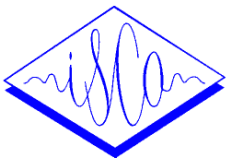

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

LETTER-TO-SOUND RULES FOR THE WELSH LANGUAGE

Briony Williams

Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, UK.

ABSTRACT

A set of letter-to-sound rules was written, in the form of context-sensitive rewrite rules. A publicly-available program was used that converted this style of rule to a C program. Three sets of rules were written, reflecting three separate passes through each input word. The first set added epenthetic vowels, while the second set located the vowels and lexical stress. The third set carried out grapheme-to-phoneme conversion. The rules were evaluated over text from a magazine article, and found to have a 96% success rate.

***Keywords**: Text-to-speech synthesis, Welsh, letter-to-sound rules.*

1 INTRODUCTION

In developing a text-to-speech (TTS) synthesiser for the Welsh language, it is necessary to include a set of letter-to-sound rules, to convert words from orthographic form to a phonemic transcription. Letter-to-sound (LTS) rules have been part of TTS systems for English for many years (e.g. [1], [2]): the work reported here seems to be the first such set of rules for Welsh.

Certain aspects of Welsh present more of a challenge to LTS rules than do others. In the case of consonant graphemes, there is almost a one-to-one mapping between grapheme and phoneme. In polysyllabic words, the penult is stressed (of the few exceptions, most are stressed on the final syllable, or 'ultima'). Irregular stress is often indicated by an accent mark in the spelling.

In the case of vowels, Welsh has phonological vowel length, determined by the type and number of the following consonants [3]. The contrast between long and short vowels exists only in stressed syllables. Orthographic 'y', as well as having long and short forms like other monophthongs (/ii/ and /i/), may also be realised as schwa (/@/) in non-final syllables. It may also form part of a diphthong. Thus the LTS rules must do a certain amount of parsing within each word.

Welsh permits *epenthetic* vowels: vowels which are pronounced but not shown in the orthography. They occur in certain contexts: for example, *cefn*, `back', is pronounced /k ee* v e n/. This occcurs in only some dialects of Welsh.

The greatest difficulties are posed by the graphemes 'i' and 'w'. The grapheme 'i' may represent either a vowel (long /ii/ or short /i/) or the palatal glide /j/. The grapheme 'w' may represent either a consonant (/w/), a monophthong (long /uu/ or short /u/), the first half of a diphthong, or a labialisation marker on an alveolar liquid after /g/ or zero. The grapheme 'w' is by far the most difficult grapheme to handle, accounting for a great number of the rules.

The rule software used for formulating the LTS rules is publicly available. The main C program of this software is known as `phon.pro': its English instantiation is known as `eng'. It was originally written by Greg Lee of Hawaii University (e-mail: lee@uhunix.uhcc.hawaii.edu), who used it to realise a set of LTS rules for English. The software allows the linguist to write critically-ordered context-sensitive rewrite rules in a form familiar to linguists. It converts the rules into a program header, and compiles a C program which runs the rules. This approach combines the advantages of user-friendliness in writing the rules with speed of running. Some modifications were made by the author to allow for Welsh-specific contexts. For example, `D' was introduced to refer to `one and only one consonant grapheme or digraph that can condition phonological length in the preceding vowel', i.e. one of `b, d, g, f, dd, ff, th, ch'. The rules were developed and run on a Sun 4 workstation, but it is also possible to run them on an IBM-compatible PC.

Welsh exists in many local dialects. These fall into two broad groupings: North and South Welsh dialects. The LTS rules are based on a South Welsh accent, but it would be simple to edit them to handle a Noth Welsh accent instead. More specifically, the rules for insertion of epenthetic vowels would be needed only in some South Welsh accents.

A Machine-Readable Phonetic Alphabet for Welsh (MRPAW) was designed. As a South Welsh accent had been chosen, there were 32 consonants (including the three labialised consonants /lw, nw, rw/, and also the phonemes /z/ and /jh/ that are used only in loan-words from English). With the 19 vowels, there were a total of 51 phonemes.

For the input of orthographic forms, it was decided to assume no special capability on the part of the computer terminal. Therefore, acute accent was input as a plus sign (+) after the vowel, grave accent as a backward slash (\) after the vowel, and circumflex (^) and diaeresis (") as the appropriate symbols after the vowel.

The LTS rules implemented by Lee for English (using his `phon' software) were based on those in [4], and only one pass through each input word was required. For Welsh, however, three passes are required, with a different set of rules for each pass. The first adds any epenthetic vowels to the input word. The second locates the stressed syllable, identifies the vowels, and differentiates between the vocalic and consonantal forms of `w' and `i'. The third implements the grapheme-to-phoneme rules proper.

2 FORMULATION OF THE RULES

2.1 Epenthetic vowels

In accents from the south central part of Wales, an epenthetic vowel occurs after one of /b, d, g, p, t, k, f, v, dh, s, x, th/ when followed by one of /l, n, r/, or between /m/ and /l/, at the end of a word. The vowel inserted is identical to that of the preceding syllable. If that contains a diphthong, then the epenthetic vowel takes on the quality of the second half of the diphthong. Examples follow:

(1) *aml*, `frequent', /a* m a l/
(2) *llwybr*, `footpath', /lh ui* b i r/

Since the rule formalism does not permit the specification of an output symbol from a null input, it is necessary to supply rules for all possible consonants that can precede an epenthetic vowel. Since the quality of the epenthetic vowel depends on that of the preceding vowel, it is also necessary, for each possible consonant, to specify all possible preceding vowels. These restrictions explain the large number of rules required for a relatively simple task, namely 170 rules (including rules for puctuation characters).

(3) e[f]R# = fe

The rule in (3) applies to the grapheme `f' when preceded by `e' and followed by one of `l', `n', `r' (represented by the single `R' in the rule), which precedes a non-alphanumeric character (represented by `#' in the rule), which would signal end-of-word. If the rule applies, the sequence `fe' is output in the place of the input `f'. Thus, for instance, input *cefn*, `back', would be output as *cefen*.

2.2 Stress and vowel location

The output of the first set of rules forms the input to the second set. These locate the stressed syllable and the vowels. They also differentiate between the various realisations of orthographic 'w' and 'i'. There are 731 rules altogether (including rules for punctuation characters)

2.2.1: Parsing task: The rules fall into several `blocks' of rules, each dealing with the same kind of context. As the rules are critically ordered, the rules with the most specific contexts appear first, while those with the least (or no) context appear last, serving as default rules.

In the first block of rules, vowels that are orthographically marked as stressed (with an accent mark) are output as capitalised graphemes (capitalisation being used as a marker of stress in the output forms). Such a syllable may form the ultima, the antepenult, or a preantepenultimate syllable of a polysyllabic word. An example follows:

(4) [a+u] = AU.
eg. *nesáu* = *nesa+u*, /n e s ai*/,
`to approach'

In this rule, the capitalisation of the output indicates stress, while the dot after the output sequence indicates the presence of a vowel. There is no context specification in this rule.

The next block of rules refers to monosyllabic words. The rules assume that all monosyllables are stressed, since stressless monosyllables, forming a limited set of function words, would already be in the dictionary. Included here are rules for penult vowels where the ultima vowel is `w' (phonemically /u/), since such a pattern would appear to be a single syllable to the rules.

The third block of rules handles unstressed penults before ultimas that are orthographically marked as stressed, or unstressed preantepenults before stress-marked antepenults. The fourth block of rules, similarly, handles unstressed penults after antepenults that are orthographically marked stressed. The fifth block handles stressed penults (the default case in polysyllables), where no orthographic marking of stress is present. An example follows:

(5) [iw]QCVC# = IW.
eg. *diwrnod*, /d iu* r n o d/, `day'

This rule states that input `iw' is output as `IW.' (ie. a stressed vowel) in the specified context. Table 1 glosses all one-character aliases used in the rules, while table 2 glosses some characters used in the output.

The sixth block of rules handles the remaining vowel cases, ie. vowels in unstressed ultimas and antepenults (both of which are normally unstressed), and unstressed penults (normally stressed). The seventh block of rules allows

consonant graphemes to pass through unchanged, while a final block does the same for punctuation marks.

Q	1 consonant (includes w)
S	1 non-w consonant
C	0 or more cons's (includes w)
G	0 or more non-w consonants
V	1 vowel (includes w)
H	1 non-w vowel
R	1 of {l, n, r}
K	1 of {c, g, ngh}
#	1 non-alphabetic (end-of-word)
D	1 of {b,d,g,f,dd,ff,th,ch}
P	1 of [s, ll}

Table 1: One-character aliases used in context specification in rewrite rules.

J	palatal glide
M	consonantal 'w'
W.	stressed vocalic 'w'
w.	unstressed vocalic 'w'

Table 2: Some characters used in the output of the rules.

2.2.2 Orthographic `i': Orthographic 'i' can be realised as either a vowel or a palatal glide, as in:

(6) #C[iai]C# = JAI.
eg. *iaith*, /j ai* th/, `language'

This states that the sequence `iai' becomes `J AI.' (ie. a palatal glide followed by the stressed diphthong `ai') in the specified context.

*2.2.3 Orthographic `w':* Orthographic `w' is the cause of many of the rules in most blocks of rules. Since `w' cannot be classed as either a consonant or a vowel before the operation of the rules, it is necessary to have separate sub-blocks of rules to deal with the special cases it presents. An example follows:

(7) [w]SGHG# = W.
eg. *bwled*, /b u* l e d/, `bullet'
(8) [w]CVC# = W.
[not one of the rules used]

The rule in (7) states that `w' becomes `W.' (identified as a vowel by the following dot, and as stressed by capitalisation) in the specified context. The simpler rule in (8), while seemingly adequate to cover the same case, would not in fact do so. This is because it is necessary to specify that at least one consonant must intervene, in order to cut out words such as: *dwyn*, /d ui* n/, `to bear', where the `w' forms part of a diphthong; and *chwarae*, /x w aa* r ai/, `to play', where the `w' is a consonant. The rule in (8) is also not adequate to cover cases such as *cwrw*, /k uu* r u/, `beer', where the second vowel is the grapheme `w' which has not yet been assigned consonantal or vocalic status.

2.3 Grapheme-to-phoneme conversion

The output of the second set of rules forms the input to the third (and final) set of rules. These rules number 356 (including the punctuation rules) and are less complex than the second set, though more complex than the first. The rules for consonants are very straightforward, while those for vowels must handle vowel length and the variant realisations of `i', `y' and `w'.

2.3.1 Vowel length: In monophthongs, phonological length is determined by the type and number of the following consonants [3]. Where a vowel is followed by one of orthographic `b, d, g, f, dd, ff, th, ch' or zero, before the next vowel (or end-of-word), then the vowel is long if it is stressed (assuming a South Welsh accent). The vowel is also long if followed by one of orthographic `s, ll' plus end-of-word. Where the vowel is followed by any other consonant, or by two or more consonants, then it is short (even if stressed). The following are examples of vowel length rules:

(9) C[E].Dwy. = ee* eg. *dedwydd*,
/d ee* d ui dh/, `happy'
(10) C[E].DwV = e* eg. *edwi*,
/e* d w i/, `to fade'
(11) C[E].V = ee* eg. *lleol*,
/lh ee* o l/, `local'

These (critically-ordered) rules use `D' to refer to the vowel lengthening consonants given above, and `P' to refer to the additional consonants `s' and `ll' which lengthen a preceding monophthong only when word-final. In (9), the single vowel-lengthening consonant is followed by a vowel and so the preceding vowel (which is stressed) is long. In (10), the vowel-lengthening consonant is followed by a second consonant and so the preceding vowel (even though stressed) is short. In (11), the stressed vowel is followed by another vowel not forming a diphthong with it, and so is long.

2.3.2 Orthographic 'i': The rules for the second pass distinguished between the vocalic and consonantal realisations of `i' by adding a dot after the former and outputting 'J' for the latter. The third set of rules then converts these to the appropriate phonemes: /i/ or /ii/ (stressed or unstressed) for the vocalic form, and /j/ for the consonantal form. An exception occurs where `s' precedes, in which case the `si' digraph takes precedence and becomes /sh/.

2.3.3 Orthographic 'y': Where it is a monophthong, orthographic 'y' becomes schwa (stressed or unstressed) when in non-final syllables of polysyllables, and /i/ or /ii/ (stressed or unstressed) in monosyllables or final syllables of polysyllables. Where it forms part of a diphthong

(orthographic `oy, wy, yw, ey'), orthographic `y' is a high front vocoid and never schwa.

2.3.4 Orthographic 'w': A final complication in the treatment of `w' is the case of labialised consonants, in words such as *gwlad*, /g lw aa* d/, `country'. In such cases, there is strong labialisation on the consonant following the /g/ (which may be /l/, /n/ or /r/). A labialised consonant may also follow word-initial /ng/ (voiced velar nasal) or be word-initial itself. One of the rules dealing with labialised consonants is as follows:

(12) #g[Mr] = rw eg. *gwraig*, /g rw ai* g/, `woman'

Since the second set of rules here outputs the `M' character, the third set of rules is provided in the input string with the information that this is consonantal 'w'.

3 EVALUATION OF OUTPUT

The next step was to evaluate the output of the rules over a representative sample of Welsh words.

3.1 Data

The data used for evaluation was taken from an article from a Welsh women's magazine. This was chosen as being representative of the level of language used by Welsh speakers in everyday life. The number of unique words was 460. Mutated forms of the same words, which were identical except for the initial phoneme, were excluded. However, inflected or derived forms of the same lexeme were included, as these often introduced a substantial difference into the pronunciation (affecting the lexical stress placement in many cases). English words that occurred in the text were also excluded. A minimal amount of pre-processing was done, in that any initial capital letters were reduced to lower-case. The rules were run over this data in three passes, as described above.

3.2 Results

An initial count of the results showed that 55 of the 460 words were incorrectly output, ie. a success rate of 88%. However, closer examination revealed various categories of error, as follows.

There were 40 monosyllabic function words among the 55 errors. These words had been assigned a stress by the rules, even though in the vast majority of cases they would not be stressed in actual use. For most of these 40 words, the fact that stress had been assigned was the only error.

There were 12 words among the 55 that were exceptions in Welsh (eg. *ymddangos*, /@ m dh a* ng g o s/, 'to appear', was output by the rules as */@ m dh a* ng o s/, without the /g/. The appearance of /g/ in this word is a lexical exception.

There were 3 words among the 55 that would have required morphological knowledge for correct specification of pronunciation. For example, *grynwraig*, /g r @* n rw ai g/, 'Quaker woman' was output by the rules as */g r @ n u* r ai g/, where the 'w' grapheme has incorrectly been taken to be a vowel. In fact, it is a labialisation marker, as a morphological boundary intervenes between 'n' and 'w' (the constituent morphemes being the mutated forms of *cryn*, from the verb 'to quake', and *gwraig*, 'woman', with labialised /rw/).

3.3 Discussion

Any set of letter-to-sound rules for any language will encounter the second and third categories of error above, as LTS rules cannot handle exceptions and have no access to morphological knowledge. However, the limited set of function words in a language would never be passed to the LTS rules in the first place. This is because these words would be included in the lexicon in any TTS system. Therefore, a fairer test of the functioning of these LTS rules as part of a TTS system would entail excluding the function words from the analysis. Given this revised interpretation of the data, the results show only 15 errors (the exception words and polymorphemic words) out of 420 input words, ie. a 96% success rate. This rate is slightly better than the 93% achieved with LTS rules for English ([2]), and probably represents a realistic upper limit on accuracy.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the support of a BP Research Fellowship, awarded by the Royal Society of Edinburgh and funded by British Petroleum.

REFERENCES

[1] J Allen, M S Hunnicutt & D Klatt (1987), *From text to speech: The MITalk system.* Cambridge: CUP.

[2] J M McAllister (1988), `CSTR Text-to-Phoneme Project Status Report'. CSTR internal Report.

[3] G M Awbery (1984), `Phonotactic Constraints in Welsh', in [5].

[4] H S Elovitz, R W Johnson, S McHugh & J E Shore (1976), `Automatic translation of English text to phonetics by means of letter-to-sound rules'. US Naval Research Laboratory Report 7948.

[5] M J Ball & G E Jones (1984), *Welsh Phonology.* Cardiff: University of Wales Press.