Building and annotating corpora of spoken Welsh and Gaelic[1]

Andrew Wilson, Celia Worth
Department of Linguistics and Modern English Language, Lancaster University

1. Introduction

The indigenous minority languages of the British Isles and Ireland (or "BIMLs") are becoming increasingly widely used in both public and private life, in some cases within a statutory framework. Speech and language technology applications for these languages are therefore now becoming an urgent need. These are needed not only for monolingual content management, but also to aid translators and interpreters since the BIMLs are nearly always used in bilingual contexts alongside English. To develop such applications, basic language resources (such as corpora of machine-readable texts, machine-readable dictionaries, speech databases and so on) are required.

The LER-BIML project at Lancaster is helping to address this need. Two recent EPSRC-funded projects at Lancaster (MILLE and EMILLE[2]) have provided a great service to the ***non-indigenous*** minority language communities in the UK by locating existing resources, investigating end-user needs and wants, examining basic technical issues and beginning to generate appropriate resources. However, no such consolidated survey or examination of issues has yet been undertaken for the BIMLs.

Within the project we have recently undertaken two surveys, one to locate existing machine-readable resources and tools for these languages, the other to investigate end-user needs and wants. In our paper, we will present the results of these surveys. Secondly, we are building two pilot corpora of spoken Welsh and Scottish Gaelic, focussing in particular on “task-oriented” spoken activity types. We will discuss in our paper the issues raised in sampling, building and annotating our corpora, including our work on developing a set of EAGLES-conformant part-of-speech tagsets for all the BIMLs with which we shall then tag the two corpora.

2. The Languages

The indigenous minority languages of the British Isles and Ireland (BIMLs) are those languages other than English that are autochthonous to the territory of the United Kingdom, the Republic of Ireland and the Isle of Man. They comprise Cornish, Scottish Gaelic, Irish, Manx, Scots, Ulster Scots (Ullans) and Welsh.

Along with most of the other major languages of Europe, these languages all belong to the Indo-European language family. However, they belong to different sub-branches of Indo-European. Cornish, Scottish Gaelic, Irish, Manx and Welsh belong to the Celtic branch of Indo-European. Celtic, in turn, divides into two distinct subgroups: P-Celtic (or Brythonic) and Q-Celtic (or Goidelic). Cornish and Welsh are P-Celtic languages, whilst Scottish Gaelic, Irish and Manx are Q-Celtic languages. The names for the two subgroups arise from the reflexes of the Proto-Indo-European *q^w. In P-Celtic this became a bilabial stop (/p/), whilst in Q-Celtic it became a velar stop (/k/): compare the word for ‘four’ in Welsh (*pedwar*) and Scottish Gaelic (*ceithir*), both descended from Proto-Indo-European **q^wetwor*. There also exists a further P-Celtic language, Breton, which is spoken in north-western France. As this is not indigenous to the British Isles or Ireland, it does not fall within the scope of our project. For the same reason the Channel Island languages Jeriaise and Guernsiaise are not included.

In contrast, the two varieties of Scots are not Celtic languages, but belong instead to the Germanic branch of Indo-European. Although they are very closely related to some dialects of English, they are recognized as separate languages. Welsh is widely used in all sectors of life in Wales and has joint official status with English. Although only 18.7% of the total population consider themselves to be Welsh-speaking (1991 Census figures), this figure rises to 43.7% in Dyfed and 61% in Gwynedd.

[1] This research is funded by the Engineering and Physical Sciences Research Council, GR/R63301/01.

[2] http://www.emille.lancs.ac.uk

Welsh is a compulsory subject in the national curriculum in Wales. Furthermore, as a consequence of the Welsh Language Act 1993, the public sector must offer its services bilingually in Wales. Irish and Ullans have both been given increased recognition in Northern Ireland under the Northern Ireland Agreement, and the Northern Ireland Executive has pledged to promote them. According to the 1991 Census, only 1.4% of the population are speakers of Scottish Gaelic, although this figure will be rather higher regionally since almost all Gaelic speakers live in the highlands and western islands. Following devolution the Scottish Executive has given considerable priority and funding to maintaining and encouraging the use of Gaelic. For instance, the report of the Executive's Gaelic Taskforce (*Gaelic: Revitalising Gaelic a National Asset*) states that "as a foundation stone in the building of the new Scotland, the Gaelic language will be an integral and dynamic component of a robust and self-assured community with economic and social stability and pride in its linguistic and cultural identity" (Education Department, Scottish Executive 2000). Both Manx and Cornish have undergone revival, and there are now once again some native speakers who are bilingual in these languages and English. In terms of direct transmission, the last native speaker of Cornish died in the 18th century and the last native speaker of Manx in 1974. However, the use of Manx is receiving substantial support from the Isle of Man government and it is being taught again in schools. Cornish, at present, has no such public-sector support, although there is a strong pressure group (Agan Tavas) seeking to acquire this.

3. Surveying existing resources

The BIMLs are presently in various states of health as regards extent of current use. This initial survey of electronic resources was concerned with the manner and extent of their distribution and the level of current activity surrounding them.[3]

3.1 Methodology

The BIMLs as noted fall into the two language families of Celtic and Scots. As it has not previously been general practice to regard and treat these two families as one individual cohesive group, there is no one central governing body responsible for coordinating and making their resources and facilities widely available. Consequently work on these languages is widespread and yet sparsely distributed. It was evident from the outset that the Internet was going to be the main starting point for locating material and suitable contacts. It was intended that our main focus should be on text corpora and machine-readable texts, but also concentrating on locating speech databases, term banks, lexicons and language analysis tools such as taggers and parsers. The Internet has the obvious benefit of supplying all texts in electronic form. Search engines proved a useful initial means for generating a general idea of the volume of material available for the BIMLs, and from here subsequent leads and contacts were then followed up. What was apparent was that most of this material would require sifting; whilst there is a wealth of information *in English* about the BIMLs, the interests of this project lie, however, with resources directly available in the languages themselves.[4]

3.2 Areas of resource

The Irish, Ulster Scots and Welsh languages are all well represented by having official Language Boards and Agencies whose websites occur entirely in bilingual format and offer useful links.[5] Whilst no such official bodies exist for the remaining BIMLs, there are organisations whose work is invaluable to promoting their respective language such as *Agan Tavas* for Cornish, *Cli* for Scottish Gaelic and *Mannin.Org.Im* for Manx. These sites include histories of the language, manuscripts, reference materials, news items, and merchandise. Most of these pressure groups offer discussion fora and mailing lists in English and in the language in question where reports are archived for public reading. Webzines are particularly popular and are a very good source of BIML resources as they form the focus of special interest groups with a targeted loyal audience. Examples are *An Gannas* for Cornish speakers, *Beo* for Irish, *Wir Leid* for Scots and *Ullans.com*. They offer comment, stories, puzzles, quizzes, jokes, polls, and reviews amongst others. The *Mercator* project which serves as an

[3] See Working Paper 1 (Wilson and Worth 2002) for full details of all URLS cited.

[4] To avoid confusion the term "BIML data/material" will refer exclusively to material written in the languages themselves, rather than in English.

[5] http://www.bnag.ie/index.htm; http://www.ulsterscotsagency.com/; http://www.bwrdd-yr-iaith.org.uk/

information network for minority languages of the European Union profiles Cornish, Scottish Gaelic, Irish and Welsh amongst these languages and directs towards associated resources.

3.2.1 Current Projects

Aside from MILLE and EMILLE there are several other various projects in progress sharing themes with those of LER-BIML. These include CELT[6], an online database of ancient and contemporary Irish literary and historical texts, the National Corpus of Irish incorporating 15 million words from a variety of contemporary books, newspapers, periodicals and discourse, marked up in accordance with the PAROLE encoding standards and MELIN which has produced dictionaries, grammars, spellcheckers and terminology lists for the initial four EU minority languages, Irish, Welsh, Catalan and Basque. Their sites offer good links to BIML data and other websites. The Oxford Text Archive[7] has a catalogue of several thousand electronic texts and linguistic corpora in a range of languages including standard reference works and mono and bilingual dictionaries. The Universities of Edinburgh and Glasgow have recently begun collaborative work on the SCOTS[8] project that aims to build a collection of electronic spoken and written texts for the languages of Scotland incorporating Scots, Scottish English and Gaelic but focussing primarily on Scots. Previous work in this field has included a one million word lexical database and frequency count for Welsh (CEG), developed at Bangor from a broad range of modern Welsh text types (Ellis et al 2001), and two speech databases for Welsh produced at Edinburgh and Swansea (Williams 1998, 1999; Jones et al 1998).

3.2.2 Education

In assessing the relative volume of BIML resources, it was clear that Welsh has the most widely available material. This is a direct result of the Welsh Language Act 1993 which states that the public sector must offer its services bilingually, something that much of the private sector has now also adopted. In locating BIML data we therefore took specific note of whether the material appeared solely in the original language, or in bilingual format. One particular good example of Welsh parallel text is to be found on university homepages. Sabhal Mór Ostaig, a Further Education College on Skye, offers a comprehensive index to key resources in Scottish Gaelic and is a primary gateway for BIML data. ACCAC has a parallel site of exhaustive Welsh language and bilingual educational sites for ages 4-18. Dublin City University has a centre, *FIONTAR*, which administers academic programs entirely through the medium of Irish, and has a bilingual website outlining this. The Centre for Manx Studies obliged in sending an extensive list of Manx resources, which included the main resources page developed by the Manx Language Officer at the Department of Education. This page offers links to short stories, the Manx Language Society's newsletter *Dhooraght* and the magazine *CARN*, dictionaries, grammars and glossaries.

There are interactive and self-teaching language courses available for all the BIMLs, ranging in style and intensity from the colloquial to the more grammatically orientated instruction, but which cater for all levels of learner. Most sites also supply various vocabulary and phrase lists and glossaries.

3.2.3 Government

All Welsh council and parliamentary sites including the National Assembly are legally required to be presented in bilingual format, and the Scottish Parliament and Northern Ireland Executive are following their lead.[9] Health Authorities in Wales that have developed their own websites also present them in this bilingual format. There is obviously much more "official" material available for Welsh than any of the other BIMLs, but Gaelic and Irish are certainly increasing their profile.

[6] http://www.ucc.ie.celt

[7] http://ota.ahds.ac.uk

[8] http://www.scottishcorpus.ac.uk/

[9] http://www.wales.gov.uk/index.htm; http://www.scottish.parliament.uk/; http://www.nics.gov.uk/

3.2.4 Media

Media resources are strong in BIML data with various online newspapers, radio broadcasts and television schedules. Newspapers vary from presenting their entire content in the original language, like the Welsh weeklies *Y Cymro* and *Golwg*, and the Irish weeklies *Foinse* and *Lá*, to producing special reports like *An Phoblacht*, Ireland's leading weekly Republican newspaper (archived).[10] BBC Online is available in Welsh, BBC Scotland has pages in Gaelic, BBC Cornwall produces a weekly audio five minute news bulletin and the Welsh and Irish stations S4C and TG4 respectively have bilingual sites.[11] The recently launched BBC4 channel broadcasts programmes in Gaelic. *RTE*, the national Irish radio service, provides bulletins online and *Raidió na Gaeltachta* has an extensive bilingual site which supplies audio downloads of its broadcasts 24 hours a day. There is a text and audio weekly review of *Manx Radio*. Search engines and web browsers can be used in Welsh, Gaelic and Irish: the *Opera* Web Browser is available in Irish, Scottish Gaelic and Welsh as well as Breton, and a detailed guide to Welsh software can be found at *Meddal*.

3.2.5 Arts and Literature

Much of the literary corpora is reproduced on various Internet sites. Poetry and song lyrics are favourites, some with audio recordings, and are often used alongside language lessons. Short stories have been written by contributors to the webzines. There are several online book inquiry services and bookshops including the Welsh Books Council, and the National Library of Wales has a bilingual website. Details and adverts for film and music festivals and local events are often posted in the respective BIML and linked through webzines and special interest groups. The National Museums and Galleries of Wales website is in parallel format.

3.2.6 Religion

As regards religious texts, the Cornish Language Board (in conjunction with the Bishop of Truro's Ecumenical Advisory Group for sevices in Cornish) has translated several books of the Bible into Cornish. Various excerpts have also been translated into Manx and there are Manx, Scottish Gaelic and Welsh versions of the Book of Common Prayer.[12]

3.2.7 Tools

Online dictionaries are widely available for all the BIMLs and spellcheckers have been developed for Cornish, Scottish Gaelic, Irish, Manx and Welsh. Canolfan Bedwyr at the University of Wales, Bangor has published extensively in the area of specialist terminology dictionaries, and include amongst others glossaries of finance and education terms produced for the National Assembly of Wales.

3.3 Conclusion

There is a reasonably healthy volume of BIML resources available on the Internet, but as predicted Welsh is the most prominent and prolific of these languages. This prominence will be a direct result of the Welsh Language Act 1993. As there is no such official legislation as regards the other BIMLs there is a strong bias amongst these languages towards popular entertainment, particularly webzines, and archaic literature as they still rely primarily on specialist interest. They are, however, increasing their profile within more official and political contexts. The paucity in particular of Scots and Ulster Scots resources is probably best accounted for by the difficulty in distinguishing the boundaries between what is a dialect of English and what is something entirely recognisable as "Scots". This is a much disputed issue.[13] Bilingual text tends to be the most favoured format of presenting BIML data, as it caters for the interests of the native BIML speaker, whilst also recognising the need to extend its

[10] http://www.ycymronow.co.uk/; http://www.foinse.ie/; http://www.irlnet.com/aprn/

[11] http://www.s4c.co.uk/; http://www.tg4.ie/

[12] www.justus.anglican.org/resources/bcp/bcp.htm

[13] In 2001 the UK Government signed and ratified the European Charter for Regional or Minority Languages. As a result in Scotland and in Northern Ireland, Scots is recognised as a regional language, except that in Northern Ireland it is referred to as Ulster Scots.

resources to the non-BIML speaker. Much material exists on the subject of these languages, but in English. The positive results of this survey therefore support the project's claims of the increasing profile of the BIMLs in the UK and Ireland today.

4. Surveying end user needs

4.1 Methodology

Previous experience has taught that the most effective way of ensuring as wide a scope as possible of potential end-users would be by means of a web questionnaire posted on the project website. Notice of this questionnaire was emailed to over fifteen Internet bulletin boards and mailing lists including HUMANIST, CORPORA, TERMCELT and CELTLING. This secured the questionnaire being disseminated to all the BIML linguistic regions, and also outside of the British Isles and Ireland to groups working with the BIMLs as non-indigenous minority languages. The questionnaire focuses on the response to language engineering resources and corpus construction for the BIMLs by varying groups of users and as such forms the basis of the project's Working Paper 2 (Wilson and Worth 2002).

4.2 Results

There were 127 responses, 57 of which were interested in receiving feedback from the survey. This would be done by emailing out a copy of the report.

Scottish Gaelic had the highest demand for corpus resources; there was no demand at all for Ulster Scots.[14] There was strong interest in seeing more availability of resources for Breton and Shetlandic, with individual requests for the Channel Island languages and Romany amongst others.

BIML	No. responses
Scottish Gaelic	86
Irish	65
Welsh	63
Scots	47
Cornish	43
Manx	42
Ulster Scots	0

Table 1 Responses showing the areas of demand for (more) BIML corpus resources.

A bilingual corpus was the most favoured corpus type:

- to contain English alongside the BIMLs in question,
- to contain sentence-aligned translations of the same texts in each language.

Most wanted to see an equal balance of written and spoken data built for the BIMLs, and for this to be done within general balanced corpora rather than in genre specific corpora. For genre specific corpora news, history and fiction proved the most popular areas of interest. Whilst people thought it important to envisage the ideal of all types of genre being made available for the individual languages, suggestions other than those proposed on the questionnaire included arts and music, youth culture, environment, travel, technology, oral literature, folklore, food and drink, media and advertising and religion and ethics.

As regards linguistically annotating the data, most would prefer just plain text. However, of the methods of annotation on offer, part-of-speech was the next most popular. The respondents would be

[14] This zero return could be explained by the presumption that the respondents identified more with the category Scots than Ulster Scots, or from a more negative point of view, that there is simply a lack of interest in this area.

happy with anything that could be made available, although there was special mention of IPA and metaphorical and dialect annotation. The question of textual mark-up returned the highest number of nil responses, but amongst those who were interested in seeing mark-up, html was the favourite.

The Internet was the favourite medium for receiving corpus data, with the CD a close second.

On the issue of the listed features and their perceived importance within a corpus, the general consensus was that there was ‘no opinion’ on their preferred status. The only features which received a majority rating of ‘essential’ were the header elements ‘author’, ‘source of data’ and ‘extent of language usage’. The spoken data features attracted the only number of ‘not wanted’ responses.

The majority of respondents were linguists rather than language engineers. Applications which the language engineers envisaged using the BIML data to build included frequency tables, speech synthesis and recognition, spelling, style, syntax and grammar checkers, bilingual dictionaries and lexica and pedagogical tools. Questions the linguists wanted to explore with the data included effects of linguistic shift and borrowings; frequencies and variations of syntactic structures, dialect, registers and discourse; patterns of code switching across genres; reported versus actual usage; patterns of growth and decline; reception by young people. Suggested support tools included concordances, checkers, search and recognition tools, glossaries, taggers, text aligners and audio and video files.

The optimistic end result was that there was an overwhelming majority (63%) that people were very likely to be working with the BIMLs in the future.

Future activity	**No. responses**
Very likely	81
Possibly	25
Unsure	13
Probably not	6
Very unlikely	2

Table 2 Response to how many linguists plan to work with BIML corpus data in the future.

4.3 Conclusion

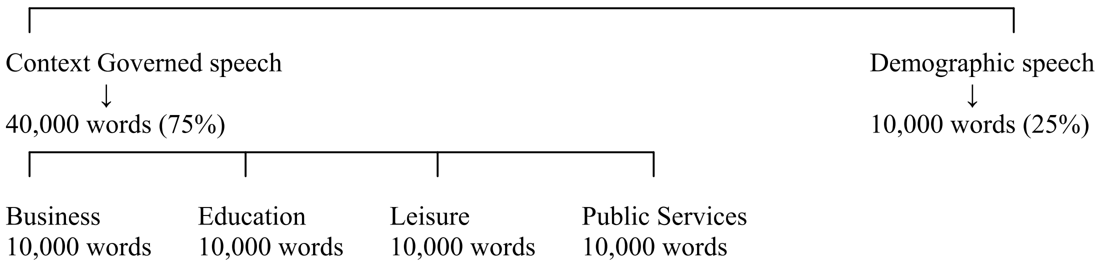
The encouraging number of responses to the survey indicated that work in progress and current activity regarding the BIMLs is healthy and positive. The higher demand for Scottish Gaelic, Irish and Welsh most likely correlates with the more specialised mailing lists that cater specifically for these languages and are in much wider use. An amendment for a future survey of this type would be to ask the respondents to indicate from which mailing list they received details of the questionnaire to secure a better overview of the distribution of the respondents. It would also be beneficial to determine the professional status of the respondents to indicate from what type of linguistic background they are approaching the data. These factors could help resolve why these results might appear to go somewhat against expectation in light of the results of the previous survey of existing resources.

5. Creating the corpora

The BIMLs, especially the spoken languages, have been relatively overlooked within a corpus-based framework and in order to investigate the practical and technical issues that they raise for language engineering development it is necessary to build corpora as a testbed for such research. Two corpora are currently under development, one for Welsh and one for Gaelic, each of 40,000 – 50,000 words. Size is clearly not the aim of this phase of work; rather the focus is on practical and technical feasibility. We are following the framework of the British National Corpus World Index for spoken data collection with a bias towards task-orientated data collection. 75% of the data for each corpus is being drawn from context governed “task-oriented” spoken activity types, and the other 25% from general everyday conversational speech.

LER BIML Spoken Corpus

40,000 – 50,000 words each for Welsh and Scottish Gaelic



Context Governed speech

↓

40,000 words (75%)

Demographic speech

↓

10,000 words (25%)

Business
10,000 words

Education
10,000 words

Leisure
10,000 words

Public Services
10,000 words

It was the aim to include, as far as is possible for the language in question, at least one sample of each of the main activity types identified by Crowdy (1993) for the British National Corpus covering business, education, leisure and public services. A range of private and public companies and organisations were then contacted inviting them to participate in the project and supply appropriate speech data. Material has consequently been offered by medical and dental practices, cathedrals and churches, schools and universities, television networks and various individuals for the conversational speech. The speech data will be recorded onto mindiscs (although in some cases data has also been supplied on audio and video tape) and released as sound files. This will not only ease the transfer of the data onto computer, but minidiscs have been found to produce a high standard of audio recording.

The ethical issues of confidentiality and consent are key priorities in assembling data of this type. We have followed Sampson's (2000) advice for anonymisation practices. Information sheets have been issued to all who have been approached to participate, explicitly stating that anonymity is assured and all confidential information will be bleeped out, and names and personal details duly altered to protect identity. Informed consent must be given by all who agree to be recorded, giving permission for the material to be used and published as the project demands. It is our intention to release the stored digitised sound waves alongside the final corpora. The additional option is given to all participants that they may ask that their recording not be released as a sound file should they prefer this. In accordance with copyright regulations (Ward 1995 and 2000), anyone who participates and supplies data has copyright of their own words and must also give signed consent for the recordings to be made publicly available. For the medical data, consent had to be granted by the dental defence union or society and the Local Research Ethics Committee for the relevant NHS Trust, and parental consent has to be granted by all those who participate under the age of 16.

6. Transcription and annotation

The data will be transcribed orthographically (though not phonetically) and will be encoded in a TEI-conformant format. The header and transcription guidelines for the LER-BIML corpora are based on the standards for spoken language encoding developed in the EMILLE project (see http://www.emille.lancs.ac.uk/spoken.htm), which, in turn, incorporate the recommendations of several user surveys and standards initiatives, including the EAGLES recommendations on dialogue (Leech, Weisser, Wilson & Grice 2000).

The two corpora will then be annotated with parts of speech using tagsets which conform to the EAGLES recommendations for morphosyntactic annotation of corpora (Leech & Wilson 1999). Two tagsets exist for each language: a base tagset implementing the EAGLES obligatory features (i.e. major parts of speech only) and an extended tagset which also implements EAGLES recommended and language-specific features. (Since the EAGLES project did not address the Celtic languages, necessary language-specific features have been identified in consultation with experts on the various languages.) The Welsh extended tagset was devised within LER-BIML (Wilson 2002), whilst, for Gaelic, we will be using the PAROLE tagset developed for the National Corpus of Irish at the Linguistics Institute of Ireland.[15] In view of the close relationship between Irish and Scottish Gaelic, it has been determined that this tagset is also appropriate for the annotation of the latter. Although we will not be collecting

[15] http://www.ite.ie/pos.htm

and annotating corpora of Cornish or Manx, extended tagsets have also been devised for these languages (Mills 2002, Phillips 2002).

For the annotation of the Welsh corpus, an automatic tagger (Hepple 2000) is currently being re-trained using the annotated CEG corpus of written Welsh (Ellis et al 2001). It is envisaged that this tagger will apply only the base tagset for Welsh. This will allow more scope for the Gaelic corpus to be tagged by hand using the extended tagset. However, we also envisage re-training the tagger for Gaelic in the near future.

7. Conclusion

As has already happened to a great extent in Wales, the BIMLs look set to become increasingly important as the devolved governments, and consequent ethnic identities become more firmly established. Therefore although some BIMLs are presently spoken only by small numbers, it is nevertheless important strategically to begin to provide language engineering solutions for them. This fact has already been recognized at a policy level by EPSRC, who state that “the multilinguality of language engineering applications is of growing importance if these applications are to be equal use to all users of the many different languages and accents in the UK” and, furthermore, that “in EPSRC terms, the languages of particular importance are the indigenous minority languages of the UK (Welsh and Scots Gaelic), and the non-indigenous minority languages (such as Urdu, Hindi, Chinese, etc)...” (EPSRC 1999). Our project is the first step towards providing data for these applications in the Celtic languages.

8. References

Crowdy S 1993 Spoken Corpus Design. *Literary and Linguistic Computing 8(4), 259-65.*

Education Department, Scottish Executive 2000 Gaelic: Revitalising Gaelic a National Asset. Report by the taskforce on public funding of Gaelic. Retrieved February 2003 from URL http://www.scotland.gov.uk/library3/heritage/gtfr-00.asp

Ellis NC, O'Dochartaigh C, Hicks W, Morgan M, Laporte, N 2001 Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh. Retrieved February 2003 from URL http://www.bangor.ac.uk/ar/cb/ceg/ceg_eng.html

EPSRC 1999 People and Interactivity: Incorporating the Human Factors Review 1998/1999: The Results of the Consultation and Plans for the Future.

Hepple M 2000 Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (ACL-2000), Hong Kong, October 2000.

Jones R, Mason J, Jones R, Helliker L, Pawlewski M 1998 SpeechDat Cymru: A large-scale Welsh telephony database, *Proceedings of the LREC Workshop: Language Resources for European Minority Languages*.

Leech G, Weisser M, Wilson A, Grice M 2000 Representation and annotation of dialogue. In Gibbon D, Mertins I, Moore R (Ed.), *Handbook of multimodal and spoken dialogue systems*, pp.1-101. Kluwer, Dordrecht.

Leech G, Wilson A 1999 Standards for tagsets. In van Halteren H (Ed.), *Syntactic wordclass tagging*, pp.55-80. Kluwer, Dordrecht.

McEnery A, Sebba M, Burnard L 1999 Minority Language Engineering (MILLE) – Final report. Report to EPSRC, Lancaster University.

Mills J 2002 Suggestions for a morphosyntactic tagset for Cornish, based on the EAGLES obligatory and recommended attributes. Retrieved February 2003 from URL http://www.ling.lancs.ac.uk/biml/cornish_tags.htm

Oxford University Computing Services 2001 Second Edition British National Corpus World Edition. Retrieved February 2003 from URL http://www.hcu.ox.ac.uk/BNC/

Phillips JD 2002 A tagset for Manx. Retrieved February 2003 from URL http://www.ling.lancs.ac.uk/biml/manx.txt

Sampson G 2000 CHRISTINE Corpus, Stage 1 Documentation Release 2. Retrieved February 2003 from URL http://www.cogs.susx.ac.uk/users/geoffs/ChrisDoc.html

Ward A 2000 Copyright and Oral History, [Internet], Oral History Society. Retrieved February 2003 from URL http://www.nmgw.ac.uk/~ohs/ohs/copyright.html

Ward A 1995 Copyright, Ethics and Oral History, Colchester: The Oral History Society.

Williams B 1998 Levels of annotation for a Welsh speech database for phonetic research. Paper from conference proceedings: *Workshop on Language Resources for European Minority Languages (B. Williams, C. Nadeu, A. Monaghan, eds.) [27th May 1998]* Granada, Spain.

Williams B 1999 A Welsh speech database: preliminary results. In *Eurospeech 99 (European Conference on Speech Communication and Technology) [5-10 September 1999]*, Budapest, Hungary.

Wilson A 2002a Suggestions for a morphosyntactic tagset for Welsh, based on the EAGLES obligatory and recommended attributes. Retrieved February 2003 from URL http://www.ling.lancs.ac.uk/biml/welsh_tags.html

Wilson A, Worth C 2002b LER-BIML Working Paper 1: Surveying existing resources for the indigenous minority languages of the British Isles and Ireland, Lancaster University. Retrieved February 2003 from URL http://www.ling.lancs.ac.uk/biml/bimls3reports1.htm

Wilson A, Worth C 2002 LER-BIML Working Paper2: Surveying end user needs for the indigenous minority languages of the British Isles and Ireland, Lancaster University. Retrieved February 2003 from URL http://www.ling.lancs.ac.uk/biml/bimls3reports2.htm