5 Building Bilingual Corpora[1]

Margaret Deuchar, Peredur Davies, Jon Russell Herring, M. Carmen Parafita Couto and Diana Carter

Introduction

The aim of our corpus-based research programme was to investigate code-switching across three language pairs, in order to determine how bilingual individuals in contrasting communities manage to use both their languages within the same conversation. In the process of this investigation we hoped to determine which model of code-switching might best account for the data, and what general statements could be made about the relation between code-switching patterns and structural or extralinguistic factors (see Chapter 6). The three languages we chose to focus on, in three different paired combinations, were Welsh, English and Spanish. Welsh and English were chosen because of their use in our local community in North Wales, UK, and then two other communities that use each of those languages with another were chosen: Miami, USA, where both English and Spanish are used, and Patagonia, Argentina, where Spanish is used in combination with Welsh. The use of the same three languages in different permutations would make it possible for us to compare the use of these languages in different contexts and contribute to the process of unravelling the effect of linguistic structure from that of social context. The location of the three communities is shown in Figures 5.1–5.3.

As there were no pre-existing corpora involving these language pairs in the public domain, we needed to collect data to build three new corpora: Welsh-English, Spanish-English and Welsh-Spanish. We planned to make the corpora available publicly in order to provide a new resource for the wider code-switching research community. This chapter describes the process of building these three corpora, and includes information on data collection and dissemination. The section on data collection will include a discussion of the recruitment and recording process as well as the administration of background questionnaires, whereas the section on dissemination

Figure 5.1 Map showing location of Wales
Source: Lonely Planet/Lonely Planet Images/Getty Images

Figure 5.2 Map showing location of Miami
Source: Map from Ezilon.com (Copyright Reserved)

Figure 5.3 Map showing location of Patagonia
Source: NgMaps/National Geographic Creative

will include information about the process and method of transcription as well as how readers may obtain access to the corpora. In addition, this chapter will provide a brief overview of research that has already been completed using these corpora.

Data Collection

In all three communities we aimed to collect spontaneous data based on informal conversations between pairs of bilingual speakers. We judged that we were more likely to obtain the kind of data we required from conversations between acquainted speakers rather than from interviews with a stranger. This was because we expected communication between bilinguals to be more likely to be conducted in just one language (without code-switching) if the situation were formal. This expectation was based on observations in the communities but also the literature on code-switching (see e.g. Jones, 1995; Zentella, 1990). The amount of conversation we aimed for varied depending on the community and its distance from our home base in Wales.

The targets were 40 hours in Wales (Welsh-English), 30 hours in Miami (Spanish-English) and 20 hours in Patagonia (Welsh-Spanish). Although most of the data collection in Wales took place over a two-year period (2005–2007), it was achieved over a period of two months in Miami (February–April 2008) and one month in Patagonia (October–November 2009). Each conversation recorded lasted about half an hour, and the final count was 151 speakers in Wales, 85 in Miami and 92 in Patagonia.

We aimed to recruit a wide range of bilingual speakers, the main criterion being that participants considered themselves to be bilingual in the two languages associated with each community. Beyond that we wished to record both men and women, of a wide range of ages (but mostly adults), with varying proficiency in the two languages. For reasons of time, proficiency was self-assessed as part of questionnaires administered after the recordings. We also gathered information on a wide range of other extralinguistic variables to be described below.

Participants recruited

In order to recruit participants, letters in both community languages were written, to be sent to speakers known to the researchers or known to a contact of theirs. This followed the social network, or 'friend of a friend', approach adopted by Milroy (1987). The project was described as concerning bilingual communication, and we mentioned that we were seeking bilingual people to record them having an informal conversation with a bilingual member of the family or friend. Recipients were invited to choose their own conversation partner and the place of recording, whether at home or work, for example. Although this freedom of choice meant that we could not control the environmental sound in the recordings, it helped to ensure informality and in the event led to recordings which were mostly highly intelligible. In Wales, the researchers were themselves Welsh-English bilinguals living locally who could draw to some extent on their own social networks. In Miami and Patagonia, however, the data were collected by fieldworkers from Wales who were outsiders to the community. Nevertheless, all fieldworkers were first or second-language speakers of the minority language in each community (just like the participants) and as they were not present for the recordings we do not consider that their language status had an affect on the recordings. In Miami two local assistants were enlisted to help with recruitment and recording, and in Patagonia names of bilingual (Welsh-Spanish) speakers were sought from local contacts in advance of the fieldworkers' visit. In addition to letters being sent to potential participants, posters were placed in universities and in public places in Wales and Miami. These methods enabled lists of potential speakers to be drawn up and participants were then contacted by telephone or e-mail to arrange a time and place for the recording.

In addition to the set of dialogues between pairs of bilingual speakers in the three locations, in one location (Miami) we had the opportunity to also collect a set of data 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. The project benefited in various ways from the inclusion of this second set of data. First, as a case study it complemented the snapshot nature of the dialogues, and could demonstrate intra-speaker variation and differences dependent on having different interlocutors in a way that the shorter thirty minute segments could not. Second, it gave us access to conversational code-switching in the workplace in a way that the dialogues, mostly arranged between friends in a non-work situation, did not. Third, it served as a kind of control for any possible effects of the Observer's Paradox (the problem of the observer affecting the data, *cf.* Labov, 1972). Even though, as set out below, considerable steps were taken to reduce this in the dialogue recording study, the 'María' recordings were longer, and carried out over an extended period, and so the potential for her and the people she interacted with to 'forget' the presence of the microphone was far greater. By comparing the amount and types of code-switching in the 'María' data it would be possible to determine whether or not the circumstances in which the half-hour recordings were made inhibited the use of her two languages in any way.

Background questionnaires

Before beginning the recording process, background questionnaires were prepared in order to obtain information about independent variables which could be used to examine variation in the data. The same questionnaire was used in all three communities, although it was translated into the local languages and slightly adapted for the local context. There were 20 questions altogether and they covered a wide range of information ranging from the more conventional categories of age, gender and occupation to detailed questions about exposure to each language in the family and education, the nature of the participants' social networks, their attitudes to each of their languages, and to code-switching.

Recording procedure

Briefing of participants

As outlined above, the recordings were of conversations between pairs of speakers who already knew each other. At the appointed time, the participants were met by one of the data collectors and given a short briefing about the project: they were told that we were studying how bilinguals communicate

with each other, although no mention was made of mixing languages or code-switching, and that we would record them having a conversation for 35–40 minutes. Before the recording it was explained that their anonymity would be protected by using pseudonyms for them and anyone they mentioned in the course of the conversation, and that they would be able to ask for anything they said to be deleted if they subsequently changed their mind (see the section on ethical considerations below). At this point, too, the researcher offered to suggest topics of conversation if the participants thought that they might require it.

Recording equipment and procedure

The recording equipment used in Wales for most recordings was a Marantz hard disk recorder. This was located in a different room from the recording and received signals from two radio microphones worn by the speakers. The separate microphones allowed the speakers to be recorded on two separate audio tracks, a process which would later facilitate transcription. The researcher was able to monitor the recording via headphones attached to the hard disk recorder. One disadvantage of this recording machine was its physical size and weight, which made it more practical to use it in the university than in external recording venues. For this reason, where transport was a problem, a portable Sony minidisk recorder was used. This operated with a stand-alone microphone placed between the speakers, and the conversation was recorded directly onto the minidisk. Whilst the advantage of this equipment was its portability, its disadvantage was its inability to record on dual stereo audio tracks, making transcription of data recorded on the minidisk recorder potentially more difficult than that recorded using the Marantz recorder.

By the time we collected our data in Miami (2008) and Patagonia (2009) we had obtained portable digital recorders, which achieved comparable acoustic quality to the Marantz hard disk recorder. In making most of the recordings the researcher attached two lapel microphones, worn by the speakers, on long leads to the portable digital recorder, which would be placed on a table or chair. One microphone was connected to the left and the other to the right channel so that each speaker's audio track could be isolated for ease of comprehension during the transcription process. This type of recording was made using either a Marantz portable digital recorder or a Microtrack recorder. In Patagonia some recordings were made with a Zoom recorder which had an external bi-directional microphone. Using the Zoom therefore did not require the participants to wear individual microphones.

A different recording procedure was used for the 'Maria' conversations in Miami. 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. For the most part she recorded two-hour stretches (the storage

limit of the recording device) of herself at work and at home. 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.

Although a considerable proportion of the 'Maria' recordings could not be used because of environmental noise, almost all the other recordings were made in indoor, quiet surroundings, leading to acceptable sound quality. Where there was background noise that interfered with transcription this was minimised digitally. In two cases recordings which had been made out of doors were not of a good enough quality to be used.

Minimising the effects of the Observer's Paradox

Several steps were taken to reduce as much as possible any effect of the Observer's Paradox. The speakers were recorded with partners whom they already knew, in most cases, very well. Audio-recording without video was used so as to protect the anonymity of the speakers. Wherever possible the researcher left the room or house so that their presence would not influence the language choices made by the participants or inhibit code-switching because of any self-consciousness. The pair was also left to talk for several minutes longer than the length that would become the final edited version in the corpus. This was because, following each recording, the first five minutes of each recording was removed in case the participants' speech might have been affected while they became accustomed to the recording equipment. These precautions proved to be highly successful in eliciting the naturalistic data sought. For example, it is noticeable in many of the recordings that both through the relaxed way in which the speakers interact, and the potentially sensitive topics that they discuss, that they did not seem to feel observed. In the extract reproduced as (1), Iris talks about her medical history with her friend James, and this is taken from the very first minute of the edited recording (where the first five minutes have been cut):

(1)	Iris:	eso	lo	que	tengo	que
		that.PRON.DEM[2]	the.DET.DEF.NT.SG	that.PRON.REL	have.V.1S.PRES	that.CONJ
		I wanna … I wanna do more natural things because I really … I'm already on medication, I'm already on Effexor and I've been on anti-depressants since I was seventeen.				

'This is what I have to do … … I wanna … I wanna do more natural things because I really … I'm already on medication, I'm already on Effexor and I've been on anti-depressants since I was seventeen.' (Herring17)

Profiles of participants

The administration of questionnaires to the participants following the recordings provided the information shown in Figures 5.4–5.6.

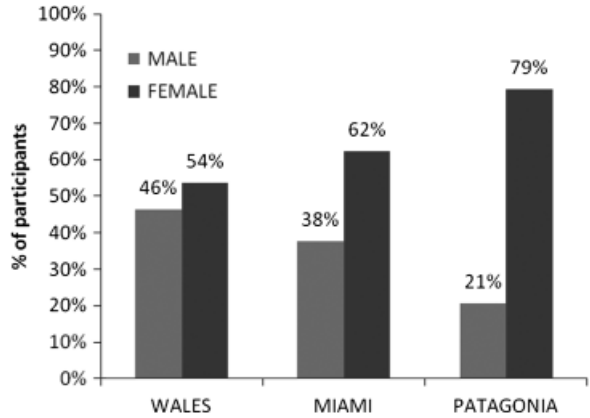

Figure 5.4 Gender distribution of the participants in the three corpora

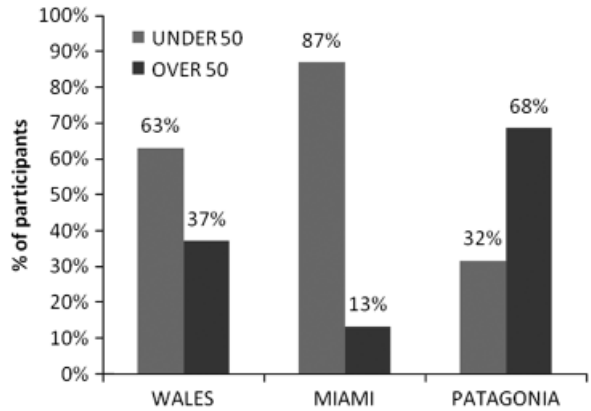

Figure 5.5 Distribution of participants according to their age over or under 50

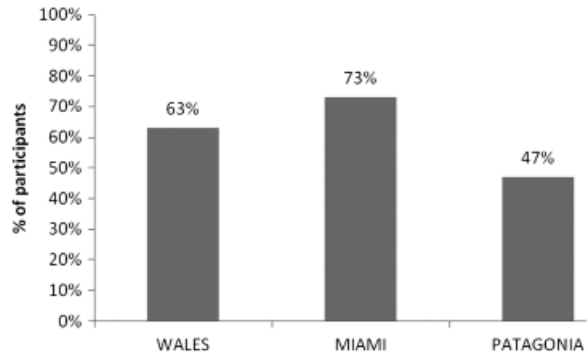

Figure 5.6 Proportion of participants showing balanced proficiency in their two languages

Gender of participants

As shown in Figure 5.4, the gender of participants was most balanced in the Wales sample, with 46% of the 151 speakers being male, and 54% female. In Miami 38% of the speakers were male and 62% female. This imbalance may have been influenced by a number of factors: two of the three data collectors were female; they also provided many of the network contacts for the third collector, the male researcher who was the community outsider; and by chance, two of the primary sites for recruiting speakers (the linguistics department at the university campus, and María's workplace) had a higher percentage of female employees. In Patagonia, the majority of the speakers recruited (79%) were female. This can be partly explained by the skewed age distribution of the Patagonia participants, most of whom were over 50 as shown in Figure 5.2. It was not easy to recruit participants under 50 who considered themselves to be bilingual, mostly because the transmission of Welsh as a first language appears almost to have ceased from the 1960s onwards. And of the over 50s group, 20% of the Patagonia participants were over 80, and only one of this group was male. So the predominance of women participants could be partly attributed to the greater longevity of women but also to the fact that our sample approached exhaustiveness of the available speakers and so we had little choice regarding gender of participants.

Age of participants

Figure 5.5 shows the distribution of the participants according to their age. As is shown in the figure, the Miami data have the largest proportion of younger speakers, whereas the Patagonia corpus has the largest proportion of older speakers, with the Wales corpus being somewhere in between the two. What all three corpora have in common is that adult speakers over the full range of ages have been recruited in each corpus.

Self-reported language proficiency of participants

Figure 5.6 shows the proportion of speakers in each of the three communities that showed a similar (high) level of (self-assessed) proficiency in speaking their two languages. We can see that this proportion was highest in Miami, at 73%, followed by Wales (63%), followed by Patagonia (47%). In Miami, of those who did not consider themselves to have the same level of proficiency in both languages, about two thirds were more proficient in English and one third were more proficient in Spanish. In Wales, the opposite pattern was found: there were twice as many people who considered themselves to be more proficient in Welsh than English than people who considered themselves to be more proficient in English than Welsh. We used self-assessment rates to measure bilingual proficiency. We relied on the capacity of the individuals to self-report accurately, a roughly equivalent sense among individuals of what self report means and an unbiased willingness to communicate their proficiency levels (see Gathercole & Thomas, 2007

for information on self-reported proficiency). However, Miami bilinguals self-rate themselves as having a higher degree (73%) of balanced bilingualism than Wales bilinguals (63%) despite the fact that bilinguals from Wales generally acquired both languages at a younger age than Miami bilinguals. Possible explanations for this result are discussed in Chapter 6. As Figure 5.3 shows, Patagonia was the only community that showed a minority of speakers (47%) with balanced proficiency. This proportion would have been even lower were it not for the predominance of older speakers in the data, who were more likely to report balanced proficiency than the younger speakers. The younger speakers usually reported a higher ability in Spanish than Welsh.

Ethical considerations

While in the process of building all three corpora, we were always mindful of ethical considerations, relating both to data collection and to making the corpora available to others. This involved obtaining ethical approval from the University's Ethics Committee, and gathering data in compliance with the legal requirements of the Data Protection Act. As required, consent forms were obtained from all participants, including from visitors who entered the recording area and spoke for a little with the participants, or people who arrived partway through and joined in the conversation. In case of participants under the age of 16, parental or guardian consent was obtained. Participants who signed the consent form agreed to allow researchers attached to the project to do the following:

(1) use the information provided on the questionnaire anonymously for research and/or teaching purposes only;
(2) make available the recorded data (sound and transcripts) on the internet, provided that fictitious names are used in the transcripts;
(3) allow access to the recorded data by other researchers, on the condition that they follow the appropriate code of ethics;
(4) allow the researchers to present some of the data as part of their work in written and oral form.

Participant anonymity was maintained in the transcription stage by the use of pseudonyms. Pseudonyms were chosen at random with the only condition being that the pseudonym reflected the participant's gender and in most cases reflected the language background of the corpus in question – for example Welsh names for participants in the Siarad (Welsh-English) corpus, Hispanic names for participants in the Miami and Patagonia corpora. Within the CHAT (MacWhinney, 2000) transcriptions a three-letter abbreviation of the pseudonym was used to prefix each main tier (see the next section). Pseudonyms for other non-participants mentioned during the recorded conversation, such as friends or family, were also used in order to ensure privacy

for people mentioned who had not taken part in the study, and who therefore could not give their own consent. An additional means of ensuring that participants were happy with their contributions was the offer, made by the researcher collecting the data, to remove any part of their contribution that they did not, in retrospect, want to be included in the corpus. This could include, for example sensitive information about the private lives of friends or family that came out during the conversation. In practice, however, very few participants voiced any objection to their entire contribution being incorporated into the corpus data.

Data Dissemination

In this section we describe the process of transcribing the data and making it available in the public domain.

Transcription method

The data were transcribed before being made available, following the CHAT transcription system and its associated software CLAN (see MacWhinney, 2000 and http://childes.psy.cmu.edu/manuals/CHAT.pdf). This particular transcription system was chosen so that our corpus could be made publicly available on *Talkbank*, where CHAT is the standard software system.

Features of CHAT

The fundamental features of CHAT notation are that utterances are placed on tiers: minimally, a main tier that consists of an orthographic representation of the words in the utterance. There are also optional tiers which may contain phonological and/or phonetic representations, word by word glosses of non-English material, a translation of the utterance, discourse level mark-up, comments and contextual notes that may help in the interpretation of the transcript by the general researcher, and so on. The main tier also has a detailed set of transcription conventions that allow the inclusion of features of natural speech that are not usually provided for by the standard orthography of the language, such as pauses, repetitions, interruptions, overlaps between speakers, false starts and 'retracings' or reformulations.

For our corpus, a further aspect encoded in the main tier is the source language of each word. When we initially transcribed the Welsh-English corpus we followed the LIDES (see the LIPPS group (2000)) system for marking the source language. Welsh words were tagged '@1' and English ones '@2'. Place names that were the same in both languages were tagged '@0', so we would encode *Bangor@0* and *Conwy@0* but *London@2* and *Llundain@1*. Words that were found in the monolingual dictionaries of both languages, for example *clown* in the Welsh-English data (*clown* appears in both Welsh and

English monolingual dictionaries) were also coded with @0' for example *clown@0* unless the pronunciation made the language membership of the word clear. Similar neutral language marking was also used with place names and some interactional markers that we considered to belong to both language systems (and found in both language dictionaries, for example *ah* and *ajá/aha* in Spanish-English, Welsh-English or Welsh-Spanish).

Language marking

Once we started transcribing the Miami corpus and had agreed to submit all of the corpora to *Talkbank*, we needed to make changes to our language marking system in order to comply with the new requirements of CHAT and *Talkbank*. These changes included the assignment of a default language to the overall transcript. This decision was made so that the transcriber would only be required to mark words used in an additional language with the code '@s' throughout the transcript, rather than marking every word. In order to indicate that a word might belong to both languages (formerly marked as '@0'), we now use a combination of ISO 639-2 alpha-3 language codes: *eng* for English, *spa* for Spanish, and *cym* for Welsh. Thus, the place name *Bangor* would be given the tag '@s:cym&eng'. In the Miami corpus, for example, the place name *Miami* would be tagged as *Miami@s:*eng&spa. The order of the language codes is determined alphabetically.

In example (2) below a fragment of a transcript[3] is given with glosses and a translation, but otherwise not using the CHAT format:

(2) Carolina: y estuvimos esquiando en *New Hampshire* porque…
and.CONJ be.V.1P.PAST ski.V.PRESPART in New Hampshire because.CONJ
'And we were skiing in New Hampshire because…'

Amelia: *oh,* qué rico!
oh.IM how.ADV nice.ADJ.M.SG
'Oh, how lovely!'

Carolina: **my dad had one of those umtownshares.** (Zeledon 1)

In Figure 5.7 the same fragment is reproduced in the CHAT format, showing the use of language tags. In addition to the language markers outlined above it includes standard CHAT markings for interruption (+/.) and pauses (.).

The assignment of language tags to words from bilingual speech is by no means simple, and the research team held regular workshops to discuss contentious examples, refine the criteria and ensure inter-transcriber agreement. The documentation of the finished corpus will include lists of transcribed words that are not currently in a reference dictionary (neologisms, or very frequent forms that have not yet been recognised by lexicographers) or those that merit attention because of the difficulty in assigning source language.

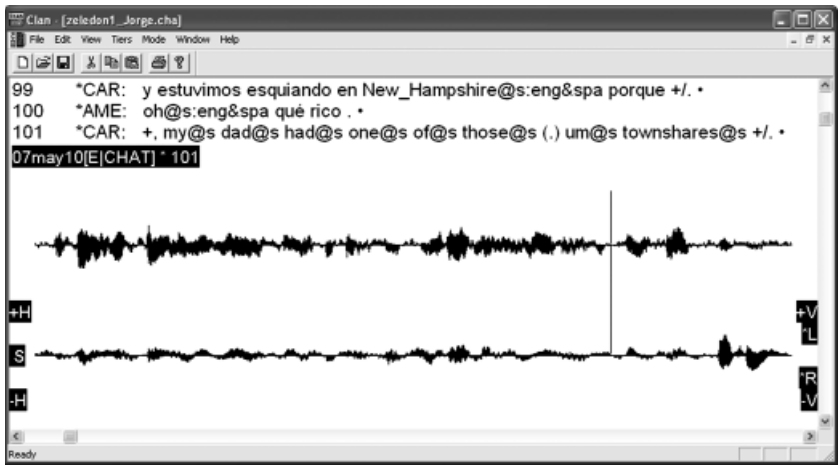

Figure 5.7 Screen shot of transcription in CHAT format

Glosses

In addition to the main tier in the CHAT transcript, we decided to include word-by-word glosses of all non-English material as well as a translation tier. These additional tiers are intended to facilitate the use of the data by members of the public who are not familiar with Welsh or Spanish. The translation tiers were added by the transcribers either while transcribing the main utterance tier, or they were added once the transcript had been finished. For the *Wales* corpus, the gloss tier was manually inserted by the transcribers. However, after consultation with a computational linguist, it was determined that an innovative auto-glossing system could be put in place for the Miami and Patagonia corpora. The auto-glossing procedure works as follows. First, the lines of a CHAT file are loaded into a database, after which each line is segmented into individual words. The words are then looked up in a digital dictionary and are disambiguated using the application of *Constraint Grammar* (cf. Karlsson *et al.*, 1995). Finally, the results are written into a gloss tier, following the *Leipzig*[4] glossing conventions.

Linking the transcriptions to sound

While transcribing, the transcriber also included a sound bullet at the end of each main tier. This links the transcript to the sound and makes it possible to listen to each tier individually while following along with the text. It is also possible to use the continuous play feature and listen to several tiers consecutively. Further information on the technical procedure for inserting sound bullets may be found in the CLAN manual (see http://childes.psy.cmu.edu/manuals/CLAN.pdf).

Transcription reliability

Numerous transcribers worked on the data transcription process over the course of several years. Although they regularly checked queries with each other during this period, it is natural that, in the process of working over a long period of time on a large amount of data, transcribers develop individual strategies for dealing with the phenomena they encounter. Such strategies need to be continually assessed and realigned.

Although all transcribers underwent similar training in the CLAN software and CHAT transcription system, and in most cases worked in the same building and could therefore communicate easily, we decided that a quantitative means of measuring the inter-reliability of transcribers was desirable. We therefore randomly selected ten per cent of recordings from which one minute (taken from the middle of the conversation) was transcribed independently by two researchers and measured the extent of their agreement by an innovative method using plagiarism software. The two resulting independent transcriptions were submitted as separate documents to Turnitin (http://www.turnitin.com), a commercial plagiarism detection service, which compares the two versions and calculates a similarity metric, given as a percentage indicating the overall similarity between the two texts. Turnitin also returns the documents with highlighted annotations, showing the passages in which similarities and differences occur. These highlighted differences can then be checked by the transcribers to see how and why their versions diverge. Any disparity in their general transcription methods can subsequently be harmonised, and any substantial differences found in the two independently transcribed sections can be discussed and resolved.

The use of this anti-plagiarism software does not replace the manual checking of inter-transcriber agreement, but rather provides a quantitative indication of the reliability of their transcription. In our case, the average reliability scores for the three corpora were 74% (Welsh-English), 94% (Spanish-English) and 88% (Spanish-Welsh), where for all corpora th transcribers were a mix of native and non-native speakers of the languages involved. It can be seen that the score for the Welsh-English corpus is slightly lower than the Spanish-English and Spanish-Welsh corpora. This may be explained by the fact that the Welsh-English scores take into account glosses and translation whereas the scores for the other corpora only take into account the main tier. Thus there is more text being compared in the Welsh-English corpus. Furthermore, many of the differences between the transcribers of the Welsh-English corpus were found in the translation tier, where two transcribers sometimes provided a slightly different translation of an utterance even though their transcription of the actual utterance was identical.

We believe our use of Turnitin for this purpose is a logical extension of the originally intended purpose of the software, and is an innovative research tool for strengthening inter-transcriber unity when building corpora. Final

checks were made on the quality of all transcriptions before submitting them, both by using error-checking software and by each one being manually proof-read by someone who had not transcribed it.

Availability of transcriptions

The three corpora will be available on *Talkbank* (see http://talkbank.org/data/BilingBank/Bangor) and also on our own site for bilingual conversational corpora (http://www.bangortalk.org.uk). As outlined above, this will allow other researchers to make use of the data for their own purposes.

Initial Analyses of the Data

The primary purpose of creating a linguistic corpus is, of course, to use its data to respond to research questions. Analyses involving data from all three of our corpora has already begun, and reports of several studies have been published, whereas others are work in progress. In this section we briefly discuss these studies so that the reader can get an idea of what research these corpora make possible.

The research questions addressed in our work include the following: (1) What can code-switching tell us about the effect of language contact on a minority language? (2) What implications do our data have for the debate about the relation between code-switching and borrowing? (3) To what extent can we evaluate competing models of our data? (4) To what extent do extralinguistic factors account for contrasting code-switching patterns in our three corpora?

Code-switching and language contact

We addressed our first research question using our Welsh-English data from Wales, where Wales is a minority language spoken by only about one fifth[5] of the population across the country as a whole. In Deuchar & Davies (2009) we discussed two similar models of the relation between code-switching and language death, examining a sample of our Welsh-English data in order to determine whether or not the linguistic conditions favouring language shift or language death could be found. We concluded that although English words may be inserted in a Welsh grammatical frame, we would need evidence of the Welsh grammatical frame shifting towards English to be concerned about language death. Davies & Deuchar (2010) extend the analysis presented in the previous paper, using data from six speakers from the Welsh-English corpus to measure the extent of word-order convergence found. Almost no clauses are found to show word-order interference from English, and so again the authors conclude that the 'danger' of English's

grammatical influence on Welsh morphosyntax is slight. More details of the analysis in these papers are provided in Davies (2010), a PhD thesis which focuses on the identification of word-order convergence in data from the Welsh-English corpus. Convergence is measured in two ways, firstly by means of the Matrix Language Frame model (Myers-Scotton, 2002) and second by the analysis of auxiliary verb deletion in data from 28 speakers. Little convergence is identified by the first method, which shows that the matrix language or morphosyntactic frame is overwhelmingly Welsh in bilingual clauses. The analysis of auxiliary deletion, however, (also discussed in Davies & Deuchar, 2014) shows age-related variation which can be interpreted as indicating a change in progress. If this were to continue, Welsh word order would change from being auxiliary-subject-verb to subject-verb, thus converging towards English.

Code-switching and borrowing

Stammers (2010) is the published version of a PhD thesis which uses an analysis of the insertion of English verbs in Welsh to address the controversy regarding the distinction between code-switching and borrowing. Using an analysis of the occurrence of soft mutation on the initial consonant of both English and Welsh verbs, he shows a strong effect of frequency on mutation for all categories. (Mutation is a morphophonological process which applies to the initial stop consonants and fricatives of Welsh words under certain syntactic conditions: see Borsley *et al.*, 2007.) The frequency of mutation is nevertheless lower for English verbs not listed in the Welsh dictionary. These could be candidates for switches as opposed to borrowings therefore. Stammers & Deuchar (2012) argue that the data from English verbs provide evidence against the nonce borrowing hypothesis (cf. Sankoff *et al.*, 1990), which they interpret as predicting that there is no difference between frequent and infrequent donor-language items in terms of their degree of integration. As their analysis of English verbs shows that frequency plays an important role in integration measured by the application of soft mutation where expected, Stammers & Deuchar conclude that their evidence refutes the hypothesis and indeed suggests that the existence of a category of nonce borrowings (infrequent donor-language items which are integrated just as well as frequent items) is questionable.

Competing models of our data

Herring *et al.* (2010) evaluate the competing predictions of a Matrix Language Frame (MLF) approach and a Minimalism approach regarding the regularities governing switches between determiners and their noun complements in our Welsh-English and Spanish-English data. They find that the Minimalist approach allowed a higher level of coverage of the data, because

unlike the MLF approach it was able to make predictions regarding nominal constructions occurring without a surrounding clause. Regarding accuracy, the MLF approach appeared to be slightly more accurate, but not significantly so.

Extralinguistic factors influencing code-switching

Carter *et al.* (2010) predicted that speakers' choice of matrix language (morphosyntactic frame) in bilingual clauses would be affected by relative language proficiency levels, the language used in education, the language of social networks and the social identity of the speakers. An analysis of bilingual clauses produced by speakers in our Wales and Miami corpus showed that proficiency did not have as important a role as we expected, but that the other three factors did have an important influence. Chapter 6 discusses the question of whether community-based norms or speaker-based variables have the greatest impact on code-switching patterns, measured in terms of the choice of the matrix language (ML) in bilingual clauses. Because of the uniformity in the data in Wales, speaker-based variables are argued to have little influence on the choice of ML, whereas community norms relating to the factors discussed above appear to be correspondingly uniform and related to the uniform choice of ML (Welsh). In Miami the code-switching patterns are more variable than in Wales, and community norms are shown to be correspondingly variable. Establishing the influence of speaker-based variables on the data in the Miami corpus will await an analysis of a larger set of data. Carter *et al.* (2011) extend the question of the role of community norms to our third set of data, collected in Patagonia. Here we show that although the uniformity of the ML in Patagonia parallels that in Wales and may be predicted on structural grounds, this could not be related to community norms. We speculate as to whether the speakers in Patagonia, comprising such a small minority, can be considered a community in the same way as in Wales and Miami.

Summary and Conclusion

In this chapter, we have provided an account of the methods used to design and build three corpora of bilingual communication in Welsh-Spanish, Spanish-English, and Welsh-English. We have outlined the steps taken to maximise the naturalness of the conversational data collected, and minimise the anticipated effects of the Observer's Paradox. We have also described our methods of transcription which included innovative methods of auto-glossing and checking inter-transcriber reliability. Finally, we have included an overview of recent and current research using the data available in the public domain.

The next chapter explores the relationship between bilinguals' two languages in code-switching further by evaluating the relative roles of intra- and extra-linguistic factors in determining language choice.

Notes

(1) Thanks are due to Kevin Donnelly for his computational expertise and to Jonathan Stammers for his assistance in the preparation of this paper. We would also like to acknowledge the help of the following colleagues who took part in the data collection process for the three corpora: Marika Fusser, Jon Herring, Siân Wynn Lloyd, Elen Robert, Nesta Roberts, Lergia Sastre, Gary Smith and Jonathan Stammers and Marilyn Zeledón.

(2) The glosses in examples (1) and (2) contain abbreviations to be understood as follows: 1P = 1st person plural, 1S = 1st person singular, ADJ = adjective, ADV = adverb, CONJ = conjunction, DEF = definite, DEM = demonstrative, DET = determiner, IM = interactional marker, M = masculine, NT = neuter, PRES = present, PRESPART = present participle, PRON = pronoun, REL = relative, SG = singular, V = verb.

(3) Spanish material is in normal type, English material in bold, and ambiguous material in italics.

(4) The Leipzig glossing rules were developed by the Max Planck Institute for Evolutionary Anthropology: see http://www.eva.mpg.de/lingua/resources/glossing-rules.php

(5) This is based on the results of the 2001 Census: see http://www.byig-wlb.org.uk