SpeechDat Cymru: A Large-scale Welsh Telephony Database

Rhys James Jones (1)**, John S. Mason** (1)**, Robert Owen Jones** (2)
Louise Helliker (3)**, Mark Pawlewski** (3)

(1) Speech Research Group, Department of Electrical and Electronic Engineering,
University of Wales, Singleton Park, Swansea SA2 8PP, Wales UK
(2) Canolfan Dysgu Cymraeg, 42 Park Place, Cardiff CF1 3BB, Wales UK
(3) Speech Technology Unit (MLB 3/19), BT Laboratories, Ipswich IP5 7RE, Suffolk UK

Project website [http://www.speech.cymru.org/], email [info@speech.cymru.org]

Abstract

We describe the collection of SpeechDat Cymru, a 2000-speaker speech recognition database for the Welsh language, recorded over the public switched telephone network (PSTN). It is collected as part of SpeechDat(II), an ELRA project which deals with the creation of databases in over 20 different European languages and dialects. Design issues common to all SpeechDat(II) databases are discussed, in addition to Welsh-specific considerations. Speaker recruitment, database recording and annotation are also detailed. It is our belief that with the current interest in multilingual speech recognition, SpeechDat Cymru will lead to an increased awareness of minority languages within the European speech community. It will also increase the status of Welsh in the global technology marketplace.

1 Introduction

It is estimated that there are 6,000 languages currently spoken world-wide. As many as 90% of these may become extinct during the 21st century (Pinker, 1994). The role of technology in killing or cultivating these endangered languages is an important one. Many commentators have pointed towards an imminent technology convergence, as the Internet consumes current forms of mass-media and telephone systems and computer networks begin to share the same infrastructure.

Speech technology plays a key role in technology convergence. Telephony speech recognition in English may be said to have reached maturity, and integration of speech technology and the Web (Lau et al., 1997) is an active area of current research.

While the bulk of speech research has focused on English (American and British), Chinese and Japanese, the collection of resources in other languages continues apace. In particular, the 1990s have seen the collection of a set of multilingual databases. The COCOSDA initiative, and its American and European participants LDC[1] and ELRA[2], deals with the cross-language concertation of speech-related databases.

The SpeechDat(I) and (II) projects[3] are regarded as the cornerstones of ELRA's work. They deal with the creation of telephone databases for a total of 22 European languages and dialects. Being *Polyphone*-like, the databases are mainly read rather than spontaneous. Callers are given unique prompt sheets for such items as isolated words, phonetically rich words/sentences, numbers and spellings.

The specific goal of SpeechDat(II) is to provide recordings of between 500 and 5000 speakers for each of 21 EU languages. The recordings will be used for speech and speaker recognition in fixed telephony and mobile telephony environments. SpeechDat(II) is sponsored by the EC under contract number LE2-4001.

Welsh is one of the languages collected as part of SpeechDat(II). Its roots are Indo-European and Celtic, and it is arguably Europe's oldest living language (Jones, 1992). Approximately 20% of the 2.5 million population of Wales speak Welsh: the vast majority of Welsh-speakers are also fluent in English. There is a substantial international network of Welsh speakers and an enthusiastic and growing community of learners[4].

We refer to the Welsh SpeechDat(II) collection as SpeechDat Cymru (SDCymru), Cymru being the Welsh word for the country of Wales. SDCymru is a 2000-speaker database collected over the fixed telephone network (PSTN). A demographic distribution of participants is required in terms of accent, age and sex, as described in Sections 2 and 3.

Quality assurance is inherent to all SpeechDat(II) databases. They must pass a rigorous validation exercise, which is dealt with by SPEX[5], a member of both the SpeechDat(II) consortium and ELRA. Databases are only passed if they meet the criteria laid down by the SpeechDat(II) documentation.

The database has been specified, collected and annotated by the University of Wales Swansea Speech Research Group, under contract to the SpeechDat(II) consortium partner, BT (British Telecommunications plc). It represents the world's first large-scale language resource for Welsh. We hope that our experiences in its creation will aid the collectors of speech/speaker recognition resources in other minority languages, particularly Celtic ones, for which considerations similar to those for Welsh apply.

We believe that SDCymru represents an unique opportunity to develop new technologies for the Welsh language, and to make a positive contribution to the day-to-day use of Welsh. Though designed with speech

[1] [http://www.ldc.upenn.edu/]
[2] [http://www.icp.grenet.fr/ELRA/]
[3] [http://speechdat.phonetik.uni-muenchen.de/]
[4] Cymdeithas y Dysgwyr (Welsh Learners' Society): [http://www.aber.ac.uk/~welwww/]
[5] [http://lands.let.kun.nl/spex/]

recognition in mind, it will also yield interesting results in the fields of Welsh phonetics, speech synthesis and applied linguistics. Furthermore, the profile of Welsh is significantly raised by the dissemination of a Welsh-language database among the major speech recognition companies of Europe.

2 Database design

Particular emphasis is paid to uniformity of databases across SpeechDat(II). Each database has a common specification of 40 or so items, and a small number of additional items. The language-independent design issues are dealt with in Section 2.1, and Section 2.2 considers the Welsh-specific components of the database.

2.1 Language-independent design issues

The information below deals specifically with SpeechDat(II) speech recognition databases collected over the PSTN (these are referred to as FDBs). MDBs, mobile telephony databases for the European digital standard GSM (Vary et al., 1989), and speaker recognition databases (SDBs) are also collected as part of SpeechDat(II). Information on their design is obtainable from [http://speechdat.phonetik.uni-muenchen.de/].

2.1.1 Collection overview Each caller is asked to make one free telephone call. Two thousand calls are thus made, each call and speaker being unique. Each caller is given a prompt sheet, and the majority of the 40 or so items per call are read from the sheets; only a small number are elicited spontaneously.

Table 1 below shows the structure of an FDB having between 500-4000 speakers, of which SDCymru is one. There are minor differences between the structure of the 500-4000 speaker databases and the ones with 4000-5000 speakers. Most items are read; those highlighted in grey are spontaneous.

Description		**Number per call**
application words		6
sequence of 10 isolated digits		1
4 connected digits	sheet number (5+ digits)	1
	telephone number (9-11 digits)	1
	credit card number (14-16 digits)	1
	PIN code (6 digits)	1
3 dates	spontaneous date, e.g. birthday	1
	prompted date, word style	1
	relative and general date expression	1
word spotting phrase using an embedded application word		1
isolated digit		1
3 spelled word (letter sequences)	spontaneous, e.g. own forename	1
	spelling of direct. city name	1
	real/artificial for coverage	1
currency money amount		1
natural number		1
5 directory assistance names	spontaneous, e.g. own forename	1
	city of birth / growing up (spontaneous)	1
	most frequent cities (set of 500)	1
	most frequent company/agency (set of 500)	1
	“forename surname”	1
2 questions, including “fuzzy” yes/no	predominantly “yes” question	1
	predominantly “no” question	1
2 time phrases	time of day (spontaneous)	1
	time phrase (word style)	1
phonetically rich words		4
phonetically rich sentences		9
additional application word (specific to SDCymru)		1

Table 1: Structure of FDBs having between 500-4000 speakers

The frequencies of occurrence of each item are designed to provide "full... uniform coverage of all lexical items that are most commonly associated with their productions" [page 9 of Winski (1997)]. Furthermore, there should be enough examples "to support training of isolated words wherever possible".
Phonetically rich sentence (PRS) and word (PRW) sets are developed as part of all SpeechDat(II) databases. These items should provide the following [page 18 of Winski (1997)]:

- sufficient training examples of all phoneme-like units in the language, including the rarest
- good coverage of biphones and common triphones, over the entire corpus

2.1.2 Lexicon Central to the design of the database is a comprehensive lexicon for the language being collected. SpeechDat(II) requires that a lexicon be delivered with the final databases, giving a phonetic transcription for each unique word uttered by the database's speakers.

A knowledge of the phonetic content of the PRW/PRS set is essential during database development, to ensure that criteria on phoneme frequencies are met. Before any major database specification can take place, then, a comprehensive lexicon of phonetically transcribed words must be built. The design of the Welsh lexicon is discussed in Section 2.2.2.

2.1.3 Phonetically rich sentences The phonetically rich sentences are designed to produce adequate coverage of each phoneme in the language, in as wide a variety of contexts as possible. They facilitate adequate training of phoneme models when recognition experiments are carried out on the database. 'Phonetically rich' is not the same as 'phonetically balanced'. Although phonetic balancing (i.e. mirroring the language's phoneme distribution) is advantageous in linguistic databases, it would mean that some of the least common phoneme-like units could not be adequately modelled during the development and training of automatic speech recognition systems.
For databases of less than 5000 speakers, which includes SDCymru, it is assumed that models for phoneme-like units can be trained effectively with

a minimum number of examples in the PRS set
= (number of speakers) / 10

This criterion does not apply to the rarest phoneme-like units in a language - those which have frequency less than 1% in typical usage. It was felt that applying the criterion to these phonemes would degrade the representation of the more frequently occurring ones.
Each sentence should be repeated no more than 10 times in the database, to ensure a diversity of sentences in the prompt sheets.

2.1.4 Phonetically rich words While PRSes provide good lexical and phonetic coverage of a language, they may not model the start and end silence contexts of all phoneme-like units to the same thoroughness. To circumvent this, SpeechDat(II) requires that PRWs (phonetically rich words) be included in the database. The constraint on the PRW set is

phoneme-like units to have minimum no. of examples
= (number of speakers) / 5

Each PRW should occur no more than five times in the database.

2.1.5 Demographic distribution SpeechDat(II) databases aim to have sufficient demographic coverage of their respective languages. Demographic factors must, then, be considered in the design of a SpeechDat(II) database, and a balanced corpus must be achieved. Sex and age are the main such factors.

Sex: The sex of a speaker greatly affects their voice quality. Women are known to have a higher pitch average than men, and have a more 'breathy' overall voice characteristic. According to some studies, women are also more likely to adhere to standard pronunciation.

Approximately equal numbers (45-55% each) of male and female speakers are collected for the SpeechDat(II) databases.

Age: The speaker's age also has an effect on voice quality and the pronunciation/syntax of a given speaker. Younger speakers, and particularly children, have a high pitch average.

The SpeechDat(II) age distribution of speakers is seen in Table 2 below.

Age	Midpoint	Minimum %
8-15	12	Optional
16-30	23	20
31-45	38	20
46-60	53	15
61+	78	Optional

Table 2: SpeechDat(II) speaker age distribution

2.2 Design issues specific to Welsh

2.2.1 Welsh phonetics The monophones used in the following descriptions are denoted in Table 3 below. These are SAM-PA (Wells, 1989) compliant.

SAM-PA symbol	Welsh example	
a	c**a**dw	(short vowel)
a:	t**a**d	(long vowel)
E	**e**nw	(short)
e:	p**e**th	(long)
I	c**i**nio	(short)
i:	c**i**	(long)
Q	t**o**n	(short)
o:	t**ô**n	(long)
U	c**w**m	(short)
u:	gwdihŵ	(long)
1	can**u**	(short)
1:	c**u**l	(long)
j	**i**ach	(consonantal 'i')
@	m**y**ned	
v	a**f**on	
p	**p**am	
k	**c**ael	
D	a**dd**o	
n	**n**aw	
d	**d**yn	
T	sai**th**	
r	**r**as	
p	**p**am	
K	**Ll**anel**l**i	

Table 3: SAM-PA monophones used in this paper

The main phonetic features which distinguish the linguistic regions of Wales are as follows:

- Different realisations of the 'u' and 'ŷ' graphemes in certain contexts. Below, their North Welsh realisation is transcribed as /1/ and /1:/ for short and long vowels respectively, the South Welsh realisation as /I/ and /i:/
- epenthetic vowels
- In monosyllabic words, lengthening of the initial element of the graphemes:
    - ae (realised as the long /a:1/, /a:I/; the short /aI/; or just /a:/ where monophthongisation occurs)
    - oe (long /o:I/ and, in certain Pembrokeshire dialects /u:E/ or /Ue:/; short /QI/; or monophthongised /o:/)
    - wy (long /u:1/ and /u:I/, short /UI/)
    - ew (long /e:U/; short /eU/)
    - aw (long /a:U/, short /aU/)
- realisation of 'a' and 'e' in the final syllable of a word

An introduction, in Welsh, to the above phonetic features may be found in Thomas and Thomas (1989).
SDCymru, as a database, cannot afford to model each individual dialect in Welsh, many of which are restricted to a few hundred of the language's 500,000+ speakers. The main accent areas are listed below. The target to be collected from each area is calculated from the distribution of Welsh speakers 16 years old and over, according to the 1991 census (OPCS, 1994).
The boundaries between regions are open to debate. Studies on the peripheries of these boundaries include Jones (1967) for the South-West/North boundary, Thomas (1901) for the South-West/North-East boundary, and Thomas (1993) for parts of the South-West/South-East boundary.

South-West Wales
- Target: 820 speakers (41% of total)
- Region: Pembrokeshire, Carmarthenshire, Cardiganshire, Swansea, Neath, Torfaen, Blaenau Gwent, Monmouthshire, Mid and South Powys

Main linguistic features:
- South Welsh realisation of the ‘u’ and ‘ŷ’ graphemes
- epenthetic vowels in the majority of areas
- no lengthening of the initial elements of graphemes ‘ae’/‘oe’/‘wy’/‘ew’/‘aw’: monophthongisation also occurs in the majority of areas
- certain diphones in the final syllables of words having ‘a’ as an initial element are realised as ‘e’ (e.g. ‘llyfr**au**’ is realised as ‘llyfr**e**’, /K @ v r E/, ‘cad**ai**r’ is realised as ‘cad**e**r’, /k a d E r/, ‘llenyddi**ae**th’ is realised as ‘llenyddi**e**th’, /K E n @ D j E T/)

North-West Wales
- Target: 580 speakers (29% of total)
- Region: Caernarfon and Meirionydd, Aberconwy and Colwyn, Ynys Môn (Anglesey)

Main linguistic features:
- North Welsh realisation of the ‘u’ and ‘ŷ’ graphemes
- lengthening of the initial elements of graphemes ‘ae’/‘oe’/‘wy’/‘ew’/‘aw’
- The grapheme ‘e’ is realised as the phoneme ‘a’ in the final syllable of words (e.g. ‘cap**e**l’, realised as ‘cap**a**l’, /k a p a l/)
- in final stressed syllables, the grapheme ‘e’ is realised as the phoneme ‘a’
- certain diphones in the final syllables of words having ‘a’ as an initial element are realised as ‘a’ (c.f. their realisation as ‘e’ in South-West Wales)

South-East Wales
- Target: 320 speakers (16% of total)
- Region: Port Talbot, Bridgend, Vale of Glamorgan, Cardiff, Caerphilly, Rhondda Cynon Taf, Merthyr Tydfil

Main linguistic features:
- bilingual education giving rise to the adoption of features from the South-West Wales dialect amongst 16-30 and 31-45 year-olds
- epenthetic vowels in the majority of areas
- amongst older speakers (46-60 and 61+), ‘e’ in the final syllable of words is realised as ‘a’ (see North-West Wales)
- South Welsh realisation of the 'u' and ‘ŷ’ graphemes
- no lengthening of the initial elements of graphemes ‘ae’/‘oe’/‘wy’/‘ew’/‘aw’

North-East Wales
Target: 280 speakers (14% of total)
Region: Denbighshire, Flintshire, Wrexham, North Powys

Main linguistic features:
- North Welsh realisation of the ‘u’ and ‘ŷ’ graphemes
- lengthening of the initial elements of graphemes ‘ae’/‘oe’/‘wy’/‘ew’/‘aw’
- certain diphones in the final syllables of words having ‘a’ as an initial element are realised as ‘e’ (see South-West Wales)

2.2.2 Lexicon Though a set of letter-to-sound rules (Williams, 1993) already exist for the Welsh language, it was decided to utilise a process which could aid lexicon developers in other minority languages, for which no such rules currently exist. A rapid development approach was deemed desirable.
The convention used for transcription was SAM-PA (Wells, 1989) compliant.

Rapid development of lexicon
A list of common Welsh words was taken. To enable fast grapheme-to-phoneme conversion, we exploited the fact that Welsh is a phonetic language, certainly more so than English. Hence, in Welsh, a grapheme can usually be mapped into an unique phoneme.
A script was written to realise this rough-and-ready grapheme-to-phoneme conversion. It operated in the following manner:
- identification of phoneme boundaries within a word - as well as the vowel diphones common to many languages, Welsh contains consonant clusters ('ch', 'ngh' etc) which should be realised as one phoneme.
- the conversion of the bounded graphemes into individual phoneme-like units.

Although the script uses a one-to-one grapheme-to-phoneme conversion technique, it was found to be surprisingly effective at transcribing Welsh words, and required only minimal amendments of the transcribed words. Following the strategy of Lamel and Gilles (1996), each transcription was checked manually.

Lexicon searching
In common with other Celtic minority languages, Welsh exhibits mutation. It occurs, in Welsh, for nine liquids and plosives, present as initial phoneme-like units of words. These initial phonemes can change in certain well-defined contexts. We can take the word 'cadair' (chair, /k a d aI r/) as an example. Table 4 below shows the three possible mutations of the word in three different contexts.

Mutation context	English equivalent	Mutation type exhibited
Dy gadair	Your chair (familiar form)	Soft
Fy nghadair	My chair	Nasal
Ei chadair	Her chair	Aspirate

Table 4: Welsh mutations

A detailed discussion of Welsh mutations is found in Jones (1992). It is sufficient to realise that mutated forms of words, though a grammatical feature of spoken and written Welsh, will not appear in the initial lexicon. Hence, support for mutation is a prerequisite for any lexicon searching software.
Subject-dependent verb endings also occur in Welsh, to a greater extent than in English. As only infinitive forms of verbs appear in the initial lexicon, verb endings must also be handled by the lexicon searching procedure.

The lexicon search routine operates as follows:
- Search for word in its given form: if it exists, output transcription and end.
- If word does not exist, convert word into an unmutated form if possible, and search for transcription of unmutated form. If it exists, alter first phoneme of transcription to convert it to the mutated form, output transcription and end.
- If word still not found, assume that word is a verb, attempt to convert verb into infinitive form, and search for this infinitive form:
    - If it exists, alter final phoneme-like units of transcription to include the verb ending, output this transcription and end.
    - If it does not exist, search for unmutated form of infinitive verb, and if this exists, alter the initial and final phoneme-like units of the transcription to include the mutation and the verb ending, output this transcription and end.
- If word is still not found, report that word is not in lexicon.

The procedure adopted by the program is found to be very effective in correctly identifying and transcribing words, even those which have been both mutated and have a verb ending. It finds approximately 95% of the words in a typical piece of text, and of those a very high proportion are correctly transcribed. Most of the errors and missing words are due to irregular verb endings, which are not present in the lexicon. The output of the program is always checked by hand.
With a reliable form of gaining phonetic transcriptions available, the spoken database items are specified.

2.2.3 Phonetically rich sentences The specification of the 2000 phonetically rich sentences (PRSes) for Welsh, set out by the SpeechDat consortium (Winski, 1997), is that they should lead to at least 200 utterances of each phoneme in the final database. Each sentence is repeated 9 times, and so if 200 utterances are to take place, there must be at least 200/9 = 23 occurrences of each phoneme in the PRSes.

Speaking style considerations
The question of speaking style is non-trivial in Welsh. The past three decades of stratification caused by such factors as broadcast Welsh (Ball et al., 1988) and bilingual Welsh/English education make it difficult to focus on the 'ideal speaker-listener', if indeed one can be defined at all. Diglossia in Welsh has to be considered in the design of a PRS set "with readability in mind" (page 19 of Winski, 1997).
It is possible to define 'literary' (Jones, 1988) and 'broadcast news' (Ball et al., 1988) styles in Welsh. Of the 2000 PRSes, about 40% achieved a 'literary' style, the remainder being akin to 'broadcast news'. It should be noted that both styles are significantly more formal than most day-to-day Welsh speech.

Sentence creation
In order to create the first 750 sentences, a set of phonetically rich sentences for English used by BT in existing databases, such as Subscriber (Simons and Edwards, 1992) were translated and adapted. The remaining 1250 sentences were taken from a Welsh-language corpus of news stories designed to be read aloud. It was felt that the Welsh used within this corpus would be relatively colloquial and suitable for use in the phonetically rich sentences.
After extracting and checking the text of the news items, a set of 800 words were added or substituted at relevant points in the sentences. This enabled the vocabulary contained in the PRSes to represent a wide range of linguistic features and dialectical variations.
Eventually, a set of sentences was found which met the requirements of the SpeechDat specification. The frequency distribution of phoneme-like units in these sentences was approximately Gaussian.

2.2.4 Phonetically rich words A set of 1600 phonetically rich words was prepared, of which four appear on each sheet. Each phonetically rich word is thus uttered five times in the 2000-speaker database.
Due to the large number of phoneme-like units in Welsh - about 60 if all accents are taken into consideration - it was not possible to ensure that each phoneme-like unit was uttered 400 times (i.e. occurred 80 times or more within the phonetically rich word set). However, the phoneme-like units for which this condition is not satisfied are the rarest ones in Welsh, which occur for far

less than 1% of the time in the language. Hence, this issue is thought to be relatively unimportant.

2.2.5 Links to other databases Some items within SDCymru complement the SpeechDat(II) speaker identification databases (SDBs). These are:

- the 6 digit PIN codes (drawn from a set of 150)
- the credit card numbers (drawn from a set of 150)

Although no Welsh SDB is recorded as part of the current SpeechDat project, it is felt that these links can avoid the need to record a large impostor database in future.

3 Database collection

3.1 Recording site and platform

The recordings were made at BT Labs, Ipswich, via an ISDN-30 circuit. The recording platforms were accessible to volunteers via a free 0800 number.

The recording hardware was a set of Pentium PCs running Consensys UNIX. An Aculab E1/PRI 30 channel ISDN card and Aculab speech card were used. Silence detection was used, with up to three time-outs allowed on the same utterance before the system assumed the caller had cleared and the call was ended. Four recording channels were provided for SDCymru and four for the English mobile database, recorded during the same period.

3.2 Speaker recruitment

Speakers were initially recruited through the 60 bilingual secondary schools in Wales. These educate children between the ages of 11 and 18 through the medium of Welsh. The school was sent information letters and response slips to be handed to the children to give to the parents. A small cash incentive was given to the school for each slip returned. Those respondents who met the demographic requirements of SpeechDat were then chosen, and a prompt sheet sent to each of these. For each individual subsequently making a phone call to the recording platform, a further, larger, donation was made to the school.

During the later stages of the database, a 'snowball' method was used to gain the final volunteers. This involved sending each volunteer a sheet with room for the names of five speakers. They were encouraged to collect the names of potential volunteers, and return the completed name slips to the University of Wales Swansea. As with the school recruitment, a small cash incentive was given to the school or charity nominated by the respondents.

3.3 Practical factors

2000 prompt sheets were created. It was decided not to use an oversampling method, as the method of collection enabled a relatively tight control on the numbers of people phoning the recording platform.

Similar items on the prompt sheet (such as digit strings, directory assistance names, phonetically rich sentences and words) were grouped together. It was felt that using this method would be more likely to result in correct utterances than if disparate items followed each other, as the speaker would not continually have to consider how to say each item.

4 Annotation

It is necessary to listen to each SpeechDat(II) call manually. This ensures quality, and also enables an accurate word-level transcription of each utterance to be obtained. The transcription software used is Vox!, produced by CSELT (Constantinescu et al., 1997). A preliminary check is made before transcription to ensure that each call has been recorded to completion, and that the call was made by a co-operative speaker.

A team of six annotators is used. Efforts are made to ensure that more than one annotator is working at the same time, in order to aid communication between annotators and achieve uniformity in the transcribed speech. A random sample of transcribed calls is taken, and checked to ensure that the transcriptions are correct and consistent. The transcriptions follow the conventions of Francesco et al. (1997). Mispronunciations, unintelligible stretches of speech, and non-speech acoustic events are all marked.

5 Progress to date

At the time of writing (mid-April 1998), 1525 calls have been recorded, 1000 of which have been annotated. Approximately 25 of the calls have been rejected before annotation, due to such factors as excessive line noise, obtrusive intermittent background noise, or uncooperative speakers.

Recruitment of women in the linguistic North and South-West of Wales is complete. Our recruitment efforts are currently aiming mainly at the collection of male voices throughout Wales, and females from the linguistic South-East.

Dissemination exercises include a website with project information, aimed primarily at non-professionals in the field of speech recognition. SDCymru has also gained a substantial amount of national press coverage, in printed media, on radio and on television.

6 Conclusions and further work

The present resources for multi-lingual databases represent a unique opportunity for benchmark experiments in cross-language recognition. Resources such as SpeechDat(I) and (II) have made 'universal speech recognition' for multiple languages an active area of current research. This could be accomplished through cross-language phoneme mapping - see Chollet et al. (1997) as an overview, and specifically Andersen et al. (1993), Sovka (1997), and Constantinescu & Chollet (1997).

There is no technical reason why speech recognition for minority languages could not reach rapid fruition. Phoneme mapping, in particular, represents a fast bootstrap method for building speech recognisers in 'minority' languages from existing resources in other 'majority' languages. However, comprehensive speech databases in the target minority languages are essential for reliable recognition.

We believe, therefore, that the collection of speech resources such as SDCymru is vital for future technology developments in these fields. The completion of

SDCymru will lead to an increased awareness of minority languages within the European speech community. Most importantly of all, the potential end-products developed as a result of the database will enable Welsh speakers to use their language more widely, and increase the status of Welsh in the global technology marketplace.

References

All SpeechDat public deliverables are obtainable from [http://speechdat.phonetik.uni-muenchen.de/]

Andersen, O. et al. (1993). Data-driven identification of poly- and mono-phonemes for four European languages. In Proc. EuroSpeech '93 (Vol. 2, pp. 759-762), Berlin, Germany.

Ball, Martin J. et al. (1988). Broadcast Welsh. In Ball. Martin J. (Ed.), The Use of Welsh. Clevedon and Philadelphia: Multilingual Matters Ltd.

Chollet, Gerard et al. (1997). Multilingual speech databases: Beyond SpeechDat. In IWSSIP '97, Poznan, Poland.

Constantinescu, Andrei et al. (1997). Report on developed tools. SpeechDat(II) public deliverable LE2-4001- SD3.1.2.

Constantinescu, Andrei and Chollet, Gerard (1997). On cross-language experiments and data-driven units for ALISP. In Proc. IEEE Automatic Speech Recognition and Understanding Workshop 1997 (pp. 606-613). IEEE Press: Santa Barbara, CA.

Darlington, Thomas (1901). Some dialectical boundaries in Mid Wales. In Transactions of the Honourable Society of Cymmrodorion. 1901, pp. 13-39.

Jones, Dafydd Glyn (1988). Literary Welsh. In Ball, Martin J. (Ed.), The Use of Welsh. Clevedon and Philadelphia: Multilingual Matters Ltd.

Jones, Robert Owen (1967). A structural phonological analysis and comparison of three Welsh dialects. MA thesis. University of Wales Bangor.

Jones, T.J. Rhys (1992). Teach Yourself Welsh. Hodder and Stoughton.

Lamel, Lori and Adda, Gilles (1996). On designing pronunciation lexicons for large vocabulary, continuous speech recognition. In Proc. ICSLP '96 (pp. 6-9), Philadelphia.

Lau, Raymond et al. (1997). WebGalaxy - Integrating spoken language and hypertext navigation. In Proc. Eurospeech '97 (Vol. 2, pp. 883-886), Rhodes, Greece.

OPCS/Office of Population Censuses and Surveys (1994). 1991 Census: Welsh language/Cyfrifiad 1991: Yr iaith Gymraeg. London: OPCS.

Pinker, Steven (1994). The Language Instinct. Allen Lane/Penguin, p. 259.

Senia, Francesco et al. (1997). Specification of orthographic transcription and lexicon conventions, SpeechDat(II) public deliverable LE2-4001-SD1.3.2.

Simons, Alison and Edwards, Keith (1992). Subscriber - a phonetically annotated telephony database. In Proc. Institute of Acoustics, Windermere 1992 (Vol. 14 Part 6, pp. 3-15).

Sovka, Pavel (1997). Cross-language experiments verifying the approach of converting French flexible vocabulary to Czech. ENST Paris research report: private communication.

Thomas, Beth and Thomas, Peter Wynn (1989). Cymraeg, Cymrâg, Cymrêg… Cyflwyno'r Tafodieithoedd. Cardiff: Gwasg Taf.

Thomas, Ceinwen H. (1993). Tafodiaith Nantgarw: astudiaeth o Gymraeg llafar Nantgarw yng Nghwm Taf, Morgannwg. Vol.1. Cardiff: University of Wales Press.

Vary, P. et al. (1989). Speech codec for the European mobile radio system. In IEEE Globecom, November 1989.

Wells, John (1989). Computer-coded phonetic notation of individual languages in the European Community. In Journal of the IPA. Vol. 19, pp. 37-53.

Williams, Briony (1993). Letter-to-sound rules for the Welsh language. In Proc. Eurospeech '93 (Vol. 2, pp. 937-940), Berlin, Germany.

Winski, Richard (1997). Definition of corpus, scripts and standards for Fixed Networks. SpeechDat(II) public deliverable LE2-4001-SD1.1.1.