Celtic Corpora

This is a guide to CHILDES data on the acquisition of Celtic languages. For a general introduction to the CHILDES database, please consult intro.pdf. The links in the table below are clickable, as are the thumbnails to the left.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Age Range</th>
<th>N</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irish - Guilfoyle</td>
<td>6</td>
<td>2–3</td>
<td>Single sessions from Irish-speaking children</td>
</tr>
<tr>
<td>on page 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Welsh 1 - Jones</td>
<td>18–30</td>
<td>7</td>
<td>Longitudinal study of natural speech from children in the early stages of acquisition of Welsh as a first language</td>
</tr>
<tr>
<td>on page 3</td>
<td>1;6–2;6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>on page 14</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. Irish – Guilfoyle

Eithne Guilfoyle
Dun Laoghaire Institute
Kill Avenue, Dun Laughaire
Co. Dublin, Ireland
guilfoyle@hotmail.com

This corpus is made up of 5 recordings made in Ireland in 1992, each approximately 45 minutes long. The children are native Irish speakers, aged between 1;7 and 2;9 years including one pair of twins. The children were recorded on audio and videotape in the presence of a parent and a researcher. The conversation is entirely in Irish, however all of the data is translated into English, and the child’s data is glossed on a %mor tier. The children played with a number of toys and the conversation primarily focuses on the toys.

The data were gathered by E. Guilfoyle, University of Calgary, under the auspices of a grant from the Social Sciences and Humanities Research Council of Canada (grant 410-91-1956). The transcripts were typed up by Nóra Welby and verified and translated by Fiona Coll. The transcriptions were glossed and checked by Eithne Guilfoyle and Síle Harrington, assisted by Leah Bartolin, Hooi Ling Soh, Erica Thrift, Grace Randa, and Sean Mac Lennan.

The data were coded with a view to studying the children’s syntactic and morphological development, however, some of the child and adult data may be of interest to people interested in borrowing, language change as all the adults are bilingual, and the children have varying exposure to English.

Table 1: Guilfoyle Children

<table>
<thead>
<tr>
<th>File</th>
<th>Child</th>
<th>Sex</th>
<th>Birthdate</th>
<th>Recording</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CAI</td>
<td>F</td>
<td>01-DEC-1990</td>
<td>5-JUL-1992</td>
<td>1;7.4</td>
</tr>
<tr>
<td>2</td>
<td>SEA</td>
<td>M</td>
<td>01-FEB-1990</td>
<td>16-JUL-1992</td>
<td>2;5.15</td>
</tr>
<tr>
<td>3</td>
<td>MAI</td>
<td>M</td>
<td>14-FEB-1990</td>
<td>17-JUN-1992</td>
<td>2;5.3</td>
</tr>
<tr>
<td>4</td>
<td>CIA</td>
<td>M</td>
<td>06-OCT-1989</td>
<td>16-JUL-1992</td>
<td>2;9.10</td>
</tr>
<tr>
<td>4</td>
<td>LAO</td>
<td>F</td>
<td>06-OCT-1989</td>
<td>16-JUL-1992</td>
<td>2;9.10</td>
</tr>
<tr>
<td>5</td>
<td>RON</td>
<td>M</td>
<td>05-JUL-1992</td>
<td>05-AUG-1992</td>
<td>2;1.0</td>
</tr>
</tbody>
</table>
2. Irish – Gaeltacht

Dr Thea Cameron-Faulkner, (University of Manchester)
Dr Tina Hickey, (University College, Dublin)
Professor Paul Fletcher, (University College, Cork)

Thea Cameron-Faulkner
School of Psychological Sciences
The University of Manchester
Oxford Road, Manchester, M13 9PL
t.cameron@manchester.ac.uk

The Gaeltacht Corpus consists of data of one native Irish-speaking family from the west coast of Ireland. The target child (Eoin) was being raised as a first language speaker of Irish. At the time of the study, Eoin was the younger of two children and aged 1;5 at the onset of recording. Eoin's mother was a native speaker of Irish and his father had native-like proficiency. Irish was the language of the home and consequently Eoin's first language. The members of his extended family living in the neighborhood were also native speakers of Irish. The project followed Eoin over the period of one year and aimed to collect one hour of recording every week.

Data collection and transcription
The mother was equipped with a DV-recorder incorporating a wide-angled lens and asked to conduct one hour of recording per week. The hour of recording could consist of one session or multiple sessions dependent on the situation (i.e. time limitations, the activity, or mood of the child). This approach to data collection was adopted in order to elicit the most representative and natural sample of interaction as opposed to set hour-long sessions in the presence of a research assistant. The speech of Eoin and his mother was coded for pragmatic function using the Inventory of Communicative Acts-Abridged (INCA-A) (Ninio et al. 1994).

Acknowledgement
We would like to take this opportunity to thank the family of the project for their time and patience. Also our thanks go to a team of dedicated transcribers; Catherine Muldoon, Caitriona Ryan, Louise Keegan, Emma Gleason, Sláine Cahillane, Diarmuid O Gruagáin, Diarmuid Clifford, Maire Treasa Ni Cheallaigh, Darina Ni Shithigh, and Ciara O’Toole. The collection and transcription of the corpus was funded by ESRC grant number RES 000-22-1125.
3. Welsh – CIG1

Bob Jones
Department of Education
University of Wales
Aberystwyth
Wales SY23 2AX
bmj@aber.ac.uk
http://users.aber.ac.uk/bmj/

This project, (C.I.G for Caffael yr Iaith Cymraeg “Acquisition of the Welsh Language”) was funded by a grant from the ESRC. It was based in the Linguistics Department of the University College of Wales, Bangor and Aberystwyth. The Principal Researchers were Dr. Robert Borsley, Dr. Michelle Aldridge, Prof. Ian Roberts (in Bangor), and R. Morris-Jones (in Aberystwyth).

The full-time research assistant in Bangor was Susan Clack and, in Aberystwyth, Gwennan Creunant was employed on a part-time basis. The project was initially to run for 12 months (from January 1996) but a subsequent extension on the basis of unspent monies gave another three months employment for both research assistants. This documentation file was written by Susan Clack on July 17, 1997.

The aims of the project as outlined in the ESRC grant application were:
1. To gather a substantial corpus of natural speech from children in the early stages of acquisition of Welsh as a first language. There were weekly tapes of 30 to 45 minutes over 9 months.
2. To investigate the development of syntax, especially clause structure, and to map out the typical pattern of development.
3. To use recent work on the development of syntax to illuminate the early stages of the development of Welsh syntax.
4. To use the Welsh data to evaluate ideas about parameter setting and functional categories that have emerged in recent work on the development of syntax.
5. To make the corpus available to other researchers through the CHILDES database.

Transcriptions were taken from 45 minute (usually) audiotapes only. Roughly, 9 months of six children (plus 4 months of another) are recorded from approximately 18–21 months to 28–30 months. The purpose was to tape naturalistic, spontaneous utterances and not to do specific elicitation.

Factors of sex of child and position in the family were not considered in choosing participants. The area from which the children were drawn is predominantly Welsh speaking (60–70% and higher in some of the villages). All the parents, (apart from one who learned Welsh from 3;0) were first language speakers of Welsh.
There are two components to this corpus. One is the Bangor dataset and the other is the Aberystwyth dataset. In Bangor, three participants (one female first child, one male second child, one female second child) were forthcoming. The first dropped out after 4 months. Tapes of a further child (first-born male) previously recorded in 1994 and roughly transcribed for a pilot study (Borsley and Aldridge) were totally redone in CHAT format. The Bangor files were prepared with the CHILDES editor.

Transcripts of a further three children (all first-born female) were prepared simultaneously in Aberystwyth by Gwennan Creunant. These files have been converted to CHAT format by the Principal Researcher in Aberystwyth who had previously worked with his own computer programs (for morphological tagging and glossing) for the analysis of the speech of older children. Tapes were made fortnightly in the initial stages (mainly one wordish) and weekly (holidays and illness permitting) in the later stages.

In Bangor there are 27 tapes in the ALAW Corpus, 26 in the RHYS Corpus. Taping of ELIN ceased after 11 sessions. It was decided to redo 29 of the 42 transcriptions of DEWI into CHAT. This means there are 93 transcriptions of Bangor area children. Transcriptions on the whole follow standard orthography (see below for exceptions) with occasional phonographic representations. Standard Welsh orthography is basically phonetic which makes a phonographic representation feasible. All speakers spoke basically Northern dialects.

The Aberystwyth corpus comprises 75 files. Transcriptions are generally phonographic. In Aberystwyth, on the mid-Wales coast, Southern and Northern dialects are heard and this is reflected in these phonographic transcriptions.
3.1 The Bangor Dataset

All the Bangor children live in the Arfon area Gwynedd, North Wales. In this area Welsh is spoken by approximately 70% of the population but this figure is higher in some of the villages where the children live. The education policy of the area is largely monolingual Welsh until at least 7;0. Otherwise, bilingual policies of administration are the norm in the public sector.

For the Bangor tapes the same toys were used in all sessions. These included a suitcase of animals, cars, a Fisher-Price™ parade of shops, Barbie dolls and Action Men with clothes, plus odds and sods. In some transcripts, child’s mother, grandmother, grandfather, father and/or siblings are present. The investigator was present on all occasions (but one of Elin and a few of Dewi which were mainly tapes made by parents for various reasons). The participants vary from corpus to corpus and day to day. In the majority the investigator and the child are alone for most of the time. In the case of Dewi, the same toys were also used but the taping was usually done in the investigator’s home, Dewi living close by in the same village.

All Bangor transcripts represent at least 30 minutes of tape time (45 minutes of most Kevin tapes). Some are longer for a variety of reasons, such as quality, ease of transcription, unusual quietness of child, sibling dominance in parts of tape, and so forth. All but a handful of tapes were made between 9:30 and 10:30 A.M. Times of the few that were not are noted in the initial headers.

All but a handful of tapes were transcribed on same day as taping. This means that context was fresh. Generally, an attempt has been made to add background and contextual information especially where utterances may be ambiguous. Efforts have also been made at making specific remarks about potentially ambiguous or odd utterances on %com lines, for example, the shaking of a head (when recalled) for negation where the form is declarative.

About 70% of the tapes have been listened to by an independent checker with the transcript available. Comments were made on transcripts. The tapes were then totally listened to again by the transcriber with checked transcripts. Amendments and corrections were then made, elevating best guesses to full status, interpreting xxx’s, adding and reiterating %COM especially in relation to intonational status of an utterance.

All transcripts (except Dewi which were second transcriptions) not independently checked have been checked with tape for a second (or third in some cases) by the transcriber. Here we have tried to be true to what is heard rather than what we know to be prescriptively correct. This comment is particularly relevant with regard to mutations (on the nonchild lines).

In the documentation of each corpus, which follows at the end of this file, there are comments for all children except Dewi. These comments were usually made straight after
transcribing and are of a general background nature with some impressions as to development.

The Bangor research assistant would like to thank the following: First and foremost the parents of the children for their unstinting cooperation. Also, Bill Hicks of Cysyll for installing the Welsh Spellchecker, Professor Cathail O’Dochartaigh (formerly of Cysyll, Bangor University and now of Glasgow University), Dr. Margaret Deuchar for advice in the early stages, Dr. Michelle Aldridge for her patience and good sense, Gwennan in Aberystwyth for sharing the lows of transcription work, Vivienne Pritchard for cheerfully checking, and many others including the members of the Manchester/Bangor reading group into Child Language, in particular Ginnie Gathercole, Marilyn Vihman, and Elena Lieven.
3.2 CHAT Usage

Generally, CHAT conventions have been followed (hopefully) but there are some divergences. These are noted here together with some general points.

1. The text replacement symbol [: text] has been used for a phonographic representation. These should not be considered as text replacements.

2. Words are not morphemicized. However, elided forms have been broken up. For example, “dwim isio” is represented by: dw i øm isio > copula1PS+pronoun+neg isio=want.

3. The trigrams ^w^ and ^y^ have been used for the circumflexed w and y of Welsh, e.g., dw^r = water and ty^= house. All other accents were available in ASCII.

4. The notation [?] has been used for best guess with 80–90% certainty. For less certainty alternative transcriptions [=? text] have been used or otherwise xx or xxx. A single x has also been used sometimes to denote something less than a ? word.

5. Time locations are not consistently marked on all transcripts. One dictaphone did not function correctly for a while and then was replaced by another with different readings. In general, the pattern of notation of @Time Locations reflect a pattern of either pause in the discourse or difficult/interesting utterances (for easier retrieval on tape) and not any regular pattern (although many have been added on checking for ease of retrieval).

6. %com lines have sometimes been used for a target language gloss in the case where there is a degree of uncertainty as well as other comments relating to interpretation. If the %gls is used, the target language gloss is quite certain. These are more common at a stage where there are more two word utterances.

7. Utterances/sentences have been delimited by # on the nonchild lines. ## on these lines would indicate a pause. # on the child lines is used for pauses.

8. Repeats on words by (/) were used in the early stages. In the later stages, as retraced utterances became more common, these were eliminated and all repeats are marked as retracings as they seemed more appropriate.

9. English words are followed by @s (s for Saesneg = English) or are represented by xs, the latter so that they can be included in MLU scores on the child lines. Chunks of English (which can then be excluded from MLU’s with -s) are represented by xxs. There will be [= English text] following all xs and xxs, or, if these are not clear [=? English text]. Many English words are spelled with Welsh phonography where it seems appropriate. There are other words where the orthography would be identical. The English words marked with @’s are generally those where the Welsh orthography does not seem appropriate.

10. The only postcodes that have been used are: [+ imit] , [+ part imit] and [+ prompt].
### 3.3 Disambiguation Devices

All Welsh/Welsh and Welsh/English homonyms have been disambiguated in the Bangor corpus (for the purpose of aiding glossing in Aber and making word lists). A variety of ways have been adopted to do this and details can be found in the following introduction to the lexicon and in the lexicon itself. The only disambiguation that has not been done is that of the predicate marker “yn” on the non-CHAT lines. The following disambiguation codes for common words are indicated below:

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>yn</td>
<td>verb/noun follows</td>
</tr>
<tr>
<td>yn1</td>
<td>prepositional use</td>
</tr>
<tr>
<td>yn2</td>
<td>predicate adjective/noun follows</td>
</tr>
<tr>
<td>yn4</td>
<td>ambiguous as to yn/yn2</td>
</tr>
<tr>
<td>i</td>
<td>first person singular pronoun following finite preposition or verb</td>
</tr>
<tr>
<td>i’</td>
<td>preposition = to/for</td>
</tr>
<tr>
<td>o</td>
<td>preposition = of/from.</td>
</tr>
<tr>
<td>øo</td>
<td>third person singular masculine pronoun</td>
</tr>
<tr>
<td>øna2</td>
<td>elided “dyna” as in “dyna fo”: there it/he is</td>
</tr>
<tr>
<td>na1</td>
<td>comparative word equivalent to “than”</td>
</tr>
<tr>
<td>na3</td>
<td>equivalent to neu = or</td>
</tr>
<tr>
<td>na7</td>
<td>“reduced” “mai” for focussed subordination</td>
</tr>
<tr>
<td>øta1</td>
<td>“reduced” “ynte”= or</td>
</tr>
<tr>
<td>do</td>
<td>“roof”</td>
</tr>
<tr>
<td>do</td>
<td>yes (past) word</td>
</tr>
<tr>
<td>(d)dyn</td>
<td>man</td>
</tr>
<tr>
<td>(d)tyn</td>
<td>tight</td>
</tr>
<tr>
<td>dy</td>
<td>second person singular pronoun (possessive type)</td>
</tr>
<tr>
<td>ødy</td>
<td>elided copula “ydy”</td>
</tr>
<tr>
<td>di</td>
<td>second person singular pronoun</td>
</tr>
<tr>
<td>ødi</td>
<td>elided aspect marker/preposition “wedi”= after</td>
</tr>
<tr>
<td>øna</td>
<td>elided locative and existential marker “yna”= there</td>
</tr>
<tr>
<td>na</td>
<td>negative (generalised)</td>
</tr>
</tbody>
</table>
3.4 Representation of English

The representation of English words in this corpus has posed some problems. Many English words have orthographic Welsh forms that might well be considered part of the Welsh language, for example, doli>doll. Welsh orthography is used where it is seems appropriate. In this sense the transcripts are not strictly phonographic. For example, ice cream is represented by ice+cream@s and not eis+crim (see Aber corpus) whether or not it is pronounced in the English way or the Welsh way.

The specific conventions adapted to address these concerns were as follows:
1. The English derivative appears after “>” e.g., doli>doll, ffrwnt>front.
2. Some forms have identical orthography to the English in which case the English derivative appears after “<=”, as in, mat<=mat, top<=top. There is no marking in the text to distinguish such a word as an English word. They may be words which are very common. This is a shortcoming but the emphasis is on trying to retain the phonographic integrity of Welsh.
3. English words with English orthography are marked with @s. In some cases Welsh orthography is inappropriate. For example, we use wee+wee@s and not wi+wi but pi+pi and not pee+pee@s. In some cases it is personal preference of the researcher. e.g., flat@s and not flat. In other cases distinguishing in this way aids in the elimination of homophonous forms e.g., go@s vs go=adjectival intensifier.
4. Some English will be marked as xs or xxs. Use of these varies from corpus to corpus but in general these strings are used for chunks of language rather than individual words. In these cases the English translation appears in a bracketed string as [= text] or [=? text] in the files.
5. Welsh words with English plurals where not the norm: nple eg blodaus= flowers (where English and Welsh plural is compounded).
6. English words with Welsh plurals where not the norm (whatever that is): nplw.

This notation has been used partly to eliminate English for the purposes of analysis, to identify chunks of code-switching and to eliminate homophonous (with Welsh) forms. It will be the case that some words marked with @s will also appear in xs or xxs strings. The main motivation for the policy adopted here is to maintain a constant representation of Welsh phonography/orthography/phonology in contrast to that of English. In mind is the fact that not everyone who may look at this corpus will be as well versed in English as Welsh speakers tend to be in Wales. Often, decisions as to how mark words have felt arbitrary but the preceding guidelines have been followed as far as possible.
3.5 Notations and Orthography

The symbols @s, xs, xxs, @o (onomatopoeic words) and markings for homonyms have been added after initial transcription in the later stages. As far as is possible the context in each case has been checked for accuracy. Sometimes something like woofwoof may appear marked with either @o or @c. This is relevant for categorization purposes. The same applies to the words “bang” and “bwm.”

Conventional spellings are used in the most part. There are a very few exceptions: isda>eistedd=sit; isio>eisiau=want; plus verbal and prepositional forms noted below.

In the lexicon, the usual dialect form or spoken form occurs after “@”. Most of these cases are subject to regular rules: 1) words with “e” in final syllable going to “a”; 2) dropping of silent “f” in words like nesaf, af, and so forth.

In the case of inflected prepositions (especially inflected forms of “gan”), there are different orthographic representations. The same applies to a few verbs such as rhoi/rhoid/ rhaed = give.

In the cases where conventional spellings are not used the conventional form (or root form) occurs after a dash. Alternatives, either English or Welsh, occur after slashes. English translations occur after equals signs. Alternative Welsh forms for English words appear after >. If they do not occur in the corpus they are marked with *. If they occur the % follows. Welsh words that appear with the English “s” plural appear in a category: nple. This notation is usually used where a Welsh plural could (and may otherwise) be expected. There will be a handful of plurals where the English plural morpheme is well established. The categories used are: n= noun, vn=verb noun, a=adjective, av=adverb, wh=wh word, fb=finit be, p=preposition, ip=inflected preposition, fv=finit verb, loc=locative adverb, g=greeting (or like). Soft mutations are indicated by ^, nasal by ^^, and aspirate by ^^^.
3.6 Layout

Number of repeats on words appear first. In later files these have almost totally been replaced with retracing symbols. This seemed more appropriate after the early one and two word stages. Proper nouns appear with capital letters. English forms have not been marked on these yet.
3.7 Pronunciation Notation

In the CHAT files, only pronunciation forms follow target form in brackets [:]. This is not the usual (text replacement) use of these brackets with CHILDES. These phonographic representations, made possible because Welsh orthography is phonetic to a high degree, are not consistently done but have been added to add a flavor of the child’s phonological competence.

Publications using these data should cite:

4. Welsh – CIG2

Bob Jones
Department of Education
University of Wales
Aberystwyth
Wales SY23 2AX
bmj@aber.ac.uk
http://users.aber.ac.uk/bmj/

This database in Childes format was produced by a project which was funded by the Economic and Social Research Council (ESRC) of the UK with an award of £60,611 (R000237978). The project ran from the 1st of July 1999 until the 30th of June 2000. It was directed by Bob Morris Jones and staffed by two researchers, Merris Griffiths and Mared Roberts, in the Department of Education, University of Wales, Aberystwyth, Ceredigion SY23 2AX, Wales, UK.

The data is based on the spontaneous recordings of children between the ages of three and seven years of age, speaking Welsh. They were recorded in schools throughout Wales in undirected play situations, mainly playing in pairs with various toys in a box of sand. The children are from different school, socio-economic, regional, and linguistic backgrounds.

The original recordings were collected during the period 1974-1977 by a project which was located in the same department, funded by the Welsh Office, directed by Professor C.J. Dodson, run by Bob Morris Jones, and staffed at various times by Brec'hed Piette, Hefin Jones, John Jones, Wyn James, Christine James, and Nesta Dodson.

There are two cohorts: children from three to five, and children from five to seven. The first digit in the names of the files which make up the database gives the age of the children. The file names of the five year olds of the older cohort are distinguished by the letter 'a' after the first digit. The remaining digits complete the file name in all cases.

The scale of the database can be indicated by the following summary:
three year olds: 25 files (c3001 - c3025), 418kb, 42 children
four year olds: 31 files (c4001 - c4031), 498kb, 62 children
five year olds: 39 files (c5001 - c5039), 859kb, 77 children
five 'a' year olds: 44 files (c5a001 - c5a044), 855kb, 87 children
six year olds: 48 files (c6001 - c6048), 1.00mb, 96 children
seven year olds: 52 files (c7001 - c7052), 1.14mb, 104 children

Personal names, local place-names, and local places-of-work have been made anonymous by using random nonsense-strings of letters: all begin with an initial capital, and the place names have a final 0. The names of public figures, fictional characters, and more
distant places have been retained. Making names anonymous loses some information about word-forms, especially about mutations - where they occur - and word-play.

The children produced many noises while playing, and some attempt has been made to transcribe these, although they are not intended to capture the phonetic details. They have the suffix @i. Nonsense forms, in word-play for instance, have the suffix @wp.

English is also spoken by various children to different degrees in the database. Single English words - either by themselves or within a Welsh utterance - are not marked. But phrases or sentences of English words are enclosed in scope symbols < ... >, and are followed by the comment [% Saesneg] - 'Saesneg' being the Welsh word for 'English'.

Similarly, phrases and sentences which are from songs, nursery rhymes, and similar material are enclosed within < ... > and are followed by the comment [% ca:n] - 'ca:n' (or 'c,n', to use the circumflex - see below) is the Welsh for 'song'.

Unfinished words (that is, fragments and not shortened words) are indicated by an initial &.

There are many homonyms, many of which come about through phonological processes of elision and assimilation in spontaneous speech. Digits and the apostrophe are used to distinguish different word-forms which otherwise have the same spelling. The lexicon gives the lexeme to which they belong. The apostrophe is declared in the 00depadd.cut file to cater for word-initial occurrences.

In spontaneous speech, patterns of a Welsh copula followed by a personal subject pronoun occur as a pronoun only. Such pronouns are indicated by a final apostrophe. There are instances, mainly of directive-like utterances within the context of a game, where it is not entirely clear what the pattern is. But these instances have likewise been give a final apostrophe.

The data files contain utterances by children and adults. The former are identified as Target_child or Child on the @Participant header line in the data files; the latter are identified as Investigators and Teachers. The utterances of the adults have been transcribed in full, but not as painstakingly as those of the children; in particular, homonyms have not all been disambiguated through transcription.

The lexicon contains the word-forms produced by the children. It does not contain word-forms produced by adult participants. The lexicon contains all the Welsh words and single English-words which occur within a Welsh utterance or by themselves. It does not contain English words which are in English phrases or sentences. It does not contain proper names, the spellings of noises or nonsense words - they can be identified in the data by an initial capital, the suffix @sn, and the suffix @gl, respectively. Neither does it contain xxx (for indecipherable material), and unfinished fragments which begin with &.

The categories and their codes in the lexicon are as follows:
?? = multi-category form which is ambiguous in context
a1 = pro-form place adjuncts like FANNA 'there', FAMA 'here', FANCW 'yonder'
ab = conjuncts and disjuncts like HEFYD 'also', FELLY 'therefore'
ad = other adjuncts
ag = apsect markers YN 'progressive', WEDI 'perfective'
an = adjectives
ar = prepositions
as = adverbs ALLAN 'out', YMLAEN 'onwards'. I-FFWRDD 'away', I-LAWR 'down',
etc.
at = adverbs beginning with TU - TU-ALLAN 'outside', TU-OL 'behind', etc.
b4 = Welsh finite verb with English inflection
bd = English verbs in "-ed", "-en" or equivalent e.g. 'crashed', 'drunk'
be = verb-noun forms (compare English plain infinitive) including auxiliaries
but not BOD 'be'
bf = finite-verb forms (including the imperative forms) except BOD 'be'
bg = English verbs in "-ing"
bp = English plain infinitive forms
cd = co-ordinating conjunctions
ce = verb-noun (compare English plain infinitive) of BOD 'be'
cf = finite forms of BOD 'be'
cm = MWY 'more' as a comparative particle before adjectives
cn = greetings and farewells
cy = subordinating conjunctions like ACHOS 'because'
eb = standard exclamations like AA 'ah', OO 'oh'
en = nouns
er = the post-modifying words ARALL 'other' and ERAILL 'others'
es = EISIAU 'wants, needs' - a nominal form
g1 = nominal wh- words - BETH 'what', PWY 'who'
g2 = adverbial wh- words - PRYD 'when', PAM 'why', SUT 'how'
g3 = the wh- word PA 'which'
g4 = compounds involving wh- words like BETH+BYNNAG 'whatever',
PRYD+BYNNAG
'whenever'
g5 = the wh- word FAINT 'how much/many'
ga = grammatically invariant answer words IE 'yes', NAGE 'no', DO 'yes' a
NADDO 'no'.
gc = the comparative particle NA 'than'
gd = demonstrative words DYNA 'there/that is', DYMA 'here/this is', DACW
'yonder is'
gg = intensifiers like RHY 'too', GO 'gairly', MOR 'so'.
gm = quantifiers like DIGON 'enough', LLAWER 'much/many, MWY 'more'
gr = preverbal particles like MI, FE, NI and focussing particles like MAI, AI
gt = the predicatival particle YN
ll = pro-form adjuncts YNA 'there', YMA 'here' and ACW 'yonder'
ly = letters of the alphabet
mo = words indicating epistemic modality EFALLAI 'perhaps', HWYRACH 'perhaps'
ne = the negator DIM 'no/not' both as quantifier and adverb
on = onomatopoeic-type forms
pa = politeness expressions
pe = determiners
pi = forms of PIAU, used to indicate ownership
qq = for obscure forms
r1 = personal pronouns
r2 = demonstrative pronouns
r3 = indefinite pronouns like RHYWUN 'someone'
r4 = negative pronouns
r5 = reflexive pronouns
r6 = reciprocal pronouns
r7 = conjunctive pronouns like FINNAU 'me too'
r8 = prefixed (possessive) pronouns
r9 = the 'alternative' pronoun LLALL 'other', LLEILL 'others'
rd = RHAID 'must, necessity'
ri = numbers
rp = universal pronouns like PAWB 'everyone'
rq = indefinite phrases like BETH+'NA 'thingie', LLE+'NA, BE+TT+'N+GALW
's what do you call it'
sg = standard verbal pauses like YMM 'uhm'
sy = standard paralinguistic forms like HY-HY 'uh-uh', MM-MM 'uhm-uhm'
ya = manner-adverbial particle YN e.g. YN GYFLYM 'quickly'

Multi-membership, if found in the corpus, is indicated by the Childes convention for
this, that is, a backward slash after the first entry, followed on the succeeding line(s) by
another entry.

These categories serve only to identify data which can be recovered for analysis. They are not intended to represent probing analyses.

This latter point applies to all transcriptional conventions in this database - they serve
as ways of recovering data for analysis.

The files supplied for this database are as follows:

data files:    c3001 - c3025
              c4001 - c4031
              c5001 - c5039
              c5a001 - ca5044
              c6001 - c6048
              c7001 - c7052
lexicon files: welsh3_7.lex (the main lexicon)
               gl.lex (nonsense words)
sn.lex (noises)
others:
   00depadd.cut
   00readme.cdc (this file)