts as the WISPR representative in South Wales, making contact with potential users and other interested parties, and giving an occasional presentation of the work.

3.Building up the research base

The challenges mapped out in the previous section have given rise to a determination to overcome then, while also laying the foundations for an ongoing infrastructure for speech technology research in Welsh and Irish. To that end, the following steps have been taken.

3.1.People

Recruiting the right people was exceedingly difficult, as there were very few people with existing expertise in speech technology for Welsh and Irish. For other minority languages, it may not be possible to locate any researchers at all in this category. In the case of people with existing expertise for English, it was likewise difficult to recruit staff. For other minority languages, it may be hard to recruit these researchers, and tele-researching would need to be explored.

In the case of people with no existing speech technology expertise, the emphasis has been on training. Two post-graduate students are part of the Irish WISPR team, and occasional training events are held. The student members are encouraged to visit Bangor for joint sessions when the tele-workers are present, to aid in the transfer of expertise.

3.2.Facilities and equipment

It has been necessary to build up research equipment and facilities. In the case of the Welsh WISPR team, this has been done from scratch (the Irish team, based at Trinity College, Dublin, already had the use of a recording studio).

At Bangor, a pre-fabricated recording booth has been installed, together with associated equipment (microphone, high-quality sound card, and laptop computer). In addition, specialist equipment (an electroglottograph) has been borrowed from a member of the Irish WISPR team.

The project has acquired a wide range of free specialist software. This includes the following:

- Festival and the Edinburgh Speech Tools library [10];
- JspeechRecorder for managing recording sessions [13];
- Praat software for annotating speech signals [14];
- Emu software for the hierarchical labelling of speech databases [15];
- HTK (Hidden Markov Model toolkit) for use in automatically labelling speech phonemically[16];
- Sphinx/Train for possible use in auto-segmentation [17].

The above software represents a good basis for development, and would be an obvious choice for minority languages beginning to set up a speech technology capability.

3.3.Publicity and networking

Given the fragmented nature of research in minority language technology, it is all the more important to engage in the widest possible dissemination of information about the project. This serves the dual purpose of forming links with other researchers who may be able to offer help, and avoiding the duplication of effort that might ensue.

To this end, research papers have been written and presented ([1], [18]). In addition, a printed newsletter has been circulated to user groups and other interested parties, to keep them up-to-date with developments. Also, a talk was given in March 2005 to a "Business Breakfast" in South Wales. A database of small Welsh IT companies has been compiled, as these companies are potential value-added resellers of WISPR software. Projected activities include publicity at the annual "Eisteddfod", a Welsh national cultural festival, which in 2005 takes place at Bangor.

At least as important as these formal activities, however, has been the ongoing networking, whereby the project attempts to create links with other projects and researchers (such as the Local Languages Speech Technology Initiative [19]) in the hope of sharing resources and best practice.

4.Research training

Training forms an important part of the WISPR project, and indeed for any new initiative in speech technology for a minority language with no existing tradition of such research.

4.1.On-the-job training

Informal on-the-job training is an ongoing feature of any research environment, but is particularly vital in a minority language context where there is no accumulated tradition of research "folklore". It involves such things as:

- Familiarisation with specialist software tools (for such tasks as phonetic hand-labelling, autosegmentation, recording speech, compiling and running Festival, etc).
- Coaching in best practice when writing scientific papers, for those who are new to this task.
- Supervision in specialist tasks, such as the hand-labelling of Welsh speech at the phonetic level.
- Guidance in software-related procedures, such as compiling and installing new software.

4.2.Training workshops

More formal training has been provided in the form of workshops, led by knowledgeable experts in their field.

- A WISPR training workshop was held for a few days in May 2004 at Dublin, led by one of the Festival developers.
- A workshop on Hidden Markov Models and autosegmentation was held in Dublin on two days in March 2005, led by one of the WISPR members.
- A talk was given in Bangor in April 2005, giving an overview the treatment of prosody in Festival.

It is also planned to initiate a programme of internal talks, in order to give young researchers practice in presenting their work to others, especially as the presentations may be in a language (English) that may be inconvenient for them to use.

In general, a strong emphasis on training will help minority languages to build up the necessary "critical mass" of overall expertise, and new researchers, that is needed for ongoing continuity of research work beyond the end of any given project.

5.Concluding comments

5.1.Some lessons from experience

The process of building a research infrastructure in speech technology takes longer than one might expect. Even in the case of Welsh, where the "e-Gymraeg" centre has an extensive track record of work in lexicography and terminology, the extra requirements of speech technology have necessitated a great deal of time, effort and money. In the case of a minority language with less of an existing foundation in digital language work, the task may be even more daunting. It is hoped that this account of the WISPR experience will help other researchers avoid mistakes and follow best (or at least, less bad) practice.

It has become clear that the most vital resource (and the scarcest) is skilled people. In the context of a chronic shortage of expertise for a minority language, the emphasis has to be placed on an ongoing training programme.

It is also clear that continuity of funding is just as important as amount of funding, since only continuity of funding can avoid the loss of crucial work from the field. This problem is particularly acute for minority languages.

5.2.Commercial potential

Just as vital to the ongoing health of the field is the need to build up a viable commercial speech technology capability, probably through small local IT companies. The Welsh WISPR project has keenly nurtured links with such companies, and will take active steps to assist them in including Welsh TTS in products. This may include such activities as "Open Days" for familiarization with the software, and individual assistance with the software.

The ensuing value-added products (such as a Welsh screen-reader, or a Welsh-speaking document scanning device) will be of direct benefit to Welsh-speaking end users. However, it will indirectly benefit the field as a whole, in opening up a previously untapped potential for development. Therefore, any similar project for other minority languages will need to take into consideration the importance of commercial exploitation of any software produced. This exploitation may be at the level of the "micro-business", and profits may be relatively modest, but it is nevertheless vital for the long-term viability of speech technology research.

5.3.Future aspirations

From a technical point of view, the research of the WISPR project forms a good springboard for subsequent work in speech recognition for Welsh and Irish. This is due to the initial work carried out in training HMM's for the automatic segmentation of speech.

From a more strategic point of view, the future aims for the WISPR project, and any subsequent project, certainly include the hope that these experiences might act as a model for other minority languages in setting up a speech technology capability. To this end, we aim to continue with networking activities among minority language researchers and others.

6.Acknowledgements

This project is funded by INTERREG, an initiative of the European Union to facilitate co-operation between adjacent regions. Additional funding has been provided by the Welsh Language Board.

7.References

[1] Prys, D., Williams, B., Hicks, B., Jones, D., Ní Chasaide, A., Gobl, C., Berndsen, J., Cummins, F., Ní Chiosáin, M., McKenna, J., Scaife, R., and Uí Dhonnchadha, E., "WISPR: Speech Processing Resources for Welsh and Irish". *Pre-Conference Workshop on "First Steps for Language Documentation of Minority Languages*", Language Resources and Evaluation Conference, May 2004, Lisbon, Portugal.

[2] Williams, B., "Diphone synthesis for the Welsh language", *Proc. 1994 International Conference on Spoken Language Processing*, 1994, Yokohama, Japan.

[3] Williams, B., "Text-to-speech synthesis for Welsh and Welsh English", *Proc. Eurospeech 1995*, vol. 2, pp. 1113-1116, Madrid, Spain.

[4] Black, A. and Lenzo, K., "Building voices in the Festival speech synthesis system". Carnegie-Mellon University, USA, 2000. Online at http://www.festvox.org/bsv/.

[5] Williams, B., "A Welsh speech database: preliminary results". *Proc. Eurospeech 1999*, vol. 5, pp. 2283-2286, Budapest, Hungary.

[6] Prys, D. and Morgan, M, "E-Celtic language Tools: The Latest Developments from Wales", *Proc. 6th Annual Conference of the North American Association for Celtic Language Teachers*, 2000. See: http://www.naaclt.org

[7] MELIN project: http://www.ite.ie/melin.htm

[8] Irish Speech Group: http://isg.eeng.may.ie

[9] "Festival" speech synthesis development and run-time system: http://www.cstr.ed.ac.uk/projects/festival/

[10] "Festvox" project: http://www.festvox.org

[11] Wolters, M., "A Diphone-Based Text-to-Speech System for Scottish Gaelic". MSc thesis, 1997. See abstract and text at http://citeseer.ist.psu.edu/309369.html

[12] Jones, R.J., Mason, J.S., Jones, R.O., Helliker, L., Pawlewski, M, "SpeechDat Cymru: A large-scale Welsh telephony database". *Workshop on speech and language technology for minority languages, Language Resources and Evaluation Conference*, 1998, Granada, Spain. See http://galilee.swan.ac.uk/homepages/Home/data/speechdat.htm

[13] JspeechRecorder from: http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html
[14] Praat from: http://www.fon.hum.uva.nl/praat/
[15] Emu from: http://emu.sourceforge.net
[16] HTK from: http://htk.eng.cam.ac.uk
[17] Sphinx/Train from: http://cmusphinx.sourceforge.net
[18] Williams, B., Prys, D. and Jones, D., "Speech technology in Welsh and Irish: the WISPR project". *ELRA Newsletter*, vol. 9, no. 4, Oct-Dec 2004.
[19] LLSTI: http://www.llsti.org