Building and Using the Siarad *Corpus*

Bilingual conversations in Welsh and English

Margaret Deuchar
Peredur Davies
Kevin Donnelly

Studies in Corpus Linguistics 81

JOHN BENJAMINS PUBLISHING COMPANY

Building and Using the *Siarad* Corpus

Studies in Corpus Linguistics (SCL)

ISSN 1388-0373

SCL focuses on the use of corpora throughout language study, the development of a quantitative approach to linguistics, the design and use of new tools for processing language texts, and the theoretical implications of a data-rich discipline.

For an overview of all books published in this series, please see
http://benjamins.com/catalog/books/scl

Volume 81

Building and Using the *Siarad* Corpus.
Bilingual conversations in Welsh and English
by Margaret Deuchar, Peredur Webb-Davies and Kevin Donnelly

Building and Using the *Siarad* Corpus

Bilingual conversations in Welsh and English

Margaret Deuchar
University of Cambridge & Bangor University

Peredur Webb-Davies
Bangor University

Kevin Donnelly
Independent Researcher, Llanfairpwll

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Cover design: Françoise Berserik
Cover illustration from original painting *Random Order* by Lorenzo Pezzatini, Florence, 1996.

DOI 10.1075/scl.81

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2017045523 (PRINT) / 2018004380 (E-BOOK)

ISBN 978 90 272 0011 2 (HB)
ISBN 978 90 272 6458 9 (E-BOOK)

John Benjamins Publishing Company · https://benjamins.com

Table of contents

Preface

This book has had a long germination process, the seeds being initially sown in 2005, when I was awarded a grant[1] from the Arts and Humanities Research Council (AHRC) for a project on 'code-switching and convergence in Welsh'. This made it possible to offer two PhD studentships on the topic and to employ several research assistants during the course of the three-year project. They were joined by other colleagues working on related research with the establishment of the ESRC Centre for Research on Bilingualism at Bangor University in 2007. Without the team work of a considerable number of people, this book could never have been written, and grateful thanks are due to the following for their contributions:

Emma Bierings, Dirk Bury, Diana Carter, Peredur Webb-Davies, Kevin Donnelly, Marika Fusser, Jon Herring, Alexandra Hindley, Ellen Kimpton, Sian Lloyd-Williams, Nesta Roberts, M. Carmen Parafita Couto, Myfyr Prys, Elen Robert, Gary Smith and Jonathan Stammers.

Missing from this list of names are of course those of our anonymous bilingual participants who kindly agreed for their conversations to be recorded. *Diolch yn fawr iawn i chi*!

This book is also accompanied by the electronic resources of the BangorTalk website <bangortalk.org.uk>, set up and maintained by Kevin Donnelly. Here all our recordings can be listened to and our transcripts read. Additional information about the data can also be found there.

Margaret Deuchar
Cambridge, April 2017

1. Reference no. 112230/1

CHAPTER 1

Introduction

This book has two parts. In the first part we describe the method used to build a corpus of informal conversational data collected from Welsh-English bilinguals, and in the second part we describe the linguistic analysis of data taken from this corpus. We hope that the book will serve as a 'how to' manual on building a bilingual spoken corpus, including methods of data collection, transcription, automatic glossing and analysis. It will also report our findings from the analysis of the first spoken corpus collected from speakers of one of Britain's oldest language, Welsh. It will be of interest to those working in corpus and socio-linguistics, as well as to those interested in bilingualism, language contact and minority languages.

What is a bilingual speech corpus and how is it useful?

A bilingual speech corpus is a collection of (normally transcribed) recordings of bilingual speakers which can be used for research on bilingualism. In order to address various questions about how bilinguals use their two languages, it is useful to have extended samples of speech by bilinguals produced in a natural way. The speech needs to be transcribed into a written form to make it conducive to analysis by qualitative and quantitative methods, and it needs to be translated and glossed (see Chapter 3) to make it accessible to people who speak English but who do not speak the bilinguals' other language, in this case Welsh.

In collecting our corpus, we had one particular research question in mind: how bilingual speakers manage to combine their two languages in the same conversation (*code-switching*), given that their two languages may not have the same structure, as is the case with Welsh and English. As acknowledged by Backus (2008) many corpora are collected for a specific research project and may not be accessible to those not involved in the project. However, making a corpus available publicly has two advantages: (1) transparency, in that the data on which results are based are verifiable, and (2) utility, in that the data can be used by other researchers and practitioners who may find it useful for their own purposes. They may have research questions which are different from our own, or they may have practical uses for the data, for example using it for listening comprehension by learners of Welsh. One reason for writing this book is to direct readers to our corpus and to make it maximally accessible for them to use.

What is code-switching?

The term *code-switching* is used in this book to mean the use of more than one language in the same conversation, and the *Siarad* corpus contains many examples of code-switching (or *switching*, for short) between Welsh and English, or stretches of speech which contain both Welsh and English words.

Linguists usually differentiate between intraclausal (or intrasentential) and interclausal (or intersentential) code-switching (see Deuchar, 2012 for a discussion of this terminology). Intraclausal code-switching is illustrated by Examples (1) and (2) below and interclausal switching by (3).

(1) dw i 'n **love**-o 'r gwlad **though**. [Fusser27-LIS]
be.1S.PRES PRON.1S PRT love-VBZ DET countryside though
"I love the countryside, though."

(Examples from the *Siarad* corpus are identified by the name of the file they come from, followed by the pseudonym ID for the speaker who produced the example. In Example (1) the intraclausal switches from Welsh to English are shown in bold. Words in bold are English whereas words in normal font in examples are Welsh. Words in italics in the examples are found in both English and Welsh dictionaries. Glosses in smaller font are placed below each word of the utterance and are listed with their expansion in Section 4.2.1 of Appendix 1. Pauses and dysfluencies are usually excluded from examples but included in the original transcripts: see <bangortalk.org.uk>).

(2) mae o **on standby**. [Fusser29-LOI]
be.3S.PRES PRON.3SM on standby
"He's on standby"

(3) *so* bosib hwnna ydy o **I don't know**. [Fusser25-HUN]
so possible that be.3S.PRES PRON.3SM I don't know
"So it's possibly that, I don't know"

As can be seen in Examples (1) and (2) some of the English insertions are individual words (*love* and *though* in (1)) while others are phrases (*on standby* in (2)). As in previous studies of code-switching (see Poplack, 1980, p. 602, Deuchar, 2005, p. 250) the most common insertions are nouns. In *Siarad* the insertions of single English words account for 71% of the English word and phrase insertions.

As will become apparent in what follows, the building of our corpus has benefitted considerably from developments in corpus and computational linguistics, including the use of increasingly sophisticated tools which can be accessed by all researchers and allow access to much larger sets of data. Not surprisingly, monolingual English is the language best represented in current corpora (see McEnery &

Hardie, 2012, pp. 71–92), but corpora have also been developed in languages such as French, Dutch, Italian, Spanish, Arabic, British Sign Language, and Chinese. McEnery & Hardie (2012) mention the existence of bilingual and multilingual corpora, but these are usually 'parallel' corpora and tend to either involve one language with translations into another or two or more monolingual corpora side by side. More relevant to our research aims is the type of bilingual corpus in which the data represent communication between speakers who are bilingual and which often involves code-switching between two or more languages. From about the year 2000 onwards many such bilingual corpora have become available in the public domain (see e.g. <talkbank.org/data/BilingBank> and the appendix to Gardner-Chloros, 2009 on the LIDES project).

Before we collected the data for our own bilingual Welsh-English corpus (known as *Siarad*, 'to speak' in Welsh) some Welsh corpora (which in many cases also included English) were already available.[2] One of the earliest public corpora of Welsh to be established was the Welsh Acquisition Database (or CIG1), set up in 1996. This consists of about 300,000 words or 84 hours of transcribed recordings representing a longitudinal study of children aged 18–30 months and their parents or other family members. The recordings were made in Bangor in north-west Wales and Aberystwyth in mid-Wales. Although the project is described as dealing with the acquisition of Welsh, English words are included in the transcriptions. A related corpus (CIG2), consists of 120 hours of transcribed recordings from 469 children from across Wales aged 3–7. The recordings were collected in 1974–7, and transcribed in 1999–2000.

As for sources of specifically adult language, one of the best known Welsh corpora before our study was a corpus of written rather than spoken Welsh called *Cronfa Electroneg o Gymraeg* (CEG) ('Electronic database of Welsh'). This was developed by Ellis, O'Dochartaigh, Hicks, Morgan, & Laporte (2001). It is described on the website[3] as consisting of "1,079,032 words of written Welsh prose, based on 500 samples of approximately 2000 words each" but in fact a search reveals that it also includes some English words, e.g. *operation* and *politics*. Another written corpus called *Ein Geiriau Ni*[4] ('our words'): *Corpus of Children's Literature in Welsh* was published in 2005 and contains 3,000 words from books written for children.

There were very few publicly available corpora of adult spoken Welsh before *Siarad*, although in 2002–2003 Lancaster University collected, as part of the Language Engineering Resources for British Isles Indigenous Minority Languages (LER-BIML) project, a small sample corpus of spoken Welsh, which is available in

2. For a detailed list please see Appendix 2.

3. <bangor.ac.uk/canolfanbedwyr/ceg.php.en>

4. Available at <egni.org>

the form of transcripts at <lancaster.ac.uk/fass/projects/biml/bimls3corpus.htm>.[5] According to Wilson & Worth (2003, p. 914), the aim was to include at least one sample of each of the activity types included in the British National Corpus. Wilson & Worth report that informed consent was obtained for the use of all of the data, but do not provide information on the demographic characteristics of the speakers involved. As in the case of other corpora described as representing the Welsh language, there are many examples in the data of English insertions. For example, in a recording unusually made in a cathedral (cathedral.txt) we find the English word *apathy* inserted in an otherwise Welsh sermon, and in an informal conversation (chat-1.txt) we find the inserted English phrase *royal institution*. There are also instances of whole clauses in English. For example, in a recording made in a dentist's surgery (dentist-1.txt) the dentist asks the patient in Welsh about the age of their baby, and receives a response in Welsh. He repeats the response in Welsh (*pymtheg mis* 'fifteen months'), but then goes on to ask in English 'How old's Billy now, the same age?' The patient responds in English this time, but then the dentist reverts to Welsh in the next utterance.

The fact that considerable amounts of English are found in pre-existing corpora which are nevertheless described as Welsh suggests that the use of English in otherwise Welsh conversation is a common practice, although one which had not been previously examined in any detail.[6] We realised that in order to draw conclusions about the nature of code-switching between Welsh and English we needed to collect a sizeable corpus with a reasonably large number of speakers. Before collecting the *Siarad* corpus specifically to examine the use of both Welsh and English in bilingual speakers' conversations, a pilot project was conducted at Bangor during which a small corpus was collected as part of a British Academy-funded project entitled "Structural aspects of Welsh-English code-switching". The data consist of recordings of informal conversations involving groups or pairs of speakers in north-west Wales and excerpts from programmes broadcast by the Welsh-language radio station, BBC Radio *Cymru*. The data are known as the Bangor Pilot data and can be found along with their transcripts at <talkbank.org> (navigate to BilingBank). The conversational data were collected by Bangor students and researchers who recorded the informal conversations within their social networks and kindly gave us permission to use the recordings. The transcripts were done in 2004–5, and the transcript headers provide information about the researchers and speakers (identified by pseudonyms) involved. Permission was also obtained from the BBC

5. For more information about this and other similar corpora, please see Appendix 2.

6. Deuchar (2006, p. 1990) briefly reviews early references to Welsh-English code-switching by Thomas (1982), Lindsay (1993), and Jones (1995, 2000).

to use the radio programmes and these were also transcribed. One conversation from the data (williams.cha) was used in a quantitative analysis to test the predictions of the Matrix Language Frame (MLF) model of code-switching proposed by Myers-Scotton (2002), and this was published by Deuchar (2006), (see Chapter 5 for further information). Chapter 5 will also describe how the findings from our pilot were replicated with the much larger amount of data that the *Siarad* corpus made available.

The historical and social context of our corpus

Welsh is a Celtic language belonging (like Breton and Cornish) to the Brythonic group. According to Davies (1993, p. 6), it is generally agreed that Brythonic speakers arrived in Britain from central Europe in the centuries following 600BC. Then after the invasion of Britain by the Romans in 43 AD, Latin became established alongside Brythonic, which evolved into what we know today as modern Welsh and became established in the British Isles in a way that Latin did not. Welsh thus has a longer history in the British Isles than English, which descended from the Anglo-Saxon brought by its speakers invading between 500 and 700 AD. The Anglo-Saxon invaders established kingdoms in eastern Britain in which their language became dominant (Davies, 1993, p. 9). Meanwhile Brythonic survived in non-Anglo-Saxon kingdoms, as Cumbric in southern Scotland and north-west England, as Welsh in the west (now Wales), and as Cornish in the south-west. The inhabitants of these areas were referred to by the Anglo-Saxons as *Welsh*, meaning 'foreigner'. Davies (1993, p. 10) describes how the Welsh and Cumbric speakers adopted the name *Cymry* (from the Brittonic *Combrogi* 'fellow countryman') and called their language *Cymraeg*. *Cymraeg* is still the Welsh term for the Welsh language, while Cumbric is extinct and Cornish died out after the eighteenth century (see Davies 1993, p. 11) although it has recently been revived. In France, Brythonic survives in the form of modern Breton.

In 1066, following the Norman Conquest, French became established as the language of the English court, and was adopted by some Welsh rulers in addition to their Welsh (see Davies, 1993, p. 17). Then in 1284, following the defeat of the Welsh prince Llewelyn by Edward I, Wales was formally annexed to England, and the *Statute of Rhuddlan* established English law in Wales. The dominance of England over Wales was consolidated in 1536, when the Statute of Wales (also known as the Act of Union) imposed English as the official language of Wales. Since 1284 the ruling classes had increasingly begun to speak English in addition to or instead of Welsh, but Deuchar (2005, p. 621) argues that "the Statute of Wales can be seen as the final stage in the linguistic colonisation of Wales".

In the fifteenth and sixteenth centuries, the Protestant Reformation provided something of a respite for Welsh as a result of the Protestant emphasis on worship in the vernacular. In 1567 a Welsh translation of the New Testament by William Salesbury was published, followed by a full translation of both the Old and New Testament by Bishop William Morgan in 1588, and the process of translation as well as the subsequent 1620 revision provided the basis for the modern Welsh literary language (cf. Jones, 1988, p. 128).

The industrial revolution between 1750 and 1850 led to the doubling of the population in Wales, from under 500,000 to more than a million, but Davies (1993, p. 36–7) reports that the proportion of Welsh speakers dropped dramatically from about 80% in 1801 to about 67% in 1851. Government attitudes to Welsh were mostly negative, as shown by a report (Lingen, Symons, & Johnson, 1847) investigating the role of Welsh in education which concluded that "[t]he Welsh language is a vast drawback to Wales, and a manifold barrier to the moral progress and commercial prosperity of the people" (Lingen et al., 1847, Part II, p. 66[7]). The report is sometimes known as the "Treachery of the Blue Books" (see Jones, 1993, p. 547). Its conclusion provides the backdrop for the 1870 Education Act, which established English-medium education in Wales, and is associated with the notorious practice of punishing children for speaking Welsh (Jones, 1993, p. 548).

Jones (1993, p. 548) reports a "slight shift in opinion" in the 1880s when the Society for the Utilization of the Welsh Language in Education advocated the teaching of Welsh in schools, but he points out that this was mainly to facilitate competence in English and would have the effect of achieving bilingualism in Welsh and English. Jones reports that by 1900 69.8% of Welsh speakers were indeed bilingual in Welsh and English, but that the proportion of Welsh monolinguals was to fall drastically in the twentieth century.

Figure 1.1, based on census data[8] and information reported by Jones (1993, p. 550), shows how the overall percentage of Welsh speakers has declined from 50% of the population in 1901 to 19% in 2011. This decline in proportions was accompanied by a decline[9] in the absolute numbers of Welsh speakers, from 929,800 in

7. See <digidol.llgc.org.uk/METS/SEW00003b/frames?div=66&subdiv=0&locale=en&mode=reference>

8. See <iwa.wales/click/wp-content/uploads/5_Factfile_Language.pdf>

9. The slight increase in the proportion of the overall population speaking Welsh in 2001 has been attributed not to a reverse in the decline in the north-western 'heartlands', but to the growth in the teaching of Welsh as a second language.

1901 (Jones, 1993, p. 550) to 562,016 in 2011.[10] In addition, Figure 1.1 also shows a gradual decline in the proportion of Welsh speakers who were monolingual. As Penhallurick (2007, p. 152) points out, by the 1960s not only had monolingualism in Welsh disappeared, but monolingualism in English had taken over and become characteristic of three quarters of the population (Penhallurick, 2007, p. 152). After 1981 census data were not collected on Welsh monolinguals, as it was assumed all respondents to the census would be able to speak English.

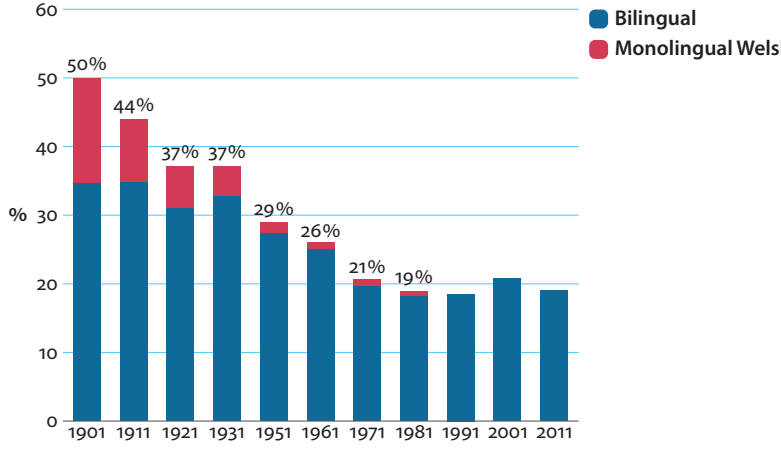


Figure 1.1 Proportion of Welsh speakers in population of Wales, 1901–2011

It should be pointed out that the results shown in Figure 1.1 mask considerable regional variation. Figure 1.2 shows the distribution of Welsh speakers (and hence of bilingual Welsh-English speakers) in 2011, based on the census results. As can be seen, the percentage of Welsh speakers is higher in northern and western areas than elsewhere, although this too declined during the twentieth century.

As Jones (1993, p. 536) says, "It is a minor miracle that Welsh has survived to this day". As he points out, "Throughout fourteen centuries of its existence the Welsh language has been under siege and during that period, whenever bilingual and linguistically mixed communities have come into being, linguistic erosion has occurred with a resultant rejection of Welsh as the primary language." Nevertheless, as we shall see in this book, Welsh is still the primary language for many Welsh speakers addressing other bilinguals, and Welsh grammar is still very much in evidence even when Welsh is combined with English.

10. <ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/2011censuskeystatisticsforwales/2012–12–11>

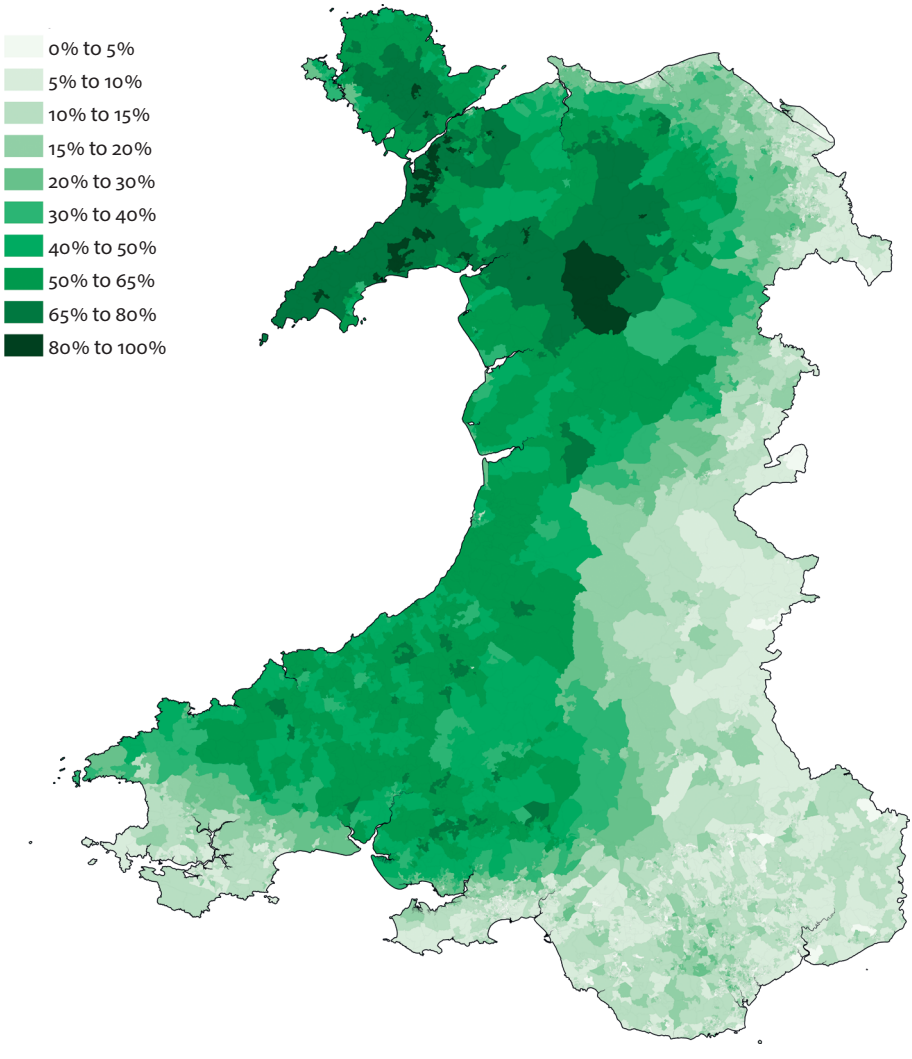


Figure 1.2 Distribution of Welsh speakers in 2011[11]

Although Welsh has been a minority language in terms of its number of speakers in Wales since the early twentieth century, it has had increasing government support, starting with the Welsh Language Act of 1967, according to which, for example, anyone wishing to use Welsh in a court of law should have the right to do so. Also, as Deuchar (2005, p. 623) reports, a separate Welsh television channel (S4C) was

11. <en.wikipedia.org/wiki/welsh_language>

established in 1982 after a long and bitter campaign, and from 1988 it was compulsory for all school pupils in Wales to be taught Welsh as either a first or a second language. Lewis (2008, p. 75) reports that in 2007 20% of primary school children were taught through the medium of Welsh. Since 1999 the Welsh Government (established that year as part of the process of devolution) has been responsible for language policy in relation to Welsh. Williams (2013, p. 142) states that the dominant strategy guiding language policy in post-devolution Wales has been based on the document *Iaith Pawb: National Action Plan for a Bilingual Wales* (Welsh Assembly Government, 2003). The main thrust of the document is the aim to promote the Welsh language in a context of Welsh-English bilingualism. For example, it states that "We want Wales to be a truly bilingual nation, by which we mean a country where people can choose to live their lives through the medium of either Welsh or English and where the presence of the two languages is a visible and audible source of strength and pride to us all" (Welsh Assembly Government, 2003, p. 11). However, Williams (2004) argues that the strategy has not been supported by funding for implementation and Coupland (2010, p. 16) suggests that the document is essentially prescriptive and that the Welsh government's approach reflects an assumption of what he calls "code integrity", whereby each language is kept separate from the other. The *Iaith Pawb* ('everyone's language') strategy has now been superseded by another outlined in the document *Iaith Fyw: Iaith Byw* ('A living language: a language for living') (Welsh Government 2011) which focuses on the development of the Welsh language specifically, once again assuming "code integrity". However, as we shall see in our examination of the *Siarad* corpus, Welsh-English bilingual speakers do not always keep their two languages completely separate, and we shall report on our findings as to how the two languages are used together.

What is the *Siarad* corpus?

Siarad (/ʃarad/) is the Welsh word for 'to speak' or 'speaking'. The *Siarad* corpus is a collection of 69 naturalistic recordings of conversations between bilingual speakers of Welsh and English. The recordings have been transcribed, glossed and translated, and can be found with links to the sound on <bangortalk.org.uk> and <talkbank.org>. The total corpus consists of about 450,000 words, or 40 hours, from 151 speakers. Information about the conventions for transcription as well as basic information about the speakers involved is included in Appendix 1 at the end of this book. The recordings were made 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. The practical aim, on the other hand, was to make the corpus available, so that other researchers and users could access it.

As Table 1.1 shows, the majority (84%) of the words in the corpus are in fact Welsh (defined as "appearing in a reference dictionary of the language"), but a proportion (4%) are English (using the same definition). A substantial minority (13%) are "indeterminate" (defined as words appearing in reference dictionaries for both Welsh and English).

Table 1.1 Number and proportion of words[12] in *Siarad* by language category

Language category	No of words	Percentage of total words
Welsh	373727	83.5%
English	16660	3.7%
Indeterminate	56502	12.6%
TOTAL	447507*	

Note: Not included in the table (but represented in the total) are 573 'mixed words', consisting almost entirely of English verbs with a Welsh suffix but including three instances of Welsh words with English suffixes[13] and 45 words from other languages such as French and Latin.

However, as shown by Stammers (2010, p. 127–128), the actual percentage of English words varies considerably from recording to recording, the lowest percentage being 0.1% and the highest 38.4% according to his calculations. The English words in our recordings appear mostly as single or phrasal insertions into otherwise Welsh clauses, although there are some clauses entirely in English, sometimes representing reported speech in English. One reason that the overall percentage of English words is so low is that Welsh includes a considerable number of English loans, as listed in Welsh dictionaries. For example, the Welsh word for 'shop' appears in Welsh dictionaries as *siop* and is pronounced in exactly the same way as the English word. This means that when it occurs in our recordings it is marked as indeterminate or ambiguous between Welsh and English.

Of course, the language source of individual words in the corpus is not the only indication of the role each language plays in bilingual conversations. In Chapters 5 and 6 we shall describe analyses which take the clause as a unit of analysis, and which differentiate between monolingual and bilingual clauses. Monolingual clauses contain words from only one language, whether Welsh or English, whereas bilingual clauses contain words from both languages. The analysis reported in Chapter 6 found that 10% of clauses were bilingual while 90% were monolingual.

12. Unintelligible words were excluded from the count. A simplified version of this table can be seen at <bangortalk.org.uk>.

13 Two tokens of *saith-ish* 'seven-ish' and one of *coch-y* 'reddy'.

Of the monolingual clauses, 88% were Welsh and only 2% were English. So we can see that the predominance of Welsh words in the corpus overall is also paralleled by the predominance of Welsh monolingual clauses. In Chapter 5 we shall report on a further indication of the importance of Welsh in the corpus, the fact that Welsh grammar is predominant in the bilingual clauses.

Summary of Chapter 1

In this chapter we have introduced the *Siarad* corpus, explained what we mean by a bilingual speech corpus, described the historical and social context, and provided some initial information about the nature of the corpus. In the next chapter we shall describe our method of data collection and the social and demographic characteristics of the speakers whose recorded conversations make up the corpus.

PART 1

Building the corpus

CHAPTER 2

Data collection and profile of the speakers in our corpus

Introduction

Between 2005 and 2008, 40 hours of spontaneous data based on informal conversations between pairs or groups of 151 Welsh-English bilingual speakers[14] were collected. Here we describe our method of data collection and provide an overview of the profile of our speakers. We aimed to recruit a wide range of speakers who considered themselves to be bilingual in Welsh and English. We wanted to record both men and women of a wide range of ages (but mostly adults), with varying proficiency in the two languages, from a variety of backgrounds and places.

Quantity and type of data

Our aim was to collect spontaneous data based on informal conversations between pairs of bilingual speakers. This was because one of our main motivations was to use the data to study code-switching, or the use of Welsh and English in the same conversation, and since communication between bilinguals is more likely to be conducted in just one language (without code-switching) if the situation were formal, we aimed for informal situations, involving conversations between pairs of already acquainted speakers rather than interviews between a bilingual speaker and an unacquainted researcher. This was based not only on observation in the communities, but also on the literature on code-switching (see Jones, 1995). A conversation between two speakers was considered to be more manageable for recording and transcription than one consisting of a group of speakers (Stammers, 2010, p. 57), although a minority of the conversations ended up including more than two speakers, either because there were more than two at the outset, or because other speakers joined the conversation after its beginning.

---

14. The corpus includes a small amount of speech from seven additional speakers (ART, CAI, CIG, CYW, HAF, MAD, SAN). These were unscheduled participants who happened to be there or who arrived during the conversation. They gave permission for the use of their data but we do not have background information for them from speaker questionnaires.

Participants

We aimed to recruit a wide range of speakers, the main criterion for inclusion in the study being that participants considered themselves to be bilingual. In addition, we wished to record both men and women, of a wide range of ages (but mostly adults), with varying proficiency in the two languages. Information regarding speakers' (self-assessed) proficiency was obtained from the background questionnaire as described later in this chapter. Although we would ideally have collected data from all regions of Wales equally, budgetary constraints meant that most of our data collection took place in north-west Wales. Nevertheless, as described below (pp. 22–23), a quarter of our speakers had been brought up outside this area.

Method of recruitment

In order to recruit participants, a letter[15] was written in both Welsh and English by the first author of this book, and was sent to speakers known to the researchers or contacts of theirs. The project was staffed by a research assistant and two PhD students. Since all of these were Welsh-English bilinguals residing in the local area, they were able to make use of their own social networks in identifying potential participants. This followed the social network, or 'friend of a friend', approach adopted by Milroy (1987). Stammers (2010, p. 58), writing as a PhD student member of our team, mentions that this was the most successful of all our recruitment methods. Other methods we used, however, included contacting participants from a previous project and advertising the project at a class for advanced Welsh learners.

The letter sent to potential participants described the project as concerning bilingual communication, and mentioned that we were seeking bilingual people in order 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. While this freedom of choice regarding the location of the recording meant that we could not control the environmental sound in the recordings, it helped to ensure informality and in the event led to recordings of reasonable quality.

In order to recruit additional participants outside the research team's social networks, a set of posters and advertisements was also used to target both university students and people from outside the university. To target university students, informally worded monolingual Welsh and bilingual English/Welsh posters, using the Welsh second person singular pronominal forms associated with informal

15 The text of the letter is in Appendix 5.

or familiar speech, were placed in university buildings, on university computer network bulletin boards, and circulated via e-mail. The Welsh-only posters were intended to help to recruit participants who were fluent in Welsh but who might otherwise ignore bilingual posters. To target people from outside the university, monolingual and bilingual posters with similar wording, but using the more formal Welsh second person plural pronominal forms were placed in shops, libraries, schools, community centres etc. The posters were used locally and also in other parts of Wales where the project team had contacts, in order to widen the search as much as possible. People interested in participating in the project's research were invited to contact the project team by e-mail or telephone. A list of potential speakers was created from these respondents and they were then contacted directly to arrange a recording.

The letter and posters made clear that participants would be paid £10 for taking part. This was possible thanks to AHRC funding and provided an incentive to encourage people to participate.

Method of recording

For each recording, the participants were met by one or more of the researchers at the agreed venue for the recording and were given a short briefing about the project. The briefing was usually in Welsh, since use of this language in Wales tends to cue listeners in to a bilingual mode (cf. Grosjean, 2001), while the use of English is more often associated with a monolingual mode. Participants were told that we were studying how bilinguals communicate with each other and that we would record them having a conversation for 35–40 minutes. No mention was made of mixing languages or code-switching, but they were invited to speak in any language they wished. 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 wished (see the section on ethical considerations below, pp. 18–19). At this point, too, the researcher offered to suggest topics of conversation if the participants thought that they might require it.

The recording equipment used 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 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. In later research conducted in Miami and Patagonia we were able to use portable digital recorders which achieved sound quality which was comparable to that of the Marantz hard disk recorder (see Deuchar, Davies, Herring, Parafita Couto, & Carter, 2014).

After starting the recording, the researcher(s) withdrew to the room where the Marantz recorder was located in order to monitor the sound quality. After about 35 minutes the researchers would return, stop the recording and administer the consent form to participants.

The recordings were later transferred to the computer in .wav format. They were given a filename based on the surname of the researcher who collected the data and a number, e.g. Davies1, Davies2 etc. Before being transcribed, the first five minutes were deleted on the assumption that speakers may have been particularly self-conscious at the beginning of the recording.

Several aspects of the procedure outlined above were designed to maximise the naturalness of the speech recorded by minimising the effects of the Observer's Paradox. These included the absence of the researcher during the recording, and the fact that participants were encouraged to be recorded with people they knew well and would presumably feel relaxed with. Although we expected speakers to be self-conscious at first in the presence of the recording equipment (and for that reason deleted the first five minutes of the recording), we considered that the choice of audio- rather than video-recording would minimise this self-consciousness and help to protect their anonymity. On listening to the recordings our impression is that we achieved a high degree of naturalness in most cases. Speakers appear to interact in a relaxed manner and sometimes to discuss highly sensitive topics which they might have avoided if they had been particularly conscious of being observed. For example, in Davies1 one of the speakers recounts an experience with illness and blood tests, while in Fusser19 the speakers get into an argument about the relative dangers of alcohol and smoking.

Ethical considerations

While collecting the data we were mindful of ethical considerations, relating both to data collection and to making the corpora available to others. Signed 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.

A copy of the consent form is provided in Appendix 3 (Welsh) and Appendix 4 (English).

Questionnaire

In order to obtain information on the background and characteristics of our speakers, we devised a questionnaire to be administered to participants after we had recorded their conversations. The questionnaire was available in both Welsh and English, and a copy can be seen in Appendix 3 (Welsh) and Appendix 4 (English). As reported by Stammers (2010, p. 61), "A balance had to be struck between collecting as much useful detailed information as possible, and keeping the questionnaire as simple and quick to fill in as possible", and the final draft was arrived at through discussion among the team members.

The detailed responses from all participants to the questionnaire were coded and are shown in a spreadsheet which is available at <bangortalk.org.uk>.

Profile of our speakers based on their completion of background questionnaires

There were 20 questions in total, covering 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 their views concerning code-switching.

Gender

The first question on the questionnaire requested participants to tick a box indicating 'Male' or 'Female'. Out of our 151 speakers, 70 were male and 81 were female. This represents 46% males compared with 54% females, or roughly half and half. We found it slightly easier to recruit female participants, and had more of an imbalance in their favour in the early part of the data collection period. But we managed to reduce this and other imbalances in the later part of the data collection period, as Stammers (2010, p. 59) reports. The fairly even balance that we achieved not only ensured that our sample reflected the population from which it was drawn in terms of gender distribution but also allowed for the possibility that the data could be used to investigate gender differences in language use.

Age

Participants were asked for their date of birth so that we could determine their age on the date of recording. We aimed for a wide distribution of ages in order to ensure maximum representativeness, but also to allow for the later investigation of the role of age in language use. The distribution of ages in our data can be seen in Figure 2.1.

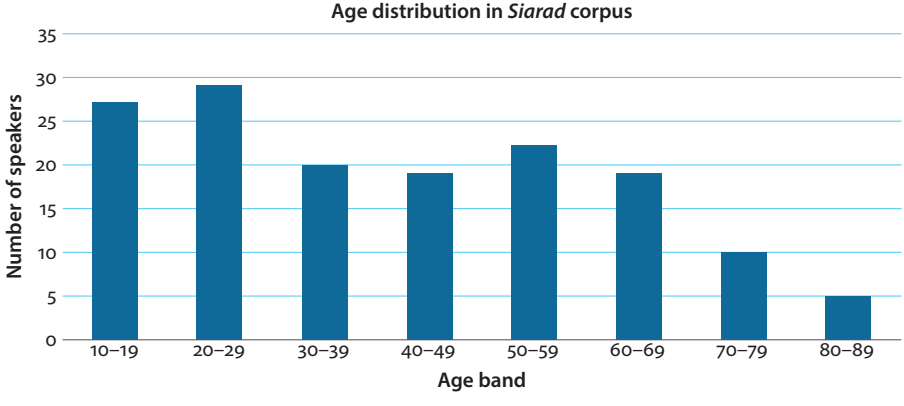


Figure 2.1 Distribution of ages of participants in the *Siarad* corpus

As Figure 2.1 shows, the age band with the largest number of speakers is 20–29, the next largest group being aged 10–19. We intended to include mostly adults in our sample, and half the speakers in the latter group are 18 or 19. However the mean age of the 10–19 age band is 17, reflecting the fact that our youngest speaker is aged 10 and that there are five other speakers under 16. In the twenties age band, where there are 29 speakers, there is a more or less even distribution of ages across the

decade, with the average age being 24. We have roughly 20 speakers in each of the thirties, forties and fifties decades, with numbers being lower only for speakers in their seventies and eighties.

The role of age can also be important in determining whether there is ongoing change in the patterns of language use. This approach adopts an analytical construct known as 'apparent time' (cf. Bailey et al., 1991). In this approach, as Tagliamonte (2012, p. 43) explains, "Age differences are assumed to be temporal analogues, reflecting historical changes in the progress of the change". In general, the speech of younger people is taken as an indication of the way language patterns are changing. In Chapters 6, 7, and 8 the role of age as an indicator of change will be discussed.

Information about both the age and gender of our speakers is included in the 'ID' headers of each transcript. This is illustrated in Figure 2.2, which shows the beginning of the transcript of the recording Fusser9, including some of the headers. Lines 4 and 5 in Figure 2.2 show, among other information, the fact that the speakers are both male and that one is 57 while the other is 58.

```
@Begin
@Languages: cym, eng
@Participants: BAG Baglan Adult, ABE Abel Adult
@ID: cym, eng|siarad|BAG|57;|male|||Adult||
@ID: cym, eng|siarad|ABE|58;|male|||Adult||
@Situation: informal conversation at researcher's house
```

Figure 2.2 Headers in the transcript of recording Fusser9

Occupation

As can be seen from question 3 in the questionnaire included in Appendices 3 and 4, speakers were asked about their present or, if retired or unemployed, their most recent occupation. This was an open question with no predetermined categories, but the responses that emerged fell into the following most common categories: Administrator, Civil Servant, Community Worker, Secondary School Student, University Student and Teacher. Rather few participants reported manual occupations (e.g. only one reported working as a mechanic).

The reason for questioning people about their occupation was to provide the necessary information for an analysis of the effect of social class on their speech. Occupation has been one of the main ways of measuring social class in the UK since the 1970s, although as Savage, Devine, Cunningham, Taylor, Li, Hjellbrekke, Le Roux, Friedman & Miles (2013, p. 220) argue, it should probably be replaced by a more sophisticated, multidimensional model. Savage et al. (2013) identify

seven classes using measures of social, economic and cultural capital, but they nevertheless admit that "there are clear occupational profiles which map onto our seven classes" (Savage et al., 2013, p. 244). As suggested by the examples of our participants' occupations listed above, the distribution of social classes in our sample probably did not reflect that in the population as a whole, as we shall also see below in relation to level of education. In this book we shall not be reporting on the relation between language patterns and speakers' occupations or social classes, but the detailed information available on the BangorTalk website will make it possible for anyone who wishes to conduct an analysis of this kind.

Geographical area

In the third item of the questionnaire participants were asked about the geographical areas in which they had spent significant periods of their lives. This was in order to make it possible for investigators to identify any linguistic features which might be associated with a particular area. Thomas (1992) has outlined the mainly lexical and phonological differences between northern and southern Welsh, and Penhallurick (2004) surveys northern and southern Welsh English. Figure 2.3 shows the distribution of our speakers according to the area[16] in which they were brought up. Willis (2016) was able to use the geographical background of speakers in *Siarad* to investigate the innovation and diffusion of the northern second person singular pronoun *chdi* ('you').

As the map in Figure 2.3 shows, the majority (74%) were brought up in north-west Wales, the area of our base at Bangor, and also the area which Aitchison and Carter (2004, p. 12) refer to as the 'heartland' of Welsh, or the area where it has traditionally been the strongest. The rest of the speakers were either brought up in north-east Wales (7%), mid or south Wales (15%), or outside Wales (6%). Of those speakers brought up outside Wales, eleven were brought up in England[17] and three abroad.[18]

16. For information about the general area (e.g. NW Wales, England) where specific speakers were brought up, please see the questionnaire data at <bangortalk.ac.uk>.

17. The locations in England were the cities of Birmingham, Liverpool, London, Manchester, the counties of Lancashire (NW England) and Sussex (SE England) and the area of the Wirral (NW England).

18. Two people were brought up in the Netherlands and one in the USA.

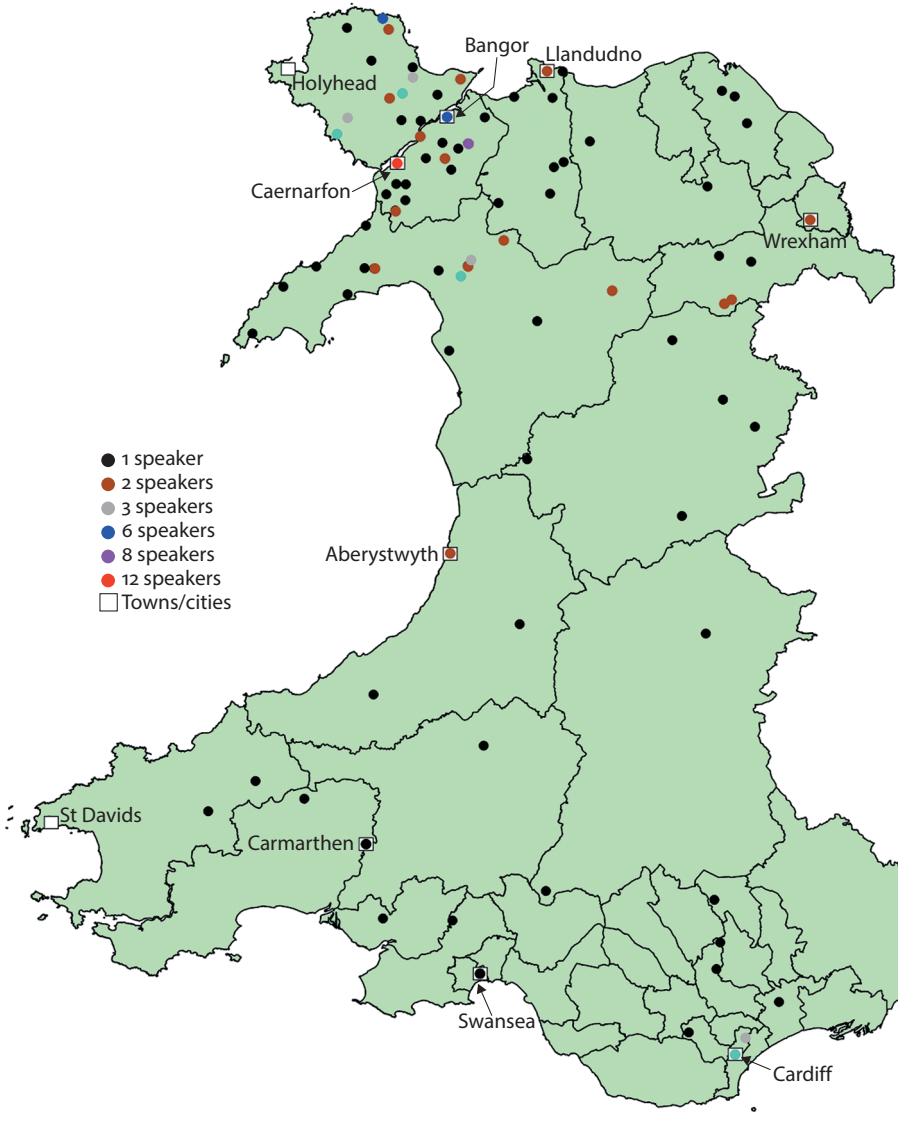


Figure 2.3 Map showing locations in Wales where speakers were brought up

Education

In question 5 of the questionnaire, speakers were asked about their highest level of formal education. This was in order to facilitate analyses of the relation between speakers' language usage and their level of education. Educational success is part of the notion of cultural capital in the analysis by Savage et al. (2013) mentioned

above. Figure 2.4 shows the distribution of our speakers according to their level of education, as reported by Stammers (2010, p. 59).

As can be seen in Figure 2.4, almost half of the participants had at least a Bachelor's degree or equivalent, and 11% had a postgraduate degree. We recognise that graduates are over-represented in our sample. This probably reflects the nature of our research team's social networks, although we did manage to recruit speakers with a lower level of qualifications also, since 25% of our sample achieved a maximum educational attainment of GCSEs or below.

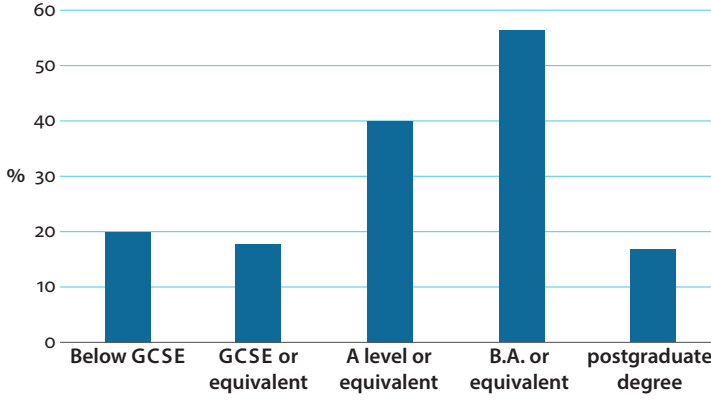


Figure 2.4 Distribution of speakers according to level of education

Age of acquisition

The next two questions (6 and 7) seek information about when speakers first acquired both Welsh and English. Although most speakers acquire at least one language from birth, there is wide variation regarding when the second language is acquired, whether simultaneously with the first language or starting later, whether in early or late childhood or indeed in adulthood. The age of acquisition of the second language as a factor affecting language proficiency and usage has been widely studied and discussed (cf. Montrul, 2008). It has been argued that the simultaneous acquisition of two languages from infancy may lead to different outcomes from the sequential acquisition of the two languages in childhood (cf. Meisel, 2010). Furthermore, as will be discussed in Chapter 8, there is some evidence that simultaneous acquisition of two languages is related to different switching patterns between the two languages as compared with the patterns of speakers who have acquired one language first and then another. Figure 2.5 uses the responses to the questions on age of acquisition to show what proportion of our speakers (a) acquired Welsh first; (b) acquired English first and (c) acquired both simultaneously.

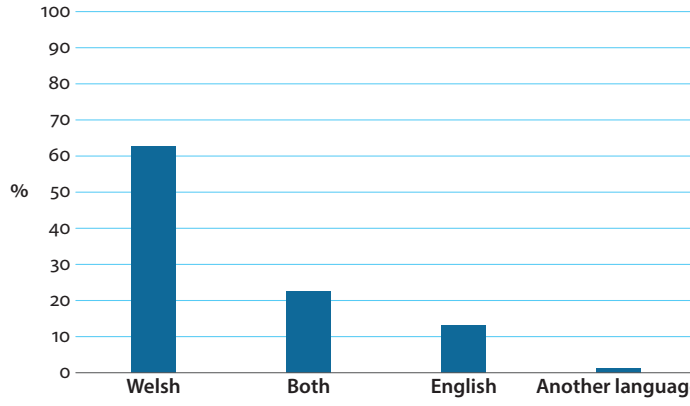
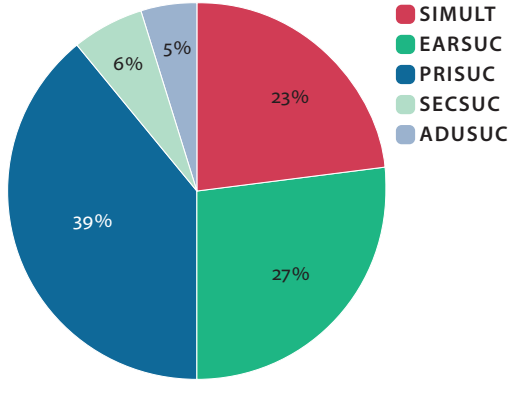


Figure 2.5 First language acquired

As Figure 2.5 shows, the majority (63%) of our participants acquired Welsh before English, just under a quarter acquired both languages simultaneously, and the remainder (13%) acquired English before Welsh. As we shall see in Chapter 8, those who acquired both languages simultaneously combined their languages in a clause more frequently than those who had not. Figure 2.6 shows the distribution of our participants in terms of when they acquired their second language, whether simultaneously with the first or at a later stage. This information was obtained by combining participants' answers to questions 6 and 7 about their age of acquisition of Welsh and English.



Key
SIMULT = simultaneous acquisition of both languages from birth
EARSUC = successive acquisition (second language acquired by age four)
PRISUC = successive acquisition (second language acquired in primary school)
SECSUC = successive acquisition (second language acquired in secondary school)
ADUSUC = successive acquisition (second language acquired in adulthood)

Figure 2.6 Patterns of bilingual acquisition for our speakers

As shown in Figure 2.6, about a quarter of our speakers acquired the two languages simultaneously from birth. Roughly the same proportion acquired the second language very soon after the first, by age four, while the largest proportion (39%) acquired their second language at primary school, showing the importance of education in Wales for bilingual acquisition. Smaller proportions of speakers acquired their second language in secondary school (6%) or adulthood (5%).

Self-reported proficiency in Welsh and English

The next two questions (8 and 9) dealt with participants' own assessment of their proficiency in Welsh and English. Proficiency is an important factor in language production and comprehension, and there is considerable debate (cf. Grosjean, 2013, p. 13) about how it relates to age of acquisition, discussed above. Measuring proficiency by self-report is not ideal since one cannot know whether all participants use the same criteria when choosing to tick the option "Confident in basic conversations", for example. However, a more objective method was not an option for us since it would have required a detailed formal assessment of each participant's proficiency. Alderson (2005, pp. 97–118) reports on a detailed study of the value of self-assessment, albeit in relation to foreign languages, and De Bot (1992) reports that migrant children's self-assessment of their proficiency in their home language appeared to be reliable and generally to correlate with the results of language proficiency tests and teachers' ratings.

One reason for asking the participants to provide assessments of their proficiency in both Welsh and English was to obtain information on the extent to which our speakers could be described as 'balanced bilinguals', or speakers who are "approximately equally fluent in two languages across various contexts" (Baker, 2011, p. 8). This information was obtained[19] by scoring speakers' responses on questions 8 and 9 and then subtracting the score for English from the score for Welsh. If the result was zero this suggested that speakers were equally fluent in both languages and they were classified as balanced bilinguals. If the result was a positive number (only 1 occurred), however, this suggested that the speaker was dominant in Welsh, and if a negative number (only −1 and −2 occurred), this suggested dominance in English. The results are shown in Figure 2.7. As can be seen in this figure, 63% of our speakers turned out to be balanced bilinguals, with 21% rating themselves as slightly more fluent in Welsh than in English. Only 14% considered their English to be more fluent than their Welsh, and most of these speakers considered their Welsh to be only slightly less fluent. This overall profile of the relative fluency of

19. Thanks are due to Diana Carter for the calculations.

our speakers in the two languages suggests that most of them have a high level of fluency of both Welsh and English. Given that Gathercole and Thomas (2009) find that fluency in Welsh is very much related to input in that language in the home, this suggests that most of our participants were exposed to Welsh at home in childhood, something which is borne out by the responses to the next set of questions.

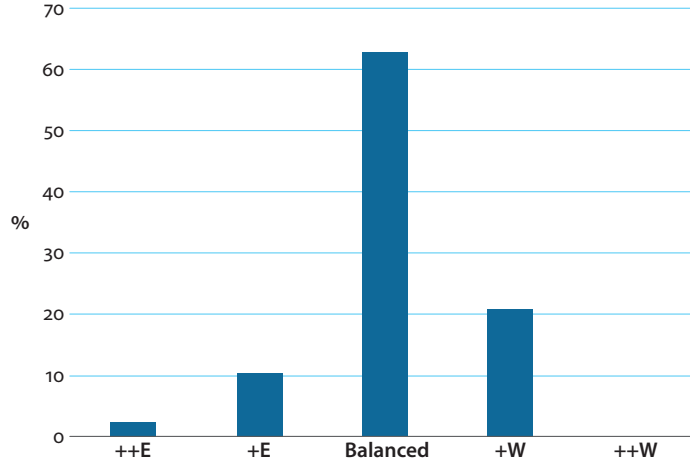


Figure 2.7 Relative language proficiency of our speakers (Based on Figure 5 in Carter, Deuchar, Davies & Parafita Couto (2011, p. 173). 'E' = English, 'W' = Welsh, and the number of pluses indicate the rough degree of dominance in the relevant language.)

Language input in the home

Questions 9, 10 and 11 deal with language input in the home during childhood, and ask in particular about the language spoken by the respondent's mother, father, and "guardian or caregiver". The last term is rather general, and about 20% did not provide any information about this category. Figure 2.8 shows how home language input was distributed across our speakers, taking into account language input from a guardian or caregiver where this information was provided.

As can be seen in Figure 2.8, the largest proportion of speakers (62%) were exposed to only Welsh at home, the next largest (27%) having been exposed to both Welsh and English, and the smallest proportion (10%) having been exposed to only English. These results are consistent with those reported above on self-reported proficiency in that it seems that the high level of Welsh input at home (reported by 89% of Welsh speakers) contributes to a generally high level of proficiency in Welsh and high level of balanced bilingualism.

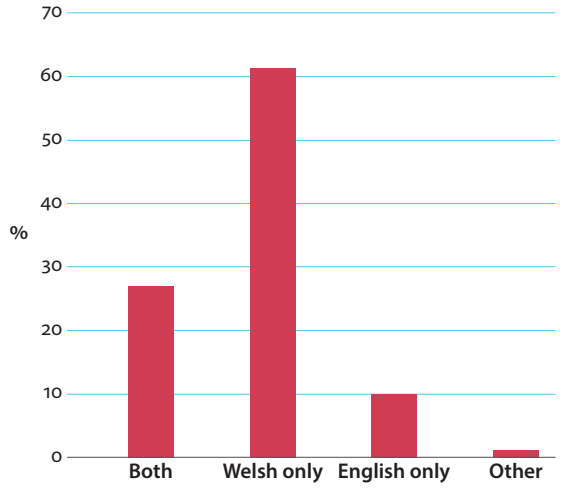


Figure 2.8 Distribution of home language input across speakers

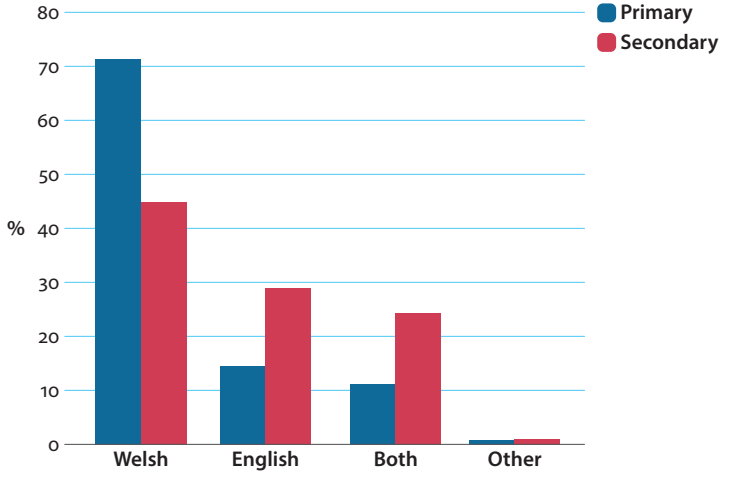


Figure 2.9 Language of education at primary and secondary school

Language of education

The next two questions (13 and 14) deal with the participants' language of education in both their primary and secondary school. In a study using some of the *Siarad* corpus, Lloyd (2008) found that speakers who had received both their primary and secondary education through the medium of Welsh tended to insert more English than those who had had their education in both Welsh and English. Furthermore,

Gathercole and Thomas (2009) found that the language of schooling had an impact on pupils' command of Welsh. Figure 2.9 shows the distribution of the *Siarad* speakers in terms of their language of education at both primary and secondary school.

As Figure 2.9 shows, the vast majority of the speakers had their primary education through the medium of Welsh, but this had dropped to less than half by the time they moved on to secondary education. This pattern is reflected in the (2007) government statistical analysis reported by Lewis (2008, p. 75), showing that a higher proportion of pupils are taught through the medium of Welsh at the primary than the secondary stage.

Languages of social network

The design of question 15 was inspired by the work of Gal (1978, 1979), who defined social networks in her study of language shift in an Austrian village in terms of "all the people (contacts) an individual spoke to in the course of a unit of time" (Gal, 1978, p. 8). Her participants were asked who they had spoken to in approximately the last seven days. In our study, instead of using a unit of time we decided to ask people to consider which five people they spoke to most frequently in general, and to identify the language spoken with that person. The responses would then give us an idea of whether people used mostly Welsh in their everyday life, mostly English, or both. Milroy (1980) was one of the first to demonstrate the importance of social network (defined in terms of scores representing network density and multiplexity) as an extralinguistic variable affecting language use in Belfast, and this has been followed by several other sociolinguistic studies (outlined by Tagliamonte, 2012, p. 37). Regarding code-switching, the main focus of our own studies of the *Siarad* corpus, Li and Milroy (1995, p. 155) argue that "it [social network] is capable of accounting more generally than any other single variable for patterns of code-switching language choice". We may note that the notion of social network is also drawn upon in the new approach to social class by Savage et al. (2013) mentioned above, where it is part of their conceptualisation of social capital (Savage et al., 2013, p. 223).

The participants' responses were scored according to whether they indicated that they spoke Welsh (1), both languages (2) or English (3) with each contact, and then an average score (rounded) was calculated. The responses are displayed in Figure 2.10.

As Figure 2.10 shows, the vast majority of the participants (60%) spoke mainly Welsh with their selected closest contacts, whereas a tenth of that number (6%) spoke mainly English. 34% reported using both Welsh and English equally. In Chapter 6 we report on work (Carter et al., 2011) suggesting that this high level of reported use of Welsh is reflected in the highly frequent choice of Welsh morphosyntax in speakers' conversations.

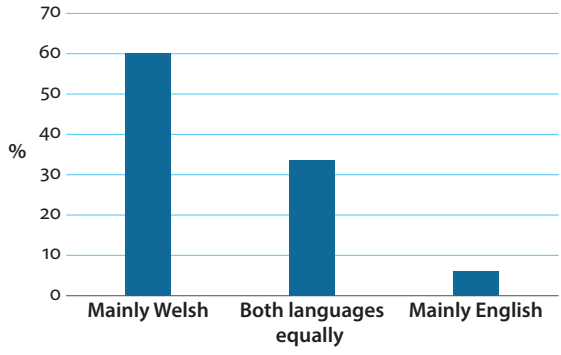


Figure 2.10 Languages spoken with members of speakers' social networks

Attitudes to language

Questions 16 and 17 were designed to obtain information on participants' attitudes to both Welsh and English. We recognise that the issue of the relation between attitudes and behaviour is highly complex (cf. Garrett 2010, p. 28) but we wanted to leave open to researchers the possibility of investigating it. Our questions were designed so that we could find out about two particular types of attitude to language, instrumental and affective. Although the notion of instrumental attitudes to language has mostly been associated with second language learning (cf. Gardner & Lambert 1972), we use the term to represent participants' perception of the usefulness of the Welsh and English languages. This is an aspect of what Garrett (2010, p. 23) calls cognitive as opposed to affective attitudes, the latter involving feelings or emotional reactions. Participants' responses in relation to the descriptions 'modern, influential and useful' in the questionnaire were considered to represent their instrumental attitudes to Welsh and English whereas their reactions to 'friendly, inspiring and beautiful' were considered to represent their affective attitudes. Figure 2.11 shows how the participants' average scores were distributed. The figure shows that speakers' attitudes to Welsh differed more on the affective than the instrumental components. A *t* test established a significance difference for the affective components only. This suggests that whereas speakers do not consider there to be a great deal of difference between the importance of Welsh and English, Welsh is more positively evaluated than English from an affective point of view, presumably because of its association with Welsh culture and identity. These results chime with those reported by Garrett (2010) who found that "Despite the general lack of positivity in the perceptions of the Wales group regarding the present and future vitality of Welsh", they "reported a positive personal commitment to the

Welsh language" (Garrett, 2010, pp. 168–169). Both our study and that reported by Garrett represent the use of direct methods of eliciting attitudes, although there are also more indirect methods. One indirect study of attitudes to Welsh and English was conducted by Bourhis and Giles (1976), who found that the language variety used in theatre announcements made a difference to the degree of co-operation by audiences.

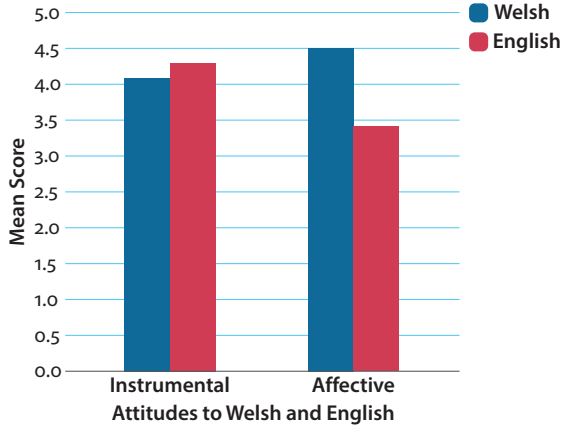


Figure 2.11 Speakers' attitudes to Welsh and English

Self-reported identity

Question 15 dealt with self-reported identity, giving participants a choice between considering themselves Welsh, English or British among others. Giles et al. (1977, p. 169) report the finding that "language is the most important dimension of identity for Welsh bilinguals". Indeed, Bourhis et al. (1973) had already argued that the Welsh language in Wales is more closely related to identity than the French language in Quebec, Canada. They also found that Welsh learners considered themselves to be just as Welsh as fluent Welsh speakers, and significantly more so than monolingual English speakers residing in Wales. This suggests that Welsh is an important marker of Welsh identity at all levels of competence in the language. Given these findings, it is perhaps not surprising that the overwhelming majority of our participants (90%) considered themselves to be Welsh, as shown in Figure 2.12.[20]

20. Based on the figures reported by Stammers (2010, p. 59).

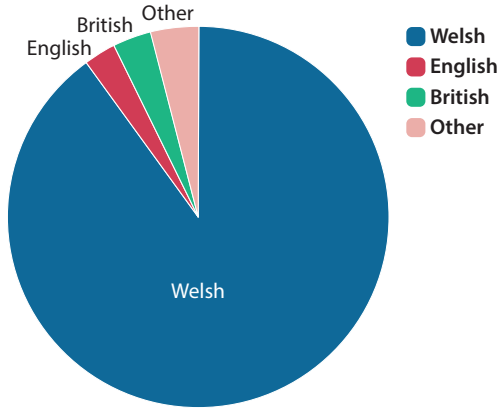


Figure 2.12 Distribution of participants' responses to questions on self-reported identity

In Chapter 6 we report on some of our research which suggests that the homogeneity of our participants' identity can help to account for the uniformity of the grammatical frame of their utterances.

Self-report of own code-switching

Question 19 asks participants indirectly to report on the extent of code-switching in their own speech by asking them whether they keep their languages separate. We cannot tell to what extent this provides us with information about the participants' actual behaviour, since previous research varies in its results regarding the accuracy of speakers' self-reports regarding their own behaviour. For example, Goodz (1989) observed code-switching addressed to children by parents who claimed to use only one language with their child and therefore not to code-switch. On the other hand, Toribio (2002) in a case study of four Spanish-English bilinguals in Santa Barbara, California, was able to show that her participants' descriptions of their own code-switching practices corresponded well with the data she collected from them. Furthermore, a semi-experimental study of Welsh-English code-switching by Parafita Couto et al. (2015), using the same questionnaire as that developed for the *Siarad* corpus, discovered a correlation between the extent of code-switching reported by participants and their actual use of code-switching in the semi-experimental task. Figure 2.13 shows our *Siarad* participants' degree of agreement with the statement "In everyday conversation I keep the Welsh and English languages separate". The Figure shows that more participants disagree than agree with the statement, suggesting that most recognise that they do code-switch.

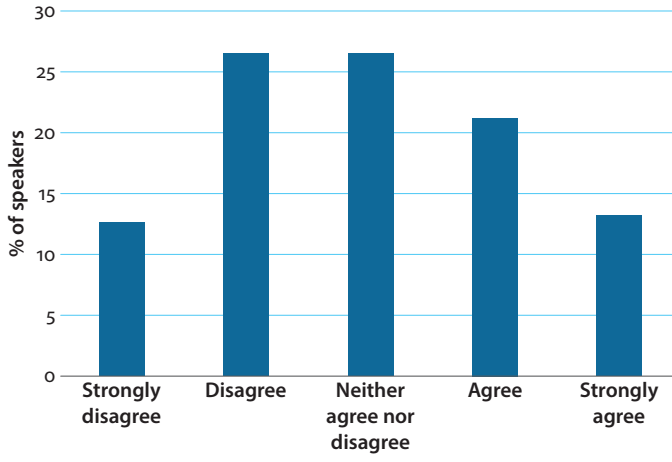


Figure 2.13 Distribution of participants' answers to the question "To what extent do you agree with the following statement: 'In everyday conversation, I keep the Welsh and English languages separate' ".

Attitudes to code-switching

Question 20 probed participants' attitudes to code-switching by asking how strongly they agreed or disagreed with the statement "People should avoid mixing Welsh and English in the same conversation". Although it is not clear how far attitudes to code-switching have an effect on behaviour (cf. Garrett, Coupland, & Williams, 2003), Montes-Alcalá (2000) predicted a correlation between positive vs. negative attitudes to code-switching and its presence or absence in the speech of Spanish-English bilinguals in California. She did not find evidence for such a relationship, but Redinger (2010) on the other hand was able to report a statistical link between attitudes and behaviour in a study of classroom code-switching in Luxembourg. In line with Redinger's findings, Parafita Couto et al. (2015) found in the semi-experimental task outlined above that speakers "who most strongly agree that the languages should be kept separate use fewer mixed determiner phrases than speakers with different attitudes". Figure 2.14 shows the extent to which the *Siarad* speakers agree that the languages should be kept separate. We can see that their opinions cluster around the middle ground, with speakers almost evenly divided between those who agree and disagree. Only a minority of speakers strongly agree or disagree. The diversity of opinions is interesting given purist attitudes to Welsh as expressed in negative statements about code-switching.[21]

21. Jones (1981: 49) describes what he considers the "indiscriminate use of English words and phrases in Welsh utterances" in the speech of "speakers below the age of fifty with a low level of formal education in the language".

As one might expect, the distribution of answers in Figures 2.13 and 2.14 is quite similar. Speakers who report avoiding code-switching are quite likely to think that others should do so as well.

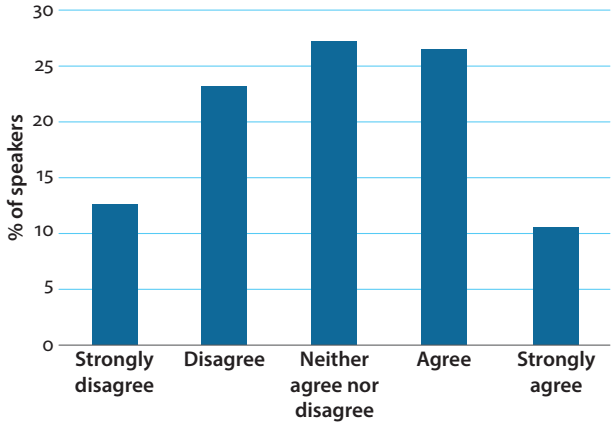


Figure 2.14 Distribution of participants' answers to the question "To what extent do you agree with the following statement: 'People should avoid mixing Welsh and English in the same conversation' ".

Summary of Chapter 2

In this chapter we have provided an overview of the characteristics of our 151 speakers, based on their answers to the questionnaire which we administered after the recordings. These characteristics and their distribution should be borne in mind while reading the rest of this book, and some of them will be the focus of particular analyses, such as that reported in Chapter 6.

In the next chapter we shall describe the methods we used in the transcription of our data.

CHAPTER 3

Transcription of the data

Introduction

In this chapter we will provide an overview of the methods we used for the transcription of our corpus. This will include a description of our methods in converting the audio recordings to a written record and making them available to the general public. It will also include a description of an automatic method of enhancing the information included in our transcriptions and of analysing the data in large quantities.

Choice of transcription system

Although transcription is often thought of as a purely practical process, it involves theoretical assumptions and choices, as Ochs (1979) pointed out. For example, a study focusing on phonetics will require a phonetic as well as an orthographic transcription, and will involve a decision as to the units of analysis, whether phonetic segments (individual sounds) or suprasegmental phenomena such as prosody and intonation. Bilingual data of the kind we collected also required decisions and conventions involving identification of the language source of each unit of analysis. In our study our focus was code-switching, or the use of words from two different languages in the same conversation. This entailed identifying the language source of each word we transcribed, and since we also aimed to distinguish intraclausal code-switching (within the clause) from interclausal code-switching (between clauses) our transcription system needed to facilitate the identification of clause boundaries.

At the time when we selected our transcription system, the CHAT system developed for CHILDES (MacWhinney, 2000) was one of the most suitable available, partly thanks to the pioneering work by the LIDES project (see Barnett, Codó, Eppler, Forcadell, Gardner-Chloros, van Hout, Moyer, Torras, Turell, Sebba, Starren, & Wensing, 2000 and Gardner-Chloros, Moyer, & Sebba, 2007). Gardner-Chloros et al. (2007) explain that although the CHAT system was initially not ideal for adult, multilingual data, having been designed primarily for monolingual child data, it had the advantage of having an institutional support base and of the potential for elaboration and additions. The LIDES group extended the CHAT system in a way which enabled it to deal with bilingual and multilingual data, in particular by language tagging and by recommending dedicated dependent tiers dealing separately with glosses and translations. Barnett et al. (2000, p. 159) describe the gloss tier as

a "word-related tier that presents a word-by-word translation of the main tier into English or another widely-used language. This tier enables researchers to work with data from languages they are not necessarily familiar with". Gardner-Chloros (2009, p. 193) describes the gloss tier slightly more broadly as "giving a word-by-word or even morpheme-by-morpheme gloss of the utterance".

Although we were familiar with the CHAT system from previous work (see Deuchar & Quay, 2000), it is no longer the only system available for multilingual data. For example, a widely used system is ELAN, which was developed by the Max Planck Institute for Psycholinguistics (see <tla.mpi.nl/tools/tla-tools/elan/>). In the CHAT manual (see <talkbank.org/manuals/CHAT.pdf>) MacWhinney points out that since CHAT can be translated to XML (which he describes as "a language used for text documents on the web"), it is compatible with ELAN and many other programs as well.

Features of CHAT

The CHAT manual describes the three major components of a CHAT transcript as being the file headers, the main tier and the dependent tier.

File headers

Each CHAT transcript begins with compulsory headers prefaced by @, including '@Begin' to indicate the beginning of the transcript, '@Languages', '@Participants' and '@ID'. There is also a compulsory header, '@End' which is placed at the very end of the transcript to indicate its termination and optional headers providing additional information. Figure 3.1 shows an initial fragment[22] of the *Siarad* file robert4.

The compulsory header '@Begin' is followed by the header '@Languages', which contains the international codes[23] for the languages found in the transcript. 'Cym' in Figure 3.1 represents *Cymraeg* or 'Welsh' and 'eng' represents English. The next header, '@Participants', introduces information about the people who were present in the recordings. This information is the speaker ID (the first three letters of their pseudonym), their full pseudonym and their role (usually 'Adult'). We can see that there are two speakers in the conversation, Kath and Kim, that their speaker IDs are KAT and KIM, and that they are both adults. There is then one '@ID' header per speaker, and this includes specific information about the speaker, some of which may already have been provided in relation to the headers above. The '@ID' headers are important for the use of some of the CLAN programs which MacWhinney

22. For full version see <bangortalk.org.uk> or <talkbank.org>

23. From the international ISO 639–2 standard: see <loc.gov/standards/iso639–2/php/code_list.php>

(2000) developed for automatic analysis. In the ID lines 4 and 5 in Figure 3.1 we can see that there is information about the speakers' languages, the corpus (*Siarad*) to which the recording belongs, the speaker ID, the speaker's age, their gender and their role (adult). The '@Situation' header is optional, and provides general information about the setting of the recording, in this case an "informal conversation at the university between friends who are studying on the same course". We have provided this information from the notes made by researchers when making the recordings in order to give users some idea of the context of each conversation.

The '@Comment' header is described by MacWhinney as an "all-purpose comment line" (MacWhinney, 2000, p. 17). As can be seen in Figure 3.1, we have used it to provide information about where the speakers were brought up or are living now, to indicate the name of the researcher who collected the data in this transcript, the names of the transcript and linked sound files, and information about the language markers and glosses. Comments can be placed anywhere throughout the CHAT file: general comments applying to the whole transcript, like those in the headers, are introduced by '@Comment', while those referring to a particular utterance are placed on a dependent tier for that utterance, and introduced by '%com'. (See below regarding dependent tiers.)

'@Date' is an optional header which provides information about the date of the recording, in this case 6th November 2006. '@Coder' introduces the name of the person who did the transcription and '@Time Duration' indicates the length of the transcribed conversation, in this case 32 minutes and 42 seconds.

Main tier

Following the headers the transcription of the actual conversation begins. The content of the utterance forms what is called the "main tier" in the transcript, an orthographic representation of the words in the utterance. This is introduced by an asterisk and the first three letters of the speaker's pseudonym, and provides us with information in orthographic form as to what the speaker actually said. So in line 16 of Figure 3.1 we can see the first utterance by KAT, 'faint o gloch wnest ti ddod i Gaerdydd?' This is followed by an utterance by KIM which appears in line 20. Researchers vary as to whether and how they divide the utterance when transcribing, but in our case we decided to put each main clause on a separate line, together with any subordinate clauses. This means that a sequence of main tiers may have the same speaker. We can see this in lines 28, 32 and 36 of Figure 3.1, all of which represent speech by KIM. There are also transcription conventions for use in the main tier that allow the inclusion of features of natural speech that are not usually provided for by the standard orthography of the language. For example, in line 20 the symbol '#' indicates a pause between words. That line also illustrates the use of the symbol '&' preceding phonetic symbols. This is used where the transcriber was able to recognise

```
@Begin
@Languages:	cym, eng
@Participants:	KAT Kath Adult, KIM Kim Adult
@ID:	cym, eng|siarad|KAT|24;|female|||Adult||
@ID:	cym, eng|siarad|KIM|25;|female|||Adult||
@Situation:informal conversation at the university between friends who are studying on the same course
@Comment:	KAT was brought up in north-west Wales and has lived there ever since
@Comment:	KIM was brought up in north-west Wales, where she is living now, after spending a few years in Cardiff
@Date:	06-NOV-2006
@Comment:	Researcher: Elen Robert, Bangor University
@Coder:	Elen Robert, Bangor University
@Time Duration:	00:32:42
@Comment:	Filename: Robert4.CHA; Robert4.WAV
@Comment:	Language markers: @s:eng = English, @s:cym&eng = Undetermined, @s:eng+cym = word with first morpheme(s)
	English, second morpheme(s) Welsh, @s:cym+eng = word with first morpheme(s) Welsh, second morpheme(s)
	English. Untagged words are Welsh except where part of an utterance headed [- eng], in which untagged words are English.
@Comment:	Gloss (in the %aut tier) generated by Bangor Autoglosser (on Linux): 20 May 2012 - there may be an error rate of up to 2%
	in the generated glosses.
*KAT:	faint o gloch wnest ti ddod i Gaerdydd ?
%aut:	size.N.M.SG+SM of.PREP bell.N.F.SG+SM do.V.2S.PAST+SM you.PRON.2S come.V.INFIN+SM to.PREP Cardiff.NAME.PL
%gls:	how_much of clock do.2S.PAST PRON.2S come.NONFIN to Cardiff
%eng:	what time did you come to Cardiff?
*KIM:	um@s:cym&eng o'n i yn Gaerdydd erbyn # tua &ts [//] # chwarter wedi tri # <dydd &s> [/] # dydd Sadwrn .
%aut:	um.IM be.V.1S.IMPERF I.PRON.1S in.PREP Cardiff.NAME.PLACE+SM by.PREP towards.PREP quarter.N.M.SG after.PREP
%gls:	IM be.1S.IMP PRON.1S in Cardiff by about quarter after three.M day day Saturday
%eng:	um, I was in Cardiff by about a quarter past three on Saturday
```

Figure 3.1 Initial fragment of *Siarad* file Robert4

```
22 %gls:  IM be.1S.IMP PRON.1S in Cardiff by about quarter after three.M day day Saturday
23 %eng:  um, I was in Cardiff by about a quarter past three on Saturday
24 *KAT:  +< ah@s:cym&eng dydd Sadwrn .
25 %aut:  ah.IM day.N.M.SG Saturday.N.M.SG
26 %gls:  IM day Saturday
27 %eng:  ah, Saturday
28 *KIM:  ac o'n i isio watsiad y game@s:cym&eng .
29 %aut:  and.CONJ be.V.1S.IMPERF I.PRON.1S want.N.M.SG unk the.DET.DEF game.N.SG.[or].came.AV.PAST+SM
30 %gls:  and be.1S.IMP PRON.1S want watch.NONFIN DET game
31 %eng:  and I wanted to watch the game
32 *KIM:  es i i t ŷ brawd fi de .
33 %aut:  go.V.1S.PAST I.PRON.1S TO.PREP house.N.M.SG brother.N.M.SG I.PRON.1S+SM be.IM+SM
34 %gls:  go.1S.PAST PRON.1S to house brother PRON.1S TAG
35 %eng:  I went to my brother's house, right
36 *KIM:  a (doe)s gynno fo (ddi)m t_v@s:eng licence@s:cym&eng .
37 %aut:  and.CONJ be.V.3S.PRES.INDEF.NEG with_him.PREP+PRON.M.3S he.PRON.M.3S nothing.N.M.SG+SM.[or].not.ADV+SM unk
38 %gls:  and be.3S.PRES.NEG with.3SM PRON.3SM NEG t_v licence
39 %eng:  and he hasn't got a TV licence
```

Figure 3.1 (*continued*)

the sounds produced by the speaker but not the corresponding word. Line 20 also illustrates the use of a double slash '[//]' to indicate that the speaker is 'retracing' and correcting what she is saying. Other symbols inserted on main tiers in this fragment include '+<' at the beginning of line 24, which indicates an overlap between KAT's speech and that of the preceding utterance by KIM in line 20.

The first word in line 20, 'um', is followed by a language marker '@s:cym&eng'. The 's' indicates that the word is in a language that is different from the majority of the words in the transcription. The majority or 'default' language of a transcription is the one with the higher number of words occurring in that transcription and its label appears first in the list following the '@Languages' header (see line 2 in Figure 3.1). In the case of this transcription it is 'cym' or Welsh. The label cym&eng following '@s:'in line 20 indicates that the hesitation word 'um' could be either Welsh or English. A word like 'um' of indeterminate language source has both 'cym' and 'eng' in its tag, in alphabetical order. Words tagged '@s:cym&eng' are written in English orthography. The label 'eng' is used for unambiguously English words, as shown in line 36 where we see 't_v@s:eng' indicating that the speaker has inserted the English word 'TV' ('television') in her utterance.

Words tagged as English are those that appear only in the English reference dictionaries listed in the documentation file, while those considered to be Welsh appear only in the Welsh reference dictionaries. Words which appear in dictionaries from both languages (many of these being loans from English into Welsh, e.g. *siop/shop*) are marked with the tag '@s:cym&eng' as illustrated above[24], indicating 'undetermined language'. Words of this kind are spelled with English orthography for reasons of accessibility, e.g. 'shop@s:cym&eng'. Similar neutral language marking was also used with place names (e.g. 'Bangor@s:cym&eng') and some interactional markers that we considered to belong to both language systems, e.g. 'um@s:cym&eng' mentioned above.. 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 corpus available at <bangortalk.org.uk> includes 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.

Each utterance in a main tier ends with an obligatory full stop or question mark, and after that a solid round circle (known as a 'sound bullet' - not shown in the excerpt above) indicates a link to the sound. Clicking on this while the CHAT software is open leads to information about the exact location of the sound. Although a transcript in CHAT could consist just of a sequence of utterances in main tiers, we

24. Unless their pronunciation unambiguously determines their language source: for more details see the documentation file in Appendix 1.

make use of the optional insertion of dependent tiers providing more information about the utterance in the main tier.

Dependent tiers

Dependent tiers are introduced with the initial symbol '%' and relate to the utterance in the main tier immediately above them. For example, the dependent tiers in lines 21, 22 and 23 provide information relating to the utterance in the main tier in line 20. Line 21 begins with the label '%aut', meaning that this dependent tier contains information about the automatic glosses of the main tier. The production of these glosses will be explained in more detail below. Each word in the main tier corresponds to a gloss in the '%aut' tier, where the sequential position of each gloss corresponds to the same position in the main tier above. So the gloss in the first gloss position of line 21, 'um.IM', corresponds to the first word of the utterance in line 20, and indicates that *um* is an interaction marker ('IM') in English. Glosses in the Bangor autoglossing system (cf. Donnelly & Deuchar, 2011a, 2011b) are conceived of systematically as consisting of the English equivalent of the word being glossed if not in English and a part-of-speech ('POS') tag. In order to maintain the required one-to-one correspondence between the main tier and the gloss tiers, we followed Rule 4 of the Leipzig system (Comrie, Haspelmath & Bickel, 2008) in avoiding spaces within a gloss for a given word, but using full stops to separate lexical and various kinds of morphological information. Any lexical information always precedes morphological information, with a full stop separating the two as in the example of the next gloss 'be.V.1S.IMPERF'. The first part of the gloss ('be') indicates that the lexical meaning of 'o'n' (equivalent to *oeddwn*, first person singular imperfect form of *bod* 'to be') corresponds to that of the English verb *to be*, while the second part of the gloss ('V') provides the grammatical information that 'o'n' is a verb. The third part of the gloss ('1S') indicates that the subject is first person singular, and the fourth part ('IMPERF') that the verb is in the imperfect tense. Where glosses contain more than one type of morphological information (as in 'V.1S.IMPERF') they are always in the same sequence and separated by a full stop. Where more than one lexical item is needed in the gloss to provide the meaning in English of a Welsh word, the lexical items are linked by an underscore as in the gloss 'with_him' for the Welsh word 'gynno' in line 36.

The second dependent tier corresponding to the utterance in line 20 is found in line 22, and is introduced by the label '%glo' where 'glo' stands for 'gloss'. This tier contains the result of manual glossing which was conducted before the automatic glossing system was developed. As can be seen from line 22, it provides slightly less information than the automatic glossing system. The conventions for both the manual and automatic glossing systems are described in the corpus documentation file (see Appendix 1). The manual glossing system was designed by the research

team to balance the needs of the non-Welsh-speaking user with the necessity of transcribing and glossing the data as rapidly as possible, hence the inclusion of less grammatical information.

In line 23 we can see the third dependent tier relating to the utterance in line 20, and it is introduced with the label '%eng' indicating that this tier contains a translation of the utterance into English. The translation tiers in our transcripts provide a freer, more idiomatic translation into English than the glosses can provide. Translations were added by the transcribers either while transcribing the main utterance tier, or once the transcript had been finished.

Our transcription conventions generally follow those outlined in the CHAT manual, but our interpretations of these are described in detail in the corpus documentation in Appendix 1.

Our autoglossing system

The CLAN system which we adopted for transcription using the CHAT format includes some codes for grammatical morphemes (see MacWhinney 2000, p. 102–104) of the kind needed in glosses, and CLAN also includes a MOR[25] program which provides morphological tagging for eleven languages. However, Welsh is not among these languages so far. Furthermore, as pointed out by Donnelly and Deuchar (2011a) MOR requires a separate pass over a file to tag each language, and the output of MOR tags needs to be run through a 'POST' or disambiguation program in order to choose the correct tag where more than one possible tag has been selected. To date POST is only available for five of the eleven languages.

A program to gloss the transcriptions automatically (an "autoglosser") was developed after the publication of *Siarad* in 2009. It was originally used for our Patagonia corpus, and was then applied retrospectively to *Siarad*. Whereas manual glossing is a labour-intensive process, automatic glosses can be produced at about 1,000 per minute using a typical desktop PC. Newer versions of the software (see <autoglosser.org.uk> allow much faster glossing: As we shall describe below, the automatic glossing system has the enormous advantage of not only being faster, but also of making some automatic analysis of the data possible, given the detailed, consistent format used in its glosses.

The development of computer technology and computational linguistics have made it possible to enhance the speed of data transcription and analysis. The limits of automatic speech recognition techniques mean that human transcribers are still needed to listen to recordings and convert them to a written transcription system, but as we shall describe below, computer assistance is increasingly useful in the process of glossing.

---

25. <talkbank.org/morgrams>

We decided to create a new glossing system to handle the *Siarad* files, being guided by the following criteria: (1) open-source software and resources would be used; (2) accessible and easily editable online dictionaries would be used; (3) simultaneous multi-language tagging would be possible; (4) the system would be suitable for use with informal speech involving false starts and repetitions.

In line with the first criterion, the following open-source software was selected for use with the autoglosser:

a. The *PHP* scripting language.[26] This handles the data files and triggers the actions of other programs on them.
b. The *PostgreSQL* database management system.[27] This is where the contents of the data files and the online dictionaries are stored. The accessibility of the online dictionaries in the database satisfies the second criterion, and the method of storage of each word with its language tag (cf. Donnelly & Deuchar 2011a) satisfies the third criterion.
c. A constraint grammar parser which allows the correct choice from among possible multiple glosses found in the online dictionaries. The one chosen was *VISL-CG3*, [28] which was developed at the University of Southern Denmark (cf. Donnelly & Deuchar, 2011b).
d. The *LateX* typesetting system,[29] which provides neat PDF output of the files.

We shall describe how the process of automatic glossing works with reference to an example utterance from the file Stammers4, shown in (1) below:

(1) *ALN: ond # dw i (dd)im actu(ally)@s:eng[?] isio mynd i wrando ar y stuff@s:cym&eng.
%eng: but I don't actually want to go and listen to the stuff

The example in (1) includes the main tier, consisting of the utterance spoken by ALN, and the dependent English translation tier. So far no gloss tier has been added. As can be seen, the words of this utterance are predominantly Welsh, but there is one English word included, *actually*, and a word which could be either Welsh or English, *stuff*.

The first step in the production of automatic glosses for this utterance involves importing the utterance in the main tier into a database table created for that purpose. Table 3.1 gives a simplified representation of the database record for this utterance, with not all fields in the utterances table being shown.

---

26. <php.net>

27. <postgresql.org>

28. <visl.sdu.dk/constraint_grammar.html>

29. <latex-project.org>

Table 3.1 Example (1) in the utterances table

Utterance ID	203
Filename	Stammers4
Speaker	ALN
Surface	ond # dw i (dd)im actu(ally)@s:eng [?] isio mynd i wrando ar y stuff@s:cym&eng.
Gloss	but be.1S.PRES PRON.1S NEG actually want go.NONFIN to listen.NONFIN on DET stuff
English	but I don't actually want to go and listen to the stuff

All utterances are given a unique number for identification, and this appears in the first row of Table 3.1. The filename and name of the speaker follow, and then the field 'Surface' identifies the actual utterance as it appeared in the main tier of the CHAT file. The 'Gloss' field lists the manual glosses that were produced before the automatic glosses as described above, and the 'English' field gives the English translation.

Once the utterance has been stored in the utterances database table, the next step is to import the contents of the 'Surface' field (from the original main tier) into another database table created to hold words. As part of this process non-word symbols specific to CHAT (e.g. the pause symbol '#') are removed and the language tags for each word are stored in a separate field labelled 'Langid' (for 'language ID'). The language ID of each word will determine which electronic dictionary the autoglosser should use to look up the word. Table 3.2 below shows some of the fields in the words table into which the words from the utterance in (1) are imported.

Table 3.2 The words from Example (1) in the words table

Location	**Surface**	**Langid**
1	ond	cym
2	dw	cym
3	i	cym
4	ddim	cym
5	actually	eng
6	isio	cym
7	mynd	cym
8	i	cym
9	wrando	cym
10	ar	cym
11	y	cym
12	stuff	cym&eng
13	.	999

In Table 3.2 the column headed 'Location' indicates the sequential position of each word in the utterance, the column headed 'Surface' indicates the actual word, and the 'Langid' column lists 'cym' for Welsh words, 'eng' for English words, and 'cym&eng' for words which could come from either language. The entry '999' in the 'Langid' column is used to indicate punctuation.

The next step is to look up each entry in the words table in a digital dictionary for the appropriate language, using the language assigned to the word by the transcriber. The dictionary table for each language takes the form of a database table which is based on the Eurfa online dictionary for Welsh (Donnelly, 2016) and on Kevin Atkinson's Moby list, now included in the *12dicts* collection of word lists (Beale, 2016) for English. Table 3.3 shows the layout of the Welsh digital dictionary table.

Table 3.3 Welsh digital dictionary table layout

Surface	Lemma	English	pos	Gender	Number	Tense
bara	bara	bread	N	M	SG	
cathod	cath	cat	N	F	PL	
mynd	mynd	go	V			INFIN
aeth	mynd	go	V		3MS	PAST
hapus	hapus	happy	ADJ			
rhywsut	rhywsut	somehow	ADV			
heb	heb	without	PREP			

The first column (headed 'Surface') shows the word itself, the second column the lemma or basic form on which the word is based, the third column the English translation of the lemma, the fourth column the part-of-speech (POS) tag, the fifth column the word's gender, the sixth column its number and the seventh column its tense if a verb. Table 3.4 shows the layout of the English digital dictionary table. The first two columns are as in the Welsh table, while the third (POS) column gives possible alternative tags which will later be disambiguated. This approach is taken in the English table because of the higher number of homophonous forms in English. For example, the POS *SV* for *walk* means that it can be either a singular noun or a verb.

Table 3.4 English digital dictionary table layout

Surface	Lemma	pos	Number	Tense
walk	walk	SV		INFIN
break	break	SV		INFIN
broke	break	AV		PAST
broken	break	AV		PASTPART
car	car	N	SG	
quick	quick	ADJ		
which		REL		

The look up process involves some modification of each word: for example mutation in Welsh is removed, elisions are expanded and regular verb endings removed in English. Where a word is not found in the digital dictionary, the gloss 'unk' for 'unknown' is generated, leading either to the addition of the word to the digital dictionary where appropriate or the correction of a misspelling. This identification of misspellings is a useful spinoff of the automatic glossing process. All matching entries in the dictionary table are written out to a file which will provide the input to the constraint grammar parser (Didriksen, 2016). Figure 3.2 shows what this would look like for Example (1).

```
"<ond>"
      "ond" {203,1} [cym] conj :but: [201902]
"<dw>"
      "bod" {203,2} [cym] v 1s pres :be: [209292]
"<i>"
      "mi" {203,3} [cym] pron 1s :I: [204322]
      "i" {203,3} [cym] prep :to: [202346]
"<ddim>"
      "dim" {203,4} [cym] n m sg :nothing: [208789] + sm
      "dim" {203,4} [cym] adv :not: [204176] + sm
"<actually>"
      "actual" {203,5} [en] adj :actual: [55812] # adv
"<isio>"
      "eisiau" {203,6} [cym] n m sg :want: [201655]
"<mynd>"
      "mynd" {203,7} [cym] v infin :go: [198448]
"<i>"
      "mi" {203,8} [cym] pron 1s :I: [204322]
      "i" {203,8} [cym] prep :to: [202346]
"<wrando>"
      "gwrando" {203,9} [cym] v 2s imper :listen: [36391] + sm
      "gwrando" {203,9} [cym] v infin :listen: [207781] + sm
"<ar>"
      "ar" {203,10} [cym] prep :on: [204320]
"<y>"
      "fy" {203,11} [cym] adj.poss 1s spoken :my: [209735]
      "y" {203,11} [cym] pron.rel :that: [200652]
      "y" {203,11} [cym] det.def :the: [204199]
"<stuff>"
      "stuff" {203,12} [in] n sg :stuff: [200908]
      "<.>"
```

Figure 3.2 Example (1) formatted for the application of constraint grammar

The use of constraint grammar (cf. Karlsson, 1990, Karlsson, Voutilainen, Heikkilä, & Anttila, 1995) is necessary because some words appearing in the data have more than one entry or more than one POS in the dictionary table. In this case constraint grammar is used to identify the correct POS tag for a specific word in a specific utterance. This will be illustrated with reference to Example (1) above.

In Example (1) the Welsh word *i* appears twice. The word *i* can either be a first-person singular pronoun or a preposition. Two constraint grammar rules[30] are relevant to disambiguating which POS is correct for each word:

a. select ([cym] "mi" pron) if (−1("bod" 1s pres))
b. select ([cym] "i" prep if (−1 v infin)) (not −2 (pron 1s))

Rule (a) says that the pronoun should be chosen if the preceding word is the first-singular present tense of the verb 'to be'. The first occurrence of *i* in Example (2) above is preceded by *dw*, which is the first person singular present tense form of the verb 'to be'. Thus rule (a) applies and *i* is glossed as a first person singular pronoun (I.PRON.1S).

Rule (b) says that the preposition should be chosen in utterances where the preceding word is an infinitive and the word before that is not a pronoun. This applies to the second occurrence of *i* in *isio wrando i*, resulting in the glossing of the second *i* as a preposition (PREP).

The choice between other alternative POS tags shown in Figure 3.2 is made in a similar way. After looking up the words in the appropriate digital dictionary and disambiguating the POS tags as necessary, the words from the utterance are stored in a words table with their correct POS tag or automatic gloss. This is shown in Table 3.5.

Table 3.5 Words table with disambiguated glosses

Location	Surface	Autogloss	Langid
1	ond	but.CONJ	cym
2	dw	be.V.1S.PRES	cym
3	i	I.PRON.1S	cym
4	ddim	not.ADV + SM	cym
5	actually	actual.ADJ + ADV	eng
6	isio	want.N.M.SG	cym
7	mynd	go.V.INFIN	cym
8	i	to.PREP	cym
9	wrando	listen.V.INFIN + SM	cym
10	ar	on.PREP	cym
11	y	the.DET.DEF	cym
12	stuff	stuff.N.SG	cym&eng

30. For further information about the constraint grammar rules used by the Bangor Autoglosser for Welsh, see <github.com/donnekgit/autoglosser/blob/master/grammar/en_es_grammar>, and for a tutorial on how to use constraint grammar, see <kevindonnelly.org.uk/2010/05/constraint-grammar-tutorial>

The autoglosser then writes the contents of the words table out to create a new autogloss tier generated from the glossed words. This new tier is used to create a new version of the CHAT file. Example (1) with the new autogloss tier is shown in Figure 3.3.

ALN: ond # dw i (dd)im actu(ally)@s:eng [?] isio mynd i wrando ar y stuff@s:cym&eng.
%aut: but.CONJ be.V.1S.PRES I.PRON.1S not.ADV + SM actual.ADJ + ADV want.N.M.SG go.V.INFIN to.PREP listen.V.INFIN + SM on.PREP the.DET.DEF stuff.N.SG
%eng: but I don't actually want to go and listen to the stuff.

Figure 3.3 Example (1) with the new autogloss tier

Figure 3.3 shows a format approximating that used in CHAT, but the autoglosser can provide a more readable pdf version of the CHAT file with the glosses aligned with the words to which they correspond. This is illustrated in Figure 3.4.

aln :	**ond**	**dw**	**i**	**ddim**	**actually**^{E}
aut:	*but.conj*	*be.v.1s.pres*	*i.pron.1s*	*not.adv + sm*	*actual.adj + adv*
	isio	**mynd**	**i**	**wrando**	**ar**
	want.n.m.sg	*go.v.infin*	*to.prep*	*listen.v.infin + sm*	*on.prep*
	y	**stuff**$^{C}_{E}$			
	the.det.def	*stuff.n.sg*			

but I don't actually want to go and listen to the stuff.

Figure 3.4 A more readable version of Figure 3.3

Use of automatic glosses for analysis

The automatic glosses in the '%aut' tier have another function in addition to providing comprehensive information about the meaning and grammatical features of the words in the main tier: they allow automatic analysis to be performed on the data. For example, in an analysis to be reported in Chapter 6 on the external factors influencing intraclausal code-switching, it was possible to use the automatic glosses indicating finite verbs to identify clause boundaries and extract all finite clauses. It was also possible to use the language markers to identify monolingual vs. bilingual clauses and thus identify the proportion of bilingual clauses in the speech of each person. This meant that much larger amounts of data could be processed than in the past.

The automatic glossing system is calculated to be 98% accurate. Using a sample of two *Siarad* files (Stammers7 and Stammers9) Donnelly, Cooper, & Deuchar (2011) compared the coverage and accuracy of both the manual and the automatic

glosses. They reported that the autoglosser had 98.3% coverage and 97.9% accuracy. While the error rate was not quite as good as the manual rate of 99.9% for both coverage and accuracy, it is nevertheless impressive. Furthermore, the vastly superior amount of data that can be covered in an automatic vs. a manual analysis has to be balanced against this slight loss of accuracy.

For more information about the use of the automatic glossing system for transcription and analysis, see Carter, Broersma, Donnelly, & Konopka (2017).

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 <talkbank.org/manuals/CLAN.pdf>). Figure 3.5 shows a screenshot of a transcript including the sound wave of a highlighted utterance.

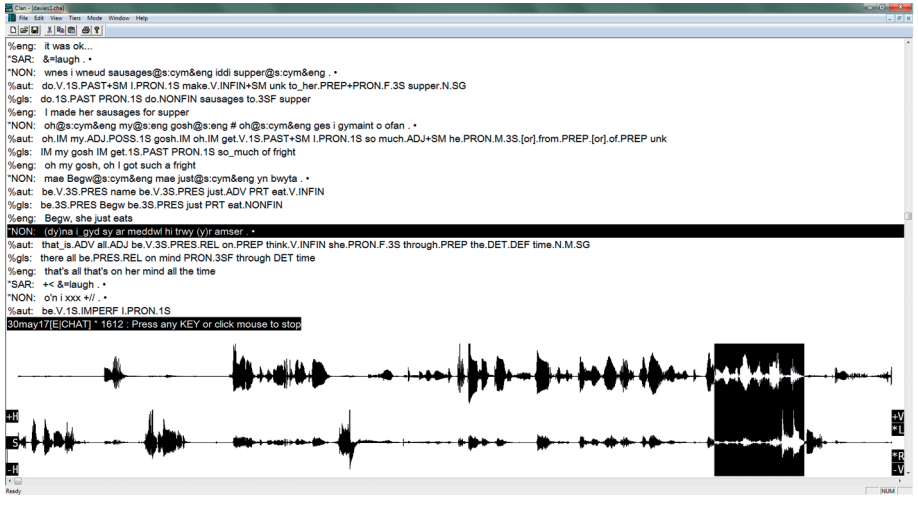


Figure 3.5 Screenshot of a transcript fragment

Transcription reliability

Numerous transcribers worked on the data transcription process over the course of several years. While 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.

While 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-transcriber reliability 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. We used two sample transcripts to measure the extent of their agreement by a new method of using plagiarism software. The two independent transcripts were submitted as separate documents to Turnitin (<turnitin.com>), a commercial plagiarism detection service. This 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 score for the *Siarad* corpus was 74%. This score takes into account detailed comparison of the main, manual gloss ('%gls') and translation tiers of the transcription. We found that many of the differences were 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 a useful research tool for strengthening inter-transcriber reliability when building corpora. Before finalising the transcriptions, other checks were also 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 the transcripts and recordings

The *Siarad* transcripts were contributed to Talkbank in 2009 and can be found at <talkbank.org/data/BilingBank/Bangor/>. The transcripts were updated in 2010 following a change in the CHAT language marking system. Since 2010 they have also been available along with other Bangor corpora on our own website <bangortalk.org.uk>. This website has the advantage of providing access not only to the CHAT files, but in addition to a more user-friendly version of the transcripts which are easier to read for the uninitiated. BangorTalk also hosts the documentation file for *Siarad* as indicated above, and other information including licensing, tools, summary data and links to publications. It is hoped that the *Siarad* data will also be available in the UK data archive (see <data-archive.ac.uk>) in the future.

Summary of Chapter 3

In this chapter we have described the process involved in the collection and transcription of our data, which are publicly available under an open licence. We have dealt with the question of the quantity and type of data, the recruitment and recording of participants, ethical considerations, the transcription of the data using CHAT and our language marking system, our system of manual as well as automatic glossing, and the advantages of the automatic glossing. We hope that this information will be useful to others who wish to build corpora, as well as to those who want to evaluate our findings. In Part Two of this book we shall present our findings from various analyses of the corpus.

CHAPTER 4

Code-switching vs. borrowing

New implications arising from our data

Introduction

This chapter will deal with the issue of how to distinguish between switches and borrowings in insertions from one language into another. While most linguists agree that there are clearcut examples of switches vs. borrowings, they disagree on whether all other-language items can be unequivocally assigned to one category vs. the other, and whether each category has specific defining characteristics. In English, the word *restaurant* is a borrowing from French, and few would argue that it is normally used as a switch into French, since many speakers of English are monolingual and unable to speak French. However, if used by a French-English bilingual speaker there might be some doubt as to its status. This kind of problem is particularly evident in dealing with Welsh-English data, because all Welsh speakers also speak English. We can illustrate the issue with reference to the bolded English words in Example (1) from our data:

(1) pan dach chi 'n defnyddio **wide-angle lenses**
when be.2PL.PRES PRON.2PL PRT use.NONFIN wide-angle lenses
dach chi 'n **emphasize**-io **'r**
be.2PL.PRES PRON.2PL PRT emphasize-VBZ DET
foreground. [Fusser17-BEN]
foreground
"When you use wide-angle lenses, you emphasize the foreground."

The longer the stretch of other-language material (here English is the 'other language'), the easier it generally is to identify that material as a switch. So in relation to (1) we may readily say that the phrase *wide-angle lenses* is a switch into English, but we might have more doubts about the status of the single words *emphasize* and *foreground*. The vocabulary of Welsh includes many well established English borrowings, so one might wonder whether *foreground* is one of these (although it is not in fact). As for *emphasize*, some investigators would consider it to be a borrowing because of the *–io* suffix, although we would not take this position as we shall explain below. In our discussion we shall focus particularly on single-word insertions, since their status is particularly controversial. The reason for our focus is not only that it is a theoretical issue but also a practical one for those seeking to

delineate the field of code-switching. If borrowings belong to a category that can be distinguished from switches, then any theory of single-word other-language insertions code-switching must clearly delineate the differences between the two categories. If on the other hand we find that there is a continuum between switches and borrowings, then any theory of single-word other-language insertions must allow for such a continuum.

Weinreich (1953, p. 11) was one of the first to draw a distinction between code-switching and borrowing, though using different terminology. He uses the term 'interference' to describe a situation where a bilingual is speaking language X but produces an 'on-the-spot borrowing' or 'nonce borrowing' from language Y (which we refer to here as 'code-switching'). This is different, he argues, from a situation where a bilingual produces a word from Y which is an established element in language X (which we refer to here as 'borrowing'). Pfaff (1979, p. 295) reports that most investigators find it important to make a distinction between code-switching and borrowing because "the two terms are usually construed as making vastly different claims about the competence of the individual speaker". Code-switching requires knowledge of two languages, she says, whereas borrowing does not. However, she reports "little agreement as to how the distinction is to be made" (Pfaff, 1979, p. 296), going on to discuss the processes undergone by English words when they are used in conversation by Spanish-English bilinguals.

In her pioneering article on Spanish-English code-switching, Poplack (1980) identifies code-switches as English words in Spanish which are not completely integrated phonologically, morphologically and syntactically. Her examples of switches include multiword switches which are not integrated at all, as well as switches which are integrated either syntactically (by following the word order of the host language) or phonologically (which applied to English words pronounced with a Spanish 'accent'.) However, words which show full phonological, morphological and syntactic integration do not qualify as switches but as borrowings into Spanish. Poplack's (1980) criteria for identifying borrowings become less restrictive in her later publications, and in her preface to a reprint of her 1980 article she states that "many of the lone other-language forms operationally classified as code-switches in *Sometimes I'll start a sentence…* would today be identified as borrowings" (Poplack, 2000, p. 222). In this preface, she also states that she no longer considers the phonological criterion to be reliable because phonological integration is in fact highly variable.

Poplack & Sankoff (1984) is a sophisticated study of the process of linguistic and social integration of English loanwords into Puerto Rican Spanish. They recognise the importance of the role of frequency, something our own study of *Siarad* will show, stating that "an important diagnostic for the incorporation of a form into the native lexicon is the increased frequency of its usage. Even the degree to which the loanword is linguistically integrated has been attributed to the frequency of its

use within the community" (Poplack & Sankoff, 1984, pp. 101–2). They report that frequency of use has been used as a criterion for identifying borrowings, along with native-language synonym displacement, linguistic integration and acceptability. They argue that it is "reasonable to assume that as a borrowed word is more and more used, it tends to become phonologically and morphologically integrated, to displace competing recipient language forms, and at least eventually, to be accepted by its native speakers" (Poplack & Sankoff 1984, p. 105). Their results support this assumption in that they discovered a correlation between frequency of use and phonological and morphological integration. They also recognise in this paper that integration is not categorical but that "The position of a borrowed element within a host language system….is a matter of degree: it may be completely integrated, partially integrated, or not integrated at all" (Poplack & Sankoff 1984, p. 106).

The study by Poplack, Sankoff, & Miller (1988) also investigates the relation between frequency of use and integration. Here a measure of frequency is used to distinguish four categories of 20,000 English words produced in mainly French discourse in Canada. The English words are categorised by how many times they are produced and/or by how many speakers. In their study their focus is on two main categories: 'nonce borrowings', which are other-language (English) words used once by only one speaker, and 'widespread' words which are used by more than ten speakers. They find that morphological and syntactic integration of these items into the recipient language occurs at a high level for both types of other-language word, and that "integration on the phonological level is… the only clear linguistic differentiator of nonce from widespread borrowing…" (Poplack et al., 1988, p. 95). With hindsight, this paper probably heralds the introduction of the nonce borrowing hypothesis by Sankoff, Poplack, & Vanniarajan (1990), where the role of frequency is downplayed, and the authors argue that there is "no difference between nonce borrowings and established loans… with respect to their morphological and syntactic integration into host language contexts" (Sankoff, Poplack, & Vanniarajan, 1990, p. 94).

They compare lone English-origin nouns with Tamil nouns when used as the objects of Tamil verbs, concluding that the English-origin nouns "behave morphologically and syntactically exactly as do established borrowings and native Tamil forms" Sankoff et al. (1990, p. 95). (It should be noted, however, that according to Stammers (2010, p. 27) differences in the case-marking of English-origin as compared to Tamil nouns suggest that some of the English-origin nouns are not fully integrated.) In the Sankoff et al. (1990) paper Poplack thus moves away from her earlier position that linguistic integration can be a matter of degree and that frequency is a relevant factor. Stammers & Deuchar (2012) and Deuchar & Stammers (2012) discuss the nonce borrowing hypothesis at length, arguing that it denies a role for frequency in the integration of borrowings, and yet frequency, as we demonstrate below, is key in influencing linguistic integration.

Sankoff et al. (1990)'s method of comparing other-language with recipient-language words in order to identify borrowings is further developed in a special journal issue edited by Poplack & Meechan (1998). The studies in the special issue are summarised by Stammers (2010, p. 26) and mostly focus on assessing the linguistic integration of lone English nouns in another recipient language by comparing them with recipient-language nouns. The most common conclusion is that the English words are borrowings on the grounds that they are generally integrated into the recipient language. However, little attention is paid to degrees of integration or to the role of frequency in integration, unlike in the earlier work by Poplack and colleagues in the 1980s.

While Poplack & Meechan thus consider most lone other-language words to be borrowings, multiword other-language insertions are considered to be clear examples of code-switches. They argue that within multiword switches the grammar of the donor language is preserved, whereas this is not the case with the insertion of lone other-language words. This argument is later repeated in Poplack & Dion (2012) as we shall see below, where we shall consider some counter-evidence from *Siarad* as well from Poplack's own work.

Myers-Scotton (1992, p. 20) differs from Poplack & Meechan's view of a clear distinction between the categories of code-switching and borrowing. Instead she argues that a "continuum of relationships exists between borrowing and all forms of code-switching material so that codeswitching and borrowing are not distinct phenomena, as some have suggested". Myers-Scotton recognises, like Poplack and colleagues, that lone other-language items are often linguistically integrated by the recipient language, but she accounts for this integration with her Matrix Language Frame (MLF) model, according to which both borrowed and code-switched forms from the non-matrix or 'embedded' language are expected to be morphosyntactically integrated into the matrix language. Rather than use linguistic integration as a criterion to distinguish borrowings from switches as Poplack and colleagues do, she suggests that frequency or predictability is a better criterion. She argues that "most researchers agree that borrowed forms and codeswitching forms differ in regard to predictability…a borrowed form will reoccur…because it has a status in the recipient language…The codeswitching form may or may not reoccur; it has no predictive value" (Myers-Scotton, 2002, p. 41). Thus, switches are defined negatively as those forms which occur with very low frequency – in another study she uses the admittedly arbitrary criterion of three occurrences: "any form occurring at least three times in a relatively large corpus is a B[orrowed] form" (Myers-Scotton, 1997, p. 207).

Jones (2005) adopts this criterion in her study of code-switching between Jersey Norman French and English, and finds that borrowings are less likely to be 'flagged' (by hesitation, metalinguistic commentary or self-correction) than switches, thus providing some evidence for a difference between the two categories. Using the frequency

criterion in this way does not assume a categorical difference between switches and borrowings but allows objective investigation of any differences that there might be.

However, Poplack (2012, p. 644) considers the "'three-occurrence' metric for the loanword/code-switch divide" as used by Jones and Myers-Scotton to be arbitrary, and argues that her own earlier studies have instead used "objective quantitative measures of frequency and diffusion". As indicated above, frequency was certainly important in her studies published in the 1980s, and it is important again in the study by Poplack & Dion (2012), where they test the assumption that English-origin items in Quebec French increase in frequency over time. One reason for testing this assumption was the general belief (attributed to others) that other-language items usually originate as code-switches, "but by virtue of being repeated often and widely enough, gradually assume more and more characteristics of the recipient language until they eventually become indistinguishable from it – bona fide *loanwords*" (Poplack & Dion 2012, p. 279). Poplack & Dion find that very few other-language items in fact become widespread, and that they are overwhelmingly ephemeral. Furthermore, they find that linguistic integration (by e.g. French verb inflection, plural marking and gender consistency) is abrupt rather than gradual in that it occurs at the first mention of the English-origin item in Quebec French.

As in work published by Poplack in the 1990s, Poplack & Dion (2012) consider all other-language items which are linguistically integrated to be borrowings, while items appearing as part of multi-word stretches in English without integration into French are considered to be unambiguous code-switches. Although Poplack & Dion allow for the possibility that lone other-language items could be unintegrated and therefore code-switches, they find only one example out of 601 which they consider to be an unintegrated code-switch because it retains its English plural marking. Thus Poplack & Dion (2012, p. 296) conclude that "lone other-language items tend to be borrowed". The reason for this generalisation is that they consider all items which show linguistic integration into the recipient language to be borrowings. They argue that linguistic integration is avoided in phrasal code-switches, since, "within multiword fragments of English only English grammar is operative" (Poplack & Dion, 2012, p. 299). They demonstrate this with the adjective-noun order of their example *free-lance politician* (Poplack & Dion, 2012, p. 302) which is inserted in an otherwise French utterance. This is contrasted with the French noun-adjective order which is apparent in the phrase *deux records anglais* ('two English records') (Poplack & Dion, 2012, p. 303) where *records* is a lone English insertion and classified as a borrowing.

Although multiword other-language insertions often do follow the grammar of the donor language, this is not always the case. Indeed Myers-Scotton (2002, p. 139) cites the example of the multiword English sequence *building high-rise* ('high-rise building') where the word order is French rather than English and which is found in Poplack's own data (Poplack 1987, p. 59). This is reproduced as Example (2) below.

(2) à côté il y en a un autre gros *building high-rise*.
"Next-door there's another big high-rise building."

In this case, instead of a multiword insertion *high-rise building*, following English adjective-noun order, the two English words, *building* and *high-rise*, have been inserted following French noun-adjective order. We have a similar example in *Siarad*, shown in the utterance in (3), where a speaker inserts the English words *massive* and *list* in noun-adjective order, following Welsh rather than English grammar:

(3) a fel like y **list massive** **'ma** [Fusser27-LIS]
and like like DET list massive here
"And like, like, this massive list."

Examples like *building high-rise* and *list massive* are in fact ignored in Poplack & Dion's (2012) bipartite distinction between lone other-language items and multiword code-switches, defined as "fragments drawn unaltered from a donor language and…. therefore governed by the grammar of that language" (Poplack & Dion, 2012, p. 298). The bipartite distinction is again assumed in their proposal as to how the process of code-switching or borrowing works. They suggest that "when speakers go to access an other-language item, they make an instantaneous decision about how to treat it. They may opt to *borrow* it, in which case they assign it all the appropriate recipient-language grammatical structure…The other alternative is to simply leave the other-language item as is…incorporating it *along with* its associated grammatical properties, a process to which we have been referring as code-switching". This suggests that Poplack & Dion's notion of borrowing goes beyond the selection of a lexical item to include the process of linguistic integration. However, for Myers-Scotton, borrowings and switches are both items to be found in the mental lexicon, the main difference being in the way they are tagged: "Presumably, lemmas underlying codeswitching forms are only tagged for the Embedded Language, while borrowed forms have lemmas tagged for *both* the donor and the recipient language, at least in the mental lexicon of those languages" (Myers-Scotton 2002, p. 153). Despite these differences of tagging, Myers-Scotton expects the process of linguistic integration to apply to both types of form, not only to borrowings as Poplack assumes.

Poplack and Myers-Scotton differ not only in how they define borrowing, but also in the hypotheses they propose. Having defined borrowings as linguistically integrated items, Poplack & Meechan (1998) predict that it will be more common for other-language items to be integrated than not, and thus that lone borrowings will be more frequent than lone switches in the data. For Myers-Scotton, however, frequency is part of the definition of borrowings rather than being a hypothesis to test. And just as Poplack & Meechan's hypothesis is Myers-Scotton's definition,

the reverse is also true. Myers-Scotton (1993) hypothesises that borrowings will have a higher degree of linguistic integration than switches, but this is Poplack & Meechan's definition of borrowing in the first place. Myers-Scotton's hypothesis about borrowing in fact distinguishes between 'central' and 'peripheral' morphological marking, (Myers-Scotton, 1993: 183) suggesting that while both code-switches and borrowings will show 'central' integration, borrowings will show more 'peripheral' integration. While central integration relates to morphosyntactic characteristics of the matrix language, peripheral integration includes morphological marking with a lighter functional load. An example of this would be plural marking in Shona data from Zimbabwe. This was analysed by Bernsten (1990), who investigated the tendency for foreign plural nouns to be prefixed with *ma-*. This had the effect of integrating them morphologically into the Shona plural noun class 6. Distinguishing between switches and borrowings on grounds of frequency, Bernsten found that switches from English that were prefixed with *ma-* as a plural indicator tended also to have English *-s* plural marking as well, whereas English borrowings had *ma-* alone to mark plural. This led her to conclude that switches into English were less well integrated into Shona than borrowings from English.

The inverse relationship between Poplack & Meechan's and Myers-Scotton's definitions and hypotheses is illustrated in Table 4.1 below. Comparing the italicised text in the table with the text in normal font will show how Poplack & Meechan's hypothesis is a definition for Myers-Scotton, and *vice versa*.

Table 4.1 Definitions vs. hypotheses regarding code-switching vs. borrowing: Poplack & Meechan compared with Myers-Scotton

Lone other-language items	Definitions of borrowing/ switches	Hypotheses
Poplack & Meechan	Linguistically integrated/not integrated	*Borrowings more frequent than switches*
Myers-Scotton	*More frequent/less frequent*	Borrowings more integrated than switches

(Based on Table 1, Deuchar & Stammers, 2016, p. 7)

Although Poplack's and Myers-Scotton's definitions of borrowing are different, we can see from Table 4.1 that the two scholars are both interested in both frequency and linguistic integration. We shall show how a study of the relation between frequency and linguistic integration in the *Siarad* data can help to adjudicate between their contrasting positions.

Stammers (2010, p. 24) summarises the contrasting positions which we shall consider in this chapter in the two quotations reproduced below:

a. "code-switching and borrowing are two distinct phenomena" (Poplack & Meechan, 1998, p. 132)
b. "code-switching and borrowing fall on a continuum" (Myers-Scotton, 1993, p. 176).

Our aim in this chapter will be to show how work using *Siarad* has contributed to a resolution of this debate.

Stammers (2010) focused on English-origin verbs which are productively inserted into Welsh conversation with the addition of an *–io* or *–o* verbalising suffix. Examples of these verbs as found in *Siarad* are given in his Table 6 (Stammers, 2010, p. 77) and include examples like *protestio*, *stopio*, *panic-io* and *cope-io*. For Poplack the addition of the verbalising suffix to these English verbs may well have been sufficient to class them as borrowings into Welsh, but from the point of view of Myers-Scotton one might interpret the use of the verbalising suffix as integration into the matrix language of the clause, i.e Welsh. For Myers-Scotton, this integration process would apply to all English words, whether borrowings or switches.

Stammers studied the linguistic integration of these English verbs using various criteria: the addition of the verbalising suffix, the use of the verbs in periphrastic vs. synthetic constructions and the application of the process of soft mutation where possible. As part of this analysis he noted whether or not each item was listed in the Welsh dictionary (*protestio* and *stopio* are listed whereas *panic-io* and *cope-io* are not) but he made no assumptions in advance regarding which items were switches and which were borrowings.

Regarding morphological integration of the verb by affixation, Stammers (2010, p. 78) found that 97.3% "were clearly morphologically integrated by means of a Welsh derivational verbal suffix (*-io* or *–o*). As he states, "From the perspective of Poplack and her associates, this fact alone would presumably be sufficient for them all to be counted as borrowings" (Stammers, 2010, p. 80). However, for Myers-Scotton he points out that the verbal suffix would be classified as an 'early system morpheme'[32] belonging to the matrix language, Welsh. Its presence would be expected on all English-origin verbs inserted into Welsh, and would thus not help identify borrowings vs. switches.

Stammers (2010) went on to use another measure of integration, this time syntactic, and investigated the participation of English-origin verbs in periphrastic vs. synthetic verbal constructions. This time he compared the integration

32. Early system morphemes are those like determiners and plural morphemes which "include the relevant *phi*-features of person, number and gender in relevant languages" (Myers-Scotton 2002: 75). They usually belong to the matrix language. For more information on the Matrix Language Frame (MLF) model see Chapter 5.

of English-origin verbs listed in the Welsh dictionary with unlisted English-origin verbs, and compared both with native Welsh verbs as a basis of comparison. His results showed that whereas 12.6% of the tokens[33] of native Welsh verbs appeared in synthetic constructions, "*none of the English-origin tokens* were found in synthetic constructions *whether or not* they were listed in a dictionary" (Stammers 2010, p. 84). He observes that if syntactic integration were the sole criterion for identifying borrowings in a Poplack-style approach, then English-origin verbs would be identified as switches rather than borrowings, whereas using the morphological integration criterion as outlined above they would be identified as borrowings rather than switches.

Given this puzzling situation, Stammers decided to use a morphosyntactic measure of integration, i.e. soft mutation. This process is outlined by Stammers (2010, p. 89) and also by Stammers & Deuchar (2012, p. 638) who include the reproduced below as Table 4.2, showing the changes that are undergone in word-initial consonants when they are subject to soft mutation:

Table 4.2 Initial consonant changes in soft mutation in Welsh

Initial consonant (phonetic)	p	t	k	b	d	ɬ	r̥	m	g
Initial consonant (orthographic)	**p**	**t**	**c**	**b**	**d**	**ll**	**rh**	**m**	**g**
	↓	↓	↓	↓	↓	↓	↓	↓	↓
Mutates to (phonetic)	b	d	g	v	ð	l	r	v	(dropped)
Mutates to (orthographic)	**b**	**d**	**g**	**f**	**dd**	**l**	**r**	**f**	

(Adapted from Stammers & Deuchar, 2012, p. 638, Table 4)

Stammers & Deuchar (2012, p. 638) distinguish between lexically and syntactically triggered mutation. They state that "In lexically triggered soft mutation, the non-finite verb is directly preceded by a preposition, clitic or other particle causing soft mutation". They give an example (reproduced as Example (4) below) where the preverbal particle *i* (translated as 'to' in English) triggers soft mutation of the initial consonant of the following verb *costi*, which becomes *gosti*.

(4) *well* mae mynd i gosti pres. [Fusser6-Ant]
well be.3s.PRES go.NONFIN to cost.NONFIN money
"Well, it's going to cost money."

Other environments involving lexically triggered mutation include where the non-finite verb is preceded by a second person or third person masculine possessive pronoun, the preposition *am* ('for/about'), *ar* ('on/about to'), *gan* ('by/while/with').

33. Stammers (2010: 101–102) reports that almost all tokens came from the 11 most frequently occurring verb types, many of which were irregular.

According to Stammers & Deuchar (2012), in syntactically triggered soft mutation[34] the non-finite verb is expected to mutate because it follows the grammatical subject (cf. King 2016, p. 14). This is illustrated in Example (5) below, where *drio* is a soft-mutated version of *trio* ('try').

(5) wnest ti drio? [Stammers5-Rho]
do.2S.PAST PRON.2S try.NONFIN
"Did you try?"

Stammers & Deuchar provide additional examples of environments for mutation, including the one in Example (6) below where mutation does not apply as expected:

(6) well i ti bwco diwrnod *off*. [Robert6-Eir]
better to PRON.2S book.NONFIN day off
"You'd better book a day off."

In this example they point out that *bwco* 'to book' is "uttered in its unmutated form, whereas the mutated form *fwco* [vʊko] would have been expected in this environment" (Stammers & Deuchar, 2012, p. 639).

Although soft mutation is the most robust mutation type in Welsh (cf. Comrie, 2000, p. 81), Ball (1988, p. 72) reported that it did not always apply where expected. This variation is advantageous for our analysis since it provided us with a fine-grained measure to compare the integration of English-origin verbs with the level of mutation found in native Welsh verbs, following the methodology advocated by Poplack & Meechan (1998).

In comparing the application of mutation to English-origin verbs with Welsh we distinguished between two categories of English-origin verbs as in Stammers' (2010) analysis of syntactic integration: those listed in a dictionary of Welsh and those not listed there. This was in order to determine whether, in addition to frequency, there is a factor of 'listedness[35]' (cf. Muysken, 2000, p. 71) which impacts on linguistic integration. Muysken (2000) suggests that the dimension of 'listedness' is important in distinguishing switches from borrowings. He says that "listedness refers to the degree to which a particular element or structure is part of a memorized list which has gained acceptance within a particular speech community" (Muysken 2000, p. 71). Listedness is thus a property of words in the mental lexicon rather than what is concretely in the community's dictionary, but the community dictionary may be expected to be a guide to listedness.

---

34. There has been considerable controversy about the best account of syntactically triggered soft mutation: for a useful summary see Borsley, Tallerman, & Willis (2007), Chapter 7.

35. i.e. being 'listed' in the mental lexicon: see later in this chapter for more discussion.

Stammers' analysis of mutation in verbs involved extracting all of the non-finite verb tokens found in the *Siarad* corpus that (i) ended in the *-(i)o* verbalising suffix; (ii) began with a consonant susceptible to soft mutation (subject to certain exclusions); and (iii) occurred in an environment where soft mutation could be expected to apply. For more details see Stammers (2010, 107–109). Each of a total of 506 tokens[36] was classified according to whether or not mutation actually applied. The application vs. non-application of mutation is shown in Figure 4.1.

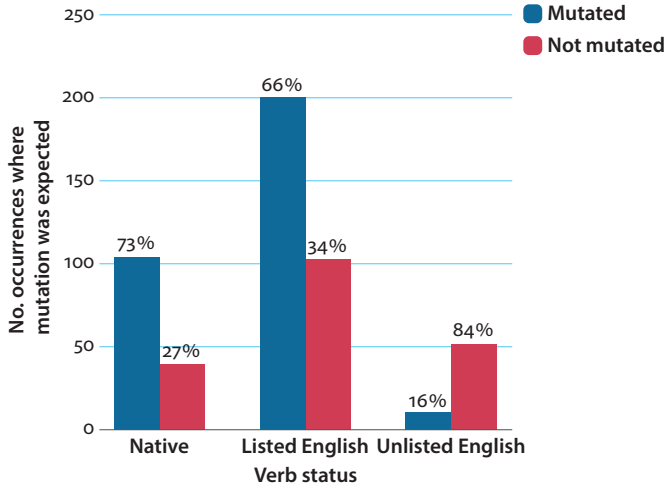


Figure 4.1 Results of analysis of soft mutation based on verb status (Based on Figure 1 in Stammers & Deuchar, 2012, p. 639)

If we compare the light ('mutated') with the dark ('not mutated') bar for each of the three verb categories in Figure 4.1 we can see that the native and listed English-origin verbs behave fairly similarly to one another, whereas the unlisted English-origin verbs behave differently. For the native and listed verbs, mutation applies more often than not, whereas the reverse is the case for the unlisted verbs. What is clear from the results is that linguistic integration as measured by soft mutation is by no means abrupt, in contrast to Poplack & Dion's finding regarding the morphosyntactic integration of English items into their Quebec French corpus. Instead, as Stammers (2010, p. 98) argues, there appears to be a continuum of integration.

36. It would have been ideal to compare the mutation rate of individual verbs, but few had more than ten tokens, and many only one token. The raw data and can nevertheless be viewed in Appendix 7 of Stammers (2010). A comparison of mutation in individual verbs awaits a larger corpus.

What is not clear from Figure 4.1, as Stammers & Deuchar (2012, p. 639) point out, is "whether the difference between the behaviour of the unlisted English and the other verbs can be explained by their unlisted status …or by a factor such as frequency". So, the next task was to investigate the role of frequency and to conduct a frequency analysis of all the verbs represented in Figure 4.1. The results (Figure 4.2) which is reproduced from Figure 2 in Stammers & Deuchar (2012, p. 640) show a clear relationship between the frequency of the verbs in the corpus and the rate of mutation in environments where this could be expected to occur. This relationship is log-linear or logarithmic rather than linear, which is why the verbs are grouped into 1–9, 10–99, 100–999 and 1000–9999 per million[37] words (Stammers & Deuchar, 2012, p. 640). This parallels the methods used in psycholinguistic studies involving word frequency (see e.g. Lewis, Gerhand, & Ellis, 2001). Using the logarithmic values, Stammers & Deuchar report a correlation of 0.99 between rate of mutation and frequency (see also Stammers, 2010, p. 105).

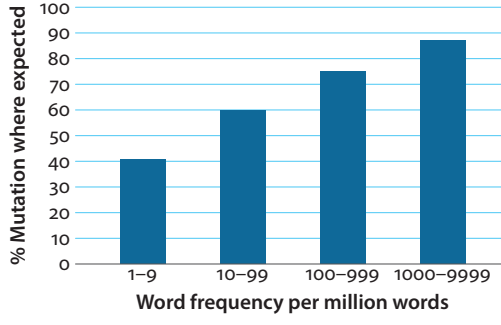


Figure 4.2 Results for mutation rate by word frequency groupings (Based on Figure 12 in Stammers & Deuchar, 2012, p. 640)

Figure 4.3 shows the relation between frequency and each of the verb categories.

As Stammers & Deuchar point out, the native Welsh and listed English verbs are shown to behave quite similarly with respect to the rate of mutation: as the frequency of the verb increases, so does its rate of mutation where expected. The unlisted verbs also show a clear relation between frequency and mutation, but they are different from the other categories of verbs in two ways: (1) they do not occur at all at a frequency of more than 99 per million words; (2) the rate of mutation for the two categories of frequency which exist (1–9 and 10–99 per million) is lower than for the other categories of verbs. As Stammers & Deuchar (2012: 642) state, "This suggests that not only frequency but also 'listedness' is a factor". Although the nature of the bilingual lexicon is still unclear (cf. Kroll, Sumutka, & Schwartz,

37. See Stammers (2010: 105) for explanation of the methodology.

2005) we know that there is cross-language competition but that language-specific selection is possible when bilinguals choose to speak just one language. So we assume that words are 'tagged' as either Welsh or English, or possibly both. What we know little about is how an item which is 'listed' for language A also becomes 'listed' for language B in the individual's lexicon.

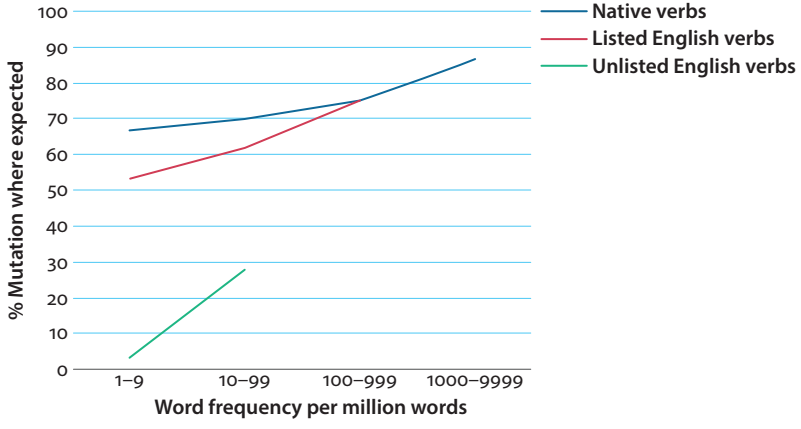


Figure 4.3 Rate of mutation for verbs in the three groups across frequency bands (Based on Table 13 in Stammers, 2010, p. 109)

Figure 4.3 suggests that there is indeed a category of English-origin verbs which are not only not listed in the Welsh dictionary but also not 'listed' in bilingual speakers' lexicon for Welsh. The evidence is that they behave differently with regard to integration in Welsh in the form of soft mutation. The role of listedness can be made even clearer by conflating the categories of listed Welsh ('native') and listed English-origin verbs and comparing their rate of mutation with that of unlisted English-origin verbs at two levels of frequency. The results are shown in Figure 4.4.

Figure 4.4 shows a similar relation between frequency and the application of mutation for both categories of verbs, but also shows that the frequency of mutation is lower for unlisted verbs. For further discussion see Deuchar & Stammers 2016. These results go against the idea that there is a continuum from code-switching to borrowing. This is because listed verbs of the same (low) frequency as unlisted verbs receive a higher level of integration than the unlisted verbs. These unlisted verbs would seem to be good candidates for identifying as code-switches, and this incidentally gives support for our transcription system which marks unlisted words as belonging to English rather than Welsh. Code-switches in our data thus appear to be distinguishable from borrowings on the grounds of low level of both frequency and integration. Borrowings may or may not occur with low frequency but will have levels of integration comparable to those of recipient- or host-language items. Our conclusions are summarised in Table 4.3 below.

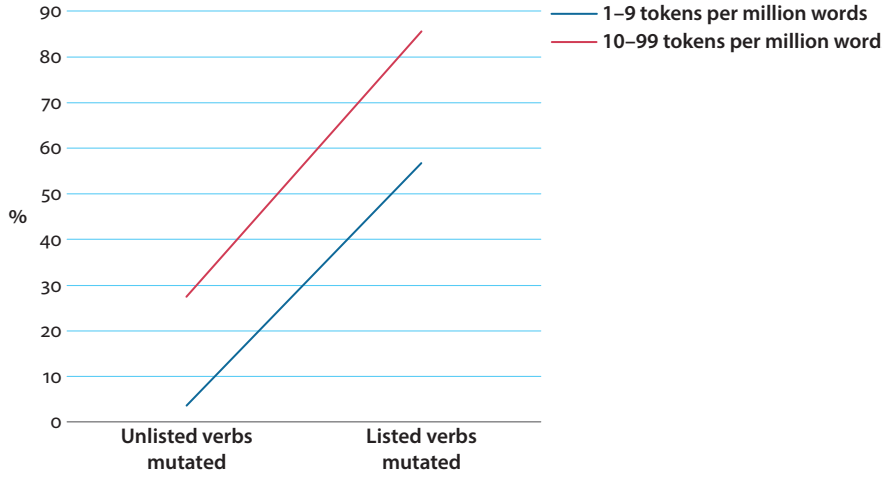


Figure 4.4 Application of mutation to listed and unlisted verbs at two levels of frequency (Based on Figure 5, Deuchar & Stammers, 2016, p. 13)

Table 4.3 Role of integration and frequency in identifying code-switches and borrowings

Donor-language items	Frequency	Integration
Code-switches	Low	Low
Borrowings	Variable	Native-like

Stammers & Deuchar (2012) also use the results represented in Figure 4.3 to argue against the nonce borrowing hypothesis (NBH) originally formulated by Sankoff, Poplack, & Vanniarajan (1990, p. 94) as predicting "no difference between nonce borrowings and established loans….with respect to their morphological and syntactic integration into host language contexts". Nonce borrowings as defined by Sankoff et al. as other-language words which are produced by bilinguals but which "differ from established loanwords in that they are not necessarily recurrent, widespread or recognized by host language monolinguals" (Sankoff et al., 1990, p. 71). On the basis of this definition, Stammers & Deuchar (2012, p. 632) reformulate the NBH in a form which is empirically testable: "There is no difference between frequent and infrequent donor-language items in terms of their degree of integration". As Stammers & Deuchar argue, the results presented in Figure 4.1 already provide evidence against the NBH. This is because the unlisted verbs in this Figure would be the best candidates for nonce borrowing status, but in order to achieve it they would need to show similar rates of the application of mutation to those found in listed and Welsh verbs. However, as we have seen, this is clearly not the case. In considering whether the difference between the behaviour of the unlisted English

and the other verbs could be explained by their unlisted status or by a factor such as frequency, Stammers & Deuchar did find a clear effect of frequency but also of their unlisted status. They thus argue that "the category of nonce borrowings is redundant" (Stammers & Deuchar, 2012, p. 643).

In a reply to Stammers & Deuchar, Poplack (2012, p. 647) argues that soft mutation manifests itself phonetically and that "phonological integration is a poor diagnostic for code-switching or borrowing status". She compares mutation to the plural marking of English words in French discourse which she says also manifests itself phonetically, and which she says "showed gradual integration according to frequency" in the study by Poplack, Sankoff, & Miller (1988). In contrast to phonological integration, Poplack (2012, p. 647) considers morphosyntactic integration to be a "nearly fail-safe indicator that a LOLI [= 'lone other-language item] has been borrowed". In their rejoinder to Poplack, Deuchar, & Stammers (2012) point out that plural inflection is actually listed by Poplack, Sankoff, and Miller (1988) as one of their indicators of morphological integration. As Deuchar & Stammers (2012, p. 650) say "it is hard to imagine a measure of morphosyntactic integration which is not manifested phonetically". Thus, the fact that soft mutation in Welsh and plural inflections have phonetic manifestations does not in our view disqualify them from being measures of morphosyntactic integration.

Summary of Chapter 4

As mentioned earlier, our aim in this chapter has been to show how work using *Siarad* has made a contribution to the debate regarding whether or not code-switching and borrowing can be differentiated. Our conclusion (summarised in Table 4.3) on the basis of the data we have examined is that they can. In this we agree with Poplack & Meechan (1998, p. 132) that "code-switching and borrowing are two distinct phenomena" and disagree with Myers-Scotton's idea (1993, p. 176) that there is a continuum from code-switching to borrowing. However, we do not agree Poplack & Meechan's definition of borrowings as just those linguistically integrated other-language items. Code switches can also be linguistically integrated, albeit at a relatively low level. We have suggested that the defining characteristics of switches involve both low frequency and low levels of integration. Borrowings are different in that they may have low or high frequency, and will have high levels of integration.

Although we do not find a continuum between code-switching and borrowing this does not exclude the possibility that switches could become borrowings over time. We suggest that this could happen with an increase in frequency accompanied by an increase in integration. We may speculate that if words with a frequency of 10–99 per million words should join the category of 100–999 (as shown

in Figure 4.3) then their level of integration as measured by mutation rate might reach that of the other words in that category, i.e. just over 70%. On the other hand, if frequency is as important as we suggest, we still need to account for the relatively high level of integration of infrequent (1–9 in a million) listed English words (again, see Figure 4.3). Perhaps these infrequent listed words were once more frequent, their high frequency going hand in hand with their relatively high level of integration, but perhaps they are now less frequent. On that interpretation, integration is a one-way process so that an increase in frequency can lead to an increase in integration, but not the reverse. These ideas clearly need testing in future research.

PART 2

Using the corpus

CHAPTER 5

The grammar of code-switching

As indicated in Chapter 1, the primary purpose of collecting the data for *Siarad* was to allow researchers to conduct analyses of code-switching in the natural speech of Welsh-English bilinguals. In Chapter 4 we used an analysis of English verbs switched into Welsh to throw light on the issue of how to distinguish switches from borrowings, and in this chapter, we shall focus on the grammar of code-switching within the clause (intraclausal code-switching). We saw in Chapter 4 that our data shows the regular use of the verbalizing suffix *-io* to integrate English-origin verb stems in line with Welsh morphology, thus suggesting that code-switching follows specific morphosyntactic patterns. Our task in this chapter will be to describe what we know of the patterns speakers use when they switch from one language to another within the same sentence or clause.

This chapter will provide an overview of the code-switching patterns found in the *Siarad* data and will thus provide an insight into the grammar of the informal speech of Welsh-English bilinguals. In describing our data we will make use of the framework known as the Matrix Language Frame (MLF) model (Myers-Scotton 2002; Myers-Scotton & Jake 2015). Carter et al. (2011) explains our rationale for selecting this framework, which we found to meet three important criteria: (i) it is designed to deal with production data, (ii) it is suitable for analysing individual clauses; (iii) it applies to both monolingual and bilingual clauses, i.e. it can be used to analyse clauses entirely in one language as well as clauses which include lexical items from more than one language. After describing the model and illustrating its application, we will review some studies which have made use of the MLF model, before using it ourselves to discuss some data from *Siarad*. Finally, we will evaluate the success of the MLF model in accounting for our data.

The concept of the matrix language in code-switching

Joshi (1985) is usually credited with the introduction of the notion of the matrix language to capture the fact that code-switching often involves an asymmetry between the two languages involved, such that one provides the main morphosyntactic frame and the other, which he labelled the embedded language, provides

material that is inserted into that frame.[38] Joshi recognised that, in clauses containing code-switching, "the two language systems are systematically interacting with each other" (Joshi 1985, p. 191) but suggested that speakers and hearers can usually agree on which is the matrix or main language of a sentence. Joshi formulated a 'switching rule' to capture the possibility of switching from the matrix to the embedded language but not vice versa. He formulated specific constraints on the switching rule, for example that "switching of a category of the matrix grammar to a category of the embedded grammar is permitted, but not vice versa" (1985, p. 192). What this means is that the two languages of a bilingual play different grammatical roles in a clause which includes code-switching. The matrix language provides the same range of grammatical categories as in a monolingual clause, while switches to the embedded language can only involve items which are not closed class, i.e. open class items like nouns, the most frequently switched category in most code-switching data. We can illustrate this with an example from our data introduced in Chapter 4, repeated below as Example (1):

(1) pan dach chi 'n defnyddio **wide-angle lenses**
when be.2PL.PRES PRON.2PL PRT use.NONFIN wide-angle lenses
dach chi 'n **emphasize**-io 'r
be.2PL.PRES PRON.2PL PRT emphasize-VBZ DET
foreground. [Fusser17-BEN]
foreground
"When you use wide-angle lenses, you emphasize the foreground."

In this example, an English noun (*foreground*) and an English verb (*emphasize*) have been inserted into an otherwise Welsh clause, as well as a noun phrase (*wide-angle lenses*). All of these English words belong to open class categories. As we shall explain in more detail later, the word *dach* (a finite verb with tense-marking) comes from the matrix language, Welsh, while only open class items come from the embedded language, English. Myers-Scotton (1993, 2002) developed Joshi's ideas of a matrix language into her theory of code-switching, known as the Matrix Language Frame (MLF) model. We discuss this model in more detail in the next section.

38. Auer and Muhamedova (2005, p. 36) point out that this asymmetry was actually observed as long ago as 1989 by Hermann Paul (1898, p. 366).

Myers-Scotton's Matrix Language Frame model

The MLF model as designed by Myers-Scotton (2002) is based on three related principles (see also Deuchar, 2006). The Matrix Language Principle states that only one language acts as the source of the morphosyntactic structure of a clause containing code-switching, and that this language is called the Matrix Language (ML). The Asymmetry Principle states that the ML will be unambiguously identifiable in such clauses, and the Uniform Structure Principle predicts that structural elements in a bilingual clause are more likely to be sourced from the ML than the other language (where a choice is possible). Myers-Scotton's preferred unit of analysis is the Complementiser Phrase (CP), which is roughly a clause, and is the highest level of structure within a syntactic tree representing a clause, and can contain other structures, such as Verb Phrases, Noun Phrases and dependent CPs (cf. Myers-Scotton, 2002, p. 8). Apart from the ML, the other language participating in a bilingual CP or clause is called the Embedded Language (EL), a term which Myers-Scotton adopted from Joshi (1985) along with the notion of Matrix Language. The above three principles define what Myers-Scotton (2002: 9) calls "classic code-switching", as opposed to "composite code-switching" (discussed in Chapter 7) which does not comply with the Matrix Language Principle. The theory which underpins the MLF model stems from the finding that the two languages in a bilingual clauses have different roles, and only one of these will be the source of the matrix language. This asymmetry is apparent in empirical evidence for "the morphosyntactic dominance of one variety in the frame" (Myers-Scotton, 2002, p. 9). This evidence includes the finding that the EL can only contribute certain types of morpheme to the clause if well-formedness is to be maintained. The permitted EL morphemes, which include 'open class' or content morphemes, will be further discussed below.

According to the MLF, in order to identify which language is the ML of any given clause, two principles may be applied: the System Morpheme Principle and the Morpheme Order Principle (Myers-Scotton, 2002: 59). In Myers-Scotton's model neither principle has primacy, and the requirements of both principles need to be satisfied in a code-switching clause, with both principles pointing to the same language as the source of the ML. We will describe each principle in turn and show how they can be used to identify the ML of a bilingual clause.

The System Morpheme Principle (SMP) states that certain types of morpheme will come from only one of the two languages, the ML: these are morphemes which Myers-Scotton says "must look outside [their] immediate maximal constituents for information about its form" (2002, p. 88). Such morphemes are labelled "outside late[39] system morphemes" in Myers-Scotton's work, and a good example of this is

39. They are called "late" because Myers-Scotton proposes that they are activated later in the sentence-building process than other types of morphemes.

subject-verb agreement, which is often indicated by an affix, inflection, or finite form of the verb. Thus the language of the morpheme which agrees with the subject of the verb is one of the indications of the ML of a clause. This is the case whether or not the entire verb is in the same language or not. In Example (2) below from Congo Swahili-French code-switching (where the French words are in bold) we can see that the entire finite verb *étais* is in French.

(2) Siku hile **j'étais** **sur** **le** **point** ya
CL9/day CL9/DEM 1S/PAST/BE at DET/M point CL9.ASSOC
ku-**renvoy**-er **mon** **épouse** kwa wa-zasi wa-ke.
INF-return-INF POSS/M wife LOC CL2-parent CL2-POSS
"That day I was at the point of returning my wife to her parents."
(Kamwangamalu, 1987, p. 172, cited with glosses by Myers-Scotton 2002, p. 93)

This finite verb *étais* in Example (2) above agrees in number and person (at least in the written form) with the first person subject *j(e)* 'I'. On the other hand, in the Swahili-English data in (3) below – in which the English item is in bold – the verb *si-ku-comment* has the subject agreement prefix *si* and negation marker *ku* in Swahili while the verb stem comment is in English. In both cases the part of the verb which includes subject-agreement belongs to the ML, French in Example (2) above, and Swahili in Example (3).

(3) Ile m-geni, hata si-ku-**comment**
DEM/CL9 CH/S-visitor even 1S.NEG-PST.NEG-comment
"That visitor, I did not even comment." (Myers-Scotton 2002, p. 89)

The Morpheme Order Principle (MOP) states that surface morpheme order in a code-switching sentence will be from one language, the ML. Myers-Scotton (1993, p. 90) uses a Marathi-English example (shown in (4) below) from the work of Pandharipande (1990, p. 18) to illustrate this (the English word is in bold).

(4) **building**-cyā samor ubhā rahā.
building-of in front stand keep
"Stand in front of the building."

As she points out, Marathi has postpositions rather than prepositions – used in English – as reflected in *building-cyā* 'of (the) building'. So, applying the Morpheme Order Principle, the Marathi word order to Example (4) points to the ML of the clause being Marathi rather than English.

The application of the MLF model to Welsh-English data

The application of the MLF model to our Welsh-English data was first presented in Deuchar (2006) and further discussed by e.g. Deuchar & Davies (2009) and Davies (2010). It requires us to identify (a) subject-verb agreement morphemes in the two languages, for the purposes of testing the System Morpheme Principle (SMP), and (b) word order differences, for the purposes of testing the Morpheme Order Principle (MOP). Below we will briefly identify the relevant aspects of the grammars of the two languages which may be used to identify the ML when applying the MLF model to a clause.

As noted above, the SMP predicts that subject-verb agreement will come from the ML. In the Example (5) below the third person singular verb *oedd* agrees with the 3rd person singular subject pronoun *hi*.

(5) oedd hi 'n edrych yn **stunning**. [Fusser30-LON]
be.3S.IMP PRON.3SF PRT look.NONFIN PRT stunning
"She looked stunning."

In applying the MLF model to Welsh-English data, then, we thus look at finite verb morphology to help identify the ML for a Welsh-English clause.

While the identification of *oedd* as Welsh illustrates the application of the SMP in (5), the word order differences between Welsh and English allow us to apply the MOP. While English is an SVO language, Welsh is VSO. The verb-first nature of Welsh means that, in both synthetic and periphrastic constructions, the finite verb is normally the first constituent. This is illustrated by the Welsh sentences in (6a), which is a synthetic construction, and (6b), which is a periphrastic construction, and by the English sentence in (6c). The finite verb is in capitals in each sentence.

(6) a. GWELODD Gareth y castell.
see.3S.IMP Gareth DET castle
"Gareth saw the castle."
b. WNAETH Gareth weld y castell.
do.3S.IMP Gareth see.NONFIN DET castle
"Gareth saw the castle."
c. Gareth SAW the castle.

In both (6a) and (6b) the finite verb (*gwelodd* and *wnaeth*) precedes the subject *Gareth*, whereas in (6c) the verb *saw* follows the subject *Gareth*. In typical declarative constructions, then, Welsh and English subject/verb order is clearly distinct. Note, however, that Welsh can also have subject-initial word order where there is emphasis, as in (7).

(7) Gareth welodd y castell.
Gareth see.3S.IMP DET castle
"It was Gareth who saw the castle."

Here the word order is overtly similar to the English sentence in (6c), and so it could be argued that the Morpheme Order Principle would be unable to identify the language source of the word order in a sentence like (7). In this case the SMP would disambiguate the identification of the ML in favour of Welsh, since the agreement on *welodd* is Welsh.

Another salient word order difference between Welsh and English occurs in NPs where a noun is modified by an adjective or noun. The typical Welsh word order for NPs is that the modifier follows the noun, as in (8a), whereas in English the modifier precedes the noun, as in (8b).

(8) a. castell hardd
castle beautiful
"(a) beautiful castle"
b. (a) beautiful castle

The positioning of a modifier in a code-switching clause is therefore a possible indicator of the ML in Welsh-English code-switching according to the MOP. However, sometimes adjectives in Welsh do not follow the noun, but precede it. Such adjectives fall into three groups: those which must precede the noun, e.g. *prif* 'main, prime', those which usually precede the noun but do not always do so, e.g. *hen* 'old', and adjectives which come from a set of words which convey a slightly different meaning depending on whether they precede or follow the noun, e.g. *unig* 'only/lonely'. These are illustrated in (9a) to (9e) below.

(9) a. prif weinidog
prime minister
"(a) prime minister"
b. hen ddyn
old man
"(an) old man"
c. dyn hen
man old
"(an) old man"
d. unig blentyn
only child
"(an) only child"
e. plentyn unig
child lonely
"(a) lonely child"

However, since adjectives patterning in this way come from a limited set, adjectival modifiers will most often be useful in testing the MOP on a clause. Nevertheless, we may note that since an 'English-like' modifier-head word order in NPs does exist in Welsh, albeit highly restricted, it is possible that the prevalence of this word order in English could influence Welsh speakers to extend their use of it in their Welsh. This is discussed further in Chapter 7, as well as in e.g. Davies & Deuchar (2010).

Numerous other word order differences exist between Welsh and English, but for the purposes of identifying the ML we can conclude that both subject/verb and head/modifier word orders are viable for the application of the MOP.

Applying the model to data

Having pointed out some salient differences between Welsh and English grammar, we now turn to demonstrating how these differences can be used to identify the ML in a clause containing Welsh-English code-switching. Example (10) below is an utterance from *Siarad*, and we shall use this to illustrate how the ML of a clause can be identified.

(10) oedd 'na fath â ryw **alley** bach yna. [Davies6-HEC]
be.3S.IMP there kind with some alley little there
"There was kind of a little alley there."

Applying the SMP, the finite verb *oedd* points to Welsh as the ML. With regards to the MOP, the word order available for examination here is the relative position of the subject and the inflected verb as well as the relative position of the noun and adjective in the NP *alley bach*. Since (10) is a declarative sentence, and the finite verb *oedd* is the first clausal element, we can identify the source of this verb position as Welsh. In addition, the NP *alley bach*, which has a Welsh adjective modifying an English code-switched noun, also shows the expected Welsh word order of a modifier following its head noun. The MOP therefore points to Welsh as supplying the ML for (10). Since the SMP and the MOP both point to Welsh, we can confidently identify Welsh as the ML.

In (10) we have a case where there is sufficient evidence to check, and satisfy, both principles. In fact, here either principle on its own would have been sufficient to identify the ML, but it is important to check both principles where possible, since as we will discuss in Chapter 7 there are cases where both principles do not identify the same language as ML, and such cases would deviate from the "classic code-switching" (cf, Myers-Scotton 2002, p. 105) that is described by the MLF.

Sometimes in naturally occurring data there will be clauses without enough evidence to check both principles. The MOP, for instance, can be applied where a clause includes a noun and adjective in different languages as in (10) above, even if

there is no finite verb allowing application of the SMP. We can illustrate this with Example (11) below, which the speaker produces an exclamation about the pressures of long distance commuting to work:

(11) arglwydd **pressure** diawledig [Davies4-OSW]
lord pressure diabolical
"God, diabolical pressure!"

In the phrase *pressure diawledig* ('diabolical pressure') the modifier *diawledig* follows the head noun *pressure* as in Welsh grammar, suggesting that the ML of this (incomplete) clause is Welsh. In this case the MOP acts as a standalone check for ML in the absence of subject-verb agreement morphology.

While the MLF was designed to describe the grammar of code-switching, the model can also be applied to clauses without code-switching, to check whether the ML is as one would predict from the language of the lexical items. Example (12) below shows a Welsh monolingual finite clause with a Welsh ML and Example (13) shows an English monolingual finite clause with an English ML.

(12) wnes i 'm cael hi tro
do.1S.PAST PRON.1S NEG get.NONFIN PRON.3SF time
cynta de. [Davies6-DAN]
first TAG
"I didn't get it the first time, eh."

(13) *oh* **there's nothing wrong with you**. [Fusser6-ANT]

In (12), the finite verb *wnes* precedes the subject pronoun *i* 'I', indicating Welsh as the provider of morpheme order; this indication is supported by the order in the head/modifier NP *tro cynta* 'first time', where the adjective *cynta* follows the head noun *tro*, showing Welsh-predominant modifier-head word order. The inflection on *wnes* matches the person of the subject and thus indicates that Welsh also provides the outside late system morphemes in the clause. As both MOP and SMP thus point to Welsh, that language is identified as the ML. In (13), the subject *there* precedes the verb *(i)s*, indicating English-predominant SV order. The morphology of the finite verb (*is*) is also English. Thus the MOP and the SMP both point to English as the source of the ML.

Although it might seem obvious that the ML of (12) should be Welsh and of (13) English, since both are monolingual without code-switching, the ML may not always match the source language of the lexical items In Examples (12) and (13) the ML and the language of the lexical items do match, but this is not always the case.[40]

40. See e.g. the discussion of the Dakkhini language by Afarli, Grimstad, & Subbaroa (2013), outlined in Chapter 8.

Previous studies of Welsh-English data using the MLF model

Previous studies examining Welsh-English code-switching – many of which have been conducted by our research group at Bangor – have made use of the MLF model, and we review them here in chronological order of publication.

Deuchar (2006) presents the results of an analysis of Welsh-English code-switching using the MLF model, where she used a small corpus of consisting of five hours of conversational data from 30 Welsh-English bilinguals; all but one were face-to-face conversations, while one was a telephone conversation. The participants of that corpus were from both north and south Wales and consisted of both males and females, aged from about 20 to middle age. In that paper, Deuchar attempted to identify whether Welsh-English code-switching data could be described as "classic" code-switching according to Myers-Scotton's definition i.e. whether or not a single language could be identified as being the source of the ML in bilingual clauses. Deuchar tested the three main principles of Myers-Scotton's theories, the Matrix Language Principle, the Asymmetry Principle and the Uniform Structure Principle.

Deuchar manually extracted 163 bilingual clauses from one of the conversations in the corpus, and tested each one as to whether they conformed to the principles described above. Possible options for clausal ML in Deuchar's analysis are given as "Welsh" (Welsh word order and Welsh agreement morphemes), "English" (English word order and English agreement morphemes), "either" (word order compatible with Welsh or English and no agreement morphemes present) and "neither" (word order compatible with neither Welsh nor English and no agreement morphemes present). Of the 163 bilingual clauses analysed, 141 clauses (87%) had a Welsh ML, 4 (2%) had an English ML, 17 clauses (10%) had an "either" ML, and one clause (1%) had a "neither" ML. Thus the great majority of bilingual clauses had a Welsh ML, and English was rarely the ML – in fact, more clauses had an "either" ML than had an unequivocally English ML. From the point of view of testing how well the MLF could be applied to Welsh-English data, Deuchar points out that 89% of clauses had an identifiable ML (Welsh ML or English ML). The 18 clauses which did not have a single identifiable language as the ML source were similar, in that (i) they lacked evidence of subject-verb agreement to satisfy the System Morpheme Principle, and (ii) word order was compatible with both Welsh and English. Deuchar (2006) interpreted the Asymmetry Principle empirically as suggesting that the ML would be unambiguously identifiable in most clauses. This was found to be the case, as indicated above.

To test the Uniform Structure Principle on her data, Deuchar checked to see whether the Welsh definite article or determiner *y/yr/'r* was used in clauses which had a Welsh ML. Welsh definite articles do not (unlike case-marking determiners in German) participate in agreement outside the nominal construction and thus are not late outsider morphemes. Therefore according to the MLF model they

would not be required to come from the ML. However, the Uniform Structure Principle would predict that Welsh (ML) determiners are preferred over English (EL) determiners, and so we would expect to find Welsh definite articles in clauses with Welsh ML. Deuchar analyses 61 definite articles in her dataset, 57 of which were Welsh (*y/yr/'r*) and 4 of which were English (the). She found that English determiners were only found either in clauses with an English ML or where they preceded an English "Embedded Language island" (two or more English words following English word order in a clause with Welsh ML). Primarily, however, Welsh determiners were used, even (in fact, preferentially) when the determiner preceded EL lexical items. Deuchar therefore argues that the Welsh-English data upholds the Uniform Structure Principle. Overall, since all three principles are upheld by her analysis, Deuchar describes Welsh-English data as being a case of "classic" code-switching according to Myers-Scotton's criteria.

Davies (2010), a PhD thesis, conducted the first analysis applying the MLF model to our *Siarad* data. Some of his results are also reported in Deuchar & Davies (2009) and in Davies & Deuchar (2010). His analysis uses three conversations from *Siarad* in which a total of six speakers are involved. We report on the findings of Davies (2010) in detail here, first showing the ML distribution across all types of clauses, then the results for finite clauses only, followed by a comparison of ML distribution in monolingual and bilingual clauses. We will also compare the ML patterns of the six speakers who were analysed.

Figure 5.1 illustrates the results of applying the MLF model to all clauses in his sample. The data examined include finite, non-finite and verbless clauses and both monolingual and bilingual clauses. Of the 3275 clauses uttered by all six speakers, 1874 or 57.22% had an identifiable ML (i.e. they were identified as having either a Welsh, an English or a dichotomous ML). The concept of a "dichotomous" ML (Deuchar & Davies, 2009; Davies & Deuchar, 2010) indicates clauses which show conflicting information as to the source of the ML. Clauses with dichotomous ML contrast with unidentifiable ML clauses, as the latter do not have enough information to use as evidence of the clause's ML, whereas the former has enough evidence but that evidence is ambiguous. We discuss dichotomous ML clauses in greater detail in Chapter 7, where we argue that in our data they are indicative of influence from contact with English.

Of the 1874 clauses in which an ML was identifiable, 1788 (95.41%) had a Welsh ML, 84 (4.48%) had an English ML and two (0.11%) had a dichotomous ML. The remaining 1401 clauses (42.78% of all clauses) analysed could not have an ML identified for them as there was insufficient information within those clauses on which to test the MOP and SMP. In clauses where an ML was identifiable, Welsh was the most frequent provider of morphosyntax, being the ML in 95.41% of such clauses in these data. Note that there were more clauses in the data which had an identifiable ML than those which lacked an identifiable ML.

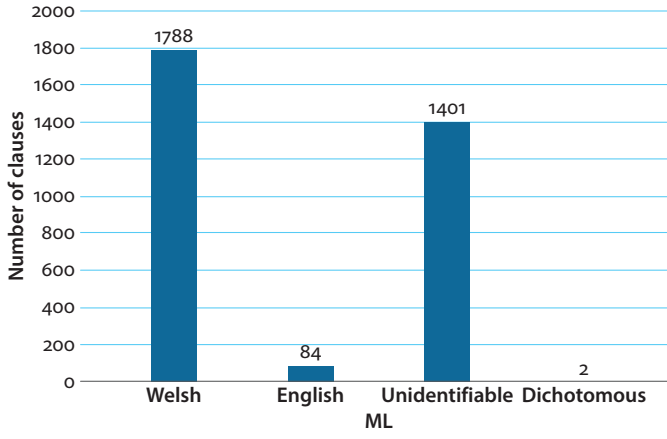


Figure 5.1 The Matrix Language distribution across all the clauses in the dataset

Unequivocally identifying the ML in a Welsh-English clause without finite verb agreement markers presents difficulties (cf. Deuchar 2006). This means that the most suitable clauses for testing the MLF model on Welsh-English data are finite clauses. Davies analysed all 1862 finite clauses in his dataset, representing 56.85% of the total number of clauses. An ML was identifiable in all of those clauses. 1778 (95.49%) finite clauses had a Welsh ML, accounting for 54.92% of the clauses in the dataset overall. English was the ML in 82 (4.40%) finite clauses, accounting for only 2.50% of all clauses in the dataset. There were two finite dichotomous ML clauses (the only two dichotomous ML clauses in the data), accounting for 0.11% of finite clauses and 0.06% of all clauses in the dataset. The ML distribution in finite clauses, both monolingual and bilingual, is illustrated in Figure 5.2.

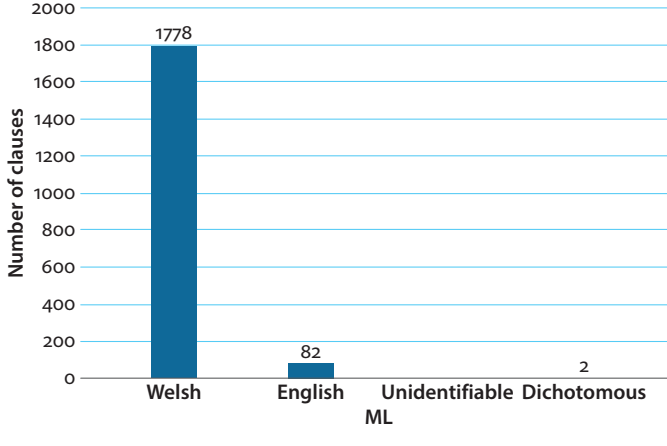


Figure 5.2 The Matrix Language distribution across all the finite clauses in the dataset (monolingual and bilingual combined)

One question raised by Davies (2010) was whether or not ML distribution varied depending on whether a clause contained code-switching ('a bilingual clause') or not ('a monolingual clause'). Focusing on finite clauses, a comparison of monolingual and bilingual clauses in Davies' analysis is shown in Figure 5.3 below.

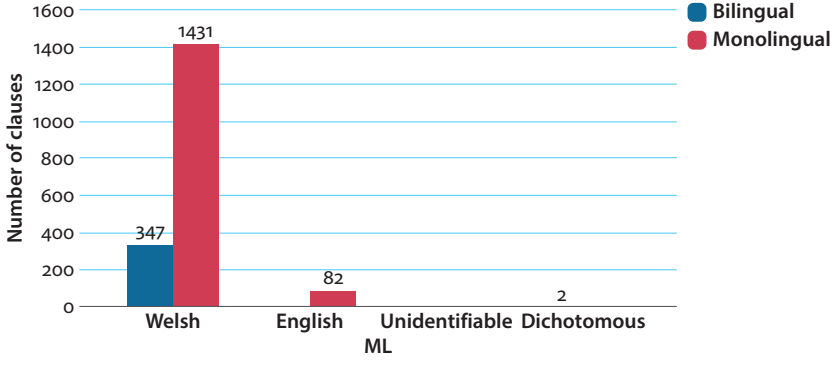


Figure 5.3 The Matrix Language distribution across all finite clauses in the dataset, comparing the distribution in bilingual clauses to the distribution in monolingual clauses

Of the 1862 finite clauses in the data sampled for this study, 1513 (81.26%) were monolingual whilst 349 (18.74%) were bilingual. The monolingual clauses were distributed in terms of ML as follows: 1431 (94.58% of monolingual clauses) clauses had a Welsh ML while 82 (5.42%) had an English ML. Since code-switching is the focus of our interest, we will discuss only the bilingual clauses in detail here.

As Figure 5.3 above shows, 347 out of the 349 bilingual finite clauses in the dataset (99.43% of the total bilingual finite clauses) had a Welsh ML and none had an English ML; the remaining two (0.57%) bilingual finite clauses had a dichotomous ML. Examples (14) and (15) below show two bilingual finite clauses from *Siarad* as an example of bilingual clauses with Welsh ML.

(14) mae raid bod gynno fo ryw fath o
be.3S.PRES necessity be.NONFIN with.3SM PRON.3SM some kind of
attitude *problem* de. [Davies6-HEC]
attitude problem TAG
"He must have some kind of attitude problem, eh."

(15) mae **Americans** yn mwy **commercial**. [Fusser27-LIS]
be.3S.PRES Americans PRT more commercial
"Americans are more commercial."

In (14) the finite verb *mae* precedes the subject *raid* 'necessity', indicating Welsh VS morpheme order. The verb morphology of *mae* is also Welsh, and its third person singular form agrees with the subject. Welsh is thus identified as the source of the

ML. The head/modifier NP *attitude problem*, which has English order (with the modifier coming before the head) and contains only English morphemes (though *problem* is also available in Welsh), is an EL island. In (15) the subject *Americans* follows the finite verb *mae*, again indicating Welsh VS order; *mae* is a singular verb whilst Americans is a plural subject, following the obligatory agreement rule of Welsh syntax for pronominal subjects (e.g. Roberts 2005, p. 44), a rule which does not occur in English, reinforcing the identification of Welsh as the ML of (15).

It is possible to infer two things from the results represented in Figure 5.3. First, Welsh is clearly the main source of the ML in bilingual finite clauses produced by these speakers. Second, English is hardly ever the source of the ML in bilingual clauses – there were no examples found in Davies' (2010) analysis, and they constituted only 2% of the clauses analysed by Deuchar (2006).

Davies (2010) also looked at variation between different speakers in terms of their ML preference. Table 5.1 below compares the ML in both bilingual and monolingual finite clauses produced by the six speakers, comparing here the number of clauses they produced with Welsh ML and English ML.

Table 5.1 Number of clauses with a Welsh or English matrix language, by speaker

Speaker	Welsh ML	English ML	Total	% Welsh ML	% English ML
Amranwen	83	1	84	99	1
Hector	298	6	304	98	2
Mabli	467	10	477	98	2
Antonia	351	12	363	97	3
Lisa	342	23	365	94	6
Daniel	237	30	267	89	11
Total	1778	82	1860	96	4

When we examine the data for each speaker separately we find that the general pattern found in the dataset as a whole is apparent. Welsh was the ML in the great majority of clauses produced by each speaker and no speaker used English as the ML more than Welsh as the ML. Amranwen produced the greatest proportion of clauses with a Welsh ML in her total output of finite clauses (98%), while Daniel produced the lowest proportion of clauses with a Welsh ML (89%), which is still a high proportion. The speaker who produced the largest proportion of clauses with an English ML in the dataset was Daniel: 30 out of 267 (11%) had an English ML, which is still a small proportion overall. The other speakers produced very few clauses with English ML in comparison to their total output – 6% or less It will be recalled from the data presented in Figure 5.3 that all clauses with English ML were monolingual, containing no code-switching. Davies (2010) explains that the higher proportion of English ML clauses produced by Daniel in his recording stemmed from his frequent use of direct quotations in English in his conversation,

which occurred because he (and, to a lesser extent, his speaking partner Hector) was reading aloud from an English newspaper during parts of the recording.

In the analysis reported by Davies (2010) we have seen that there was a clear predominance of Welsh ML over English ML. Previous studies of code-switching morphosyntax (e.g. Smith 2006, who looked at Spanish-English bilinguals' speech) confirm that bilinguals commonly show asymmetry in their speech, such that they produce bilingual clauses with one language as the source of morphosyntactic structure more frequently than the other. In a paper discussing code-switching between typologically-distinct languages (of which Welsh and English are examples), Chan (2009, p. 197) suggests that "there seems to be a universal tendency to select one morpho-syntactic rule". This is certainly reflected in our data by the overwhelming tendency for speakers to choose Welsh as the ML in both monolingual and bilingual clauses. Comparing the data analysed by Davies (2010) with two other datasets – Spanish-English and Welsh-Spanish bilingual data – Carter et al. (2011) suggest that the uniformity of the ML in the Welsh-English data may be related to the fact that Welsh and English have contrasting orders (SVO vs. VSO). In this case they suggest that the optimal option is to select the word order of one of the languages rather than both. Conversely, they suggest that where two languages have similar word order (e.g. English and Spanish are both SVO), one would expect a more equal balance of ML choice by speakers, which is indeed is what they find in their sample of Spanish-English data collected in Miami. This suggestion is supported by evidence from syntactic priming in bilingual data (see e.g. Torres Cacoullos & Travis, 2016; Fricke & Kootstra, 2016) which demonstrates that the structure and lexical items of utterances previously heard or produced by a speaker can influence the structure and lexicon of a new utterance.

If, as suggested by our data, speakers of two contrasting languages do tend to select one as ML across the discourse, what factors determine which one they select? According to Myers-Scotton (2002, p. 25), a necessary condition for speakers' choice of a language as ML is full proficiency in that language, whereas they can have "anywhere from limited to full proficiency in the other language". We reported in Chapter 2 that 77% of our participants reported being confident speakers of Welsh, 68% confident speakers of English, and that 63% appeared to be balanced bilinguals. This suggests that speakers do have a high enough level of proficiency in both Welsh and English to choose both as an ML. However, Welsh is chosen overwhelmingly more frequently[41] as an ML by all speakers. Why should Welsh be chosen as the main ML rather than English, despite the fact that the majority of speakers are proficient

41. However, we do have some evidence that speakers with lower confidence in Welsh produce more (mostly monolingual) clauses with English ML than other speakers. But clauses with Welsh ML are still in the majority in the data from the least confident speakers.

in both? If both Welsh and English meet Myers-Scotton's proficiency condition for choice of ML, then other factors may help to explain why Welsh and not English is chosen. Carter et al. (2011) suggest that the predominant choice of Welsh as ML may be a way of communicating membership of the Welsh community.

In Figure 2.12 in Chapter 2, we show how 90% of our participants chose Welsh as their identity, in contrast to English, British or Other. According to Myers-Scotton (1998, p. 98), the ML selected may be "indexical of solidarity" and associated with ingroup membership. In fact, the indigenous, minority language seems to be used as ML in many communities.[42] It is thus not surprising if Welsh is used as ML to represent the strong Welsh identity of the majority of our participants.

We suggested above that syntactic priming, or the syntax of utterances already heard or spoken, may influence the maintenance of the same structures. If so, it is possible that the structure of the language spoken with one's most frequent interlocutors could have an effect also. We reported in Chapter 2 that we asked our participants which five people they spoke to most frequently, and which language they spoke with those five. The responses allowed us to assign social network scores to each participant, based on the extent to which they spoke Welsh, English or both languages with their closest contacts. Figure 2.10 showed that 60% spoke mainly Welsh, indicating that the syntactic structures of Welsh were most frequent in their conversations. So both speakers' sense of Welsh identity as well as the influence of their habitual language choice may be factors that explain why Welsh is chosen as ML.

Although we have found the MLF to be a fruitful framework for analysing our data, there has been some criticism of it (see e.g. Auer & Muhamedova, 2005), and it has been argued (see e.g. MacSwan, 2009) that a theory of code-switching which does not appeal to mechanisms specific to code-switching is superior to one which does. MacSwan argues that the generative Minimalism Program can provide a simpler and more elegant account. An example of a Minimalist approach is provided by Moro Quintanilla's (2014) account of determiner phrases in Spanish-English code-switching, in which she argues that the existence of a gender feature on the Spanish determiner compared with its absence on the English determiner means that while Spanish determiners can occur with English nouns, the reverse pattern has not been considered grammatical by previous authors. Since Welsh, like Spanish, has grammatical gender, Herring, Deuchar, Parafita Couto, & Moro Quintanilla (2010) analysed determiner phrases (i.e. DPs that contain a determiner and a noun from different languages) in Welsh-English (from *Siarad*) and Spanish-English (from our Miami corpus) code-switching data. The aim was to compare the effectiveness of the

42. Bhatt (2015) argues on the basis of code-switching data from India that the relation between the 'base' (ML) and 'intrusive' (EL) is one of power, with the ML being an indigenous variety with relatively low status while the EL is English, the language of prestige.

MLF model and the generative Minimalist Program (MP) in predicting the language of determiners in mixed determiner phrases. The prediction of the MLF was that the determiner will be sourced from the ML, and as we have reported above, evidence had already been reported by Deuchar (2006) for the accuracy of this prediction. The prediction of the MP, on the other hand, was that the determiner would be sourced from the language with grammatical gender, i.e. Welsh or Spanish but not English. Thus, for Welsh-English code-switching, the MP would predict that speakers would say *y television* but not **the teledu* 'the television'. The data analysed by Herring et al. from both language pairs showed that their speakers did make more use of determiners from the language with gender (Welsh or Spanish), consistently with the predictions of the MP, but it was also the case that Welsh and Spanish were also used more frequently than English as the ML. Since Welsh was the ML in all the mixed determiner phrases, it is therefore not possible to conclude that the preference for Welsh determiners was caused by the fact that Welsh has grammatical gender. Instead, it may be due to the fact that the ML was Welsh. Future research from communities where English is used as an ML in code-switching data may throw light on whether grammatical gender as well as the ML has a role in the structure of mixed determiner phrases, but for the moment we can conclude that the MP is not a convincing alternative account of Welsh-English code-switching when compared with the MLF. Furthermore, we note that Fricke & Kootstra (2016) found the ML to be a useful construct in their study of priming in our Spanish-English data, and we take this to be a preliminary indication of its potential use in accounts of bilingual production.

To summarise, the studies of code-switching in *Siarad* which have been published so far all identify a similar pattern, where Welsh is almost always the ML chosen for bilingual clauses, and speakers tend to behave similarly in this regard. However, those analyses only looked at the speech of a small number of participants, primarily because the method of data extraction was manual and highly time-consuming. In the next section we will present the results of an automated analysis of all the speakers in *Siarad*, to see whether or not the findings of Davies (2010) etc. are replicated using the whole corpus.

Automatic analysis of Matrix Language distribution in the *Siarad* corpus

The analysis we have reported on so far was conducted manually and was labour-intensive, but the autoglossing system described in Chapter 3 has allowed us to deal with the entire corpus simultaneously and identify patterns across a large number of speakers. As we shall show, this has allowed us to verify our findings

from a small number of speakers by comparing them with the automatically produced findings from the entire corpus.

The automatic glossing system (see Chapter 3) makes it possible to identify the language of the finite verb within a clause, which is an implementation of the System Morpheme Principle outlined above. The other Principle of the MLF, the Morpheme Order Principle, requires us to look at the relative position of the finite verb and the subject as an additional way of identifying the source of the ML, but we discovered that this criterion was less conducive to automatic analysis. For the automated analysis we therefore decided that we would only use the System Morpheme Principle as a means of identifying the ML. This decision can be justified on the grounds that we found a high degree of agreement between the two Principles in identifying the ML. In other words, we found that the word order used by speakers almost always matched the language identified by the finite verb of the clause; exceptions tend to involve the relative position of a modifier and a noun and will be discussed in Chapter 7. So for our Welsh-English data, at least, it seems sufficient to use only the System Morpheme Principle to give us a generally accurate overview of the ML distribution of the clauses.

Before coding the clauses according to their matrix language it was necessary to split our corpus into clauses. In fact, only 24% of the utterances in the corpus were longer than one clause and therefore required this. Welsh is the predominant language of the corpus (only 4% of words are unambiguously English), but since no parser is as yet available for Welsh, we used a relatively unsophisticated method to segment these utterances. (A similar approach was used for English and mixed utterances.) This involved (i) using the autogloss to mark all finite verbs, (ii) moving the marker leftwards as required onto conjunctions, relatives or interrogatives where these preceded the verb and (iii) dividing the utterance at the marker. In (16) the mark points are underlined, and the segmentation points are marked with '/'. Note that the automatic glossing system is used in the examples in this section rather than the manual glossing system used elsewhere in the book.

(16) mae (y)r hogan (y)na / oedd ar
be.V.3S.PRES the.DET.DEF girl.N.F.SG there.ADV be.V.3S.IMPERF on.PREP
Eastenders / mae hi (we)di gwneud
Eastenders be.V.3S.PRES she.PRON.F.3S after.PREP make.V.INFIN
un / dydy. [Roberts2-IRW]
one.NUM be.V.3S.PRES.NEG
'That girl from Eastenders, she's done one, hasn't she?'

In (17) and (18), the mark point has been moved leftwards from the finite verb marked with an asterisk:

(17) mae (y)n sure / eith hi
be.V.3S.PRES PRT sure.ADJ go.V.3S.PRES she.PRON.F.3S
rywbryd timod / ond mae*
at_some_stage.ADV + SM know.V.2S.PRES but.CONJ be.V.3S.PRES
(h)i (y)n mwynhau yn fa(n) (y)ma
she.PRON.F.3S PRT enjoy.V.INFIN PRT place.N.MF.SG + SM here.ADV
wedyn… [Davies2-GRE]
afterwards.ADV
"She'll probably go sometime, but she's enjoying it here, so …"

(18) fasai fo (y)n gwybod / (ba)sai
be.V.3S.PLUPERF he.PRON.M.3S PRT know.V.INFIN be.V.3S.PLUPERF
medru rheoli (we)dyn / be mae*
be_able.V.INFIN manage.V.INFIN afterwards.ADV what.INT be.V.3S.PRES
(y)n fwyta wedyn / basai? [Fusser13-ANN]
PRT eat.V.INFIN + SM afterwards.ADV be.V.3S.PLUPERF
"He'd know he could then control what he eats then, couldn't he?"

To test the accuracy of the segmentation of clauses in Welsh, the predominant language, 1318 Welsh-only utterances were collected, and every tenth utterance was examined to check whether the clauses had been correctly segmented. Since all those utterances contained four or more clauses, this resulted in a sample of 528 clauses, in which there were 35 errors (7%). There were 30 instances of a split where none was required, four of a required split not being made, and one where a clause had been marked as finite when it contained no verb. Although utterances consisting of four clauses or more (as in the test) make up only 2% of the corpus, they make a particularly rigorous test sample because their length increases the number of possible places for segmentation errors to occur. Thus the error rate for these longer utterances is likely to be an upper limit on the overall error rate, and one would expect the error rate to be lower overall. This expectation was tested manually using a sample from Stammers4. The first 200 utterances of the transcript of Stammers4 were split by hand and compared to the output from the clause splitter. In these 277 clauses there was only one error (a split where none was required) – an error rate of less than one per cent. We therefore conclude that the automatic segmentation is a sufficiently accurate basis on which to draw conclusions.

The segmentation process produced 81,352 candidate clauses from 158[43] speakers represented in *Siarad*. For each of these, the glosses assigned to the words by the autoglosser were scanned in software to pick out the word in the clause that was

43. This number included the seven for whom we do not have questionnaire data, as mentioned in Chapter 2.

likely to be a finite verb. Then in order to assign an ML to the clause, the part-of-speech tag assigned by the autoglosser to each word was used to pick out the most likely finite verb in the clause, and this was then combined with the language tag assigned by the transcriber to that word to give a suggested matrix language for the clause. If there was no language tag for the verb, this determined the assignment of the default language, Welsh, as ML, whereas if the putative verb was tagged as *@s:eng*, this would identify English as the ML of the clause. So for example the occurrence of the Welsh finite verb *mae* ('is, are') in the clause would mean that the ML in that clause was determined as Welsh, whereas the occurrence of *is@s:eng*, tagged as English, in the clause would mean that the ML in that clause was determined as English. In a minority of cases the suggested ML is incorrect, due to deficiencies in this version of the autoglosser, particularly in relation to English.

Table 5.2 provides the preliminary results for clauses for which an ML was automatically suggested. Although we shall provide some revised figures below as a result of manual corrections, we can already safely conclude that Welsh is overwhelmingly the most common ML in the corpus.

Table 5.2 Matrix Language distribution across finite clauses in *Siarad* (automatic analysis)

Matrix Language	Number of clauses	% of all finite clauses
Welsh	65,134	98
English	1,295	2
Total	66,429	100

The results of this automatic analysis show that the overall matrix language distribution patterns identified in previous analyses of (i) a subset of *Siarad* (Davies 2010, etc.) and (ii) of data from a pilot corpus (Deuchar, 2006) are repeated when the whole of *Siarad* is analysed. Welsh is chosen as ML overwhelmingly frequently while English ML is very rare.

The advantage of the automated analysis is that it allows us to identify with a high degree of confidence the speech patterns of a large number of speakers over tens of thousands of clauses. There are, admittedly, a low number of errors, but these need to be balanced against the large amount of data that can be dealt with using this approach.

As well as being coded for the language of the finite verb, the clauses in our data were also coded for linguality, or whether they were monolingual (all words from one language) or bilingual (words from both languages). In our manual analysis we found that while clauses with a Welsh ML are the most frequent type found in both monolingual and bilingual clauses, we almost never found bilingual English

ML clauses – all clauses containing code-switching had a Welsh ML in Davies (2010) and Davies & Deuchar (2009), whereas Deuchar (2006) found very few. The automated analysis allows us to see if this pattern holds for the whole corpus. The distribution of the ML in monolingual clauses vs. bilingual clauses for those clauses for which an ML was suggested is shown in Table 5.3. Table 5.3 includes the results following some manual checking which suggested that the number of bilingual clauses with English ML was incorrect. This resulted from the fact that the autoglosser was not always able to identify English finite verbs and split clauses containing them.

Table 5.3 ML distribution across monolingual and bilingual clauses in *Siarad* (modified automatic analysis)

	Monolingual	**Bilingual**	**Total**
Welsh	59,309	5,908	65,217
English	1,158	53	1,211
Total	60,467	5,960	66,428
% Welsh ML	98	99	98

As can be seen, in both monolingual and bilingual clauses, the slightly corrected automated analysis shows that Welsh is the ML in the overwhelming majority of clauses (98% in monolingual clauses, 99% in bilingual clauses). Of clauses with a Welsh ML, 91% are monolingual and 9% bilingual. Note that although in the manual analysis by Davies (2010) no bilingual clauses were found to have an English ML, the sample used by Davies was relatively small. Now that we have a larger sample we can see bilingual clauses with English ML do occur, but they make up less than 1% of the total data analysed and only 4% (53/1211) of the total number of clauses with English ML. This suggests that bilingual clauses with English ML are dispreferred in relation to bilingual clauses with Welsh ML.

As shown in Table 5.3, Welsh is the ML in the majority (5908/5961 or 99%) of the bilingual clauses in our data. In these clauses, Welsh provides the main morphosyntactic frame into which English words or phrases are inserted. But in the very small number of bilingual clauses with English ML, the reverse is the case: Welsh words or phrases are inserted into an English morphosyntactic frame. Some examples are provided below. The most frequent type involves the insertion of a Welsh tag like *de* at the end of an otherwise English utterance, as in (19):

(19) *oh* **no everything stops for the** *Bill* de. [Roberts3-LER]
IM no everything stops for the Bill TAG
"Oh no, everything stops for the Bill, eh."

This kind of 'tag switching' (cf. Poplack 1980, p. 589) is peripheral to the clause as pointed out by Muysken (2000, p. 99) and has little impact on the clause structure. Peripherality is also a feature of the next most frequent type of switch into Welsh, involving adverbials. Muysken (2000, p. 100) notes that there are many examples of adverbial switches in Treffers-Daller's (1994) French-Dutch data from Brussels. Adverbials are peripheral in the sense that they could be deleted without affecting the argument structure of the clause. This is illustrated in Example (20) below where *o gwbl* 'at all' is an adverbial phrase, a kind of 'add-on' to the English statement *it's not on*.

(20) **it's not on** o gwbl. [Fusser3-ALY]
it's not on at all
"It's not on at all."

Example (21) illustrates the use of an adverbial phrase EL island in Welsh even though the finite verb and rest of the word order is in English:

(21) efo *project* iechyd a lles **it's wider than** *just*
with project health and welfare it's wider than just
health. [Fusser22-WYN]
health
"With the health and well-being project, it's wider than just health."

We also have some examples where the main clause is in English and the subordinate clause is in Welsh. This is illustrated in (22), where a conditional clause in Welsh follows the main clause in English:

(22) **no it's a court appearance** os y fi cael
no it's a court appearance if be.1S.PRES PRON.1S get.NONFIN
yn ddal am hwnna. [Davies9-MOS]
POSS.1S catch.NONFIN for that_one
"No, it's a court appearance if I get caught for that."

In all the above examples the inserted Welsh material is relatively peripheral to the clause and might be considered to exemplify what Muysken (2000, p. 26) calls the alternation pattern of code-switching, in which "the two languages in the clause remain relatively separate".

The other main category of bilingual clauses with English ML is where a complement to the English verb *be* is inserted in Welsh. Two clauses of this kind occur in the utterance in (23):

(23) **it's** cinio Pasg, **it's** yr un fath. [Lloyd1-JEA]
it's lunch Easter it's DET one kind
"It's Easter lunch, it's the same thing"

In this utterance, material in Welsh appears as the complement of the English verb in both clauses. Note that *cinio Pasg* 'Easter lunch' is an Embedded Language island since the word order within the phrase is Welsh.

It is important to emphasize that the examples above of bilingual clauses with English ML come from a subset of no more than 1% of the clauses in our data to which an ML could be assigned. The fact that bilingual clauses with an English ML are so few supports our previous findings based on much smaller samples, but the very large size of the automatically analysed sample has allowed us to identify some interesting patterns. In general we can say that although Welsh insertions into an English morphosyntactic frame are definitely dispreferred, they can occur, but usually in clause-peripheral positions. This peripherality suggests that the English grammar of our speakers is in some sense more impervious to the influence of Welsh than their Welsh grammar is to English. However, more research is needed before we can speculate further, and this could include investigation of the type of speakers who are found to use Welsh insertions in otherwise English clauses.

Reviewing the MLF

In this chapter we have presented the results of applying the MLF model to a set of data from Welsh-English bilinguals, based on Myers-Scotton's (2002) version of the model. Having identified that Welsh is the predominant ML for Welsh-English bilinguals' clauses, we now turn to evaluating how effective the MLF model is in describing the Welsh-English data.

One way of measuring our success in applying the MLF model to the data could be to consider its scope, or the extent of coverage of the data – that is, how many clauses can in the data for which the MLF can be conclusively used to identify an ML. In the analysis of a subset of the data by Davies (2010), which looked at a subset of *Siarad* data, an ML could be identified in all finite clauses analysed, indicating that the MLF has a very broad scope when applied to finite clauses produced by Welsh-English bilinguals. However, these were only 57% of all the clauses in the dataset (1862 out of 3275). When the model is applied to clauses without a finite verb, its scope is much narrower. 14% of the clauses (450 out of 3275) in the dataset analysed for Davies (2010) were non-finite, and only four such clauses could have an ML identified; in such clauses naturally only the Morpheme Order Principle could be tested, since they contain no finite verb agreement morphology. Similarly, verbless clauses comprised 29% of the dataset (963 out of 3275), and only eight such clauses had an identifiable ML; again only the Morpheme Order Principle could be applied. In total, an ML could be identified in 57% of clauses in that dataset when all clause types are considered. It looks as though the MLF is very effective in

identifying the ML in finite clauses in Welsh-English data, but relatively ineffective in identifying the ML in any other type of clause.

However, we may have taken an unduly conservative approach in limiting our study to finite clauses and ignoring incomplete or verbless clauses. Eppler, Luescher, & Deuchar (2017) point out that Myers-Scotton (2002, p. 55) suggests that a clause may have null elements, and in their study involving the comparative evaluation of three syntactic frameworks (including the MLF) for mixed determiner-noun constructions they take account of assumed null elements in assigning matrix languages to the clauses in their data. This alternative approach could also have been applied to the Welsh-English data also. For example, the utterance shown in (24) below contains a non-finite clause *a syrffio fel* 'and surf like' which was not assigned a ML by Davies (2010) because the SMP and MOP cannot be applied to this clause.

(24) aethon ni aros lawr yn *Whitesands* a
go.1PL.PAST PRON.1PL stay.NONFIN down in Whitesands and
syrffio fel [Fusser27-LIS]
surf.NONFIN like
"We went to stay down in Whitesands and surf, like."

However, there is arguably ellipsis of the finite verb also found in the previous conjoined clause *aethon ni aros lawr yn Whitesands*, so that *a syrffio fel* could be considered to be a finite clause containing the ellipted elements indicated in square brackets in (25):

(25) a [aethon ni] syrffio fel.
and go.1PL.PAST PRON.1PL surf.NONFIN like
"And we surfed, like."

If one interprets the tenets of the MLF more liberally, one could simply label clauses like *a syrffio fel* as a Welsh ML clause based on presumed ellipsed elements, and in doing so the scope of the model would increase considerably.

Under the more conservative interpretation of the MLF which we have assumed, we acknowledge that the model may have a wider range of coverage in pairs of languages which not only have contrasting word order but also relatively complex verb morphology in the language chosen as ML. This is the case, for example, in the Swahili-English data exemplified in Myers-Scotton (1993). On the other hand, pairs of languages with limited morphology and similar order will be less easy to deal with, as demonstrated by Wang (2017) in relation to Southern Min and Mandarin as spoken in Taiwan. Wang was obliged to reintroduce a morpheme frequency criterion to identify the ML in clauses where the MOP and SMP turned out not to be sufficient. The morpheme frequency criterion was described by Myers-Scotton

(1993) (and abandoned in Myers-Scotton 1997) but whereas Myers-Scotton applied it to the whole discourse, Wang applies it to individual clauses. We did not take this approach with our Welsh-English data, but it would have been another way of increasing the scope of the MLF for our data.

In summary, the application of the MLF model to our Welsh-English data shows that the model has considerable explanatory power in describing the code-switching patterns found. Its principles are not as effectively testable on non-finite or verbless clauses but, as suggested, its scope could be further extended as suggested above by assuming that null elements in a clause are the result of ellipsis, or by (re-)introducing the criterion of morpheme frequency.

Summary of Chapter 5

This chapter has demonstrated that code-switching in Welsh-English data is highly homogeneous. Using the MLF model, which identifies the language which is the source of the grammatical matrix of a bilingual clause, we have shown that Welsh is almost always the Matrix Language of clauses produced by Welsh-English bilinguals. This means that the word order of Welsh is usually used, regardless of the language of the content words used, and that Welsh finite verb morphology is prevalent, a finding which suggests that Welsh is not undergoing radical morphosyntactic change (cf. Davies & Deuchar, 2009) We shall return to this theme in Chapter 7.

CHAPTER 6

Code-switching and independent variables

Introduction

As described in Chapters 2 and 3, the corpus collection process included the administration of a detailed background questionnaire which provided information about the characteristics of our speakers, including their age, gender, occupation, education, age of acquisition of Welsh and English, self-reported proficiency, and a number of other factors or variables. Such information has been increasingly used over the last few decades to study whether or how variables like these are related to linguistic behaviour, an area of research known as variationist sociolinguistics (cf. Tagliamonte 2012).

Review of the literature

The quantitative study of the relation between extralinguistic and linguistic variables was pioneered by Labov from the 1960s onwards. Some of his best-known work is included in his 1972 text *Sociolinguistic Patterns* which includes, for example, his famous study of the relation between the social class of speakers and the pronunciation of postvocalic (r).

Although Labov's early work focused on variation in English, Poplack (1980) pioneered extension of the variationist method to bilingual code-switching data collected among Puerto Rican Spanish-English bilinguals in New York. Her dependent linguistic variable was type of code-switching ('intrasentential' vs. 'extrasentential') and her extralinguistic factors included reported bilingual ability, sex of speaker, age of acquisition of second language, social network membership, ethnic identity and workplace. Some of these factors seemed to be related to type of code-switching when considered separately, but in order to exclude the possibility that these effects were caused by the factors being correlated with one another, she adopted multivariate analysis of the kind advocated by Labov (see e.g. Sankoff & Labov, 1979). The results of this multivariate analysis showed that four extralinguistic factors contributed to the occurrence of intrasentential code-switching: speaker sex, age of L2 acquisition, language dominance and workplace. Specifically, women, those who acquired L2 early, those who were balanced bilinguals and those who worked within the city block which she labels 'El Barrio' were more likely to

produce intrasentential code-switching than men, those who acquired L2 later, or were dominant in Spanish, or who worked away from El Barrio.

Speaker sex or gender has been shown to be an important variable in studies of variation in English, some of which are argued to provide evidence for a "sociolinguistic gender pattern" (see e.g. Fasold, 1990, Labov, 1990) in which women tend to use more standard English than men. Given this finding, Cheshire & Gardner-Chloros (1998) investigated whether, in line with Poplack's results, women might code-switching less than men "in order to conform with a more purist or socially acceptable speech style" (Cheshire & Gardner-Chloros, 1998, p. 14). However, they were able to find little evidence for this pattern, reporting for example that Treffers-Daller (1992) had found no significant difference between men's and women's use of intrasentential switching and that Gardner-Chloros (1992) had found no significant difference in the switching rates of male and female Cypriot Greek-English bilingual speakers. Overall, they conclude that "although a consistent pattern of sex differentiation is assumed to exist in [language use in] monolingual communities, there is no evidence of any consistent patterning of this kind in bilingual communities" (Cheshire & Gardner-Chloros, 1998, p. 28).

Treffers-Daller (1992), in a study of Dutch-French code-switching in Brussels, found that local background, language of education, self-rated proficiency in each language and degree of puristic attitudes were all significant predictors of intraclausal code-switching, although there was some interaction between local background and language of education. The code-switching of speakers over the age of 60 was compared with those under 60, and though no significant difference was found, Treffers-Daller reports a "trend that older informants switch more within sentences than younger informants" (Treffers-Daller, 1992, p. 148). She suggests that intraclausal code-switching is actually disappearing in Brussels owing to the influence of purism in Dutch.

In monolingual variation studies the role of the age of speaker is important not only as a social variable comparable to age, but also because of the 'apparent time paradigm' (cf. Bailey, 2002), according to which the speech of younger speakers may be indicative of language change. The same considerations apply to the study of bilingual speech, but in the case of bilingualism the age of acquisition of the two languages is also an important variable. For monolinguals the age of acquisition is from birth in normal cases, but bilingual acquisition can involve a time lag between the acquisition of one language and another. If this time lag is several years, with the second language being learned later than in early childhood, then there is normally an effect on proficiency, and the bilingual speaker may not be 'balanced' or equally fluent in the two languages. Poplack (1980) found that speaker age was not significant in predicting code-switching, but age of acquisition of English was. Speakers who had been born in the USA or arrived there in early childhood acquired English

at a younger age than those who arrived in adulthood, and were found to produce significantly more intrasentential code-switching, presumably because of their fluency in the two languages.

The study by Backus (1996) of Turkish-Dutch code-switching provides some information about the effect of age of acquisition of the two languages in an immigrant context in the Netherlands. Backus classifies his speakers into three groups based on their age of arrival in the Netherlands. Those belonging to the "first generation" arrived in the Netherlands and so were first exposed to Dutch when they were older than 12, the "intermediate generation" arrived at between 5 and 12 years old, and the "second generation" were either born in the Netherlands or were under 5 at the age of arrival. He found different patterns of code-switching in the three groups. The first generation generally produced Dutch insertions within a Turkish morphosyntactic framework, while the intermediate generation produced frequent interclausal[44] code-switching as well as the same type of intraclausal code-switching as the first generation. The second generation produced mostly interclausal code-switching with infrequent intraclausal switching in which either language could provide the morphosyntactic frame. While the three groups doubtless differed from one another in their patterns of acquisition, we do not have sufficient detail about the bilingual acquisition of the second generation to determine whether they acquired Turkish in the home first and Dutch later, or whether they acquired both Turkish and Dutch simultaneously from birth.

Nortier (1990) focuses on the role of language proficiency in code-switching between Dutch and Moroccan Arabic in the Netherlands. She reports on a sophisticated method of measuring bilingual proficiency in Dutch and Moroccan Arabic which uses self-report of the speaker's preferred language, calculation of the percentage of speaker time taken up by each language, native speaker judgment of proficiency in each language, and error analysis of the Dutch in the corpus. She finds that "a high degree of bilingual proficiency is related to the use of relatively many intrasentential and single word switches" (Nortier, 1990, p. 115), supporting Poplack's findings on the significance of proficiency

Finlayson, Calteaux, & Myers-Scotton (1998) also examined the effect of proficiency on code-switching in their quantitative study comparing the code-switching of two groups of South African township residents who had been educated to Grade 10 or above versus Grade 9 or below. The more highly educated group were assumed to be more proficient in English, and their insertion of English phrases versus single words into either Zulu or Sotho was compared with the other group. Their results showed that the first group produced English insertions consisting of significantly

44. See Chapter 1 for exemplification of the terms 'interclausal' and 'intraclausal'.

more phrases than the second group, which tended to produce single word insertions. As Finlayson et al. (1998, p. 415) state, "The more educated group would be expected to be more proficient in English…and therefore more able to produce well-formed constituents in English". They also show that the more educated group produced more monolingual clauses in English.

A similar finding is reported by Adalar & Tagliamonte (1998), who compared the bilingual speech of two generations of Turkish-English bilinguals who were living in northern Cyprus but who had previously lived in London. The first generation was born in Cyprus and acquired English on emigration to England in their teens, whereas the second generation was born in England and acquired both languages from birth. In a comparison of the insertion of single English words vs. multiple English word stretches into otherwise Turkish speech, the authors found that the second generation used more multiple-word stretches, or what Myers-Scotton (2002) would call 'Embedded Language islands'. Adalar and Tagliamonte's findings regarding the greater tendency of the fluent English-speaking second generation to produce multiple English word insertions directly parallels Finlayson et al.'s (1998) findings regarding a similar tendency on the part of the more fluent and well-educated English speakers in their study.

The studies reviewed above have demonstrated in particular the effect on code-switching of proficiency and/or a related measure, the effect of the age of acquisition of the second language in the case of immigrants whose heritage language is different from the language of the country in which they settle. However, no previous study seems to have compared the effect of the age of acquisition of both languages in simultaneous and successive bilinguals. Deuchar, Donnelly, & Piercy (2016), summarised below, not only report on this aspect, but base their results on the entire *Siarad* corpus.

Few studies since Poplack's (1980) seem to have used multivariate analysis to establish the independent effect of extralinguistic factors on code-switching, possibly because of the work involved in transcribing and coding an entire corpus before one can even embark on analysis. Automatic processing of the data, however, brings rigorous statistical analysis of large corpora within the bounds of feasibility.

The limitation imposed by manual analysis conducted in our earlier work on *Siarad* meant that the coded subsamples of the corpus which resulted were not large enough for multivariate analysis (cf. Parafita Couto, Davies, Carter, & Deuchar 2014). So in this earlier work we instead conducted comparative analyses of code-switching patterns in samples from our three bilingual corpora collected in Wales, Miami (USA) and Patagonia (Argentina) and attempted to account for the differences in terms of community-specific factors (see Carter, Deuchar, Davies, & Parafita Couto 2011). They found that whereas in Wales (using the *Siarad* data) the matrix language was predominantly Welsh as reported in Chapter 5, it varied more

between English and Spanish in the Miami data. They suggested that this might be due to the contrasting word order of Welsh and English (VSO vs. SVO) compared with the similar word order of English and Spanish (SVO). In other words, speakers may have a preference for maintaining a relatively uniform word order in their conversations. Carter et al. (2011) also identified the matrix language of each bilingual clause, and found that this was most uniform in the sample from Wales, where 100% of the clauses had Welsh as the matrix language. Carter et al. (2011) also suggested that two extralinguistic variables (the commonest language of the speaker's social network, and self-perceived identity) might be relevant to the choice of matrix language. In Wales, speakers tended to have a mainly Welsh-speaking social network (see Figure 2.10) and mainly perceive themselves as Welsh (see Figure 2.12). These characteristics are consistent with their overwhelming choice of Welsh as a matrix language. In contrast, the tendency of Spanish-English speakers in Miami to have a more bilingual social network and more diverse perceptions of their identity may be related to their more diverse choice of Spanish and English as alternative matrix languages. In Patagonia the relation between these factors and the matrix language was unclear, possibly because of the small number of Welsh speakers in that community.

Study by Deuchar et al. of factors influencing intraclausal code-switching

In Chapter 3 we described the development of a system of automatic glossing, and it was this that made it possible for Deuchar, Donnelly, & Piercy (2016) to extract the data needed for a multivariate analysis of the whole corpus. They aimed to examine to what extent speaker characteristics were related to the use of intraclausal code-switching. In particular, the study focused on the relation between the speakers' varying patterns of bilingual acquisition, speaker age and the use of intraclausal code-switching.

Method

They focused on the clause as a unit of analysis in order to examine the influence of extralinguistic factors on both intraclausal and interclausal code-switching. Intraclausal code-switching was measured in terms of the number of bilingual clauses as a proportion of the total number of clauses. By definition, intraclausal code-switching can occur only in bilingual clauses. Example (1) below, cited in Chapter 1 as an example of intraclausal code-switching, was coded as bilingual because it contains words from both English and Welsh.

(1) mae o **on standby** [Fusser29-LOI]
be.3S.PRES PRON.3SM on standby
"he's on standby"

Interclausal code-switching, on the other hand, exemplified by (2) below, was measured in terms of the number of finite clauses which involved a change of verb language from the previous clause. *Bod* ('to be') clauses were treated as finite in line with Borsley et al. 2007. This limitation to finite and *bod* clauses for the analysis of interclausal code-switching was necessary because the parsing algorithms used are at an early stage of development, and currently lack the ability to characterise non-finite verb forms in detail. Example (2) consists of two consecutive monolingual clauses, the first in Welsh and the second in English. Since the two clauses are in two different languages, there is an interclausal switch from the first to the second clause.

(2) [*so* bosib hwnna (y)dy o] [**I don't**
so maybe that be.3S.PRES PRON.3SM I don't
know]. [Fusser25-HUN]
know
"So possibly that's it, I don't know."

Words tagged as indeterminate (see Chapter 1), which are part of the lexical stock of both languages, were not counted as code-switches.

Each clause was coded as either monolingual or bilingual, as described in Chapter 5. This allowed us to quantify the amount of intra-clausal code-switching by speakers in terms of its presence (in bilingual clauses) versus absence (in monolingual clauses). The categories 'bilingual clause' and 'monolingual clause' mentioned in Chapter 5 were treated as variants of a variable labelled 'linguality': bilingual ('biling'), monolingual Welsh ('monoW') or monolingual English ('monoE') – examples illustrating this coding are given in Table 6.1, which also shows the language of the verb ('verblg'), whether Welsh ('cym') or English ('eng').

Table 6.1 Examples showing the linguality of extracted clauses

File-name	Utterance ID	Speaker	Clause	Linguality	Verblg
fusser17	1257	AET	oedd o yn dechrau diflannu	monoW	cym
fusser25	148	HUN	**because they're leaving**	monoE	eng
robert2	267	RIS	achos mae gynna chdi **spellchecker** Cymraeg arno fo	biling	cym
lloyd1	720	GRG	in Cymru we recycle	biling	eng

67,515[45] automatically extracted monolingual and bilingual clauses from 148 speakers were used for this analysis. This sample excluded data from the two speakers who had neither Welsh nor English as their first language and the data from a third speaker whose questionnaire data was missing information on first language acquired. 11,601 clauses consisting of only one word had also been excluded as at least two words are needed to study code-switching. Table 6.2 indicates how the clauses were distributed in terms of their linguality. The number of clauses per speaker ranged from 47 to 1,106, with a mean of 456. As can be seen in the table, the majority of the clauses (88%) are monolingual Welsh and a very small fraction (2%) are monolingual English. Bilingual clauses (which contain intraclausal switches) make up 10% of all the clauses, however. Of the monolingual clauses, 97% are Welsh, so the vast majority.

Table 6.2 Linguality of clauses analysed

Linguality	Number	Per cent
monolingual Welsh	59,152	88
monolingual English	1,656	2
bilingual	6,707	10
Total	67,515	100

(Adapted from Deuchar, Donnelly, & Piercy, 2016, p. 227, Table 8.3)

In order to analyse the data Deuchar et al. (2016) used Rbrul (Johnson, 2009). This is a new version of the variable rule program originally developed in the 1970s and used in Poplack's (1980) study. As Johnson (2009, p. 359) explains: "A variable rule program evaluates the effects of multiple factors on a binary linguistic 'choice' – the presence or absence of an element, or any phenomenon treated as an alternation between two variants. The factors can be internal (linguistic), such as phonological or syntactic environment, or external (social), for example, speaker gender or social class. The program identifies which factors significantly affect the response variable of interest, in what direction, and to what degree". The earliest variable rule program was Varbrul (cf. Cedergren & Sankoff 1974), with the more recent Goldvarb (cf. Sankoff, Tagliamonte, & Smith 2005) being the most widely-used currently. According to Johnson (2009), however, Goldvarb suffers from the limitation that it treats each linguistic token as if it were an independent observation, whereas in

45. The figure of 67,515 here differs from that of 66,428 in Chapter 5 because the latter excludes clauses where no identification of the ML is possible on the basis of the verb, whereas the analysis reported in this chapter included clauses both with and without verbs and covered a slightly different number of speakers.

fact linguistic tokens produced by the same speaker should not really be treated as independent. Rbrul overcomes this limitation by using mixed effects modelling to distinguish between 'fixed' or external effects like gender or age and the random effects of individual speakers.

The dependent (response) variable in the analysis reported in Deuchar et al. was the 'linguality' of each clause: whether it was bilingual or monolingual.

Results

The results relating to pattern of bilingual acquisition are shown in Table 6.3.

Table 6.3 Mixed effects logistic regression predicting bilingual clauses with speaker as a random effect

Pattern of bilingual acquisition	Number of clauses	% of bilingual clauses	Centred factor weight	Log-odds
Both Welsh and English from birth	15572	14.7	0.6	0.407
L2 by age four	19006	10.3	0.487	−0.053
L2 at primary school	26501	7.8	0.478	−0.087
L2 at secondary school	3710	6.6	0.485	−0.059
L2 in adulthood	2726	5.6	0.448	−0.209

(Adapted from Deuchar et al., 2016, p. 229, Table 8.4)

The results in Table 6.3 are laid out so as to make them accessible to both Goldvarb and Rbrul users. Goldvarb uses the term 'factor groups' to refer to extralinguistic variables like pattern of bilingual acquisition, a non-continuous (categorical) variable which contains five 'factors' (categories): 'Both Welsh and English from birth' 'L2 by age four', 'L2 at primary school', 'L2 at secondary school' and 'L2 in adulthood'. Rbrul uses the terminology 'factors' instead of 'factor groups' and 'levels' instead of 'factors' (Johnson, 2009, p. 361). The multivariate analysis selects those factor groups which have a significant effect on the dependent variable, and computes a 'factor weight' for each factor, which indicates the probability of the dependent variable occurring in that context. A factor weight (probability) of 0.5 indicates that there is only a 50/50 chance of the factor influencing the dependent variable, so only factor weights above 0.5 allow the factor in question to be considered as having an effect. In the table above, the acquisition pattern "Both Welsh and English from birth" has a factor weight of 0.6, and can therefore be interpreted as influencing the dependent variable: the production of bilingual clauses. However, the factor weights for the four other patterns of acquisition are below 0.5 and hence are not considered to have an effect on the production of bilingual clauses.

The log-odds derive from the Rbrul rather than the Goldvarb version of the variable rule program, and, like factor weights, indicate probabilities. Johnson (2009, p. 361) explains how log-odds values are calculated and provides a table showing that they correspond directly to factor weights. Tagliamonte (2012, p. 141) summarises the difference between the two types of value as follows: "Goldvarb factor weights range between 0 and 1, while log-odds can take on any positive or negative value from negative infinity to positive infinity and are anchored around zero". As Tagliamonte points out, zero is neutral for log-odds values while a positive value indicates a favouring effect and a negative value a disfavouring effect. In the table above we can see that only the factor 'Both Welsh and English from birth' has a positive log-odds value. The other four all have negative values, but (with the possible exception of 'L2 from adulthood') are very close to the neutral value of zero.

Deuchar et al. (2016) also found a negative correlation between age and the production of bilingual clauses. This is reflected in Figure 6.1, which shows that as speaker age increases, the percentage of bilingual clauses produced decreases. Deuchar et al. suggest that this is evidence for an ongoing change in language norms, supporting the idea that code-switching is becoming more common and acceptable, at least in informal contexts. This interpretation might be questioned if an interaction between age and pattern of bilingual acquisition existed, such that younger speakers were more likely to have acquired the two languages simultaneously than older speakers. However, no such interaction was found, suggesting that age of speaker has an independent effect on code-switching.

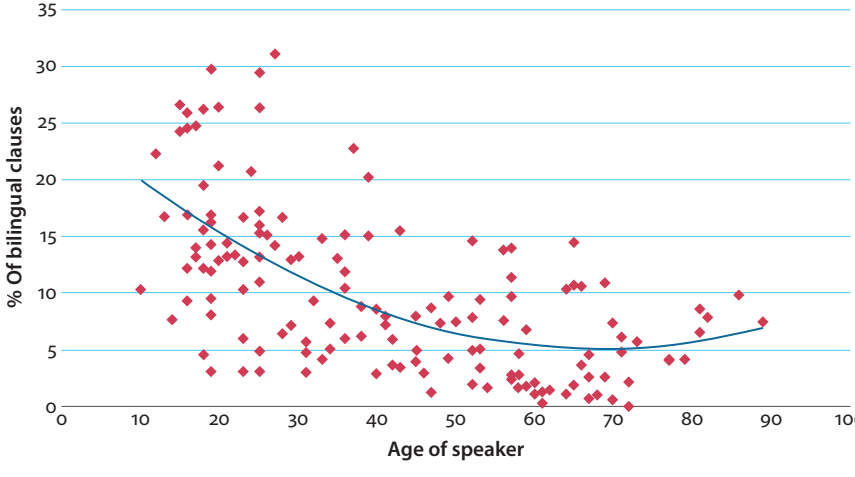


Figure 6.1 Percentage of bilingual clauses by speaker age (Based on Figure 8.2 of Deuchar et al., 2016, p. 229)

From their presentation of the quantitative results shown in Table 6.3 and Figure 6.1 above, Deuchar et al. (2016) went on to consider whether the varying patterns of bilingual acquisition outlined might lead to qualitative as well as quantitative differences in code-switching. In order to do this, they divided the bilingual clauses into two types: single-word insertions and multi-word insertions. As has been indicated in Chapter 5, most clauses coded as bilingual consisted of a Welsh morphosyntactic frame with the insertion of one or more English words. Single word insertions were defined as being single words in otherwise monolingual Welsh clauses as in Example (3).

(3) ti (e)rioed yn **serious**. [Davies6-DAN]
PRON.2S never PRT serious
"You're never serious."

Some investigators consider single-word insertions to be problematic because they might be borrowings (see Chapter 4 for discussion of this issue), so the additional analysis to be described allowed Deuchar et al. to see whether the patterns they found were similar or different with multiword insertions as compared to single word insertions. Example (4) shows a multiword insertion:

(4) dylet ti fod yn gallu gwrando
should.2S.CONDIT PRON.2S be.NONFIN PRT can.NONFIN listen.NONFIN
(ar)no fe *top* **to bottom**. [Davies9-MOS]
on.3SM PRON.3SM top to bottom
"You should be able to listen to it top to bottom."

Deuchar et al. (2016) compared the use of clauses with single words and clauses with multiword insertions in the speech of three groups of speakers: those who acquired English and Welsh simultaneously, those who acquired English first, and those who acquired Welsh first. The results in Figure 6.2 show that single-word insertions are used more frequently than multi-word insertions by all groups, but that speakers who learned both English and Welsh simultaneously use more of both types of insertions.

As can be seen in Figure 6.2, all speakers use more single word than multiword insertions, but those who acquired both Welsh and English simultaneously use more of both. These results are similar to those presented in Table 6.3. In both analyses, early simultaneous bilingualism is shown to favour code-switching, and this applies to both single-word and multi-word insertions. Single-word insertions are certainly more frequent in the data (71% compared to 29% according to Deuchar et al.) but there is a correlation between single-word and multi-word insertions, as shown in Figure 6.3. The correlation means that speakers who use more single-word insertions also use more multi-word insertions.

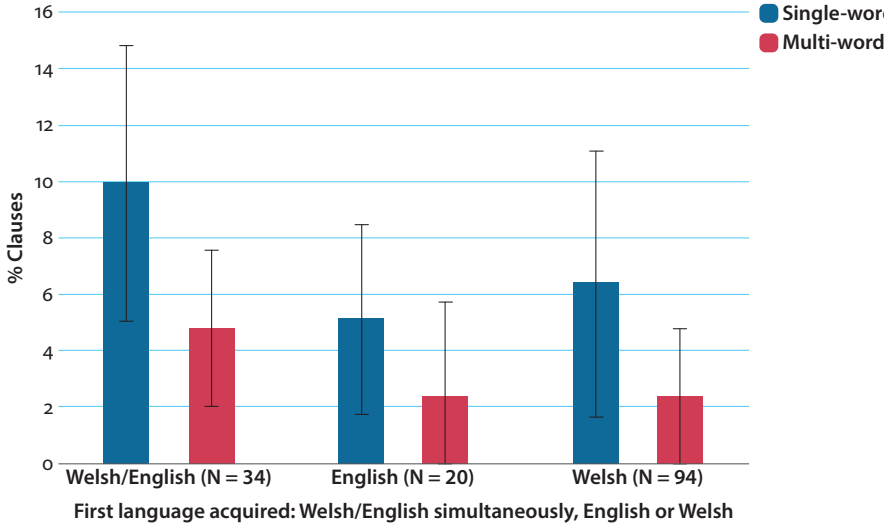


Figure 6.2 Single-word vs. multi-word insertions by first language acquired (Based on Figure 8.3 of Deuchar et al., 2016, p. 231)

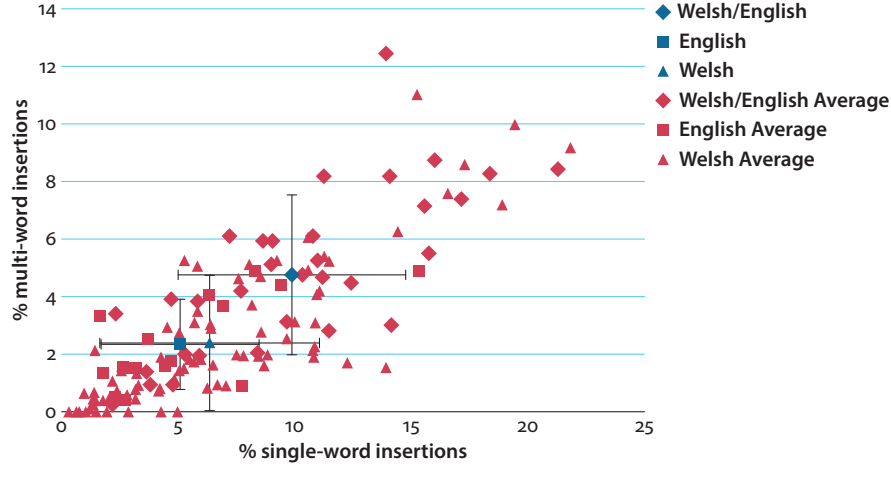


Figure 6.3 Correlation between single-word and multi-word insertions (Based on Figure 8.4 of Deuchar et al., 2016, p. 232)

Study of factors influencing interclausal code-switching[46]

As we reported above, Poplack (1980) found that intraclausal switching (termed 'intrasentential switching' by her) was favoured by balanced bilinguals. This finding has sometimes been interpreted to suggest that 'interclausal switching' (termed 'intersentential' by Poplack) was not associated with early acquisition of L2 or with balanced bilingualism. For example, Berk-Seligson (1986, p. 314) reports that Poplack's results "have shown that frequent intrasentential code-switching is associated with high bilingual ability, whereas use of intersentential switching is associated with non-fluency or dominance in one language over the other". (See also Bentahila & Davies (1995, p. 77) for a similar statement.) However, a careful perusal of Poplack's comparison of the switch types produced by Spanish-dominant versus (balanced) bilingual speakers (see Poplack, 1980, p. 607, Figure 2) shows that the balanced bilinguals not only produced a higher percentage of intrasentential switching than the Spanish-dominant speakers, but also appear from Poplack's Figure 2 to have produced a higher proportion of switches consisting of full sentences. This suggests that balanced bilinguals may produce not only more intrasentential switches than less balanced bilinguals, but also more intersentential switches consisting of full sentences. Indeed, Poplack states that "those who claim to be bilingual…..favour large amounts of the switches hypothesized to require most knowledge of both languages, sentential[47] and intra-sentential switches" (Poplack 1980, p. 606). Treffers-Daller's (1994) results are consistent with this statement since she finds a correlation between intraclausal and interclausal code-switching in her study of code-switching between French and Dutch. This is not surprising if both are related to balanced bilingual proficiency.

Until now we have paid little attention to interclausal code-switching in the *Siarad* data. In fact, interclausal switching is much less frequent than intraclausal switching, both in absolute and relative terms, as Table 6.4 shows. Whereas intraclausal switching occurred in 10% of eligible clauses, interclausal switching occurred only in 2%. In fact over a third of our speakers did not do any interclausal switching at all.

Table 6.4 Intraclausal and interclausal switching in *Siarad*

Type of switch	Eligible clauses	Number showing switching
Intraclausal	67,515	6,707 (10%)
Interclausal	39,704	674 (2%)

46. This study was done in collaboration with Caroline Piercy, and we thank her for her work on the statistical analysis.

47. Poplack uses this term to refer to switching between sentences, also known as 'intersentential' switching. We use the term 'interclausal' instead.

In Table 6.4 the clauses eligible for consideration of intraclausal code-switching include all monolingual and bilingual clauses enumerated in Table 6.2 above. For interclausal switching, eligible clauses consist of all two-clause sequences where each clause contains a finite verb. The number is lower than that for intraclausal switching, because in many cases a finite clause is followed by a non-finite clause.

As a follow-up to the study by Deuchar et al. (2016) on intraclausal code-switching, we investigated whether interclausal switching in *Siarad* was favoured by some of the same factors as intraclausal code-switching. The results showed that despite its relative infrequency for most speakers, interclausal code-switching did pattern with speaker attributes in the same manner as the factors that affected intraclausal code-switching. Both pattern of bilingual acquisition and age were significant. Figure 6.4 shows the percentages of interclausal code-switching produced by each of three groups categorised according to first language acquired. (N = 34 Welsh/English, N = 20 English, N = 94 Welsh).

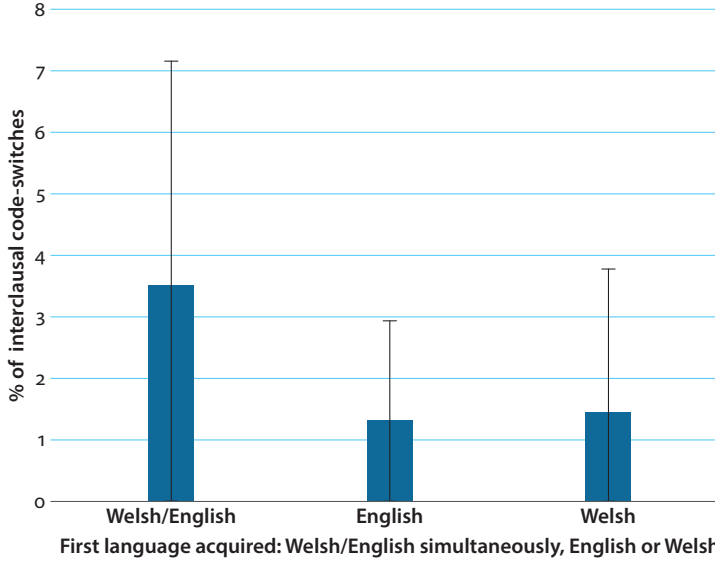


Figure 6.4 Percentage of interclausal code-switches by first language acquired

Two-tailed t-tests assuming equal variance showed that those that had learned both languages simultaneously used significantly more interclausal switches than those that learnt English ($p = 0.01$) or Welsh ($p < 0.001$) first. Those that had learnt Welsh or English first were not significantly different from each other ($p = 0.81$).

Figure 6.5 shows the relation between percentage of interclausal switching and age, which showed a weak negative correlation ($r = -0.23$, $p < 0.01$). This suggests

that younger speakers are more likely to produce not only intraclausal but also interclausal code-switching. A regression line has been included to guide the eye. (N = 148; mean = 1.89%).

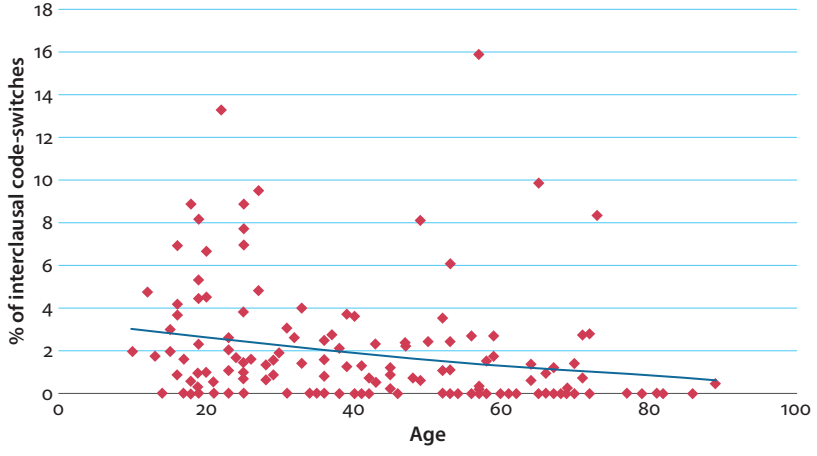


Figure 6.5 Percentage of interclausal switches by age of the speaker

Discussion

In this chapter we have so far reviewed previous work examining the relation between extralinguistic factors and code-switching, and we have presented our own findings which suggest that both intraclausal and interclausal code-switching are favoured by early simultaneous acquisition of both Welsh and English. There has recently been considerable research on varying patterns of bilingual acquisition, and a distinction is usually drawn between simultaneous and successive bilingual acquisition (see e.g. Meisel 2004). There is evidence that these different patterns of bilingual acquisition have different outcomes: for example Meisel (2010) found that sequential German-French bilinguals who had begun acquiring French at age 3 in Hamburg produced errors in the production of French finite verb forms even after six years of exposure to the language, whereas errors of this type were virtually never produced by simultaneous German-French bilinguals. What we know little about as yet is how contrasting patterns of bilingual acquisition affect the language usage of these bilinguals as adults, but our findings suggest that early simultaneous acquisition may indeed contribute to proficiency in both languages and facilitate code-switching.

As yet we know little about how facilitation of code-switching may work, but Weber-Fox & Neville (1999) explored how the age of acquisition of a second language affects the neural subsystems involved in language processing. The participants in their study were Chinese-English bilinguals who had acquired English at five different age categories similar to those used in our study. ERPs (Event-related potentials) elicited by phrase structure violations showed "increased bilateral distribution with increased second language immersion" (Weber-Fox & Neville, 1999, p. 30). These and some behavioural results showing slower syntactic processing with increased age of second language acquisition led the authors to conclude that "the development of at least some neural subsystems for language processing is constrained by maturational changes, even in early childhood" (Weber-Fox & Neville, 1999, p. 36). This conclusion suggests that the timing of bilingual acquisition may indeed affect that facility with which speakers switch back and forth between two languages with different syntactic structures, and thus the frequency with which they will choose to code-switch.

Summary of Chapter 6

Our multivariate analysis of 67,515 bilingual and monolingual clauses from 40 hours of Welsh-English conversational data collected from 148 speakers showed that intraclausal code-switching was produced more frequently by those who had acquired Welsh and English in infancy than those who had acquired the two languages sequentially. A similar pattern was found for the much less frequent interclausal code-switching. We suggested that recent developments in research showing how the timing of bilingual acquisition could affect brain structure might help to explain our results. We also found a tendency for younger speakers to code-switch more than older speakers, and suggested that there is a change in progress related to more permissive attitudes to code-switching.

As we have emphasized throughout this book, our corpus is relatively large compared with some of those used in previous research. We suggest that this allows considerable confidence in our results.

CHAPTER 7

Change in Welsh grammar

So far in this book we have discussed how the *Siarad* corpus can be used to describe the nature of the informal speech patterns of Welsh-English bilinguals, and the results we have reported on up to now have been primarily synchronic, since the corpus itself was collected over a short time-span and is thus itself a synchronic representation of Welsh at the beginning of the 21st century. However, the speech of Welsh-English bilinguals is not static, and in particular it is relevant to consider to what extent Welsh is undergoing language change, particularly, perhaps, considering the potentially strong influence that English exerts as a socially-dominant language in most of Wales. How can the *Siarad* corpus help us in identifying what language change may be happening in Welsh, and how can we analyse its data in order to predict the future of the Welsh language, including its grammar?

In this chapter we review three studies based on using the *Siarad* corpus to study change in various aspects of Welsh grammar. In the next section we will review Deuchar & Davies (2009) and consider how code-switching (see Chapter 5) can be analysed to predict language change in the Welsh context. We will then focus on the concept of grammatical convergence in discussing the findings of Davies and Deuchar (2010). In that section we will also present a new automated analysis to look for signs of convergence in our code-switching data. Finally, we will turn from code-switching to look at a different aspect of Welsh grammar by reviewing Davies & Deuchar (2014), which argues that auxiliary verb deletion in informal Welsh can be viewed as morphosyntactic change in progress.

Code-switching and language change

In Chapter 5 we described an analysis of code-switching data from *Siarad* using the Matrix Language Frame model (Myers-Scotton 2002) as a theoretical framework. Building on the method and process presented in Davies (2010) and Davies and Deuchar (2010), we undertook an automated analysis of the matrix language distribution of clauses from the whole *Siarad* corpus. Our main finding was that Welsh-English code-switching in the *Siarad* corpus is highly homogeneous, with speakers using a Welsh matrix language in the majority of clauses produced, even when those clauses are bilingual (i.e. contain code-switching). However, there were

a small number of exceptions to this trend which were not discussed in detail in Chapter 5 – namely rarely occurring clauses where the two principles of the MLF model point to either language as being the source of the grammatical frame of those clauses. While clearly they were very infrequent in our data, we argue that such clauses act as clues to language change, and thus reflect a possible link between code-switching and change.

Deuchar & Davies (2009) approached the question of the role of code-switching and bilingualism in word-order change by using Welsh-English code-switching as a test case. They examined the future of Welsh from a linguistic perspective, building on existing theories of grammatical change relating to code-switching, with the aim of seeing whether or not the Welsh language is undergoing change, and whether there is any sign that English rather than Welsh might be used as the grammatical frame of bilingual utterances. Davies and Deuchar point to work by Welsh linguist Alan Thomas (1982), who suggested that code-switching – at least in the Welsh-English context – could be indicative of language shift. Language shift is the social process whereby a language is used with decreasing frequency in certain linguistic domains, leading to reduced proficiency among the speaker population, and potentially leading ultimately to the death of that language.

Thomas (1982, p. 218) sees code-switching as a possible indicator of language decay and death in two stages. The first stage is when a speaker code-switches not only in informal contexts (which all speakers may do) but also in formal contexts, which "more accomplished speakers" would not do. At the first stage only vocabulary is affected by means of the insertion of English words, but a second stage would be indicated if speakers should adopt English grammar as well as vocabulary. Thomason and Kaufman (1988, p. 102) comment on Thomas's ideas, suggesting that the type of speech described by Thomas could indeed be interpreted as "first steps" on the continuum of language death.

Deuchar & Davies (2009) inferred from Thomas' ideas the following three-stage framework linking code-switching and language death:

I. Speakers exhibit Welsh grammar in their speech but they use some English words or phrases.
II. Speakers exhibit English grammar as well as Welsh grammar in their speech.
III. Speakers exhibit only English grammar.

The transition between the second and third stages would involve speakers no longer using Welsh grammar in their speech, and so the third stage would follow the death of Welsh as a spoken language, at least for those speakers.

Deuchar & Davies (2009) pointed out that Thomas' ideas had coincidental parallels with a process called "matrix language turnover" proposed by Myers-Scotton (1998), relying on her MLF model (discussed in Chapter 5). According to Myers-

Scotton, matrix language turnover involves code-switching that is not of the "classic" type (see Chapter 5 for a definition), but instead the ML consists of structure from both participating languages – what she calls a "composite ML" – and is an indicator of structural change and language decay or death. Deuchar & Davies (2009, pp. 22–23) interpret Myers-Scotton's view of the three stages in matrix language turnover as follows:

I. "Classic" code-switching, with Language A as matrix language and Language B as embedded language.
II. Composite code-switching, with a composite of Languages A and Language B as matrix language.
III. Monolingualism in Language B follows a shift or "turnover" to Language B as matrix language.

Composite code-switching (the second stage), according to this model, could be taken as a sign that ML turnover is in progress. A completed turnover (the third stage) would mean that Language B has taken over from Language A as the predominant ML in the speech community. Myers-Scotton (1998) proposes that this could, accompanied by the requisite social changes of language shift, then result in speakers abandoning Language A entirely (even as an EL) and become monolinguals in Language B. A community where ML turnover does not appear to be in progress – i.e. a stable situation – would usually exhibit just "classic" code-switching.

Myers-Scotton's model and Thomas' implied model share some striking similarities. First, both models highlight the significance of code-switching as a potential indicator of language shift. Second, they both claim that major grammatical change, in the form of a change in the grammatical frame for sentences produced by speakers, features in situations of language shift. Third, they both posit an intermediary stage in such a change, where speakers use grammar from both the waning and the waxing language when speaking. These similarities were pointed out in Deuchar & Davies (2009) and were a spur to test the predictions of both authors for the case of Welsh by analysing code-switching data from the *Siarad* corpus to see whether or not the grammatical frames used by speakers indicated that Welsh was undergoing language shift and/or ML turnover.

As described in Chapter 5, Deuchar & Davies (2009) used the Matrix Language Frame (MLF) model as a basis for their analysis. They also proposed a new category of matrix language (ML) – "dichotomous" – to identify clauses which did not adhere to "classic" code-switching (i.e. the kind the MLF model focuses on in Myers-Scotton's work). A dichotomous ML is where the two principles of the model, the Morpheme Order Principle and the System Morpheme Principle, each point to a different language as providing the source for the ML. It builds on the concept of a "composite" ML, a term which Myers-Scotton (1993, 2002, etc.) uses to describe a

situation where the clause appears to contain some structure from both participating languages, as in the following sentence shown in (1), from Schmitt's (2000) study of Russian-English child language data (the English word is in bold).

(1) odin byl **pitcher**. (from Schmitt, 2000, p. 24)
one be.3SM.PAST pitcher
"one was a pitcher."

The SMP would point to this clause as having a Russian ML on the basis of the form of the finite verb *byl*. However, if Russian were the ML, then the SMP would suggest that the English morpheme *pitcher* should carry the appropriate Russian instrumental case marking (*pitcher-om*) required by the verb *byl*, yet in (8) *pitcher* is a bare form. Schmitt accounts for the data by suggesting that in this case of language attrition (Russian is the child's L1 but he has moved to the USA), "the level of activation of the EL rises so that it competes with the ML in projecting the frame and…some of the grammatical frame is projected by information in the EL lemma" (Schmitt, 2000, p. 24). In this case, she argues that English is providing the grammatical frame for the noun *pitcher* which as a result is bare as it would be in an equivalent English sentence. In this and other examples she suggests that both Russian and English contribute to providing a composite ML.

Davies & Deuchar (2010) suggested that in a clause containing composite code-switching the MLF model fails to identify a single language as being the ML source, and instead suggests that such a clause has an ML which is a composite of two languages' grammars, perhaps due to some type of language change (e.g. convergence: see below) where the speaker has used structure from both languages instead of one. Davies & Deuchar (2010) formalise the concept of a composite ML (which occurs in bilingual clauses only) by proposing the concept of a dichotomous ML (which can occur in both bilingual and monolingual clauses).

The criteria for identifying a dichotomous ML are as follows. When the MLF model is tested on a clause, if both the SMP and MOP principles point to language A, then language A is the ML. If both principles point to language B, then language B is the ML. If either or both principles point to different languages, then that clause has a dichotomous matrix language. Including an option to identify clauses as having a dichotomous ML when conducting an analysis of a corpus using the MLF model can allow the researcher to identify clauses where a speaker is not producing 'classic' code-switching, and can be a valuable indicator of possible grammatical change in a dataset.

Only one clause with a dichotomous ML (discussed below) was found in the subset of *Siarad* data analysed by Deuchar & Davies (2009). The vast majority (936 out of 986 or 94.93%) of the finite clauses analysed in the data – from the speech of four participants across two conversations – were identified as having a Welsh ML.

The remaining 49 clauses (4.97%) had an English ML, containing exclusively English words and thus no code-switching. It was therefore clear that the four Welsh-English bilinguals analysed produced grammar that fits into the first stage of Thomas' paradigm, where Welsh grammar is primarily used especially when code-switching, and the first stage of Myers-Scotton's model, where speakers prefer Welsh ML and their speech has not undergone ML turnover. Thus in that sample of the data the influence of English grammar on the Welsh of bilinguals is not intense enough for it to be used as the grammatical frame for speakers' clauses. In Chapter 5 we noted that although an analysis of a total of 66,428 clauses identified a few examples of bilingual clauses with English ML, these constituted less than 1% of the data.

Deuchar & Davies (2009) conclude that their results indicate that Welsh-English code-switching is still indicative of the first stage in Myers-Scotton's (1998) model, with Welsh providing the main grammatical frame and some English content words and phrases being inserted. They argue that this stage can be associated with stability, especially if this is favoured by the socio-political circumstances. While these were not favourable at the time Thomas (1982) was writing, and doubtless affected his interpretation, the slight upturn in the number of Welsh speakers demonstrated by the census in 2001[48] together with recent efforts at revitalisation combine to provide a more favourable context for the survival of Welsh.

Code-switching and convergence

The study presented in Davies & Deuchar (2010) extends the analysis presented in Deuchar & Davies (2009), in analysing six speakers from the *Siarad* corpus instead of four. The primary aim of this paper was to consider the link between a dichotomous ML and word-order convergence. The overall results of the matrix language analysis presented in Davies & Deuchar (2010) were essentially as we reported in Chapter 5. Here we will focus on what Davies & Deuchar (2010) say about convergence and grammatical change in the Welsh-English bilingual context. We will first give a summary of the literature on convergence and then describe how much word-order convergence, and what type(s), Davies & Deuchar found in the *Siarad* data, and what this might tell us about language change in Welsh.

We consider convergence to be a form of language change in which languages in contact become structurally more similar to one another (Backus, 2004, p. 179), via the enhancement of structural similarities that exist between those languages

48. This pattern was not maintained in 2011, as the UK Census revealed that the number of Welsh speakers in Wales had decreased slightly since 2001.

(Bullock & Toribio, 2004, p. 91). Toribio (2004, p. 167) describes "the preferential use [by speakers] of some structures over other options" as potentially leading to convergence, and Backus (2005, p. 333) considers convergence to occur when there is a change of distribution in the usage of a pattern in one language because of the influence of another. There is controversy in the literature concerning whether convergence is the result of a process or the process itself (see e.g. Backus 2004). We follow Silva-Corvalán (1994) in assuming it is the result of a process.

Evidence of word-order convergence may be found where morphemes from Language A (L_A) which are expected also to have structure (i.e. word order) from L_A will instead have structure that is more expected of Language B (L_B), but still available in L_A. We will provide examples of such constructions below. In these instances, the L_AL_B bilingual has adjusted the usual L_A word order associated with those L_A morphemes based on the existence of that word order in L_B. Myers-Scotton (2002) refers to a similar phenomenon when she states that a clause in which word-order convergence has occurred will be one where the morphemes come from one language but the grammar includes structure from both participating languages (2002, p. 164).

Several linguistic studies have demonstrated convergence occurring in various languages due to language contact (e.g. Sandalo, 1995; Toribio, 2004; Bullock & Gerfen, 2004; Montrul, 2004; Schmitt, 2001; see Davies, 2010 or Davies & Deuchar, 2010 for a review), and we now review studies which discuss contact-induced change in Welsh, whether or not the change is described as convergence.

Jones (1998) discusses whether or not Welsh is undergoing language obsolescence, noting that "[m]odern spoken Welsh is displaying reduction, simplification, increased linguistic transparency and quite prolific lexical borrowing" (1998, p. 257). Her study examines sociolinguistic variation in two Welsh dialects in Anglicized areas of Wales: Rhymney in south Wales and Rhosllannerchrugog in north-east Wales. Jones draws attention to contact-induced phenomena such as calquing (of which she says there are "numerous instances", particularly in the speech of speakers younger than sixty years old [1998, p. 83]). Crystal (2008, p. 64) defines *calque* as "a type of borrowing, where the morphemic constituents of the borrowed word or phrase are translated item by item into equivalent morphemes in the new language". Jones gives the example of a phrasal verb construction using the non-finite verb *troi* 'turn' plus the preposition *i ffwrdd* 'off' in (2) below, where the historical Welsh form would be *diffodd* 'extinguish'. Jones points to the similarity between such examples and the equivalent English pattern, arguing that the Welsh construction is influenced by the English.

(2) troi e i ffwrdd. (Jones, 1998, p. 83, including her glosses)
VN-to turn him to off
"Turn it off."

Calquing of this sort, Jones (1998, p. 86) argues, is "an indication of language obsolescence", whereby a foreign construction takes the place of a native one over time. Jones argues from comparing the two corpora indicate that there is some degree of language obsolescence and dialect convergence (to what she calls 'Standard Oral Welsh') in spoken Welsh from both north and south Wales (Jones, 1998, p. 289). Of course, interpretations of the implications of convergence may vary, since one could argue that examples like (2) are an indication of language vitality since speakers seem to be inventing knew ways to express concepts in Welsh while making use of structures from English.

Deuchar (2006) examines Welsh-English bilingual utterances which appear to show word-order convergence. An example is given in (3).

(3) fi 'di bod i 'r **bus** lle.
PRON.1S PRT be.NONFIN to DET bus place
"I have been to the bus place." [taken from Deuchar 2006, p. 1996]

This is a bilingual clause in that it contains lexical items from both Welsh and English. All morphemes are Welsh except for the code-switching insertion *bus* from English, but the word order in the NP *bus lle* is not the expected Welsh word order. The Welsh word order for these morphemes would be *lle bus*, where the modifier *bus* follows the head noun *lle*. Instead, the modifier is found preceding the head. Deuchar notes that the morphemes in this clause have a one-to-one correspondence with the equivalent English, and suggests that it could be identified as convergence of Welsh towards English word-order.

The inference from these studies is that there is some evidence for word-order convergence in Welsh, but it is seemingly restricted to certain constructions (e.g. phrasal verbs, word order within NPs). Whilst English therefore appears to have occasional influence on Welsh word order, Welsh is the usual source of morpho-syntax for the clauses produced by the speakers these studies analysed analysed, whether or not they contain code-switching.

Davies & Deuchar (2010) reported on only one clause as having a dichotomous ML, shown in (4) below.

(4) roedd drws-nesa # pobl yn wneud # **sloe** *gin* [Fusser6-AMR]
be.3S.IMP next-door people PRT make.NONFIN sloe gin.
"The next-door people made sloe gin."

Applying the System Morpheme Principle to this clause shows that the verbal morphology of *roedd* is from Welsh, and so this principle would identify Welsh as the ML. The relative position of the verb *roedd* and the subject *drws-nesa pobl* 'next-door people', where verb precedes subject, is typically Welsh. However, the order of the modifier *drws-nesa* 'next-door' preceding the head noun *pobl* 'people' reflects an

order more associated with English (which is almost always modifier-head order) than Welsh (which is typically head-modifier order). The Morpheme Order Principle, in this case, cannot identify just one language as the source of the ML, since the verb and subject order point to Welsh and the head and modifier order point to English. This lack of 'agreement' between the two principles means that this clause can be labelled as having a dichotomous ML. (This classification depends on assuming that *drws-nesa* 'next door' modifies the noun rather than being an adverbial referring to the location of the action.)

Davies (2010) discusses an additional clause with a dichotomous ML, shown in (5). The speaker is describing the type of dress she wants to buy.

(5) **and** sydd ddim yn *Cinderella* **type** # peth. [Fusser27-MAB]
and be.PRES.REL NEG PRT Cinderella type thing
"And which isn't a Cinderella type thing."

This is a bilingual clause – the words *and* and *type* are English[49] while the other words are Welsh; *Cinderella* is a proper noun and therefore we do not consider it to be language-specific. The finite verb *sydd*, a present relative form of bod 'be', is Welsh, so the SMP indicates Welsh as being the source of the ML. There is no overt subject in the clause, so the verb/subject word-order criterion cannot be tested. The word order within the noun phrase *Cinderella type peth* 'Cinderella type thing' is modifier-head, since the two modifiers *Cinderella* and *type* precede the head noun *peth* 'thing'. This word order would suggest an English ML according to the MOP. Since the SMP points to Welsh but the MOP points to English as supplying clause word order, Davies (2010) identified this clause as having a dichotomous ML, since there is 'disagreement' between the two principles of the MLF.

The main interpretation by Davies and Deuchar of the word-order in clause (4) and by Davies of (5) is that they are examples of convergence to English. As noted in Chapter 5, modifier-head order is *available* in Welsh, but is *limited* to certain prenominal adjectives (e.g. *hen* 'old', *prif* 'prime, main') or poetic usage. Modifier-head order is the norm in English, on the other hand, apart from a few infrequent exceptions.[50] The dichotomous ML clause in (4) has the speaker using Welsh lexical items (*drws-nesa* and *pobl*) with morphosyntax more expected of English (modifier-head);

49. Note that while *type* has historically been borrowed into Welsh as *teip* [teip], here the speaker pronounces it [taip], as if it were the English word, so we have classified it as a code-switch rather than a borrowing here.

50. Some exceptions exist, such as *court martial*, *queen consort*, *battle royale* – many of which are due to the influence of the French grammar of the original borrowing – but this word order is not frequent in English speech otherwise.

the clause in (5) has the speaker using lexical items from both Welsh (*peth* 'thing') and English (*type*) but with word order more associated with English (modifier-head again). These can be classified as convergence in the same vein as the examples cited at the beginning of this section. Of course, we could interpret just two clauses showing convergence out of the speech of six speakers as a speech error of some kind, and indeed we found no other examples of this type of convergence in either of the two speakers' speech. Almost all of the data analysed fit the definition of "classic" code-switching rather than composite code-switching, and word-order convergence is very infrequent. Davies & Deuchar conclude that this is a sign of continuity rather than change in Welsh grammar from a morphosyntactic point of view, according to the aspects of morphosyntax that can be identified using the MLF model.

The analyses presented in the reviewed studies by Deuchar and Davies are particularly revealing as indicative of the homogeneity of the code-switching produced by the speakers they analysed. However, as described in Chapter 5, the analysis was limited – in terms of the number of speakers and clauses analysed – by the fact that it had to be done manually. As such, while the analyses are presumed by the authors to be representative of the language in the corpus as a whole, a very large proportion of the speech in *Siarad* was not represented in the analyses reported in the above studies. A search described in Davies (2010, pp. 220–222) of the whole of *Siarad* found only twelve clauses with a dichotomous ML across the speech of all 151 participants; this search was based on listening to the corpus and noting down examples as they were heard. Thus, while the analyses presented in Deuchar & Davies (2009), Davies & Deuchar (2010) and Davies (2010) suggest that dichotomous ML clauses – and therefore convergence – are very rare in *Siarad*, how sure can we be that the small number of examples in the samples mentioned represents the true frequency in the corpus as a whole?

One major benefit of a fully glossed and tagged corpus like *Siarad* (which was not available to the researchers when they were preparing the earlier analyses) is that it is possible to make automated analyses of the whole corpus in much less time than it takes to do a manual analysis. As reported in Chapter 5, we have since performed an automated analysis of the whole of the *Siarad* data, identifying via verbal morphology the ML of every finite clause in the data. However, using verbal morphology alone is not a detailed enough analysis of a clause for the purposes of identifying whether or not it has a dichotomous ML, since the definition of a dichotomous ML in our data is where the language origin of clausal word order does not match the language of the verbal morphology. A second automated analysis was therefore undertaken to indicate the frequency of dichotomous ML clauses in *Siarad*.

Although Examples (4) and (5) above include noun phrases with nouns as modifiers of noun heads, adjectives can of course also modify nouns. In an automatic

analysis, the relative order of head and modifier is easier to determine where the modifier is adjectival, since adjectives, unlike nouns, are unambiguously modifiers when juxtaposed with nouns. For this reason, our automatic analysis focused on sequences containing adjectives and nouns. In order to identify clauses with a dichotomous ML, we searched the entire corpus for sequences of English or Welsh adjectives preceding a noun from a different language in clauses with Welsh finite verbs. The type of sequences we were searching for are illustrated by the following fictitious examples, where the relevant NP is underlined:

i. Clause with Welsh verb and noun phrase with ENGLISH adjectival modifier + WELSH noun head: e.g. *Mi wnes i weld y **ferocious** ci* "I saw the ferocious dog".
ii. Clause with Welsh verb and noun phrase with WELSH adjectival modifier + ENGLISH noun head: e.g. *Mi wnes i weld y ffyrnig **dog*** "I saw the ferocious dog".
iii. Clause with Welsh verb and noun phrase with WELSH adjectival modifier + WELSH noun head: e.g. *Mi wnes i weld y ffyrnig ci* "I saw the ferocious dog".

All of these examples have a dichotomous ML in that the relative order of adjective and noun is as in English rather than Welsh, but the language of the verb is Welsh.

The automated analysis for dichotomous ML clauses found 17 clauses of type (i), 9 of type (ii) and none of type (iii). However, manual checking of the 17 apparently type (i) clauses revealed that only 8 met the requirements of a dichotomous ML in that the finite verb was Welsh while the adjective/noun order was as in English rather than as in Welsh. Five were excluded because they include a prenominal English expletive (i.e. swear-word), which is a normal morphosyntax for such items in Welsh-English code-switching – an example from the data is given in (6).

(6) …fydd off ei **fucking** ben erbyn un-ar-ddeg. [Stammers6-BLW]
be.3s.FUT off POSS.3s fucking head by eleven
"…He will be off his fucking head by eleven."

An additional item (*every blwyddyn* 'every year') was excluded because the Welsh equivalent of *every* appears prenominally as in English. One further item was excluded because there was no finite verb to potentially clash with the word order, and two items were excluded because the adjective appearing before the noun was not modifying that noun but had another function. Of the apparently type (ii) items, all were excluded on the grounds either that the Welsh prenominal adjective (*hen* 'old') regularly occurred in that position before not only English but also Welsh nouns, or that there was no finite verb. The 8 mixed type (i) NPs (all from clauses with a Welsh finite verb) which provide evidence that the clause has a dichotomous ML are listed in Table 7.1 below.

Table 7.1 Noun phrases from *Siarad* which suggest a dichotomous ML

Recording	Noun phrase	English translation
Davies5	whole byd	'whole world'
Davies7	massive tŷ[51]	'massive house'
Fusser25	subsidiary ysgol	'subsidiary school'
Fusser27	honorary fath	'honorary sort'
Robert3	actual waliau	'actual walls'
Roberts3	raw peth	'raw thing'
Stammers6	numerous pobl	'numerous people'
Stammers7	actual cwestiynau	'actual questions'

For illustration, here are some of the above examples in the context of the clause in which they appear. The relevant NP is underlined in each example.

(7) a wnes i gael paned efo **numerous** pobl
and do.1S.PAST PRON.1S get.NONFIN cuppa with numerous people
yn cerdded pasio lawr fan 'na. [Stammers6-BLW]
PRT walk.NONFIN pass.NONFIN down place there
"And I had a cuppa with numerous people walking down past there."

(8) timod fath â mae # fath â **subsidiary** # ysgol
know.2S kind with be.3S.PRES kind with subsidiary school
neu… [Fusser25-ALB]
or
"You know, like… it's… like a subsidiary… school or…"

The automatic analysis was successful in finding some examples of dichotomous ML which had not been found previously, but both the manual analysis reported by Davies (2010) and the automated analysis agree in their conclusion that the frequency of dichotomous ML clauses relative to the overall number of clauses in *Siarad* is very small. While challenges remain in finding a comprehensive and reliable way of identifying this kind of convergence in a large corpus of data, the indication is that word-order convergence of this type is rare in Welsh-English speech.

51. This noun phrase co-occurred only with a tag *yndy* 'it is', and no other finite verb.

Auxiliary deletion in the speech of Welsh-English bilinguals

We now turn to consider auxiliary verb deletion (AuxD), a feature of colloquial Welsh which we identified while building the *Siarad* corpus and which has implications for the future of Welsh grammar. Davies & Deuchar (2014) describe how auxiliary deletion is increasing in frequency and discuss the cause of this change in progress including the relative role of internal and external factors.

Periphrastic constructions in Welsh normally require a finite auxiliary verb, which is often an inflected form of the verb *bod* 'be'. An example of a periphrastic construction is given in (9).

(9) a wedyn wyt ti mynd i chwith. [Davies11-OWA]
and then be.2S.PRES PRON.2S go.NONFIN to left
"And then you go to the left."

(Note that this speaker omits the expected aspectual particle *yn* before the non-finite verb *mynd*, but that is not of immediate relevance here.) In (9) the auxiliary is *wyt*, a 2nd person singular present tense form of *bod* 'be', preceding the subject *ti* 'you'. Davies and Deuchar (2014), following e.g. Jones & Thomas (1977), King (2016) and Borsley et al. (2007), point out that in spoken Welsh the construction can also be found *without* this auxiliary, as in (10).

(10) ti 'n jocian. [Davies6-DAN]
PRON.2S PRT joke.NONFIN
"You're joking."
[form with auxiliary: WYT ti'n jocian.]

There is no apparent semantic difference between (10) here, which uses an auxiliary (which Davies & Deuchar label +A), and (9) above, which has a deleted auxiliary (–A), although the +A form can be considered more formal (cf. Jones & Thomas, 1977). Thus the distinction between the two forms seems primarily stylistic, meaning that the –A form might be expected to be common in *Siarad*, since it is a corpus of informal speech.

Davies & Deuchar, who focus on constructions where the auxiliary is a 2nd person singular present tense form of *bod* 'be', *wyt*, find many examples of –A clauses in different kinds of constructions, like the types seen in Examples (11) through (13).

(11) ti 'di siarad ybyty fo gyda
PRON.2S PRT.PAST talk.NONFIN about PRON.3SM with
fi. [Fusser27-LIS]
PRON.1S
"You've talked about him with me."
[+A version: WYT ti 'di siarad ybyty fo gyda fi.]

(12) ti 'n gwybod lle mae hwnna? [Stammers5-SIO]
PRON.2S PRT know.NONFIN place be.3S.PRES that one
"Do you know where that is?"
[+A version: WYT ti'n gwybod lle mae hwnna?]

(13) ti 'm yn meddwl 'sa fo 'n byd
PRON.2S NEG PRT think.NONFIN be.3S.CONDIT PRON.3SM PRT world
boring iawn? [Fusser19-OLW]
boring very
"Don't you think it would be a very boring world?"
[+A version: (D)WYT ti'm yn meddwl 'sa fo'n byd **boring** iawn?]

They also find some examples of + A clauses, where *wyt* is not deleted, as in Examples (14) and (15).

(14) wyt ti 'n **agree**-io efo be maen nhw
be.2S.PRES PRON.2S PRT agree-VBZ with what be.3PL.PRES PRON.3PL
'n trio wneud? [Fusser19-OLW]
prt try.NONFIN do.NONFIN
"Do you agree with what they're trying to do?"

(15) a wedyn wyt ti mynd i chwith. [Davies11-OWA]
and then be.2S.PRES PRON.2S go.NONFIN to left
"And then you go to the left."

Analysing a subset of 28 speakers' speech from *Siarad* (643 clauses where *wyt* is the auxiliary and where the auxiliary is either retained or deleted), Davies & Deuchar find that speakers overwhelmingly delete the auxiliary (–A): 598 (93%) of the clauses analysed lacked an overt auxiliary. The speakers analysed ranged in age from 12 to 81 (mean age was 43). Participants were born in different areas of Wales, although the majority (18 out of 28) came from North Wales. 14 speakers were male and 14 were female.

Davies & Deuchar also analysed age variation in the frequency of AuxD. Speakers were divided into age bands (10–19 years old, 20–21 etc. up to 60 plus). Figure 7.1 shows how the proportion of auxiliaries deleted varied according to the age band of the speaker and illustrates the report by Davies & Deuchar (2014, p. 236) that "there is a sharp increase in the production of overt auxiliaries between the 40–49 and 50–59 age groups".

A Chi-square test of the raw figures showed that there was a significant relationship between age and AuxD at the 5% level (χ^2 =13.817, df = 5, p = .017; Cramer's V = .147). Post-hoc statistical tests[52] we have performed since the publication of that paper have highlighted more specifically where the significant differences

52. We are grateful to Sarah Cooper (Bangor University) for her advice and assistance with the statistical analysis discussed here.

between the age groups lie. Accordingly, the adjusted residual values for each age band shown in Table 7.2 along with the raw figures show how far from the expected number each actual number lies. The asterisked values for the 40–49 and the 50–59 age group indicate that the former group used significantly fewer auxiliaries than expected whereas the latter group used significantly more. This supports Davies and Deuchar's observation cited above. Although the adjusted residuals for the 60–69 age group do not show a significant difference from those of the other groups, this may because the data for this group are sparse.

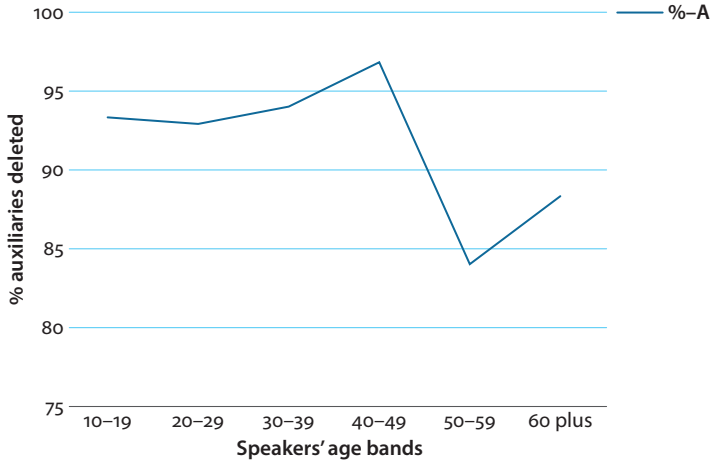


Figure 7.1 Percentage of auxiliaries deleted by age band of speaker

Table 7.2 Distribution of auxiliary presence (+A) and absence (–A) (N = 643) by age band.

Age band	10–19			20–29			30–39			40–49			50–59			60+		
	no.	%	AJ	no.	%	AJ	no.	%	AJ	no.	%	AJ	no.	%	AJ	no.	%	AJ
+A	5	6.67	–.1	14	7.04	.0	7	5.98	–.5	5	3.13	–2.2*	12	16	3.3*	2	11.76	.8
–A	70	93.33	.1	185	92.96	.0	110	94.02	.5	155	96.88	2.2*	63	84	–3.3*	15	88.24	–.8
Total	75	100		199	100		117	100		160	100		75	100		17	100	

(Adapted from Davies & Deuchar, 2014, p. 236, Table 6)

Nevertheless, the overall pattern is still that all age groups delete *wyt* with high frequency in Davies & Deuchar's data.

Davies & Deuchar point out that age variation can be a synchronic indicator of diachronic change via apparent time (e.g. Bailey, Wikle, Tillery, & Sand, 1991; Labov, 2001; Chambers, Trudgill, & Schilling-Estes, 2004), where differences in

the speech of different generations of speakers is taken as being indicative of the language having changed across those generations, since, as Chambers (2003) suggests, once an individual's speech patterns are established during young adulthood, "those features remain relatively stable for the rest of their lives" (2003, pp. 202–203). Davies & Deuchar (2014) interpret their findings that older speakers delete the auxiliary *wyt* in 2nd person singular *bod* constructions less frequently than younger speakers as a sign that a change is in progress in Welsh whereby deleting the auxiliary is becoming more frequent. Based on these data, Davies & Deuchar (2014) suggested that AuxD in Welsh may have first appeared during the early 20th century, but new research by Willis (2016) finds that AuxD can actually be found in texts from 1850, in particular in some Welsh translations of Harriet Beecher Stowe's novel *Uncle Tom's Cabin*. In several of these texts AuxD is used in the translated speech of the black slave characters to represent their distinct speech variety. Willis points to AuxD being employed by Welsh authors from the 19th through to the 20th century to reflect how non-fluent second-language learners of Welsh were perceived to speak, which Willis argues indicates that it was viewed as change coming from outside of Welsh, with AuxD later spreading to the speech of native Welsh speakers. Thus the change identified in the *Siarad* data seems to be the tail end of a change long in the making, which began in the 19th century. This is reflected by the high frequency of –A in the speech of the over 50s in our data (more than four in five clauses show –A in that age group overall), and that the low value of Cramer's V (.147) for the statistical analysis indicates that the difference in –A frequency according to age is statistically significant but weak. This is further indication that deletion of *wyt* was already an established feature of Welsh when the speakers aged 50 and over acquired the language in the 1940s and 1950s, even though it has become more frequent in the intervening generations.

In trying to explain why AuxD occurs Davies & Deuchar draw on a combination of both internal and contact-induced explanations – what Thomason & Kaufman (1988) call "multiple causation", where there is more than one contributory factor in a given change. First the authors draw attention to the similarity between a –A clause which has an initial subject pronoun *ti* 'you' and normal English main clause word-order (SVO): in both cases the subject is the first overt element. Davies & Deuchar suggest that this could be indicative of a trend in Welsh for speakers to prefer S-initial clauses under the influence of English word order. Since Welsh already has S-initial constructions for focus, this could be explained as convergence (see the discussion of the analysis in the previous section), where speakers are more likely to delete the auxiliary to produce a S-initial clause since such a construction is already available in Welsh. A focus-initial construction from *Siarad* is shown in (16) below, where *mam* 'Mum' is the focus of the construction:

(16) *mam* mynd yn **nuts** efo fi. [Davies15-NEL]
mum go.NONFIN PRT nuts with PRON.1S
"Mum went nuts with me."

Davies & Deuchar's quantitative analysis only considered constructions where 2nd person singular *wyt* was the auxiliary, but AuxD has been identified as occurring with some first and third person subjects too, albeit primarily in southern Welsh dialectal speech as opposed to in northern varieties (see Borsley et al., 2007). An example of a –A clause where the subject is 1st person singular is shown in (17), where the speaker is discussing what word she uses instead of the Welsh words for 'toes'.

(17) fi *just* yn dweud '**toes**' [Fusser27-LIS]
PRON.1S just PRT say.NONFIN toes
"I just say 'toes'."
[+A version: dw i *just* yn dweud **'toes'**.
be.1S.PRES PRON.1S just PRT say.NONFIN toes
"I just say 'toes'."]

In the version with the first person auxiliary *dw*, note that the first person pronoun *i* (an alternative to *fi*, cf. King, 2016, p. 106) is used by most of our *Siarad* speakers rather than *fi*. Davies & Deuchar, following Jones (2004), note that in both northern and southern dialects it is precisely those auxiliaries beginning with a vowel which tend to be deleted. Furthermore, one could point out that another Welsh auxiliary, *gwneud* 'do' does not appear to be subject to deletion. This could be because it begins with a consonant or perhaps because of its functional load as a past tense marker, whereas deletion of forms of the auxiliary *bod* occur in the present tense, which could be considered the default. However, even if phonological factors are involved in triggering the deletion of bod as auxiliary, the fact remains that the resulting constructions approximate English in their order, and Davies & Deuchar argue that this similarity to English acts as a catalyst with the result that the change is "boosted by intensified language contact via bilingualism" (Davies & Deuchar, 2014, p. 239). Thus both internal and external factors can be seen to contribute to change. Note that Davies and Deuchar's (2014) analysis was a manual analysis of the speech of 28 speakers. There are 151 speakers in *Siarad*, so it would be beneficial to perform an analysis of the whole corpus to see if the AuxD pattern found by Davies and Deuchar (2014) is representative of the whole corpus. However, *Siarad* was not transcribed with a gloss tier which represents syntactic sentence properties like null elements, including deleted auxiliary verbs, so a machine analysis to extract such sentences is challenging and time-consuming. Nevertheless, an advantage of *Siarad* is that other researchers can enhance the corpus by adding content in the form of more detailed transcription oriented to specific research questions. This would allow future automated analysis of phenomena like AuxD to be conducted on the *Siarad* data.

Summary of Chapter 7

In this chapter we have suggested, based on our own research using *Siarad*, that internal change, as in the case of auxiliary deletion, may be accelerated by the influence of English structure, so that the change may lead to some convergence in the grammar of the two languages. However, our study of head/modifier order in the search for clauses with dichotomous ML indicated that there is little evidence for convergence in the contrasting Welsh and English word order inside noun phrases.

We have also shown how a searchable corpus can provide evidence for our conclusions, and how automatic and manual analysis can complement one another. Although we have focused here on our own research, we shall show in the next chapter how *Siarad* has also provided a resource for other researchers.

CHAPTER 8

Additional research using *Siarad*

Introduction

In this chapter we provide an overview of the use that has already been made of the *Siarad* corpus and which we have not mentioned previously in this book. Some of this research has been done by us and/or our colleagues, whereas other studies have been conducted by researchers from outside our group. We will focus first on work which has made use of the corpus and furthers our understanding of code-switching, an important motivation for collecting our corpus in the first place. We will then we will go on to consider how it has also been used as a resource for the study of various aspects of the Welsh language, including phonetics and phonology, morphosyntax, grammar and acquisition.

Quantitative analysis of code-switching

In Chapter 6 we described how the quantity of code-switching produced by our speakers was calculated as a proportion of the total finite clauses that were bilingual (with material from both languages) as opposed to monolingual (material from one language only). We were able to relate this to specific external variables to determine their effect on the quantity of code-switching. Lloyd (2008) had a similar aim and used data from 121 of the speakers in the *Siarad* corpus. All these speakers had been brought up in North Wales. Instead of focusing on the clause as a unit of analysis she chose the word, and quantified code-switching in terms of the percentage of English words used by speakers. As described in Chapters 5 and 6, the conversations in *Siarad* are mostly in Welsh, with intraclausal switching being far more frequent than interclausal switching to English, so this alternative method of quantifying code-switching can be justified. As in the results of our study reported in Chapter 6, she found that age had a significant effect, so that younger speakers used more English than older speakers. In addition, medium of education had an effect in that those speakers who had received both their primary and secondary education through the medium of Welsh used more English than those who had received their education through the medium of both languages. Although it may sound paradoxical that those who are exposed to more Welsh in an educational setting may actually insert more English into their Welsh than those who are exposed to less

Welsh, it is less surprising in the light of our results (reported in Chapter 6) that it is those who acquired both languages very early and can be assumed to be proficient in both languages who produce the most code-switching, their proficiency apparently allowing them to go back and forth relatively effortlessly between the two languages.

A similar method to Lloyd's of quantifying code-switching was used by Prys (2016), who used the entire *Siarad* corpus to investigate the effect of the external variables age, gender, educational level attained and attitudes. Only age had a significant effect, and the pattern found closely replicates Lloyd's results in that code-switching peaks in the speech of young adults and then gradually diminishes until it is lowest for speakers in their sixties, rising a little for those in their seventies. It is striking that these results are similar to ours depicted in Figure 6.1, even though our method of measuring code-switching differs. As Figure 6.1 shows, the quantity of switching again peaks in young adults, diminishing over age until there again appears to be an upturn in the quantity used by the oldest speakers. Although these oldest speakers appear to buck the trend of older speakers code-switching less then younger speakers, this may be because of the narrowing[53] of social networks in older age, as suggested by Chambers and Trudgill (1980, p. 79). Prys (2016, p. 276) interprets his graph showing the relation between code-switching and age as an "apparent time profile for code-switching", an approach which is consistent with our suggestion in Chapter 6 that there is a change in progress such that the quantity of code-switching used by speakers is increasing.

Prys (2016) also makes the innovative suggestion that code-switching is a marker of style in the sense that it encodes stylistic as well as social information (cf. Labov 1972). This suggestion is compatible with his general finding that there is less code-switching in the more formal setting of Welsh radio programmes than in the informal *Siarad* conversations, and is supported more specifically by the further finding that code-switching in the radio programmes correlates with indices of (in-) formality which include laughter, overlapping speech and retracing.

Qualitative analysis of code-switching: Conflict sites

A puzzle of great interest to theorists of code-switching since the 1980s is what happens to code-switching at 'conflict sites' in code-switching, i.e. where the grammar of the two languages involved would be in conflict in some way. For example, the relative position of noun and adjective is different in English and Welsh, so an interesting question arises as to which word order a mixed language noun/adjective construction

53. In other words, older speakers may have a smaller number of interlocutors with whom they are relatively familiar, thus encouraging the use of code-switching.

would follow. Early work on code-switching by Poplack (1980) proposed the 'equivalence constraint', according to which code-switching would not occur at conflict sites of this kind, but much subsequent work has shown that this constraint does not always hold (cf. e.g. Bentahila & Davies, 1983; Berk-Seligson, 1986; Nartey, 1982). To pursue this question in relation to Welsh-English code-switching, Parafita Couto, Fusser, & Deuchar (2015) used the *Siarad* data as one of three methodological approaches in their evaluation of competing theoretical accounts of the relative position of noun and adjective at 'conflict sites' in Welsh-English bilingual speech. Mixed Welsh-English nominal constructions including nouns and adjectives are conflict sites in the sense that English grammar leads one to expect placement of the adjective before the noun whereas Welsh grammar leads to the normal expectation that the adjective will follow the noun. The MLF predicts that word order will follow that of the matrix language of the clause, whereas a Minimalist approach pursued by Cantone & MacSwan (2009) in relation to Italian-German code-switching suggested that the language of the adjective should determine the relative word order of adjective and noun. In order to adjudicate between the two approaches in relation to Welsh-English data, Parafita Couto et al. draw on naturally occurring mixed constructions found in *Siarad*, an elicitation task, and an auditory judgment task. Using the automatic glossing system described in Chapter 3, 137 examples of mixed constructions containing an adjective and a noun were extracted from the data. All of these constructions occurred in clauses where the morphosyntactic frame or matrix language was Welsh. In about two-thirds of the data the noun was English and the adjective was Welsh, whereas in the remainder of the data the noun was Welsh and the adjective English. Then, in a semi-experimental 'Director-Matcher task' (cf. Gullberg, Indefrey, & Muysken, 2009), a further 238 mixed nominal constructions were elicited. Again, all turned out to occur in clauses with a Welsh matrix language. The most common combination, as in the *Siarad* corpus, was an English noun with a Welsh adjective, with a less common combination again being Welsh noun and English adjective. The third type of data collection related to acceptability judgments and involved presenting oral stimuli which included some mixed nominal constructions. Interestingly, all examples of mixed constructions containing a noun in one language and an adjective in the other were considered unacceptable, whereas some 'filler' stimuli containing code-switching were considered acceptable. This negative evaluation of mixed noun/adjective constructions may reflect a relative reluctance on the part of speakers to produced mixed constructions which lack syntagmatic congruence in the two languages (cf. Deuchar, 2005), and this may also be reflected in the low number of mixed/noun constructions in the *Siarad* corpus: only 137 in a corpus of about half a million words. Overall the data in the *Siarad* corpus and those elicited semi-experimentally led Parafita Couto et al. to the conclusion that the MLF account of word order in mixed nominal constructions was superior to

the alternative, Minimalist account. This study shows how naturalistic data can be an important component in a study using other methods also.

Recently, neuroscientific methods have also become available to linguists, and another study by Parafita Couto, Boutonnet, Hoshino, Davies, Deuchar, & Thierry (2017) used an innovative approach with event-related potentials (ERPs) to identify whether the MLF or a Minimalist account provided a better account of the acceptability of various adjective/noun orders in mixed Welsh/English nominal constructions. The study provided some further evidence for the greater empirical accuracy of the MLF approach, particularly in the results from contrasting pairs of stimuli where the two competing frameworks made opposite predictions regarding grammaticality. This is the case in Example (1) below. The MLF would predict the sentence in (1) to be grammatical because English provides the morphosyntactic frame of the sentence, and at the same time the relative position of the noun and adjective, i.e. adjective preceding the noun, also follows English grammar.

(1) **The bear chased one** gwyn **horse**.
the bear chased one white horse
"The bear chased one white horse."

On the other hand, the Minimalist account of Cantone & MacSwan (2009), outlined above, would predict the relative order to follow the language of the adjective, Welsh, and therefore for the noun to precede the adjective.

Contrasting predictions from the two frameworks would also be made for Example (2) where the morphosyntactic frame of the sentence is Welsh this time.

(2) Helodd yr arth un **white** ceffyl.
chase.3S.PAST DET bear one white horse
"The bear chased one white horse."

The Minimalist account would predict 8.2 to be grammatical because the English adjective *white* precedes the noun. On the other hand, the MLF would predict it to be ungrammatical. This is because the morphosyntactic frame is Welsh and thus the adjective would be expected to follow the noun, whatever the language of the adjective.

The ERP results provided data based on brain activity which (unbeknown to the participants) suggested that stimuli of type (1) were grammatical whereas those of type (2) were ungrammatical, thus in line with the predictions of the MLF and leading to the implication that the MLF offered a better account of the data. However, for pairs of stimuli where both accounts agreed on their (un)grammaticality, the ERP results did not show a clear difference between the brain activity reacting to the members of such pairs. This led to the suggestion that methodological improvements were needed before definitive conclusions could be drawn about the implications of the ERP data.

Triggers of code-switching

Another avenue which has been explored using the *Siarad* corpus is the question of what causes a speaker to code-switch. Following work by Broersma & De Bot (2006) on code-switching between Moroccan Arabic and Dutch, Carter, Broersma, & Donnelly (2016) explored the idea that lexical items which are cognates in the two languages concerned may actively trigger the occurrence of code-switching. Cognate words are those which share a common origin, and while this is most obvious in languages of the same family (cf. the Germanic cognates *house* in English and *Haus* in German) the term 'cognate' is also applied to words in pairs of languages of different historical origins but where the words have been borrowed from one language to the other. This is the case with many words in Welsh which have been borrowed from English, resulting in cognates. Examples of cognate words in Welsh and English are *siop/shop, car/car*, and *jas/jazz*. Carter et al. used corpus data from *Siarad* to examine the role of cognates in triggering code-switching. The results showed that clauses containing cognates were more likely to contain switches within the same clause than clauses without cognates, and that clauses containing cognates were also more likely to be followed by interclausal switches than clauses without cognates. This suggests that cognate words may act as bridges from one of the bilingual's languages to the other one.

Accommodation in code-switching

Prys, Deuchar, & Roberts (2012) used the *Siarad* corpus to develop a method for measuring accommodation in code-switching which they then applied to data collected from interviews between pharmacists and patients about the use of regularly prescribed medication. The method involved calculating the relative proportion of Welsh and English words in segments of the transcripts and using a formula to measure how much the speaker changed their proportion of Welsh versus English words over time in relation to the proportion being used by the interlocutor. This was piloted using six files from *Siarad* before being used for the analysis of the pharmacist-patient communication. The overall results showed an interesting disparity in accommodation strategies between the pharmacist and his patients. While the pharmacist tended to accommodate by converging in quantity towards the patient's use of code-switching, the extent of accommodation shown by patients towards the pharmacist was variable: patients were almost equally divided between those who tended to converge towards the speech of the pharmacist and those who diverged from it. The ability to accommodate shown by the pharmacist may have useful implications for the training of healthcare professionals.

Welsh language

Since the *Siarad* corpus consists predominantly (84%) of Welsh rather than English (4%)[54] words, and since the vast majority of the clauses produced by speakers are entirely in Welsh, the corpus is a good source not only of data on bilingual speech but also of data on the use of the Welsh language. Several investigators have already drawn on it for their work on the phonetics (Cooper, 2011; Carter & Cooper, 2012) and grammar (Breit, 2012; Willis, 2017) of Welsh, on sociolinguistic variation in morphosyntax (Davies 2016) and for information on the likely nature of the adult input to children (Thomas, Williams, Jones, Davies, & Binks, 2014).

In a study of the acoustic nature of laterals in Welsh, Carter and Cooper (2012) compared lateral consonants in our *Siarad* and Patagonia corpora. They analysed data from ten female speakers producing a total of 425 fricatives and 989 approximants. Their results replicated previous work in that the fricatives were found to be 'clearer' than the approximants, but Patagonian fricatives were 'darker' than those produced in Wales. The distinction between 'clear' and 'dark' laterals refers to the extent of velarisation, or the raising of the back part of the tongue during the articulation of the lateral. Velarised laterals are 'dark' whereas non-velarised laterals are 'clear'. The study by Cooper (2011) went beyond segmental phonetics in investigating turn-taking cues used by Welsh speakers in conversation. Cooper followed French & Local (1983) in distinguishing between competitive and non-competitive turns, the former being turns which attempt to interrupt the current speaker and the latter which do not. Cooper found that, as in conversations in English, competitive turns tended to be characterised by relatively high frequency and high volume.

In addition to the variationist study of code-switching mentioned above, Prys (2016) also examined the effect of the same extralinguistic variables on the mutation of Welsh words in *Siarad*. This analysis specifically excluded English words or words which might be English, since as discussed in Chapter 4, there is evidence that such words may be mutated at a different rate from native Welsh words. Prys considered to what extent the extralinguistic variables of age, gender, educational level attained and attitudes were related to the rate of standard mutation in third person possessive pronouns and following certain mutation triggers such as prepositions. As in the code-switching analysis, age had a significant (but opposite) effect on mutation, such that older speakers mutated more than younger speakers (while younger speakers code-switched more than older speakers). Gender had an effect in only some categories of mutation, whereas the effect of level of education was more general: in almost all categories of mutation more educated speakers tended

54. See <bangortalk.org.uk>

to use more standard variants than less educated speakers. Attitudes, however, did not have a consistent effect on mutation. Prys also discovered that an internal factor had an important effect on mutation: whether or not the word was a place name and how frequent the place name was. For example, the pattern of mutation following locative (*yn*) varied according to whether or not place names where included in the analysis and when they were included, mutation occurred more with more frequent place names. An additional contribution of his study on mutation was to suggest that some mutation triggers act, like code-switching, as stylistic markers (in particular the conjunction (*a*) 'and' (*â/gyda*) 'with') while others are mere indicators in that they provide social information but do not mark style-shifting by individual speakers. The trigger *fy* 'my' is an example of an indicator as it varies according to age but not style.

Turning to studies of the grammar of Welsh, Willis (2017) investigated variation and change in the form of the second person pronoun in northern varieties of Welsh. Data from *Siarad* on the use of the innovative second person pronoun *chdi* is used, together with data collected in a sentence repetition task, to contribute to the development of new methods to identify the role of geographical location in linguistic innovations. Detailed evidence is presented for the importance of the north-western region in the origin and propagation of the innovation.

Breit (2012) is a dissertation focusing on auxiliary deletion in Welsh syntax, and this work is an excellent complement to the work by Davies & Deuchar (2014) reported in Chapter 7. Whereas Davies & Deuchar (2014) focus on phonological and social constraints on auxiliary deletion in the speech of 28 speakers from *Siarad*, including examples of both auxiliary presence and absence, Breit focuses specifically on constructions showing auxiliary deletion in the speech of 143 speakers. He is able to deal with this large amount of data by using the autoglossing system (described in Chapter 3) to extract constructions exhibiting auxiliary deletion. One interesting aspect of his work is that he tests the contrasting predictions of Borsley et al. (2007) and Jones (2004) regarding which types of pronominal subjects occur with auxiliary deletion. He finds that both scholars are correct in predicting deletion to be possible with second person singular subjects, and not possible with third person singular subjects. The implication from Jones (2004)'s work that deletion cannot occur with third person subjects is found not to be supported by the data, however. Table 8.1 below summarises Breit's results regarding his comparison of the predictions with the data in *Siarad*.

In addition to using the corpus data, Breit also conducted an experimental survey of acceptability using oral stimuli, which provided support for Borsley et al.'s predictions in that auxiliary deletion with first person singular turned out to be less acceptable than with first and second person plural.

Table 8.1 Breit's (2012) results on auxiliary deletion and grammatical person/number in *Siarad* corpus compared to predictions from Borsley et al. (2007) and Jones (2004)

Pronominal subjects	Borsley et al. (2007) on auxiliary deletion	Jones (2004) on auxiliary deletion	Found in *Siarad* corpus
1st person singular	limited	limited	yes
2nd person singular	yes	yes	yes
3rd person singular	no	no	no
1st person plural	yes	limited	limited
2nd person plural	yes	yes	limited
3rd person plural	limited	no	limited

(Based on Table 3.2 in Breit, 2012, p. 28).

Davies (2016) used the *Siarad* corpus to investigate the relation between variation in the use of 1st person plural and 3rd person singular possessive constructions in Welsh and the age of the speaker. He found that there was a tendency for younger speakers to use the less traditional or 'literary' forms and instead to use a new form modelled on the genitive construction in Welsh, in which the possessed is followed by the possessor, e.g. *gwaith Bob* ('his work'), which he argues is a sign of change in progress. This innovation was particularly clear in the third person singular, and shows that it is no longer appropriate to assume, following previous authors (e.g. Jones, 1990), that this construction occurs primarily in child speech, but is instead found increasingly in the speech of speakers of all ages.

Thomas, Williams, Jones, Davies, & Binks (2014) is a study of the acquisition of plural morphology in Welsh which focuses on the effect of the quantity and quality of the input. In this study the *Siarad* corpus is used to explore the use of plurals in adults as a guide to the kind of input that children acquiring Welsh will be receiving. Thomas et al. selected 12 recordings from *Siarad* to analyse, totalling 6 hours and 45 minutes overall. These included 27 speakers, all of whom spoke Welsh as their first language and English as their second language. 1,030 instances of plural forms were extracted, and coded according to whether or not the forms followed prescriptive norms. They found that the most frequent type of plural involved the addition of a suffix (most frequently *–au*), while the second most frequent type had an English *–s* ending. They found very few forms that were deviant from prescriptive norms, although 22% of forms had an English-origin *–s* which is nevertheless acceptable in colloquial Welsh. (In fact, the English *–s* plural is listed in the most authoritative Welsh dictionary (*Geiriadur Prifysgol Cymru*) as an English borrowing.) They concluded that "adults' productive command of the plural system in Welsh is extremely uniform across speakers, rendering the quality of input to children consistent and reliable" (Thomas et al., 2014, p. 484). This conclusion was further supported by the results of a plural production task given to Welsh-speaking adults,

in which they found generally uniform production, especially among L1 speakers. Having established a general uniformity in the quality of the input to the children, they administered a plural production task to children from Welsh-speaking, English-speaking and bilingual homes. Their results showed that the children from Welsh-speaking homes clearly outperformed the children from bilingual homes as well as those from English-speaking homes. This suggests that not only is early exposure to Welsh important in achieving adult-like command of plural formation, but also that the quantity of input makes a difference. It may not be enough to only be exposed to some Welsh in early childhood, if the quantity of the Welsh input is limited by simultaneous exposure to English as in the bilingual homes.

The studies by Willis, Breit, and Thomas et al. all illustrate how *Siarad* can be useful in studies of Welsh which combine both corpus-based and experimental methods. The same point can be made in relation to the study of code-switching as illustrated by the study by Parafita Couto et al. (2015) outlined above. As Gullberg, Indefrey, & Muysken (2009, p. 39) conclude in their review of research techniques for the study of code-switching, "There are benefits to be gained from integrated studies that seek to validate experimental methods and data against naturally occurring code-switching".

Summary of Chapter 8

In this chapter we have provided an overview of some of the research that has been conducted using the *Siarad* corpus not only as a resource for work on code-switching, but also for analysis of spoken Welsh.

CHAPTER 9

Conclusion and future directions

This book is divided into two main sections, the first relating to how we built the *Siarad* corpus, and the second to the results of some of the analyses we and others have conducted using the corpus.

Both sections were preceded by an Introduction, in which we expressed the hope that the book would be of interest to researchers in the areas of corpus linguistics and sociolinguistics as well as to people interested in bilingualism, language contact and minority languages.

In the Introduction we also addressed the question of the nature and usefulness of a bilingual corpus. We suggested that extended samples of speech produced by bilinguals in a natural way can help to address various questions about how bilinguals use their two languages. Of course, these samples need to be made accessible to users in the form of transcriptions, and the transcriptions need to be glossed and translated so that they can be used by people who do not necessarily speak both languages involved. We explained that our own particular research question involves asking how bilingual speakers manage to combine their two languages in the same conversation, and whether the patterns of combination follow those that have been described for other speech communities.

We acknowledged our debt to corpus and computational linguistics in making available new tools for analysing data as well as new ways of making the data available to the public. It is gradually becoming the norm to make available the data on which one's analysis is based, something which was either impossible or uncommon in the past. For example, Chiarcos, Hellmann, and Nordhoff (2011) describe the activities of a group working to promote the idea of open linguistic resources and to propose ways of making these resources available.

In our Introduction we also acknowledged pre-existing corpora including both Welsh and English, although we noted that these corpora were often described as consisting only of Welsh. In general, it was clear that there was as yet little spoken data from Welsh-English bilinguals in the public domain, and this led to our collection of a pilot corpus and then the larger *Siarad* corpus.

We also provided an overview of the historical and social context of our corpus. We described how Welsh monolingualism was replaced by majority monolingualism in English during the twentieth century, and how the proportion of the population made up by Welsh-English bilinguals has fallen to its current figure of

about 19% of the population, although the percentage is higher in northern and western areas. However, bilingualism in Welsh and English is now part of government policy, and this is reflected in education.

Although educational practice tends to assume that the two languages should be kept separate, actual usage in informal situations does not exhibit 'code' or language separation, and this is demonstrated by the *Siarad* corpus in which, although 84% of the words are Welsh, 4% are unambiguously English and 13% could be either Welsh or English. A clause-based analysis has shown 90% of the clauses to be monolingual (mainly monolingual Welsh) while 10% of the clauses are bilingual.

Following the Introduction, Part One (Chapters 2 and 3) deals with the details of how we built the corpus. Chapter 2 covers the profiles of the 151 speakers who were recorded, and Chapter 3 with how the data were collected and transcribed.

Chapter 2 uses the questionnaire data we collected from our 151 speakers to describe their general profile. Both Welsh and English versions of the questionnaire as given to participants can be found in Appendix 1. The answers have been coded and recorded in the form of a spreadsheet which can be found at <bangortalk.org.uk>. As indicated in Chapter 2, the 20 questions 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.

About half of our speakers were male and half female, with a roughly equal distribution across the decades until the seventies and eighties, where we have a lower number of speakers. Regarding occupation, these have been recorded verbally rather than with any quantitative coding, but the information will make it possible for other researchers to conduct analyses relating the linguistic data to social class, for example. About three-quarters of our speakers were brought up in north-west Wales, the rest having been brought up north-east Wales, mid or south Wales, or outside Wales. Regarding education, we acknowledge that graduates (about 50% of our participants) are probably over-represented in our sample.

Chapter 2 also provides information about the patterns followed by our speakers in acquiring their two languages. We have seen that about a quarter of our speakers acquired Welsh and English simultaneously, and that where one language was acquired before the other, more speakers acquired Welsh first and English second than English first and Welsh second. Chapter 6 explores the relation between pattern of bilingual acquisition and the production of code-switching.

The results from self-assessment by the speakers of their proficiency in both languages showed that 63% were balanced bilinguals, meaning that they assessed their proficiency in Welsh and English to be at the same level. We suggested that the generally high level of proficiency in Welsh probably reflected exposure to Welsh in

the home. The answers to the questions about language input in the home showed that 89% had indeed been exposed to Welsh at home in childhood and 62% only to Welsh. The high level of Welsh input continued for most speakers, since the majority had their primary education through the medium of Welsh, but fewer than half of the speakers had Welsh-medium education at secondary school. This situation is a cause for concern among those committed to the role of education in the revitalisation of Welsh. Baker & Prys Jones (2000) report that as many as 40% of children educated through the medium of Welsh at primary school move on to English-medium secondary school. Redknap (2006, p. 12) describes what she calls the "language loss" which may occur following the transition from primary to secondary school. She suggests that some children actually become "second language" Welsh speakers when they join secondary school.

The answers to our question on speakers' social networks were entirely compatible with other indicators of the importance of Welsh in the home and in early education: 89% spoke Welsh with their five closest contacts. Although not anticipated when we designed the questionnaire for the collection of the *Siarad* corpus, the answers to this question have proved useful methodologically in evaluating the authenticity of our recordings for another corpus, that collected in Patagonia <bangortalk.org.uk>. This methodological benefit will be described below in connection with an overview of how our experience with the *Siarad* corpus led to further research with other corpora.

Our questions on attitudes to Welsh and English led to the finding that on average, speakers found English more 'useful' etc. (instrumental evaluation) than Welsh, but on the affective dimension ('friendly' etc.) they evaluated Welsh more highly than English.

Regarding self-reported identity, Chapter 2 described how 90% of our participants reported this to be Welsh, and in Chapter 6 we suggested that this homogeneity of identity could help to explain the overwhelming choice of Welsh as matrix language. Because there was so little variation in self-reported identity we were unable to include identity as an extralinguistic factor in the variationist study by Deuchar et al. (2016) reported in Chapter 6.

Finally, the results from self-report of code-switching and attitudes to code-switching suggested that most speakers recognise that they do it and that they are divided as to whether they think it is an acceptable practice.

Chapter 3, the second of the two chapters in Part One of this book, describes the methods we used in collecting and transcribing our corpus. We described how we recruited individual bilingual speakers using letters and posters and then asked them to find a partner with whom they were willing to be recorded during an informal conversation. The method of recording conversations was also described along with the equipment used. We described how we tried to maximise the naturalness

of the speech recorded and thus to resolve the Observer's Paradox, and how we were generally satisfied with the result. We also described our adherence to ethical norms and the details of the consent we obtained from participants.

We explained our reasons for choosing the CHAT system of transcription, which included our plan to make our data public. We provided a brief summary of the features of CHAT, including our use of the main, glossing and translation tiers, and the language marking system used to tag words. We described the automatic glossing system we developed and how it works, using an online dictionary and constraint grammar to produce glosses rapidly. We explained how the transcriptions are linked to the computer sound files, and how we checked them for reliability using a new method. We also provided details in Chapter 3 of where the corpus is hosted, and plans for the future.

Whereas Part One of this book deals with how the corpus was built and what it contains, Part Two deals with how it has been used in research by us and others. Chapter 4 deals with the thorny issue of how code-switching can be differentiated from borrowing. As we explained in Chapter 4, this is a theoretical issue which also has practical implications for how we define the object of our research, code-switching. We review previous work on this issue and show how contrasting approaches have been taken by key scholars such as Poplack and Myers-Scotton. We describe how code-switching and borrowing are distinct phenomena for Poplack whereas they are on a continuum for Myers-Scotton. We then go on to describe how work using *Siarad* has contributed to the resolution of this debate. We report on work by Stammers (2010) and Stammers & Deuchar (2012). They used mutation as a diagnostic test of the degree of integration of English verbs into Welsh and found that this was closely related to the frequency of those verbs. The conclusion from their research was that code-switches are distinguishable from borrowings on the grounds of a low level of both frequency and integration. They argued that their finding was evidence against the nonce borrowing hypothesis of Sankoff et al. (1990) and that the category of nonce borrowings was therefore redundant.

Chapter 5 deals with the grammar of code-switching, and evaluates how far Myers-Scotton's Matrix Language Frame accounts for code-switching in our data. The notion of a matrix language is outlined, and we describe the criteria we have used for identifying the matrix language of a clause in our data, reviewing our earlier work in this process. We show how we have found Welsh to be the predominant source of the matrix language in our data, and for the code-switching patterns in the data to be highly homogeneous. Most clauses are monolingual with a Welsh grammatical frame or matrix language, very few are monolingual English, and virtually all bilingual clauses have a Welsh matrix language. On the basis of a much smaller amount of data showing the same pattern, Deuchar & Davies (2009, reviewed in Chapter 7) argued that code-switching does not necessarily present a threat to the

future of Welsh as others have suggested, but that the overwhelming presence of the Welsh matrix language may be an indicator of stability in the status of Welsh, especially since this is favoured by the current socio-political circumstances.

In Chapter 6 we outline the variationist method of linking characteristics of speakers to the variation in their speech and report on a study which investigates the extralinguistic characteristics influencing code-switching in the *Siarad* corpus. We explain how the development of our automatic glossing system made it possible to extract the data we needed for a multivariate analysis of the whole corpus. The results showed that intraclausal code-switching was more frequent in younger than in older speakers' speech, and also in that of speakers who had acquired Welsh and English simultaneously rather than sequentially.

Chapter 6 also reports that although we found interclausal code-switching to be much less frequent than intraclausal code-switching, nevertheless the same extralinguistic factors were found to be significant. Younger speakers and those who had acquired the two languages simultaneously also produced more interclausal code-switching. We suggested that our findings on the relation between the early simultaneous acquisition of Welsh and English and the production of code-switching complement recent findings from the study of how brain structure is affected by the timing of bilingual acquisition. The fact that verbal fluency and the speed of syntactic processing appear to be enhanced by the early acquisition of an additional language suggested to us that the timing of bilingual acquisition may affect speakers' facility to code-switch and thus its frequency in adult speech.

Chapter 7 is a little different from the others in Part Two of this book in that it does not focus specifically on code-switching, but rather on what implications we can draw from the *Siarad* corpus for the likely future development of Welsh grammar. We focus on two aspects of the data which may be indicative of how the grammar of Welsh could be changing. The first is the small number of clauses with a 'dichotomous matrix language', where it cannot be said unequivocally that the matrix language is Welsh or English. We build on work published earlier (Deuchar & Davies, 2009) by adding an automatic analysis yielding such clauses. We conclude they are so rare that they cannot be interpreted as indicating ongoing change in the most common matrix language, which remains Welsh. Anomalous data in corpora can be difficult to interpret, as it may not be clear whether or not they represent 'performance errors'. However, our view is that constructions or phenomena which are low in frequency (like clauses with dichotomous matrix language) are more likely to be errors than more frequent constructions.

The second type of data we examine in Chapter 7 is clauses where the second person auxiliary appears to have been deleted clause-initially, with the effect that the word order is SV rather than AuxSV. SV is of course more similar to English, an SVO language, and so we raised the question of whether these constructions could

reflect developing convergence towards English-influenced grammar in Welsh. The fact that auxiliary deletion is more common in younger than older speakers led us to suggest that there is a change in progress. We suggested that the change may be due both to internal factors as well as to contact with English

Chapter 8 provides an overview of the use that has already been made of the *Siarad* corpus and which had not already been discussed in this book. The overview included both work which has used *Siarad* to increase our understanding of Welsh-English code-switching and work which has used the corpus to increase our knowledge of how the Welsh language is used.

In relation to code-switching we reviewed the work relating to conflict sites and showed how it provided support for the MLF framework and for the importance of taking the whole clause into account in analyses of the structure of code-switching. We also reviewed evidence for the role of cognates in triggering code-switching, and for how conversational partners may accommodate to one another (or not) in their use of code-switching, as well as work on how code-switching seems to be an indicator of style in Welsh.

In relation to the use of Welsh we covered the work on the topics of phonetics, turntaking, mutation, the use of the pronoun *chdi*, further work on auxiliary deletion, and change in the use of possessive constructions. Finally we reviewed a study of the acquisition of Welsh plural morphology which used *Siarad* as a source of information regarding the probable linguistic input to children.

Future directions

Turning now to the possibilities of future research using *Siarad*, we suggest that further contributions may be made in areas which include code-switching, the study of Welsh, the study of English, and the study of minority languages in general.

Code-switching

In Chapter 5 we briefly discussed criticisms of the MLF model which we have used for the analysis of code-switching in *Siarad*. However, to date we have not come across a satisfactory alternative model to account for the data, but we await the contributions of researchers working in other frameworks such as Minimalism (cf. González-Vilbazo & Lopez, 2011) Word Grammar (cf. Eppler, 2010), or Distributed Morphology (cf. Alexiadou, Lohndal, Åfarli, & Grimstad, 2015).

Another possible avenue for future research on code-switching involving *Siarad* would be to compare the corpus with aspects of other corpora which have

been studied. For example, Fricke, Kroll, & Dussias (2016) used our Miami corpus to investigate whether spontaneous code-switching data support the findings of experimental studies of language switching (e.g. Meuter & Allport, 1999; Costa & Santesteban, 2004) which indicate a processing cost associated with language switching. To address their question they created a dataset consisting of sets of utterances where one member of the set contained code-switching and the others did not. All utterances in each set were produced by the same speaker in the same conversation, in the same (main) language and were the same length in words. The syntactic category of the switched word in the utterance containing code-switching was matched by a non-switched word in a similar category in the utterance not containing code-switching. The investigators then compared speech rate in each code-switched utterance with the comparable utterances in the set not containing code-switching. Their results showed that speech rate slowed down before a code-switch. In discussing the reasons for this, they point out that their results seem to support the idea that "even when highly balanced, proficient codeswitchers retain full control over the choice to switch languages this switch does not come without a cost[55]". It is interesting to consider this suggestion in the light of our findings, reported in Chapter 6, that the speakers in *Siarad* who code-switch the most are also those who acquired the two languages at a very early age and are therefore likely to be highly fluent in these two languages. It would be interesting to replicate Fricke et al.'s study on data from *Siarad*. If code-switching appears to affect speech rate as in the Fricke et al. study of the Miami data, this could explain why such a high degree of proficiency (or early acquisition) of both languages is needed to manage frequent code-switching.

In order to find further evidence that speakers anticipate code-switches, Fricke et al. also measured the VOT (voice onset time) of English words with initial voiceless stops produced in close proximity to Spanish words. Voice onset time is generally shorter in Spanish than in English stop consonants (see Lisker & Abramson, 1964). Fricke et al. found a tendency for English words following Spanish words (but not preceding them) to have shorter VOTs than when they were produced in unilingual utterances. They concluded that "the effect of cross-language phonological activation on English VOT is anticipatory in nature", which was entirely in line with their results on slower speech rate anticipating code-switching. Again, this kind of study could be replicated using *Siarad* and analysing any Welsh-influenced phonetic characteristics of English words inserted in Welsh utterances as opposed

55. They do however mention two alternative explanations for the slowed speech they find: (1) difficulties in lexical retrieval and (2) differences in the prosodic organisation of the two languages.

to occurring in English utterances. VOT differences between Welsh and English (cf. Ball 1984, p. 15 vs. Lisker & Abramson, 1964, p. 394) may not be sufficient to examine cross-linguistic influence of the kind detected by Fricke et al., but other phonetic characteristics could be examined, e.g. the pronunciation of English words such as GOAT with a monophthong [o:] (arguably Welsh-influenced) as opposed to a diphthong [əʊ]. For a study of Welsh-influenced vs. standard pronunciation in Welsh English speech, see Wilson & Deuchar (2017).

An additional focus of research that could be applied to *Siarad* is in the area of structural priming. As described by Fricke & Kootstra (2016, p. 181) "Structural priming in language production typically refers to speakers' tendency to re-use the syntactic structure of recently processed sentences". There has been some work on structural priming in bilingual speech, but until recently based only on experimental tasks and not on spontaneous conversational data. In another analysis of data from our Miami corpus, they find that factors such as code-switching in a particular utterance tend to prime (or facilitate) code-switching in the following utterance. Their results suggest that it might be worthwhile to redesign a variationist study of the kind reported in Chapter 6 so that it incorporates factors specific to the conversation in progress.

Fricke & Kootstra also report evidence for the priming of the matrix language in the Miami data. They found that the selection of a matrix language for a clause was influenced by the matrix language of the previous clause. In our own analyses of *Siarad* (see especially Chapter 5) we have discovered an overwhelmingly uniform pattern of code-switching such that Welsh provides the main grammatical frame and English words and phrases are inserted in this frame. As described above, Carter et al. 2011) found that there was more variability of morphosyntactic frame in the Miami than in the *Siarad* data. Carter et al. pointed out that this greater variability occurred where the two languages involved (Spanish and English) had similar word orders (SVO). Uniformity, on the other hand, occurred with a pair of languages (Welsh and English) with very different word orders (VSO vs. SVO). Interpreting this pattern in terms of syntactic priming we could suggest that the sequential repetition of the same VSO matrix language (Welsh) that we find in the *Siarad* data is an example of speakers tendency to "re-use the syntactic structure of recently processed sentences" (Fricke & Kootstra, 2016, p. 181, quoted above.).

This book has concentrated so far on just two of the following approaches to code-switching outlined by Gardner-Chloros (2009, p. 10) in her textbook: (1) "Sociolinguistic/ethnographic descriptions of code-switching situations"; (2) "Pragmatic/conversation analytic approaches"; (3) "Grammatical analyses of samples of code-switching and the search for underlying rules, models and explanations to explain the patterns found". So far we have not considered the second approach, but the *Siarad* data is highly suitable for conversation analytic approaches, and it

is hoped that other researchers may use it for this purpose. Musk (2010) is an example of a conversation analytic approach to Welsh-English code-switching based on data which he collected from informal discussions by young people attending a bilingual school in Wales. His data lead him to argue, for example, that "When these bilinguals speak Welsh, for the most of the time the language boundary between Welsh and English is only loosely maintained" (Musk, 2010, p. 179). In line with methods established in the approach known as conversation analysis (CA[56]) he takes a participant-oriented approach to the conversation of bilinguals. He does this by distinguishing (Musk, 2010, p. 186) between what he calls *code-mixing* "where code-alternation is the default or neutral choice" and *code-switching* "where code-alternation is used to display additional meaning" as in the use of English for quoting the speech of others. As he nevertheless acknowledges (Musk, 2010, p. 195), the analyst's perspective is also necessary to identify which "*codes* (languages or language varieties") are used at each stage of the conversation. It would be interesting for an approach of this kind to be extended to a larger sample of data such as that available in the *Siarad* corpus. Another aspect of code-switching which has been shown to lend itself to both the participant-oriented approach of CA and an analyst-oriented approach is the phenomenon of flagging, which Poplack (1988) introduced to refer to discourse phenomena which draw attention to a switch made in their proximity (cf. Rosignoli, 2011, p. 1). According to Poplack (1988, p. 230) flagging can include "repetition, hesitation, intonational highlighting, explicit metalinguistic commentary". Example (1) from Fusser26 of *Siarad* includes metalinguistic flagging:

(1) fel mae 'r Sais yn deud, '**ignorance is bliss**'!
like be.3S.PRES DET Englishman PRT say.NONFIN ignorance is bliss
"as the English say, 'ignorance is bliss'!" [Fusser26-IOL]

Before this example occurs in the conversation, IOL has been teasing her husband TEC about the fact that he cannot find things in a kitchen drawer. He responds that she is always moving things, whereupon she suggests in this utterance by saying "ignorance is bliss" that he may prefer not to know where things are. Their conversation is typical of the *Siarad* data in that it is predominantly in Welsh, but she flags the use of this proverb in English by prefacing it with the comment *fel mae'r Sais yn deud ('as the English say')*.

Rosignoli (2011) conducted an analysis of his English-Italian code-switching data from both an analyst-oriented and a participant-oriented perspective. From the participant-oriented perspective (deriving from CA) he shows how the presence

56. See Levinson, 1983, Chapter 6, for a description of CA, and Auer, 1998 for the application of CA to bilingual conversation.

of flagging before a switch indicates a retrieval problem, whereas flagging afterwards indicates whether or not the switch was perceived as suitable in relation to the medium of the conversation tacitly agreed upon. Rosignoli's study reflects the importance of sequence in a CA approach. Then from an analyst-oriented perspective he presents quantitative results demonstrating that there is a correlation between the frequency of an item in the data and its production with or without flagging. Less frequent items, such as adjectives and verbs, are more likely to be flagged than more frequent items such as nouns. Both qualitative and quantitative research of this kind could be fruitfully conducted on the *Siarad* data.

The study of Welsh

Up till now we have made suggestions mostly about future research on code-switching which could be conducted with the use of the *Siarad* corpus. However, as should be clear from our review in Chapter 8 of completed research which has used *Siarad*, the data can contribute to our knowledge regarding the use of Welsh, particularly since the majority of the corpus consists of material in Welsh. Further studies could be conducted like that of auxiliary deletion described in Chapter 7, using a different linguistic variable. Morris (2013) is an example of a study using data collected on the phonological variables (l) and (r), in which he shows the influence of the extralinguistic variables home language, gender, and area of residence on the choice of variants by speakers. Similar studies could be conducted on the *Siarad* data, since the detailed questionnaire data is available.[57] Further work could also be done building on the study by Carter & Cooper (2012) described in Chapter 8 on fricatives and approximants. While they were only able to study the speech of ten female speakers without reference to any other demographic characteristics, a variationist study could draw on all the variables described in Chapter 2, in order to study their influence on phonetic production.

At the time of writing we anticipate that the quantity of easily accessible data on Welsh will be considerably augmented by the work of the publicly funded project *Corpws Cenedlaethol Cymraeg Cyfoes* (CorCenCC[58]), or 'National Corpus of Contemporary Welsh'. The aim is for this to consist of a total of 10 million words, 4 million in speech, 4 million in print, and 2 million 'e-language'. Some of the data will be contributed by crowdsourcing methods.

57. See <bangortalk.org.uk>

58. See <corcencc.org>

The study of English

In addition to the *Siarad* corpus being of use for research on the Welsh language, Leimgruber (2013) has suggested that this kind of corpus may actually be able to contribute to research on world varieties of English by providing evidence for the ways in which English combined with other languages. He argues that our kind of data may help to compensate for the fact that corpora (like the International Corpus of English or ICE) are designed to represent geographical varieties of English but have the drawback that they exclude other languages with which English is in contact. Our analysis of *Siarad* so far has suggested that English is used either as a source of lexical items to be inserted in otherwise Welsh utterances, or where English is used for a complete clause or utterance, this tends to be in quoting the speech of others. However, the function of English insertions in our data could be further explored.

Minority languages

In Chapter 6 we reported on the development of our techniques in automatic glossing to identify the factors favouring the production of code-switching. Our method of autoglossing allowed us to quantify the relative proportion of bilingual and monolingual clauses, and also to identify the language of the finite verb in each clause. Our experience shows that the language of the finite verb is a good indicator of the morphosyntactic frame or matrix language of the clause, and we have shown (Deuchar & Davies, 2009) that this is predominantly Welsh in our data. We interpreted this finding to indicate that the prognosis for the future of Welsh is good, whereas if we had found evidence that English was replacing Welsh as the main matrix language, this would have led to a more pessimistic prognosis. A future research project could develop a diagnostic tool to measure the health of minority languages in comparable corpora to this one. Key measures to be used by this tool would be those already piloted, i.e. calculation of (1) the overall percentage of monolingual and bilingual clauses and (2) the percentage of monolingual and bilingual clauses with a minority vs. majority matrix language. The results of the calculations would be used for the prognosis in relation to the minority language concerned. Specifically, we would expect a positive prognosis where there is a high percentage of monolingual clauses in the minority language, and where the majority of the bilingual clauses have a minority matrix language. Conversely, the prognosis for the survival of the minority language would be poor where the percentage of clauses in the minority language is low and/or there are few bilingual clauses with a minority matrix language.

One good reason for conducting research on minority languages is to discover whether or not they are endangered and need special protection in order to avoid

language or dialect death. Crystal (2000, p. 19) argues that "Some sort of classification of endangerment needs to be made" so that effective action can be prioritised. However, he says that "Comparing levels of endangerment is very difficult, in view of the diversity of language situations around the world, and the lack of theoretical models". He notes that some classifications of endangerment are based on the number of speakers of a minority language (where this is known or can be estimated) while others use additional criteria. He notes Wurm's (1998) definition of endangered languages, for example, as those which "have few or no children learning the language, and the youngest good speakers are young adults" (Crystal (2000, p. 21). In his well-known work on *Reversing Language Shift*, Fishman (1991) outlines a classification of minority languages in terms of the existing threat to their survival. He calls this GIDS (A Graded Typology of Threatened Statuses) and likens it to a Richter scale for earthquakes The typology has eight stages on a scale from 1 (best) to 8 (worst), where 1 represents "some use of Xish in higher level educational, occupational, governmental and media efforts" and 8 corresponds to "most vestigial users of Xish are socially isolated old folks" (Fishman, 1991, p. 87–109).

Fishman discusses the difficulty of collecting suitable data to apply this typology. He suggests that in order to assess whether or not a minority language is endangered, one needs information about language competence, language use and attitudes. However, he says that as it may be time-consuming and expensive to collect such data "there is usually no practical alternative to either collecting self-report data about them via 'scales' or 'questionnaires', on the one hand, or, on the other hand, to letting trustworthy and informed observers report on their impressions as well and as uniformly as they can" (Fishman, 1991, p. 49). However, we would argue that the collection and analysis of a corpus as described in this book tools would in fact be a practical and reasonable alternative to the methods described by Fishman. An early version of our approach, which could be further developed, is exemplified in the work by Deuchar & Davies (2009) as outlined above. Myers-Scotton (1998) had argued that matrix language turnover, or the change of the main morphosyntactic frame from the minority to the majority language, could be an indication of language shift or death, so we concluded that this looked very unlikely in the case of Welsh, at least on the basis of our data. However, work by Wang (2007) on the Tsou language as spoken by Mandarin-Tsou bilinguals in Taiwan, demonstrated that whereas Tsou was the main matrix language in bilingual clauses produced by older speakers, this role had been replaced by Mandarin in the speech of younger people. This was compatible with other indicators of Tsou as an endangered language.

Application of the approach begun by Deuchar & Davies (2009) to a range of minority languages could make it possible to address the following questions:

1. What do the corpus data for minority languages tell us about their prospects for survival?
2. What factors are associated with good vs. poor prospects for survival?

Assuming the availability of data from a minority and majority language similar to that found in *Siarad*, we can suggest the following steps in analysis:

1. Extract all clauses with finite verbs from the data (preferably using an automatic glossing system if available).
2. Code each clause as either monolingual or bilingual,
3. Code the verb of each clause as belonging to either the majority or the minority language (e.g. English or Welsh in Wales). This identifies the matrix language or morphosyntactic frame of the clause
4. Calculate the overall percentage of monolingual and bilingual clauses
5. Calculate the percentage of monolingual and bilingual clauses with a minority vs. majority matrix language.

The results from the above steps could be used to develop a prognosis regarding the survival of the minority language in question. We suggest that this would be a positive prognosis where there is a high percentage of monolingual clauses in the minority language, and where a high percentage of the bilingual clauses have a minority matrix language. Conversely, the prognosis for the survival of a minority language would be poor where the percentage of clauses in the minority language is low and/or there are few bilingual clauses with a minority matrix language.

Summary of Chapter 9

In this chapter we have summarised the main content of this book and have made suggestions for new research using *Siarad* and comparable corpora. We hope that others will take up the challenge and contribute to pushing back the frontiers of our knowledge of bilingual communication and its implications.

APPENDIX 1

Documentation file for the *Siarad* corpus

(also at bangortalk.org.uk)

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.

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*.

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

---

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

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 Webb-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 chronologically). The sound and transcription files for each conversation share the filename, but have different file extensions (*.mp3 and *.cha respectively). For example, Davies3.cha is the transcription of Peredur Webb-Davies' third recording (sound file Davies3.mp3). 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 gender. 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[4] 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 Webb-Davies, Marika Fusser, Sian 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

4. childes.talkbank.org/manuals/CHAT.pdf

independently from the original transcriber of that particular transcript. Transcripts were then compared and a rate of similarity was calculated. The average reliability score[5] 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 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.

---

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

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
syth_bín	syth bín	straight away
ta_waeth	'ta waeth	anyway
un_ai	un ai	either

Table 3.1 (*continued*)

Our transcription	Standard orthography	English equivalent
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[6] 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).

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.

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

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 interjections and interactional markers, e.g. *ah*, *aha*, *oh*, *ooh*, *um* which are found in the reference dictionaries for both languages. 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, pp. 14–20, 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 nonstandard 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 [r]. Otherwise it is transcribed as 'r', even when the standard orthography prescribes 'rh'. Some speakers do not have [r] 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 Welsh colloquial forms which are not 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
(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
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'[7]
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	diciau	tuberculosis	northern variant attested in "diciau" article in GPC

7. See youtube.com/watch?v=Z-x7zLAZdLM

Table 3.2 (*continued*)

Our transcription	Standard equivalent	English equivalent	Comments
(b) Colloquial verb forms			
adnabodais i	adnabyddais i	I recognised	
aethai hi, aethen ni, aethan nhw	ai, aem, aent	she/we/they go	would
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	
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'

(*continued*)

Table 3.2 (*continued*)

Our transcription	Standard equivalent	English equivalent	Comments
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*
(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
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)
FOCUS	item with focus
FUT	future
GB	's – genitive or auxiliary 'be' elision

Table 4.2 (*continued*)

Abbreviation	Representing
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
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

(*continued*)

Table 5.2 (*continued*)

Transcription	Manual gloss
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)	Gender
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
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

(*continued*)

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

APPENDIX 2

List of corpora containing contemporary Welsh language before the collection of the data for *Siarad*

CIG1 (1996)

Type: Spoken
Size: 300k words
Website: <users.aber.ac.uk/bmj/abercld/cronfa18_30/sae/intro.html>; <childes.talkbank.org/browser/index.php?url=Celtic/Welsh>; <cymraeg.org.uk/kig>
Description: Child language acquisition material based on conversations between Welsh children aged 18–30 months and their caregiver or another adult.
Licence: CC-BY-SA[1], GPL3[2]
Downloadable: Yes

SpeechDat Cymru (1998)

Reference: Jones, R. J., Mason, J. S. D., Jones, R. O., Helliker, L., & Pawlewski, M. (1998). SpeechDat Cymru: A large-scale Welsh telephony database. In *Proc. LREC Workshop: Language Resources for European Minority Languages.*
Type: Spoken
Size: Unspecified
Website: <catalog.elra.info/product_info.php?products_id=557>
Description: Words and phrases read aloud by 2000 different speakers over a public telephone network, as part of the European SpeechDat project aimed at providing speech recognition data.
Licence: Proprietary
Downloadable: Yes, on payment of a license fee.

---

1. Creative Commons Attribution-ShareAlike

2. General Public License version 3

Welsh Speech Database (1999)

Reference: Williams, B. (1999). A Welsh speech database: preliminary results.
Type: Spoken
Size: Unspecified
Description: Audio recordings of scripted speech (from Welsh periodicals), read aloud in a studio, aimed at investigating the acoustic characteristics of Welsh speech sounds and prosody.
Licence: Unspecified
Downloadable: No

CIG2 (2000)

Type: Spoken
Size: 570k words
Website: <users.aber.ac.uk/bmj/abercld/cronfa3_7/sae/intro.html; <childes.talkbank.org/browser/index.php?url=Celtic/Welsh>; <cymraeg.org.uk/kig>
Description: Child language acquisition material, originally collected 1974–7, based on conversations between Welsh children aged 3–7 years and their caregiver or another adult.
Licence: CC-BY-SA, GPL3
Downloadable: Yes

Cronfa Electroneg o Gymraeg (CEG) (2001)

Reference: Ellis, N. C., O'Dochartaigh, C., Hicks, W., Morgan, M., & Laporte, N. (2001). Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh.
Type: Written
Size: 1m words
Website: <bangor.ac.uk/canolfanbedwyr/ceg.php.en>; <corpws.cymru/ceg/?lang=en>
Description: A corpus of 500 samples of written Welsh prose (mainly post-1970) of 21 text types, aimed at research in psychology and psycholinguistics, child and second language acquisition, and general linguistics.
Licence: Non-commercial
Downloadable: Yes

LER-BIML (Language Engineering Resources for British Indigenous Minority Languages) (2002)

Type: Spoken
Size: 45k words
Website: <lancaster.ac.uk/fass/projects/biml>

Description: The Welsh component includes data from a range of spoken contexts including sermons, dental appointments, TV shows, sports commentary, school lessons, and domestic conversations.
Licence: Unspecified
Downloadable: Yes

Historical Corpus of the Welsh Language (2004)

Type: Written
Size: 420k words
Website: <people.ds.cam.ac.uk/dwew2/hcwl/menu.htm>
Description: Samples of around 15,000 words from 30 texts written between 1500 and 1850, aimed at assisting linguistic, literary and historical research on the Welsh language.
Licence: Non-commercial
Downloadable: Yes

Crúbadán Welsh Corpus (2004)

Reference: Scannell, K. P. (2007, September). The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop* (Vol. 4, pp. 5–15).
Type: E-text
Size: 24m words
Website: <crubadan.org>
*Description: W*ebsite text harvested by a web crawler, designed to support grammar checking, language recognition, machine translation and lexicography research.
Licence: CC-BY
Downloadable: Word-frequencies and n-grams only

Ein Geiriau Ni – Corpus of Children's Literature in Welsh (2005)

Type: Written
Size: 3m words
Website: <egni.org>
Description: Drawn from 430 Welsh books for children, and designed to help develop literacy materials and curricula and to support translators.
Licence: Publisher's copyright.
Downloadable: Yes

UAGT-PNAW Parallel Welsh-English Corpus (2006)

Reference: Jones, D., & Eisele, A. (2006). Phrase-based statistical machine translation between English and Welsh. *Strategies for developing machine translation for minority languages*, 75.
Type: Written
Size: 23m words
Website: <cymraeg.org.uk/Jones-Eisele>
Description: Aligned sentences from the proceedings of the National Assembly for Wales, aimed at testing a Welsh-English machine translation system and linguistic analysis.
Licence: Crown copyright
Downloadable: Yes

APPENDIX 3

Participants' questionnaire and consent form (Welsh versions)

Holiadur

Byddem ni'n ddiolchgar pe baech chi'n gallu rhoi'r wybodaeth gefndir ganlynol i ni i'n helpu ni â'n hastudiaethau.

1. Ai dyn ynteu dynes ydych chi? Dyn ☐ Dynes ☐
2. Dyddiad geni:……
3. Beth yw eich swydd bresennol (neu os ydych chi wedi ymddeol neu'n ddi-waith, beth oedd eich swydd ddiwethaf cyn ymddeol neu golli'ch gwaith)?
   ……
4. Nodwch yr ardaloedd ble rydych chi wedi byw am gyfnodau sylweddol yn eich bywyd: e.e.:

Lle: *Llandegfan, Ynys Môn*	Dyddiadau: *1975–93*
Lle: *Lerpwl*	Dyddiadau: *1993–99*
Lle: *Melbourne, Awstralia*	Dyddiadau: *1999–2002*
Lle: *Bethesda, Gwynedd*	Dyddiadau: *2002–05*
Lle: ……	Dyddiadau: ……
Lle: ……	Dyddiadau: ……
Lle: ……	Dyddiadau: ……
Lle: ……	Dyddiadau: ……
Lle: ……	Dyddiadau: ……
Lle: ……	Dyddiadau: ……

5. Beth yw'r lefel uchaf o addysg ffurfiol y gwnaethoch chi ei gorffen?
   ☐ TGAU, lefel-O/TAU, Tystysgrif Ysgol, NVQ lefel 1 neu 2 neu gyfwerth
   ☐ Lefel A/AS, Tystysgrif Ysgol Uwch, GNVQ, Diploma Cenedlaethol BTEC, NVQ lefel 3 neu gyfwerth
   ☐ Gradd Fagloriaeth, Diploma Addysg Uwch/Bellach, TAR, HND, NVQ lefel 4 neu gyfwerth
   ☐ Gradd Feistr, Doethuriaeth / PhD, NVQ lefel 5 neu gyfwerth
   ☐ Dim un o'r uchod
6. Ers pryd ydych chi'n gallu siarad Cymraeg?
   ☐ Ers pan oeddwn i'n 2 flwydd oed neu'n ifancach
   ☐ Ers pan oeddwn i'n 4 mlwydd oed neu'n ifancach
   ☐ Ers yr ysgol gynradd
   ☐ Ers yr ysgol uwchradd
   ☐ Dysgais i Gymraeg yn oedolyn

7. Ers pryd ydych chi'n gallu siarad Saesneg?
   ☐ Ers pan oeddwn i'n 2 flwydd oed neu ifancach
   ☐ Ers pan oeddwn i'n 4 mlwydd oed neu ifancach
   ☐ Ers yr ysgol gynradd
   ☐ Ers yr ysgol uwchradd
   ☐ Dysgais i Saesneg yn oedolyn
8. Ar raddfa 1 i 4, pa mor dda yn eich barn chi ydych chi'n siarad Cymraeg?
   ☐ **1** Rwy'n gwybod rhai geiriau ac ymadroddion yn unig
   ☐ **2** Rwy'n hyderus mewn sgyrsiau syml
   ☐ **3** Rwy'n eithaf hyderus mewn sgyrsiau estynedig
   ☐ **4** Rwy'n hyderus mewn sgyrsiau estynedig
9. Ar raddfa 1 i 4, pa mor dda yn eich barn chi ydych chi'n siarad Saesneg?
   ☐ **1** Rwy'n gwybod rhai geiriau ac ymadroddion yn unig
   ☐ **2** Rwy'n hyderus mewn sgyrsiau syml
   ☐ **3** Rwy'n eithaf hyderus mewn sgyrsiau estynedig
   ☐ **4** Rwy'n hyderus mewn sgyrsiau estynedig
10. Pa iaith (ieithoedd) oedd eich mam yn eu siarad â chi pan oeddech chi'n blentyn (os yw hyn yn berthnasol)?
    ☐ Cymraeg
    ☐ Saesneg
    ☐ Cymraeg a Saesneg
    ☐ Arall (Rhowch fanylion)……
    ☐ Amherthnasol
11. Pa iaith (ieithoedd) oedd eich tad yn eu siarad â chi pan oeddech chi'n blentyn (os yw hyn yn berthnasol)?
    ☐ Cymraeg
    ☐ Saesneg
    ☐ Cymraeg a Saesneg
    ☐ Arall (Rhowch fanylion)……
    ☐ Amherthnasol
12. Pa iaith (ieithoedd) oedd unrhyw warcheidwad neu roddwr gofal arall yn eu siarad â chi pan oeddech chi'n blentyn (os yw hyn yn berthnasol)?
    ☐ Cymraeg
    ☐ Saesneg
    ☐ Cymraeg a Saesneg
    ☐ Arall (Rhowch fanylion)……
    ☐ Amherthnasol
13. Ym mha iaith (ieithoedd) oeddech chi'n cael eich dysgu'n bennaf yn yr ysgol gynradd?
    ☐ Cymraeg
    ☐ Saesneg
    ☐ Cymraeg a Saesneg
    ☐ Arall (Rhowch fanylion)……
14. Ym mha iaith (ieithoedd) oeddech chi'n cael eich dysgu'n bennaf yn yr ysgol uwchradd?
    ☐ Cymraeg
    ☐ Saesneg
    ☐ Cymraeg a Saesneg
    ☐ Arall (Rhowch fanylion)……

15. Gwnewch restr isod o bump o'r bobl y byddwch yn siarad â nhw fwyaf yn eich bywyd bob dydd, naill ai wyneb yn wyneb neu dros y ffôn, e.e. eich partner, eich plentyn, ffrind, cydweithiwr etc. Yna nodwch ym mha iaith (ieithoedd) byddwch yn siarad â'r person hwnnw gan amlaf, fel mae'r tabl enghreifftiol yn dangos.

Enw'r person, neu berthynas	**Yr iaith byddaf yn ei siarad â'r person hwnnw gan amlaf: (Rhowch dic mewn un blwch isod ar bob llinell)**			
	Cymraeg	**Saesneg**	**Yr un faint o Gymraeg a Saesneg**	**Iaith arall**
1. *Siân*	✓			
2. *Mam*		✓		
3. *Bos*			✓	
4. *Jacques*				✓
5. *Chwaer*		✓		

Llenwch y tabl isod

Enw'r person, neu berthynas (Defnyddiwch enwau gwneud os yw'n well gennych)	**Yr iaith byddaf yn ei siarad â'r person hwnnw gan amlaf: (Rhowch dic mewn un blwch isod ar bob llinell)**			
	Cymraeg	**Saesneg**	**Yr un faint o Gymraeg a Saesneg**	**Iaith arall**
1.				
2.				
3.				
4.				
5.				

16. Pa sgôr byddech chi'n ei rhoi i'r iaith Gymraeg ar raddfa o 1 i 5 mewn perthynas â'r nodweddion canlynol? Rhowch gylch o gwmpas un rhif ar bob llinell.

	←				→	
hen ffasiwn	1	2	3	4	5	modern
ddim yn gyfeillgar	1	2	3	4	5	cyfeillgar
dim dylanwad	1	2	3	4	5	dylanwadol
di-fflach	1	2	3	4	5	yn ysbrydoli
da i ddim	1	2	3	4	5	defnyddiol
hyll	1	2	3	4	5	prydferth

17. Pa sgôr byddech chi'n ei rhoi i'r iaith Saesneg ar raddfa o 1 i 5 mewn perthynas â'r nodweddion canlynol? Rhowch gylch o gwmpas un rhif ar bob llinell.

	←				→	
hen ffasiwn	1	2	3	4	5	modern
ddim yn gyfeillgar	1	2	3	4	5	cyfeillgar
dim dylanwad	1	2	3	4	5	dylanwadol
di-fflach	1	2	3	4	5	yn ysbrydoli
da i ddim	1	2	3	4	5	defnyddiol
hyll	1	2	3	4	5	prydferth

18. Ydych chi'n eich ystyried eich hun yn fwy na dim yn……?
    - ☐ Gymro/Cymraes
    - ☐ Sais/Saesnes
    - ☐ Albanwr/Albanes
    - ☐ Gwyddel/Gwyddeles
    - ☐ Prydeiniwr
    - ☐ Arall (nodwch):……
19. I ba raddau ydych chi'n cytuno â'r datganiad canlynol:
    "Wrth sgwrsio bob dydd, byddaf i'n cadw'r iaith Gymraeg a'r iaith Saesneg ar wahân."
    - ☐ **1** Anghytuno'n gryf
    - ☐ **2** Anghytuno
    - ☐ **3** Ddim yn cytuno nac yn anghytuno
    - ☐ **4** Cytuno
    - ☐ **5** Cytuno'n gryf
20. I ba raddau ydych chi'n cytuno â'r datganiad canlynol:
    "Dylai pobl osgoi cymysgu Cymraeg a Saesneg yn yr un sgwrs."
    - ☐ **1** Anghytuno'n gryf
    - ☐ **2** Anghytuno
    - ☐ **3** Ddim yn cytuno nac yn anghytuno
    - ☐ **4** Cytuno
    - ☐ **5** Cytuno'n gryf

Caniatâd y Siaradwr

Trwy hyn rwyf yn rhoi fy nghaniatâd i'r wybodaeth rwyf wedi ei rhoi ar yr holiadur uchod gael ei defnyddio at ddibenion ymchwil ac/neu ddysgu yn unig (gan gynnwys cyhoeddiadau ac/neu adroddiadau ymchwil) ar yr amod bod fy enw'n cael ei gadw'n gwbl gyfrinachol.

Trwy hyn rwyf hefyd yn rhoi caniatâd i'm recordiad (sain a thrawsgrifiad) gael ei gyfrannu i'r archifau ieithyddol, gan gynnwys archif ryngwladol *LIDES (*<http://talkbank.org/data/LIDES>) ar y rhyngrwyd, ar yr amod bod enwau ffug yn cael eu defnyddio yn y trawsgrifiad yn lle enwau iawn y siaradwyr a phersonau eraill sy'n cael eu henwi yn ystod y sgwrs.

Trwy hyn rwyf hefyd yn cytuno i ganiatáu mynediad llawn i'r data hwn i ymchwilwyr ar yr amod eu bod yn ymrwymo i'r cod moeseg perthnasol (i weld y *Talkbank Code of Ethics* ewch at http://talkbank.org/share/ethics.html). Rwyf yn deall hefyd, trwy lofnodi'r ffurflen ganiatâd hon, fy mod i'n rhoi caniatâd i'r ymchwilwyr a enwyd uchod i gyflwyno detholiadau o'r data hwn fel rhan o'u gwaith ar ffurf ysgrifenedig ac/neu lafar heb gael caniatâd pellach gennyf.

Rwyf trwy hyn yn trosglwyddo'r hawlfraint ar fy nghyfraniad i Gyfarwyddwr y Project, yr Athro Margaret Deuchar.

Enw……

Cyfeiriad……

……

…… Cod Post……

Llofnod …… Dyddiad ……

Diolch yn fawr iawn i chi am eich amser a'ch cydweithrediad.

APPENDIX 4

Participants' questionnaire and consent form (English versions)

Questionnaire

We would be grateful if you could give us the following background information to help us with our studies.

1. Are you: Male ☐ Female ☐?
2. Date of birth:……
3. What is your present occupation (or if retired or unemployed, what was your last occupation before retiring or becoming unemployed)?
   ……
4. Please indicate the areas where you have lived for significant periods of your life:
   e.g.:

Place: *Llandegfan, Ynys Môn*	Dates: *1975–93*
Place: *Liverpool*	Dates: *1993–99*
Place: *Melbourne, Australia*	Dates: *1999–2002*
Place: *Bethesda, Gwynedd*	Dates: *2002–05*
Place: ……	Dates: ……
Place: ……	Dates: ……
Place: ……	Dates: ……
Place: ……	Dates: ……
Place: ……	Dates: ……
Place: ……	Dates: ……

5. What is the highest level of formal education you have completed?
   ☐ GCSE, O-level/CSE, School Certificate, NVQ level 1 or 2 or equivalent
   ☐ A/AS level, Higher School Certificate, GNVQ, BTEC National Diploma, NVQ level 3 or equivalent
   ☐ Bachelor's Degree, Diploma of Higher/Further Education, PGCE, HND, NVQ level 4 or equivalent
   ☐ Master's Degree, Doctorate, NVQ level 5 or equivalent
   ☐ None of the above
6. Since when have you been able to speak Welsh?
   ☐ Since I was 2 years old or younger
   ☐ Since I was 4 years old or younger
   ☐ Since primary school
   ☐ Since secondary school
   ☐ I learned Welsh as an adult

7. Since when have you been able to speak English?
   - ☐ Since I was 2 years old or younger
   - ☐ Since I was 4 years old or younger
   - ☐ Since primary school
   - ☐ Since secondary school
   - ☐ I learned English as an adult
8. On a scale of 1 to 4, how well do you feel you can speak Welsh?
   - ☐ **1** Only know some words and expressions
   - ☐ **2** Confident in basic conversations
   - ☐ **3** Fairly confident in extended conversations
   - ☐ **4** Confident in extended conversations
9. On a scale of 1 to 4, how well do you feel you can speak English?
   - ☐ **1** Only know some words and expressions
   - ☐ **2** Confident in basic conversations
   - ☐ **3** Fairly confident in extended conversations
   - ☐ **4** Confident in extended conversations
10. Which language(s) did your mother speak to you while you were growing up (if applicable)?
    - ☐ Welsh
    - ☐ English
    - ☐ Welsh & English
    - ☐ Other (Please specify)……
    - ☐ N/A
11. Which language(s) did your father speak to you while you were growing up (if applicable)?
    - ☐ Welsh
    - ☐ English
    - ☐ Welsh & English
    - ☐ Other (Please specify)……
    - ☐ N/A
12. Which language(s) did any other guardian or caregiver speak to you while you were growing up (if applicable)?
    - ☐ Welsh
    - ☐ English
    - ☐ Welsh & English
    - ☐ Other (Please specify)……
    - ☐ N/A
13. Through which language(s) were you predominantly taught at primary school?
    - ☐ Welsh
    - ☐ English
    - ☐ Welsh & English
    - ☐ Other (Please specify)……
14. Through which language(s) were you predominantly taught at secondary school?
    - ☐ Welsh
    - ☐ English
    - ☐ Welsh & English
    - ☐ Other (Please specify)……

15. Make a list below of five of the people you speak to most in your everyday life, either in person or on the phone, e.g. your partner, your child, a friend, a workmate etc. Then note which language(s) you mostly speak with that person, as shown in the sample table.

Name of person, or relationship	**Language mostly spoken with that person: (place a tick in one cell below for each line)**			
	Welsh	**English**	**Equally Welsh & English**	**Another language**
1. *Sian*	✓			
2. *Mother*		✓		
3. *Boss*			✓	
4. *Jacques*				✓
5. *Sister*		✓		

Please fill in table below

Name of person, or relationship (use fictitious names if you prefer)	**Language mostly spoken with that person: (place a tick in one cell below for each line)**			
	Welsh	**English**	**Equally Welsh & English**	**Another language**
1.				
2.				
3.				
4.				
5.				

16. How would you rate the Welsh language on a scale of 1 to 5 regarding the following properties? Circle one number in each line.

	←				→	
old-fashioned	1	2	3	4	5	modern
unfriendly	1	2	3	4	5	friendly
uninfluential	1	2	3	4	5	influential
uninspiring	1	2	3	4	5	inspiring
useless	1	2	3	4	5	useful
ugly	1	2	3	4	5	beautiful

17. How would you rate the English language on a scale of 1 to 5 regarding the following properties? Circle one number in each line.

	←				→	
old-fashioned	1	2	3	4	5	modern
unfriendly	1	2	3	4	5	friendly
uninfluential	1	2	3	4	5	influential
uninspiring	1	2	3	4	5	inspiring
useless	1	2	3	4	5	useful
ugly	1	2	3	4	5	beautiful

18. Do you consider yourself to be mainly……?
    - ☐ Welsh
    - ☐ English
    - ☐ Scottish
    - ☐ Irish
    - ☐ British
    - ☐ Other (please specify):……
19. To what extent do you agree with the following statement:
    "In everyday conversation, I keep the Welsh and English languages separate."
    - ☐ **1** Strongly disagree
    - ☐ **2** Disagree
    - ☐ **3** Neither agree nor disagree
    - ☐ **4** Agree
    - ☐ **5** Strongly agree
20. To what extent do you agree with the following statement:
    "People should avoid mixing Welsh and English in the same conversation."
    - ☐ **1** Strongly disagree
    - ☐ **2** Disagree
    - ☐ **3** Neither agree nor disagree
    - ☐ **4** Agree
    - ☐ **5** Strongly agree

Speaker's Consent

I hereby give my permission for the information I have given on the above questionnaire to be used for research and/or teaching purposes only (including research publications and/or reports) subject to strict preservation of my anonymity.

I also hereby give my permission for my recording (sound and transcript) to be contributed to linguistic archives, including the international archive *LIDES (*http://talkbank.org/data/LIDES) on the internet, provided that the names of speakers and other persons named during the conversation are replaced by fictitious names in the transcript.

I also hereby agree to permit full access to these data to researchers provided they subscribe to the relevant code of ethics (for the *Talkbank Code of Ethics* see http://talkbank.org/share/ethics.html). I also understand that, by signing this consent form, I give the aforementioned researchers permission to present excerpts of these data as part of their work in written and/or in oral form, without further permission from me.

I hereby assign the copyright in my contribution to the Project Director, Professor Margaret Deuchar.

Name……..
Address……...……
……..……
…… Post Code……...……
Signature …… Date ……...……
Thank you very much for your time and co-operation.

APPENDIX 5

Text of letters to potential participants (Welsh and English)

Cyfathrebu Dwyieithog yng Nghymru

Rydym yn ysgrifennu atoch i ofyn a fyddech chi'n fodlon cymryd rhan mewn project ymchwil ar sut mae pobl ddwyieithog yn cyfathrebu â'i gilydd yng Nghymru. Yr Athro Margaret Deuchar o Brifysgol Cymru, Bangor sy'n rhedeg y project, ynghyd â thîm o ymchwilwyr.

Yr hyn yr hoffem ni ei wneud yw eich recordio'n cael sgwrs anffurfiol ag aelod dwyieithog o'ch teulu neu â ffrind dwyieithog. Mae croeso i chi ddewis y person dwyieithog yr hoffech chi gael eich recordio gydag ef neu hi, a ble'r hoffech i ni eich recordio. Gallwn ni ddod i'ch cartref neu i'ch gwaith, neu os byddai'n well gennych gallwch chi ddod i'r brifysgol. Neu gallwch gael eich recordio yn sgwrsio dros y ffôn gyda'ch partner; ni fyddai rhaid i ymchwilydd ddod i'ch cartref. Bydd y gwaith recordio'n cymryd tua 45 munud a byddwn ni hefyd yn gofyn i chi lenwi holiadur byr. Ni ddylai'r sesiwn gyfan gymryd mwy nag awr, a byddwn yn talu £10 yr un i chi a'ch ffrind neu aelod o'ch teulu.

Dylai'ch sgwrs fod yn un naturiol, hamddenol. Ond byddwn ni'n rhoi rhai syniadau i chi ynghylch pynciau y gallech chi siarad amdanynt fel man cychwyn. Ar ôl gorffen byddwch chi'n cael cyfle i wrando ar y recordiad rhag ofn bod yna unrhyw ran ohono nad ydych chi eisiau i ni ei gadw. Unwaith y byddwch chi wedi rhoi caniatâd i ni gadw'r recordiad, byddwn ni'n ei ddadansoddi'n ddienw at bwrpasau ymchwil.

Bydd un o staff ein project yn eich ffonio cyn bo hir i weld a fyddech chi'n fodlon cymryd rhan. Yn y cyfamser, cysylltwch â ni os oes gennych chi unrhyw gwestiynau neu bryderon trwy lythyr, ffôn neu e-bost.

Cyfeiriad ar gyfer llythyrau: Yr Athro M Deuchar, Adran Ieithyddiaeth, Prifysgol Cymru, Bangor, Gwynedd, LL29 7ED.

Rhif ffôn: (xxxxx) xxxxxx E-bost: xxxxxx@bangor.ac.uk

Rydym yn edrych ymlaen at siarad â chi cyn bo hir.

Yr eiddoch yn gywir,
Yr Athro Margaret Deuchar.

Bilingual Communication in Wales

We are writing to you to ask if you would be willing to participate in a research project on how bilingual people communicate with each other in Wales. This is being conducted by Professor Margaret Deuchar of the University of Wales, Bangor, together with a team of researchers.

What we would like to do is to make a recording of you having an informal conversation with a bilingual member of your family or a friend. You are welcome to choose the bilingual person you would like to be recorded with, and the place you would like us to make the recording. We can come to your home or your workplace, or if you prefer you can come to the University. Alternatively you can be recorded over the telephone with your partner; there would be no requirement for a researcher to visit you. The recording will take about 45 minutes and we will also ask you to fill in a short questionnaire. The whole session should not take more than an hour, and we will pay £10 each to you and your friend or family member.

Your conversation should just be a natural, relaxed chat. We will however give you some ideas of topics you can talk about to start you off. After the recording you will be given the opportunity to listen to the recording in case there is any part of it you do not want us to keep. Once you have given us your permission to keep the recording, we will analyse it anonymously for research purposes.

One of our project staff will telephone you soon to see whether you would be willing to participate. In the meantime, please contact us if you have any questions or concerns by letter, telephone or email:

Address for letters: Professor M Deuchar, Dept of Linguistics, University of Wales, Bangor, Bangor, Gwynedd LL29 7ED

Telephone number: (xxxxx) xxxxxx E-mail: xxxxxx@bangor.ac.uk

We look forward to talking to you soon.

Yours sincerely,
Professor Margaret Deuchar.

References

Adalar, N., & Tagliamonte, S. (1998). Borrowed nouns; bilingual people: The case of the Londrali in Northern Cyprus. *International Journal of Bilingualism*, 2(2), 139–159. doi:10.1177/136700699800200202

Afarli, T. A., Grimstad, M. B., & Subbarao, K. V. (2013). Dakkhini and the problem of a matrix language frame in sustained language contact. In *International Conference on Language Contact in India* (pp. xliv–lvii). Pune, India: Deccan College Post Graduate and Research Institute.

Aitchison, J. W., & Carter, H. (2004). *Spreading the Word: The Welsh Language 2001*. Wales: Y Lolfa.

Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency*. London: Continuum.

Alexiadou, A., Lohndal, T., Afarli, T. A., & Grimstad, M. B. (2015). Language mixing: A Distributed Morphology approach. In T. Bui & D. Özyildiz (Eds.), *NELS 45: Proceedings of the Forty-Fifth Annual Meeting of the North East Linguistic Society* (Vol. 1, pp. 25–38). Cambridge, MA: MIT, CreateSpace Independent Publishing Platform.

Auer, P. (1998). Introduction: Bilingual conversation revisited. In P. Auer (Ed.), *Code-Switching in Conversation: Language, Interaction and Identity* (pp. 1–24). London: Routledge.

Auer, P., & Muhamedova, R. (2005). ' Embedded language' and 'matrix language' in insertional language mixing: Some problematic cases. *Rivista Di Linguistica*, 17(1), 35–54.

Backus, A. (1996). *Two in One. Bilingual Speech of Turkish Immigrants in the Netherlands*. Tilburg: Tilburg University Press.

Backus, A. (2004). Convergence as a mechanism of language change. *Bilingualism: Language and Cognition*, 7(7), 179–181. doi:10.1017/S1366728904001567

Backus, A. (2005). Codeswitching and language change: One thing leads to another? *International Journal of Bilingualism*, 9(4), 307–340. doi:10.1177/13670069050090030101

Backus, A. (2008). Data banks and corpora. In *The Blackwell Guide to Research Methods in Bilingualism and Multilingualism*. Oxford: Blackwell pp. 232–248. doi:10.1002/9781444301120.ch13

Bailey, G. (2002). Real and apparent time. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *Handbook of Language Variation and Change* (pp. 312–332). Oxford: Blackwell pp.232–248.

Bailey, G., Wikle, T., Tillery, J., & Sand, L. (1991). The apparent time construct. *Language Variation and Change*, 3(3), 241–264. doi:10.1017/S0954394500000569

Baker, C. (2011). *Foundations of Bilingual Education and Bilingualism* (5th ed.). Clevedon: Multilingual Matters.

Baker, C., & Jones, M. P. (2000). Welsh language education: a strategy for revitalization. In C. Williams (Ed.), *Language Revitalization: Policy and Planning in Wales* (pp. 116–137). Cardiff: University of Wales Press.

Ball, M. (1984). Phonetics for phonology. In M. J. Ball & G. E. Jones (Eds.), *Welsh Phonology: Selected Readings* (pp.5–39). Cardiff: University of Wales Press.

Ball, M. J. (1988). *The Use of Welsh: A Contribution to Sociolinguistics*. Clevedon: Multilingual Matters.

Barnett, R., Codó, E., Eppler, E., Forcadell, M., Gardner-Chloros, P., van Hout, R., … others. (2000). The LIDES Coding Manual: A document for preparing and analyzing language interaction data. *International Journal of Bilingualism*, 4(2), 131–271. doi:10.1177/13670069000040020101

Beale, A. (2016). *Version 6 of the 12dicts word lists*. Retrieved from <wordlist.aspell.net/12dicts-readme/>

Bentahila, A., & Davies, E. E. (1983). The syntax of Arabic-French code-switching. *Lingua*, 59, 301–330. doi:10.1016/0024-3841(83)90007-4

Bentahila, A., & Davies, E. E. (1995). Patterns of code-switching and patterns of language contact. *Lingua*, 96, 75–93. doi:10.1016/0024-3841(94)00035-K

Berk-Seligson, S. (1986). Linguistic constraints on intrasentential code-switching: A study of Spanish-Hebrew bilingualism. *Language in Society*, 15, 313–348. doi:10.1017/S0047404500011799

Bernsten, J. (1990). *The Integration of English Loans in Shona: Social Correlates and Linguistic Consequences* (Unpublished PhD Dissertation). Michigan State University.

Bhatt, R. M. (2015). *Bilingual situations in India: Power relations between languages analysed through code-switching*. Unpublished

Borsley, R. D., Tallerman, M., & Willis, D. (2007). *The Syntax of Welsh*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511486227

Bourhis, R. Y., & Giles, H. (1976). The language of co-operation in Wales. *Language Sciences*, 42, 13–16.

Bourhis, R. Y., Giles, H., & Tajfel, H. (1973). Language as a determinant of Welsh identity. *European Journal of Social Psychology*, 3(4), 447–460. doi:10.1002/ejsp.2420030407

Breit, F. (2012). Constraints on auxiliary deletion in colloquial Welsh. Unpublished BA dissertation, Bangor University. <florian.me.uk/resources/files/files/publications/breit-ba2012.pdf>

Broersma, M., & Bot, K. de. (2006). Triggered code-switching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and Cognition*, 9, 1–13. doi:10.1017/S1366728905002348

Bullock, B., & Gerfen, C. (2004). Phonological convergence in a contracting language variety. *Bilingualism: Language and Cognition*, 7(2), 95–104. doi:10.1017/S1366728904001452

Bullock, B., & Toribio, A. J. (2004). Introduction: Convergence as an emergent property in bilingual speech. *Bilingualism: Language and Cognition*, 7(7), 91–93. doi:10.1017/S1366728904001506

Cacoullos, R. T., & Travis, C. E. (2016). Two languages, one effect: Structural priming in spontaneous code-switching. *Bilingualism: Language and Cognition*, 19(4), 733–753. doi:10.1017/S1366728915000127.pdf

Cantone, K. F., & MacSwan, J. (2009). The syntax of DP-internal codeswitching. In L. Isurin, D. Winford, & K. de Bot (Eds.), *Multidisciplinary Approaches to Codeswitching* (pp. 243–278). Amsterdam: John Benjamins. doi:10.1075/sibil.41.14can

Carter, D., Broersma, M., & Donnelly, K. (2016). Applying computing innovations to bilingual corpus analysis. In E. V. A. Alba de la Fuente & C. Martínez-Sanz (Eds.), *Language Acquisition Beyond Parameters: Studies in Honour of Juana M. Liceras* (pp. 281–301). Amsterdam: John Benjamins. doi:10.1075/sibil.51.11car

Carter, D., Broersma, M., Donnelly, K., & Konopka, A. (2017). Presenting the Bangor autoglosser and the Bangor automated clause splitter. *Digital Scholarship in the Humanities*, 33(1), 21–28. doi:10.1093/llc/fqw065

Carter, D., Deuchar, M., Davies, P., & Couto, M.-C. P. (2011). A systematic comparison of factors affecting the choice of matrix language in three bilingual communities. *Journal of Language Contact*, 4, 1–31. doi:10.1163/187740911X592808

Carter, P., & Cooper, S. (2012). Variation in the acoustics of laterals in Welsh. Poster presented at the 2012 Colloquium of the British Association of Academic Phoneticians, University of Leeds, 26th-28th March 2012.

Cedergren, H., & Sankoff, D. (1974). Performance as a statistical reflection of competence. *Language*, 50, 333–355. doi:10.2307/412441

Chambers, J. K. (2003). *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Oxford: Blackwell.

Chambers, J. K., & Trudgill, P. (1980). *Dialectology* (1st ed.). Cambridge: Cambridge University Press.

Chambers, J. K., Trudgill, P., & Schilling-Estes, N. (Eds.). (2004). *Handbook of Language Variation and Change*. Oxford: Wiley-Blackwell. doi:10.1002/9780470756591

Chan, B. H.-S. (2009). Code-switching between typologically distinct languages. In B. E. Bullock & A. J. Toribio (Eds.), *The Cambridge Handbook of Linguistic Code-Switching* (pp. 182–198). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511576331.012

Cheshire, J., & Gardner-Chloros, P. (1998). Code-switching and the sociolinguistic gender pattern. *International Journal of the Sociology of Language*, 129(1), 5–34. doi:10.1515/ijsl.1998.129.5

Chiarcos, S., Hellmann, C., & Nordhoff, S. (2011). Towards a linguistic linked open data cloud: The Open Linguistics Working Group. *TAL*, 3, 245–275.

Comrie, B. (2000). From potential to realization: An episode in the origin of language. *Linguistics*, 38, 989–1004. doi:10.1515/ling.2000.019

Comrie, B., Haspelmath, M., & Bickel, B. (2008). *Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-Morpheme Glosses*. Retrieved from <eva.mpg.de/lingua/resources/glossing-rules.php>

Cooper, S. (2011). Frequency and loudness in overlapping turn onset by Welsh speakers. In *Proceedings of the 17th International Congress of Phonetic Sciences* (Vol. 4, pp. 516–519). Hong Kong.

Costa, A., & Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language*, 50, 491–511. doi:10.1016/j.jml.2004.02.002

Coupland, N. (2010). Welsh linguistic landscapes 'from above' and 'from below'. In A. Jaworski & C. Thurlow (Eds.), *Semiotic Landscapes: Language, Image, Space*. (pp. 77–101). London: Continuum.

Crystal, D. (2000). *Language Death*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139106856

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell. doi:10.1002/9781444302776

Davies, J. (1993). *The Welsh Language*. Cardiff: University of Wales Press.

Davies, P. (2010). *Identifying Word-order Convergence in the Speech of Welsh-English Bilinguals* (Unpublished PhD dissertation). Bangor University.

Davies, P. (2016). Age variation and language change in Welsh: Auxiliary deletion and possessive constructions. In M. Durham & J. Morris (Eds.), *Sociolinguistics in Wales*. (pp. 31–59). London: Palgrave Macmillan. doi:10.1057/978-1-137-52897-1_2

Davies, P., & Deuchar, M. (2010). Using the Matrix Language Frame model to identify word order convergence in Welsh-English bilingual speech. In A. Breitbarth, C. Lucas, S. Watts, & D. Willis (Eds.), *Continuity and change in Grammar* (pp. 77–96). Amsterdam: John Benjamins. doi:10.1075/la.159.04dav

Davies, P., & Deuchar, M. (2014). Auxiliary deletion in the informal speech of Welsh-English bilinguals: A change in progress. *Lingua*, 143, 224–241. doi:10.1016/j.lingua.2014.02.007

de Bot, K. (1992). Self-assessment of minority language proficiency. In L. Verhoeven & J. H. A. L. de Jong (Eds.), *The Construct of Language Proficiency: Applications of Psychological Models to Language Assessment* (pp. 137–146). Amsterdam: John Benjamins. doi:10.1075/z.62.15bot

Deuchar, M. (2005). Congruence and code-switching in Welsh. *Bilingualism: Language and Cognition*, 8, 255–269. doi:10.1017/S1366728905002294

Deuchar, M. (2006). Welsh-English code-switching and the Matrix Language Frame model. *Lingua*, 116(11), 1986–2011. doi:10.1016/j.lingua.2004.10.001

Deuchar, M. (2012). Code switching. In C. A. Chapelle (Ed.), *Encyclopedia of Applied Linguistics* (pp. 657–664). New York, NY: Wiley Online Library. doi:10.1002/9781405198431.wbeal0142

Deuchar, M., & Davies, P. (2009). Code switching and the future of the Welsh language. *International Journal of the Sociology of Language*, 195, 15–38. doi:10.1515/IJSL.2009.004

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

Deuchar, M., Donnelly, K., & Piercy, C. (2016). *Mae pobl monolingual yn minority*: Factors favouring the production of code-switching by Welsh-English speakers. In M. Durham & J. Morris (Eds.), *Sociolinguistics in Wales* (pp. 209–239). London: Palgrave Macmillan. doi:10.1057/978-1-137-52897-1_8

Deuchar, M., & Quay, S. (2000). *Bilingual Acquisition: Theoretical Implications of a Case Study*. Oxford: Oxford University Press.

Deuchar, M., & Stammers, J. R. (2012). What IS the 'Nonce Borrowing Hypothesis' anyway? *Bilingualism: Language and Cognition*, 15(3), 649–650. doi:10.1017/S1366728911000563

Deuchar, M., & Stammers, J. R. (2016). English-origin verbs in Welsh: Adjudicating between two theoretical approaches. *Languages*, 1(1), 1–16. doi:10.3390/languages101007

Didriksen, T. (2016). *Constraint Grammar Manual*. Retrieved from <visl.sdu.dk/cg3/chunked/>

Donnelly, K. (2016). *Eurfa: a free (GPL) dictionary for Welsh, v0.3*. Retrieved from <eurfa.org.uk/>

Donnelly, K., Cooper, S., & Deuchar, M. (2011). Glossing CHAT files using the Bangor Autoglosser. In *8th International Symposium for Bilingualism*. Oslo. Retrieved from <kevindonnelly.org.uk/resources/words/Donnelly2011_ISB8.pdf>

Donnelly, K., & Deuchar, M. (2011a). The Bangor Autoglosser: A multilingual tagger for conversational text. Wrexham, Wales: ITA11.

Donnelly, K., & Deuchar, M. (2011b). 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. Retrieved from <hdl.handle.net/10062/19298>

Ellis, N. C., O'Dochartaigh, C., Hicks, W., Morgan, M., & Laporte, N. (2001). *Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh*. Retrieved from <bangor.ac.uk/canolfanbedwyr/ceg.php.en>

Eppler, E. D. (2010). *Emigranto: The syntax of German-English code-switching*. Vienna: Braumüller.

Eppler, E. D., Luescher, A., & Deuchar, M. (2017). Evaluating the predictions of three syntactic frameworks for mixed determiner-noun constructions. *Corpus Linguistics and Linguistic Theory*. 13(1), 27–63. doi:10.10.1515/cllt-2015-0006

Fasold, R. (1990). *The Sociolinguistics of Language*. Oxford: Blackwell.

Finlayson, R., Calteaux, K., & Myers-Scotton, C. (1998). Orderly mixing and accommodation in South African codeswitching. *Journal of Sociolinguistics*, 2(3), 395–420. doi:10.1111/1467-9481.00052

Fishman, J. A. (1991). *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages*. Clevedon: Multilingual Matters.

French, P., & Local, J. (1983). Turn-competitive incomings. *Journal of Pragmatics*, 7, 701–715. doi:10.1016/0378-2166(83)90147-9

Fricke, M., & Kootstra, G. J. (2016). Primed codeswitching in spontaneous bilingual dialogue. *Journal of Memory and Language*, 91, 181–201. doi:10.1016/j.jml.2016.04.003

Fricke, M., Kroll, J. F., & Dussias, P. E. (2016). Phonetic variation in bilingual speech: A lens for studying the production–comprehension link. *Journal of Memory and Language*, 89, 110–137. doi:10.1016/j.jml.2015.10.001

Gal, S. (1978). Peasant men can't get wives: Language change and sex roles in a bilingual community. *Language in Society*, 7(1), 1–16. doi:10.1017/S0047404500005303

Gal, S. (1979). *Language Shift: Social Determinants of Linguistic Change in Bilingual Austria*. New York NY: Academic Press.

Gardner, R. C., & Lambert, W. E. (1972). *Attitudes and Motivation in Second-Language Learning*. Rowley, MA: Newbury House.

Gardner-Chloros, P. (1992). The sociolinguistics of the Greek-Cypriot community in London. In M. Karyolemou (Ed.), *Plurilinguismes: Sociolinguistique du Grec et de la Grèce* (Vol. 4, pp. 112–136). Paris: CERPL.

Gardner-Chloros, P. (2009). *Code-switching*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511609787

Gardner-Chloros, P., Moyer, M., & Sebba, M. (2007). Coding and analysing multilingual data: The LIDES project. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora, Volume 1: Synchronic Databases* (pp. 91–120). London: Palgrave Macmillan. doi:10.1057/9780230223936_5

Garrett, P. (2010). *Attitudes to Language*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511844713

Garrett, P., Coupland, N., & Williams, A. (Eds.). (2003). *Investigating Language Attitudes: Social Meanings of Dialect, Ethnicity and Performance*. Cardiff: University of Wales Press.

Gathercole, V. C. M., & Thomas, E. M. (2009). Bilingual first-language development: Dominant language takeover, threatened minority language take-up. *Bilingualism: Language and Cognition*, 12(2), 213–237. doi:10.1017/S1366728909004015

Giles, H., Taylor, D. M., & Bourhis, R. Y. (1977). Dimensions of Welsh identity. *European Journal of Social Psychology*, 7(2), 165–174. doi:10.1002/ejsp.2420070205

González-Vilbazo, K., & López, L. (2011). Some properties of light verbs in code-switching. *Lingua*, 121(5), 832–850. doi:10.1016/j.lingua.2010.11.011

Goodz, N. S. (1989). Parental language mixing in bilingual families. *Infant Mental Health Journal*, 10(1), 25–44. doi:10.1002/1097-0355(198921)10:1<25::AID-IMHJ2280100104>3.0.CO;2-R

Grosjean, F. (2013). Bilingualism: A short introduction. In Grosjean, F. & Li, P. *The Psycholinguistics of Bilingualism* (pp. 5–25). Oxford: Wiley-Blackwell.

Grosjean, F. (2001). The bilingual's language modes. In J. L. Nichol (Ed.), *One Mind, Two Languages: Bilingual Language Processing* (pp. 1–22). Oxford: Blackwell.

Gullberg, M., Indefrey, P., & Muysken, P. (2009). Research techniques for the study of code-switching. In B. E. Bullock & A. J. Toribio (Eds.), *The Cambridge Handbook of Linguistic Code-switching* (pp. 21–39). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511576331.003

Herring, J. R., Deuchar, M., Couto, M. P., & Moro Quintanilla, M. (2010). 'I saw the madre': Evaluating predictions about codeswitched determiner-noun sequences using Spanish-English and Welsh-English data. *International Journal of Bilingual Education and Bilingualism*, 13(5), 553–573. doi:10.1080/13670050.2010.488286

Johnson, D. E. (2009). Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects Variable Rule analysis. *Language and Linguistics Compass*, 3(1), 359–383. doi:10.1111/j.1749-818X.2008.00108.x

Jones, B. L. (1981). Welsh: Linguistic conservation and shifting bilingualism. In J. D. M. E. Haugen & D. Thomson (Eds.), *Minority Languages Today* (pp. 40–51). Edinburgh: Edinburgh University Press.

Jones, B. M. (1990). Variation in the use of pronouns in verbnoun phrases and genitive noun phrases in child language. In M. J. Ball, J. Fife, E. Poppe & J. Rowland (Eds.), *Celtic Linguistics/Ieithyddiaeth Geltaidd: Readings in the Brythonic Languages. Festschrift for T. Arwyn Watkins*. Amsterdam: John Benjamins. doi:10.1075/cilt.68.09jon

Jones, B. M. (2004). The licensing powers of mood and negation in spoken Welsh: Full and contracted forms of the present tense of bod 'be'. *Journal of Celtic Linguistics*, 8, 87–107.

Jones, D. G. (1988). Literary Welsh. In M. J. Ball (Ed.), *The Use of Welsh: A Contribution to Sociolinguistics* (pp. 125–171). Clevedon: Multilingual Matters.

Jones, K. (1995). Code-switching, intertextuality and hegemony: Exploring change in bilingual discourse. *Proceedings of the Summer School, 'Code-Switching and Language Contact'* (pp. 108–118). Ljouwert/Leeuwarden: Fryske Akademy.

Jones, K. (2000). Texts, mediation and social relations in a bureaucratized world. In M. Martin-Jones & K. Jones (Eds.), *Multilingual Literacies: Reading and Writing Different Worlds* (pp. 209–228). Amsterdam: John Benjamins.

Jones, M. C. (1998). *Language Obsolescence and Revitalization*. Oxford: Clarendon Press.

Jones, M. C. (2005). Some structural and social correlates of single word intrasentential code-switching in Jersey Norman French. *Journal of French Language Studies*, 15(1), 1–23. doi:10.1017/S0959269505001894

Jones, M., & Thomas, A. R. (1977). *The Welsh Language: Studies in its Syntax and Semantics*. Cardiff: University of Wales Press.

Jones, R. O. (1993). The sociolinguistics of Welsh. In M. J. Ball & J. Fife (Eds.), *The Celtic Languages* (pp. 536–604). London: Routledge.

Joshi, A. (1985). Processing sentences with intrasentential code switching. In D. R. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural Language Parsing* (pp. 190–205). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511597855.006

Kamwangamalu, N. M. (1987). French/vernacular code mixing in Zaire: Implications for syntactic constraints on code mixing. In B. Need, E. Schiller, & A. Bosh (Eds.), *Chicago Linguistics Proceedings* 22 (pp. 166–180). Chicago, IL: University of Chicago Linguistics Department.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. In *COLING 90 Proceedings of the 13th Conference on Computational Linguistics* (pp. 168–173). Retrieved from <dl.acm.org/citation.cfm?id=991176> doi:10.3115/991146.991176

Karlsson, F., Voutilainen, A., Heikkilä, J., & Anttila, A. (Eds.). (1995). *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*. Berlin: Mouton de Gruyter. doi:10.1515/9783110882629

King, G. (2016). *Modern Welsh: A Comprehensive Grammar* (3rd ed.). London: Routledge.

Kroll, J. F., Sumutka, B. M., & Schwartz, A. I. (2005). A cognitive view of the bilingual lexicon: Reading and speaking words in two languages. *International Journal of Bilingualism*, 9(1), 27–48. doi:10.1177/13670069050090010301

Labov, W. (1972). *Sociolinguistic Patterns*. Pennsylvania, PA: University of Pennsylvania Press.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, 2, 205–254. doi:10.1017/S0954394500000338

Labov, W. (2001). *Principles of Linguistic Change, Volume II: Social Factors*. Oxford: Blackwell.

Leimgruber, J. R. E. (2013). The trouble with World Englishes: Rethinking the concept of 'geographical varieties' of English. *English Today*, 29(3), 3–7. doi:10.1017/S0266078413000242

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Lewis, M. B., Gerhand, S., & Ellis, H. D. (2001). Re-evaluating age-of-acquisition effects: Are they simply cumulative-frequency effects? *Cognition*, 78(2), 189–205. doi:10.1016/S0010-0277(00)00117-7

Lewis, W. G. (2008). Current challenges in bilingual education in Wales. *Aila Review*, 21(1), 69–86.

Li, W., & Milroy, L. (1995). Conversational code-switching in a Chinese community in Britain: A sequential analysis. *Journal of Pragmatics*, 23(3), 281–299. doi:10.1016/0378-2166(94)00026-B

Lindsay, C. F. (1993). Welsh and English in the city of Bangor: A study in functional differentiation. *Language in Society*, 22(01), 1–17. doi:10.1017/S0047404500016894

Lingen, R. R. W., Symons, J. C., & Johnson, H. R. V. (1847). *Reports of the Commissioners of Enquiry into the State of Education in Wales, Appointed by the Committee of Council on Education*. William Clowes. Retrieved from <llgc.org.uk/discover/digital-gallery/printed-material/the-blue-books-of-1847>

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3), 384–422. doi:10.1080/00437956.1964.11659830

Lloyd, S. W. (2008). *Variables that affect English language use within Welsh conversation in North Wales* (Unpublished MA Thesis). Bangor University.

MacSwan, Jeff. (2009). Generative approaches to code-switching. In B. Bullock & A. Toribio (Eds.), *The Cambridge Handbook of Linguistic Code-Switching* (pp. 309–335). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511576331.019

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, Volume II: The database. *Computational Linguistics*, 26(4), 657–657. doi:10.1162/coli.2000.26.4.657

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Meisel, J. M. (2004). The bilingual child. In T. K. Bhatia & W. C. Richie (Eds.), *The Handbook of Bilingualism* (pp. 91–113). Oxford: Blackwell.

Meisel, J. M. (2010). Age of onset in successive acquisition of bilingualism: Effects on grammatical development. In M. Kail & M. Hickmann (Eds.), *Language Acquisition across Linguistic and Cognitive Systems* (pp. 225–248). Amsterdam: John Benjamins. doi:10.1075/lald.52.16mei

Meuter, R., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory and Language*, 40, 25–40. doi:10.1006/jmla.1998.2602

Milroy, L. (1980). *Language and Social Networks*. Oxford: Blackwell.

Milroy, L. (1987). *Observing and Analysing Natural Language: A Critical Account of Sociolinguistic Method*. Oxford: Blackwell.

Montes-Alcalá, C. (2000). Attitudes towards oral and written codeswitching in Spanish-English bilingual youths. In A. Roca (Ed.), *Research on Spanish in the U.S.* (pp. 218–227). Somerville, MA: Cascadilla Press.

Montrul, S. A. (2004). Subject and object expression in Spanish heritage speakers: A case of morpho-syntactic convergence. *Bilingualism: Language and Cognition*, 7(2), 125–142. doi:10.1017/S1366728904001464

Montrul, S. A. (2008). *Incomplete Acquisition in Bilingualism: Re-examining the Age Factor*. Amsterdam: John Benjamins. doi:10.1075/sibil.39

Moro Quintanilla, M. (2014) The semantic interpretation and syntactic distribution of determiner phrases in Spanish-English code-switching. In J. MacSwan (Ed.), *Grammatical Theory and Bilingual Codeswitching* (pp. 213–226). Cambridge, MA: The MIT Press.

Morris, J. (2013). *Sociolinguistic Variation and Regional Minority Language Bilingualism: An Investigation of Welsh-English Bilinguals in North Wales* (Unpublished PhD Dissertation). University of Manchester.

Musk, N. (2010). Code-switching and code-mixing in Welsh bilinguals' talk: Confirming or refuting the maintenance of language boundaries. *Language, Culture and Curriculum*, 23(3), 179–197. doi:10.1080/07908318.2010.515993

Muysken, P. (2000). *Bilingual Speech: A Typology of Code-mixing*. Cambridge: Cambridge University Press.

Myers-Scotton, C.. (1992). Comparing code-switching and borrowing. *Journal of Multilingual and Multicultural Development*, 13(1–2), 19–39. doi:10.1080/01434632.1992.9994481

Myers-Scotton, C. (1993). *Duelling Languages: Grammatical Structure in Codeswitching* (1st ed.). Oxford: Clarendon Press.

Myers-Scotton, C. (1997). *Duelling Languages: Grammatical Structure in Codeswitching* (2nd ed.). Oxford: Clarendon Press.

Myers-Scotton, C. (1998). A way to dusty death: The matrix language turnover hypothesis. In L. A. Grenoble & L. J. Whaley (Eds.), *Endangered Languages: Language Loss and Community Response* (pp. 289–316). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139166959.013

Myers-Scotton, C. (2002). *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780198299530.001.0001

Myers-Scotton, C., & Jake, J. (2015). Cross-language asymmetries in code-switching patterns. In J. W. Schwieter (Ed.), *The Cambridge Handbook of Bilingual Processing* (pp. 416–458). Cambridge: Cambridge University Press. doi:10.1017/CBO9781107447257.019

Nartey, J. (1982). Code-switching, interference or faddism? Language use among educated Ghanaians. *Anthropological Linguistics*, 24(2), 183–192.

Nortier, J. (1990). *Dutch-Moroccan Arabic Code Switching*. Dordrecht: Foris.

Ochs, E. (1979). Transcription as theory. *Developmental Pragmatics*, 10(1), 43–72.

Pandharipande, R. (1990). Formal and functional constraints on code-mixing. In R. Jacobson (Ed.), *Codeswitching as a Worldwide Phenomenon* (pp. 15–31). Bern: Peter Lang.

Parafita Couto, M. C., Fusser, M., & Deuchar, M. (2015). How do Welsh-English bilinguals deal with conflict? Adjective-noun order resolution. In G. Stell & K. Yakpo (Eds.), *Code-switching between Structural and Sociolinguistic Perspectives* (pp. 65–84). Berlin: De Gruyter.

Parafita Couto, M. D. C., Boutonnet, B., Hoshino, N., Davies, P., Deuchar, M., & Thierry, G. (2017). Testing alternative account of code-switching using event-related potentials: a pilot study on Welsh-English. In F. Lauchlan & M. C. Parafita Couto (Eds.), *Bilingualism and Minority Languages in Europe: Current Trends and Developments* (pp. 240–254). Cambridge: Cambridge Scholars Publishing.

Parafita Couto, M. C., Davies, P., Carter, D., & Deuchar, M. (2014). Factors influencing code-switching. In E. M. Thomas & I. Mennen (Eds.), *Advances in the Study of Bilingualism* (pp. 111–138). Clevedon: Multilingual Matters.

Paul, H. (1898). *Principien der Sprachgeschichte* (3rd ed.). Halle: Max Niemeyer.

Penhallurick, R. (2004). Welsh English: Phonology. In E. W. Schneider, K. Burridge, B. Kortmann & C. Upton (Eds.), *A Handbook of Varieties of English, Volume 1: Phonology* (pp. 98–112). Berlin: Mouton de Gruyter.

Penhallurick, R. (2007). English in Wales. In D. Britain (Ed.), *Language in the British Isles* (pp. 152–72). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511620782.010

Pfaff, C. W. (1979). Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English. *Language*, 55(2), 291–318. doi:10.2307/412586

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching. *Linguistics*, 18(7–8), 581–618.

Poplack, S. (1987). Contrasting patterns of code-switching in two communities. In E. Wande, J. Anward, B. Nordberg, L. Steensland, & M. Thelander (Eds.), *Aspects of Bilingualism: Proceedings from the Fourth Nordic Symposium on Bilingualism, 1984* (pp. 51–76). Uppsala: Borgström, Motala.

Poplack, S. (1988). Contrasting patterns of codeswitching in two communities. In M. Heller (Ed.), *Codeswitching. Anthropological and Sociolinguistic Perspectives* (pp. 215–244). Berlin: Mouton De Gruyter. doi:10.1515/9783110849615.215

Poplack, S. (2000). Sometimes I'll start a sentence in Spanish *y termino en español*: toward a typology of code-switching. In L.Wei (Ed.) *The Bilingualism Reader* (pp. 221–256). London and New York: Routledge.

Poplack, S. (2012). What does the Nonce Borrowing Hypothesis hypothesize? *Bilingualism: Language and Cognition*, 15(3), 644–648. doi:10.1017/S1366728911000496

Poplack, S., & Dion, N. (2012). Myths and facts about loanword development. *Language Variation and Change*, 24(3), 279–315. doi:10.1017/S095439451200018X

Poplack, S., & Meechan, M. (1998). How languages fit together in code-mixing. *International Journal of Bilingualism*, 2(2), 127–138. doi:10.1177/136700699800200201

Poplack, S., & Sankoff, D. (1984). Borrowing: the synchrony of integration. *Linguistics*, 22(1), 99–135. doi:10.1515/ling.1984.22.1.99

Poplack, S., Sankoff, D., & Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26(1), 47–104. doi:10.1515/ling.1988.26.1.47

Prys, M. (2016). *Style in the Vernacular and on the Radio: Code-switching and Mutation as Stylistic and Social Markers in Welsh* (Unpublished PhD Dissertation). Bangor University.

Prys, M., Deuchar, M., & Roberts, G. (2012). Measuring bilingual accommodation in Welsh rural pharmacies. In *Multilingual Individuals and Multilingual Societies* (pp. 419–436). Amsterdam: John Benjamins. doi:10.1075/hsm.13.28pry

Redinger, D. (2010). *Language Attitudes and Code-switching Behaviour in a Multilingual Educational Context: The Case of Luxembourg* (Unpublished PhD Dissertation). University of York.

Redknap, C. (2006). Welsh-medium and bilingual education and training: Steps towards a holistic strategy. In C. Redknap, W. G. Lewis, S. R. Williams, & J. Laugharne (Eds.), *Welsh-Medium and Bilingual Education* (pp. 1–20). Bangor: School of Education, Bangor University.

Roberts, I. G. (2005). *Principles and Parameters in a VSO Language: A Case Study in Welsh*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780195168211.001.0001

Rosignoli, A. (2011). *Flagging in English-Italian Code-switching* (Unpublished PhD Dissertation). Bangor University.

Sandalo, F. (1995). *A Grammar of Kadiwé* (Unpublished PhD Dissertation). University of Pittsburgh.

Sankoff, D., & Labov, W. (1979). On the uses of variable rules. *Language in Society*, 8 (2–3), 189–222. doi:10.1017/S0047404500007430

Sankoff, D., Poplack, S., & Vanniarajan, S. (1990). The case of the nonce loan in Tamil. *Language Variation and Change*, 2(1), 71–101. doi:10.1017/S0954394500000272

Sankoff, D., Tagliamonte, S., Smith, E. (2005). GoldVarb X: a variable rule application for Macintosh and Windows. <individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm>

Savage, M., Devine, F., Cunningham, N., Taylor, M., Li, Y., Hjellbrekke, J., Le Roux, B., & Miles, A. (2013). A new model of social class? Findings from the BBC's Great British Class Survey experiment. *Sociology*, 47(2), 219–250. doi:10.1177/0038038513481128

Schmitt, E. (2000). Overt and covert codeswitching in immigrant children from Russia. *International Journal of Bilingualism*, 4, 9–28. doi:10.1177/13670069000040010201

Schmitt, E. (2001). *Beneath the Surface: Signs of Language Attrition in Immigrant Children in Russia* (Unpublished PhD dissertation). University of South Carolina.

Silva-Corvalán, C. (1994). *Language Contact and Change: Spanish in Los Angeles*. Oxford: Oxford University Press.

Smith, D. J. (2006). Thresholds leading to shift: Spanish/English codeswitching and convergence in Georgia, U.S.A. *International Journal of Bilingualism*, 10(2), 207–240. doi:10.1177/13670069060100020501

Stammers, J. (2010). *The Integration of English-origin Verbs into Welsh: A Contribution to the Debate over Distinguishing between Code-switching and Lexical Borrowing*. Saarbrücken: VDM Verlag.

Stammers, J., & Deuchar, M. (2012). Testing the nonce borrowing hypothesis: Counter-evidence from English-origin verbs in Welsh. *Bilingualism: Language and Cognition*, 15(3), 630–643. doi:10.1017/S1366728911000381

Tagliamonte, S. A. (2012). *Variationist Sociolinguistics: Change, Observation, Interpretation*. Oxford: Wiley-Blackwell.

Thomas, A. R. (1992). The Welsh language. In M. J. Ball & J. Fife (Eds.), *The Celtic Languages* (pp. 251–345). London: Routledge.

Thomas, E. M., Williams, N., Jones, L. A., Davies, S., & Binks, H. (2014). Acquiring complex structures under minority language conditions: Bilingual acquisition of plural morphology in Welsh. *Bilingualism: Language and Cognition*, 17(3), 478–494. doi:10.1017/S1366728913000497

Thomas, H. C. (1982). Registers in Welsh. *International Journal of the Sociology of Language*, 35, 87–115.

Thomason, S., & Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics*. Berkeley, CA: University of California Press.

Toribio, A. J. (2002). Spanish-English code-switching among US Latinos. *International Journal of the Sociology of Language*, 158, 89–120.

Toribio, A. J. (2004). Convergence as an optimization strategy of bilingual speech: Evidence from codeswitching. *Bilingualism: Language and Cognition*, 7, 165–173. doi:10.1017/S1366728904001476

Treffers-Daller, J. (1992). French-Dutch codeswitching in Brussels: Social factors explaining its disappearance. *Journal of Multilingual and Multicultural Development*, 13(1–2), 143–156. doi:10.1080/01434632.1992.9994488

Treffers-Daller, J. (1994). *Mixing Two Languages: French-Dutch Contact in a Comparative Perspective*. Berlin: Mouton de Gruyter. doi:10.1515/9783110882230

Wang, S.-L. (2007). *Evaluating Competing Models of Code-switching with Reference to Mandarin/Tsou and Mandarin/Southern Min Data* (Unpublished PhD Dissertation). Bangor University.

Wang, S.-L. (2017). The PF Disjunction Theorem to Southern Min/Mandarin code-switching. *International Journal of Bilingualism* 21 (5), 541–558. doi:10.1177/1367006916637677

Weber-Fox, C., & Neville, H. (1999). Functional neural subsystems are differentially affected by delays in second language immersion: ERP and behavioural evidence in bilinguals. In D. Birdsong (Ed.), *Second Language Acquisition and the Critical Period Hypothesis* (pp. 23–38). Mahwah, NJ: Lawrence Erlbaum Associates.

Weinreich, U. (1953). *Languages in Contact: Findings and Problems*. New York, NY: Linguistic Circle.

Welsh Assembly Government. (2003). *Iaith Pawb: National Action Plan for a Bilingual Wales*. Welsh Government. Retrieved from.

Welsh Government. (2012). *A Living Language: A Language for Living*. Retrieved from <gov.wales/docs/dcells/publications/122902wls201217en.pdf>

Williams, C. (2004). *Iaith Pawb*: The doctrine of plenary inclusion. *Contemporary Wales*, 17, 1–27.

Williams, C. (2013). *Minority Language Promotion, Protection and Regulation: The Mask of Piety*. London: Palgrave Macmillan. doi:10.1057/9781137000842

Willis, D. (2016). Cyfieithu iaith y caethweision yn Uncle Tom's Cabin a darluniadau o siaradwyr ail iaith mewn llenyddiaeth Gymraeg, *Llên Cymru* 39(1), 56–72.

Willis, D. (2017). Investigating geospatial models of the diffusion of morphosyntactic innovations: The Welsh strong second-person singular pronoun *chdi*. *Journal of Linguistic Geography*, 5(1), 41–61. doi:10.1017/jlg.2017.1

Wilson, A., & Worth, C. (2003). Building and annotating corpora of spoken Welsh and Gaelic. In *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 909–917). Cardiff: University Centre for Computer Corpus Research on Language Technical Papers.

Wilson, C., & Deuchar, M. (2017). Extralinguistic factors influencing the pronunciation of English by Welsh-English bilinguals. In F. Lauchlan & M. C. Parafita Couto (Eds.), *Bilingualism and Minority Languages in Europe: Current Trends and Developments* (pp. 48–69). Newcastle upon Tyne: Cambridge Scholars.

Wurm, S. A. (1998). Methods of language maintenance and revival, with selected cases of language endangerment in the world. In K. Matsumura (Ed.), *Studies in Endangered Languages: Papers from the International Symposium on Endangered Languages* (pp. 191–211). Tokyo: Hituzi Syobo.

Index

This book is a research monograph divided into two parts. The first part describes the methods used to build the first sizeable corpus of informal conversational data collected from bilingual speakers of Welsh and English: Siarad. The second part describes the linguistic analysis of data from this corpus (available at bangortalk.org.uk). The information in Part One will be useful as a 'how to' manual on building a bilingual spoken corpus, including methods of data collection, transcription, glossing and analysis. The findings reported in Part Two throw new light on the debate regarding code-switching vs. borrowing, the application of the Matrix Language Framework (MLF) to the grammar of Welsh-English code-switching, the extralinguistic factors influencing variation in quantity of code-switching, and the extent to which the grammar of Welsh is changing in contact with English. Additional findings by other researchers using the corpus are also reported, and possible future directions are discussed.

ISBN 978 90 272 0011 2

JOHN BENJAMINS PUBLISHING COMPANY