Documentation file for the Bangor Patagonia Corpus

March 2025

Queries on the corpus?

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.

Keywords:

1. Introduction

The Patagonia corpus of Welsh-Spanish bilingual speech was recorded in late 2009 and transcribed from 2010 to 2011 as part of a research project funded by the Economic and Social Research Council (ESRC). The main theoretical aim of the project was to test alternative models of code-switching with Welsh-Spanish data.

1.1. Conditions of use

The corpus is made available under the GNU General Public License, version 3 or later.[1] Researchers who use it are requested to subscribe to the TalkBank Code of Ethics[2] and acknowledge the corpus as set out below.

Please refer to the corpus as the 'Bangor Patagonia corpus', and provide a link to the website by which you accessed the corpus, either http://bangortalk.org.uk or http://www.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 addresses given above.

1.2. Canonical version of the data

The most up-to-date version of the data as well as more detailed documentation is available on http://bangortalk.org.uk.

2. The data

The corpus consists of 43 audio recordings and their corresponding transcripts of informal conversation between two or more speakers, involving a total of 94 speakers from Patagonia, Argentina. Participants were recruited via a social network approach: as only a very small percentage of the inhabitants of Patagonia are fluent in both Spanish and Welsh, names of bilingual speakers were sought from local contacts in advance of the fieldworkers' visit. In total, the corpus consists of 195,190 words of text from just under 21 hours of recorded conversation. The transcriptions (in CHAT format – see below) are linked to the digitized recordings through sound links at the end of each main tier. Most recordings were in stereo, and were made using Marantz, Zoom or Microtrack digital audio recorders.

The recordings were made at a place convenient for the speakers, e.g. at their homes or workplaces. After setting up the equipment the researcher would leave the speakers to talk freely with one another. In some cases the researcher re- entered briefly during the recording. This is noted in the transcripts and speech by the researcher is usually not transcribed. The first five minutes of all recordings after the point when the researcher left the room have been deleted, in case the participants' speech was initially affected by the presence of the recorder.

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 signed by a parent or guardian.

There are a few instances where speakers who have not given consent feature in recordings, e.g. a neighbour walking in briefly. In these cases the utterances have been transcribed as *www* and replaced by silence in the audio file. This can sometimes mean that parts of the consenting participants' speech are lost as well where there is overlap with the non-consenting speaker. In addition, beeps have been placed over the names of people about whom sensitive information is given.

The recordings in the corpus are named after the Patagonia region of Argentina where the recordings took place and are numbered in order of the sequence of recording. The sound and transcription files for each conversation share the filename, but have different file extensions (**.wav* or **.mp3* for the sound file and **.cha* for the transcription). For example, *patagonia3.cha* is the transcription of the third recording (sound file *patagonia3.wav*). Basic details regarding the context of each conversation and the speakers involved are given in the transcript headers. Some additional information about the speakers and recordings is available to researchers on request.

All recordings have been transcribed in the CHAT transcription and coding format (MacWhinney 2000), as set out in the online manual.[3]

All transcripts have been done by trained transcribers working on the project: Fraibet Aveledo, Diana Carter, Marika Fusser, Lowri Jones, M. Carmen Parafita Couto, Myfyr Prys and Jonathan Stammers. For 10% 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 score1 for independent transcriptions was 88%. Furthermore, all the transcripts were checked by another member of the transcription team and corrections made accordingly. The team of checkers included the following researchers in addition to the original transcription team: Margaret Deuchar, Lara Gil Vallejo, Jon Herring, Guillermo Montero Melis, and Susana Sabin-Fernández.

All transcripts contain at least three different tiers. In addition to the main tier, required by CHAT, we use an automatically generated gloss tier (%xaut) for the closest English equivalent for each word (including morphological information where rele-

[1] http://gnu.org/copyleft/gpl.html
[2] http://talkbank.org/0share/ethics.html
[3] https://talkbank.org/0info/manuals/CHAT.pdf

Table 1. Welsh phrases

Our transcription	Conventional form	English equivalent
cyd_ddigwyddiad	cyd-ddigwyddiad	coincidence
dim_byd	dim byd	nothing
o_hyd	o hyd	still
o_k	OK	OK
ta_beth	'ta beth	anyway
tu_ôl_i	tu ôl i	behind
un_ai	un ai	either
yn_ôl_i	yn ôl i	back

Table 2. Spanish phrases

Our transcription	Conventional form	English equivalent
ni_fu_ni_fa	ni fu ni fa	neither nor
no_más	no más	only
o_k	OK	OK
o_la_la	olalá	ooh la la
copo_de_nieve	copo de nieve	guelder rose

vant), and a translation tier (%eng), which contains a free translation into English 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 a sound link to the corresponding section of the recording.

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

3. Main tier

3.1. Layout of transcription

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.

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 1950-2004) (GPC), *Geiriadur yr Academi* (Griffiths & Jones 1995) (GyrA) and *Cysgeir* (Canolfan Bedwyr 2008) for Welsh, the *Diccionario de la Lengua Española* online from the Real Academia Española (DLE) and the *Diccionario de Americanismos* (2010) (DA) for Spanish. These are referred to as GPC, GyrA, Cysgeir, DLE and DA respectively throughout this document. Where items are entered as two hyphenated words in these reference dictionaries, they are connected by an underscore in the transcripts instead of a hyphen. 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 Other items which are treated as words are:

(i) interjections and interactional markers, e.g. *ajá*, 'aha', *ay*, 'oh', *hym*, 'hmm', etc.

(ii) proper names (including names of books, films, organisations etc.), a sequence of words being connected by underscores, e.g. *Butch_Cassidy*, *Buenos_Aires*.

(iii) abbreviations (connected by underscore), e.g. *B_B_C*.

(iv) Welsh numbers consisting of two words involving ten which translate into one English word, e.g. *un_ar_ddeg*, 'eleven', *pedwar_deg*, 'forty'. Note that other numbers such as those containing 'hundred', 'thousand' etc. are transcribed as separate words, e.g. *cant saith deg tri*, *ciento setenta y tres*, '173'.

(v) Welsh phrasal prepositions, formed using two morphemes, where separation of the two elements of the word would make any gloss of those individual elements unhelpful, were transcribed with an underscore between the two morphemes; e.g. *oddi_wrth*, which means 'from', but whose individual morphemes translate respectively as 'out of' and 'next to'.

Examples of the phrasal prepositions described in 3.1.3.v are listed in Tables 1 and 2, along with some other phrases which are similarly transcribed because they normally translate into just a single English word.

4 Contractions that do not have entries in the reference dictionaries listed above or, in the case of Welsh, 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*.

5 There are some continuous sequences of characters in the main tier which are not treated as words. These include'simple events' such as &=*laugh*, *xxx* for unintelligible sounds, or the use of an ampersand (&) plus phonetic characters for intelligible sounds without clear meaning, as in e.g. *&pfe* where the speaker produces the non-word [pfe].

6 Please note that pause markings are not used consistently in the transcripts. Additionally, pauses between utterances are generally not marked. We have used the 'lazy overlap' markings ('+<') for overlapping speech.

3.2. Language marking

1 A default language is assigned to each transcription based on the language contributing the greater number of words. The default language is the first language listed in the @Language tier in the file header, and is indicated by the ISO-639-3 abbreviation for the language: *cym* = Welsh, *eng* = English, *spa* = Spanish. Words without any language markers in the transcription are in the default language unless they are part of an utterance preceded by a precode indicating that it is in a non-default language – see 3.2. 2 for details.

2 Individual utterances in the second or third most frequent language are marked with precodes at the beginning of the main tier: e.g. *[- cym]* for Welsh, *[- spa]* for Spanish, and these utterances contain no language tags. In mixed utterances each word in the non-default language is marked by a tag consisting of *@s:* followed by the relevant ISO-639-3 abbreviation: *@s:cym* = Welsh, *@s:spa* = Spanish, *@s:eng* = English, *@s:cym&spa* = undetermined (see 3.2. 4), *@s:spa+cym* = word with first morpheme(s) Spanish, final morpheme(s) Welsh, *@s:cym+spa* = word with first morpheme(s) Welsh, final morpheme(s) Spanish.

3 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). A word or morpheme is considered to be Spanish if it or all its elements are found in either of the Spanish reference dictionaries (e.g. *principito* is considered

Table 3. Language of unlisted words

Our transcription	Language(s)	English equivalent
ah	Welsh & Spanish	ah
ajá	(Welsh &) Spanish	aha
argian	Welsh	good lord
ay	Spanish	oh
bah	Spanish	bah
bechod	Welsh	how sad
diar	Welsh	dear
eh	Welsh & Spanish	eh
ew	Welsh	oh
hym	Welsh	hmm
mm	Welsh	mm
mmhm	Welsh	mmhm
nefi	Welsh	heavens
oh	Welsh & Spanish	oh
oi	Spanish	oh
ta	Welsh	then
ta_ra	Welsh	goodbye
ta_ta	Welsh	goodbye
te	Welsh	be
w	Welsh	ooh
wel	Welsh	well
ý	Welsh	er
ych	Welsh	yuck
ym	Welsh	um

to be a Spanish word because *príncipe* and *-ito* are both listed in DLE). However, we have considered some words not listed in the dictionaries to be either Welsh or Spanish, as indicated in Table 3.

4 The language marker *@s:cym&spa* 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 respective reference dictionaries), that are pronounced in a way that is possible both in Welsh and in Spanish, e.g. [foto] (*ffoto* in Welsh or *foto* in Spanish) or [pjano] (*piano* in both languages).

5 *@s:cym&spa* also marks interjections and interactional markers that may be interpreted as ambiguous, e.g. *ah*, *oh*. 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) is marked *@s:cym* as it is only found in the Welsh-language reference dictionaries. There are also some instances where we assigned a language to an interactional marker that was not listed in any of the dictionaries – see 3.2.3.

6 Where a lexeme could belong to both languages, but its pronunciation in a specific occurrence belongs unambiguously to one language only, it will be marked *@s:cym* or *@s:spa* (and written in the orthography of that language) according to its pronunciation. For example, if *hotel* is pronounced with initial [h], it will be marked *@s:cym*, without initial [h] it will be marked *@s:spa*.

7 Proper names and titles are marked *@s:cym&spa* (undetermined) unless there are alternatives in each language in general use, e.g. *Butch_Cassidy@s:cym&spa*, *Buenos_Aires@s:cym&spa*, *Arglwydd_Dyma_Fi@s:cym&spa* (a Welsh hymn), but *Argentina@s:spa*, *Ariannin@s:cym* (the Welsh name for 'Argentina').

3.3. Orthography

1 We have used a Unicode font[4] for the transcription. Occasional non-lexical phonological fragments are spelt out following an ampersand using IPA symbols (e.g. &, ʧ, ʊ),[5] and these may not show up correctly if a Unicode font is not used.

2 Words marked as *@s:spa* (Spanish) are transcribed in conventional Spanish orthography.

3 Words marked as *@s:cym* (Welsh) are transcribed in conventional 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 in usage from conventional orthography:

(i) Colloquial second person singular verb and preposition endings not usually represented in writing are transcribed as *-a* where they are followed by the pronoun *chdi*, e.g. *oedda chdi*, 'you were', *amdana chdi*, 'about you'.

(ii) 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* (the representation of the word in the Welsh-language reference dictionaries).

(iii) 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.

(iv) Morphemes in Welsh which are usually written with an initial apostrophe, such as the marking of the ellipsis of a possessive pronoun in e.g. *'nhad*, 'my father', are transcribed without this initial apostrophe (e.g. *nhad*) owing to the constraints of CHAT.

(v) 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 initial [k]) and *yn Gaerdydd* (with initial [g]), as well as the standard form *yng Nghaerdydd* (with initial [ŋ̊]), according to what is heard, where heard. We have also transcribed the aspirate mutation of /m/ and /n/ after the 3rd singular feminine possessive adjective common in regional varieties, e.g. *ei mham*, 'her mother', with initial [m̥]), rather than standard *ei mam* (with initial [m]).

(vi) There are also quite a few instances in the corpus where speakers who are learners of Welsh use ungrammatical or unconventional forms. These include 'hypermutation', where an already mutated initial consonant undergoes another round of mutation, e.g. *tipyn*, 'a bit' is mutated to *dipyn* and nonstandardly to *ddipyn*.

4 Words whose language source is undetermined are transcribed in Spanish rather than in Welsh orthography, e.g. *avocado@s:cym&spa* rather than *afocado@s:cym&spa*.

5 When words marked as Spanish or undetermined are mutated (where the sound of an initial consonant is changed depending on the grammatical context, see for example King 1993:14-20), the initial (mutated) sound is written in Welsh orthography and the rest in Spanish, e.g. *rhyw ddoctora*, 'some (female) doctor'.

6 There is some variation in the way initial consonants in

[4] http://en.wikipedia.org/wiki/Unicode

[5] https://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds

Welsh have been transcribed. In some instances the transcriber interpreted a word to have a soft mutation where the speaker may simply have used the Spanish variant of a consonant rather than the Welsh one. This is especially true for stops, where Spanish /p/, /t/, /k/ are more similar to Welsh /b/, /d/, /g/ than to Welsh /p/, /t/, /k/. For example, the transcription may record *dâl*, 'payment', with soft mutation), where the speaker was intending to say *tâl* (without mutation).

4. Gloss tier

Each word (see 3.1. 2 and 3.1. 3) in the main tier is given a gloss in the gloss tier (%aut). The gloss tier has been produced automatically using the Bangor Autoglosser,[6] free (GPL) software developed at the Centre – for further details see Donnelly and Deuchar 2011. The transcripts were manually corrected after autoglossing to deal with the small number (less than 2%) of incorrectly-attributed glosses.

1 Non-words (see 3.1. 5) are not glossed.

2 All words are glossed with the closest English-language equivalent (in lower case) and, where appropriate, information about parts of speech. English equivalents of proper names are used where they exist (for example, *Caerdydd@s:cym* is glossed as 'Cardiff'). If there is no English-language equivalent to a name, it is glossed 'name'.

3 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 or Spanish word involves more than one word. For example, *neithiwr* is glossed as 'last_night'.

4 The English lexeme in a gloss is followed by information about parts of speech, separated by dots. Table 4 lists the part-of-speech abbreviations. Some examples:

- Spanish *hijos* is glossed 'son.N.M.PL', meaning'plural of the masculine noun *hijo*'.
- Welsh *mae* is glossed 'be.V.3S.PRES', meaning 'third person singular present of the verb *be*'.
- Spanish *me* is glossed 'me.PRON.OBL.MF.1S', meaning 'oblique pronoun, 1st person singular, masculine or femine'.
- Welsh *fan* is glossed 'place.N.MF.SG+SM', meaning 'singular of the noun *man*, 'place', which can be either masculine or feminine, with a soft mutation'.

5. References

Asociación de Academias de la Lengua Española (2010). *Diccionario de Americanismos*.

Real Academia Española. *Diccionario de la Lengua Española*. https://dle.rae.es

Canolfan Bedwyr (2008). *Cysgliad*. Prifysgol Bangor. https://www.cysgliad.com

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.

Donnelly, K. and Deuchar, M. (2011) Using constraint grammar in the Bangor Auto- glosser to disambiguate multilingual spoken text. In: *Constraint Grammar Applications: Proceedings of the NODALIDA 2011 Workshop*, Riga, Latvia. Tartu: NEALT Proceedings Series. https://dspace.utlib.ee/dspace/handle/10062/19298

[6] https://github.com/donnekgit/autoglosser

Table 4. Part-of-speech abbreviations

Abbreviation	Representing
0	impersonal
123S	1st, 2nd, 3rd person singular
13S	1st, 2nd, 3rd person singular
1P	1st person plural
1S	1st person singular
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
3P	3rd person plural
3S	3rd person singular
3SP	3rd person singular or plural
ADJ	adjective
ADV	adverb
AM	aspirate mutation
ASV	adjective, singular noun, or verb
AUG	augmentative
COMP	comparative
COND	conditional
CONJ	conjunction
DEF	definite
DEM	demonstrative
DET	determiner
DIM	diminutive
E	exclamation
EMPH	emphatic
F	feminine
FAR	far (demonstrative)
FOCUS	item with focus
FUT	future
GER	gerund
H	pre-vocalic h after 3S.F, 1P and 3P possessives
HYP	hypothetical
IM	interactional marker
IMPER	imperative
IMPERF	imperfect
INDEF	indefinite
INFIN	infinitive
INT	interrogative
INTENS	intensive
M	masculine
MF	masculine or feminine
N	noun
NEAR	near (demonstrative)
NEG	negative
NM	nasal mutation
NT	neuter
NUM	numeral
OBJ	object
OBL	oblique
ORD	ordinal
PAST	past
PASTPART	past participle
PL	plural
PLUPERF	pluperfect
POSS	possessive
PRECLITIC	accented form before clitics
PREP	preposition
PREQ	pre-qualifier
PRES	present
PRESPART	present participle
PRON	pronoun
PRT	particle
QUAN	quantifier
REFL	reflexive
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

Griffiths, B. and Jones, D.G. (eds.) (1995). *The Welsh Academy English-Welsh Dictionary / Geiriadur yr Academi*. Cardiff: University of Wales Press. https://techiaith.bangor.ac.uk/GeiriadurAcademi

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, NJ: Lawrence Erlbaum Associates.

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. https://geiriadur.ac.uk/gpc/gpc.html

6. File summary

Filename	Length (mm:ss)	Main participants	Age (years)	Sex
patagonia1	10:02	3	22, 21 ,28	F, F, M
patagonia2	29:18	2	66, 82	F, F
patagonia3	28:44	2	82, 78	F, F
patagonia4	23:33	3	48, 47, ?	F, M, M
patagonia5	22:16	5	90,82,67,72,61	F, F, M, F, F
patagonia6	27:17	2	54, 96	F, F
patagonia7	44:31	2	66, 68	F, F
patagonia8	31:37	2	84, 83	F, F
patagonia9	37:05	2	65, 69	F, F
patagonia10	25:48	2	35, 9	M, F
patagonia11	43:59	3	81,74, 86	F, F, F
patagonia12	29:55	2	78, 25	F, F
patagonia13	28:31	2	61, 42	F, F
patagonia14	31:28	2	63, 74	F, F
patagonia15	29:45	2	81, 42	F, F
patagonia16	32:10	3	54, 54, 46	M, F, F
patagonia17	30:15	2	21,18	F, F
patagonia18	30:46	3	73, 46, ?	M, M, F
patagonia19	43:10	2	38, 8	M, F
patagonia20	34:29	2	67, 58	F, F
patagonia21	33:30	2	60, 60	F, F
patagonia22	26:18	2	84, 56	F, F
patagonia23	29:54	2	63, 64	F, F
patagonia24	30:57	2	53, 88	F, F
patagonia25	29:47	4	48, 44, 13, 8	M, F, F, M
patagonia26	37:54	2	55, 27	F, M
patagonia27	27:45	2	69, 68	F, M
patagonia28	35:07	2	22, 28	F, F
patagonia29	14:26	2	18,18	M, F
patagonia30	33:10	2	74, 71	F, F
patagonia31	39:20	2	81, 70	F, F
patagonia32	30:02	2	71, 75	F, M
patagonia33	30:26	2	72, 29	F, M
patagonia34	28:42	2	58, 34	F, F
patagonia35	32:55	2	75, 71	M, F
patagonia36	33:04	1	70	F
patagonia37	29:28	2	46, 44	F, F
patagonia38	35:21	2	30, 37	F, F
patagonia39	29:30	2	71, 76	F, F
patagonia40	24:09	2	90, 92	F, F
patagonia41	22:05	4	71, 55, 76, ?	M, F, F, M
patagonia42	33:35	2	70, 87	F, F
patagonia43	11:29	2	56, ?	F, M
42	**20:55:46**	**94**		

NOTE: Three speakers feature on two recordings each.