A Corpus-Based Analysis of Codeswitching Patterns in Bilingual Communities

Diana Carter, Peredur Davies, M^a Carmen Parafita Couto, Margaret Deuchar.

ESRC Centre for Research on Bilingualism in Theory and Practice, Bangor University

INTRODUCTION

The first aim of the current study is to compare codeswitching (CS) patterns in the bilingual speech of three communities: Miami, Patagonia, and Wales. In order to classify the CS patterns found in our data, we apply the Matrix Language Frame (MLF) model (Myers-Scotton, 1993, 2002) to all bilingual clauses contained in nine recordings of natural conversation. We selected the MLF as a means of classifying our data because the model has been tested successfully on Welsh, English and Spanish data in previous studies (Deuchar, 2006; Deuchar & Davies, 2009; Davies & Deuchar, forthcoming). Our second objective is to investigate the relationship between these CS patterns and sociodemographic community characteristics. We propose that variation found in these patterns may be linked to community-wide extralinguistic factors. We predict that the choice of matrix language (ML) will be affected by relative language proficiency levels, the language used in education, the language of their social networks and the social identity of the participants. We expect the ML to be language in which most speakers are proficient (cf. Myers-Scotton 2002: 27), and for the ML to also match the language used in education, the main language of social networks, and the language most associated with their perceptions of their own identity. For example, for the Welsh-Spanish bilinguals in Patagonia, if the language of education and their social network is Spanish, and they have a higher proficiency in Spanish than in Welsh, then we would predict that Spanish would be the preferred ML. If there is variability with respect to these factors, then we expect a mixed ML distribution in the data.

MATRIX LANGUAGE FRAME MODEL

The basic proposal of the MLF model is that in CS there is a 'base' or Matrix Language (ML), which supplies the morphosyntactic frame for the clause, and an Embedded Language (EL), which provides a proportion of the content morphemes. The ML of a clause may be identified by applying two principles: the System Morpheme Principle (SMP) and the Morpheme Order Principle (MOP). According to the SMP, the ML sources outside late system morphemes, which are morphemes that have 'grammatical relations external to their head constituent' (Myers-Scotton, 2002: 59), and thus have to look outside their maximal projection for information about their grammatical form. Finite verb morphology is an example of a late outsider morpheme, since verbal morphology is dependent on the subject constituent for information about its form. The MOP states that word order will also be sourced from the ML. Example (1) below demonstrates the application of the MLF model to a bilingual Welsh-Spanish clause (Spanish is in bold type).

(1) [[bydda i (y)n wneud biotechnoleg] achos dw i ddim yn gallu
be.FUT.1SG I PRT do.INF biotechnology because be.PRES.1SG I NEG PRT can.INF
inscribir=me yn dau]]
enrol.INF=ACC.1SG in two
‘I’ll be doing biotechnology, because I can’t enrol in both.’ (Patagonia 29)

In example (1), the finite verb *dw* is in Welsh and has subject-verb agreement with the subject pronoun *i*. As subject-verb agreement is an example of an outside late system morpheme, Welsh is identified as the ML and Spanish as the EL according to the SMP. Additionally, when applying the MOP we see that the word order is Welsh, thus supporting the claim that the ML of the clause is Welsh.

THE BILINGUAL COMMUNITIES

In order to compare CS patterns across bilingual corpora we investigated three bilingual communities: 1) Spanish-English bilinguals from Miami, Florida; 2) Welsh-Spanish bilinguals from Patagonia, Argentina; and 3) Welsh-English bilinguals from Wales. In this section we provide a brief description of the main similarities and differences between the three communities. To begin with, Wales and Patagonia have been established bilingual communities since the 19th Century, whereas Miami is a relatively young community that was established in the 1960s (Gathercole, 2007). With respect to the language families, the Spanish-English bilinguals from Miami speak a Romance language and a Germanic language, both of which have a Verb-Subject-Object (VSO) word order. The Welsh-English bilinguals speak a Celtic language (VSO) and a Germanic language (SVO), and the participants from Patagonia speak a Romance language (SVO) and a Celtic language (VSO). There are differences between the three communities concerning their levels of proficiency in the two languages. In the Wales community the bilinguals tend to have native-like proficiency in both languages but in the other two communities relative proficiency in the two languages is more variable. In the Miami and Patagonia communities there are both native-like bilinguals and speakers who are more proficient in one language than the other. This is partly due to the limited availability of bilingual education in these two communities. The bilinguals in Patagonia do not have the option of attending school in Welsh, but may attend Welsh language classes. In Miami, there have been attempts at establishing bilingual schools since the 1960s; however, in 1984 only four bilingual schools remained of the original 14 created twenty years earlier. In Wales, Welsh-medium and bilingual schools have developed since 1939, and since 1999 Welsh has been compulsory in schools.

METHOD

Participants

A similar recruitment method was used for all three corpora. Potential participants were contacted through recruitment letters and the 'friend of a friend' approach (see Milroy, 1987). There were 151 Welsh-English bilingual participants from Wales, 85 Spanish-English bilingual participants from Miami, and 92 Welsh-Spanish bilingual participants from Patagonia. The ages of the participants ranged from 9 to 92 years and included both males and females. Background information on the participants was determined by administering a questionnaire, which included questions about their age, language history, language proficiency, education, social identity (ie. Welsh, Argentinean etc.), and the language of their social network (ie. the main language spoken with family, closest friends, co-workers etc.).

Data Collection

The data for the Welsh-English corpus were collected over two years in North Wales by a research team from Bangor University. The participants were recorded having natural conversations in pairs or groups of three for approximately half an hour. In order to minimize the Observer's Paradox, the investigator was not present for the duration of the conversation (Labov, 1972). A total of forty hours of digital audio was recorded and later transcribed following the CHAT method (MacWhinney, 2000). The data for the Spanish-English and Welsh-Spanish corpora were collected and transcribed using an identical methodology. The Spanish-English corpus (forty hours) was recorded in Miami and the Welsh-Spanish corpus (twenty hours) was collected in Patagonia.

Data Analysis

For our analysis we extracted all the bilingual utterances from nine transcripts either manually or with the Computerized Language Analysis (CLAN) program (MacWhinney, 2000). The utterances were divided so that the resulting units of analysis were bilingual simple clauses. We applied the MLF framework as explained above in order to identify each clause as having either a Welsh, English or Spanish matrix language.

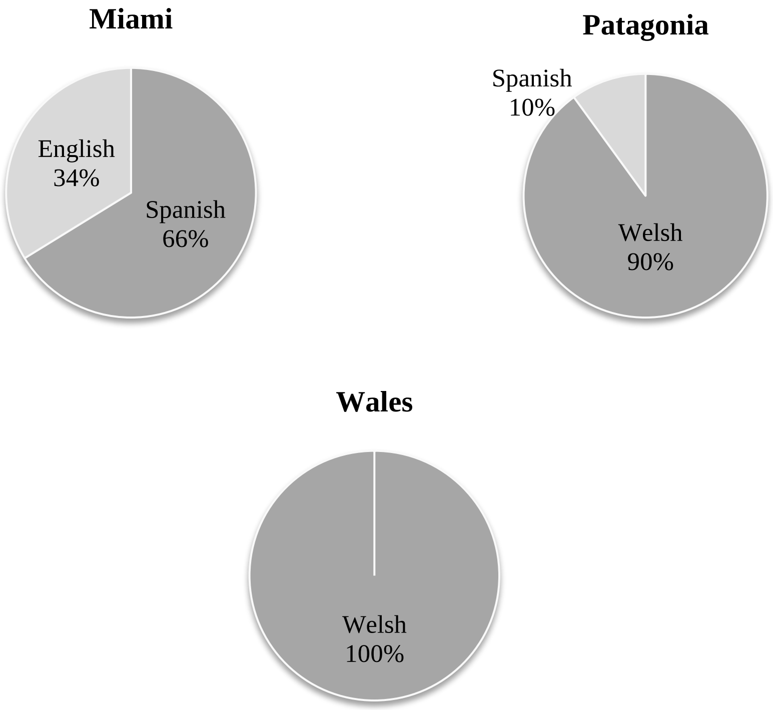
In order to examine the relationship between CS patterns and community characteristics, we analyzed the questionnaire responses from all of the participants from each corpus for the following extralinguistic variables: language proficiency, language of education, national identity, and social network. A chi-square test for independence was conducted to test for significant differences between communities.

RESULTS

Our results from the analysis of the simple bilingual clauses according to the MLF framework are as follows. 100% of the bilingual clauses from the Wales data had Welsh as the

ML (N = 347/347)[1]. The Patagonia data showed a similar trend as 90% of the bilingual simple clauses had Welsh as the ML (N = 18/20). The Miami data, on the other hand, showed more variability as 66% (N = 98/148) of the bilingual simple clauses had a Spanish ML and the remaining 34% (N = 50/148) were identified as having English as the ML. These findings are illustrated below in Figure 1.

Figure 1 MLF distribution for Wales, Patagonia and Miami



The focus of the sociodemographic analysis was on the following extralinguistic factors[2]: 1) self-rated proficiency in the majority language; 2) self-rated proficiency in the minority language; 3) language of primary school education; 4) language of secondary school education; 5) national identity; and 6) social network.

Proficiency

The analysis of the self-reported proficiency in the minority language (see Figure 2) did not show a significant difference between the Miami and Wales groups; however there was a highly significant difference between Miami and Patagonia ($\chi^2 = 22.42$, df = 3, $p < 0.0001$) and Wales and Patagonia ($\chi^2 = 44.25$, df = 3, $p < 0.0001$). The results from the analysis of the

[1] In a previous paper which presents these Welsh-English data (Davies & Deuchar forthcoming), we analyzed 336 bilingual clauses, not 347 as here. The difference between these two figures is due to a slight difference in our interpretation of a certain Welsh verb in the two studies, but the ML distribution pattern in both the analysis presented in Davies & Deuchar and in the present paper is in essence the same and should be considered homogeneous. Davies & Deuchar also discussed the data from the viewpoint of convergence, identifying certain clauses as having a "dichotomous" ML, which we do not consider here.

[2] In order to facilitate the presentation of the results, we use the term *majority language* to refer to the language spoken by the majority of the respective country, ie. English in Wales and the US, and Spanish in Argentina. The term *minority language* is used to refer to Welsh in Wales and Argentina, and Spanish in the US.

majority language (see Figure 3) revealed that there were significant differences between Wales and Patagonia ($\chi^2 = 16.75$, df = 2, $p < 0.001$) and between Wales and Miami ($\chi^2 = 8.6$, df = 2, $p < 0.05$), but not between Miami and Patagonia. Overall, there was a higher percentage of participants from Miami who reported they were *confident* in English (85%) than those who gave themselves the same rating for Spanish (74%). The participants from Patagonia also rated their proficiency level higher for the majority language, as 89% indicated they had a high proficiency level in Spanish, but only 41% reported the same high proficiency level for Welsh. In Wales we find the opposite effect. More participants have a high proficiency in Welsh (77%) than in English (69%). The higher self-reported proficiency level of the Welsh-English bilinguals in the minority language may be due to the differences we find in the primary and secondary school mediums, which are presented in the following section.

Figure 2 Proficiency in the Minority Language

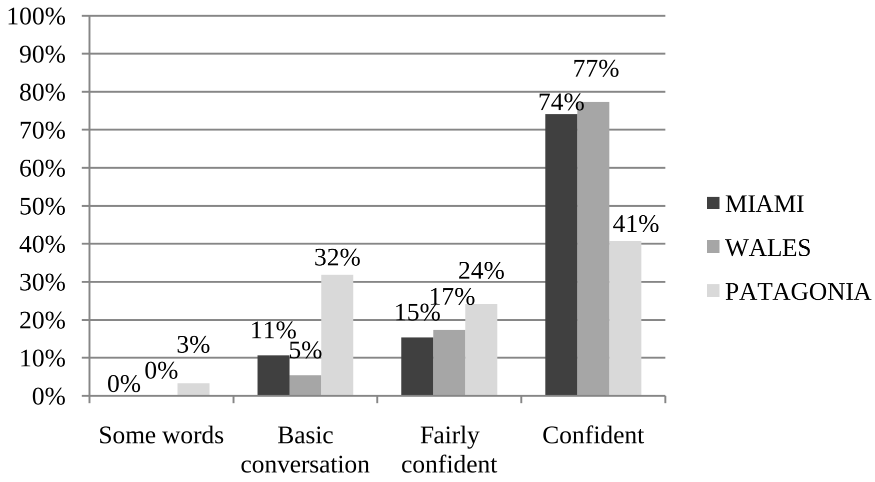
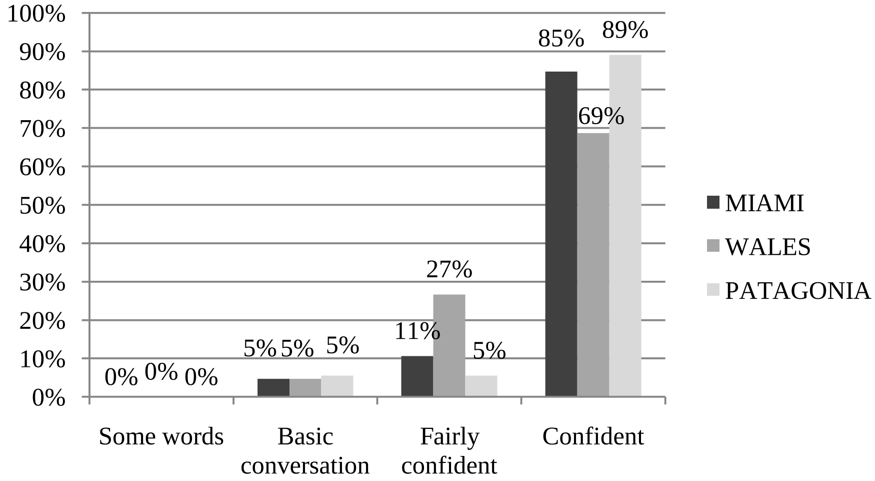


Figure 3 Proficiency in the Majority Language



Language of Education

The majority of the Wales participants (72%) had received their primary school education in Welsh and 45% had also received their secondary school education in the same language medium. In contrast, the Patagonia participants had received their primary (97%) and secondary (88%) school education in the majority language of Spanish. In Miami, over half had received their primary school (52%) and secondary school (61%) in the majority language (English). There were highly significant differences between all three communities for both the language of their primary school education ($\chi^2 = 174.65$, df = 6, $p < 0.0001$) and secondary education ($\chi^2 = 129.34$, df = 8, $p < 0.0001$). The results for these two factors are illustrated graphically in Figures 4 and 5.

Figure 4 Language of Instruction in Primary School

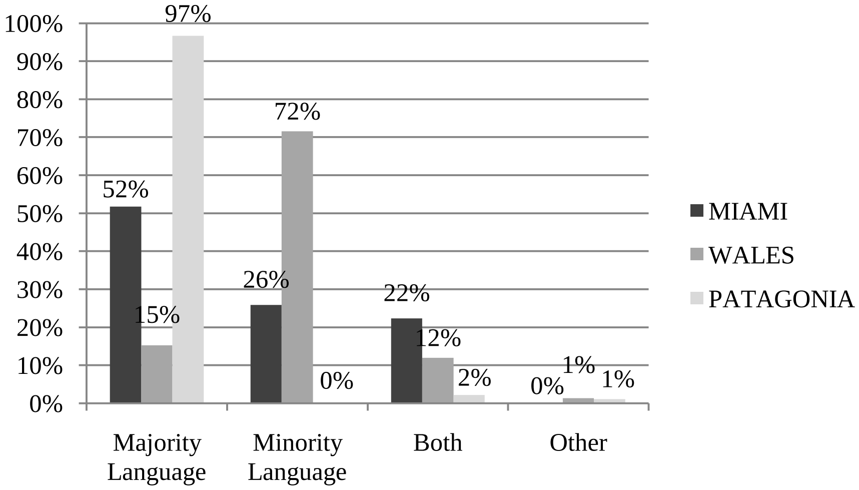
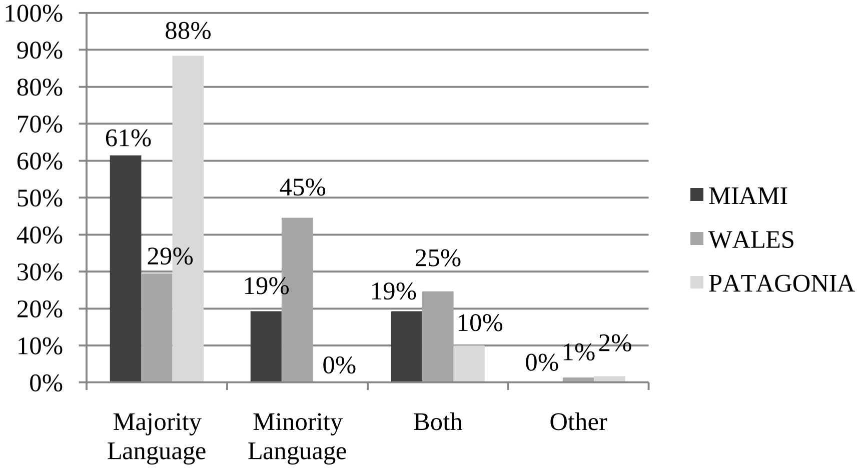


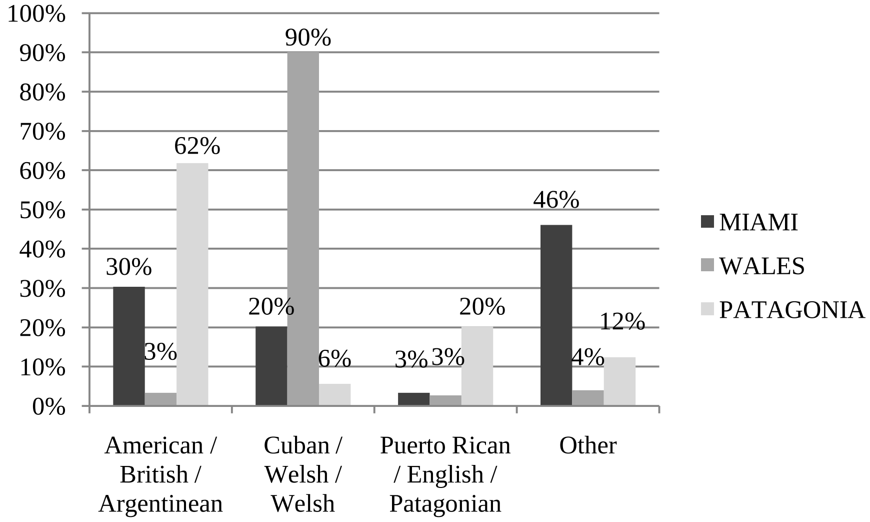
Figure 5 Language of Instruction in Secondary School



Identity

The results for the *identity* factor are presented in Figure 6. Almost all of the Welsh-English participants identity themselves as Welsh (90%). The data from the Miami group is not as homogeneous. Approximately half of the Spanish-English participants fall into the 'other' category, which includes Venezuelan, Dominican, and Cuban-American. The remaining Spanish-English participants identify themselves American (32%) and Cuban (21%). The majority of the participants from Patagonia indicated that they identify themselves as Argentinean (62%) or Patagonian (20%). The differences between the three language pairs were found to be highly significant ($\chi^2 = 166.69$, df = 6, $p < 0.0001$).

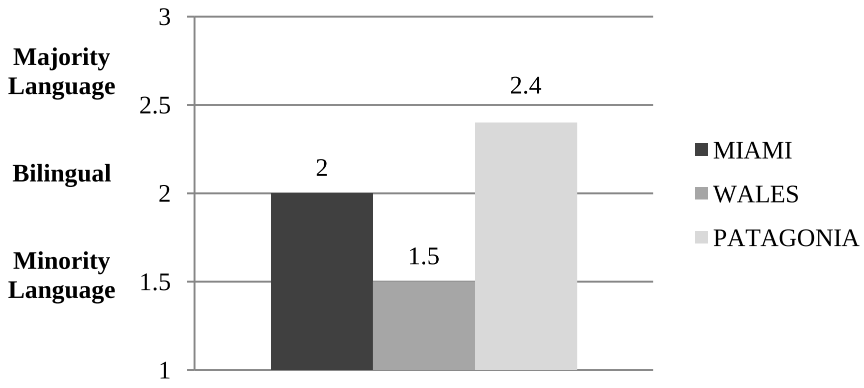
Figure 6 Identity



Social Network

In order to investigate the main social network language for each bilingual community we calculated the mean scores per participant and then from these results we calculated the mean score overall for each group (Miami, Wales, and Patagonia). The results of the social network analysis are illustrated in Figure 7. A score of 3 indicates that the main language of speakers' social networks is the majority language of the respective country, which is English in the case of Wales and Miami, and Spanish in Patagonia. A score of 2 means that speakers use both languages in their social networks, and a score of 1 indicates that the main language of the social network is the minority language (Welsh in Wales and Patagonia, and Spanish in Miami). The findings revealed that the Spanish-English participants generally have a bilingual social network (score = 2), the Welsh-English participants tend to have a more Welsh-speaking network (score = 1.5), and the Welsh-Spanish participants have a more Spanish social network (score = 2.4).

Figure 7 Social Network Mean Scores



Discussion

As outlined in the introduction, we expected the language in which speakers were most proficient to dictate their choice of ML. This was not borne out in all three communities. In Miami, participants considered themselves to speak both English and Spanish to a fairly high level, and this would seem to be reflected in their choice of both English and Spanish as the ML. In Wales, although speakers assessed their proficiency in Welsh more highly than in English on average, there was not much difference between their proficiency in the two languages, and certainly not enough to explain the virtually exclusive use of Welsh as the ML in bilingual clauses. Even more dramatically, participants in Patagonia reported being much more proficient in Spanish than Welsh, and yet in the data we analyzed, Welsh was the ML in the small proportion of bilingual clauses that we found. It seems then that there may be factors other than relative language proficiency that will determine the choice of ML. The language of education seemed to be a good predictor of the ML in Wales, where it was mostly Welsh, and in Miami, where it was both English and Spanish. In Patagonia, however, the language of education was Spanish and yet the ML was mainly Welsh. We believe that some participants spoke Welsh for the recordings in order to impress the visitors from Wales who were collecting the data, and that they may have used Spanish more in other environments. It could also be the case that Welsh-Spanish code-switching is not a common practice in the Patagonia community, compared with the communities in Miami and Wales. There was a large difference between the number of bilingual clauses found in the Patagonia data compared to the Wales data. For the Patagonia dataset, there were 20 bilingual simple clauses out of a total of 1099 clauses (1.8%), whereas for the Wales dataset, there were 347 bilingual simple clauses out of a total of 1862 clauses (18.6%).

There was uniformity in the results for identity and social network in Wales but variation in the results found in Miami. The uniformity of these extralinguistic variables in Wales seems to be related to the uniformity of ML of clauses produced by speakers from Wales, while in turn the diversity of the extralinguistic variables found in Miami seems to be related to the variation in ML of clauses produced by speakers from Miami. The Patagonia data, however, did not show such a correlation: the speakers from Patagonia expressed uniformly Argentinean identity and a predominantly Spanish social network, from which we would predict that they would show preference for Spanish ML if the Wales and Miami data are used as a model, but in fact the bilingual clauses produced by the Patagonia speakers were uniformly Welsh. Needless to say, as we have pointed out above, there are doubts about how 'natural' the Patagonian data were.

In both Patagonia and Wales the predominant ML is Welsh. Welsh is a VSO language, whereas Spanish and English are SVO. As suggested by Chan (2009), there may be a universal tendency to select one ML, unless the two languages concerned have similar word order (as is the case in the Miami data, where both English and Spanish are SVO). If so, this would explain why there is more variation in the choice of the ML in the Miami data than in the Wales and Patagonia data, since in those bilingual communities one of the languages is SVO but the other is VSO.

CONCLUSION

Our predictions about the role of proficiency in the choice of the ML were not borne out. However, the languages of education, of identity and of the social network were all shown to be related to the choice of the ML in Wales and Miami. In Patagonia these factors would have led us to expect Spanish rather than Welsh to be the predominant ML, but there may have been methodological reasons for the choice of Welsh. The uniformity of the ML was nevertheless in line with a structural explanation, as in Wales. Further research is needed to discover the relative role of structural versus extralinguistic factors in determining the choice of ML.

ACKNOWLEDGEMENTS

The support of the Arts and Humanities Research Council (AHRC), the Economic and Social Research Council (ESRC), the Higher Education Funding Council for Wales and the Welsh Assembly Government is gratefully acknowledged. The work presented in this paper was part of the programme of the ESRC Centre for Research on Bilingualism in Theory and Practice at Bangor University. We are indebted to all the transcribers for the three corpora, as well as to Lofti Ayali, Rocío Pérez-Tattam, Alberto Rosignoli, Carmen Silva-Corvalán, Gerald Stell, Jorge Valdés Kroff, Myfyr Prys, Kevin Donnelly, Adrian Luescher, Eva Eppler, and the audience of the 7th International Symposium of the Spanish Linguistics Society (Santiago de Compostela, February 2010) and the Welsh Syntax Seminar (Gregynog, July 2010) for useful discussion and invaluable input.

REFERENCES

Breitbarth, Anne, Christopher Lucas, Shiela Watts & David Willis (eds.) (forthcoming): *Continuity and Change in Grammar*. Amsterdam: John Benjamins.

Bullock, Barbara & Almeida Toribio (eds.) (2009): *Linguistic code-switching*. Cambridge: Cambridge University Press.

Chan, Brian (2009): “Code-switching between typologically distinct languages”. In Bullock *et al.* (2009).

Davies, Peredur & Margaret Deuchar (forthcoming): “Using the Matrix Language Frame model to identify word order convergence in Welsh-English bilingual speech”. In Breitbarth *et al*. (forthcoming: 77-96).

Deuchar, Margaret (2006): “Welsh-English codeswitching and the Matrix Language Frame model”. Lingua 116(11): 1986-2011.

Deuchar, Margaret & Peredur Davies (2009): “Code-switching and the future of Welsh”. International Journal of the Sociology of Language 195: 15-38.

Gathercole, Virginia (2007): “Miami and North Wales, So Far and Yet So Near: A Constructivist Account of Morphosyntactic Development in Bilingual Children”. The International Journal of Bilingual Education and Bilingualism 10(3): 224-247.

Labov, William (1972): “Some principles of linguistic methodology”. Language and Society 1(1): 97-120.

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

Milroy, Lesley (1987): *Language and Social Networks*. Oxford: Blackwell Publishing.

Myers-Scotton, Carol (1993): *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford: Oxford University Press.

Myers-Scotton, Carol (2002). *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford: Oxford University Press.