COPiL CAMBRIDGE OCCASIONAL PAPERS *in* LINGUISTICS
Volume 9, Article 8: 212–234, 2016 | ISSN 2050-5949

"*MAE POBL MONOLINGUAL YN MINORITY*": FACTORS FAVOURING THE PRODUCTION OF CODE-SWITCHING BY WELSH-ENGLISH BILINGUAL SPEAKERS*

Margaret Deuchar
Bangor University
University of Cambridge

Kevin Donnelly
7 Ty'n y Cae
Llanfairpwll
Anglesey
LL61 6UX

Caroline Piercy
Bangor University

Abstract

This paper presents new findings regarding the relationship between early bilingual acquisition and the use of code-switching in adulthood. We report on the results of an automatic analysis of 67,515 clauses from a Welsh-English corpus (www.bangortalk.org.uk and http://talkbank.org/data/BilingBank/Bangor) collected from 148 speakers. Using a clause-based analysis, we aimed to determine which extralinguistic factors appear to influence the production of bilingual clauses (i.e. containing code-switching) versus monolingual clauses. We used an innovative automatic glossing mechanism to extract clauses and analysis by Rbrul. We found that the significant factors which influenced code-switching in our data were age and pattern of bilingual acquisition. Younger speakers produced more code-switching than older speakers, and those who had acquired both Welsh and English simultaneously from birth produced more code-switching than those who had acquired Welsh first and then English, or vice versa.

Introduction

Poplack's (1980) landmark study of code-switching among Puerto Rican Spanish-English speakers in New York City provided evidence that "code-switching, rather than representing debasement of linguistic skill, is actually a sensitive indicator of bilingual ability" (Poplack 1980: 581). She found that those speakers who did the most intrasentential code-switching had acquired both English and Spanish in

* We are grateful for the financial support of the AHRC and ESRC in the collection and analysis of the data, and for the contribution of all those who provided, collected and transcribed data.

early childhood and also rated themselves as 'bilingual' as opposed to dominant in Spanish or English.

Given the evidence that code-switching appears to be facilitated by proficiency in the two languages, a question which has not yet been fully answered is how varying patterns of bilingual acquisition lead to a greater or lesser propensity to code-switch. Meisel (2004), for example, distinguishes between simultaneous acquisition of two languages, child second language acquisition, and adult second language acquisition. He argues that the differing effects of these patterns of bilingual acquisition need to be determined "in the light of empirical research investigating linguistic and neuropsychological aspects of bilingualism acquired during different age ranges" (Meisel 2004: 105). Indeed, in a study of structural plasticity in the bilingual brain, Mechelli, Crinion, Noppeney, O'Doherty, Ashburner, Frackowiak & Price (2004) report on how the timing of bilingual acquisition and proficiency attained affect the density of grey matter and structural reorganisation in the brain. It seems likely, then, that similar factors may affect code-switching behaviour.

Poplack's (1980) study was not able to deal directly with the relation between patterns of acquisition and code-switching since only two of her twenty speakers were simultaneous bilinguals. However, since the time of her study, developments in corpus linguistics mean that we can now analyse much large sets of data in a relatively short amount of time. These developments allow, among other things, the automatic extraction of data for analysis, as we shall demonstrate in our study of 148 Welsh-English bilinguals with varying patterns of bilingual acquisition.

Review of the literature

In this section we review some of the previous work which has investigated the relation between social and linguistic factors in the study of code-switching, with special emphasis on the role of early bilingual acquisition. We also review relevant work on corpus linguistics and previous work specifically on Welsh-English data.

Poplack (1980) is one of the best known early studies on the multivariate analysis of code-switching. Her data were collected in 'El Barrio', an area of New York City inhabited by a Puerto Rican community since the 1930s. Data were analysed from twenty speakers who differed from one another regarding their age of arrival in the USA, Eleven were male and 9 female. Data were collected through interviews and 'natural' recordings by a member of the community, and speakers also completed a language attitude questionnaire. Sixty-six hours of recordings yielded 1,835 instances of code-switching, all of which were coded in terms of syntactic function. A broad distinction was drawn between intrasentential and extrasentential switches[1], and the relation between these categories and extralinguistic characteristics of the speakers was studied using VARBRUL 2 (Sankoff 1975), a tool for multivariate analysis. The results showed that the factors which were related to the production of intra-sentential code-switching were gender, age of arrival/L2 acquisition, language dominance and work place. More intra-sentential code-switching

[1] This included both 'sentential' (also called 'intersentential') and 'tag' switches.

was produced by women than men, by those who had been born in the USA or arrived in early childhood, by those who were balanced bilinguals rather than Spanish dominant, and by those who worked inside the community.

Almost all of Poplack's speakers had acquired English later than Spanish, albeit at different ages, and since the age of acquisition of English corresponded perfectly with the age of speakers' arrival in the USA, age of acquisition was not considered separately. Furthermore, since only two speakers had acquired English in early childhood, the effect of simultaneous vs. successive acquisition could not be compared. Our study differs from Poplack's in that we are able to compare the effect of simultaneous and successive acquisition at different ages, and in that we are dealing with a fairly stable bilingual community which is not the result of the immigration of minority language speakers.

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. He 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 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.

Treffers-Daller (1992) reports on a study of Dutch-French code-switching in Brussels which might be considered more similar to the community in our own study in that the community is not the result of recent migration. Treffers-Daller found that the variables 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. Treffers-Daller (1994) includes details of the background questionnaire administered to participants, but information about their patterns of language acquisition in childhood is not elicited and so we cannot determine how this might be linked to their code-switching patterns. However, she did investigate the effect of age on the production of code-switching. 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: 148). She suggests that intraclausal code-switching

is actually disappearing in Brussels owing to the influence of purism in Dutch.

In studies of language variation the age of the speaker is of course an important independent variable because of the possibilities of the 'apparent time paradigm' (Bailey 2002), according to which the speech of younger speakers may be indicative of language change. Thus the extent of code-switching by younger speakers compared with older speakers may provide an indication of whether code-switching is decreasing or increasing. Poplack (1980) found that age of speaker, was not a significant variable in predicting type of code-switching. However, this may be because of the relatively small number (20) of her speakers and the fact that 75% of them were between the ages of 20 and 40. The age of our 148 speakers ranged from 10 to 89 and we shall show how age is a key variable in our study.

As mentioned above, Poplack (1980) found that gender was a significant variable and that women produced more intrasentential switching than men. In fact, over half of their switches were intra-sentential compared with only one third of men's switches. Given what are often considered robust findings regarding the differences between male and female monolingual speech in English, termed "the sociolinguistic gender pattern" by Cheshire & Gardner-Chloros (1998), these authors set out to investigate whether "other factors being equal, the general pattern appeared to hold, with women code-switching less than men in order to conform with a more purist or socially acceptable speech style" (Cheshire & Gardner-Chloros 1998: 14). They were able to find little evidence for this "general 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 Greek Cypriot-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: 28).

As well as drawing on previous work on code-switching, the current study also draws on developments in corpus linguistics which have benefitted enormously from the use of increasingly sophisticated computational tools which can be accessed by all researchers and allow access to much larger sets of data. Although corpora were used in hard copy form (e.g. on index cards) before this time, the Brown corpus of American English, built in the 1960s, was the first electronically readable linguistic corpus (see e.g. Baker 2010: 59). Not surprisingly, English is the language best represented in current corpora (see McEnery & Hardie 2012: 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 tend to either involve one language with translations into another or two or more monolingual corpora side by side. There is no mention of corpora of spoken bilingual communication including code-switching between two or more languages. However, such corpora have been available in the public domain since about 2000 (see e.g. talkbank.org/BilingBank and the appendix to Gardner-Chloros 2009 on the LIDES project).

One of the first corpora on the Talkbank website to be extensively analysed is the Eppler corpus of German-English conversation by Austrian immigrants in London, described in a monograph by Duran Eppler (2010). Duran Eppler used the CHAT system from Talkbank (MacWhinney 2000) for the transcription of her data, which means that she could also use the Talkbank CLAN programs for its analysis. She used the CLAN programs to generate quantitative analyses of her data, for example on the frequency of code-switching, but her syntactic analysis was done manually. She uses CLAN to report on the distribution of languages per speaker, but did not otherwise study code-switching patterns in relation to speakers or speaker characteristics.

McEnery & Hardie (2012: 226) consider that, since the 1980s, corpus linguistics has become established as part of the "methodological toolbox" of linguistics. According to (Gries 2011: 291) "it is probably no exaggeration to say that it is only over the last 20 years or so, that corpus linguistics has really taken off and developed into one of the most widely-used methods in linguistics". However, Baker (2010: 1) notes that "corpus linguistics has made only a relatively small impact on sociolinguistics". This is rather surprising since, as Baker points out, corpus linguistics shares with variationist sociolinguistics a quantitative approach to the study of differences between populations. Nevertheless, as Tagliamonte (2006) points out, sociolinguists were working on corpora from at least the 1980s, even if these corpora were not always in the public domain. Poplack (1989), for example, describes the process of creating her Ottawa-Hull French corpus and her use of the Oxford Concordance Program to produce word frequency lists, vocabulary statistics and an index or concordance. The advent of the personal computer not only made it possible for individual researchers to analyse their corpora automatically in this way, but the computer also allowed access to new tools for the analysis of variation. One of these was the variable rule program (cf. Sankoff 1975, mentioned above), which has undergone various developments since.

Johnson (2009) introduces a new version of the variable rule program called Rbrul, which we use in the present study. This is more versatile than the earlier version, Goldvarb, although its results can be presented in a format compatible with that of Goldvarb to allow comparison with previous work. Rbrul is an implementation in the statistics language R (Venables, Smith & the R Core Team 2013) of mixed-effects modelling, which, among other things, takes into account the random effects introduced by individual speakers (cf. Baayen, Davidson & Bates 2008). In comparing Rbrul with Goldvarb, Johnson notes that the latter treats each token as if it were independent, even though this is not the case: the tokens are not independent, since they occur in groups produced by individual speakers. There is therefore a danger of Goldvarb overestimating external effects like gender and age. However, mixed effects models can distinguish between 'fixed effects' like gender and age and 'random effects' like the effects of individual speakers. As Johnson (2009: 365) says, a mixed effects model "can still capture external effects, but only when they are strong enough to rise above the inter-speaker variation". Drager & Hay (24: 60) argue that an increase in statistical robustness is the main reason that this model should be adopted by sociolinguists, and point out that the model al-

lows the simultaneous study of both group and individual variation. Tagliamonte & Baayen (2012) demonstrate the usefulness of the mixed-effects models in the analysis of York English and also introduce "random trees" as a way of providing information about the relative importance of a range of variables, especially where the data are unbalanced and contain complex interactions.

Our study on the factors influencing the code-switching patterns of Welsh-English also builds on previous work we have done in this area. Deuchar (2005) used pilot conversational data to demonstrate that code-switching was more likely to occur where there was both paradigmatic and syntagmatic congruence between the grammatical categories of Welsh and English. Deuchar (2006) used a small sample of conversational data to argue that Welsh-English code-switching was conducive to analysis by the Matrix Language Frame (MLF) approach in that a matrix language (usually Welsh) could clearly be identified in bilingual clauses. Similar results were reported by Davies & Deuchar (2010) in a paper which argued that there was very little evidence that the speech of bilinguals was leading to convergence between Welsh and English. Similarly, Deuchar & Davies (2009) argued that although some of the clauses (16%) of a sample of speakers were bilingual in that they contained both Welsh and English words, the morphosyntactic frame of the clauses was almost always Welsh, justifying confidence in the stability of the Welsh language.

Lloyd (2008) conducted a study using some of the same data as ours in order to determine which external variables affected the percentage of English words used in otherwise Welsh conversations. She analysed the speech of 121 speakers from our Siarad corpus who had been brought up in North Wales. Using background information from our questionnaire, she found that the age of the speaker, the language of their education, and parental input were all important factors. However, she did not examine the effect of pattern of bilingual acquisition, an important variable in our study. Her results showed that older speakers used a smaller percentage of English words on average than younger speakers. Regarding language of education, Lloyd 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. This result was contrary to her predictions in that she had expected the latter category to use more English words. However, there was a confound with age in that those who had received their education in both Welsh and English tended to be older. Regarding home language, Lloyd found that speakers who had heard Welsh from at least one parent had a (statistically non-significant) tendency to use more English than those who had heard only English. She suggests that those speakers who have heard more Welsh at home may be more likely to be balanced bilinguals because of the large amount of English input in society at large. This argument might also help to explain her results relating to the language of education and are in line with Gathercole & Thomas (2009)'s findings that enhanced input in Welsh is necessary for command of Welsh to equal command of English in Wales.

Carter, Deuchar, Davies & Parafita Couto (2011) reported on a comparative analysis of the factors influencing code-switching patterns in a sample of speakers from our three bilingual corpora collected in Wales, Miami (USA) and Patagonia (Ar-

gentina). They compared the proportion of bilingual vs. monolingual clauses in each sample and identified the matrix language or morphosyntactic frame of each clause. The highest proportion of bilingual clauses (19%) was found in the Welsh-English sample collected in Wales, while the lowest proportion (3%) was found in the Welsh-Spanish sample collected in Patagonia. Regarding the matrix language of the bilingual clauses, this was found to be most uniform in the sample from Wales, where 100% of the clauses had Welsh as the matrix language. The Patagonia sample was almost as uniform, with 93% of the bilingual clauses having Welsh as a matrix language, but the Miami data showed more variability with 66% of the Spanish-English bilingual clauses having a Spanish ML and the remaining 34% having English as a matrix language. Carter et al. (2011) noted that there was uniformity in the choice of ML when the language pair had contrasting word orders, as in VSO (Welsh) vs. SVO (English and Spanish) in Wales and Patagonia. They then sought to account for the specific choice of the ML in terms of external factors. Self- reported proficiency in both languages turned out to be relatively high in both Wales and Miami, and it seems that this may have favoured the production of bilingual clauses in those two samples, whereas the lower proportion of fluent bilinguals in Patagonia may account for the smaller proportion of bilingual clauses there. Regarding the choice of the matrix language, Carter et al. predicted that the most common language of the social network would also be the most common matrix language. This prediction was fulfilled in Wales, where speakers' mainly Welsh-speaking social network could be linked to their overwhelming choice of Welsh as a matrix language. Similarly, the tendency of Spanish-English speakers in Miami to have a more bilingual social network was arguably reflected the more diverse choice of both Spanish and English as matrix languages. In Patagonia the relation between social networks and matrix language was unclear, partly because of the small number of Welsh speakers in that community.

Parafita Couto, Davies, Carter & Deuchar (2014) report on the first multivariate analysis of our Spanish-English data, in which we attempted to find a relation between external factors and the choice of Spanish vs. English as matrix language in our Miami data. A Goldvarb analysis of 2,611 clauses extracted manually from three transcripts of conversations revealed no significant relationship between the choice of matrix language and external factors, but this may have been because of the small amount of data. In the study to be reported here we were able to analyse 67,515 clauses as a result of computer-assisted glossing and clause segmentation.

Data collection and transcription

In collecting the Siarad Welsh-English corpus we obtained 40 hours of spontaneous data based on informal conversations between pairs of bilingual speakers. Most of the data were collected over a two-year period (2005–2007) and came from 151 speakers. Our aim was to recruit a wide range of bilingual speakers, the main criterion being that participants considered themselves to be bilingual in the two languages associated with each community. Beyond that we wished to record both men and women, of a wide range of ages (but mostly adults), with varying profi-

ciency in the two languages. Proficiency was self-assessed as part of questionnaires administered after the recordings. We also gathered information on a wide range of other external variables which included age, gender, occupation, age of acquisition of the two languages, language input in the family, social networks and self-report on the extent of participants' code-switching. Our method of recruitment was to send letters to bilingual speakers known to our research team or their contacts and also to place advertisements in the university and in public places. Our researchers were themselves Welsh-English bilinguals living locally who could draw to some extent on their own social networks. The project was described as concerning bilingual communication, and the letter mentioned that we wanted to make recordings of informal conversation between bilingual people. We invited letter recipients to choose a bilingual family or friend with whom they would be willing to be recorded. Recipients were invited to choose the place of recording, whether at home or work, for example. While this freedom of choice meant that we could not control the environmental sound in the recordings, it helped to ensure informality.

Once appointments had been made with participants, they were met by one of the researchers and given a short briefing about the project: they were told that we were studying how bilinguals communicate with each other, although no mention was made of mixing languages or code-switching, and that we would record them having a conversation for 35–40 minutes. Before the recording it was explained that their anonymity would be protected by using pseudonyms for them and anyone they mentioned in the course of the conversation, and that they would be able to ask for anything they said to be deleted if they subsequently changed their mind. The recording equipment used for most recordings was a Marantz hard disk recorder, while a small number were recorded with a portable Sony minidisk recorder. Several steps were taken to reduce as much as possible any effect of the Observer's Paradox. The speakers were recorded with partners whom they already knew, in most cases very well. Audio-recording without video was used so as to intrude less on the conversation. Wherever possible the researcher left the room or house so that their presence would not influence the language choices made by the participants or inhibit code-switching because of any self-consciousness. The pair was also left to talk for several minutes longer than the length that would become the final edited version in the corpus. This was so that the first five minutes of each recording could be removed in case the participants' speech might have been affected while they became accustomed to the recording equipment. These precautions proved to be highly successful in eliciting the naturalistic data sought. For example, it is noticeable in many of the recordings that both through the relaxed way in which the speakers interact, and the potentially sensitive topics that they discuss, that they did not seem to feel observed.

The transcription system selected was CHAT, and its associated CLAN software CLAN (see MacWhinney 2000 and http://childes.psy.cmu.edu/manuals/ CHAT.pdf) since it was to be made available on Talkbank, where CHAT is the standard software system. The fundamental features of CHAT notation are that utterances are placed on tiers: minimally, a main tier that consists of an orthographic representation of the words in the utterance. There are also optional tiers which may contain phono-

91 *JAQ: mi ges i heddiw # crackers@s:cym& eng # a # egg@s:eng mayonnaise@s:cym& eng.
93 %gls: PRT get.1S.PAST PRON.1S today crackers and egg mayonnaise
94 %eng: I had today crackers and egg mayonnaise

Figure 1 Tiers in transcribed utterances

logical and/or phonetic representations, word-by-word glosses of non-English material, a translation of the utterance, discourse level mark-up, comments etc. We decided that each transcribed utterance would minimally have a main tier, a gloss tier, and a tier with translation into English. These tiers are illustrated in Figure 1 from stammers2.[2] The first (line 91) is the main tier, the second (line 93) a gloss tier and the third (94) is the translation tier.

The main tier contains the actual words of the speaker's utterance, and also shows the source language of each word. Following the current norms in CHAT, words belonging to the ('default') language which has the most words in the transcript are not marked for language, but words from other languages are so marked. In *Siarad* Welsh is always the default language, and English words are marked with the tag '@s:eng' as in the English word 'egg@s:eng' in the above example. There is also a large number of words (often loans from English into Welsh) which are marked with the tag '@s:cym&eng' indicating 'undetermined language'. Words like 'mayonnaise@s:cym&eng' in the example above are originally English words but are found in Welsh dictionaries and often pronounced as in English. Words of this kind are spelled with English orthography but marked as undetermined. Similar neutral language marking was also used with place names and some interactional markers that we considered to belong to both language systems, e.g. 'ah@s:cym&eng'.

The glossing[3] of the main tier (resulting in the words in the gloss tier marked with '%gls') was initially done manually, but was later augmented by adding a further tier (%aut) containing glosses generated automatically by computer (Donnelly & Deuchar 2011), and it is these glosses which were used for the analysis reported in this paper. The automatic system splits the transcribed utterances into words, looks up the words in open source dictionaries, adds glosses to each word, uses constraint grammar to disambiguate multiple glosses and writes the final glosses into the CHAT file. It is calculated to be 97–98% accurate. Figure 2 shows the utterance from example 1 as stored in the database: the spoken words are in the column labelled 'surface', the automatic glosses in the 'auto' column, and the language origin of each word ('cym' for Welsh, 'eng' for English) is in the last column.

The example in Figure 1 can then be expanded with more detailed glossing information as shown in Figure 3. It is the automatic glosses in the '%aut' tier which allows the analysis to be performed.

[2] See www.bangortalk.org.uk.
[3] In the Siarad CHAT files available on Bangortalk (www.bangortalk.org.uk) the autogloss (marked with '%aut') is given as an additional tier to the manual gloss (%gls).

location	surface	auto	langid
1	mi	PRT.AFF	cym
2	ges	get.V.1S.PAST+SM	cym
3	i	I.PRON.1S	cym
4	heddiw	today.ADV	cym
5	crackers	cracker.N.SG+PL	cym&eng
6	a	and.CONJ	cym
7	egg	egg.N.SG	eng
8	mayonnaise	mayonnaise.N.SG	cym&eng
9	.	*NULL*	999

Figure 2 Example of utterance with automatic glosses

*JAQ:	mi ges i heddiw # crackers@s:cym& eng # a # egg@s:eng mayonnaise@s:cym& eng.
%gls:	PRT get.1S.PAST PRON.1S today crackers and egg mayonnaise
%aut:	PRT.AFF get.V.1S.PAST+AM PRON.1S today cracker.N.SG+PL and.CONJ egg.N.SG mayonnaise.N.SG
%eng:	I had today crackers and egg mayonnaise

Figure 3 Tiers in transcribed utterances with more glossing information

Data analysis

Intraclausal vs. interclausal code-switching

The terms intraclausal and interclausal correspond roughly to what are called intrasentential and intersentential code-switching elsewhere but are more precise (cf. Deuchar 2012). Intraclausal code-switching is illustrated by example (1)[4] below and interclausal code-switching by example (2).

(1) [*maen* *nhw* *(y)n* *rhoi* *e* *yn* *y*
be.V.3S.PRES they.PRON.3P PRT give.V.INFIN he.PRON.M.3S in.PREP the.DET.DEF
STEAM *ROOM* [*dw* *i* *mynd* *yn*]].
steam.N.SG room.N.SG be.V.1S.PRES I.PRON.1S go.V.INFIN in.PREP

'They put it in the steam-room I go to.' [fusser27: 139]

[4] Words in italic lower case are Welsh, in italic upper case English, and non-italics are used for words belonging to both languages. The glosses have been aligned with the words for ease of reading and are explained in the Siarad documentation file to be found at www.bangortalk.org.uk.

Examples (1) to (9) are referenced by giving the name of the file they come from, followed by the number of the utterance (called the 'main tier' in CLAN).

(2) [*so* *bosib* *hwnna* *(y)dy* *o*]
so.ADV possible.ADJ+SM that.PRON.DEM.M.SG be.V.3S.PRES he.PRON.M.3S
[*I* *DON'T* *KNOW*]
I.PRON.SUB.1S do.V.1S.PRES+NEG know.V.INFIN

'So possibly that's it, I don't know.' [fusser25: 1073]

In example (1) there is a switch within the clause to the English phrase steam room whereas in example (2) there is a switch from an entire Welsh clause to the English clause *I don't know*.

Our analysis focused on intraclausal code-switching, which was much more frequent in our data than interclausal code-switching. For the purposes of the analysis, intraclausal code-switching was considered to be manifest in clauses coded as bilingual rather than monolingual. Example (1) above would be coded as bilingual because it contains words from both English and Welsh. Example (2), however, would be considered to consist of two monolingual clauses, one in Welsh and the other in English. Words which could belong to either Welsh or English (on the grounds that they were found in dictionaries of both languages) were ignored in the process of coding. Thus English loanwords in Welsh were distinguished from switches. The extent of intraclausal code-switching was measured in terms of the number of bilingual clauses produced as a proportion of the total number of clauses.

Data preparation

Because of our focus on the clause as a unit of analysis, all utterances from the corpus had to be split 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 (3) the mark points are underlined, and the segmentation points are marked with '/':

(3) *mae* *(y)r* *hogan* *(y)na* */ oedd* *ar* *Eastenders*
be.V.3S.PRES the.DET.DEF girl.N.F.SG there.ADV be.V.3S.IMPERF on.PREP name
/ mae *hi* *(we)di* *gwneud* *un* */ dydy*
be.V.3S.PRES she.PRON.F.3 after.PREP make.V.INFIN one.NUM be.V.3S.PRES.NEG

'That girl from Eastenders, she's done one, hasn't she?' [roberts2: 800]

In (4) and (5), the mark point has been moved leftwards from the finite verb marked with an asterisk:

(4) *mae* *(y)n* *sure* */ eith* *hi* *rywbryd*
be.V.3S.PRES PRT sure.ADJ go.V.3S.PRES she.PRON.F.3S at_some_stage.ADV+S
timod */ ond* *mae** *(h)i* *(y)n* *mwynhau* *Yn*
know.V.2S.PRES but.CONJ be.V.3S.PRES she.PRON.F.3S PRT enjoy.V.INFIN PRT
fa(n) *(y)ma* *wedyn.*
place.N.MF.SG+S here.ADV afterwards.ADV

'She'll probably go sometime, but she's enjoying it here, so ...'

[davies2: 3592]

(5) *fasai* *fo* *(y)n* *gwybod* */ (ba)sai* *medru*
be.VS.PLUPERF he.PRON.M.3S PRT know.V.INFIN be.V.3S.PLUPERF be_able.V.INFIN
rheoli *(we)dyn* */ be* *mae** *(y)n* *fwyta*
manage.V.INFIN afterwards.ADV what.INT be.V.3S.PRES PRT eat.V.INFIN+SM
wedyn */ basai*
afterwards.ADV

[fusser13: 872]

To test the accuracy of the segmentation of clauses in Welsh, the predominant language, 1318 Welsh-only utterances which had been split into four or more clauses were collected, and every tenth one was examined to check whether the clauses were correctly segmented. In the 528 clauses in the sample, 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 2.4% 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 used the data to address the following research questions:

- What is the extent of intraclausal code-switching in the Siarad corpus?
- Do speaker characteristics such as age and pattern of bilingual acquisition predict the observed code-switching?

Statistical analysis

As described above, each clause was coded as either monolingual or bilingual. This allowed us to quantify the amount of code-switching by speakers in terms of its

presence (in bilingual clauses) versus absence (in monolingual clauses). The categories 'bilingual clause' vs. 'monolingual clause' were treated as variants of the dependent variable which we label 'linguality'. Table 1 illustrates the automatic coding of the linguality of each clause, whether bilingual ('biling'), monolingual Welsh (monoW) or monolingual English (monoE). Clauses coded as monolingual contain only words from one language (whether Welsh or English) whereas bilingual clauses contain one or more words from both languages. In addition, we coded the language of the verb ('verblg'), whether Welsh ('cym') or English ('eng').

File-name	Utterance ID	Speaker	Clause	Verblg	Linguality
fusser17	1257	AET	oedd o yn dechrau diflannu	cym	monoW
fusser25	148	HUN	because they're leaving	eng	monoE
roberts2	267	RIS	achos mae gynna chdi spellchecker Cymraeg arno fo	cym	biling
lloyd1	720	GRG	in Cymru we recycle	eng	biling

Table 1 Results of coding the linguality of extracted clauses

The data comprised 80,352 clauses from the 151 speakers in the Siarad corpus. However, for this analysis we removed two speakers EVA and GLA who had learned Dutch as their first language, because we wished to focus on the role of Welsh and English acquisition in early childhood as a predictor of code-switching. It was also necessary to remove a further speaker, ARD, since the data on first language acquired was missing. Removing these three speakers gives a large data set for analysis of 148 speakers and 79,116 clauses. The speakers were distributed by age and gender as shown in Table 2. The effect of speaker gender turned out not to be significant unlike that of age, on which we report below.

	Overall	Male	Female
N	148	70	78
Average age	42	43	40
Youngest	10	12	10
Oldest	89	86	89
%	100	47	43

Table 2 The speaker sample by age and gender

Before analysis of intraclausal code-switching could begin, clauses consisting of only one word were removed from the data set. This is because we considered it necessary for there to be at least two words within a clause to provide an opportunity for intraclausal code-switching to take place[5]. In total, 11,601 clauses of only one word were removed leaving 67,515 clauses in the data set distributed as shown in Table 3.

Table 3 shows that the majority of clauses (88%) are monolingual Welsh and only a tiny fraction (2%) are monolingual English. However, bilingual clauses (those containing intraclausal code-switches) make up 10% of all clauses. 147 of the 148 speakers in the analysis used a majority of Welsh monolingual clauses (range 61.7–99.7% per speaker). Contrast this with the use of English: here the range of use, excepting speaker GRG (81.8% solely English clauses) was 0–28% monolingual English clauses; indeed this analysis shows that 21 speakers used no monolingual English clauses at all. All but one speaker, DER, produced intraclausal code-switches to varying degrees; the range per speaker is 0–31.1% intraclausal code-switches per speaker.

	N	%
Total clauses	67515	100
Of which:		
Monolingual Welsh	59152	88
Monolingual English	1656	2
Bilingual (Welsh & English)	6707	10
Mean per speaker	456	
Minimum per speaker	47	
Maximum per speaker	1106	

Table 3 Distribution of clauses consisting of more than one word by language and speaker

Table 3 shows that intraclausal code-switching (as evidenced by bilingual clauses) is found frequently. We then examined whether any speaker attributes and self-reports of code-switching were correlated with the use of intraclausal code-switching. The Siarad questionnaire responses provided a rich and diverse set of social data to analyse. However, many of the questionnaire responses were designed to elicit related information, and answers to these questions were therefore often correlated. For example, speakers were asked to assess their own about their self-described ability in Welsh and English and also about when they learned both languages. In order to ensure the independence of external factors in the multivariate model, we chose to focus in the analysis to be reported here on the relation between diverse patterns of bilingual acquisition and the age of the speaker. Age was treated as a continuous variable while the factor group 'pattern of bilingual

[5] Word-internal code-switching can occur in Welsh when an English verb is given a verbal suffix, e.g. concentrate-io. There were 333 instances of this in the 11,061 clauses that we removed and thus these instances were not included in our analysis of intraclausal code-switching.

acquisition' included five factors: (1) Welsh and English were acquired simultaneously from birth; (2) The second language (L2, whether Welsh or English) was being acquired by age four; (3) L2 was acquired at primary school; (4) L2 was acquired a secondary school; (5) L2 was acquired in adulthood.

The multivariate analysis was conducted in R using Rbrul (Johnson 2009). The dependent variable was the linguality of each clause; bilingual vs. monolingual Welsh or English. The analysis used a mixed-effects model with speaker included as a random intercept. This approach has the advantage of compensating for the effects of idiosyncratic linguistic behaviour by particular speakers. The results of our analysis are shown in Table 4. Table 4 shows that age and pattern of bilingual acquisition are related to the number of intraclausal code-switches a speaker produces.

	Log-odds	**No. of Clauses**	**% of Bilingual Clauses**	**Centred Factor Weight**
Age	-0.02	67515		
Pattern of bilingual acquisition				
Both Welsh and English from birth	0.407	15572	14.7	0.6
L2 by age four	-0.053	19006	10.3	0.487
L2 at primary school	-0.087	26501	7.8	0.478
L2 at secondary school	-0.059	3710	6.6	0.485
L2 in adulthood	-0.209	2726	5.6	0.448

Table 4 Mixed effects logistic regression predicting bilingual clauses with speaker as a random effect

Regarding age, the analysis shows that as age increases the presence of bilingual clauses decreases. Details of the relation between age and code-switching are shown in Figure 4.

Table 4 also shows that speakers who learned Welsh and English simultaneously were more likely to produce intraclausal code-switches than speakers who learnt one language later than the other.

Discussion

In the introduction we pointed out that although ideas about the relation between code-switching and proficiency have been familiar since Poplack's 1980 work, little has previously been known about the impact of patterns of bilingual acquisition on adult bilinguals' speech production. In particular, we have not known how these patterns influence speakers' choice to code-switch within clauses or not to switch. Our results as reported in Table 4 show that those speakers who had acquired both Welsh and English from birth were significantly more likely to produce intraclausal code-switching than all other categories of speaker, including those who had acquired their second language as young as age four. Although the overall percentage of bilingual clauses in our data is 10%, the bilingual clauses of speakers who were

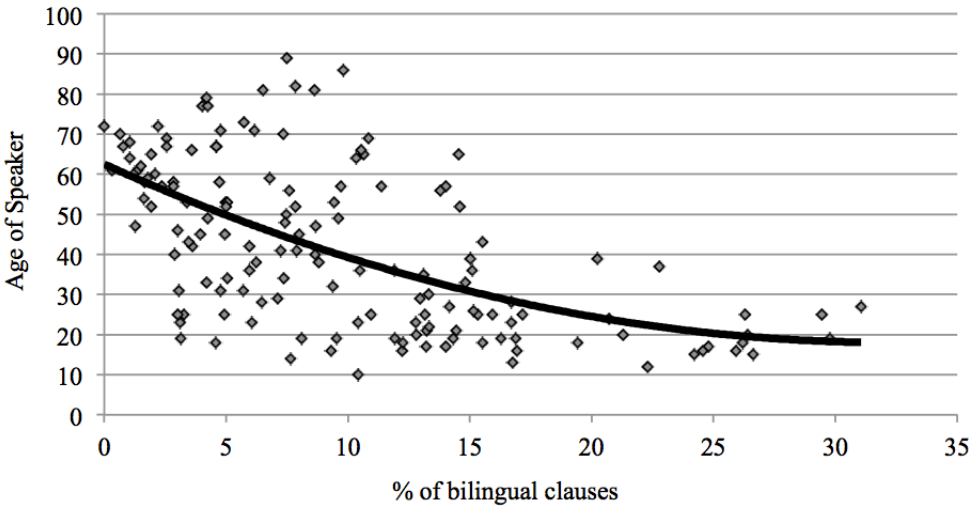

Figure 4 Percentage of bilingual clauses by speaker age
Note: N=148 speakers. Mean age = 42. Pearson's r = -0.57, $p \leq 0.01$. A regression line has been included to guide the eye.

simultaneous bilinguals as infants make up 15% of their output. This percentage drops to 6% for those who acquired their second language as adults.

It is well known that achieving native-like competence in a language or languages is very rarely possible unless the languages are acquired at a young age, but there is debate about what exactly this cut-off age is. (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 where virtually never produced by simultaneous German-French bilinguals. Meisel suggests that these differences may be explained by neural maturation, with some important changes occurring in the fourth year of life. He refers to neuroscientific studies which support his conclusion.

In the introduction we referred to the study by Mechelli et al. (2004) which showed how the timing of bilingual acquisition affected the density of grey matter in the brain. Specifically, they found that grey matter density in the inferior parietal cortex was negatively correlated with the age of acquisition in the second language. In other words, simultaneous bilinguals had the greatest density of grey matter in this area, followed by those who had acquired their second language early, followed by those who had acquired it later. Furthermore, the authors point out that "The inferior parietal region that is associated with second-language acquisition corresponds exactly to an area that has been shown by functional imaging to become activated during verbal-fluency tasks" (Mechelli et al. 2004: 757). We suggest that intraclausal code-switching is the type of activity to be particularly favoured by verbal fluency. Another relevant study is that by Weber-Fox & Neville (1999) who explore 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 elicited by phrase structure violations showed "increased bilateral distribution with increased second language immersion" Weber-Fox & Neville (1999: 30). These and some behavioural results showing slower syntactic processing with increased age of second language acquisition led them 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: 36). This conclusion suggests to us 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.

One might wonder whether simultaneous acquisition of two languages in infancy would lead to qualitative as well as quantitative differences in code-switching. Although not working with simultaneous bilinguals, Finlayson, Calteaux & Myers-Scotton (1998) found that multilinguals with a higher level of English proficiency produced more switched English phrases than those with a lower level of proficiency, who tended to switch single English words. Treffers-Daller (1992: 144) reports excluding single-word switches from her analysis of French-Dutch code-switching in case they might be borrowings. In our study we excluded borrowings (described above as loans marked in our transcription as '@s:cym&eng') from our analysis of code-switching, but decided to investigate whether simultaneous bilinguals produced more switched phrases (as opposed to switched single words) than those who had acquired one language later than the other.

To do this we classified the bilingual clauses into two types: single-word insertions and multi-word insertions. Single-word insertions were defined as being single words in otherwise monolingual Welsh clauses as seen in (6). Or they could be multiple incidences of single word insertions within an otherwise Welsh clause as seen in (7).

(6) *ti* *(e)rioed* *yn* *SERIOUS*
you.PRON.2S never.ADV PRT serious.ADJ

'You're never serious.' [davies6: 494]

(7) *well* *APPARENTLY* *well* *APPARENTLY* *mae* *MONOLINGUAL*
well.ADV apparently.ADV well.ADV apparently.ADV be.V.3S.PRES monolingual.ADJ
pobl *MONOLINGUAL* *yn* *MINORITY* *bach* *yn* *y*
people.N.F.SG monolingual.ADJ PRT minority.N.SG small.ADJ in.PREP the.DET.DEF
byd.
world.N.M.SG

'Well, apparently monolingual people are a small minority in the world.' [stammers3:339]

Multi-word insertions are those that have longer structures of the switched language. Example (8) shows a multi-word insertion of English into a clause with a

Welsh inflected verb and (9) shows a multi-word insertion of Welsh into a clause with an English inflected verb.

(8) *dylet* *ti* *fod* *yn* *gallu* *gwrando*
ought_to.V.2S.IMPERF you.PRON.2S be.V.INFIN+SM PRT be_able.V.INFIN listen.V.INFIN
(ar)no *fe* *TOP* *TO* *BOTTOM* *AND*
on_him.PREP+PRON.M.3S he.PRON.M.3S top.N.SG to.PREP bottom.N.SG and.CONJ
ENJOY *THE* *WHOLE* *THING*
enjoy.V.INFIN the.DET.DE whole.ADJ thing.N.SG

'You should be able to listen to it top to bottom and enjoy the whole thing.' [davies9:183]

(9) *YOU* *KNOW* *DOING* *USUAL* *a* *siarad*
you.PRON.SUB.2SP know.V.2SP.PRES do.V.PRESPART usual.ADJ and.CONJ talk.V.INFIN
dros *popeth.*
over.PREP+SM everything.N.M.SG

'You know, doing the usual and talking across everything.' [davies12: 3380]

Table 5 shows that in our data the majority of code-switches were single-word insertions.

	No. of bilingual clauses	**% of bilingual clauses**
Total	6707	100
Single-Word Insertions	4772	71
Multi-Word Insertions	1935	29

Table 5 Distribution of single-word vs. multi-word insertions

In this analysis we divided our speakers into three groups: those who acquired English and Welsh simultaneously, those who acquired English first, and those who acquired Welsh first. Figure 5 shows the percentage of single-word vs. multi-word insertions produced by each group. It can be seen that single-word insertions are used more frequently than multi-word insertions by all groups, but that the speakers who learnt both English and Welsh simultaneously use more of both.

Two-tailed t-tests showed that the groups who had learnt English or Welsh first were not significantly different from each other in their production of either single-word insertions ($p=0.26$) or multi-word clauses ($p=0.94$). Furthermore, as shown in Figure 4, single-word insertions and multi-word insertions were positively correlated ($r = 0.83$, $p = <0.0001$). This means that speakers who use more single-word insertions also use more multi-word insertions. Thus in our data at least we do not yet have evidence for the pattern of bilingual acquisition affecting the size of insertions in code-switching.

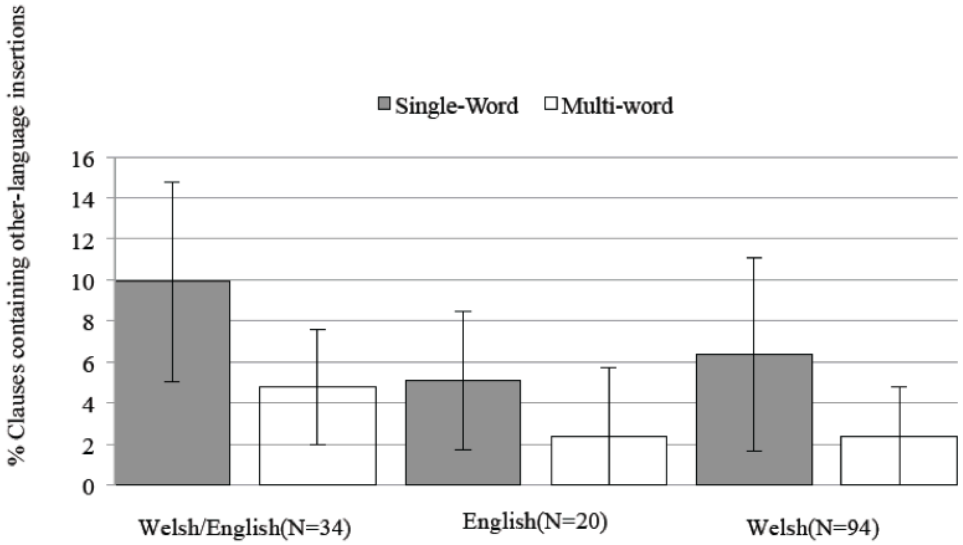

Figure 5 Single-word vs. multi-word insertions by first language acquired
First language acquired: Welsh/English simultaneously, English or Welsh

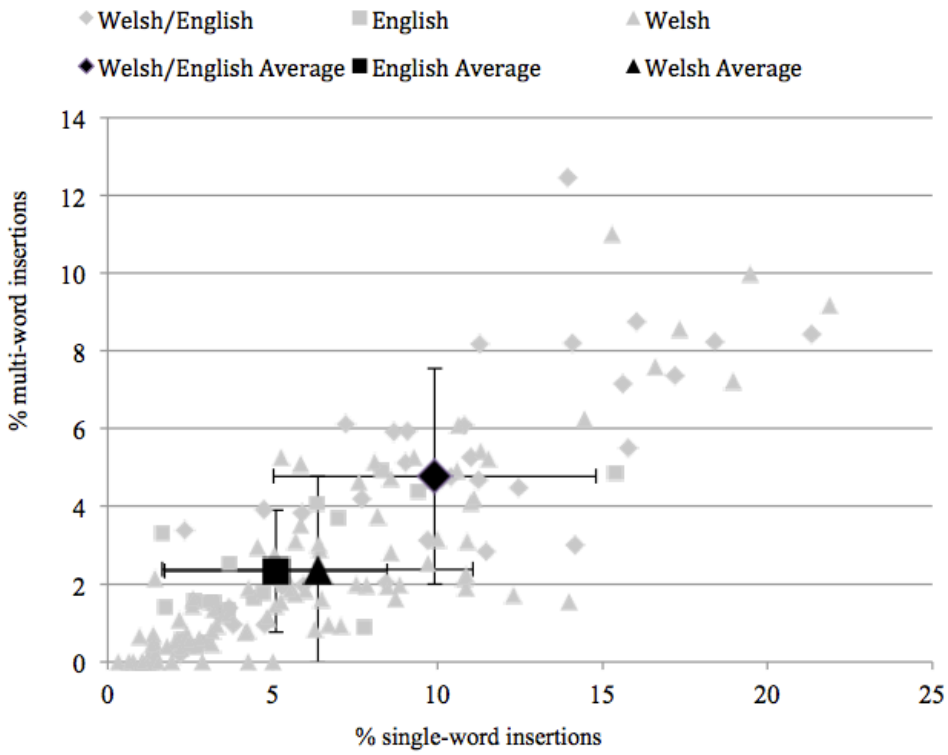

Figure 6 Correlation between single-word and multi-word insertions
Key: Language labels indicate language acquired first
Note: error bars show one standard deviation.

Although our results show a relation between simultaneous acquisition of the two languages and the more frequent production of code-switching, this does not mean that such a relationship will be found in all bilingual communities, since community norms doubtless play a role. While our own observations in Wales and the evidence of the Siarad corpus demonstrate that code-switching is a community norm in informal conversations between Welsh-English bilinguals, not all bilingual communities use code-switching. For example, it is not common in Patagonia, Argentina, where we collected a Welsh-Spanish corpus (see the Patagonia corpus at www.bangortalk.org.uk and Carter et al. 2011). However, we do predict a similar finding to ours in an analysis of our Spanish-English corpus from Miami (see www.bangortalk.org.uk and Parafita Couto et al. (2014).

Our results showed that the other important external factor was age. Figure 4 shows a negative correlation between age and code-switching, such that the older the speaker, the less frequent the proportion of bilingual clauses. The fact that there is only one speaker (aged over 70) who avoids code-switching demonstrates that despite the differences in the frequency of code-switching, it is used by participants of all ages. Nevertheless, interpreting our results within the 'apparent time' paradigm suggests that there is an ongoing change in language norms and that code-switching is becoming more common and acceptable, at least in informal contexts. This interpretation would assume no interaction between age and pattern of bilingual acquisition, such that younger speakers may be more likely to have acquired the two languages simultaneously than older speakers. We therefore checked for an interaction of this kind but found none. Another interpretation of our results would be that there is 'age-grading', or that an individual's language behaviour will change as he/she gets older. However, our observations suggest that the apparent time interpretation is the more likely. In support of this interpretation we may note an increasingly relaxed attitude to code-switching within the classroom in the literature on bilingual education both in Wales and abroad (cf. Lewis & Andrews 2014).

Our results have methodological as well as theoretical implications. Although our methods of automatic parsing and analysis can be improved further, we have shown that it is possible to extract large amounts of data with a low level of error. Using automatic glossing and data extraction methods has made it possible to deal with data from a larger number of speakers than has previously been possible in code-switching studies. This means that we can also be more confident in the validity of our results.

Conclusion

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. We speculated that this difference could be due to the timing of different patterns of bilingual acquisition in relation to neural maturation. 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. Finally, we suggest that the large size of our corpus and our automatic data extraction methods allow considerable confidence in our results.

Margaret Deuchar
Department of Theoeretical and Applied Linguistics
9 West Rd,
Cambridge CB4 1AD
md118@cam.ac.uk

References

Baayen, R. H., D. J. Davidson & D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412.

Backus, A. 1996. *Two in one. Bilingual speech of Turkish immigrants in the Netherlands.* Tilburg, the Netherlands: Tilburg University Press.

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

Baker, P. 2010. *Sociolinguistics and corpus linguistics.* Edinburgh: Edinburgh University Press.

Carter, D., M. Deuchar, P. Davies & M. C. Parafita Couto. 2011. A systematic comparison of factors affecting the choice of matrix language in three bilingual communities. *Journal of Language Contact* 4. 153–183.

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

Davies, P. & M. Deuchar. 2010. Using the matrix language frame model to measure the extent of word-order convergence in Welsh-English bilingual speech. In A. Breitbarth, C. Lucas, S. Watts & D. Willis (eds.), *Continuity and Change in Grammar*, Amsterdam and Philadelphia: John Benjamins Publishing.

Deuchar, M. 2005. Congruence and Welsh-English code-switching. *Bilingualism: Language and Cognition* 8(3). 255–269.

Deuchar, M. 2006. Welsh-English code-switching and the Matrix Language Frame model. *Lingua* 116(11). 1986–2011.

Deuchar, M. 2012. Code-switching. In C. A. Chapelle (ed.), *Encyclopedia of applied linguistics*, 657–664. New York: Wiley.

Deuchar, M. & P. Davies. 2009. Code switching and the future of the Welsh language. *International Journal of the Sociology of Language* 195. 15–38.

Donnelly, K. & M. Deuchar. 2011. Using constraint grammar in the Bangor Autoglosser to disambiguate multilingual spoken text. In E. Bick, K. Hagen, K. Müürisep & T. Trosterud (eds.), *Proceedings of the NODALIDA 2011 Work-*

shop Constraint Grammar Applications (NEALT Proceedings Series 14), 17–25 http://hdl.handle.net/10062/19298.

Drager, K. & J. Hay. 24. Exploiting random intercepts: two case studies in sociophonetics. *Language Variation and Change* 1. 59–78.

Duran Eppler, E. 2010. *Emigranto: The Syntax of German/English Code-Switching*. Vienna: Braumüller.

Finlayson, R., K. Calteaux & C. Myers-Scotton. 1998. Orderly mixing and accommodation in South African codeswitching. *Journal of Sociolinguistics* 2(3). 395–420.

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. No. 4, juin, 112–136. Paris: CERPL.

Gardner-Chloros, P. 2009. *Code-switching*. Cambridge: Cambridge University Press.

Gathercole, V. C. M. & E. M. Thomas. 2009. Bilingual first-language development: dominant language takeover, threatened minority language take-up. *Bilingualism: Language and Cognition* 12(2). 213–37.

Gries, S. Th. 2011. Introduction to this special issue. *Revista Brasileira de Linguística Aplicada* 11(2). 291–294.

Johnson, D. E. 2009. Getting off the golvarb standard: introducing rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3(1). 359–383.

Lewis, W. G. & H. Andrews. 2014. Bilingualism in Higher Education. In E. M. Thomas & I. Mennen (eds.), *Advances in the study of bilingualism*, 171–197. Bristol: Multilingual Matters.

Lloyd, S. W. 2008. *Variables that affect English language use within Welsh conversations in North Wales*. Bangor: Bangor University Unpublished Master's thesis.

MacWhinney, B. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

McEnery, T. & A. Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Mechelli, A., J. T. Crinion, U. Noppeney, J. O'Doherty, J. Ashburner, R. S. Frackowiak & C. J. Price. 2004. Neurolinguistics: structural plasticity in the bilingual brain. *Nature* 431(7010). 757–757.

Meisel, J. M. 2004. The bilingual child. In T. K. Bhatia & W. C. Ritchie (eds.), *The Handbook of Bilingualism*, 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*, 225–248. Amsterdam: John Benjamins.

Parafita Couto, M. C., P. Davies, D. Carter & M. Deuchar. 2014. Factors influencing code-switching. In E. M. Thomas & I. Mennen (eds.), *Advances in the Study of Bilingualism*, 111–138. Bristol: Multilingual Matters.

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. 1989. The care and handling of a megacorpus: The Ottawa-Hull French project. In R. Fasold & D. Schiffrin (eds.), *Language Change and Variation*, 411–

444. Amsterdam and Philadelphia: John Benjamins.
Sankoff, D. 1975. Varbrul 2. Unpublished program and documentation.
Tagliamonte, S. & R. H. Baayen. 2012. Models, forest and trees of York English: *was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.
Tagliamonte, S. A. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.
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.
Treffers-Daller, J. 1994. *Mixing two languages: French-Dutch contact in a comparative perspective*. Berlin, New York: Mouton de Gruyter.
Venables, W., D. M. Smith & the R Core Team. 2013. An introduction to R (notes on R: A programming environment for data analysis and graphics). http://cran.r-project.org/doc/manuals/r-release/R-intro.pdf.
Weber-Fox, C. M. & H. J. Neville. 1999. Functional neural subsystems are differentially affected by delays in second language immersion: ERP and behavioral evidence in bilinguals. In D. Birdsong (ed.), *Second Language Acquisition and the Critical Period Hypothesis*, 23–38. Mahwah, NJ: Lawrence Erlbaum Associates.