Documentation File
for the
Bangor *Siarad* Corpus

For queries, contact Dr. Peredur Webb-Davies, School of Linguistics and English Language, Bangor University, Gwynedd, Wales, LL57 2DG.
Email: p.davies@bangor.ac.uk, or m.deuchar@gmail.com.

May 2014

Section 1

Introduction

1.1.1 The Siarad[1] corpus of Welsh-English bilingual speech was recorded and transcribed between 2005 and 2008 as part of a research project funded by the Arts and Humanities Research Council (AHRC), entitled 'Code-switching and convergence in Welsh: a universal versus a typological approach'. The main theoretical aim of the project was to test alternative models of code-switching with Welsh-English data.

1.1.2 Please refer to the corpus as the 'Bangor *Siarad*' corpus, and provide a link to the website by which you accessed the corpus, either *bangortalk.org.uk* or *talkbank.org*. Please also cite:

> Deuchar, M., P. Davies, J. Herring, M. Parafita Couto, and D. Carter (2014). Building bilingual corpora. In: E. M. Thomas and I. Mennen (Eds.), *Advances in the Study of Bilingualism*, pp. 93–111. Bristol: Multilingual Matters.

We request that a copy of any publications that make use of this corpus be sent to us at the email address *m.deuchar@gmail.com*. For introductory information about the Welsh-speaking community see Deuchar (2005).

1.1.3 The Siarad corpus is licensed under the GNU GPL[2] if retrieved from *bangortalk.org.uk*, and under Creative Commons BY-NC-SA[3] if retrieved from *talkbank.org*.

[1] **Siarad** (/ʃarad/) is the Welsh word for 'to speak' or 'speaking'.
[2] gnu.org/licenses/gpl.html
[3] creativecommons.org/licenses/by-nc-sa/3.0

Section 2

The data

2.1.1 The corpus consists of 69 audio recordings and their corresponding transcripts of informal conversation between two or more speakers, involving a total of 151 speakers from across Wales. Participants were recruited via a variety of methods, including advertising, approaching visitors at a Welsh-language cultural event, and using the research team's extended social network. In total, the corpus consists of 452,116 words of text from 40 hours of recorded conversation. The transcriptions (in CHAT format) are linked to the digitized recordings through sound links at the end of each main tier. Most recordings were in stereo, and made using radio microphones and a Marantz hard disk recorder. A minidisk recorder was also occasionally used, meaning that some recordings are in mono mode.

2.1.2 The recordings were made at a place convenient for the speakers, e.g. at their homes, workplaces or at the university. After setting up the equipment the researcher would leave the speakers to talk freely with one another. The first five minutes of all recordings after the point when the researcher left the room have been deleted. In some cases the researcher re-entered briefly during the recording. These sections have not been transcribed, but notes have been made in the relevant parts of the transcripts.

2.1.3 At the end of each recording all participants were asked to fill in questionnaires providing background information regarding their age, gender, location of places lived, etc, in order to provide information for sociolinguistic analysis. They were also asked to sign consent forms giving permission for their recording and its transcript to be used for research purposes and to be submitted to online linguistic archives. The consent form included the provision that the names of speakers and other people named in the recording would be replaced by pseudonyms in the transcript. In the case of children of 16 years or younger, a consent form was also been signed by a parent or guardian.

2.1.4 Sound and transcription files in the corpus are named after the researcher (primarily) responsible for recording them, namely Marika Fusser, Peredur Davies, Elen Robert, Jonathan Stammers, Nesta Roberts, Gary Smith and Margaret Deuchar. Each file name is made up of the surname followed by a number (ordered chrono-

logically). The sound and transcription files for each conversation share the filename, but have different file extensions (*.wav and *.cha respectively). For example, Davies3.cha is the transcription of Peredur Davies' third recording (sound file Davies3.wav). In a few cases numbers are discontinuous. The Fusser files begin with Fusser3, for example. Also, five recordings collected (including Fusser20, Fusser24 and Davies8) ultimately had to be left out of the corpus. In three cases this was due to the lack of consent from speakers in the recording, in one case due to the researcher taking an extensive part in the conversation, and in one case due to a participant being a Welsh speaker from Patagonia who was not a Welsh-English bilingual.

2.1.5 A list of the transcript files in the corpus can be found in the Appendix. This list includes information about the length of the recording, the number of main participants, their age and sex. Details regarding the context of each conversation and a list of all speakers involved are given in the transcript headers. Some additional information about the speakers and recordings is available to researchers on request.

2.1.6 All recordings have been transcribed in the CHAT transcription and coding format (MacWhinney, 2000), in accordance with the online CHAT manual[1] from the Talkbank website. All further references to CHAT in this document are taken from this online version.

2.1.7 All transcripts have been done by trained transcribers working on the project: Peredur Davies, Marika Fusser, Siân Wynn Lloyd, Elen Robert and Jonathan Stammers. For 22% of the transcripts an independent transcription was done, in which a member of the transcription team transcribed one (randomly selected) minute of the recording independently from the original transcriber of that particular transcript. Transcripts were then compared and a rate of similarity was calculated. The average reliability score[2] for independent transcriptions was 75%.

2.1.8 All transcripts contain at least three different tiers. In addition to the main tier, required by CHAT, we use two alternative gloss tiers for the closest English equivalent for each word (including morphological information where relevant). One tier contains manually produced glosses and is labelled *%gls* while the other contains automatically generated glosses and is labelled *%aut*. There is also a translation tier (*%eng*), which contains a free translation of the main tier. A comments tier (*%com*) has also been used occasionally for comments by the transcriber that are specific to the utterance in the corresponding main tier. All main tiers include

[1] childes.talkbank.org/manuals/CHAT.pdf

[2] An innovative method was used based on Turnitin plagiarism detection software (turnitin.com). For further details see Deuchar et al. (2014).

a sound link to the corresponding section of the recording.

2.1.9 The remainder of this document outlines the conventions used in the main tier and the gloss tier.

Section 3

Main tier

3.1 Layout of transcription

3.1.1 Since the theoretical aims of the project include clause-based analysis, the transcribed data are divided into clauses where possible. Where an utterance contains two main clauses, each clause in that utterance is written on a separate main tier. Complex clauses are treated as one clause and therefore subordinate clauses are included in the same tier as their main clauses. Adverbial clauses are also written on the same main tier as their related main clause.

3.1.2 Each main tier is divided into units which we call 'words' for the purposes of these conventions. With some exceptions (see 3.1.3) a word is considered to be a continuous sequence of characters containing no spaces, as found in Geiriadur Prifysgol Cymru (Thomas, 2004), Geiriadur yr Academi (Griffiths and Jones, 1995), Cysgeir (Canolfan Bedwyr, 2008) or the Oxford English Dictionary online (OED, 2008). These are referred to as GPC, GyrA, Cysgeir and OED respectively throughout this document. Where items are entered as two hyphenated words in these reference dictionaries, they are connected by underscore in the transcripts. When one of the reference dictionaries offers more than one alternative (e.g. *minibus*, *mini-bus* or *mini bus*), or when the reference dictionaries differ from each other, the most compact alternative is chosen (*minibus* in this case).

3.1.3 Other items which are treated as words are:

1. interjections and interactional markers, e.g. *ah*, *er*, *um* etc.
2. proper names (including names of books, films, organisations etc.), a sequence of words being connected by underscores, e.g. *Elton_John*, H*ong_Kong*, *Sweet_Valley_High*
3. abbreviations (connected by underscore), e.g. *N_S_P_C_C*
4. some prepositions and adverbs, usually represented as two words, whose individual parts are meaningless or difficult to translate in isolation, e.g. *oddi_wrth*. See Table 3.1 - most of these are normally translated into just a single English word.

Table 3.1 – Phrases treated as words

Our transcription	Standard orthography	English equivalent
ar_bwys	ar bwys	next to
ar_draws	ar draws	across
ar_fin	ar fin	on the verge of
ar_gael	ar gael	available
ar_goll	ar goll	lost
ar_gyfer	ar gyfer	for
ar_ôl	ar ôl	after
au_pair	au pair	au pair
dim_byd	dim byd	nothing
ei_gilydd	ei gilydd	each other (3rd person)
eich_gilydd	eich gilydd	each other (2nd person)
ein_gilydd	ein gilydd	each other (1st person)
er_mwyn	er mwyn	for
ers_talwm	ers talwm	in the past, long ago
gwalch_y_pysgod	gwalch y pysgod	osprey
gwir_yr	gwir yr	honestly
hyd_yn_hyn	hyd yn hyn	so far
i_fewn	i fewn	in(to)
i_ffwrdd	i ffwrdd	away
i_fyny	i fyny	up
i_gyd	i gyd	all
i_lawr	i lawr	down
i_mewn	i mewn	in(to)
lefel_A	lefel A	A-level
lefel_O	lefel O	O-level
naill_ai	naill ai	either
o_dan	o dan	under
o_danodd	o danodd	beneath
o_gloch	o'r gloch	o'clock
o_gwbl	o gwbl	at all
o_gwmpas	o gwmpas	around
o_wrth	oddi wrth	from
oddi_ar	oddi ar	off
oddi_wrth	oddi wrth	from
oni_bai	oni bai	unless
pob_dim	pob dim	everything
pryf_copyn	pryf copyn	spider

Continued on next page

Table 3.1 – *Continued from previous page*

Our transcription	Standard orthography	English equivalent
syth_bín	syth bín	straight away
ta_waeth	'ta waeth	anyway
un_ai	un ai	either
wrth_gwrs	wrth gwrs	of course
y_chdi	y chdi	you (emphatic)
y_fi	y fi	I/me (emphatic)
y_nhw	y nhw	they/them (emphatic)
y_ni	y ni	we/us (emphatic)
yn_erbyn	yn erbyn	against
yn_ôl	yn ôl	back
yn_ystod	yn ystod	during

3.1.4 Contractions that do not have entries in one of the Welsh-language reference dictionaries (namely GPC, GyrA or Cysgeir) or in King (2003), are transcribed in full, but the unpronounced parts are bracketed. For example, the pronunciation of *fel yna* 'like that' as [vɛla] in speech is represented in the transcripts as *fel (yn)a*.

3.1.5 There are some continuous sequences of characters in the main tier which are not treated as words. These include simple events such as *&=laughs* (see CHAT[1] 7.8.1), *xxx* for unintelligible material, or the use of an ampersand plus phonetic characters for intelligible sounds without clear meaning (see CHAT 6.4 for both).

3.2 Language marking

3.2.1 Each word in the main tier has its language source identified. The default language is identified as that providing the greatest number of words in the transcript and is Welsh in all the transcripts. Welsh words are unmarked but English words are identified with the tag *@s:eng*. Words which could come from both Welsh and English are considered to be of 'undetermined' language and are marked *@s:cym&eng*, where *cym* represents Welsh and *eng* English. An entire utterance in English, the non-default language, is marked with a precode *[- eng]* instead of marking each word as English.

3.2.2 A word or morpheme is considered to be Welsh if it can be found in any of the Welsh-language reference dictionaries or in King (2003) or Thomas (1996).

[1]References to section numbers in the online CHAT manual are to the version dated March 4, 2014.

3.2.3 Words which contain two or more morphemes from different languages are marked as mixed-language words, e.g. *concentrate_io@s:eng+cym* 'to concentrate'. However, where a word containing at least one English morpheme and at least one Welsh morpheme is included in one or more of the Welsh-language reference dictionaries, it is marked as a Welsh word. For example, the English word *use* forms the basis of the Welsh word *iwsio* 'to use' but the entire word is considered to be Welsh (and transcribed without a language marker as *iwsio*) because it is included in one of the Welsh-language reference dictionaries.

3.2.4 The language marker *@s:cym&eng* is used with words where the language source is undetermined. It marks words that occur in the lexicon of both languages (as determined by the Welsh-language reference dictionaries for Welsh or by the OED for English), that are pronounced in a way that is possible both in Welsh and in English, e.g. [əŋkl] (*uncle* in English or *yncl* in Welsh) or [mat] (*mat* in both languages). *@s: cym&eng* also marks a specified list of interjections and interactional markers, e.g. *ah*, *aha*, *hmm*, *oh*, *ooh*, *um*. Other interjections and interactional markers are assigned language markers according to their inclusion (or not) in the reference dictionaries. For example, *ych* (a marker of disgust equivalent to 'yuk' in English) is unmarked as it is considered to be Welsh and only found in the Welsh-language reference dictionaries.

3.2.5 Where a lexeme could belong to both languages, but its pronunciation in a specific occurrence belongs unambiguously to one language only, it will be assigned a language source and marked or not accordingly, depending on its pronunciation. For example, *toast@s:eng* is used where the word is pronounced with [əʊ] as in English only, but *toast@s:cym&eng* where the word is pronounced with [o] as in Welsh or some varieties of Welsh English.

3.2.6 Proper names and titles are marked *@s:cym&eng* (undetermined) unless there are alternatives in each language in general use, e.g. *Elton_John@s:cym&eng*, *One_Flew_Over_the_Cuckoo's_Nest@s:cym&eng*, *Hong_Kong@s:cym&eng*, *Tebot_Piws@s:cym\&eng* (a Welsh-language pop group, literally meaning 'purple teapot') but *Cardiff@s:cym&eng*, *Caerdydd* (the Welsh word for 'Cardiff').

3.2.7 According to GPC, the '-s' plural ending is an established loan in the Welsh lexicon. Any plural formed with the '-s' ending is assigned the language source of the previous morpheme. For example, *pregethwrs* is unmarked like the singular form *pregethwr* 'preacher', but *dolphins@s:cym&eng* is marked undetermined like *dolphin@s:cym&eng* and *dogs@s:eng* English like *dog@s:eng*.

3.2.8 In multi-word phrases, each word is tagged separately, regardless of the phrase's internal syntax. For example, in *traffic@s:cym&eng lights@eng*, *traffic* is

coded as undetermined, although the syntax of the whole phrase is English.

3.3 Orthography

3.3.1 Words marked as *@s:eng* (English) are transcribed in standard English orthography, including contractions, such as *isn't*. Some non-standard spellings for colloquial forms such as *gonna* are used.

3.3.2 Words whose language source is undetermined are transcribed in English rather than in Welsh orthography, e.g. *acid@s:cym&eng* rather than *asid@s:cym&eng*. This is in order to make the corpus more accessible to non-Welsh-speakers who might use the data.

3.3.3 When words marked as English or undetermined are mutated (where the sound of an initial consonant is changed depending on the grammatical context, see for example King 2003, pp14–20, the initial (mutated) sound is written in Welsh orthography and the rest in English, e.g. *ei firthday@s:eng* = 'his birthday'; *ei goat@s:eng* = his coat. In the case of words that begin with 'qu' in English orthography but that are mutated in the data, the mutated sound and the following [w] are written in Welsh orthography, e.g. *question* (unmutated), *gwestion* (soft mutation), *chwestion* (aspirate mutation), *nghwestion* (nasal mutation).

3.3.4 Words marked as Welsh are transcribed in Welsh orthography. We have not represented regional variation in the transcripts, except in cases which have orthographic representation in the Welsh-language reference dictionaries or in King (2003).

There are some cases where we differ from the standard orthography:

1. We transcribe some non-standard verb-noun suffixes, e.g. *-ian* in *swnian* 'to grumble' rather than *-io* in the standard form *swnio*. We also transcribe non-standard plural forms of nouns, e.g. *cobenni* 'night-dresses' rather than the standard form *cobannau*.

2. We represent non-standard usage of inflected prepositions. Agreement markers for person and number show considerable variation in the spoken language. Thus one may, for example, find several forms for 'to you' (plural/respect form), such as *wrthoch chi* (the variant found in King (2003), *wrthych (chi)* (more formal variant, e.g. prescribed in Thomas (1996) as well as *wrthach chi* (more colloquial, northern variant). The orthography used in the transcripts is based on pronunciation.

3. Northern second person singular verb and preposition endings not usually

represented in writing are transcribed with a final *-a* where they are followed by the pronoun *chdi*, e.g. *oedda chdi* 'you were', *arna chdi* 'on you'. Where they occur in isolation, they are transcribed as *-achd*, e.g. *oeddachd* 'you were/weren't you', *arnachd* 'on you'.

4. We do not represent morpheme-final [v] when it is not pronounced. For example, [pɛntre] 'village' is written *pentre* in the transcripts rather than *pentref* (as the word is represented in the Welsh-language reference dictionaries).

5. Morpheme-initial /r/ is only transcribed as 'rh' where it is clearly heard by the transcriber to be voiceless [r̥]. Otherwise it is transcribed as 'r', even when the standard orthography prescribes 'rh'. Some speakers do not have [r̥] as part of their phonological system in any case.

6. Morphemes in Welsh which are usually written with an initial apostrophe, such as the possessive pronoun *'w*, and the marking of the ellipsis of a possessive pronoun in e.g. *'nhad* 'my father', do not have this initial apostrophe marked in the transcripts owing to the conventions of CHAT.

7. We have represented mutation (sound change to initial consonants) or its absence without following prescriptive rules as to where mutation might or might not be expected. Thus the Welsh form of 'in Cardiff' may be transcribed *yn Caerdydd* (with an initial [k] on the placename) and *yn Gaerdydd* (with initial [g]), as well as the standard form *yng Nghaerdydd* (with initial [ŋ̊], according to what is heard. We have also transcribed the aspirate mutation of /m/ and /n/ after the 3rd singular feminine possessive adjective common in northern varieties, e.g. *ei mham* 'her mother', with initial [m̥) , rather than standard *ei mam*' with initial [m].

3.3.5 In Table 3.2 we list some colloquial forms which are not represented in the Welsh-language reference dictionaries but which we have transcribed as indicated:

Table 3.2 – Colloquial forms

Our transcription	**Standard equivalent**	**English equivalent**	Comments Comments
		(a) Colloquial words	
(r)hein, (r)hain, (r)heiny etc	rhein, rhain, rheiny etc.	these, those etc.	pronounced with initial [h]
cordwellt	cordwellt	cordgrass	technical term listed on termau.org but not in our reference dictionaries
cwfwr	cyfarfod	meet	common in north west Wales

Continued on next page

Table 3.2 – *Continued from previous page*

Our transcription	**Standard equivalent**	**English equivalent**	Comments Comments
cyfryngi	cyfryngi	someone working in the media	recent coinage not yet in dictionaries
dafedd	edafedd	yarn	mutates to *ddafedd*
diwc	duwcs	gosh	
dôl, yn_dôl	yn ôl	back	common in north Wales
fannodd	dannodd	toothache	northern variant
ffluch	—	fling	heard in the north west
fformwleiddio	geirio	formulate	coinage based on English equivalent
Fictorianaidd	Fictoraidd	Victorian	wide-spread variant
gewin, gwinedd	ewin, ewinedd	claw(s)/ finger nail(s)	colloquial form listed in GPC article for *ewin*
gosa	oni bai	unless	heard in north-western varieties
hompen	homer	huge thing	form used by speaker from the south-west
jaman	—	(embarrass)	*cael jaman* = Caernarfon expression for 'being proved wrong'[2]
molchi	ymolchi	wash oneself	mutates to *folchi*
nunman	unman	nowhere	widespread
olradd	ôl-raddedig	postgraduate	heard in Welsh universities
penwsnos	penwythnos	weekend	GPC has an entry for *wsnos*
perchynu	perchen	own	form used several times by 16 year-old native speaker
pwpŵo	—	poo (verb)	*pwpŵo* is listed in GPC meaning 'talking derogatively'
socsen	sociad	soaking	heard in north-western varieties
ticiâu	diciâu	tuberculosis	northern variant attested in "diciau" article in GPC
(b) Colloquial verb forms			
adnabodais i	adnabyddais i	I recognised	
aethai hi, aethen ni, aethan nhw	âi, aem, aent	she/we/they would go	
byswn i, bysa chdi etc.	baswn i, baset ti etc.	I would, you would etc.	very common in northern varieties
cad	cadw	keep	imperative
caethet ti, caethen ni	caet, caem		
cawd	cafwyd	was had	impersonal
chwerthais i, chwerthon ni	chwarddais i, chwarddon ni	I laughed, we laughed	
cyma	cymer	take	imperative
dois i, doith o, dothon ni etc.	des i, daeth o, daethon ni etc.	I came, he came, we came etc.	
dyla fi	dylwn i	I should	

Continued on next page

[2]See youtube.com/watch?v=Z-x7zLAZdLM

Table 3.2 – *Continued from previous page*

Our transcription	**Standard equivalent**	**English equivalent**	Comments Comments
dylen i, bydden i etc.	dylwn i, byddwn i etc. common in southern varieties	I should, I would etc.	
gada	gad	leave	imperative
mag	mae	(he/she/it) is	3rd singular present form of *bod* 'to be' heard in south-western varieties
na i etc.	a i etc.	I will go etc.	heard in the Caernarfon area
oedd nhw, wneith nhw etc.	oedden nhw, wnan nhw etc.	they were, they will etc.	3rd person singular verb forms used with plural pronouns
syma	symud	move	imperative
tes i (ddi)m	es i ddim	I didn't go	some northern varieties
troeodd hi	troes, trodd	she turned	
y fi	rwy i	I am	southern Welsh
	(c) Interactional markers		
argob	argoel	gosh	interactional marker based on *arglwydd* 'lord'
asu	—	gosh	interactional marker based on *Iesu* 'Jesus'
diwc	duwcs	gosh	variant listed in GPC under *duwcs*
duwarth	duwcs	gosh	interactional marker based on *duw* 'god'
duwedd	duwcs	gosh	interactional marker based on *duw* 'god'
ewadd	ew	wow	
iargoel	argoel	gosh	interactional marker based on *arglwydd* 'lord'
iesgob	esgob	gosh	interactional marker based on *Iesu* 'Jesus'
myn_diân_i	—	by heck	interactional marker based on *diawl* 'devil'
wannwyl	duw annwyl	good lord	a contraction of *duw annwyl*
bleugh	—	—	English interactional marker indicating disgust
hehey	—	—	English interactional marker indicating approval
woohoo	—	—	English interactional marker indicating joy
wow	—	—	English interactional marker indicating surprise

Continued on next page

Table 3.2 – *Continued from previous page*

Our transcription	**Standard equivalent**	**English equivalent**	Comments Comments
		(d) Playful and ad hoc forms	
cyn_rodidenaidd	—	‘pre-rhododendric’	ad hoc adjective to describe the period before the arrival of rhododendrons in Wales
geitha fi	ges i	I got	uttered by 10-year old after a number of retracings
ruddydendrons	rhododendrons	rhododendrons	apparently a citation of local playwright W.S. Jones
		(e) Miscellaneous	
cynna fi, cynna chdi etc.	gen i, gen ti, etc.	before me, before you etc.	preposition inflected in northern varieties
dwmbo, wmbo	dw i ddim yn gwybod	I don’t know	contraction
henach, henaf	hŷn, hynaf	older, oldest	

Section 4

Gloss tier

4.1 Principles

4.1.1 Each word (see 3.1.2 and 3.1.3) in the main tier is given a manual gloss in the gloss tier (*%gls*) and an automatic gloss in a second gloss tier (*%aut*) which has been automatically generated using the Bangor Autoglosser (bangortalk.org.uk/autoglosser.php): for further details see Donnelly and Deuchar (2011).

4.1.2 Non-words (see 3.1.5) are not glossed, with the exception of *xxx*, which are represented by the same characters in the manual gloss (*%gls*), while being omitted for readability in the autogloss (*%aut*).

4.1.3 In both gloss tiers, words are glossed with the closest English-language equivalent (in lower case), with the exception of proper names (see below). In Welsh or mixed-language words, morphological information is frequently included in the gloss in upper case: see 4.2.1.

4.1.4 Proper names (including names of books, films, organisations etc.) marked as English or undetermined are glossed as they appear in the main tier. For example, *Hong_Kong@s:cym&eng* is glossed as 'Hong_Kong', *Cardiff@s:eng* is glossed as 'Cardiff' and *Tebot_Piws@s:cym&eng* is glossed as 'Tebot_Piws'. However, proper names marked as Welsh are glossed with their English-language equivalents. For example, *Caerdydd* is glossed as 'Cardiff'.

4.1.5 Lexical information always precedes morphological information in the gloss. A full stop (.) is used to separate morphological information from lexical information (e.g. go.NONFIN) and also to separate two pieces of morphological information (e.g. PRON.3SM). Some manual glosses contain only morphological information, such as POSS.2S for the 2nd singular possessive adjective *dy*.

4.1.6 The underscore is used in the gloss tier to connect more than one lexical item in a gloss, where the English translation of a single Welsh word involves more than one word. For example, *neithiwr* is glossed as 'last_night'.

4.2 Key to morphological glosses

4.2.1 Table 4.1 provides a key to the morphological abbreviations included in the manual glosses, and Table 4.2 provides a similar key to the automatic glosses.

Table 4.1 – Manual gloss abbreviations

Abbreviation	**Representing**
1,2,3	1st, 2nd, 3rd person
CONDIT	conditional/habitual past
DET	determiner
F	feminine
FUT	future/habitual present (verb *bod* 'to be' only)
IM	interactional marker/exclamation, e.g. *um*, *oh*
IMP	imperfect (verb *bod* 'to be' only)
IMPER	imperative
IMPERSONAL	impersonal
INT	interrogative
M	masculine
NEG	negative
NONFIN	nonfinite
NONPAST	nonpast tense (used for present/habitual/future)
PL	plural
PAST	past tense
PERF	perfect
POSS	possessive
POSSD	possessed
PRES	present tense (verb *bod* 'to be' only)
PRON	pronoun
PRT	particle
REL	relative
S	singular
SUBJ	subjunctive

Table 4.2 – Automatic gloss abbreviations

Abbreviation	**Representing**
0	impersonal
1P	1st person plural
1S	1st person singular

Continued on next page

Table 4.2 – *Continued from previous page*

Abbreviation	**Representing**
123S	1st, 2nd, 3rd person singular
123P	1st, 2nd, 3rd person plural
123SP	1st, 2nd, 3rd person singular and plural
13P	1st, 3rd person plural
13S	1st, 3rd person singular
12S123P	1st, 2nd person singular and 1st, 2nd, 3rd person plural
12S13P	1st, 2nd person singular and 1st, 3rd person plural
23P	2nd, 3rd person plural
23S	2nd, 3rd person singular
23SP	2nd, 3rd person singular or plural
2P	2nd person plural
2S	2nd person singular
2SP	2nd person singular or plural
2S123P	2nd person singular and 1st, 2nd, 3rd person plural
3P	3rd person plural
3S	3rd person singular
3SP	3rd person singular or plural
A.POT	adjective of potential
ADJ	adjective
ADV	adverb
AFF	affirmative
AG	agent
AM	aspirate mutation
AV	adjective or verb
ASV	adjective, singular noun, or verb
AUG	augmentative
BE	auxiliary verb ‘be’
COMP	comparative
COND	conditional
CONJ	conjunction
DEF	definite
DEM	demonstrative
DET	determiner
DIM	diminutive
E	exclamation
EMPH	emphatic
F	feminine
FAR	far (demonstrative)

Continued on next page

Table 4.2 – *Continued from previous page*

Abbreviation	**Representing**
FOCUS	item with focus
FUT	future
GB	's – genitive or auxiliary 'be' elision
GER	gerund
H	pre-vocalic h after 3S.F, 1P and 3P possessives
HAVE	auxiliary verb 'have'
HYP	hypothetical
IM	interactional marker
IMPER	imperative
IMPERF	imperfect
INDEF	indefinite
INFIN	infinitive
INT	interrogative
M	masculine
MF	masculine or feminine
N	noun
NEAR	near (demonstrative)
NEG	negative
NM	nasal mutation
NT	neuter
NUM	numeral
OBJ	object
ORD	ordinal
PAST	past
PASTPART	past participle
PERF	perfect
PL	plural
PLUPERF	pluperfect
POSS	possessive
PREP	preposition
PREQ	pre-qualifier
PRES	present
PRESPART	present participle
PRON	pronoun
PRT	particle
PV	plural noun or verb
QUAN	quantifier
REFL	reflexive

Continued on next page

Table 4.2 – *Continued from previous page*

Abbreviation	**Representing**
REL	relative
SG	singular
SM	soft mutation
SP	singular or plural
SUB	subject
SUBJ	subjunctive
SUP	superlative
SV	singular noun or verb
TAG	tag question
V	verb

4.2.2 Gender-specific adjectives in Welsh are not marked for gender in the gloss. For example, *gwyn* (used to modify masculine nouns) and *wen* (used to modify feminine nouns) are both glossed simply 'white' in the manual glosses but 'white.ADJ.M' and 'white.ADJ.F + SM' in the automatic glosses.

4.2.3 Numerals are glossed for gender where appropriate. For example, *dau* and *dwy* are glossed as 'two.M' and 'two.F' respectively. The autogloss is be 'two.NUM.M' and 'two.NUM.F' respectively.

4.2.4 Welsh collective nouns are glossed by the English plural. For example, *moch* (singular collective noun indicating 'pigs') has the gloss 'pigs'. The automatic gloss is 'pigs.N.M.PL'.

4.2.5 In the manual glosses of third person singular possessive constructions, the gender of the possessor is marked only where there is positive evidence of that gender (i.e. either when the possessed noun is mutated, or when a gender-specific pronoun follows the possessed noun, specifically referring to the possessor). The gender is marked on the possessive adjective. For example:

'her mother'
ei mam : POSS.3S mother
ei mham: POSS.3SF mother
ei mam hi : POSS.3SF mother PRON.3SF

'his mother'
ei fam : POSS.3SM mother
ei fam e : POSS.3SM mother PRON.3SM
ei mam e : POSS.3SM mother PRON.3SM

The above applies also to possessive constructions involving non-finite verbs preceded by *ei*. For example:

'he was born'
gaeth (e) ei eni: get.3S.PAST (PRON.3SM) POSS.3SM bear.NONFIN

'he/she was shot'
gaeth ei saethu: get.3S.PAST POSS.3S shoot.NONFIN

4.2.6 When a possessive construction in the first person singular is marked only by mutation of the noun, the possessed noun is followed in the manual gloss by '.POSSD.1S'. For example:

'my father'
nhad : father.POSSD.1S

(However, the automatic gloss will mark *nhad* as 'father.N.M.SG + SM'.)

Contrast this with the following:

fy nhad : POSS.1S father
nhad i : father PRON.1S
fy nhad i : POSS.1S father PRON.1S

Section 5

Tags

5.1.1 Certain phrases in Welsh, usually at the end of an utterance, but also sometimes mid-utterance, are used discursively to engage with the listener. We term these 'tags'. Tags can be agreeing (i.e. they include a verb form that agrees in person, number and tense with the finite verb in the main clause) or they can be non-agreeing. Both kinds are particularly problematic in transcription, as they are seldom seen in the written language and therefore there are no fixed conventions for their spelling. They can also be problematic for glossing.

5.1.2 Table 5.1 is an incomplete list of agreeing tags that may occur, giving the tag as represented by us in the main tier, and its manual gloss. This will serve as a pattern for other agreeing tags with different verbs, tenses and persons.

Table 5.1 – Agreeing tags

Transcription	**Manual gloss**
byddaf	be.1S.FUT
na fyddaf	NEG be.1S.FUT
yn_byddaf	be.1S.FUT.NEG
medri	can.2S.NONPAST
na fedri	NEG can.2S.NONPAST
yn_medri	can.2S.NONPAST.NEG
dylai	should.3S.CONDIT
na ddylai	NEG should.3S.CONDIT
yn_dylai	should.3S.CONDIT.NEG
ydy, yndy	be.3S.PRES
nac (y)dy	NEG be.3S.PRES
yn_dydy, yn_tydy, dydy, tydy	be.3S.PRES.NEG
oes e	be.3S.PRES there
nag oes e	NEG be.3S.PRES there
yn_does e, does e	be.3S.PRES.NEG there

5.1.3 Table 5.2 is a list of common non-agreeing tags with their spellings and

their glosses.

Table 5.2 – Non-agreeing tags

Transcription	**Manual gloss**
felly, (fe)lly	thus
wsti, ysti, sti	know.2S
wchi, (w)chi	know.2PL
yli, (y)li	see.2S.IMPER
ylwch, (y)lwch	see.2PL.IMPER
yn_de, de	TAG
yn_do, do	yes
yn_dyfe, dyfe	PRT.INT.NEG
chimod, chibod	know.2PL
chwel	see.2PL
deud	say.2S.IMPER
deuda	say.2S.IMPER
deudwch, (deu)dwch	say.2PL.IMPER
dywedwch	say.2PL.IMPER
yn_dofe, dofe	yes
dywed, dywad, dŵad	say.2S.IMPER
fel	like
gwed	say.2S.IMPER
iawn	right
na	no
naci	no
naddo	no
nag yfe	NEG PRT.INT
nage	no
ti gweld, ti weld	PRON.2S see.NONFIN
ti (y)n gweld	PRON.2S PRT see.NONFIN
timod, tibod, timbod	know.2S
twel, tiwel, tweld	see.2S
ie,ia	yes
yfe	PRT.INT
sbo	suppose.1S.PRES
wasi	mate

References

Canolfan Bedwyr (2008). *Cysgliad*. Prifysgol Bangor.

Deuchar, M. (2005). Minority language survival in northwest wales: an introduction. In J. Cohen, K. McAlister, K. Rolstad, and J. MacSwan (Eds.), *Proceedings of the 4th International Symposium on Bilingualism*, Somerville, MA, pp. 621–624. Cascadilla Press.

Deuchar, M., P. Davies, J. Herring, M. P. Couto, and D. Carter (2014). Building bilingual corpora. In E. M. Thomas and I. Mennen (Eds.), *Advances in the Study of Bilingualism*, pp. 93–111. Clevedon: Multilingual Matters.

Donnelly, K. and M. Deuchar (2011). Using constraint grammar in the Bangor Autoglosser to disambiguate multilingual spoken text. In *Constraint Grammar Applications: Proceedings of the NODALIDA 2011 Workshop, Riga, Latvia*, NEALT Proceedings Series, Tartu.

Griffiths, B. and D. G. Jones (Eds.) (1995). *Geiriadur yr Academi / The Welsh Academy English-Welsh Dictionary*. Cardiff: University of Wales Press. See also: geiriaduracademi.org.

King, G. (2003). *Modern Welsh : a comprehensive grammar* (2nd ed.). London: Routledge.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Mahwah: Lawrence Erlbaum.

OED (2008). *Oxford English Dictionary*. Oxford: Oxford University Press. See also: oed.com.

Thomas, P. W. (1996). *Gramadeg y Gymraeg*. Cardiff: University of Wales Press.

Thomas, R. J. (Ed.) (1950-2004). *Geiriadur Prifysgol Cymru : a dictionary of the Welsh language*. Cardiff: University of Wales Press. See also: welsh-dictionary.ac.uk/gpc/gpc.html.

Appendix

File Summary

Filename	Length (mm:ss)	Number of main participants	Age (years)	Sex
Davies1	35:07	2	18, 19	FF
Davies2	43:05	2	23, 23	FF
Davies3	35:32	2	13, 15	MM
Davies4	38:38	2	57, 57	MM
Davies5	35:36	3	17, 18, 18	MMM
Davies6	34:51	2	23, 25	MM
Davies7	20:04	2	14, 16	FF
Davies9	18:19	2	19, 22	MM
Davies10	25:20	3	52, 58, 63	MMF
Davies11	33:56	3	52, 67, 72	FMF
Davies12	34:09	2	19, 20	FF
Davies13	32:18	2	19, 20	MM
Davies14	27:46	2	53, 67	FM
Davies15	32:48	2	23, 26	FF
Davies16	34:24	2	16, 16	MM
Davies17	29:49	2	31, 35	MF
Deuchar1	29:49	2	64, 65	FF
Fusser3	32:36	2	31, 32	FF
Fusser4	31:46	2	54, 73	MF
Fusser5	35:25	3	29, 36, 42	MFF
Fusser6	20:10	2	36, 52	FF
Fusser7	25:45	2	36, 39	FF
Fusser8	63:53	3	59, 60, 70	FFF
Fusser9	46:22	2	57, 58	MM
Fusser10	35:31	2	53, 57	MM
Fusser11	45:40	2	52, 77	MM
Fusser12	60:32	3	18, 46, 58	FFF
Fusser13	55:20	3	60, 61, 65	FFM
Fusser14	26:43	2	43, 47	MF
Fusser15	39:46	2	40, 50	FM

Continued on next page

Continued from previous page

Filename	Length (mm:ss)	Number of main participants	Age (years)	Sex
Fusser16	38:55	2	68, 69	FF
Fusser17	49:13	2	47, 65	MM
Fusser18	34:48	2	41, 41	MF
Fusser19	33:24	2	28, 38	FM
Fusser21	37:00	2	16, 17	FF
Fusser22	27:52	2	40, 49	FM
Fusser23	36:50	2	71, 81	MF
Fusser25	32:30	2	25, 25	MF
Fusser26	35:50	2	69, 71	FM
Fusser27	33:42	2	19, 20	FF
Fusser28	20:12	2	21, 30	MM
Fusser29	31:42	2	25, 27	FF
Fusser30	34:37	2	25, 28	FF
Fusser31	36:06	2	12, 43	MM
Fusser32	34:55	3	25, 34, 64	FMM
Lloyd1	34:56	2	53, 53	MF
Robert1	33:50	2	25, 29	FM
Robert2	40:29	2	19, 19	MF
Robert3	32:06	2	15, 16	FF
Robert4	32:42	2	24, 25	FF
Robert5	41:29	2	59, 89	FF
Robert6	29:26	2	27, 56	FF
Robert7	35:31	3	34, 57, 66	MFM
Robert8	39:40	5	77, 79, 81, 82, 86	MMMMM
Robert9	30:32	2	23, 35	FM
Roberts1	35:22	2	25, 33	MM
Roberts2	40:19	2	45, 45	FF
Roberts3	40:08	2	41, 56	FF
Roberts4	40:01	2	19, 21	MF
Smith1	25:14	2	17, 45	MM
Stammers1	29:56	2	61, 72	MM
Stammers2	30:10	2	10, 38	FF
Stammers3	31:16	2	33, 37	FM
Stammers4	30:04	2	40, 42	FM
Stammers5	34:48	2	36, 39	FM
Stammers6	44:59	3	18, 48, 49	FMF
Stammers7	34:06	2	25, 31	MM

Continued on next page

Continued from previous page

Filename	**Length (mm:ss)**	**Number of main participants**	**Age (years)**	**Sex**
Stammers8	30:31	2	66, 67	MF
Stammers9	25:22	2	67, 70	FF
Totals: 69	**40:00:27**	**151**		