Automatic conversion between historical spelling, modern spelling, and IPA for Welsh

JOHN D. PHILLIPS

Abstract

本稿では、歴史的文字遣い、発音記号、そして現代文字遣いの間を自動変換する為のコンピュータソフトウェアの設計と実装について述べる。綴り法則、形態素解析、音素配列条件を併せて、不定の文字遣いの曖昧さを説き明かすことで、文字遣いと発音記号との写像を実現する。その発音記号が諸文字遣いの変換中軸に為る。

古代の文書を、現代の諸言語の為にますます利用可能になっている自然言語処理ソフトによって分析しやすくするのが目的である。これにより、言語変化や様々な歴史的段階の比較研究を容易にする。

具体的には、この研究はウェールズ語の其々の音素の機能負担量の変化に関する調査の下調べとして実装された。それで、ウェールズ語の実装を示して、中世の文字遣い、発音記号、そして現代文字遣いの間の変換に関わる問題とその解決方法について論じる。

1. Introduction

In recent years much valuable linguistic research has been done using computers and machine-readable text. Texts in one or more languages can be downloaded from the Internet and analysed or compared automatically with appropriate computer software. Software is currently available for various types of grammatical and statistical analysis for many of the more widely-spoken languages.

A language changes over time and it would be useful to be able to compare different historical stages of a language automatically. Texts in many older languages have by now been scanned or transcribed into computer-readable format and can be searched and indexed automatically by computer. However spelling conventions, even whole writing systems change over time, adding an extra layer of difficulty to automatic comparison of texts of different periods. Shakespeare's vocabulary cannot be compared directly with the vocabulary of a modern English writer because the spelling is different and the computer cannot match the words up — though in this particular case, editions of Shakespeare in modern spelling are available. The Japanese of a century ago cannot be analysed with standard software because the post-war spelling reform means that the words of a century-old text will appear in a different form in the software's modern dictionary.

It is also the case that for many languages and periods spelling is less fixed than for modern English and Japanese. Orthography often allows several possible spellings for the same word, a latitude which is at odds with the format of modern dictionaries.

The research described here has produced software which converts texts in historical orthography into modern orthography, with the aim of doing linguistic research on those texts using software designed for the modern language.

2. Writing Systems

A writing system is a symbolic system used to represent elements of a language. It comprises a set of base elements, and a set of rules and conventions which assign meaning to the base elements, termed orthography. Many current writing systems are alphabetic: the base elements are letters representing the consonants and vowels of the spoken language. For convenience the exposition below refers to alphabetic writing systems, although it is equally applicable to syllabic writing systems.

An ideal alphabetic writing system is phonemic, also described less accurately as phonetic. The term comes from the phoneme, the smallest distinctive unit in the linear sequence of sounds which make a spoken utterance. A phoneme is distinctive in the sense that changing any phoneme of a word changes the meaning, for example *pit* contains three phonemes because it contrasts with *nit, pat,* and *pin.* In a phonemic writing system, each letter corresponds to a phoneme, so that spelling corresponds letter by letter to pronunciation: spelling can be predicted from pronunciation and vice versa. Many alphabets, ancient and modern, are phonemic, or nearly so, including the ancient Greek alphabet and its variant the Roman alphabet. The Roman alphabet developed for writing Latin, for which five vowels *aeiou* and thirteen consonants *bcdfghlmnprst* were sufficient. Latin also distinguished short from long vowels, and this distinction was rarely marked in writing, so the system was not entirely phonemic in everyday use: the well-known homograph sentence *malo malo malo malo* would have been pronounced /maːlo maːloː maloː maloː/; as Benjamin Britten[1] puts it

Malo: I would rather be
Malo: In an apple tree
Malo: Than a naughty boy
Malo: In adversity

The Latin alphabet also used five other letters: the letters *k* and *q* represented the same phoneme as *c*, and the letter *x* represented the combination *cs*. Later the letters *y* and *z* were taken from the Greek alphabet to use in borrowed Greek words containing those letters: *y* was [y] in Greek, but was pronounced the same as *i* in Latin, and *z* seems to have been pronounced as though written *ds*.

Later, when the Latin alphabet was adapted to other languages, other letters were added, now for example *ð* and *þ* are used in Icelandic and *j, v, w* are used in several European languages.

Conversion between phonemic writing systems is simple, consisting of just replacing each letter by its equivalent. With the Roman alphabet (and to a lesser extent in other alphabets) there is a tradition of using a sequence of letters to represent a single phoneme. English for instance has the digraphs *ph* /f/, *sh* /ʃ/ and *th* /θ/ or /ð/, amongst others. These give rise to ambiguities in reading and transliteration, for instance in the common mispronunciation of *mishap* as /miʃəp/. As an example of transliteration: the Serbian language is written with both Cyrillic- and Roman-based scripts. The correspondence between them is regular, but involves several correspondences between single Cyrillic letters and sequences of Roman letters: Cyrillic ј corresponds to Roman *j* and Cyrillic л to Roman *l*, but the single Cyrillic letter љ corresponds to Roman *lj*. There is no problem converting Cyrillic to Roman, but ambiguity arises in converting Roman to Cyrillic: should *lj* convert to љ or to лј?

Though all alphabetic writing systems have a large degree of correspondence between written letters and spoken phonemes, many are far from the phonemic ideal. In some cases this is because the pronunciation has

[1] In the opera *The Turn of the Screw*, libretto by Myfanwy Piper. The four words are: 1st person singular present of the verb *malle* 'prefer', dative case form of the noun *mâlus* 'an apple tree', dative of the adjective *malus* 'naughty', dative of the abstract noun *malum* 'adversity'.

changed over centuries and the orthography has not kept up, as with English, French, Gaelic, Tibetan, and many other languages. In other cases it arises from an alphabet developed for one language being used to write another, and the match between letters and phonemes being poor owing to the languages having different inventories of phonemes. Old Welsh is a good example: it probably had nine vowels /a e i o u ɨ ʉ ə ɵ/ and twenty-eight consonants /p b t d k g ɸ β θ ð χ ɣ m̥ m n̥ n ŋ̊ ŋ ɬ l r̥ r β̰ s h w j/. When Old Welsh was written using the Latin alphabet, most of the letters were made to do double duty, *b* for instance representing both /b/ and /β/, *i* representing /i/, /ɨ/, /ə/ and /j/. The resulting ambiguity means that the phonological interpretation of a spelling is not straightforward.

3. Approaches to the problem

The specific problem addressed here of converting historical orthography into IPA and modern orthography, appears not to have been researched before. Repositories of historical texts typically provide search facilities in original spelling only, with no grammatical or phonological analysis. Research into the morphology or grammar of historical varieties, or comparison with the modern language, can only be done manually.

The spelling of historical texts could of course be modernised with a dictionary listing the various older spellings of each word. This would need to be a full-form dictionary of course e.g. listing not just the verb *write*, but all its forms *writes, wrote, writing, written.* However vocabularies of languages run at least to tens of thousands of words and far more for languages with productive morphology: even in Latin each verb has a hundred and twenty-odd forms. The compilation of such a dictionary would hence be at least laborious and possibly (for highly inflexional languages) impracticable. It should anyway be unnecessary when the conventions of the orthographies are well-known.

The difference between the various historical stages of a language is similar to the difference between related dialects and languages. Software for computer-assisted dialect adaptation (CADA) and computer-assisted related-language adaptation (CARLA) has been produced in recent years[2] and could be used to translate between different stages of the same language. This software uses tables of phonological and morphological equivalences between dialects of a language, assuming that the equivalences are regular and can be represented unambiguously in the dialects' writing systems. Lexical substitution is also allowed for, using tables of equivalent words. As with other machine translation software, the output is a draft to be manually corrected. This works well when the dialects concerned use the same writing system, or writing systems where graphemic equivalences can be listed unambiguously. Traditional writing systems, though, are often ambiguous so that tables of equivalences used alone will produce highly ambiguous output. CARLA software normally has no means of solving ambiguity: such ambiguous cases must be handled as lexical substitution, so that in the end there is little advantage over the dictionary method.

The Ontology for Accessing Transcription Systems (OATS)[3] addresses a related problem, supporting operations over disparate transcription systems and practical orthographies. It is a knowledge base intended to be used by external software for search, error checking, and conversion. It includes ontological description of writing systems and relations for mapping them to the International Phonetic Alphabet. At present the knowledge base contains African languages with well-designed, phonemic orthographies. It is not clear how the ontology and relations would represent irregular orthographies like those of modern English and French, or inexact ones like mediæval Welsh.

The work described here, like OATS, allows description of a mapping between orthography and the IPA, and uses the IPA as a pivot to convert between different orthographies, in this case historical and modern orthography. Indeterminacy and underspecification in the phonological interpretation of an orthography is solved by a combination of spelling rules, phonotactic rules, and morphological analysis using an off-the-shelf morphological analyser and dictionary of the modern language.

[2] The Summer Institute of Linguistics has an overview of what is available at http://carla.sil.org/carlaint.ppt

[3] Steven Moran: 'An ontology for accessing transcription systems'. *Language Resources and Evaluation*, 45 (2011), pp. 345-360.

The program was designed and implemented in the context of a research project on changes in the Welsh language over the centuries, and the program is described here with examples from mediæval and modern Welsh. The program assists research into language change: for instance in changes over time in frequency or usage of vocabulary or grammar. It was originally written to facilitate comparison of the functional load of the different phonemes of Welsh between historical periods of the language.

4. The Welsh Language

Welsh is the indigenous language of Wales, an area on the western side of the island of Great Britain, bordered by England to its east and the Irish Sea and Atlantic Ocean to its north, west, and south. Welsh is a Celtic language, closely related to Breton (spoken in Brittany, part of France), and less closely to the Gaelic of Scotland and Ireland. There are about 600,000 speakers in Wales, another 100,000-odd in England, and some thousands in the old Welsh settlement in Patagonia.

Written Welsh has a continuous history from ancient times to the present day. Welsh was probably first written down when Wales was part of the Roman empire from the first to the fifth centuries A.D., though the earliest surviving text dates from around the year 700. Surviving materials from the early centuries are few, but the consistency of their language and orthography, as well as some explicit references, suggest that the scarcity is due to subsequent destruction in warfare, rather than to a lack of production. The language of these early texts (Early Welsh and Old Welsh) is sufficiently different to modern Welsh as to make it inaccessible to a modern reader without explanation.

However there is a large corpus of written Welsh from the later mediæval period. From about the twelfth century, literary, historical, legal, medical, religious, and other texts were produced throughout Wales in a homogeneous standard language which is still intelligible to modern readers without great difficulty. All the important mediæval texts are available in printed editions, and in the last decade a series of projects have made many available in machine-readable form, some freely downloadable from the Internet.

The language of the mediæval texts changed somewhat over the centuries, and developed into the modern standard written language.

Concurrently, spoken forms of the language developed into the various regional varieties of modern colloquial Welsh. The nature of the changes, and exactly when and why they happened, is currently an active area of research.

The spelling system of Mediæval Welsh is very different to that of the modern language, and this is an obstacle to computational research on the texts. Automatic comparison with modern texts is not possible because of the different spelling. Software for processing the texts beyond simple search has not so far existed. Using the software described in this paper, old texts can be converted to phonetic symbols (of the International Phonetic Alphabet, IPA) or to modern spelling, and then analysed with existing software for grammatical analysis of modern Welsh[4] or for phonetic analysis.[5]

5. The Spelling of Mediæval Welsh

Welsh orthography changed during the twelfth century when the absorption of the Welsh church into the Catholic church, and the Europeanisation of neighbouring England subsequent to its conquest by the Normans, brought greater exposure to European scribal practices. Most surviving manuscripts are from the period subsequent to the change, belonging to the stages of the language called Middle Welsh (from the change in orthography until the fourteenth century) and early Modern Welsh (from the fifteenth century).

The orthography of Middle Welsh[6] was not standardised, and though it

[4] For example the modules for automatic grammatical analysis underlying the software applications for checking modern Welsh spelling and grammar produced by the Language Technologies Unit at Canolfan Bedwyr, University of Bangor, Wales (http://www.bangor.ac.uk/canolfanbedwyr/technolegau_iaith.php.en?); or the freely available computational grammar of Welsh produced by the Pargram project at the Department of Linguistics of the University of Essex (http://privatewww.essex.ac.uk/~louisa/esrcproj/).

[5] For example the software produced by the ASJP for comparing vocabulary; or Surendran and Niyogi's software for analysis of functional load.

[6] The orthography is described in detail by T. M. Charles-Edwards and P. Russell in 'The Hendregadredd Manuscript and the orthography and phonology of Welsh in the early fourteenth century', *National Library of Wales Journal,* 28 (1994). David Willis has a description of the phonology of the earlier stages of Welsh in his chapter 'Old and Middle Welsh' in *The Celtic languages*, edited by Martin Ball and Nicole Müller. Routledge, 2009.

is basically phonemic, the number of letters in the alphabet is inadequate so that several are made to do double duty and there is much latitude in the way phonemes without dedicated letters are spelt. Because of this the phonological interpretation of a spelling is often not obvious. The letters are those of the Latin alphabet: *abcdefghiklmnoprstuy* — *q*, *x* and *z* were not used. The letters have basic sound values as in contemporary mediæval Latin:

a	b	c	d	e	f	g	h	i	k	l	m	n	o	p	r	s	t	u	y
a	b	k	d	e	ɸ	g	h	i	k	l	m	n	o	p	r	s	t	u	i

There is considerable variation in how the Welsh phonemes not included above were spelt, depending to some extent on period and individual scribe. The following spellings were used for additional phonemes

β	θ	ð	χ	ŋ	ɬ	r̥	w	ʉ	j	ɨ	ə
u,f	t,th	d,t	ch	g	l,ll	r	u	u	i,y	y	y,e

A new Latin phoneme /v/ had developed from semivocalic /u/ early in mediæval Latin but the spelling did not change so the letter *u* came to represent both /u/ and /v/. Twelfth-century Latin had no /β/, and Welsh had always used *f* to represent bilabial /ɸ/, and this and the Latin /v/ perhaps seemed equidistant from Welsh /β/, resulting in its spelling as either *f* or *u*.

This orthography is ambiguous in both directions. The letter *f* for instance represents both /ɸ/ and /β/, but /β/ in turn can be represented by either *f* or *u*. The letter *g* represents either /g/ or /ŋ/, so that the spelling *agori* can represent /agori/ ‘to open’ or /aŋori/ ‘to anchor’. *L* can represent either /l/ or /ɬ/ though, as noted below, *ll* eventually became the standard spelling for /ɬ/.

Three digraphs with *h* originate in the Latin transcription of Greek words. *Ch*, *ph*, and *th* were originally transcriptions of the Greek aspirated stops *χ*, *ϕ* and *θ* in Classical Latin. The Classical Greek pronunciation was [k^h, p^h, t^h] so the transcription with *h* was natural, though native Latin-speakers would no doubt have pronounced /k, p, t/. In later Greek, *χ*, *ϕ* and *θ* came to be pronounced [x, ɸ, θ]. The digraphs *ch*, *ph*, and *th* were sometimes used in Old Welsh with these values, and their use continued into Middle Welsh.

Welsh has an opposition between fortis and lenis consonants. Though there are allophonic variants, the fortis series /p t k ɸ θ χ m̥ n̥ ŋ̊ ɬ r̥ s/ are typically voiceless, aspirated and long. The lenis series /b d g β ð m n ŋ l r/ are typically voiced, unaspirated and short. Note that the lenis series is lacking /ɣ/, which had existed in Old Welsh but was lost to elision or vocalisation, and /z/, which seems never to have existed. For those of the fortis series which could be written with a single consonant, /p t k ɸ ɬ s/, noticeably long allophones were sometimes written double: *pp, tt, cc, ff, ll, ss* respectively. *Ff* and *ll* eventually became the standard spellings for /ɸ/ and /ɬ/ respectively. With the nasals only, the fortis series could be distinguished from the lenis series, spelt *m, n, g*, by marking the aspiration with *h*. Fortis nasals do not occur word-finally, but word-internally the spellings *mh, nh, gh* are regular for /m̥ n̥ ŋ̊/. Word-initial fortis nasals are the result of the type of apophony called consonant mutation: a variation in the initial consonants of words in the Celtic languages according to their grammatical function. The modern dictionary form of the word for 'Wales' for instance is *Cymru* /kəmrɨ/, but in a text the word often appears as one of the mutated forms *Gymru* /gəmrɨ/, *Chymru* /χəmrɨ/, and *Nghymru* /ŋ̊əmrɨ/. Mutations are often ignored (spelt as the base form) in mediæval orthography. In particular there is rarely any attempt to notate a word-initial fortis nasal. Nasals are mutations of corresponding stops, so instead of /m̥ n̥ ŋ̊/, the respective base forms /p t k/ are written. Sometimes a morpheme-initial fortis nasal within a word is treated similarly, for example /əm̥en/ 'within', made up of two morphemes, *yn* and *pen*, is spelt *ymhen* or *ympen*.

Word-finally there was no contrast between voiced and voiceless stops in Middle Welsh. Word-final stops were probably voiceless but unaspirated. They were written with either *bdg* or *ptc*.

The system described above is a very ambiguous representation of the language. The letters *d, f, t, u, y* in particular are sufficiently ambiguous and frequent as to make reading difficult. The distinctions between /g ŋ/ (both spelt *g*), /ɬ l/ (both spelt *l*), and /r̥ r/ (both spelt *r*) carry smaller functional loads so are less important. Over the centuries the ambiguity was reduced by developing less ambiguous ways of representing some of the phonemes.

From quite early in the period, the occasional scribe uses special letter forms for /ð/ (usually transcribed *δ*) and /ɬ/ (transcribed *ll* or ỻ). These became commoner later, surviving into the age of printing and in handwriting into the twentieth century.

In the handwriting of the time, several letters had alternative shapes (allographs). The long and round forms of *s* (ſ s) and the straight and rotunda forms of *r* (r ꝛ) are examples. The choice of allograph is partly free, partly determined by orthographic context: round *s* (when used) is found only as a capital or word-finally; rotunda *r* is used after round letters such as *o* and *p*, straight *r* elsewhere. These allographs are usually not distinguished in transcription. The letters *c* and *k* too were largely interchangeable and might be considered allographs, though historically they are separate letters.

The letter *u* had a number of allographs, letter forms which are often distinguished in transcription as *u, v, 6, w* (though the form transcribed as *w* itself had several variants). In the early period these were pure allographs but over the centuries a tendency developed to reserve *u* for /ʉ/, *v* for /β/, *6* for /u/ and *w* for /w/ (or /u̯/). The distinctions were never made consistently: *u* and *v* always remained positional variants to some extent, *v* being word-initial, *u* word-internal, so that *vuud* spells /ʉβʉð/ 'obedient' (*ufudd* in modern spelling). Hence it was *f* and not *v* that eventually became the standard spelling for /β/ (modern /v/). Neither was there ever consistency in the distribution of the letters *6* and *w* over the allophones of semivowel [u̯] or [w] in falling diphthongs, rising diphthongs, and consonant clusters.

The orthography described above is underdetermined, allowing alternative spellings for many words. Usage of some letters and letter-forms changed over time. Hence details of orthography vary between individual scribes and between periods. It was usual for several scribes to work on the same manuscript, and the text was often copied from an earlier exemplar and the orthography modified to a greater or lesser degree in the process. Because of this there is rarely a high degree of consistency in the orthography of a manuscript, though the amount of inconsistency varies greatly between manuscripts.

6. The Spelling of Modern Welsh

The orthography of Modern Welsh[7] is largely phonemic. The Welsh alphabet has twenty-eight letters:

[7] Described in detail by Peter Wynn Thomas in *Gramadeg y Gymraeg*, University of Wales Press, 1996, pp. 750–798.

a	b	c	ch	d	dd	e	f	ff	g	ng	h	i	l	ll	m	n	o	p	ph	r	rh	s	t	th	u	w	y
a	b	k	χ	d	ð	e	v	f	g	ŋ	h	i	l	ɬ	m	n	o	p	f	r	r̥	s	t	θ	ɨ	u	ɨ,ə

The alphabet contains several digraphs, each counted as a single letter, with its own place in the alphabetical order.

The one-to-one correspondence between letters and phonemes is disturbed in both directions. Both *ff* and *ph* represent /f/, but the two are in complementary distribution: *ph* represents only word-initial mutated /p/, *ff* is used elsewhere. Both *u* and *y* can represent /ɨ/. This is a genuine ambiguity, due to the two earlier phonemes /ɨ/ and /ʉ/ having merged as /ɨ/ in the modern language. Modern *ty* 'house' and *tu* 'side' are pronounced exactly the same, though they were earlier distinguished as /tɨ/ and /tʉ/ respectively.

In the other direction, there are ambiguities reading sound from spelling. Though the digraph *ng* unambiguously represents /ŋ/, it is indistinguishable to the eye from the two-letter sequence *ng* which represents /ng/. The place name *Bangor* is pronounced /bangor/, though the spelling could equally well represent /baŋor/. Alphabetical order distinguishes the digraph and the two-letter sequence in a dictionary: *lleng* /ɬeŋ/ 'a legion' follows *llegach* /ɬegaχ/ 'feeble', but *llengar* /ɬengar/ 'fond of literature' follows *llên* /ɬeːn/ 'literature'. Other two-letter sequences which might be mistaken for digraphs are separated with a hyphen, e.g. *hil-laddiad* 'genocide' (cf. *teyrnladdiad* 'regicide' with no hyphen because *nl* cannot be misinterpreted as a digraph), *ufudd-dod* 'obedience' (cf. *segurdod* 'idleness'), *llygad-dyst* 'eye witness'.

The other main source of ambiguity is the letter *y* which can represent /ɨ/ or /ə/. The two are generally in complementary distribution: *y* represents /ɨ/ in a final syllable, including most monosyllables, /ə/ elsewhere, e.g. *tywyll* /ˈtəwɨɬ/ 'dark', *tywyllwch* /təˈwəɬuχ/ 'darkness'.

A minor source of ambiguity is that the three high vowels /i ɨ u/ can be either full vowels or semivowels. All three occur as semivowels in falling diphthongs, e.g. /ai̯ aɨ̯ au̯/ all occur, as in *llai* /ɬai̯/ 'lesser', *llau* /ɬaɨ̯/ 'lice', *llaw* /ɬau̯/ 'hand'. Rising diphthongs have only /i/ and /u/ as semivowels, e.g. in *iach* /i̯aχ/ 'healthy', *wedi* /u̯edi/ 'after'. Thus a sequence of high vowels may be ambiguous between a rising or falling diphthong. Tautosyllabic *wi* is always /u̯i/, but *iw* can be falling or rising: both occur in *lliwiwr* /ɬiu̯.i̯ur/ 'a dyer'. *Wy* /uɨ/ is regularly ambiguous.

Of the semivowels only /u̯/ occurs non-syllabically within consonant clusters: *gwlad*, *gwneud*, *gwraig* are monosyllabic with *w* representing /u̯/; *marwnad* and *meddwdod* are likewise bisyllabic. The spelling in such cases is ambiguous: similarly spelt *gwrol* and *gwryw* are bisyllabic with *w* representing a full vowel /u/.

7. Resolving ambiguity

Spelling is an inexact representation of spoken language. Modern Welsh orthography normally identifies words unambiguously, and the pronunciation is usually predictable from the spelling. Though older orthography is ambiguous, it is usually possible to decide which word is intended by a historical spelling, and pronunciation can mostly be deduced by comparing the historical and modern spelling. There can however be a problem when both spellings contain the same ambiguity.

7.1. Lexical ambiguity Sometimes converting mediæval to modern spelling will result in no change, e.g. the word *coch* 'red' has only one possible conversion because each of word-initial *c*, medial *o* and final digraph *ch* has only one possible phonological interpretation and each has the same modern spelling. Other words have only one possible phonological interpretation but the modern spelling is different, for instance mediæval *kar6*, modern *carw* 'a deer'. More usually phonological interpretation of the letters will produce several alternatives, e.g. in the mediæval spelling *cerdet* the letters *d* and *t* are ambiguous: the spelling could represent /kerdet/, /kerded/, /kerdeð/, /kerðet/, /kerðed/ or /kerðeð/. If we look in a modern Welsh dictionary we find only one of these: *cerdded* 'to walk', representing /kerðed/. Checking a dictionary will usually solve such indeterminacy, though not always, e.g. the mediæval spelling *bot* could represent either the verb *bod* 'to be' or the noun *bodd* 'satisfaction'.

Many words appear in text in a form other than the form listed in the dictionary: nouns have plural forms, verbs inflect for tense, number and person, adjectives for comparison, prepositions for number and person. The mediæval spelling *uu6yt* could represent /βʉuɨd/, /βʉuɨð/, /ʉβuɨd/, /ʉβuɨð/, /uβuɨd/, /uβuɨð/, etc. None of these forms will be found in a dictionary. The word is the past impersonal form of the verb 'to be' with initial-consonant mutation, *fuwyd* in modern spelling. The impersonal ending *-wyd* is added to the past form of the verb 'to be', *bu,* and the initial consonant mutates in the phrase *llawen fuwyd* 'people were merry'. To interpret such inflected word-forms correctly, we need not just a dictionary but morphological analysis, analysis of the structure of words. For each possible interpretation of the spelling *uu6yt* we need to attempt a morphological analysis. Only when the analysis succeeds is the interpretation plausible.

The program described here uses pre-existing software[8] for morphological analysis of present-day Welsh to interpret a mediæval spelling. The morphological analyser uses a set of rules of Welsh morphology along with the Collins-Spurrell Welsh dictionary to provide an analysis of the input word. To find the word represented by a mediæval spelling, each possible phonological interpretation of the spelling is converted to modern spelling and analysed by the morphological analyser. A successful analysis implies that the input was a genuine Welsh word. Usually only one possible phonological interpretation of a spelling will turn out to be a genuine word, hence disambiguation is achieved.

Occasionally more than one possible interpretation turns out to be a genuine word: the example of *bod/bodd* was given above. More frequent are ambiguities involving initial-consonant mutation. For instance mediæval *g6r* is unambiguously the word for 'man', but since *g* can represent either /g/ or /ŋ/, it could be the dictionary form or a mutated form, either *gŵr* or *ngŵr* in modern spelling. Word-initial *g-*, *d-* and *r-* are always ambiguous in this way between dictionary form and mutated form: *dyn* 'person' could be /dɨn/, identical in modern spelling, or the mutated form /ðɨn/, modern *ddyn*. *Ryuel* is 'war', either dictionary form *rhyfel* /r̥əβel/ or mutated *ryfel* /rəβel/. Mutation and inflexion occasionally interact, so that mediæval *dyvaud* could be the irregular past tense form of the verb 'to say', /dəu̯au̯d/ (modern colloquial *dywad, dŵad*), or its mutated form /ðəu̯au̯d/; or it could represent /dəβau̯ð/, modern *dyfodd*, a mutated form of the past tense of the verb *tyfu* 'to grow', /–au̯ð/ (modern *-odd*) being the regular past-tense ending.

The current implementation of the program takes individual words as input, so the only possible basis for choice between alternatives such as

[8] See John D. Phillips: 'The Bible as a basis for machine translation', in the *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, 2001, pp. 221–228.

these is textual frequency[9]: in the above example *bod* is far more frequently used than *bodd* so if we choose *bod* we will be right most of the time. If the program was redesigned to accept texts, rather than individual words, as input, a language model could be deployed to choose the most appropriate word in context in such cases. The reason this has not yet been done is that the small improvement in accuracy which would result would have only an insignificant effect on the results of the research for which the program is presently being used.

7.2. Phonotactic constraints With Welsh, some phonemic distinctions cannot be derived reliably from either modern or historical spelling. As explained above, a historical spelling is interpreted by first producing all its possible phonemic interpretations, then converting each to modern spelling and putting it through morphological analysis. Only those interpretations for which the analysis succeeds are possible words — usually only one. This reliably gives the modern spelling of the word, but will only give a determinate phonemic representation if one can be unambiguously derived by combining the textual spelling and the dictionary spelling. If both historical and modern orthography share the same ambiguity, as with the letter *y*, which represents both /ɨ/ and /ə/ in both mediæval and modern spelling, that ambiguity will remain.

Phonotactic constraints can solve some such ambiguities: the reason *y* can be used successfully for both /ɨ/ and /ə/ is that the correct pronunciation can usually be deduced from the phonological context. Certain phonological configurations are disallowed in a language, and the system allows these to be stated as phonotactic constraints on the phonemic representation.

***ng* —** /ŋ/ or /ng/? Mediæval *g* is ambiguous. The spelling *r6g* could represent /r̥ug/ or /r̥uŋ/, in modern spelling *rhwg* or *rhwng*. Only the latter appears in the dictionary, so the correct interpretation is /r̥uŋ/. This works even though mediæval *g* /g,ŋ/ and modern *ng* /ŋ,ng/ are both ambiguous,

[9] Frequency is measured by the rate of occurrence of the word (the lexeme) in a standard Welsh corpus of one million words. See N. C. Ellis, C. O'Dochartaigh, W. Hicks, M. Morgan, N. Laporte: 'Cronfa Electroneg o Gymraeg (CEG): Cronfa ddata eirfaol o filiwn o eiriau sy'n cyfrif amlder defnydd geiriau yn y Gymraeg' (2001), available from http://www.bangor.ac.uk/development/canolfanbedwyr/ceg.php.cy.

because the types of ambiguity are distinct.

More and more in later manuscripts, /ŋ/ was written *ng*, and this eventually became the standard spelling. When both the textual spelling and the dictionary spelling use *ng* ambiguously as a digraph /ŋ/ and a sequence /ng/, the phonotactic knowledge that /ng/ cannot be tautosyllabic solves the ambiguity in most cases: word-initial and word-final *ng* must represent /ŋ/, as must *ng* in some word-internal consonant clusters. This will not help in cases of intervocalic *ng*. At least one major Welsh dictionary explicitly marks the comparatively few words containing the sequence *ng* /ng/[10] and in this spirit the software here described allows phonotactic constraints to refer to wordlists, in this case a list of modern Welsh words containing the sequence *ng*. The wordlist was extracted manually from a list of dictionary words containing *ng* (digraph and sequence), but could in principle be produced automatically from a dictionary containing the information.

Vowel length was not phonemic in earlier Welsh, though vowels had long and short allophones predictable from the phonetic context. In the modern language, simplification of final consonant clusters and coalescence of originally heterosyllabic adjacent vowels, along with borrowing from English, has introduced phonemic vowel length. In modern orthography, length can be notated with a circumflex, but rarely is. The are a small number of minimal pairs distinguished only by vowel length, such as *can – cân*, *llen – llên*, *tal – tâl*, *twr – tŵr*. Most such minimal pairs are monosyllables ending in /n/ or /r/, where the word with the short vowel originally ended with a double consonant providing the context for a short vowel: the older forms were /kann kan/, /ɬenn ɬen/, etc., phonetically [kanː kaːn], [ɬenː ɬeːn], etc. While vowel length is clearly phonemic in the modern language, it is a marginal phoneme and there is considerable dialectical variation. Long vowels resulting from simplification, and most long vowels in borrowed words, are reliably long in all parts of Wales, but the details of which phonetic contexts trigger other long vowels are different in different regions.

[10] The Welsh Academy English-Welsh dictionary, by Bruce Griffiths and Dafydd Glyn Jones, published by the University of Wales Press, 2003. The publisher will not permit research using a machine-readable version of this dictionary.

Because of the regional variability and the consequent uncertainty as to exactly which vowels should be long, and also because vowel length is not phonemic in the texts being analysed, vowel length is ignored in the work described here.

***y* — /ɨ/ or /ə/?** As noted above, *y* generally represents /ɨ/ in a final syllable, including a monosyllable, /ə/ elsewhere, e.g. *tywyll* /ˈtəwɨɬ/ 'dark' vs. *tywyllwch* /təˈwəɬuχ/ 'darkness', *llyfr* /ɬɨvr/ 'book' (singular) vs. *llyfrau* /ɬəvraʉ/ 'books' (plural). Two simple phonotactic constraints restricting /ɨ/ to final syllables and /ə/ to non-final syllables capture this distribution.

However there are quite a number of exceptions to the general principle. (i) A small number of unstressed monosyllabic function words have /ə/: definite article *yr/y* , possessive pronouns *fy* 'my' and *dy* 'your', preposition *yn/ym/yng* 'in'. (ii) Before a vowel *y* represents heterosyllabic /ɨ/ e.g. *lletya* /ɬeˈtɨ.a/ 'to lodge', *cywelyes* /kəu̯eˈlɨ.es/ 'a female bedfellow'. (iii) The triphthongs *ayw* and *oyw* are /aɨ̯u̯/ and /oɨ̯u̯/ respectively. (iv) *Wy* however is ambiguous. It can represent a falling diphthong with /ɨ̯/ as in *llwyth* /ɬuɨ̯θ/ 'a load', *llwytho* /ɬuɨ̯θo/ 'to load' ; or a rising diphthong with variable *y* as in *chwyd* /χu̯ɨd/ 'vomit', *chwydu* /χu̯ədɨ/ 'to vomit'.

The words in (i) are simply exceptions, and can only be captured by a wordlist or a list of word-specific constraints. Types (ii) and (iii) are subregularities and as such amenable to simple phonotactic constraints. The ambiguity of *wy* in (iv) is not rule-governed — there is no way to predict from the spelling that the *y* in *chwydu* is /ə/ rather than /ɨ/. The spelling *wy* appears in very many words so it is not easily solved with a wordlist. Since the falling diphthong is the commoner, it is (incorrectly) generalised here, so that the *y* in *wy* is (incorrectly) always /ɨ/.

Rising and falling diphthongs Neither historical nor modern orthography distinguish reliably between rising and falling diphthongs. The same ambiguity is shared by both writing systems and there is no simple way to solve it.

There is a tendency in some historical texts, particularly later in the period, to reserve the letter-forms *i, 6* for the full vowels /i, u/, and use *y, w* to represent the semivowels /i̯, u̯/ (of course *y* also represents /ɨ/ and /ə/ in all texts). Unfortunately few if any texts are consistent enough to rely upon. In the example text below for instance, which is comparatively consistent, the identical diphthong /uɨ̯/ is written *wy* in the first word but *6y* in the seventh; the spelling *kar6* quoted above has *6* though linguistic evidence suggests that the word was monosyllabic /karu̯/. The example text below has *gweith* /gu̯eiθ/ 'occasion', but a few paragraphs later the same word is spelt *g6eith*. Modern spelling is no more helpful: it occasionally uses a diacritic to distinguish minimal pairs as in *gŵyr* /guɨ̯r/ 'he/she/it knows' vs. *gwŷr* /gu̯ɨr/ 'men', but most diphthongs are not so marked. There are also weak phonotactic clues: rising diphthongs have a slightly restricted occurrence so that phonological context and morphemic structure will sometimes disambiguate, but by no means always. It would be possible to combine these three types of clue — from historical spelling, modern spelling, and phonotactics — to distinguish between /i ɨ u/ and /i̯ ɨ̯ u̯/, but many unsolved cases would remain.

The solution adopted here is simply to ignore the distinction between semivowel and full vowel, in line with modern Welsh spelling, which consistently uses the same letters for both. IPA output has three phonemes /i ɨ u/. These occur in both rising and falling diphthongs, which are not distinguished.

7.3. Language change The Welsh language has changed comparatively little over the last few centuries; nevertheless the changes which have occurred are enough to cause problems using a modern morphological analyser on mediæval texts. The types of change which can cause problems are in vocabulary, morphology, and pronunciation.

A small proportion of words have become obsolescent, though most still appear in dictionaries, and almost all can be found in the Collins-Spurrell dictionary used here. It is in any case simple to add a supplement to the dictionary. One could argue that missing words should have been in the dictionary to start with.

Change in morphology creates more of a problem. A few verbal endings have changed over the centuries. For instance, plain past-tense forms of verbs in Old Welsh mostly ended in *-s*. The basic past-tense ending was in modern spelling one of *-as/-es/-is/-wys*, the choice determined partly lexically but mostly by the vowel of the verb stem. During the mediæval period, a new ending *-odd* spread at the expense of the others[11], so that *gwelas* ‘saw’ is now *gwelodd*, *cafas* ‘took’ is now *cafodd*, *dodes* ‘put’ is now *dododd*, *rhoddes* ‘gave’ is now *rhoddodd*, *torres* ‘broke’ is now *torrodd*. Analogical changes have affected a few other endings: second person singular imperfect *-ut* has become *-et*; first-person plural *-am* on verbs and prepositions has become *-om*; third-person plural *-unt* on prepositions has become *-ynt*; and the infix *-ys-* in plural past and in pluperfect endings has become *-as-*.

These changes could be handled either by adding extra rules to the morphological analyser, or by modernising the ending so that the morphological analyser sees the modern Welsh form. The latter course was chosen in keeping with the goals of this project, which included processing mediæval texts with existing tools designed for modern Welsh. Modernising endings using rules in the same format as the spelling rules is simple and means the morphological analyser can be used unchanged. The disadvantage is that there is a possibility of incorrect conversion because inflectional endings cannot be distinguished from coincidentally similar words. Endings must be modernised before morphological analysis, but whether the latter part of a particular word is an inflexion or not can only be known by morphological analysis. For instance a hypothetical mediæval spelling *anues* would have phonemic interpretations /anwes, anβes/. The first would be accepted by the morphological analyser as the noun *anwes* ‘a caress’, and the second, after converting the *-es* to *-odd* in case it might be an inflexion, would be accepted as the noun *anfodd* ‘dissatisfaction’. Hence both /anwes/ and /anβes/ would be accepted as genuine words, the latter spuriously since there is not and never has been a word *anfes*. Had the original form /anβes/ been passed directly to a morphological analyser containing rules to handle obsolete endings such as verbal *-es*, it would have been rejected because there is no verbal stem *anf-*. A way around this problem would be to state the morphological modernisation rules in a format which allowed for conditions to be passed to the morphological analyser. At present rules for spelling and morphology use an identical format, making for simplicity and efficiency. If, for instance, the rule converting mediæval

[11] The spread is documented by Simon Rodway in his paper ‘Two developments in medieval literary Welsh and their implications for dating texts’ in *Yr Hen Iaith: studies in Early Welsh* edited by Paul Russell for Celtic Studies Publications, 2003, pp. 67–74

-as/-es/-is/-wys to modern *-odd* could specify that morphological analysis of the word must be as a verb, the problem would be solved.

Regular sound change is handled without problems by the spelling conversion program. Long *nn* and *rr* have become short word-finally and before consonants and semivowels. This is an important change because the number of words affected is large and includes common words such as mediæval *h6nn*, modern *hwn* 'this', *penn* 'head, beginning' and *byrr* 'short'. The diphthong /eʉ/ in final syllables (including monosyllables) has changed to /aʉ/. The diphthong /ei/ in final syllables (including monosyllables) has changed to /ai/ except in words ending in a consonant cluster other than /ft, nk/ and usually /nt/. With /nt/ both forms are found, e.g. modern Welsh has *peint* 'a pint' but *braint* 'privilege'. The French borrowing 'tournament' can have either diphthong: *twrnamaint/twrnameint.*

Lexically-conditioned sound change is harder to handle neatly. The diphthong /au/ in the final syllable of most polysyllabic words has changed to /o/. Exceptions include the nominal derivational ending *-awd* in e.g. *traethawd* 'an essay', *cyfaddawd* 'a compromise' and *pedwarawd* 'a quartet'; and some transparent compounds, e.g. *anffawd* 'misfortune' (from *an+ffawd*), *gwerthfawr* 'valuable' (from *gwerth+mawr*). Other phonologically similar words have changed: modern *gwaelod* 'bottom' and *tafod* 'tongue' both used to end in /aud/. This is an important change because it affects the regular past-tense form of verbs discussed above: *gwisca6d* /gwisgauð/ becomes modern *gwisgodd* /gwisgoð/ 'wore', *kerda6d* /kerðauð/ becomes *cerddodd* /kerðoð/ 'walked'. Since exceptions to this change are not entirely predictable, both the original and the modernised forms are passed to the morphological analyser — one or the other, occasionally both, will be found in the dictionary. Where both forms are found in the dictionary, they will normally be variant pronunciations of the same word, e.g. *graslon/graslawn* 'gracious' , *Ionor/Ionawr* 'January', *union/uniawn* 'straight', *ymado/ymadaw* 'depart'. In only a handful of cases are there two different words: *drygnaws* 'a bad atmosphere' and *drygnos* 'a bad night' are literary compounds of *drwg* 'bad' and *naws* 'atmosphere' or *nos* 'night'; *pennawd* 'a heading' has the nominal derivational ending *-awd* discussed above, and *pennod* 'a chapter' has original *-o-*. Since output in both IPA and modern spelling is derived from the original spelling, not from the dictionary word, these doublets should not cause problems.

A difference which lies on the border between orthographic change and sound change is the epenthetic vowel commonly written in word-final consonant clusters in mediæval Welsh. This is not usually written in modern Welsh, though often heard in pronunciation, particularly colloquially. An example is mediæval *geiuyr* modern *geifr* ‘goats’. The modern pronunciation can be either /geivr/ or /ˈgeivir/. In mediæval poetry this word would be treated as monosyllabic: an epenthetic vowel was not counted as a syllable. In modern Welsh, the epenthetic vowel, when pronounced, has the quality of the preceding vowel, but in mediæval orthography it is usually written *y*, suggesting a neutral vowel. This and the non-syllabic treatment in poetry suggest a pronunciation something like /geißə̯r/. Other examples are *ochyr*, modern *ochr* /oχr/ ‘side’ and *kynnedyf*, modern *cynneddf* /ˈkəneðv/ ‘instinct’. The epenthetic vowel normally appears only in word-final clusters: the plural forms of the above examples are *ochrau* and *cyneddfau*, with no epenthetic vowel. Of course a *y* in a final syllable does not always represent an epenthetic vowel: *cledyf* is disyllabic, modern *cleddyf* /ˈkleðɨv/ ‘sword’.

A sound change which is ignored in the present work is the loss of rounding in the vowel /ʉ/. As noted above, with loss of rounding, the vowel came to be pronounced identically to /ɨ/, so that the words *llus* ‘bilberries’ and *llys* ‘a court’, originally /ɬʉs, ɬɨs/ respectively, came to be pronounced identically as /ɬɨs/. In parts of South Wales the change has gone further so that the three written vowels /i ɨ ʉ/ are all pronounced /i/. None of these changes are represented in any form of Welsh spelling, so they are ignored here. The IPA representation of modern Welsh /ɨ, ʉ/ may look incongruous to a native speaker, but it could be argued that at an abstract level this representation is correct, since the two vowels behave differently morphophonemically: when an ending is added /ʉ/ remains unchanged as in *llusa* /ɬʉsa/ ‘to collect bilberries’, *llusen* /ɬʉsen/ ‘a single bilberry’; /ɨ/ however changes to /ə/ as in the plural form *llysoedd* /ɬəsoeð/ and the adjective *llysol* /ɬəsol/ ‘pertaining to a court’, It is anyway a simple matter to specify that the phoneme /ʉ/ is phonetically [ɨ] (or [i] in South Wales).

8. Example text

Below is the beginning of the mediæval story *Pwyll*, the first of the collection known as the Mabinogi, from a manuscript of the late fourteenth century[12] .

> Pwyll penndeuic dyuet a oed yn argl6yd ar seith cantref dyuet. a threigyl gweith yd oed yn arberth prif lys ida6. a dyuot yn y uryt. ac yn y ved6l uynet y hela. Sef kyfeir o e gyuoeth a vynnei y hela glynn cuch.

In English: Pwyll Prince of Dyfed was lord of the seven districts of Dyfed. Once he was at Arberth, a main court of his, and it came into his mind and his thoughts to go hunting. The part of his country he wished to hunt was Glyn Cuch.

The next example has been converted automatically from the original spelling to IPA:

> /puɨɬ penndeβig dəβed a oeð ən argluɨð ar seiθ kantreβ dəβed . a θreigl gueiθ əð oeð ən arberθ priβ lɨs iðau . a dəβod ən i βrɨd . ak ən i βeðul βəned i hela . seβ kəβeir o i gəβoeθ a βənnei i hela glɨnn kʉχ ./

Converting the above from IPA to modern spelling gives the original phonemes and morphology in modern spelling:

> pwyll penndefig dyfed a oedd yn arglwydd ar seith cantref dyfed . a threigl gweith ydd oedd yn arberth prif lys iddaw . a dyfod yn i fryd . ac yn i feddwl fyned i hela . sef cyfeir o i gyfoeth a fynnei i hela glynn cuch .

The program can also output the modernised pronunciation and morphology used for dictionary lookup. This is useful to speakers of modern Welsh if the object is just to read the text. Punctuation and capitalisation are here modernised manually.

> Pwyll Pendefig Dyfed a oedd yn arglwydd ar saith cantref Dyfed. A threigl gwaith ydd oedd yn Arberth, prif lys iddo. A dyfod yn i fryd ac yn i feddwl fyned i hela. Sef cyfair o'i gyfoeth a fynnai i hela, Glyn Cuch.

Note that the third-person possessive pronoun *ei* is spelt the same as the

[12] The version in the manuscript Oxford Jesus College 111 (The Red Book of Hergest), beginning on page 175r. As of 21st September 2012, the transcription can be seen at http://www.rhyddiaithganoloesol.caerdydd.ac.uk/en/ms-page.php?ms=Jesus111&page=175r, and a photograph of the original at http://image.ox.ac.uk/show?collection=jesus&manuscript=ms111.

preposition *i* ‘to’ here — in earlier Welsh (as in modern colloquial Welsh) they were the same and cannot be differentiated by orthography alone.

Since only pronunciation and morphology are modernised in this version of the text, it could perhaps be useful for research into idiom, sentence structure, and discourse structure, using computational tools designed for modern Welsh.

9. Future work

Small incremental improvements to the spelling rules can be expected as more texts are run through the system and errors noticed in the results. However it is expected that the main improvement to the accuracy of the system will come though better choice from alternative interpretations of ambiguous spellings. As explained above, where the spelling of a word is ambiguous, the present system makes its choice by textual frequency. The example earlier was the mediæval spelling *bot*, which can be interpreted as /bod/, modern *bod*, the infinitive of the verb ‘to be’, or as /boð/, modern *bodd* ‘satisfaction’. *Bod* is far more frequently used than *bodd* so if we choose *bod* we will be right most of the time. This method is forced on the system because it takes each word separately, out of context. This is an efficient way to work because earlier results can be (and are) re-used: most of the words in any text are used several times, and the same analysis is used for each occurrence.

A more accurate method is to consider the word in context using a *language model*. Most widely used are statistical language models, which count occurrences of word collocations in large text corpora. From this the probability of different words appearing next to each other can be calculated. In the present application, the probabilities would be used to choose the most probable of alternative interpretations of a mediæval spelling in the context of the surrounding words.

As noted above, ambiguities involving initial-consonant mutation are a particular problem in interpreting mediæval Welsh spelling. Statistical language models were originally developed for English and have been applied to various other languages, but it is not clear that they would handle mutation very well. Various different approaches need trying here. Also as noted above, mutations are often ignored (spelt as the base form) in mediæval orthography. How is the phrase *y tat a y urodyr* ‘his father and his brothers' to be interpreted? The word *y* 'his' causes mutation of the following noun, so that /brodɨr/ mutates to *urodyr* /βrodɨr/. Likewise *tat* /tad/ 'father' is expected to mutate to /dad/ but here no mutation is written. We can never know for certain what the mediæval pronunciation was, but the linguistic evidence suggests that the mutation must have been obligatory and that the pronunciation was indeed /dad/. Does this justify notating the pronunciation as /i dad a i βrodɨr/, in modern spelling *ei dad a'i frodyr*?

10. Appendix: the Welsh spelling rules

The format of spelling rules is here exemplified with the rules needed to interpret the mediæval spelling *threigyl.* Each spelling rule shows the IPA and its mediæval representation.

```
mediæval("θ","th").
mediæval("t","t").
mediæval("d","t").
mediæval("ð","t").
mediæval("h","h").
mediæval("r","r").
mediæval("r̥","r").
mediæval("e","e").
mediæval("i","i").
mediæval("g","g").
mediæval("ŋ","g").
mediæval("ɨ","y").
mediæval("l","l").
mediæval("gl","gyl").
```

Software written in the programming language Prolog converts each equivalence in the above format into a Prolog clause expressing the equivalence using a pair of difference lists. The resulting collection of clauses can then be used in the manner of a Definite Clause Grammar[13] to

[13] Definite Clause Grammar is a widely-used formalism for natural-language processing. A grammar commonly contains rules of syntax and is used for grammatical analysis of strings of words. Here we treat the individual letters as the units to be analysed, and the analysis is simply the phonemic representation.

convert words or whole texts in either direction between IPA and mediæval spelling. The above collection of rules would produce twenty four possible interpretations of *threigyl*, including many obviously incorrect ones such as /ðhreigl/, /ðhr̥eigl/, /dhreigl/, /dhr̥eigl/. Equivalences can be stated to be obligatory with an exclamation mark (the Prolog 'cut'), so that

```
mediæval("θ","th"),!.
```

will ensure that *th* is always interpreted as /θ/, reducing the number of interpretations to six: /θreigɨl/, /θreiŋɨl/, /θreigl/, /θr̥eigɨl/, /θr̥eiŋɨl/, /θr̥eigl/. The first three are all reasonable interpretations of the spelling. The latter three seem phonotactically unlikely, if not impossible; but the mediæval spelling of /r̥/ is *r*, and the only reasonable phonotactic condition that comes to mind is one banning /r̥/ in word-initial consonant clusters. Stating such phonotactic conditions in the spelling rules would speed analysis up by reducing the number of interpretations to be checked, but of course phonotactically impossible words will necessarily fail morphological analysis so the final result will be the same. Anyway, /θreigl/ is the only one of the six interpretations with a successful morphological analysis: it is a mutated form of modern *treigl* 'a turn'.

The rules needed to convert /θreigl/ into modern orthography are

```
modern("θ","th"),!.
modern("r","r").
modern("e","e").
modern("i","i").
modern("g","g").
modern("l","l").
```

These rules applied to the phonemic representation produce the modern spelling *threigl*.

Lookahead is also available on spelling rules:

```
mediæval("gw","gu",next("aei")),!.
```

states that *gu* represents /gw/ before a vowel;

```
mediæval("b","p",end).
```

states that *p* can represent /b/, but only at the end of a word.

The mediæval rules are intended for texts written in Middle Welsh in general. As noted above, orthography changed to some extent with time, and individual scribes and manuscripts are often more consistent than a general description would suggest, so that accuracy can be improved by tailoring the rules to a particular manuscript or group of manuscripts. An area in which this is clear is the representation of the phoneme /ð/. In some manuscripts, and in early printed books, a special letter, transcribed *δ*, is consistently used for this phoneme and *d* represents only /d/. In other manuscripts /ð/ is spelt *d* word-initially and either *d* or *t* elsewhere, many manuscripts being consistent in using one or the other, either *d* or *t*. It may be convenient to disable spelling rules inappropriate to the text being analysed — note that this is always a matter of disabling (i.e. temporarily removing) a rule, e.g. a manuscript which never uses *t* to represent /ð/, does not need the rule

```
    mediæval("ð","t").
```