The Miami Corpus: Documentation File

Margaret Deuchar
ESRC Centre for Research on Bilingualism
Bangor University
Bangor
Gwynedd LL57 2DG
United Kingdom
m.deuchar@bangor.ac.uk

A INTRODUCTION

The Miami corpus of Spanish-English bilingual speech was recorded and transcribed between 2008 and 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 Spanish-English data.

Acknowledgments
Please refer to the corpus as the Bangor Miami corpus, and provide a link to the website by which you accessed the corpus, either http://www.talkbank.org or http://bangortalk.org.uk. We request that a copy of any publications that make use of this corpus be sent to us at the above address.

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.

B THE DATA

The corpus consists of 56 audio recordings and their corresponding transcripts of informal conversation between two or more speakers, involving a total of 84 speakers living in Miami, Florida (USA). Participants were recruited via a variety of methods, including advertising and using the research team's extended social network.

From the 56 audio recordings, 15 are transcripts of conversations from one individual, recorded over a longer period of time in conversation with more than one speaker. The participant ('María') was already known by the research team to be a balanced bilingual who frequently and

consistently code-switched in daily conversation, and so she was invited to make recordings of her interactions with colleagues, family and friends. Maria decided when and with whom to make recordings, by means of a small digital recorder worn on her belt with a moderately concealed lapel microphone. She recorded 42 conversations, 15 of which have been selected for transcription on the basis of their acoustic quality. The research team had no control over when or where the recordings were made and also did not have control over the technical aspects such as checking audio levels, environmental noise and changing batteries in the recorder. Maria's interlocutors did not sign consent forms or fill in questionnaires and so the transcripts of the15 recordings only represent Maria's speech, while utterances from other speakers are transcribed as "www".

In total, the corpus consists of 242,475 words of text from 35 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 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.

Sound and transcription files in the corpus are named after the researcher who did the recording 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'/'*.mp3' for the

sound file and '*.cha' for the transcription). For example, Sastre2.cha is the transcription of the second recording made by Sastre (sound file Sastre2.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), in accordance with the 2012 version of online manual available on http://childes.psy.cmu.edu/manuals/chat.pdf. All references to the CHAT manual in this document are to this online version.

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. Additionally, teams from Penn State University (Amelia Dietrich, Giuli Dussias, Chip Gerfen , Rosa Guzzardo, and Jorge Valdes Kroff), Australian National University (Bronwyn Wrigley, Manuel Delicado, and Jennifer Plaistowe) also collaborated in the process of transcriptions.

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 score[1] for independent transcriptions was 83%. Furthermore, all the transcripts were proofread by another member of the transcription team and corrections made accordingly. An additional team of transcribers and checkers included the following researchers in addition to the original transcription team: Margaret Deuchar, Sarah Fairchild, Marika Fusser, Lara Gil Vallejo, Guillermo Montero Melis, Esther Nuñez, Susana Sabin-Fernández, and Jonathan Stammers.

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 relevant), and 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 a sound link to the corresponding section of the recording.

The following contributed to the translation tier: Adriana Acevedo, Olga Bolaños, Vanesa Bonavota, Rubén Chapela, Magdalena Gazda, Ana Muerza, Renata Kendall, Mary Silva, Sara Viñas, and Renée Zeichen.

[1] An innovative method was used based on Turnitin plagiarism detection software (http://www.turnitin.com). Deuchar, M., Davies, P. Herring, J.R., Parafita Couto, M. & Carter, D. (in press) Building bilingual corpora: Welsh-English, Spanish-English and Spanish-Welsh. In I. Mennen and E. Thomas (eds) *Unravelling Bilingualism*. Multilingual Matters.

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

C MAIN TIER

1. Layout of transcription

1.1. Since the theoretical aims of the project included 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.

1.2. Each main tier is divided into units which we call 'words' for the purposes of these conventions. With some exceptions (see C.1.3) a word is considered to be a continuous sequence of characters containing no spaces as found in the *Webster's Dictionary for English*, and in the *Diccionario de la Lengua Española* online from the Real Academia Española and the *Diccionario de Americanismos* (2010) for Spanish. These are referred to as 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. 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).

1.3. Other items which are treated as words are:
   (a) interjections and interactional markers, e.g 'ajá' (= 'aha'), 'ay' (= 'oh'), 'mmhm' (='mhm'), etc.
   (b) propernames (including names of books, films, organisations etc.), a sequence of words being connected by underscores, e.g., 'Nueva_York'.
   (c) abbreviations (connected by underscore), e.g. 'B_B_C'
   (d) examples of phrases that are not found in the DLE and DA are listed below.

Transcribed form	**Conventional form**	**English**
ni_fu_ni_fa	ni fu ni fa	neither nor
no_más	no más	only
o_k	OK	OK
vale_turca	vale turca	it doesn't matter
o_la_la	olalá	ooh la la
copo_de_nieve	copo de nieve	guelder rose

1.4. There are some continuous sequences of characters in the main tier which are not treated as words. These include simple events such as '&=laugh' (see CHAT 7.6.1), 'xxx' for unintelligible sounds, or the use of an ampersand ('&') plus phonetic characters for intelligible sounds without clear meaning (see CHAT 6.4 for both).

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

2. Language marking

2.1. 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: spa = Spanish, eng = English. 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 next paragraph for details.

2.2. Individual utterances in the second or third most frequent language are marked with precodes at the beginning of the main tier: e.g. [- eng] for English, [- 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:spa = Spanish, @s:eng = English, @s:eng&spa = undetermined (see below, 2.4), @s:spa+eng = word with first morpheme(s) Spanish, final morpheme(s) English , @s:eng+spa = word with first morpheme(s) English, final morpheme(s) Spanish.

2.3. A word or morpheme is considered to be English if it can be found in any of the English-language reference dictionaries. 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 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 English or Spanish, as indicated in the list below.

Transcribed form	**Language**	**English equivalent**
cucu	Spanish	bottom
estrech	Spanish	stretch (jeans)

2.4. The language marker @s:eng&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 English and in Spanish, e.g [pjano] ('piano' in both languages).

2.5. @s:eng&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, 'ay' (='oh') is marked @s:spa as it is only found in the Spanish-language reference dictionaries.

2.6. 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:eng 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:eng, without initial [h] it will be marked @s:spa.

2.7. Proper names and titles are marked '@s:eng&spa' (undetermined) unless there are alternatives in each language in general use, e.g. 'Caracas@s:eng&spa', Sears@s:eng&spa but 'New_York@s:eng' 'Nueva_York@s:spa', (the Spanish word for 'New York').

3. Orthography

3.1. We have used a Unicode font (http://en.wikipedia.org/wiki/Unicode) for the transcription. Occasional non-lexical phonological fragments are spelt out following an ampersand using IPA symbols (http://www.langsci.ucl.ac.uk/ipa/ipachart.html) (e.g. &ʧʊ), and these may not show up correctly if a Unicode font is not used.

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

3.3. Words considered to be Spanish are transcribed in Spanish orthography. We have not represented regional variation in the transcripts, except in cases which have orthographic representation in the Spanish-language reference dictionaries.

3.4. Words whose language source is undetermined are transcribed in English rather than in Spanish orthography, e.g. football, internet, lunch, etc.

D. GLOSS TIER

1. Principles

Each word (see C1.2 and C.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 (http://bangortalk.org.uk/autoglosser.php), 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.

1.1. Non-words are not glossed.

1.2. 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, 'Nueva_York@s:spa' is glossed as 'New_York'). If there is no English-language equivalent to a name, it is glossed 'name'.

1.3. 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 Spanish word involves more than one word. For example, 'veinticinco' is glossed as 'twenty_five'.

1.4. The English lexeme in a gloss is followed by information about parts of speech, separated by dots. Some examples:

- Spanish 'hijos' is glossed 'son.N.M.PL', which means 'plural of the masculine noun "hijo"';
- Spanish 'me' is glossed 'me.PRON.OBL.MF.1S', meaning 'oblique pronoun, 1st person singular, masculine or femine';
- English "wouldn't" is glossed "be.V.1S.COND+NEG", meaning "the first person singular conditional tense of the verb 'be', with an attached negative marker".

2. Parts 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

****

REFERENCES

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

Diccionario de la Lengua Española. Real Academia Española. (www.rae.es)

Deuchar, M., Davies, P., Herring J.R., Parafita, M.C., and Carter, D. (in press) Building bilingual corpora: Welsh-English, Spanish-English and Spanish-Welsh. In I. Mennen and E. Thomas (eds) *Unravelling Bilingualism*. Multilingual Matters.

Donnelly, K. and Deuchar, M. (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.* Tartu: NEALT Proceedings Series.
(http://dspace.utlib.ee/dspace/handle/10062/19298)

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

APPENDIX

File summary:

File name	Length (mm:ss)	No. of main participants	Age (years)	Sex
HERRING1	0:32:18	2	24, 27	F, F
HERRING2	0:30:42	2	21, 19	M, M
HERRING3	0:31:37	2	37, 41	F, M
HERRING5	0:27:10	2	41, 40	F, M
HERRING6	0:28:14	2	43, ?	F, M
HERRING7	0:24:53	2	22, ?	M, M
HERRING8	0:29:43	2	39, 42	F, M
HERRING9	0:32:39	2	21, 20	F, M
HERRING10	0:33:52	2	33, 34	F, F
HERRING11	0:31:00	2	64, 63	M, F
HERRING12	0:33:06	2	22, 20	M, M
HERRING13	0:29:53	2	? , 32	F, F
HERRING14	0:30:04	2	20, 23	M, F
HERRING15	0:29:53	2	? , 21	M, M
HERRING16	0:30:51	2	24, 30	M, M
HERRING17	0:29:58	2	? , 25	M, F
SASTRE1	0:33:52	2	57, 44	M, F
SASTRE2	0:41:00	2	78, 55	F, M
SASTRE3	0:43:02	3	37, 43, 52	M, M, F
SASTRE4	0:31:26	2	29, 22	F, F
SASTRE5	0:29:03	2	36, 66	F, F
SASTRE6	0:30:20	2	43, 42	M, F
SASTRE7	0:29:58	2	19, 15	F, F
SASTRE8	0:33:20	2	63, 13	F, F
SASTRE9	0:40:02	2	48, 60	F, F
SASTRE10	0:39:40	2	35, 35	F, F
SASTRE11	0:40:25	2	30, 60	M, F
SASTRE12	0:30:59	2	48, 41	F, F
SASTRE13	0:29:43	2	25, 19	M, F
ZELEDON1	0:29:38	2	26, 21	F, F
ZELEDON2	0:26:53	2	22, 21	M, F
ZELEDON3	0:30:25	2	19, 11	F, M
ZELEDON4	0:21:48	2	40, ?	M, M
ZELEDON5	0:23:41	2	35, 37	F, F
ZELEDON6	0:30:25	2	21, 19	F, F
ZELEDON7	0:30:20	2	19, 21	F, M
ZELEDON8	0:37:53	2	42, 45	F, F
ZELEDON9	0:30:51	2	12, 09	F, F
ZELEDON11	0:30:40	2	21, 25	M, M
ZELEDON13	0:34:42	2	18, 19	F, F
ZELEDON14	0:33:01	2	22, 19	F, F
MAR1	0:15:02	1	45	F
MAR2	0:01:42	1	45	F

MAR4	0:17:22	1	45	F
MAR7	0:04:34	1	45	F
MAR10	0:17:32	1	45	F
MAR16	2:41:36	1	45	F
MAR18	1:38:40	1	45	F
MAR19	0:53:58	1	45	F
MAR20	0:31:50	1	45	F
MAR21	0:05:29	1	45	F
MAR24	0:41:40	1	45	F
MAR27	1:22:55	1	45	F
MAR30	0:59:58	1	45	F
MAR31	1:45:40	1	45	F
MAR40	2:25:47	1	45	F
Total	**35:11:04**	**84**		